Cours de Traitement Automatique du Langage

DI5 – 2021-22

Polytech ’Tours

J.Y. RAMEL

Au programme

- **Séances de cours puis TP**
  - CM : 6 x 2h (JYR + Thomas Bourquard)
  - Travaux pratiques : 6 x 2h (D Maurel + JYR)
    ➔ TP classiques au début puis mise en oeuvre d'un système TAL + rapport
  - Avec UniTex d'abord (D. Maurel)
  - Avec Python (jupyter Notebook ou autre) → installation préalable requise
  - Celene incrémentalement…

- **Evaluation**
  - CR des Travaux Pratiques (seul ou en binome)

- **Contenu**
  - Introduction : Organisation et objectifs (pédagogiques) + définitions
  - Pré-traitements et traitements basique
  - Modèles de langages
  - Etiquetage (syntaxique) de séquences et Word embedding
  - Analyses sémantiques : sens des mots, coréférences, sentiment analysis, …

Les sources - Quelques cours similaires

Les cours suivants ont largement servi pour la création de ce support

- Cours de TAL. 2021. JY Antoine, Agata Savary. Master BDMA de Blois; Université de Tours.
- Cours de TAL. 2021, ARIES A., Ecole nationale Supérieure d'Informatique d'Alger
- Cours Speech and Language Processing. de Jurafsky, D. and Martin, J. H; 2020. https://web.stanford.edu/~jurafsky/slp3/
- CS224n: Natural Language Processing with Deep Learning.Stanford university (2020). https://web.stanford.edu/class/cs224n/

Plus légèrement :

- Livre Deep Learning for Natural Language Processing - Creating Neural Networks with Python, Palash Goyal, Sumit Pandey, Karan Jain, Apress
- CS388: Natural Language Processing (2018). University of Texas. https://www.cs.utexas.edu/~mooney/cs388/
- Natural Language Processing S21 (2020). Carnegie Mellon University. http://demo.clab.cs.cmu.edu/NLP/
- CSE 517: Natural Language Processing (2020). University of Washington. https://courses.cs.washington.edu/courses/cse517/
- CS 294-5: Statistical Natural Language Processing (2005). University of Berkely. https://people.eecs.berkeley.edu/~klein/cs294-5/index.html
- CS 5340/6340: Natural Language Processing (2020). University of Utah. https://my.eng.utah.edu/~cs5340/schedule.html

Si besoin…

- Si vous avez des difficultés :
  - Internet est vaste : soyez autonomes (et critique) …
  - Références à des ouvrages / travaux tout au long du cours
- N’hésitez pas à demander, venir discuter :
  - ramel@univ-tours.fr - Bureau 2° étage
- Quelques conseils
  - Ne surtout pas commencer à travailler à la dernière minute
  - Travaillez en plusieurs fois - Se laisser du temps pour réfléchir aux problèmes
  - Du temps est disponible durant les cours et TP - Profitez en !
  - Evitez de trop surfer le web, réfléchissez avant …

Introduction

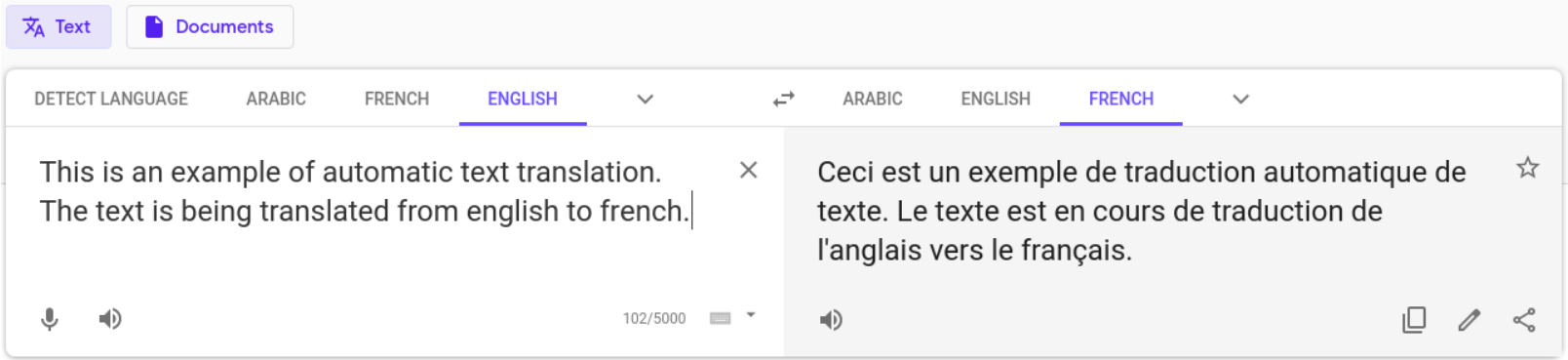
Introduction - Définition

- TALN: Traitement automatique du langage naturel (NLP)
- TAL: Traitement automatique des langues

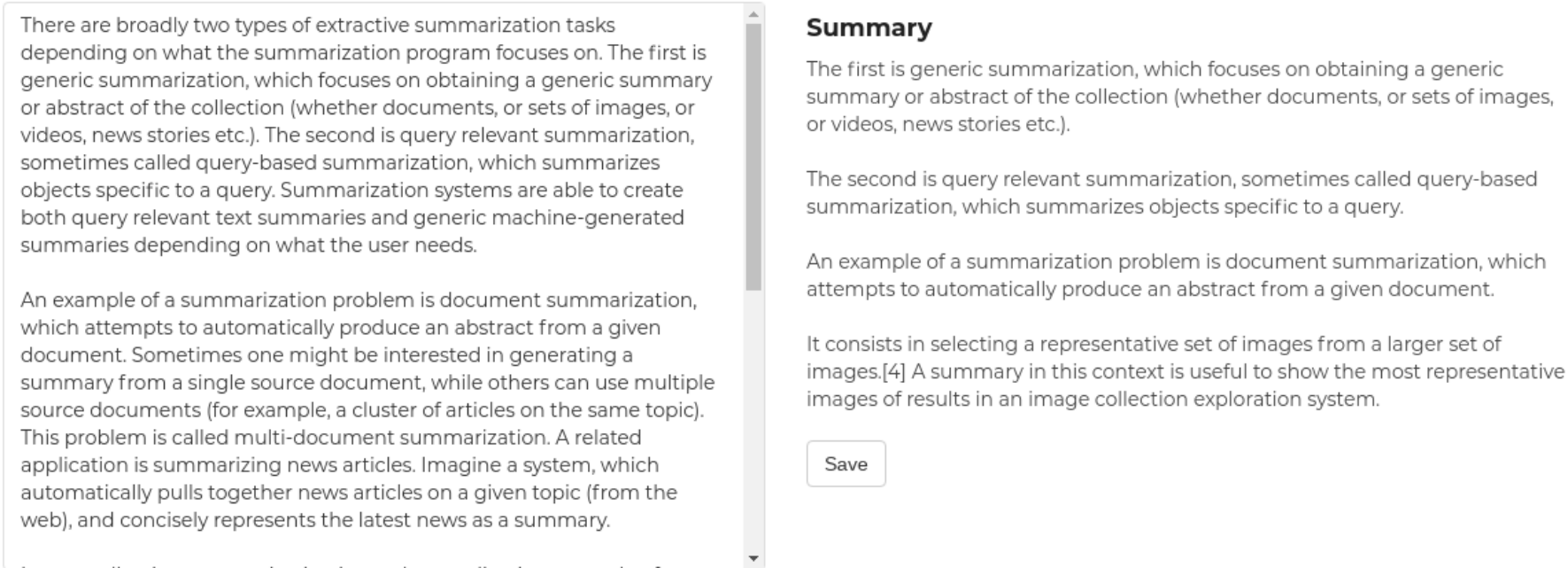
- Définition : l'ensemble des méthodes permettant de rendre le langage humain accessible aux ordinateurs.
- Un domaine multidisciplinaire
  - Linguistique: Etude du langage
  - Informatique: Traitement automatique de l'information
  - Intelligence artificielle: Ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine

Introduction – Applications du TAL

- Traduction automatique → Google Translate, DeepL

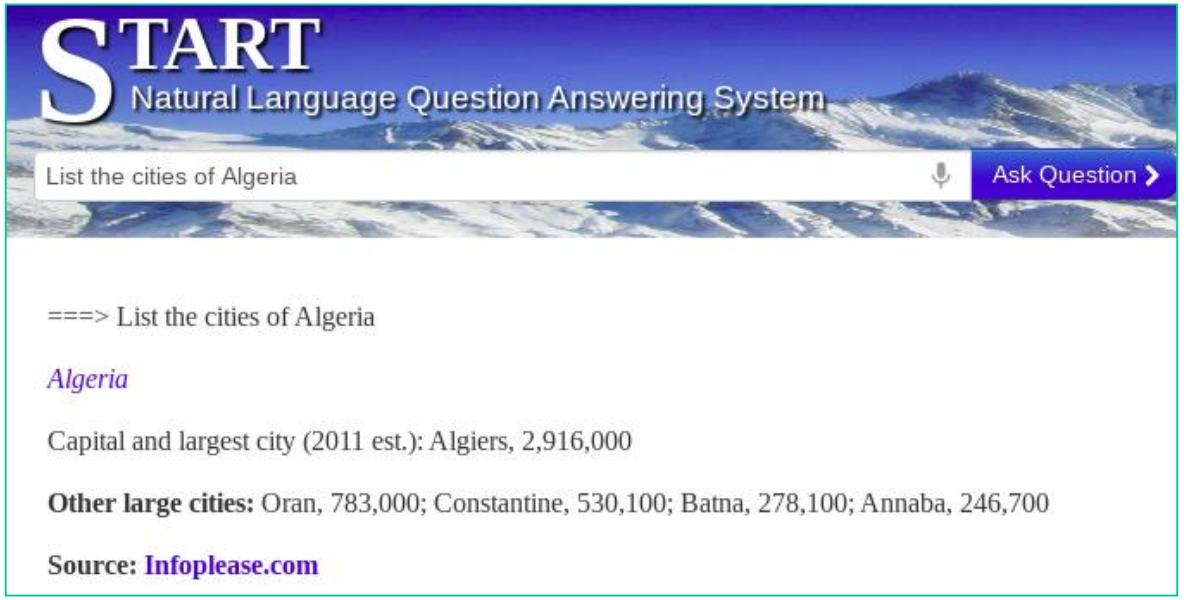


- Résumé automatique → Esummarizer



Introduction – Applications du TAL

- ChatBot, Question/Réponse → système Start



Introduction – Applications du TAL

Et toutes les autres….

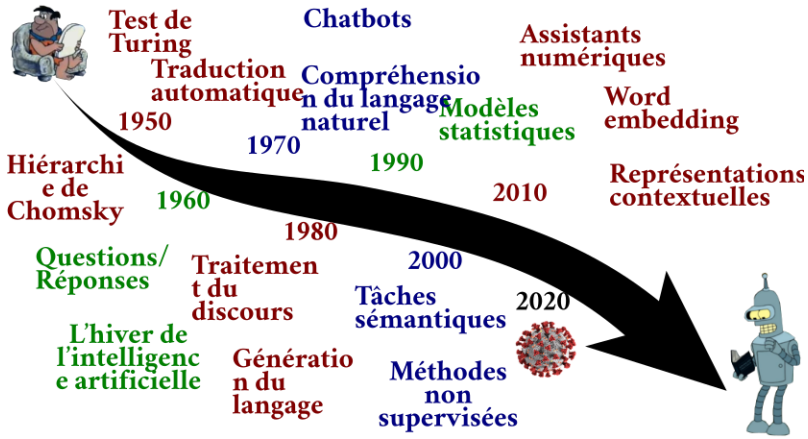
- Recherche d'informations, fouille de texte, recommandations,…
  - Connaissance du marché (Market intelligence) : surveiller les compétiteurs afin de se tenir au courant des événements liés à l'industrie.
  - Marketing (pub,…),
  - Veille de marché / d'opinion (fake news, …),
  - Santé (analyse de CR, génération de rapports, …),
  - RH (CV, …), éducation (correction auto, …)

- Analyse de sentiments (réseaux sociaux)
  - Surveillance de la réputation : on utilise l'analyse des sentiments pour savoir si les clients sont heureux avec des produits ou non.

Introduction - Histoire

Naissance de l'IA : Les années 1950

- 1951 : Shannon a exploité les modèles probabilistes des langages naturels [Shannon, 1951].
- 1954 : Expérimentation IBM pour traduire automatiquement 60 phrases du russe vers l'anglais.
- 1956 : Chomsky a développé les modèles formels de syntaxe.
- 1958 : Luhn (IBM) a expérimenté sur le résumé automatique de texte [Luhn, 1958]



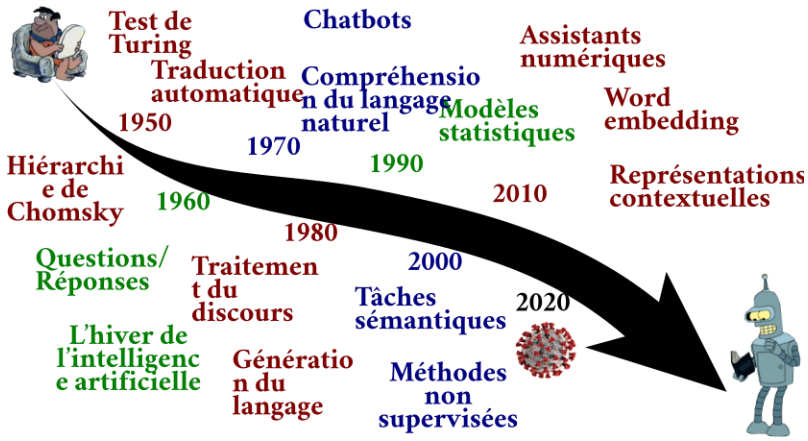
L'Age d'or de l'IA : Les années 1960

- 1961 : Développement du premier analyseur syntaxique automatique a U.Penn. [Joshi, 1961, Harris, 1962]
- 1964 : Weizenbaum a mis au point ELIZA, une simulation d'un psychothérapeute au sein du laboratoire MIT AI.
- 1964 : Bobrow a mis au point STUDENT, conçu pour lire et résoudre des problèmes d'algebre de lycée à partir d'alnalyse de texte [Bobrow, 1964]
- 1967 : Brown corpus, le premier corpus électronique

→ **Analyse probabiliste et syntaxique de texte**

Introduction - Histoire

Hiver de l'IA : Les années 1970 / 80



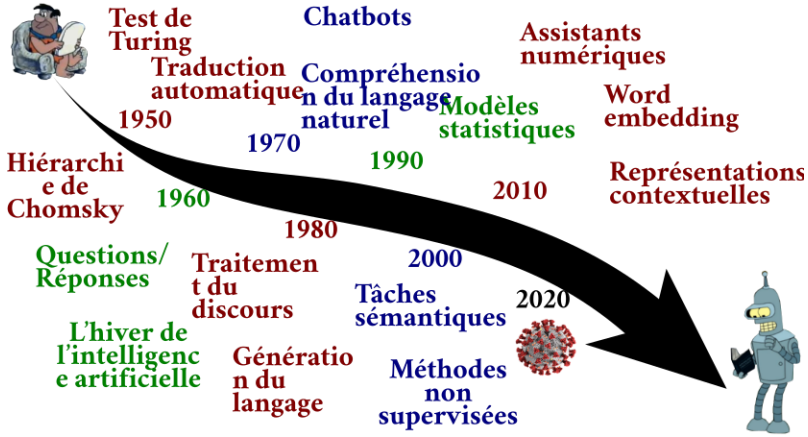
- 1971 Le MIT développe SHRDLU, un programme de compréhension du langage naturel [Winograd71]
- 1972 Colby (Stanford) cree PARRY un chatbot qui simule une personne avec la schizophrénie
- 1975 MARGIE un système qui fait des inférences et des paraphrases a partir des phrases en utilisant la représentation conceptuelle du langage.
- 1975 DRAGON, un système pour la reconnaissance de la parole en utilisant les HMM [Baker, 1975].
- 1980 KL-One, représentation de connaissance pour le traitement de la syntaxe et la sémantique [Bobrow and Webber, 1980]
- 1986 TRUMP, analyseur de langage en utilisant une base lexicale [Jacobs, 1986]
- 1987 MU, Conférence sur l'extraction des données financée par DARPA
- 1988 Utilisation des HMM dans l'étiquetage morpho-syntaxique [Church, 1988]

→ Solutions symboliques sur le traitement du discours et génération de phrases

Introduction - Histoire

Printemps de l'IA : Les années 1990-2000

- 1990 Une approche statistique pour la traduction automatique [Brown et al., 1990]
- 1993 Pen Tree-bank, un corpus annoté de l'anglais [Marcus et al., 1993]
- 1995 Wordnet, une base lexicale pour l'anglais [Miller, 1995]
- 1996 SPATTER, un analyseur lexical statistique basé sur les arbres de décision [Magerman, 1996]



→ **Popularité des méthodes statistiques, génération de corpus**

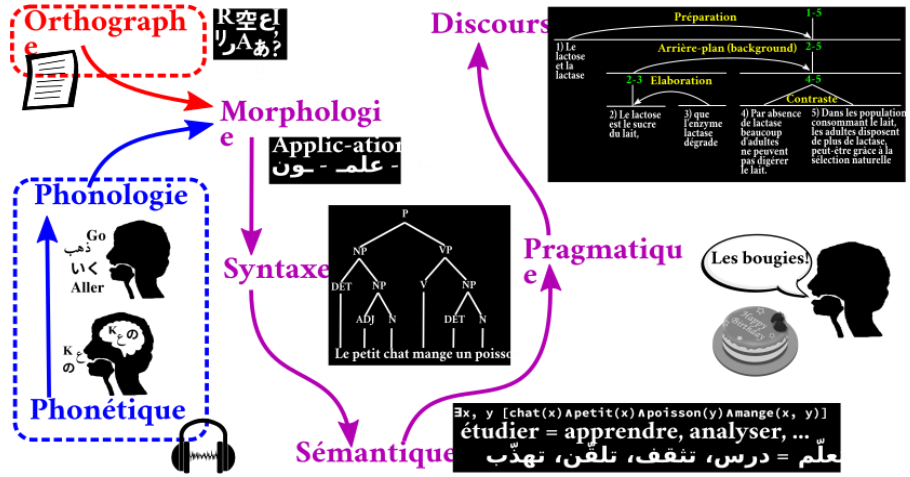
- 2003 Les modelés probabilistes de langues en utilisant les RN [Bengio et al., 2003]
- 2006 Watson : un système de question/réponse IBM

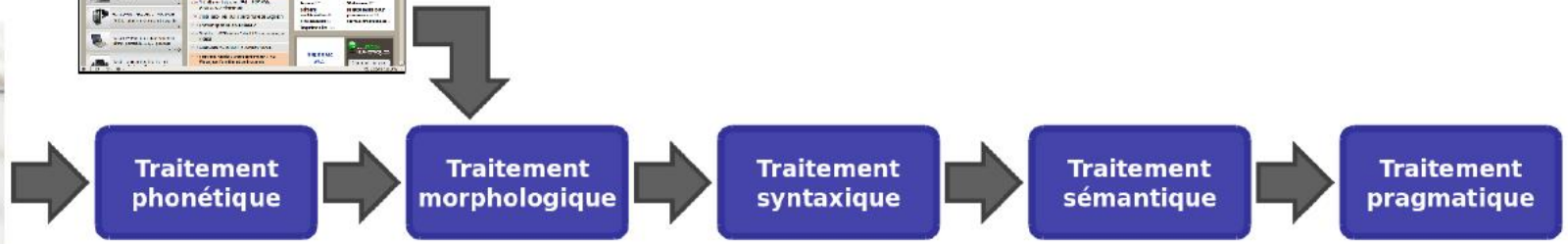
→ **Utilisation de l'apprentissage non supervisé et semi-supervisé comme alternatives a l'apprentissage purement supervisé**

- 2011 Sortie de Siri (Apple), suivi par Alexa (Amazon,2014) et Google Assistant (2016)
- 2014 Word embedding [Lebret and Collobert, 2014]
- 2018 Apparition des représentations contextuelles (des modèles de langue pré-entraınes) :
  ULMfit (fast.ai) [Howard and Ruder, 2018], ELMO (AllenNLP) [Peters et al., 2018], GPT(OpenAI) [Radford et al., 2018], BERT(Google) [Devlin et al., 2018], XLM (Facebook) [Lample and Conneau, 2019]

→ **Déplacement du focus sur les taches sémantiques**

Introduction - Niveaux de traitement d'une langue





Introduction - Niveaux de traitement

Phonetique, phonologie et orthographe : Phonologie

- Etude des sons ou phonemes d'une langue donnee
- s'interesse aux sons en tant qu'elements d'un systeme

Exemple : le phonème */r/*

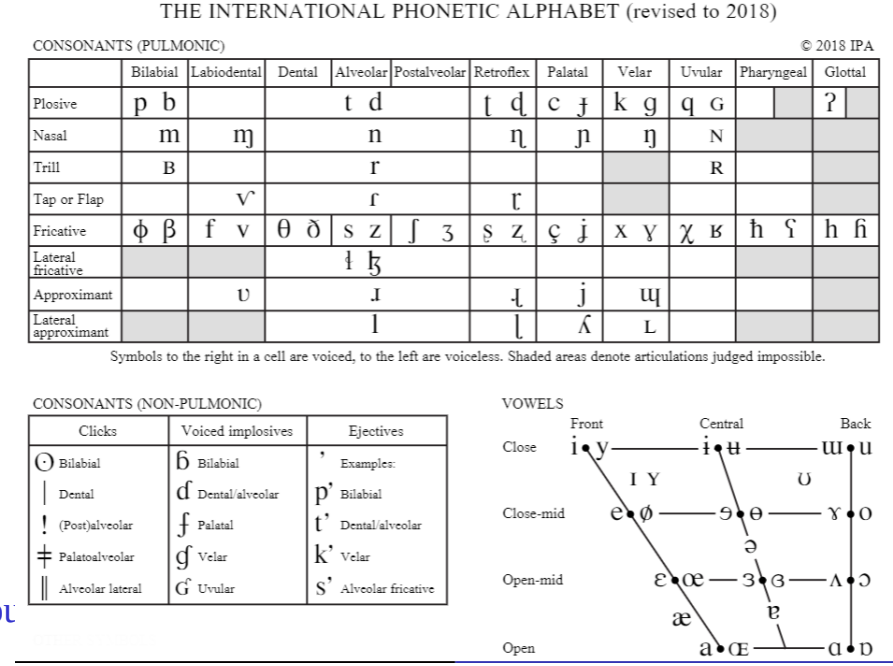
- En français, le *r* peut se prononcer (en phonétique) : roulé [r], grasseyé [ʀ], ou normal (parisien) [ʁ]
- Il est transcrit toujours de la même façon, exemple rat /rat/
- En arabe, on trouve les consonnes ر [r] et غ [ɣ] qui ont deux phonèmes différents : /r/ et /ɣ/ respectivement. Exemple, غريب /ɣæriːb/ (étranger)

Introduction - Niveaux de traitement

Phonétique, phonologie et orthographe : Orthographe

Etude des types et de la forme des lemmes / système d'écriture (selon le graphème)

- Logo-graphique: logo-grammes, chacun est un graphème unique notant un lemme (mot).
  - Exemple : Kanji (Japonais) :日,本,語
- Syllabique : symboles, chacun représente un syllabe (son vocalisé).
  - Exemple, Hiragana (Japonais) :る,た,め,の;
    Katakana (Japonais) :セ,ク;
- Alphabétique : lettres, chacune d'elles représente un phonème.
  - Exemple, Le latin : A, B, C, etc.
    L'arabe : ب, ت, ٥
- Ponctuation
- Règles d'écriture



Introduction - Niveaux de traitement

Morphologie et syntaxe : Morphologie

- Etude de la formation des mots, y compris la façon dont les nouveaux mots sont inventés dans les différentes langues
- Etude de la façon dont les formes des mots varient en fonction de leurs utilisations dans les phrases.
- Morphème: la plus petite unité de langage qui a sa propre signification.
  - Exemple, les noms propres, les suffixes, etc.
- Lexème: un ensemble de toutes les formes grammaticales qui ont le même sens.
- Lemme: un mot choisi parmi ces formes pour représenter le lexème.
- Les catégories grammaticales : classe ouverte (adjectif, nom, verbe) et classe fermée (adverbe, article, conjonction, interjection, preposition, pronom)

Introduction - Niveaux de traitement

Typologie morphologique des langues

- Langues isolantes/analytiques: chaque mot est constitué d'un et d'un seul morphème
  - Les modifications morphologiques sont peu nombreuses, voire absentes.
  - Parmi ces langues : mandarin, vietnamien, thaı, khmer, etc..
  - Exemple : 四个男孩 → "quatre garcons" (lit."quatre [entité de] masculin enfant")

- Les langues flexionnelles/synthétiques: les mots sont formés d'une racine en plus de morphèmes supplémentaires

- Langues agglutinantes: les morphèmes sont toujours clairement différentiables phonétiquement l'un de l'autre.
  - Parmi ces langues : finnois, turc, japonais, langues berberes, etc..
  - Exemple → 行く,行きます

- Langues fusionnelles: il n'est pas toujours aisé de distinguer les morphèmes de la racine, ou les morphèmes les uns des autres.
  - Parmi ces langues : anglais, français, arabe, etc.
  - Exemple: foot, feet

Introduction - Niveaux de traitement

Morphologie flexionnelle

- Formation de mots sans changer de catégorie ou créer de nouveaux lexèmes
- Modification de la forme des lexémes afin qu'ils s'adaptent à différents contextes grammaticaux
- Flexion: déclinaison ou conjugaison
- Affixation : préfixe, infixe, suffixe → étudiant (masculin-singulier),étudiantes (feminin-pluriel)
- Nombre: singulier, pluriel
- Personne: premiere, deuxieme, troisieme, etc.
- Genre: masculin, féminin, neutre, commun
- Temps: passé, présent, futur

- Voix: active, moyenne, passive, etc.
- Polarité: affirmative, négative
- Politesse: informelle, formelle, etc.

Introduction - Niveaux de traitement

Morphologie dérivationnelle

- Formation de mots en changeant de catégorie (jouer, joueur) ou en créant de nouveaux lexémes (connecter, déconnecter)
- Affixation : en utilisant des préfixes, infixes et/ou suffixes.
  - Exemple : happy (ADJ), unhappy (ADJ), unhapyness (N)
- Composition : en fusionnant des mots dans un seul
  - Exemple : porter(V) + manteau (N) = porte-manteau (N); wind (N), mill (N), windmill (N)
- Conversion : en changeant la catégorie grammaticale d'un mot sans aucune modification
  - Exemple : orange (fruit, N), orange (couleur, ADJ); visiter (V), visite (N); fish (N), to fish (V)
- Troncation :
  - Exemple : bibliographie, biblio, information, info

Introduction - Niveaux de traitement

Syntaxe

- Structure des phrases : comment les mots se combinent pour former des phrases
- Catégories grammaticales ou parties du discours (verbe, Nom, Adjectif, etc) des mots de la phrase (étiquetage morpho-syntaxique).
- Ordre des mots selon le sujet (S), le Verbe (V) et l'Objet (O)
- Fonctions grammaticales : Sujet, COD, COI, etc.

Exemple SOV [japonais]

カリムさんは日本語を勉強します。

Exemple SVO [français]

Karim apprend le français.

Exemple VSO [arabe]

يتعلّم كريم العربية .

Ordre	Proportion	Exemples
SOV	44.78%	japonais, latin, tamoul, basque, ourdou, grec ancien, bengali, hindi, sanskrit, persan, coréen
SVO	41.79%	français, mandarin, russe, anglais, haoussa, italien, malais (langue), espagnol, thaï
VSO	9.20%	irlandais, arabe, hébreu biblique, philippin, langues touarègues, gallois
VOS	2.99%	malgache, baure, car (langue)
OVS	1.24 %	apalai, hixkaryana, klingon (langue)

Introduction - Niveaux de traitement

Syntaxe – grammaire de constituants

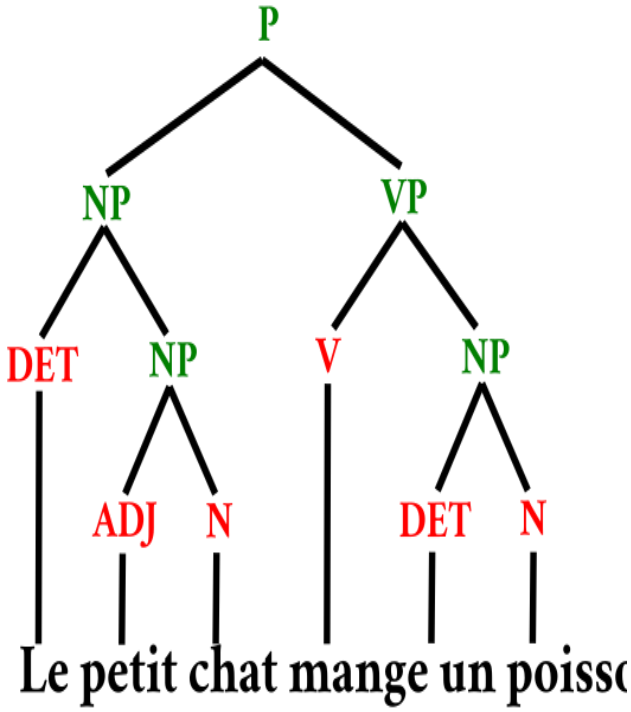
- Une phrase est constituée de plusieurs syntagmes qui sont constitues des mots et d'autres syntagmes.
- Un syntagme contient un noyau qui est l'élément central
- Selon le noyau, le syntagme peut être : nominal (NP), adjectival (AP),verbal (VP) ou prépositionnel (PP)
- Le système formel le plus utilisé pour modéliser la structure des constituants d'une phrase est la grammaire hors-contexte

Introduction - Niveaux de traitement

Syntaxe – grammaire de constituants

- La grammaire qui a généré cet arbre syntaxique peut être écrite :
  - P→NP VP
  - NP→DET NP'
  - NP→DET N
  - NP'→ADJ N
  - VP→V NP

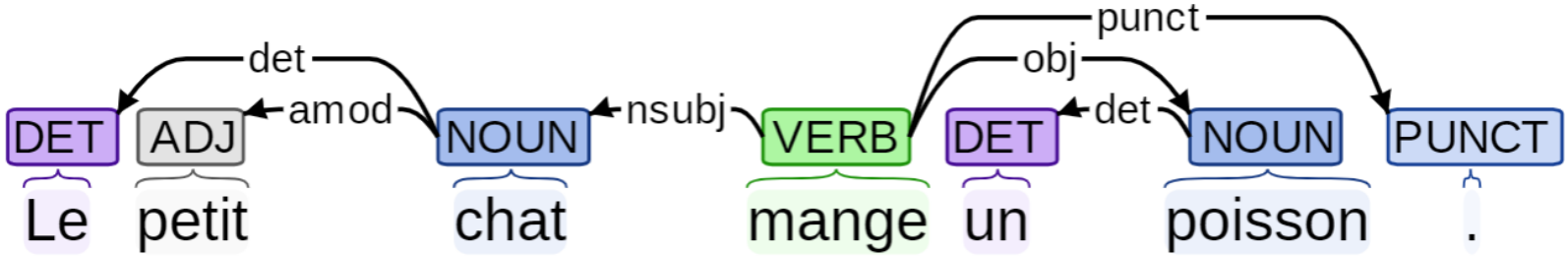
Rem : La deuxième règle n'ai pas écrite (NP→DET NP) sinon on peut avoir plusieurs déterminants pour un nom



Introduction - Niveaux de traitement

Syntaxe – grammaire de dépendances

- Grammaire de dépendances: un ensemble des relations binaires entre les mots de la phrase
- La structure syntaxique est décrite en terme de mots et pas des syntagmes
- Les relations peuvent être : un sujet nominal (nsubj), un objet (obj), un modificateur d'adjectif (amod), déterminant (det), etc



Un exemple de dépendances générée par `https://corenlp.run/`

Introduction - Niveaux de traitement

Sémantique

- Etude du sens dans les langages
  - Sémantique lexicale: sens des mots
  - Sémantique propositionnelle: sens des phrases

Un exemple de différents sens

Le terme poulet a plusieurs sens selon la phrase (https://fr.wiktionary.org/wiki/poulet)

- J'écoute les piaulements des poulets. "*Petit du coq et de la poule, plus âgé que le poussin, avant d'être adulte.*"
- Je mange du poulet. "*Viande de jeune poule ou jeune coq.*"
- Mon petit poulet. "*Terme d'affection, que l'on adresse généralement aux enfants.*"

Introduction - Niveaux de traitement

Sémantique : Le sens des mots

- Dans les systèmes logographiques, un graphème représente un ou plusieurs sens (en général, avec plusieurs prononciations)

川 (rivière), 山(montagne),
音(son, bruit) + 楽(musique, confort, facilité) =音楽(musique)

- Un mot peut avoir plusieurs sens (Polysémie)
  - Le sens d'un mot est construit d'un ensemble de "primitives sémantiques"
  - a l'aide de relations entre morphèmes (un réseau de sens)

- Sème : une unité minimale de signification (trait sémantique minimal)
- Sémème : un faisceau de sèmes correspondant a une unité lexicale

Un exemple de l'analyse sémique

Mot/Sème	animé	domestique	félin
Chat	+	+	+
Lion	+	-	+
Chien	+	+	-

Introduction - Niveaux de traitement

Sémantique : Le sens des mots

- Synonyme : si on substitue un mot par un autre dans une phrase sans changer le sens, donc les mots sont des synonymes
- Antonyme : le sens opposé. Les deux mots doivent exprimer deux valeurs d'une même propriété.
  - Exemple, grand et petit expriment la propriété taille
- Hyponyme : un mot ayant un sens plus spécifique qu'un autre.
  - Exemple, chat est l'hyponyme de félin.
- Hyperonyme : un mot avec un sens plus générique.
  - Exemple, felin est l'hyperonyme de chat

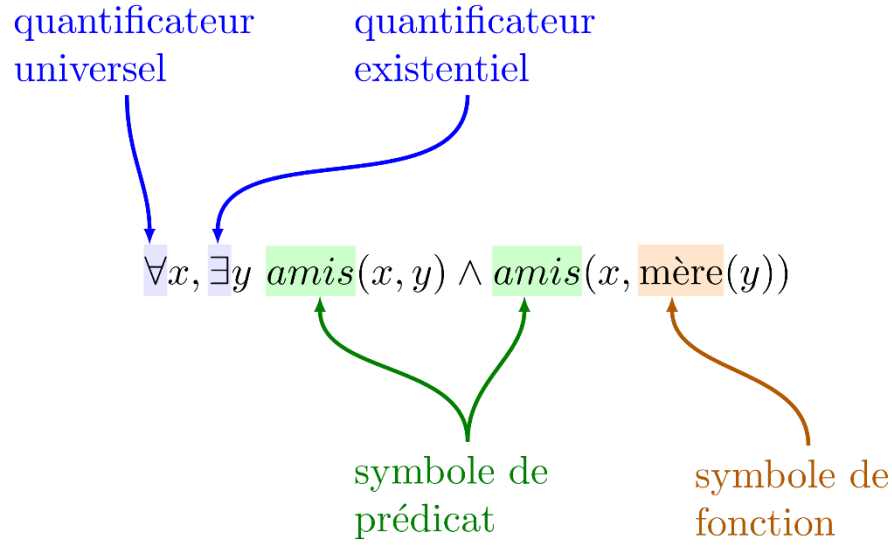
Introduction - Niveaux de traitement

Sémantique : Le sens des propositions

- Souvent représentée par une formule de logique du premier ordre (calcul de prédicat)

Logique des prédicats:

- Constante : une entité spécifique du monde (Pierre, Polytech)
- Variable : réfère a une entité anonyme : x, y.
- Fonction : résulte d'une entité spécifique sans avoir besoin de créer une nouvelle constante
  Ex. locationOf(Polytech)
- Prédicat : une relation entre des entités
  Ex. Posséder(x,y)
- Connecteurs logiques ($\neg$, $\wedge$, $\vee$, $\Rightarrow$, =, …) et quantificateurs ($\forall$, $\exists$)



Introduction - Niveaux de traitement

Sémantique : Le sens des propositions

Un exemple de la représentation sémantique d'une phrase

- Quelques étudiants possèdent deux ordinateurs
- Deux ordinateurs sont possédés par quelques étudiants
- $E = Etudiant, O = Ordinateur, P = Possede$
- $\exists x(E(x) \wedge \forall y(O(y) \Rightarrow P(x,y)))$ ☒
- $\exists x(E(x) \wedge \exists y,z(((O(y) \wedge P(x,y)) \wedge (O(z) \wedge P(x,z))))$ ☒
- $\exists x(E(x) \wedge \exists y,z(\neg y = z \wedge ((O(y) \wedge P(x,y)) \wedge (O(z) \wedge P(x,z)))))$ ☑
- $\exists x(E(x) \wedge \exists y((O(y) \wedge P(x, twoOf(y)))))$ ☑

Un exemple d'inférence

$$\frac{\text{Chaque homme est mortel. } \forall x(Homme(x) \Rightarrow Motel(x)) \qquad \text{Socrate est un homme. } Homme(Socrate)}{Motel(Socrate)}$$

Introduction - Niveaux de traitement

Sémantique : Pragmatique et discours

- Imaginer qu'on est dans un anniversaire
- Une personne crie : "Les bougie!"
- On peut directement comprendre qu'il n'y a pas des bougies sur la tarte
- Question: comment peut-on déduire ca ?
- Réponse: Savoir prendre en compte le contexte

Sémantique	**Pragmatique**
une personne **A** est arrivée tard à un rendez-vous. **B** : "A votre avis, quelle heure est-il ?"	
A annonce l'heure	**A** explique les raisons du retard

Introduction - Niveaux de traitement

Sémantique : Pragmatique et raisonnement

- Implication conversationnelle: réfère a ce que le locuteur veut dire d'une façon implicite.
- D'après [Grice, 1979], les interlocuteurs doivent respecter certaines normes (maximes) conversationnelles : quantité, qualité, pertinence et manière.
- Présupposition: réfère aux suppositions faites par les interlocuteurs lors de la communication.
- Acte de langage: référe a l'interaction linguistique du locuteur afin d'agir sur son environnement.
- Une liste des actes de langage est fournie par [Austin, 1962] : déclaration, ordre, question, interdiction, salutation, invitation, félicitations, excuses, …

Introduction - Niveaux de traitement

Sémantique : Discours et coréférences

Un exemple de coréférence

- ***The cat*** doesn't fit in the box because ***it*** is too big.
- The cat doesn't fit in ***the box*** because ***it*** is too small.

- Exemples
  - Pronoms: le chat a chassé une souris et **il** joue avec elle.
  - Syntagmes nominaux: Apple est un fabricant d'ordinateur. La firme est mondialement connue.
  - Noms: Comme Apple, IBM fabrique des ordinateurs.
  - Zero Anaphora: dans des langues comme le japonais, parfois, la référence est omise.

Introduction - Niveaux de traitement

Sémantique : Discours et cohérences

- Relation entre les énoncés d'un discours
- Rhetorical Structure Theory (RST): le modèle du discours le plus utilisé
- Deux types d'énoncé : Noyau et satellite
- L'étudiant s'est absenté hier. Il a été malade.
  - Première phrase est un noyau puisqu'elle décrit l'evenement principal
  - Deuxième phrase est un satellite puisqu'elle dépend de la première
- Il y a plusieurs relations de cohérence : raison, élaboration, évidence, attribution, narration, etc

Un exemple de la raison

[$_{NOY}$ L'étudiant s'est absenté hier.] [$_{SAT}$ Il a été malade.]

Un exemple de l'élaboration

[$_{NOY}$ L'examen est facile.] [$_{SAT}$ Il ne prend qu'une heur.]

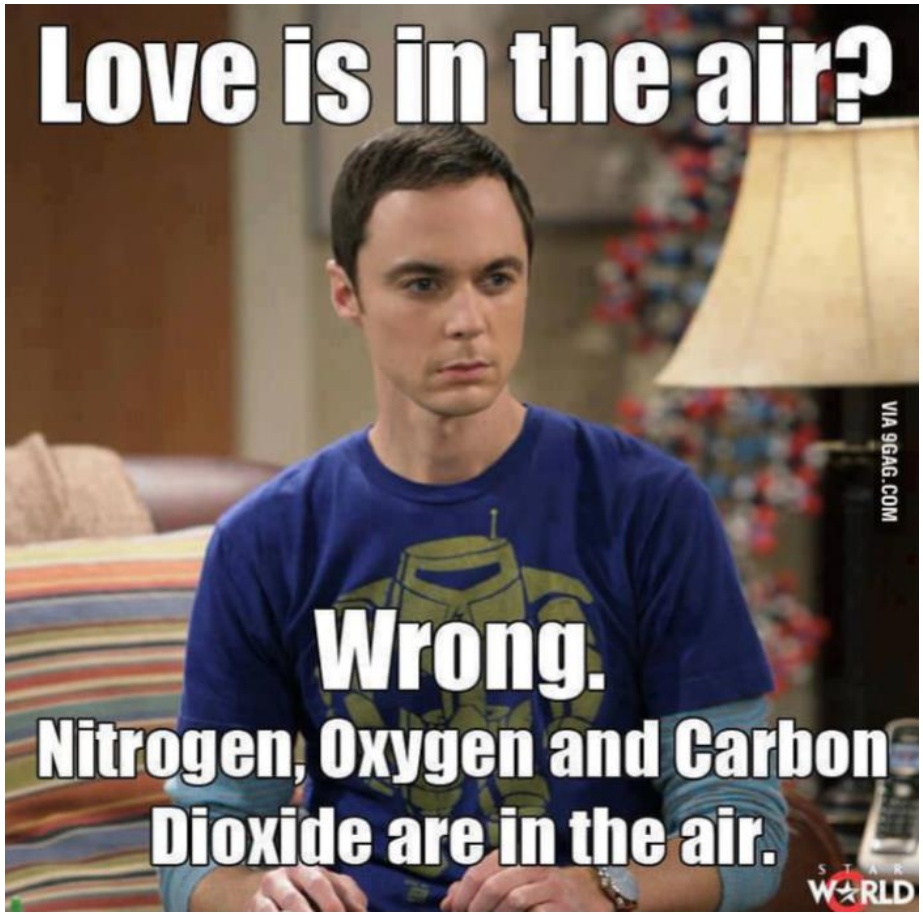
Un exemple de l'évidence

[$_{NOY}$ Kevin doit être ici.] [$_{SAT}$ Sa voiture est garée à l'extérieur.]

Introduction – Retour sur les applications

- Niveau 1 : Tache morpho-syntaxique
  - Délimitation de la phrase et séparation des mots
  - Lemmatisation et racinisation: transformer les mots en une forme standard (token)
  - Etiquetage morpho-syntaxique: trouver les catégories grammaticales des mots d'une phrase
  - Analyse syntaxique: trouver la structure syntaxique d'une phrase (comment elle est formée)
  - Extraction terminologique: rechercher des terminologies spécifiques dans un texte (E.R.)
- Niveau 2 : Tache Sémantique
  - Etiquetage des rôles sémantiques: chercher le rôle sémantique des mots et syntagmes
  - Reconnaissance d'entités nommées
  - Analyse sémantique: trouver la représentation sémantique du texte
  - Paraphrase: formuler un texte différemment → résumé automatique
  - Génération automatique de texte
- Niveau 3 : Analyse de discours
  - Résolution de coréférence: trouver les différentes références à un meme objet dans le texte
  - Analyse du discours: chercher les relations entre les phrases
  - Résolution d'ellipse: trouver les éléments omis du texte.
    Exemple d'ellipse : "Pierre mange des cerises, Paul des fraises"

Pour conclure l'introduction…



Les défis restent nombreux…

- Manque de ressources et d'évaluations
  - Selon les langues
  - Annotation manuelle des corpus d'entraınement et de test
  - Évaluation manuelle
- Prise en compte d'un contexte étendu
- Ambiguïtés et coréférences
  - polysémie: un mot ayant plusieurs sens
  - Métaphore, expression polylexicale ("It's raining cats and dogs")
  - Variations (franglais, sms,…), personnalité, émotions,
- Aspects éthiques
  - Vie privée et utilisation
  - Biais dans les corpus