

# Introduction au Traitement Automatique des Langues

Anne-Laure Ligozat

2018/2019<sup>1</sup>

---

1. librement inspiré des cours de Xavier Tannier, Aurélien Max et Dan Jurafsky, que je remercie

# Qu'est-ce que le TAL ?

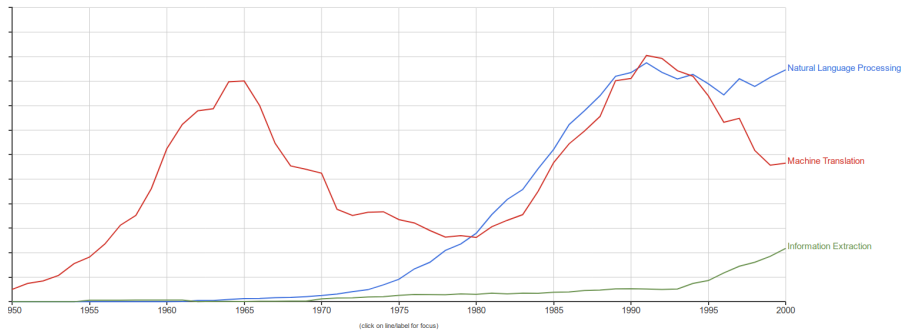
Traitement automatique des langues ou TAL (🇺🇸 Natural Language Processing ou NLP)

- discipline de l'informatique et de l'IA qui étudie les interactions entre les ordinateurs et les langues
- discipline à la frontière de la linguistique et de l'informatique
- née à peu près en même temps que l'informatique (années 50)
- ambition initiale : traduction automatique

## Difficultés de la traduction automatique

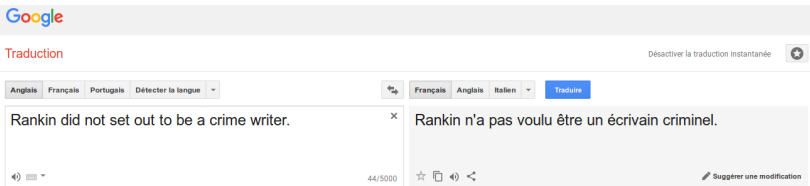
- The spirit is willing but the flesh is weak. (*L'esprit est fort mais la chair est faible.*)
- The vodka is strong but the meat is rotten. (*La vodka est forte mais la viande est pourrie.*)

# Historiquement



# À quoi sert le TAL ?

- traduction automatique
- correction orthographique
- extraction d'information
- simplification de textes
- agents conversationnels
- ...



The screenshot shows the Google Translate web interface. At the top left is the Google logo. Below it, the word "Traduction" is displayed in red. On the right side, there is a link "Désactiver la traduction instantanée" and a plus icon. The main interface has a language selection bar with "Anglais", "Français", "Portugais", and "Détection de la langue" (with a dropdown arrow). Below this, there are buttons for "Français", "Anglais", and "Italien" (with a dropdown arrow), and a blue "Traduire" button. The input text box contains "Rankin did not set out to be a crime writer." with a close button (X) and a character count "44/5000". The output text box contains the French translation "Rankin n'a pas voulu être un écrivain criminel." Below the output, there are icons for star, copy, speaker, and left arrow, along with a link "Suggérer une modification".

Google translate

# À quoi sert le TAL ?

- traduction automatique
- **correction orthographique**
- extraction d'information
- simplification de textes
- agents conversationnels
- ...

Antidote - Corrector - "Letter.doc"

Microsoft Word document Language • 60% French, 40% English Custom usage

Correction

- ✓ Language 21
- T Typography 11
- S Style 8

Revision

- ★ Pragmatics >
- ★ Semantics >
- ★ Vocabulary >
- ★ Logic >

Statistics

Nous avons le projet emballant de faire une grande fête samedi le 6 novembre à 8 hres dans notre logement pour commémorer le 10ème anniversaire de mariage du neveu de la directrice de la cousine du coiffeur du maire. Ce dernier n'a pas encore confirmer sa présence à la fête, mais, quoiqu'il fasse, il serait agréable que vous et vos bambins se joignez à notre fête, sans compter que ce sera pour vous une or joigniez Mode

Mode - Le verbe joindre se conjugue au subjonctif quand il sert de complément au verbe être dans ce contexte.

▼ Errors (18)

- French
- samedi le le s...
- confir... con...
- quoiqu' quo...
- se vous
- joignez joig...

# À quoi sert le TAL ?

- traduction automatique
- correction orthographique
- **extraction d'information**
- simplification de textes
- agents conversationnels
- ...

Google





Tous Maps Images Actualités Vidéos Plus Paramètres Outils

Environ 487 000 résultats (0,61 secondes)


**Tourisme Edimbourg | Voyageurs du Monde | VoyageursduMonde.fr**  
 Consultez les avis voyageurs et conseils de TourismePratique.com

**Tourisme Edimbourg | Millebonnesadresses.com**  
 Retrouvez les meilleures adresses de la région de la ville de Edimbourg. Restaurants à pied, à vélo, la gastronomie italienne. Escapades à la journée. Plage Edimbourg

**Activités à découvrir à Edimbourg**

 Château d'Edimbourg Forteresse spectaculaire, vue sur	 Old Town Histoire et shopping	 Palais de Holyrood Visites abbaye et palais royal du XVIIe	 Princes Street Shopping
---	---	--	---

Guide de voyage pour Edimbourg



## Edimbourg


Voyage

Située sur une colline et fort peuplée, Edimbourg est la capitale de l'Ecosse. Elle inclut une vieille ville médiévale et une élégante nouvelle ville géorgienne avec jardins et bâtiments néoclassiques. Dominant la ville, le château d'Edimbourg abrite les joyaux de la couronne d'Ecosse et la pierre du destin, utilisée lors du couronnement des monarques écossais. Arthur's Seat (le "siège d'Arthur") est un imposant pic situé à Holyrood Park et offrant une vue panoramique sur la ville. Calton Hill accueille monuments et métronaux en son sommet.

**Météo** : 17 °C, vent SE à 13 km/h, 56 % d'humidité  
**Heure locale** : vendredi 12:33

Organiser un voyage

Guide de voyage Edimbourg



# À quoi sert le TAL ?

- traduction automatique
- correction orthographique
- extraction d'information
- simplification de textes
- agents conversationnels
- ...

Having Nyong'o, with her darker skin and natural short crop, on the cover of Porter magazine's "Desire Issue" or putting trans actress and activist Laverne Cox on Variety's cover would have been unheard of years ago.

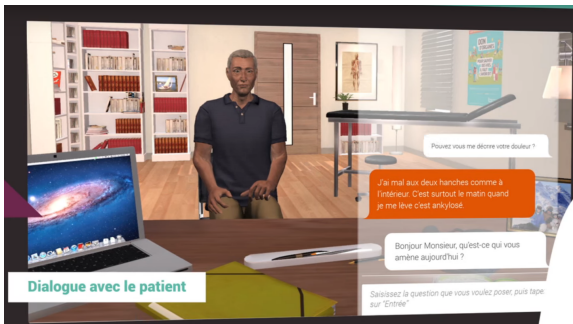
→

Nyong'o has dark skin and short hair on one magazine's cover. Laverne Cox is a transgender actress who appears on another cover. These women would not have been on magazine covers years ago.

Newsela

# À quoi sert le TAL ?

- traduction automatique
- correction orthographique
- extraction d'information
- simplification de textes
- **agents conversationnels**
- ...



PatientGenesys



# Les acteurs du domaine

- gros éditeurs
  - Facebook, IBM, Microsoft, Xerox, Apple, Toshiba, Sony, Google, Yahoo, Orange, etc.
- intégrateurs/utilisateurs
  - Ford, Symantec, EADS, Thalès/Arisem, BBN, SRI, EC, etc
- PME françaises
  - Exalead, Temis, ACapella, Lingway, Sinequa, Synapse, Systran, Reverso/Softissimo, Vecsys, Pertimm, Mondeca, etc.
- laboratoires de recherche
  - John Hopkins, Stanford, Berkeley, MIT, U. Maryland, Columbia, NYU, Cambridge, Edimbourg, Aix-la-Chapelle, Stuttgart, Paris Diderot/INRIA, Paris Sud/LIMSI etc.

# Difficultés du langage naturel

## Langage naturel

- ambigu
  - *Hospitals are sued by 7 foot doctors.*
  - *Teacher strikes idle kids.*
- implicite
  - *London is a famous writer.*
- redondant

# Concrètement...

- langue non standard
  - *Ouuuui c'est fini les horaires d'été #RERB*
  - *s est pas de votre faute après s est les siège pourris que j aime pas*
- expressions toutes faites
  - *avoir les pieds sur terre*
  - *retourner sa veste*
- néologismes
  - *uberisation*
  - *songwriteuse*
- noms d'entités
  - *il crée sa plus célèbre chorégraphie, Le Sacre du printemps*
  - *quatre années après le premier Rebus*
  - *elle sort le single Formation*

...

1 Introduction

2 Bribes de linguistique

3 Analyses TAL

# Les différents niveaux de la langue

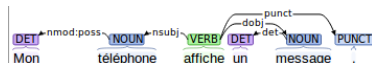
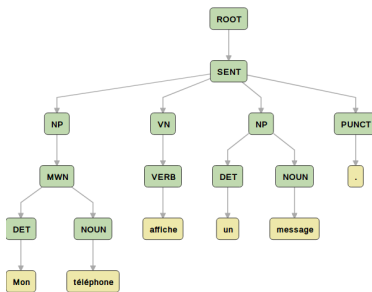
- *phonétique et phonologie, étude des sons*
- morphologie, étude des mots
- syntaxe, étude de la façon dont les mots sont agencés
- sémantique, étude du sens
- *pragmatique, étude de l'utilisation du langage en contexte*

# Morphologie

- mot = forme de mot **chantais, chantons...** ou unité lexicale **chanter**
- notions de radical et d'affixes (préfixes et suffixes) **a-ton-al**
- procédés courants :
  - dérivation : création de nouvelles unités lexicales par ajout d'affixes  
**banal** → **banaliser**
  - flexion : variation d'une unité lexicale en fonction de critères grammaticaux **oeil** → **yeux**
    - forme non fléchie d'un mot = lemme
  - composition : création de nouvelles unités lexicales à partir d'unités existantes **portefeuille**
- (morpho-syntaxe) parties du discours ou catégories/classes morpho-syntaxiques **La/DET mer/NOUN est/VERB bleue/ADJ ./PUNCT**
  - éventuellement complétées d'informations morphologiques  
**La/DET-fs mer/NOUN-fs est/VERB-3ppi bleue/ADJ-fs ./PUNCT**

# Syntaxe

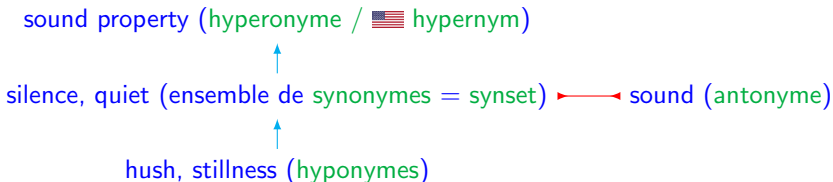
- unité de base = phrase (🇺🇸 sentence) = suite de mots
- arbres syntaxiques : analyse en constituants ou en dépendances
  - nature = groupe nominal, groupe verbal, adjectif...
  - fonction = sujet, complément d'objet, attribut...
- notion de grammaire



# Sémantique

## Sémantique lexicale

- relations lexicales : synonymie, antonymie, hyponymie
- homonymie, polysémie



relations **WordNet**

## Sémantique grammaticale

- dénotation (objet du monde désigné) vs sens **la licorne d'en face**
- construction du sens d'une phrase



1 Introduction

2 Bribes de linguistique

3 Analyses TAL

# Analyses TAL - niveau du mot

## Segmentation

### Segmentation en mots (🇺🇸 tokenization)

- problème très complexe dans certaines langues
- délimiteurs de mots (et de phrases) ambigus
  - *etc., T.A.L, 21,3, aujourd'hui, l'illusion, jusqu'à, Jean-Louis, donne-t-il, États-Unis, France Inter...*

# Analyses TAL - niveau du mot

## Lemmatisation/Racinisation

### Lemmatisation

- trouver forme canonique ou lemme (🇺🇸 lemma) d'un mot
  - chevaux → cheval
  - penseras → penser

### Racinisation (🇺🇸 stemming)

- trouver racine (🇺🇸 stem) d'un mot
  - chevaux, cheval → cheva-
  - penseras, pensais → pens-

# Analyses TAL - niveau du mot

## Étiquetage

Étiquetage morpho-syntaxique/en parties du discours (🇺🇸 POS tagging)

- attribuer une catégorie grammaticale aux mots

Je/PRON suis/VERB content/ADJ que/SCONJ ce/PRON soit/VERB  
clair/ADJ entre/ADV nous/PRON ./PUNCT

étiquettes **Universal POS tags**

# Analyses TAL - niveau du mot

## Désambiguïsation

Désambiguïsation du sens (🇺🇸 Word Sense Disambiguation ou WSD)

- rattacher une occurrence d'un mot à un sens donné (ou pas)

fenêtre = menuiserie ou interface graphique ?

# Analyses TAL - niveau du mot

## Désambiguïsation

### Désambiguïsation du sens (🇺🇸 Word Sense Disambiguation ou WSD)

- rattacher une occurrence d'un mot à un sens donné (ou pas)

fenêtre = menuiserie ou interface graphique ?

- **S: (n) window** (a framework of wood or metal that contains a glass windowpane and is built into a wall or roof to admit light or air)
- **S: (n) window** (a transparent opening in a vehicle that allow vision out of the sides or back; usually is capable of being opened)
- **S: (n) window** (a transparent panel (as of an envelope) inserted in an otherwise opaque material)
- **S: (n) window** (an opening that resembles a window in appearance or function) "*he could see them through a window in the trees*"
- **S: (n) window** (the time period that is considered best for starting or finishing something) "*the expanded window will give us time to catch the thieves*"; "*they had a window of less than an hour when an attack would have succeeded*"
- **S: (n) windowpane, window** (a pane of glass in a window) "*the ball shattered the window*"
- **S: (n) window** (an opening in a wall or screen that admits light and air and through which customers can be served) "*he stuck his head in the window*"
- **S: (n) window** ((computer science) a rectangular part of a computer screen that contains a display different from the rest of the screen)

# Analyses TAL - niveau du mot

## Plongements lexicaux

### Plongements lexicaux (🇺🇸 word embeddings)

- représenter les (formes de) mots dans un espace continu de (relativement) faible dimension et (qu'on espère) sémantiquement pertinent



(Ghannay et al., 2015)

# Analyses TAL - niveau de la phrase

## Segmentation

### Segmentation en phrase (🇺🇸 sentence segmentation)

- délimiteurs de phrases ambigus
- problème de la définition d'une phrase
  - listes à puces
  - citations
  - ...



# Analyses TAL - niveau de la phrase

Analyse syntaxique (🇺🇸 parsing)

## Chunking

- identifier les frontières de groupes

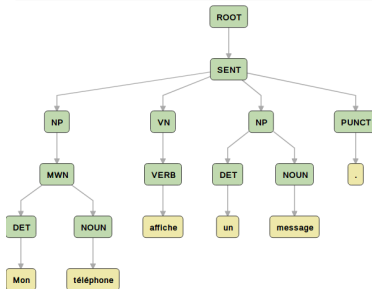
[Il] [le] [supplie] [de garder] [son nez] [anormalement grand] [en dehors]  
[des affaires] [des autres].

# Analyses TAL - niveau de la phrase

Analyse syntaxique (🇺🇸 parsing)

## Analyse en constituants

- identifier les groupes et les relations entre eux



Grammar	Lexicon
$S \rightarrow NP VP$	<i>Det</i> → that   this   the   a
$S \rightarrow Aux NP VP$	<i>Noun</i> → book   flight   meal   money
$S \rightarrow VP$	<i>Verb</i> → book   include   prefer
$NP \rightarrow Pronoun$	<i>Pronoun</i> → I   she   me
$NP \rightarrow Proper-Noun$	<i>Proper-Noun</i> → Houston   NWA
$NP \rightarrow Det Nominal$	<i>Aux</i> → does
$Nominal \rightarrow Noun$	<i>Preposition</i> → from   to   on   near   through
$Nominal \rightarrow Nominal Noun$	
$Nominal \rightarrow Nominal PP$	
$VP \rightarrow Verb$	
$VP \rightarrow Verb NP$	
$VP \rightarrow Verb NP PP$	
$VP \rightarrow Verb PP$	
$VP \rightarrow VP PP$	
$PP \rightarrow Preposition NP$	

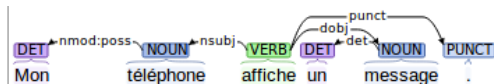
Speech and Language Processing. Daniel Jurafsky et James H. Martin

# Analyses TAL - niveau de la phrase

Analyse syntaxique (🇺🇸 parsing)

## Analyse en dépendances

- identifier les relations entre les mots



# Analyses TAL - niveau sous-phrastique

## Reconnaissance d'entités nommées

### Reconnaissance d'entités nommées (🇺🇸 named entity recognition)

- typiquement noms de personnes, lieux (pays, villes, régions...), organisations (entreprises, universités, ONG...)

Le footballeur ivoirien **Yaya Touré**<sub>PERSON</sub> a déclaré que les joueurs noirs pourraient boycotter la Coupe du monde en **2018**<sub>DATE</sub> si la **Russie**<sub>LOCATION</sub> n'aborde pas le problème du racisme dans les stades de football.

# Analyses TAL - niveau sous-phrastique

## Extraction de relations

### Extraction de relations

- relations issues de campagnes d'évaluation ou de bases de connaissances

author of

Pythagore prônant le végétarisme est une peinture de Pierre Paul Rubens

# Analyses TAL - niveau textuel

## Résolution de coréférence

### Résolution de coréférence

- Relier chaque mention d'une entité à l'entité à laquelle elle fait référence

**Sophie Marceau** est une actrice et réalisatrice française. **Elle** a été révélée à l'âge de 14 ans par le film La Boum, qui **lui** a permis de devenir d'emblée une vedette du cinéma français. Avec La Boum 2, **elle** obtient le César du meilleur espoir féminin. Le travail de **la comédienne** avec Mel Gibson sur l'épopée guerrière Braveheart en 1995 **lui** ouvre en grand les portes de Hollywood et marque le début d'une carrière internationale.

# Le TAL actuellement

- Approches robustes
  - grande quantité de textes
  - approches statistiques
- Évaluation
- Disponibilité d'outils et ressources
  - nombreuses langues mais anglais prévalent

# Technologies de la langue

## mostly solved

### Spam detection

Let's go to Agral 

Buy VIAGRA ... 

### Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.


### Named entity recognition (NER)


PERSON ORG LOC

Einstein met with UN officials in Princeton

## making good progress

### Sentiment analysis


Best roast chicken in San Francisco! 

The waiter ignored us for 20 minutes. 


### Coreference resolution

Carter told Mubarak he shouldn't run again. 


### Word sense disambiguation (WSD)

I need new batteries for my *mouse*. 

### Parsing


I can see Alcatraz from the window! 

### Machine translation (MT)

SP13GG I-PYTHOGE-4408 Y7 FF 88... 

The 13<sup>th</sup> Shanghai International Film Festival...

### Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30 

Party  
May 27  
add

## still really hard

### Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

### Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

### Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is good 

### Dialog

Where is Citizen Kane playing in SF? 

Castro Theatre at 7:30. Do you want a ticket?

Dan Jurafsky, *Introduction to NLP*