

# Modèles de Linguistique Computationnelle

## CM 1 : Introduction

### M1 Langue et Informatique

Crédits: Gaël Lejeune, Karĕn Fort, Iana Atanassova, Djame Seddah, Eleni Kogkitsidou, Olga Seminck

---

Alice Millour prenom.nom@sorbonne-universite.fr

Yoann Dupont prenom.nom@sorbonne-universite.fr

2020-2021

Sorbonne-Université

# Plan

Introduction à la linguistique computationnelle

Notions de bash

Rappels de python

- Globalement :
  - 13 séances, dont 1 de révision
  - 6 cours / enseignant
  - base de 1h CM / 1h30 TD
  - identifiant cours sur Moodle : M1SOL030
- Évaluation :
  - contrôle continu
  - contrôle terminal

# Notions vues dans le cours

- aujourd'hui : introduction à bash et rappels de python
- expressions régulières (recherche de motifs)
- formats de corpus et transformation
- normalisation de corpus textuels
- classification (de textes, de mots)
- reconnaissance d'entités nommées
- évaluation d'outils
- éventail des tâches de TAL

permet de résoudre **une tâche** grâce à une **méthode**

Pendant ce cours : on va explorer plusieurs tâches et plusieurs méthodes

Vos idées ?

## tâches « visibles »

- moteurs de recherche
- traduction automatique (par exemple : Google translate)
- correction orthographique
- suggestion automatique
- transcription automatique de la parole
- systèmes de dialogue (écrit / oral)
- etc.

tâches « invisibles » (ou sous-jacentes)

## Tour de France 2020 : peur sur les descentes

Si les ascensions et les arrivées au sommet font le sel de la Grande Boucle, les descentes jouent un rôle important dans les étapes, comme lors de la 16e, mardi. Elles tétanisent certains coureurs.

- **segmentation** (<sub>Si</sub><sub>les</sub><sub>ascensions</sub><sub>et</sub><sub>les</sub>)



tâches « invisibles » (ou sous-jacentes)

## Tour de France 2020 : peur sur les descentes

Si les ascensions et les arrivées au sommet font le sel de la Grande Boucle, les descentes jouent un rôle important dans les étapes, comme lors de la 16e, mardi. Elles tétanisent certains coureurs.

- **segmentation** (`Si les ascensions et les`)
- **lemmatisation** (le lemme de `descentes` est « `descente` », de « `font` » est `faire`)

tâches « invisibles » (ou sous-jacentes)

## Tour de France 2020 : peur sur les descentes

Si les ascensions et les arrivées au sommet font le sel de la Grande Boucle, les descentes jouent un rôle important dans les étapes, comme lors de la 16e, mardi. Elles tétanisent certains coureurs.

- **segmentation** ( `Si les ascensions et les` )
- **lemmatisation** (le lemme de `descentes` est « `descente` », de « `font` » est `faire` )
- **analyse grammaticale** ( « `Si` » = `conjonction`, « `les` » = `déterminant` )

tâches « invisibles » (ou sous-jacentes)

## Tour de France 2020 : peur sur les descentes

Si les ascensions et les arrivées au sommet font le sel de la Grande Boucle, les descentes jouent un rôle important dans les étapes, comme lors de la 16e, mardi. Elles tétanisent certains coureurs.

- **segmentation** ( $\_Si\_les\_ascensions\_et\_les\_$ )
- **lemmatisation** (le lemme de **descentes** est « **descente** », de « **font** » est **faire**)
- **analyse grammaticale** ( « **Si** » = **conjonction**, « **les** » = **déterminant**)
- **analyse syntaxique**  
( « **les ascensions et les arrivées au sommet** » est **sujet** de « **sont** » )

tâches « invisibles » (ou sous-jacentes)

## Tour de France 2020 : peur sur les descentes

Si les ascensions et les arrivées au sommet font le sel de la Grande Boucle, les descentes jouent un rôle important dans les étapes, comme lors de la 16e, mardi. Elles tétanisent certains coureurs.

- **segmentation** (`Si les ascensions et les`)
- **lemmatisation** (le lemme de `descentes` est « `descente` », de « `font` » est `faire`)
- **analyse grammaticale** ( « `Si` » = `conjonction`, « `les` » = `déterminant`)
- **analyse syntaxique**  
( « `les ascensions et les arrivées au sommet` » est `sujet` de « `sont` » )
- **reconnaissance d'entités nommées** ( « `Tour de France 2020` » est une entité qu'on peut catégoriser comme « compétition sportive » )

tâches « invisibles » (ou sous-jacentes)

## Tour de France 2020 : peur sur les descentes

Si les ascensions et les arrivées au sommet font le sel de la Grande Boucle, les descentes jouent un rôle important dans les étapes, comme lors de la 16e, mardi. Elles tétanisent certains coureurs.

- **segmentation** ( $\_Si\_les\_ascensions\_et\_les\_$ )
- **lemmatisation** (le lemme de **descentes** est « **descente** », de « **font** » est **faire**)
- **analyse grammaticale** ( « **Si** » = **conjonction**, « **les** » = **déterminant**)
- **analyse syntaxique**  
( « **les ascensions et les arrivées au sommet** » est **sujet** de « **sont** » )
- **reconnaissance d'entités nommées** ( « **Tour de France 2020** » est une entité qu'on peut catégoriser comme « compétition sportive »)
- **identification d'anaphores** ( « **elles** » réfère à « **les descentes** »)

tâches « invisibles » (ou sous-jacentes)

## Tour de France 2020 : peur sur les descentes

Si les ascensions et les arrivées au sommet font le sel de la Grande Boucle, les descentes jouent un rôle important dans les étapes, comme lors de la 16e, mardi. Elles tétanisent certains coureurs.

- **segmentation** ( $\_Si\_les\_ascensions\_et\_les\_$ )
- **lemmatisation** (le lemme de *descentes* est « *descente* », de « *font* » est *faire*)
- **analyse grammaticale** ( « *Si* » = *conjonction*, « *les* » = *déterminant*)
- **analyse syntaxique**  
( « *les ascensions et les arrivées au sommet* » est *sujet* de « *sont* » )
- **reconnaissance d'entités nommées** ( « *Tour de France 2020* » est une entité qu'on peut catégoriser comme « compétition sportive » )
- **identification d'anaphores** ( « *elles* » réfère à « *les descentes* » )
- **désambiguïsation lexicale** ( « *Tour* » correspond au sens 3/B./1./a.) de la page tour du CNRTL : « *Mouvement, déplacement (à peu près) circulaire où l'on revient au point de départ.* ».)

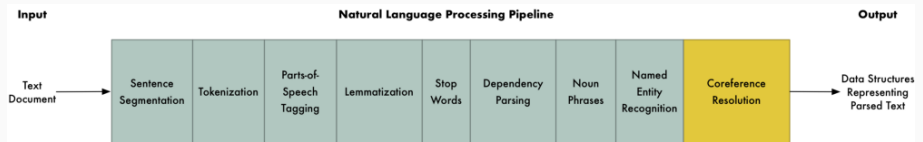
tâches « invisibles » (ou sous-jacentes)

## Tour de France 2020 : peur sur les descentes

Si les ascensions et les arrivées au sommet font le sel de la Grande Boucle, les descentes jouent un rôle important dans les étapes, comme lors de la 16e, mardi. Elles tétanisent certains coureurs.

- **segmentation** ( `Si` `les` `ascensions` `et` `les` )
- **lemmatisation** (le lemme de `descentes` est « `descente` », de « `font` » est `faire` )
- **analyse grammaticale** ( « `Si` » = `conjonction`, « `les` » = `déterminant` )
- **analyse syntaxique**  
( « `les ascensions et les arrivées au sommet` » est `sujet` de « `sont` » )
- **reconnaissance d'entités nommées** ( « `Tour de France 2020` » est une entité qu'on peut catégoriser comme « compétition sportive » )
- **identification d'anaphores** ( « `elles` » réfère à « `les descentes` » )
- **désambiguïsation lexicale** ( « `Tour` » correspond au sens 3/B./1./a.) de la page tour du CNRTL : « *Mouvement, déplacement (à peu près) circulaire où l'on revient au point de départ.* ».)
- etc.

# Exemple de chaîne de traitement



Source : [https:](https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e)

[//medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e](https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e)



# Linguistique computationnelle : quelques exemples

- traduction automatique : [google](#) vs [bing](#)
- analyse de sentiments : [lien](#)
- reconnaissance d'entités nommées : [lien](#)
- analyse syntaxique : [lien](#)
- extraction d'informations (questions / réponses) : [lien](#)
- génération de texte : [lien](#)

permet de résoudre **une tâche** grâce à une **méthode**

Vos idées ?

- approches par règles
- approches par apprentissage : *Machine learning*
  - apprentissage « traditionnel » statistique
  - approches neuronales (*Deep learning*)

diffèrent (notamment) par les **ressources** qu'elles requièrent

- approches par règles  $\Rightarrow$  règles écrites par des linguistes
- approches par apprentissage : *Machine learning*
  - apprentissage « traditionnel » statistique  $\Rightarrow$  ressources d'apprentissage
  - approches neuronales (*Deep learning*)  $\Rightarrow$  ressources d'apprentissage de très grande taille

le big data diffèrent (notamment) par les ressources qu'elles requièrent

permet de résoudre **une tâche** grâce à une **méthode**

permet de résoudre **une tâche** grâce à une **méthode**

en exploitant des **ressources linguistiques**

Introduction à la linguistique computationnelle

Notions de bash

Rappels de python



# Pourquoi utiliser bash ?

Bash = un langage intégré à linux qui permet (entre autres) :

- de naviguer dans le système de fichiers
- de visualiser / modifier les fichiers

⇒ utile pour manipuler les fichiers / ressources linguistiques utilisées en modélisation linguistique

```
$ commande -option1 -option2 -option3 argument1 argument2
```

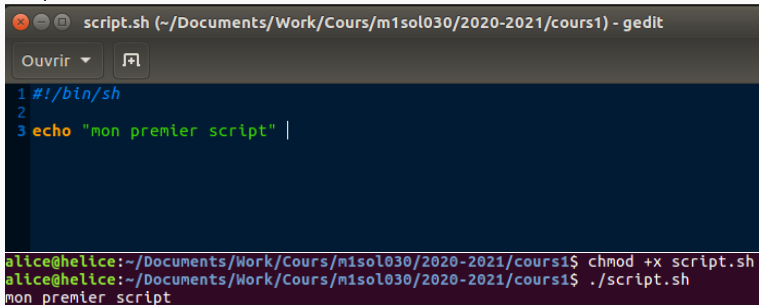
# ligne de commande vs. script

deux modes d'utilisation :

- ligne de commande :

```
alice@helice:~/Documents/Work/Cours/misol030/2020-2021/cours1$ echo "Hello world !"
Hello world !
```

- script :



The screenshot shows a terminal window titled "script.sh (~/Documents/Work/Cours/misol030/2020-2021/cours1) - gedit". The window has a menu bar with "Ouvrir" and a file icon. The editor shows three lines of code:   
1 `#!/bin/sh`  
2  
3 `echo "mon premier script" |`  
Below the editor, the terminal shows the following commands and output:  
`alice@helice:~/Documents/Work/Cours/misol030/2020-2021/cours1$ chmod +x script.sh`  
`alice@helice:~/Documents/Work/Cours/misol030/2020-2021/cours1$ ./script.sh`  
`mon premier script`

## Quelques commandes usuelles....

- se repérer dans la hiérarchie de fichiers avec **pwd**, se déplacer avec **cd**
- manipuler des répertoires et fichiers : créer un répertoire avec **mkdir**, un fichier avec **touch**, supprimer avec **rm**, déplacer avec **mv**
- lister les fichiers d'un répertoire avec **ls**
- afficher le contenu d'un fichier, avec **cat**, **less**, **head**, **tail**
- faire des modifications dans un fichier avec **tr**
- lire sur l'entrée standard avec **read**

remarque : `".."` désigne le dossier « parent » du dossier courant.

Si vous êtes dans `/home/utilisateurs/Documents` le dossier `..` correspond à `/home/utilisateurs/`

# arguments, options, entrées, sorties

une commande peut avoir 0, 1 ou plus arguments, lire sur l'entrée standard, afficher (ou non) un résultat sur la sortie standard, changer de comportement en fonction des options...

- `ls` vs `ls ..` (ajout d'un argument)
- `ls` vs `ls -l` (ajout d'une option)
- `ls` vs `ls -l ..` (ajout d'une option et d'un argument)

## Redirection : >, >> et |

On peut rediriger la sortie standard vers :

- **un fichier en l'écrasant** avec >
- **un fichier en concaténant** la sortie au contenu existant avec >>

essayez :

```
$ touch fichier.txt
$ echo "première phrase"
$ echo "première phrase" > fichier.txt
$ cat fichier.txt
$ echo "première phrase" >> fichier.txt
$ cat fichier.txt
$ echo "deuxième phrase" >> fichier.txt
$ cat fichier.txt
$ echo "troisième phrase" > fichier.txt
$ cat fichier.txt
```

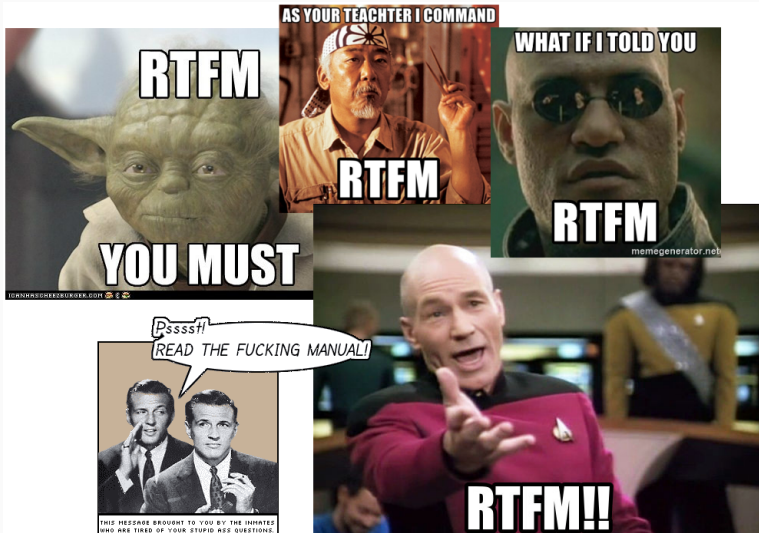
## Redirection : >, >> et |

On peut rediriger la sortie standard vers :

- **une nouvelle commande** avec | (voir TD1)

# Comment on s'en sort ? : la commande man

« Read The **Fine** Manual »





# Exemple : apprendre à utiliser la commande cp

\$ man cp

```
CP(1) User Commands
NAME
    cp - copy files and directories

SYNOPSIS
    cp [OPTION]... [-I] SOURCE DEST
    cp [OPTION]... SOURCE... DIRECTORY
    cp [OPTION]... -t DIRECTORY SOURCE...

DESCRIPTION
    Copy SOURCE to DEST, or multiple SOURCE(s) to DIRECTORY.

    Mandatory arguments to long options are mandatory for short options too.

    -a, --archive
        same as -dR --preserve=all

    --attributes-only
        don't copy the file data, just the attributes

    --backup[=CONTROL]
        make a backup of each existing destination file

    -b
        like --backup but does not accept an argument
```

\$ man man

À quoi servent les commandes :

- `wc`
- `grep`
- `sort`
- `uniq`

Combien d'arguments prennent les commandes :

- `pwd`
- `cp`
- `grep`

Quelles options utiliser :

- pour afficher sur la sortie standard un texte dont les lignes ont été classées par ordre **décroissant** (sort)
- pour afficher sur la sortie standard un texte dont les lignes ont été classées **aléatoirement** (sort)
- pour lister les fichiers du dossier courant en affichant leurs tailles (ls)
- pour lister les fichiers du dossier courant en affichant leurs tailles de manière « lisible par un humain » (ls)

à quoi sert la commande `cut` ?

que renvoie :

```
$ cut -d';' -f2 file.txt
```

si le fichier file.txt contient :

```
Prenom;Nom;Email;Age;Ville
```

```
Judith;Dreyfus;judith@exemple.com;31;Marseille
```

```
Ariane;Delaubier;ariane@exemple.com;28;Dieppe
```

```
Laura;Capitaine;laura@exemple.com;27;Paris
```

# Plan

Introduction à la linguistique computationnelle

Notions de bash

Rappels de python

Vos questions la semaine prochaine :

- ouverture / fermeture de fichiers
- formatage des chaînes de caractères
- boucles
- fonctions (différence entre print et return ?)
- listes, dictionnaires et fonctions associées