

Introduction au Traitement Automatique des Langues

- 1 - Introduction Générale**
- 2 - Les applications du TAL**

Introduction au traitement automatique des langues

Objectifs de la matière

Le traitement automatique des langues (TAL) vise l'élaboration d'outils et de méthodes capables d'appréhender leur sémantique afin d'en faciliter la prise de connaissance et plus généralement l'exploitation. Selon l'usage que l'on veut en faire, les niveaux d'interprétation peuvent être différents, allant de l'identification de termes pour extraire des mots-clés à des résumés, des traductions ou de la recherche d'informations précises en réponse à des questions. L'objectif de ce module est de présenter les problématiques posées pour le TAL et les principaux modèles pour analyser, synthétiser, exploiter et produire des documents.

Pré requis recommandés : programmation, Python, etc.

Unité d'enseignement : UM21

Crédit : 4

Coefficient : 2

Mode d'évaluation : Examen, CC TP.

Introduction au traitement automatique des langues

Contenu de la matière :

- 1) Introduction Générale
- 2) Les applications du TAL
- 3) Les niveaux de traitement - Traitements de «bas niveau»
- 4) Les niveaux de traitement - Le niveau lexical
- 5) Les niveaux de traitement - Le niveau syntaxique
- 6) Les niveaux de traitement - Le niveau sémantique
- 7) Les niveaux de traitement - Le niveau pragmatique

Interactions Homme-Machine

Contenu de la matière :

TP : Python

- Python for NLP
- Librairies : nltk, SpaCy, etc.



HUGGING FACE

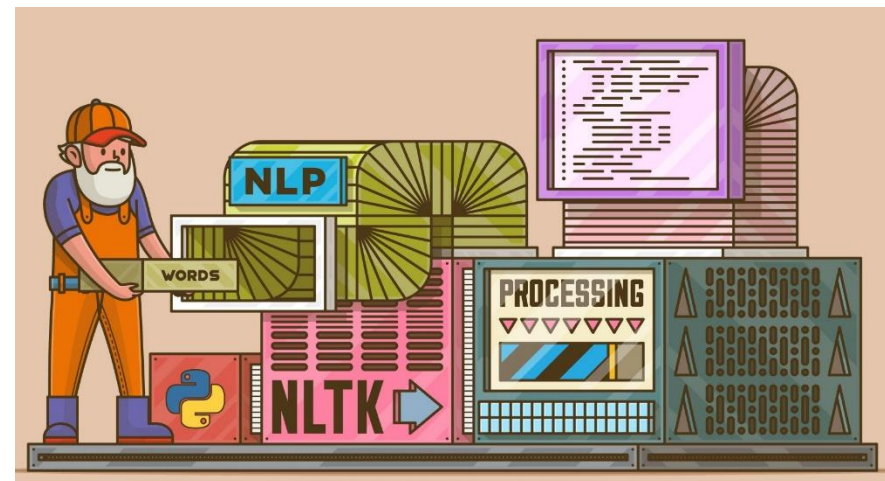


GET STARTED WITH



FOR TEXT MINING (NLP)

spaCy



Plan du cours

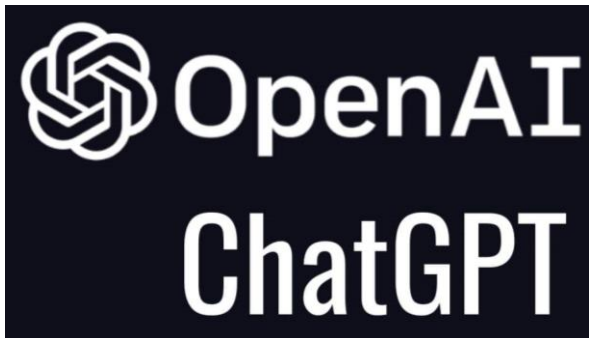
1. Définitions et généralités
2. Motivations du TAL
3. Brève histoire du TAL
4. Les difficultés du TAL : ambiguïté et implicite
5. Les applications du TAL & exemples
6. Domaines / Topics
7. Outils du TAL – Représentation de connaissances
8. Méthodes et techniques du TAL

Définitions

ChatGPT



Définitions



HE

That's great. How can I run this template via AWS CLI?



To run this CloudFormation template using the AWS CLI, you can use the ``aws cloudformation create-stack`` command. This command takes the name of your stack as a parameter, as well as the path to a file containing the CloudFormation template.

Here's an example of how you might run this template using the AWS CLI:

Copy code

```
aws cloudformation create-stack \  
  --stack-name chatgpt-demo-stack \  
  --template-body file:///path/to/template.yml
```

This command will create a new CloudFormation stack with the name ``chatgpt-demo-stack``, using the template defined in the file at ``/path/to/template.yml``.

You can also use the ``aws cloudformation validate-template`` command to validate the template before creating the stack, to ensure that it is well-formed and does not contain any syntax errors.

For more information about using the AWS CLI to manage CloudFormation stacks, see the AWS documentation: <https://docs.aws.amazon.com/cli/latest/reference/cloudformation/index.html>.

Définitions

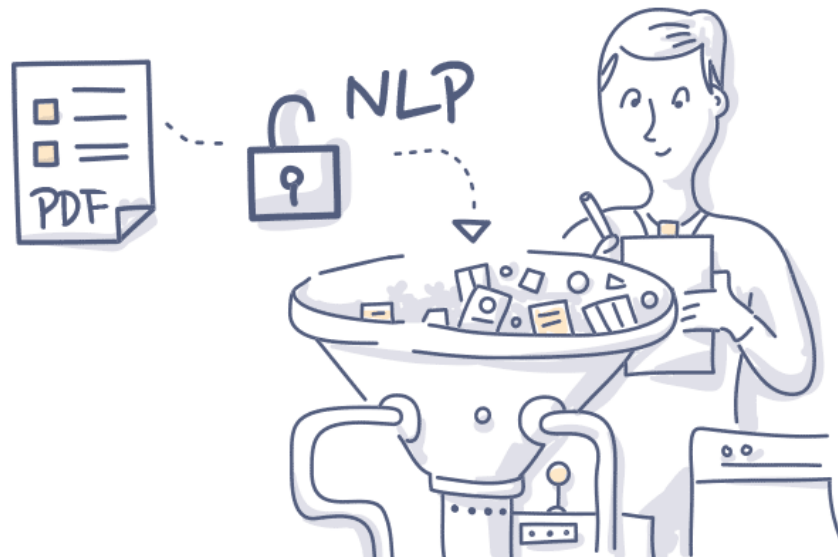
- Le traitement automatique **du langage naturel** ou de la **langue naturelle** (abr. **TALN**) ou des **langues** (abr. **TAL**)
- Ang: Natural Language Processing (abr. **NLP**)
- Ainsi, le TAL ou TALN est parfois nommé **ingénierie linguistique**.
- est une discipline à la frontière de la **linguistique**, de **l'informatique** et de **l'intelligence artificielle**, qui concerne l'application de programmes et techniques informatiques à tous les aspects du **langage humain**.
- 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**.
- Traitement du langage naturel et non pas formel (C, Java, etc).
- Traitement de la langue sous forme **écrite (texte)**, le traitement de la **parole**.

Définitions

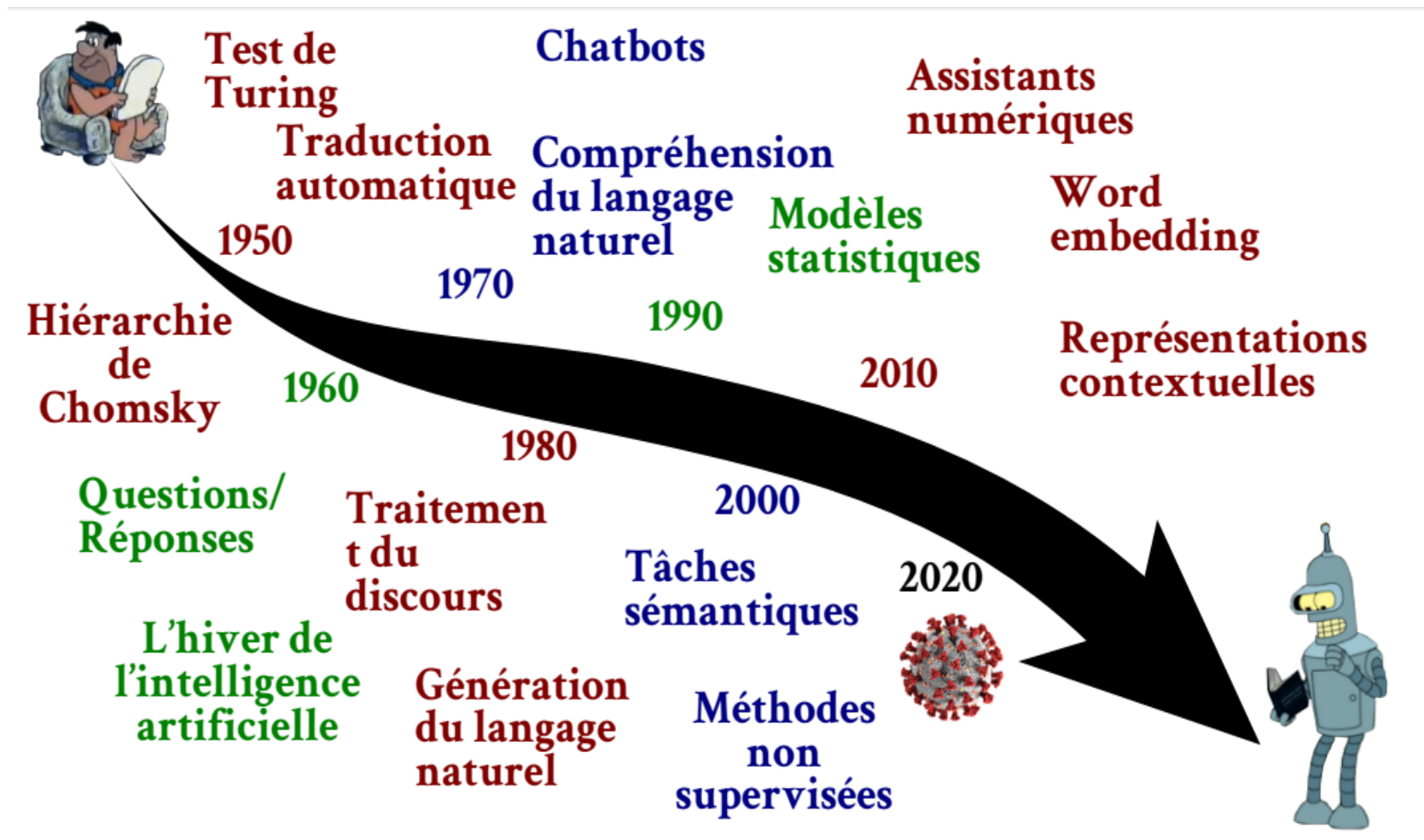
- Le traitement automatique **du langage naturel** ou de la **langue naturelle** (abr. **TALN**) ou des **langues** (abr. **TAL**)
- Traitement de la langue sous forme **écrite (texte)**, le traitement de la **parole**.
- Les données **textuelles** à traiter se déclinent à l'aune des 3 V (variété, volume, vélocité) – Big Data.
- Elles consistent en des documents écrits, **pages Web**, **emails** et autres textes « traditionnels », mais également en contenus de **blogs**, de **réseaux sociaux**, en **sms**, en **documents audio transcrits** automatiquement, ce qui correspond donc à des types et des qualités de langue très divers.
- Données **non structurées** ou **semi structurées**.
- Ces volumes énormes de données textuelles et leurs variétés ont accru le besoin au traitement automatique **de la langue**.

Motivations du TALN

- **Pourquoi** s'intéresser à l'automatisation du traitement du langage naturel ?
 1. La volonté de **modéliser** une compétence fascinante (le **langage**), afin de tester des hypothèses sur les mécanismes de la communication humaine. (**Académique**)
 2. Le besoin de **mettre en œuvre des applications** capables de traiter efficacement des quantités considérables d'informations « naturelles » aujourd'hui disponibles sous forme électronique. (**Pratique**)



Brève histoire du TALN



Brève histoire du TALN

1954 - **Traduction automatique**, la mise au point du premier traducteur automatique basique, Georgetown-IBM. Quelques phrases russes, sélectionnées à l'avance, furent traduites automatiquement en anglais.

1962 - Première conférence sur la **traduction automatique** est organisée au MIT par Y. Bar-Hillel. « le problème de la traduction automatique est probablement insoluble. »

Entre 1951 et 1954 - Zellig Harris publie ses travaux les plus importants de **linguistique** (linguistique distributionnaliste).

1957 - N. Chomsky qui publie ses premiers travaux importants sur la **syntaxe** des langues naturelles, et sur les relations entre **grammaires formelles et grammaires naturelles**.

1960's - Les étudiants de Marvin Minsky développent divers systèmes (BASEBALL (1961), SIR (1964), STUDENT (1964), ELIZA (1966) ...) mettant en œuvre des **mécanismes** de traitement simples, à base de **mots-clés**. **ELIZA**, qui simule un **dialogue** entre un psychiatre et son patient.

Brève histoire du TALN

1970's - voient le développement d'approches surtout **sémantiques** (Roger Schank, Yorick Wilks, ...) le rôle de la syntaxe étant pratiquement omis ou, tout du moins considéré comme secondaire.

L'importance du **contexte**, de représentation des **connaissances**, et le rôle essentiel d'une bonne connaissance du **domaine** traité pour comprendre un texte est ainsi mis en avant.

Les recherches ont cessé de se limiter à l'interprétation de phrases seules pour aborder le **traitement d'unités plus importantes** : les **récits** et les **dialogues**.

1971 - Winograd (MIT) a développé **SHRDLU**, un programme de compréhension du langage naturel.

1972 - Colby (Stanford) a créé **PARRY** un **chatbot** qui simule une personne avec la schizophrénie paranoïde.

1975 - **DRAGON**, un système pour la **reconnaissance automatique de la parole** en utilisant les modèles de Markov cachés.

Brève histoire du TALN

1980's - les propositions issues des **statistiques** et **l'intelligence artificielle**.

Une partie importante des travaux vise à analyser et à formaliser des mécanismes **d'acquisition automatique des connaissances**, qui permettent d'extraire directement de lexiques ou de corpus de documents, des règles de grammaire, ou encore des connaissances sémantiques.

1980 - **KL-One**, représentation de connaissance pour le traitement de la syntaxe et la sémantique.

1987 - **MUC** conférence sur l'extraction des données financée par DARPA.

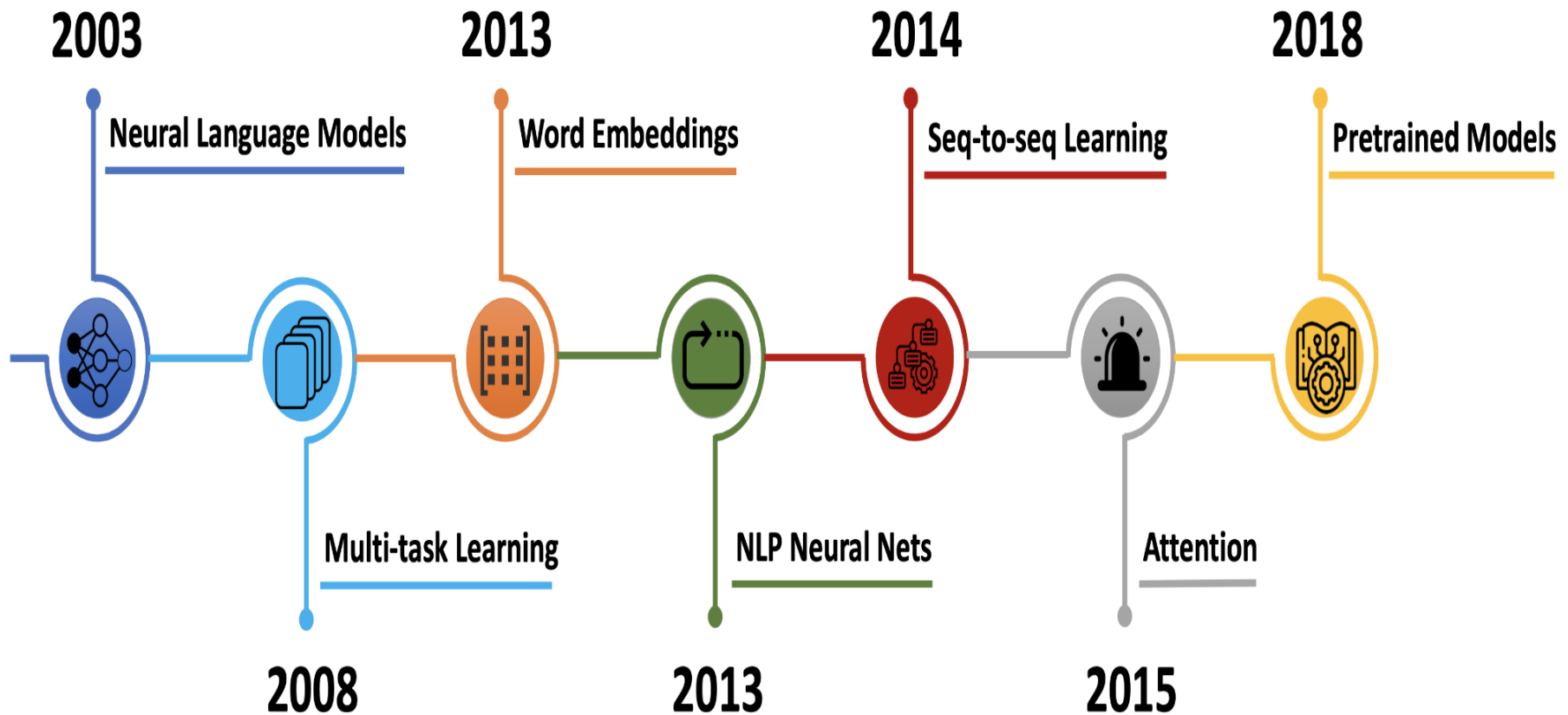
1990 – Une approche statistique pour la traduction automatique de **Brown**.

1993 - **Pen Treebank**, un corpus annoté de l'anglais.

1995 - **Wordnet**, une base lexicale pour l'anglais.

Brève histoire du TALN

Histoire récente : Méthodes et techniques IA – **Neural NLP**



ULMfit (fast.ai), ELMO (AllenNLP), GPT (OpenAI), BERT (Google), XLM (Facebook)

Les difficultés du TALN : ambiguïté et implicite

Les difficultés sont de deux ordres : **ambiguïté** du langage, et quantité d'**implicite** contenue dans les communications naturelles.

1 - Ambiguïté :

- Le langage naturel est ambigu, à tous ses niveaux (lexical, syntaxique, sémantique, contextuel, etc.)
- Elle se manifeste par la **multitude d'interprétations** possibles pour chacune des **entités linguistiques** pertinentes pour un niveau de traitement.
- *Exemple* : lexical sémantique - déterminer si le mot « **avocat** » se rapporte au domaine juridique ou au domaine alimentaire ?



Les difficultés du TALN : ambiguïté et implicite

Les difficultés sont de deux ordres : **ambiguïté** du langage, et quantité d'**implicite** contenue dans les communications naturelles.

2 - Implicite:

- L'activité langagière s'inscrit toujours dans un **contexte d'interaction** entre deux humains, dotés d'une **connaissance du monde** et de son fonctionnement.
- La machine ne dispose pas de cette **connaissance d'arrière-plan**, ce qui rend la compréhension complète de la majorité des énoncés difficile, voire impossible.
- Nécessité de disposer de bases de connaissance additionnelles, donnant accès à la fois à un savoir sur le monde (ou le domaine) en général (**connaissance statique**) et sur le contexte de l'énonciation (**connaissance dynamique**).

Les difficultés du TALN : ambiguïté et implicite

Les difficultés sont de deux ordres : **ambiguïté** du langage, et quantité d'**implicite** contenue dans les communications naturelles.

2 - Implicite:

- *Exemple* : **coréférence**, désambiguïstation du référent du pronom personnel **il** :

- L'étudiant a éteint son smartphone parce qu'**il**

... **il** était déchargé

... **il** voulait révisé

- Figures de style : paradoxes, métaphores, oxymores, antithèses, etc.
Exemple : Briser la glace.

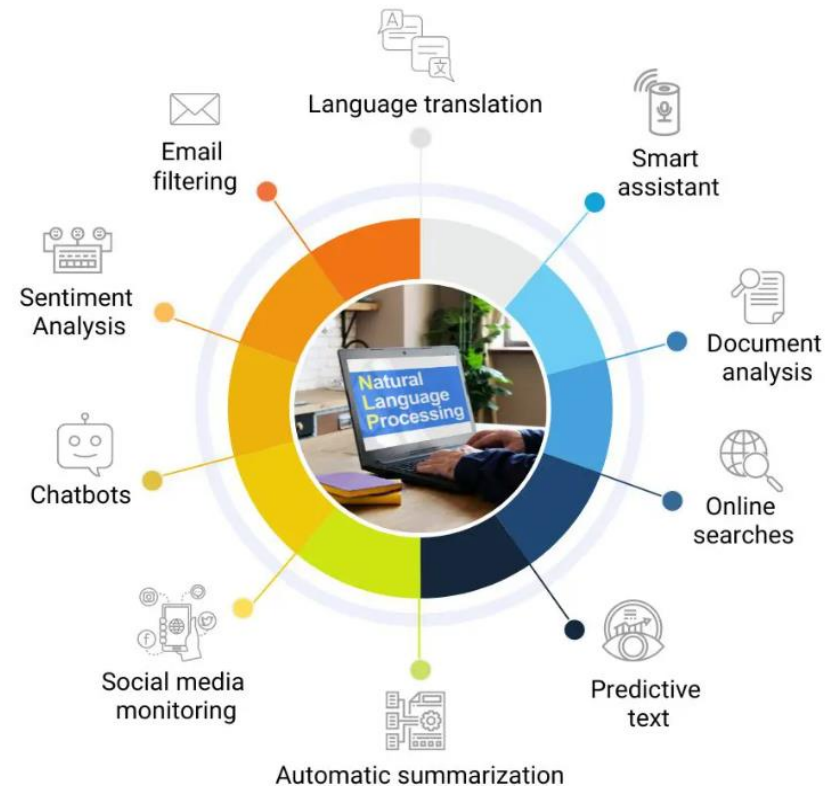
Les difficultés du TALN

- Manque de **ressources**, **outils** et **datasets** pour les langues moins parlées.
- Annotation manuelle des **corpus** d'entraînement et de test.
- Traitement des documents larges : l'apprentissage automatique est limité lorsqu'il s'agit de représenter des contextes longs.
- **Variations** et communication non standard : les utilisateurs ne respectent pas les standards d'écriture d'une langue. Ex. langue de chat, arabizi, franglais, etc.
- **Intention** : il existe des phrases qui veulent pas dire ce qu'on comprend directement ; elles veulent une autre chose.
- **Emotions** : les phrases peuvent changer de sens selon les émotions accompagnées.
- **Evaluation** manuelle et automatique.

Les applications du TALN

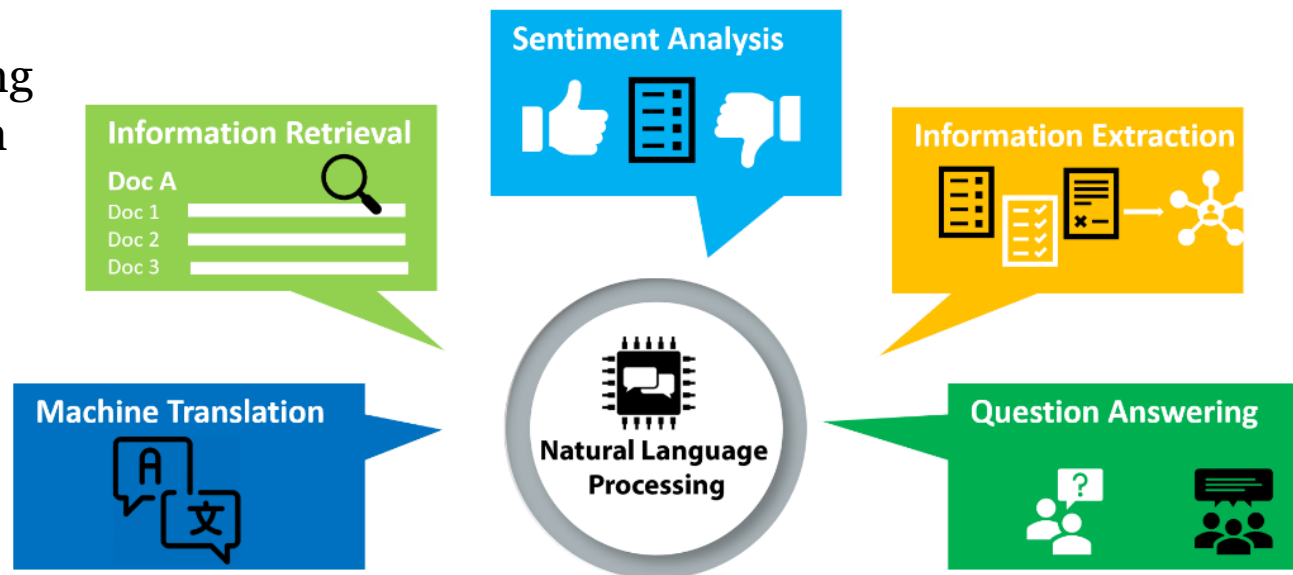


Applications of Natural Language Processing



Domaines / Topics du TALN

- Machine Translation
- Information Retrieval
- Information Extraction & Named entity recognition
- Text Categorization
- Text Mining
- Summarization
- Sentiment Analysis and Opinion Mining
- Text and Speech Classification
- Topic Modeling
- Question Answering
- Speech recognition
- Dialogue systems
- Etc.



Les applications du TALN

Les applications du TALN peuvent être regroupé en **trois** grandes familles :

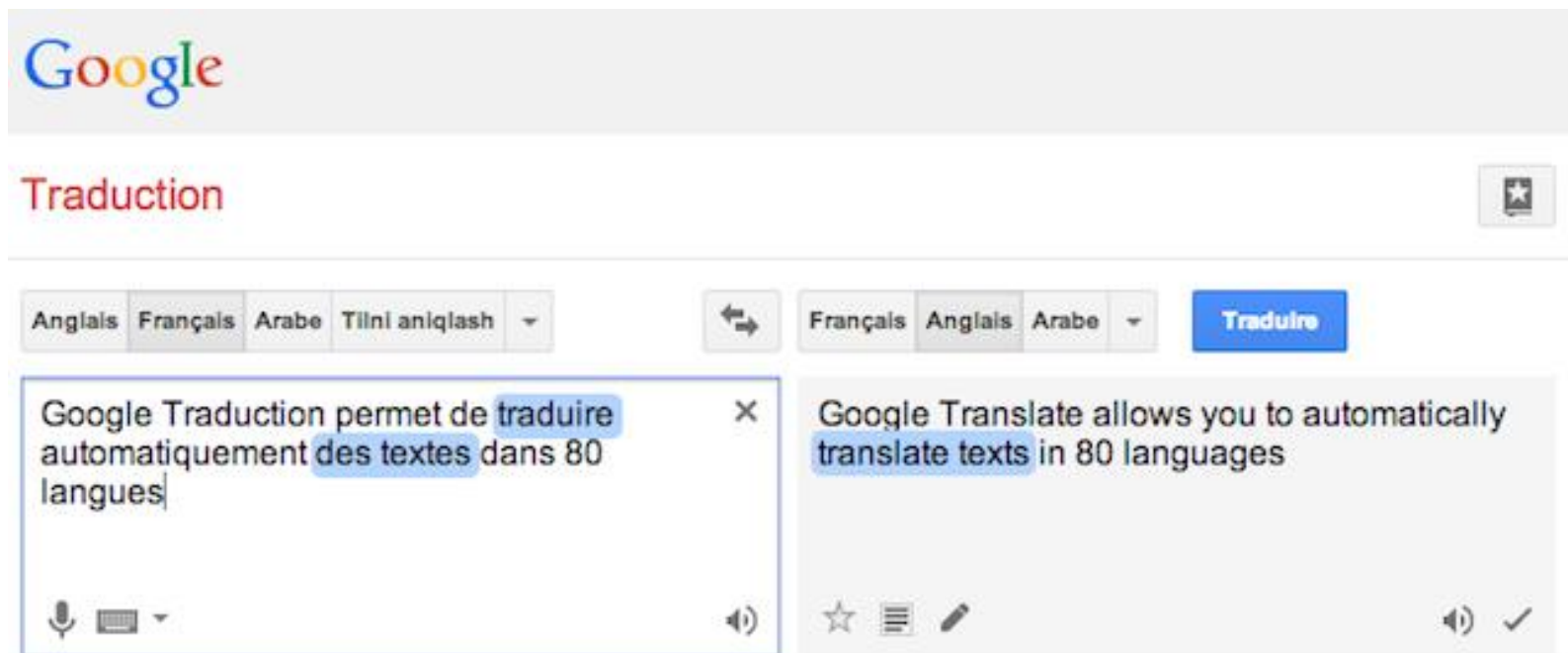
1 - Le **traitement** documentaire:

- Les applications qui visent à faciliter le traitement par l'humain des immenses ressources disponibles en langage naturel.
- Exemples:
 - La traduction automatique.
 - La recherche de documents « intéressants » dans des bases documentaires.
 - Le résumé automatique de texte.
 - Le routage, classement ou l'indexation automatique de e-documents.
 - Questions/Réponses - Correction automatique de réponses écrites.
 - La lecture automatisée de documents.
 - L'analyse d'un corpus de documents relatifs à un thème donné.
 - Des thésauri décrivant des relations entre concepts, WordNet

Les applications du TALN

1 - Le traitement documentaire:

- La traduction automatique : **Google Traduction**



Les applications du TALN

1 - Le traitement documentaire:

- Le résumé automatique de texte: **Resoomer** - <https://resoomer.com/>



The screenshot displays the Resoomer website's user interface. At the top, the 'RESOOMER' logo is on the left, and a navigation menu with links for 'Service', 'Extensions', 'Comment ça marche', 'PREMIUM', 'Contact', and 'Connexion' is on the right. A language dropdown menu is set to 'Français'. Below the navigation bar, a light blue banner contains the text: 'Allez à l'essentiel dans vos textes, résumez « pertinemment » en 1 clic'. The main content area features a light blue box with a white background. Inside this box, there are three buttons at the top: 'Exemple de texte', 'Effacer le texte', and a red 'Resoomer' button. Below these buttons, a small note reads 'Uniquement textes argumentatifs'. A large text input area follows, with the placeholder text 'Copiez-collez ici votre texte argumentatif ou l'URL de votre article'. A small icon of a notepad and pencil is visible in the bottom right corner of the input area.

Les applications du TALN

1 - Le traitement documentaire:

- Questions/Réponses - Correction automatique de réponses écrites.

QA KiS Question Answering
wiKiframework-based
System

What is largest city in USA ? [get answers](#)

DBpedia to query :

[DBpedia FR examples](#) [DBpedia EN examples](#) [DBpedia IT examples](#) [DBpedia DE examples](#)

[Results](#) [Technical details](#) [Reconciliation](#)



New York City

[DBpedia](#)
[\[more details\]](#)

Les applications du TALN

Les applications du TAL peuvent être regroupé en **trois** grandes familles :

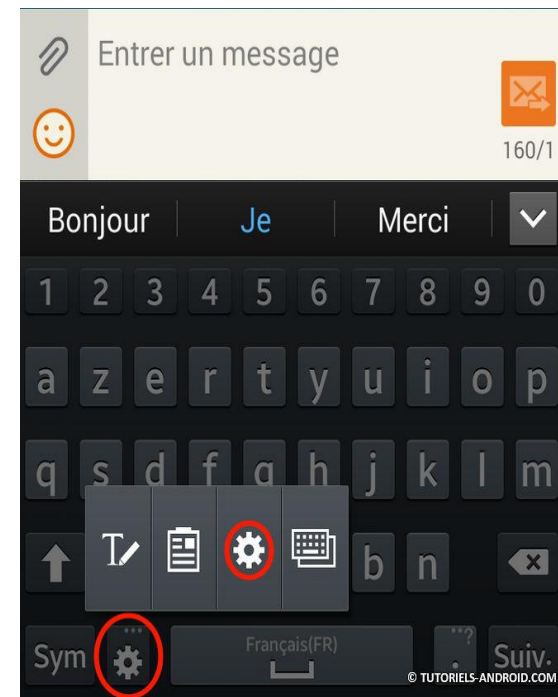
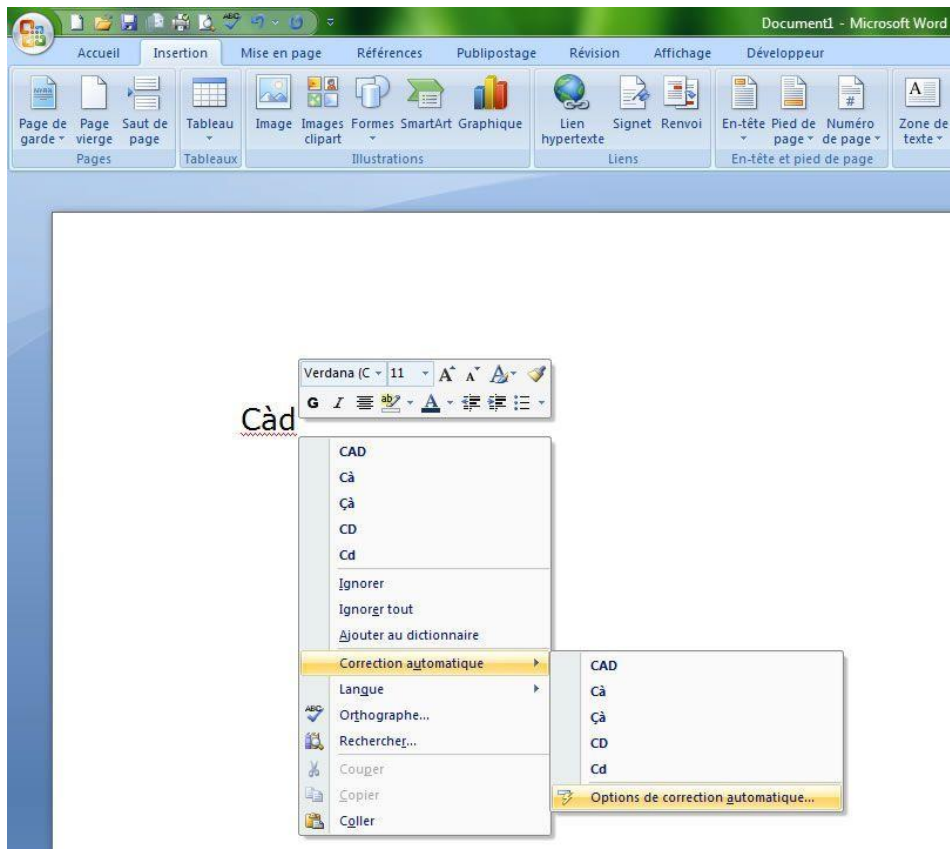
2 - La **production** de documents:

- Les applications du domaine de l'aide à la production de texte (la génération de textes), gestion de documents.
- Exemples:
 - Correcteurs automatiques,
 - Le résumé automatique de texte.
 - La reconnaissance optique de caractères,
 - Les correcteurs d'orthographe ou de syntaxe,
 - Les correcteurs « stylistiques » ou les aides intelligentes à la rédaction,
 - L'apprentissage assisté par ordinateur des langues naturelles,
 - La génération automatique de documents à partir de spécifications formelles.

Les applications du TALN

2 - La production documentaire:

- Correcteurs automatiques.



Les applications du TALN

Les applications du TAL peuvent être regroupé en **trois** grandes familles :

3 - Les **interfaces** naturelles, IHM:

- Les applications du domaine des interfaces naturelles (i.e. en langage naturel)
- Exemples:
 - Interrogation en langage naturel de bases de données ou de moteurs de recherche.
 - Les interfaces vocales, les applications des modules de reconnaissance de parole, synthèse de parole, génération et gestion de dialogue, accès aux bases de connaissance, etc.

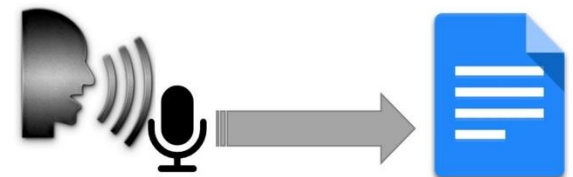
Les applications du TALN

3 - Les interfaces naturelles, IHM : Les interfaces vocales, les applications des modules de reconnaissance de parole, synthèse de parole, etc.



The screenshot shows the article page for "Intelligent Fake News Detection: A Systematic Mapping" in the Journal of Applied Security Research. A red circle highlights the audio player controls, which include a "Listen" button, a play/pause button, a stop button, and various navigation icons. The article title is "Intelligent Fake News Detection: A Systematic Mapping" by Caio V. Meneses Silva, Raphael Silva Fontes, and Methanias Colaço Júnior. The page also displays a "Free access" badge and a "Full Article" button.

DICTEE VOCALE GRATUITE



Les applications du TALN

3 - Les interfaces naturelles, IHM : Les interfaces vocales, les applications des modules de reconnaissance de parole, synthèse de parole, etc.

The screenshot shows the Cleverbot website. At the top is a blue navigation bar with links: [apps](#), [more](#), [people](#), [search](#), [snips](#), [conversations](#), [tweaks](#), and [sign in](#). The Cleverbot logo is centered below the bar. A text input field contains the placeholder text "dire à cleverbot" and a microphone icon. Below the input field are three buttons: "think about it", "think for me", and "thoughts so far". A status bar below these buttons shows "it doesn't understand" in green, "you do" in yellow, and "see what you agreed" in green. Below this is a "Sponsored Links" section. A large banner for "VERSALITY A SCI-FI NOVEL BY ROLLO CARPENTER" is displayed, featuring a background of colorful isometric cubes. In the bottom left corner of the banner, it says "help support Cleverbot". At the bottom of the page is a row of social media and utility links: [contact](#), [app](#), [api](#), [snips](#), [tweets](#), [clevs](#), [yt](#), [fb](#), and [insta](#). A small text element "iavasprintnoteshow()" is visible in the bottom left corner of the page.

Les applications du TALN

memerBot: Towards Automatic Image Meme Generation [Sadasivam et al., 2020]

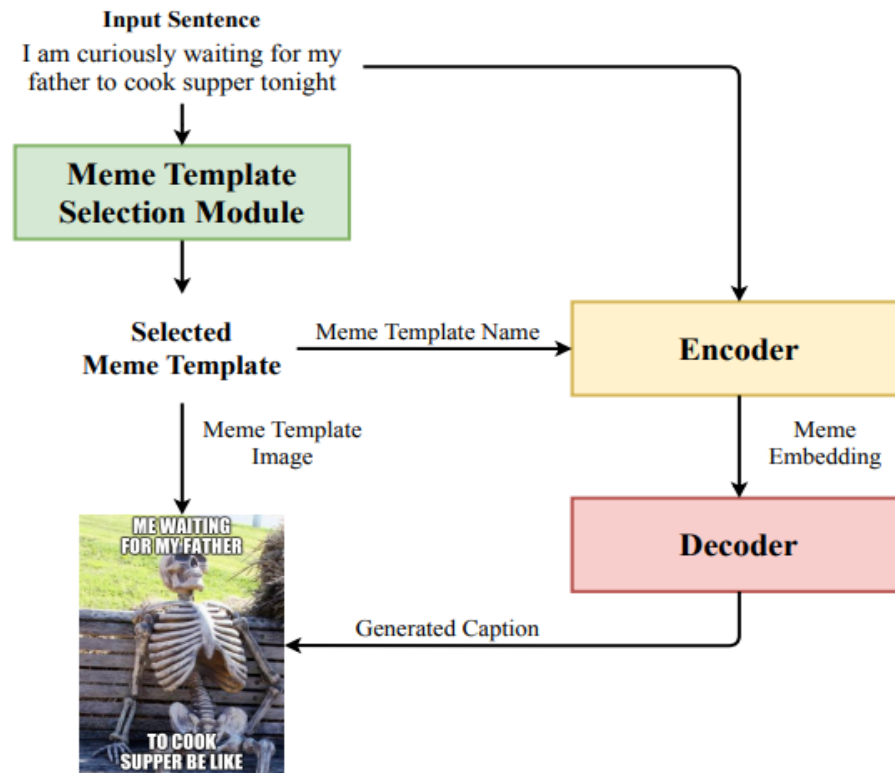


Figure 1: An illustrative figure of **memerBot**. It generates an image meme for a given input sentence by combining the selected meme template image and the generated meme caption.

Les applications du TALN



MusicLM: Generating Music From Text

| paper | dataset |

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, Christian Frank

Google Research

Abstract We introduce MusicLM, a model generating high-fidelity music from text descriptions such as "*a calming violin melody backed by a distorted guitar riff*". MusicLM casts the process of conditional music generation as a hierarchical sequence-to-sequence modeling task, and it generates music at 24 kHz that remains consistent over several minutes. Our experiments show that MusicLM outperforms previous systems both in audio quality and adherence to the text description. Moreover, we demonstrate that MusicLM can be conditioned on both text and a melody in that it can transform whistled and hummed melodies according to the style described in a text caption. To support future research, we publicly release MusicCaps, a dataset composed of 5.5k music-text pairs, with rich text descriptions provided by human experts.

Caption

The main soundtrack of an arcade game. It is fast-paced and upbeat, with a catchy electric guitar riff. The music is repetitive and easy to remember, but with unexpected sounds, like cymbal crashes or drum rolls.

Generated audio

▶ 0:00 / 0:30 — 🔊 ⋮

<https://google-research.github.io/seanet/musiclm/examples/>

Outils du TALN - Représentation de connaissances

Lexique – Lexicon

- Le lexique d'une langue est l'ensemble de ses **mots** (somme des vocabulaires utilisés), ou de façon plus précise en linguistique de ses **lemmes**.
- Dictionnaire - Vocabulaire** succinct (d'une langue, d'une science, d'un art, etc.).
- Lexique, **avec définitions**, restreint à un **domaine particulier** => **Glossaire - Glossary**

Lexique : « L'alimentation »					
Lexique : Les aliments					
Les légumes	Des haricots <u>beurre</u>	Une boisson	Du poisson	Les fruits	Les condiments
Un <u>plat</u>	Un <u>chou</u> de Bruxelles	De l'eau	Du thon		L'huile
De la soupe	Une asperge	Du thé	Des crevettes	L'abricot	Le vinaigre
Du <u>gratin</u>	Un artichaut	Du café		L'amande	La moutarde
Une courgette	Du brocoli	Du soda		La cerise	La mayonnaise
Une aubergine	Une citrouille	Du sirop		La châtaigne	Le ketchup
Un oignon	Une endive	Du jus	Les fruits exotiques	Le citron	Les épices
Une tomate	Du fenouil	Du lait		La clémentine	Le <u>sel</u>
Un concombre		De l'alcool	L'ananas	La fraise	Le poivre
Une carotte	Du pain	Du vin	La banane	La framboise	Le piment
Un champignon	Des biscottes	De la bière	Le chadeck	Le fruit de la <u>passion</u>	L'ail
Des épinards	Une tartine	Du champagne	Le citron vert	La groseille	L'oignon
Un <u>chou</u>	De la farine	Du rhum	Le corossol	Le kiwi	Le curry
Un <u>chou</u> -fleur	De la cassave		Le fruit du dragon	La mandarine	Le curcuma
Un poivron		De la viande	La goyave	Le <u>melon</u>	Le gingembre
Un poireau	Les féculents	Du poulet	La <u>lime</u>	La mûre	La cannelle
Du giraumon	Du <u>riz</u>	De la dinde	Le litchi	La myrtille	Le roucou
Un <u>cœur</u> de palmier	Des pâtes	Du canard	La mangue	La <u>poisette</u>	Les <u>herbes</u>
Des haricots verts	De la semoule	Du bœuf	Le maracudja	La <u>pois</u>	L'origan
Des haricots en grain	Du couac	Un steak	La <u> noix</u> de coco	L' <u>orange</u>	Le <u>thym</u>
Des haricots rouges	Des <u>pommes</u> de terre	Du cabri	La papaye	Le pamplemousse	Le romarin
Des petits <u> pois</u>	Des patates	De l'agneau	Le ramboutan	La pastèque	Le persil
De la lentilles	De la purée	Du mouton		La <u>pêche</u>	La coriandre
Des pois d'angole	Du manioc	Du <u>porc</u>		La <u>poire</u>	La cive
Du maïs	De l'igname	Du jambon	De la confiture	La pomme	La ciboule
De la salade	De la dachine	Des saucisses	Du coulis	La prune	La <u>chouette</u>
De la laitue	Des bananes plantain	Du saucisson	De la marmelade	Le raisin	La ciboulette
	Des patates douces	Du pâté	De la compote		

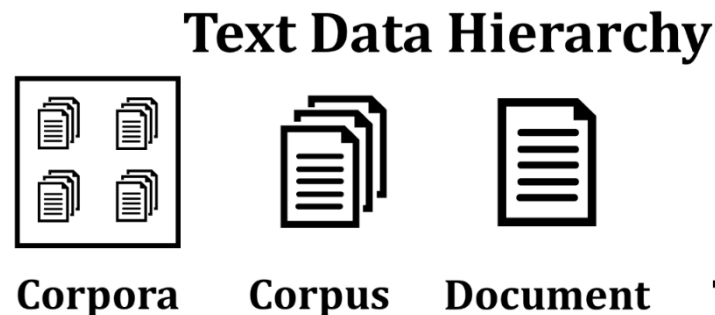
Outils du TALN - Représentation de connaissances

Corpus – Corpus

- Recueil / Ensemble fini de documents, artistiques ou non (**textes**, images, vidéos, etc.) en format électronique, regroupés comme base pour une étude précise.
- En TAL, il permet d'extraire des tendances et notamment de construire des ensembles de **n-grammes** (une séquence de n mots en TAL).



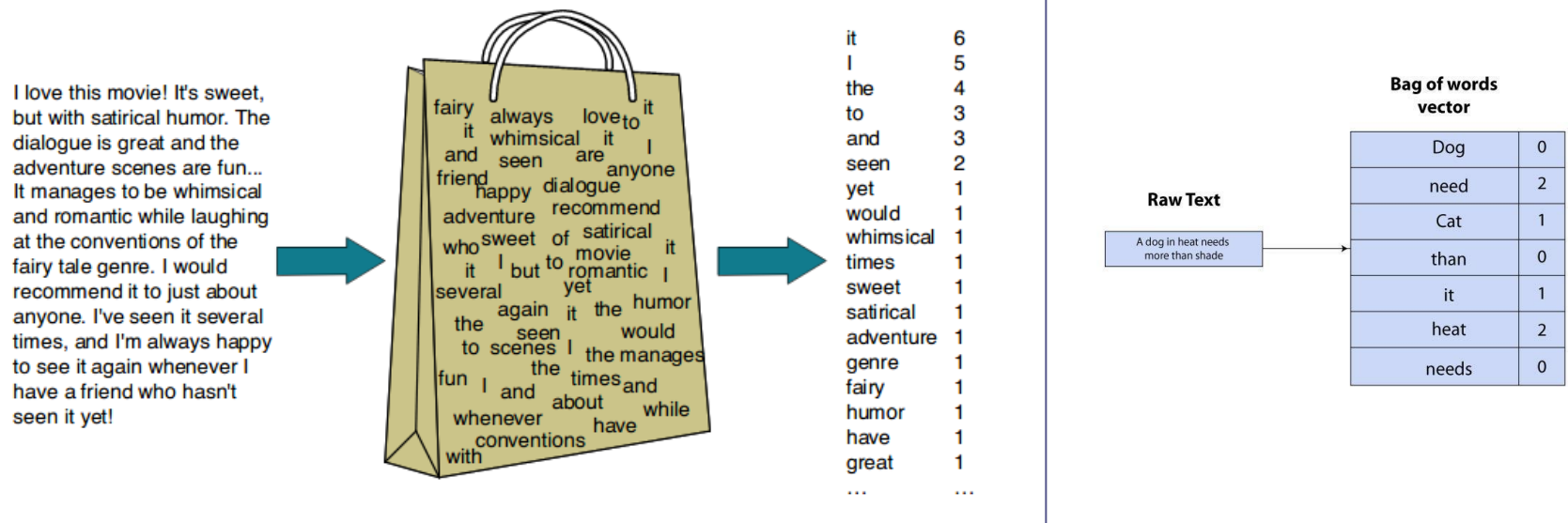
**Natural Language Processing
Corpora**



Outils du TALN - Représentation de connaissances

Sac de mots – Bag of words

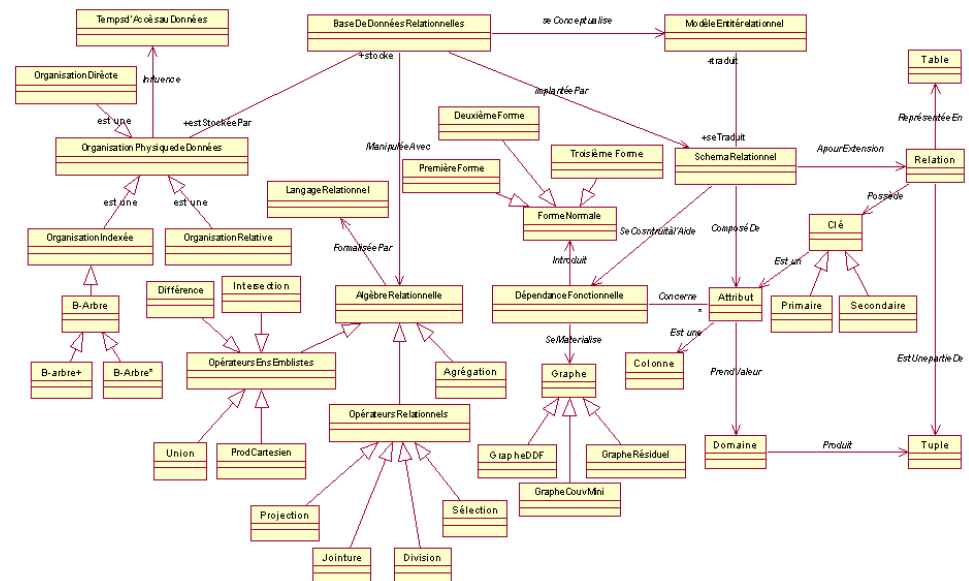
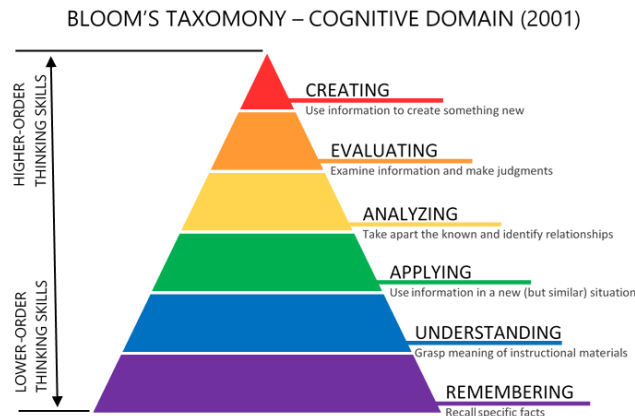
- Ensemble fini de mots (= vocabulaire d'un corpus donné).
- C'est la manière la plus simple de représenter un document par l'ensemble des mots qu'il contient.
- En pratique, ça peut être par exemple un **vecteur** de fréquence d'apparition des différents mots utilisés.

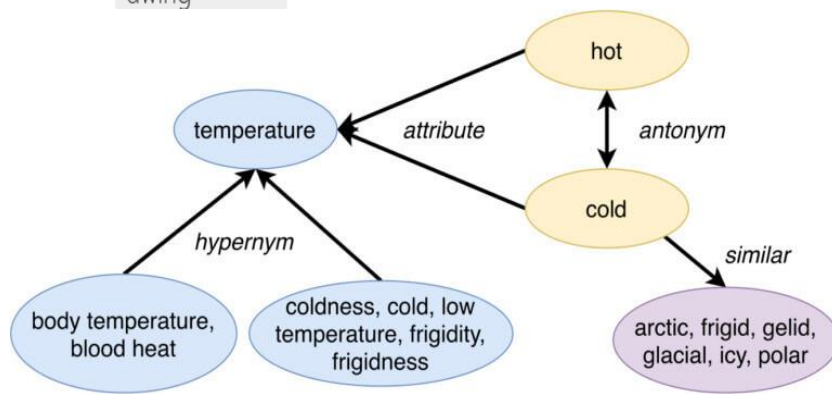


Outils du TALN - Représentation de connaissances

Thesaurus – Ontologies

- La taxonomie est la science des lois de la classification/hiérarchisation.
- Thesaurus est une taxonomie + synonymes et antonymes de mots.
- Ontologie est une sorte de taxonomie + sémantique / raisonnement / inférence. Ensemble structuré des termes et concepts représentant le sens d'un champ d'informations





Méthodes et techniques du TALN

Comment **implémenter** les systèmes TAL ?

- Les méthodes (approches) **symboliques** : méthodes fondées sur des règles fondées sur l'expertise humaine; modélisation parfois logique, automates, etc.
- Les méthodes **stochastiques**: méthodes probabilistes fondées sur des calculs statistiques effectués à partir de corpus. Statistical inference.
- Approches **empiriques** fondées sur les **données**, où la connaissance est extraite par des techniques d'apprentissage automatique :
 - ✓ Machine Learning
 - ✓ Neural networks / Deep Learning

Références

Cours - *François Yvon* – Une petite introduction au Traitement Automatique des Langues Naturelles,

<https://perso.limsi.fr/anne/coursM2R/intro.pdf>

Article – Marcel Cori - Des méthodes de traitement automatique aux linguistiques fondées sur les corpus

- <https://www.cairn.info/revue-langages-2008-3-page-95.htm>

Article - Pascale Sébillot - Le traitement automatique des langues face aux données textuelles volumineuses et potentiellement dégradées : qu'est-ce que cela change ?

- <https://hal.archives-ouvertes.fr/hal-01056396/document>

Cours - ARIES Abdelkrime - Le traitement automatique du langage naturel.

https://github.com/projeduc/ESI_2CS_TALN