

# Traitement automatique de la langue

## Principaux concepts et méthodologies

Xavier Tannier

xavier.tannier@sorbonne-universite.fr



Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions

CC BY-NC-SA



ETAL, Lannion, 14-18 juin 2021

# Quel est le problème ?

Pourquoi le langage naturel est-il différent des autres données ?

# Le langage est redondant

**Plus de 700 000 personnes ont suivi l'explosion d'une pastèque en direct sur Facebook**

**L'explosion d'une pastèque en direct fait le buzz**

**CETTE PASTÈQUE EST DEVENUE CÉLÈBRE GRÂCE À FACEBOOK LIVE**

**VIDEO. Facebook Live: Plus de 800.000 personnes regardent en direct l'explosion d'une pastèque**

**L'explosion d'une pastèque fascine le web – Vidéo**

**L'explosion d'une pastèque sur Facebook suivie par 700.000 personnes**

# Le langage est redondant



...nnes ont suivi  
...ue en direct sur  
...K

L'e  
pa  
bu

VENUE

VIDEO. Facebook Live  
personnes regardent  
d'une pastèque

L'e  
le



ne

L'explosion d'une pastèque  
suivie par 700.000 personnes

# Le langage est ambigu

YAHOO!  
FRANCE



Rechercher

Web

**Images**

Vidéo

Actualités

Shopping





# Le langage est ambigu, implicite

*Jean vend une tarte **aux pommes**.*

*Jean vend une tarte **aux clients**.*

Si votre bébé ne supporte pas le lait cru,  
faites-le bouillir



*Si le vice-chancelier invite Simone de Beauvoir, il regrettera d'avoir une féministe à sa table.*

*Si le vice-chancelier invite le président des États-Unis, il regrettera d'avoir une féministe à sa table.*

# Le langage est implicite



Le Monde

## Un Caravage a-t-il été découvert dans un grenier en France

?

Le Monde - il y a 1 heure



Le tableau représentant la décapitation d'un prisonnier dans un grenier de la région de Toulouse. Studio SEBERT. Un tableau découvert dans un grenier près de Toulouse et présenté comme une œuvre du Caravage...

Connaissance du monde

Métonymie



Challenges.fr

## Primaire républicaine: les anti-Trump peuvent-ils encore le stopper?

Challenges.fr - il y a 1 heure



Manoeuvres dans les Etats, entourloupes sur le terrain, recherche d'un candidat qui serait pas le milliardaire... Les "tout-à-fait" de la primaire de l'arrêter? De l'Orlando, le 5 mars. Joe Raedle/Getty Images

Connaissance du contexte



## 14 mai 1610 : Henri IV est assassiné | L'Histoire

[www.histoire.presse.fr/.../14-mai-1610-henri-iv-est-assassine](http://www.histoire.presse.fr/.../14-mai-1610-henri-iv-est-assassine) 14-05-201...

14 mai 2012 - Le 14 mai 1610, alors qu'il se rend à l'Arsenal pour s'entretenir avec le duc de Sully, Henri IV est...

Déduction (présupposition)

# Le langage est naturel

Ironie

Mensonges

Humour

Sarcasme

Néologismes

Synonymie contextuelle

Paraphrase

Polysémie

Niveaux de langage

Multilinguisme

Périphrase

Ellipses

Anaphores

Sentiment

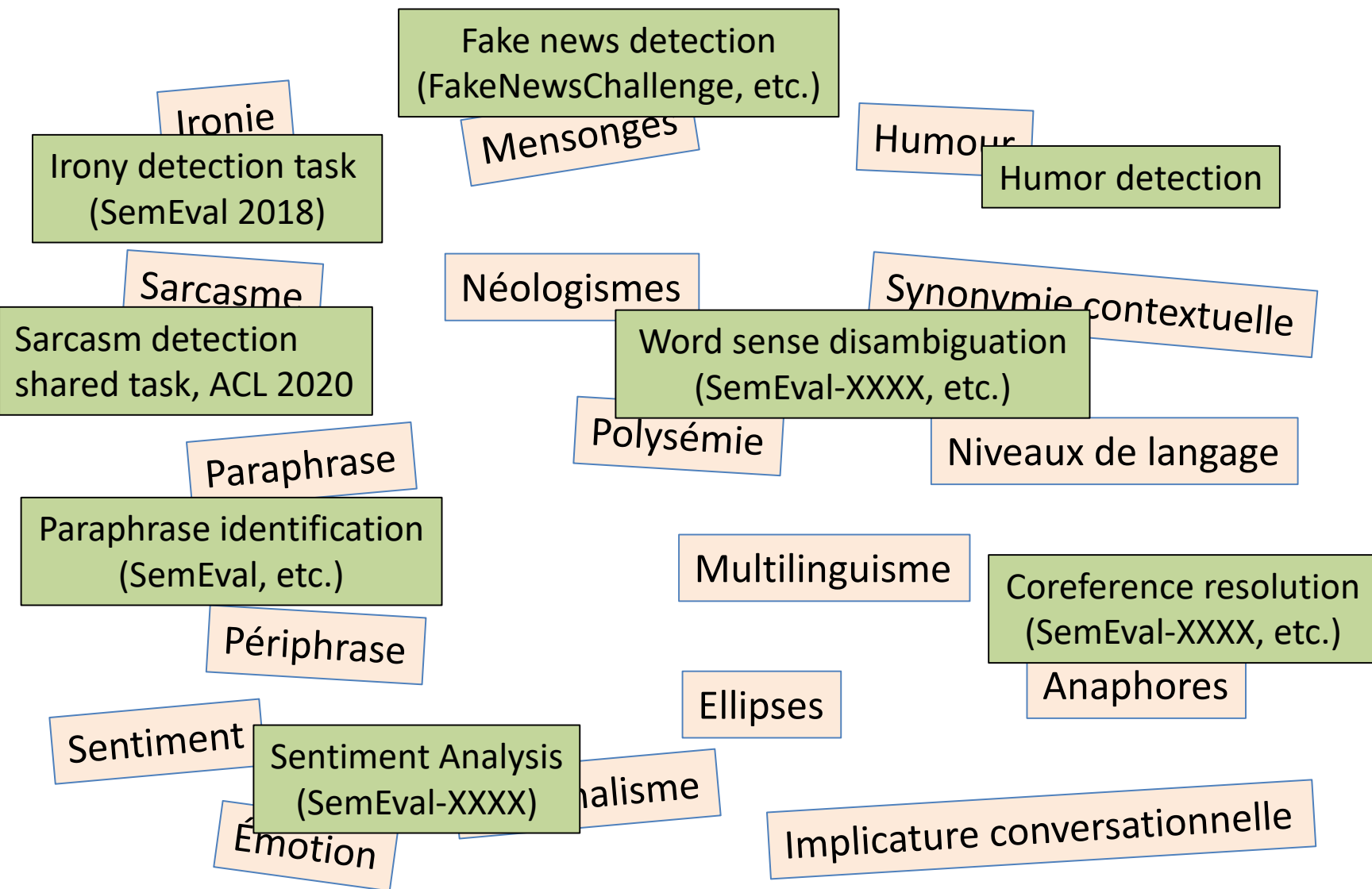
Régionalisme

Implicature conversationnelle

Émotion



# Le langage est naturel



# Le langage est naturel

Fake news detection  
(FakeNewsChallenge, etc.)

Ironie

Irony detection task  
(SemEval 2018)

Mensonges

Humour

Humor detection

Sarcasme

Néologismes

Synonymie

Sarcasm detection  
shared task, A...

Tous les phénomènes linguistiques sont interdépendants  
et la séparation en tâches n'est qu'une façon d'avancer  
sans devoir se préoccuper de tout à la fois...

Paraphrase

Paraphrase identification  
SemEval, etc.

Multilinguisme

Coreference resolution  
SemEval-XXXX, etc.

Périphrase

Ellipses

Anaphores

Sentiment

Sentiment Analysis  
(SemEval-XXXX)

Analisme

Émotion

Implicature conversationnelle

# La langue est vivante

Arthur Rimbaud  
Honte (19<sup>ème</sup> siècle)

*Tant que la lame n'aura  
Pas coupé cette cervelle,  
Ce paquet blanc, vert et gras,  
A vapeur jamais nouvelle,*

*(Ah ! Lui, devrait couper son  
Nez, sa lèvre, ses oreilles,  
Son ventre ! et faire abandon  
De ses jambes ! ô merveille !)*

*Mais non ; vrai, je cr  
Que pour sa tête la l  
Que les cailloux pou  
Que pour ses boyaux*

SMS 12 ans

Cc alor ta dmende pour ce  
soir

Jen sai rien chui pa aller ché  
le toubib

Ouai jen ai jamai vu des  
vivan mes kan il sn mor ca  
mderange pa

Noé t c se kil i a fair en  
techno

tu pe pa timaginer a quel  
poin il me manque <3

Quant l'Emperere ad faite sa justice  
E esclargiée est la sue granz ire,  
En Bramimunde ad chrestientet mise.  
Passet li jurz, la nuiz est aserie,  
Culchet s'est li Reis en sa cambre voltice.  
Seinz Gabriel de part Deu li vint dire :  
« Carle, semun les oz de tun empire,  
« Par force iras en la tere de Bire,  
« Rei Vivien si succurras en Imphe,  
« A la citet que païen unt asise.  
« Li chrestien te recleiment e crient. »  
Li Emperere n'i volsist aler mie :  
« Deus ! dist li Reis, si penuse est ma vie ! »

Chanson de Roland  
(11<sup>ème</sup> siècle)

Allocution présidentielle  
(21<sup>ème</sup> siècle)



Donald J. Trump ✓  
@realDonaldTrump

Despite the constant negative press covfefe

RETWEETS 125,672  
LIKES 160,178



12:06 AM - 31 May 2017

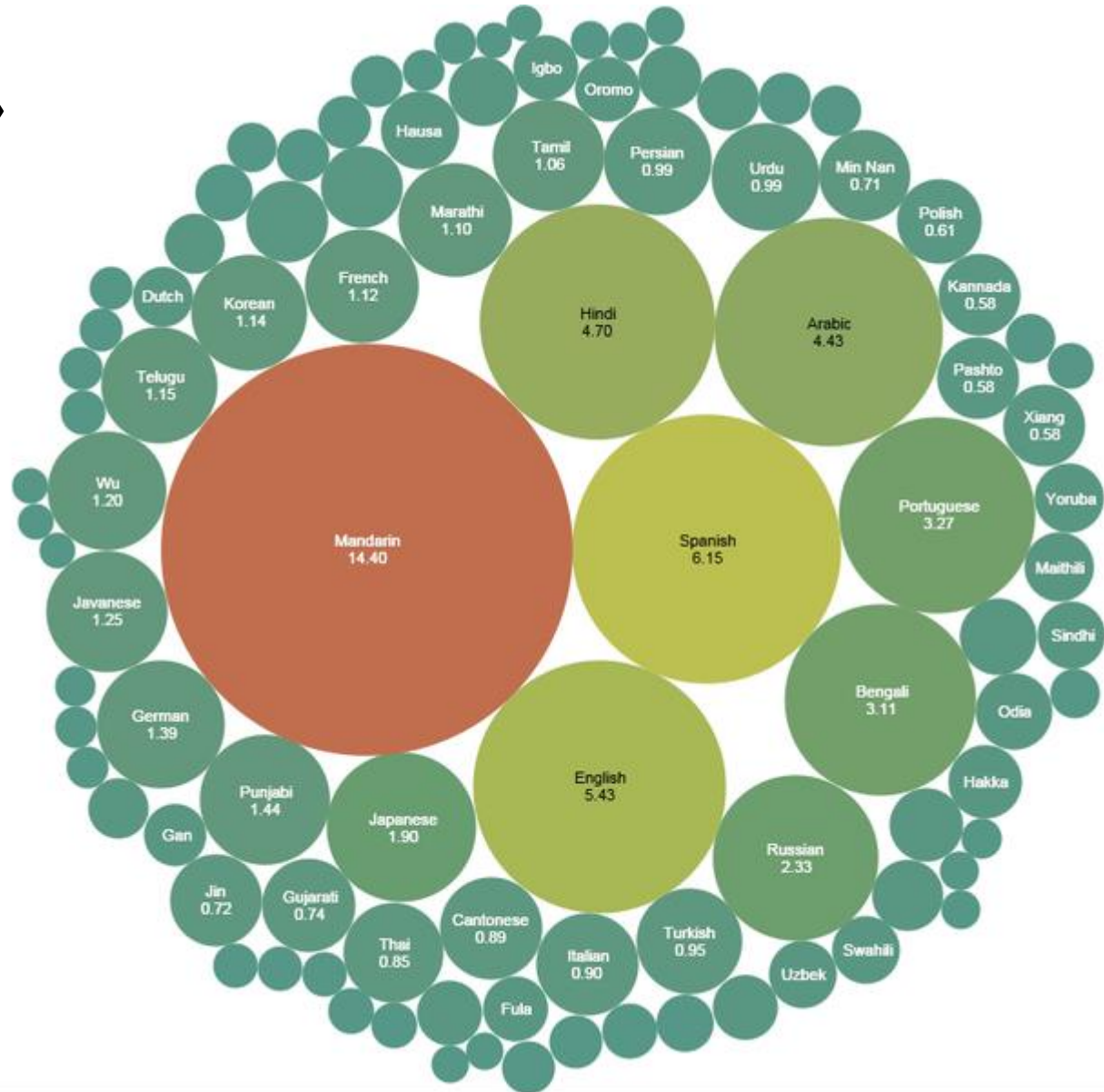
41K 126K 160K

# Les langues sont nombreuses

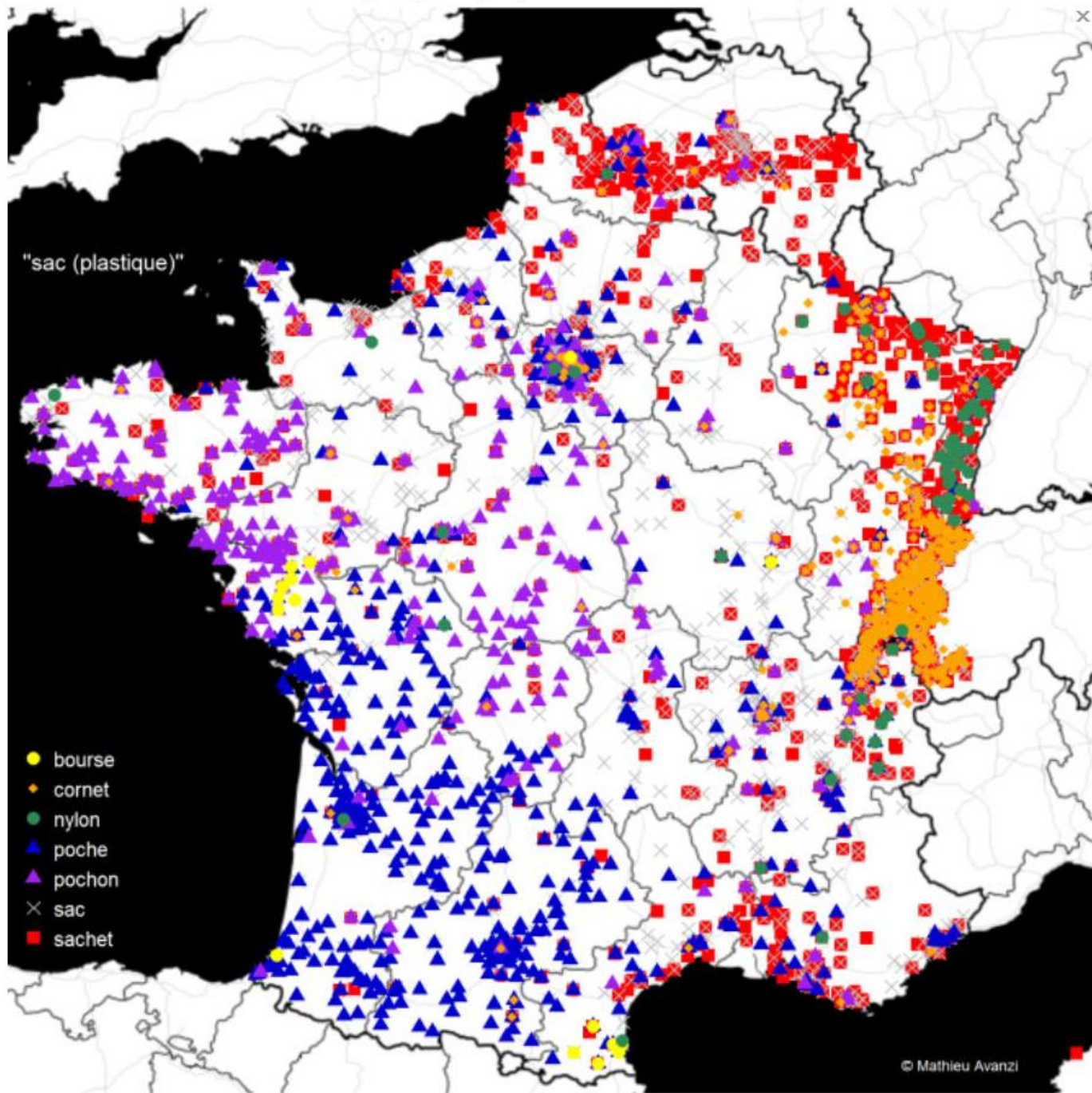
... et souvent « peu dotées »  
(*low-resource*)



Pieter Bruegel l'Ancien,  
La Grande Tour de Babel  
1563





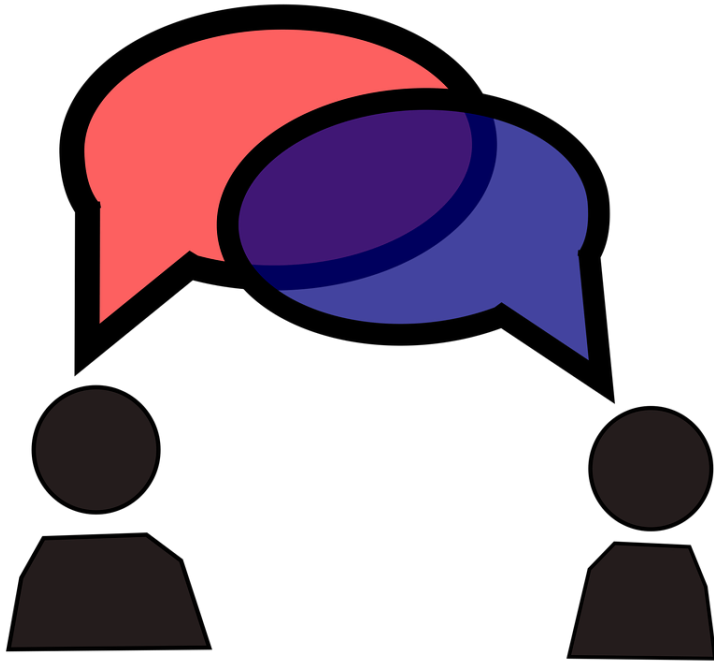


# La langue est multimodale





# Le langage est une interaction



vous avez, il y a une date de limite je pense,  
pour accéder au bureau quand on est au C.A.  
il y a une... on va dire une date  
oui, oui.  
par rapport à l'A.G.  
oui il faut il faut faire partie de de la  
ah non non absolument pas, tu peux être  
du C.A. ou non vous n'avez pas ça dans les  
statuts, parce que dans certains  
je pense pas  
statuts il doit y avoir au moins six mois

La transcription automatique : un rêve enfin accessible ?

Elise Tancoigne, Jean-Philippe Corbellini, Gaëlle Deletraz, Laure Gayraud,

Sandrine Ollinger, Daniel Valéro

Sept. 2020

# La communication n'est pas que verbale



« La République, c'est moi ! C'est moi qui suis parlementaire ! »

Source: Le Parisien, 17 octobre 2018

<https://www.leparisien.fr/politique/jean-luc-melenchon-le-cout-de-la-colere-17-10-2018-7921920.php>

# Le langage peut être utilisé pour mentir (ou exagérer, induire en erreur...)

LES DÉCODEURS

## « Gilets jaunes » : « Constitution disparue », complotisme débridé sur les réseaux sociaux

Derrière une théorie très partagée par les « gilets jaunes », qui postule que la Constitution a disparu depuis un an, se trouve Serge Petitdemange, septuagénaire youtubeur aux accents conspirationnistes.

Par Samuel Laurent • Publié aujourd'hui à 12h20

CHRISTOPHE ASSELIN - JANV. 26, 2021

## Vaccins et Covid : les Fake News les plus fréquentes sur les médias sociaux en France

Pour 17% des français, le coronavirus a été créé **intentionnellement** en laboratoire. Les théories du complot et la désinformation, générant de la méfiance à l'égard des vaccins contre la COVID-19 pourraient conduire à des taux de vaccinations en dessous des 55% nécessaires pour obtenir une immunité collective aux États-Unis et en Grande-Bretagne selon une étude de la London School of Hygiene & Tropical Medicine <sup>(1)</sup>. Et en France ? En novembre, 54% des Français déclaraient qu'ils se feraient vacciner, les plaçant en tête des pays les plus réticents. Les fake news sur la Covid et le coronavirus sont légions sur le web et les réseaux sociaux depuis le

ACCUEIL > MONDE

### VIDÉO. Donald Trump : Ses dix «fake news» les plus marquantes depuis son élection

FAKE OFF Le président des Etats-Unis aime beaucoup les « faits alternatifs »...

Mathilde Cousin | Publié le 20/01/18 à 10h45 — Mis à jour le 20/01/18 à 10h45

2 COMMENTAIRES 316 PARTAGES

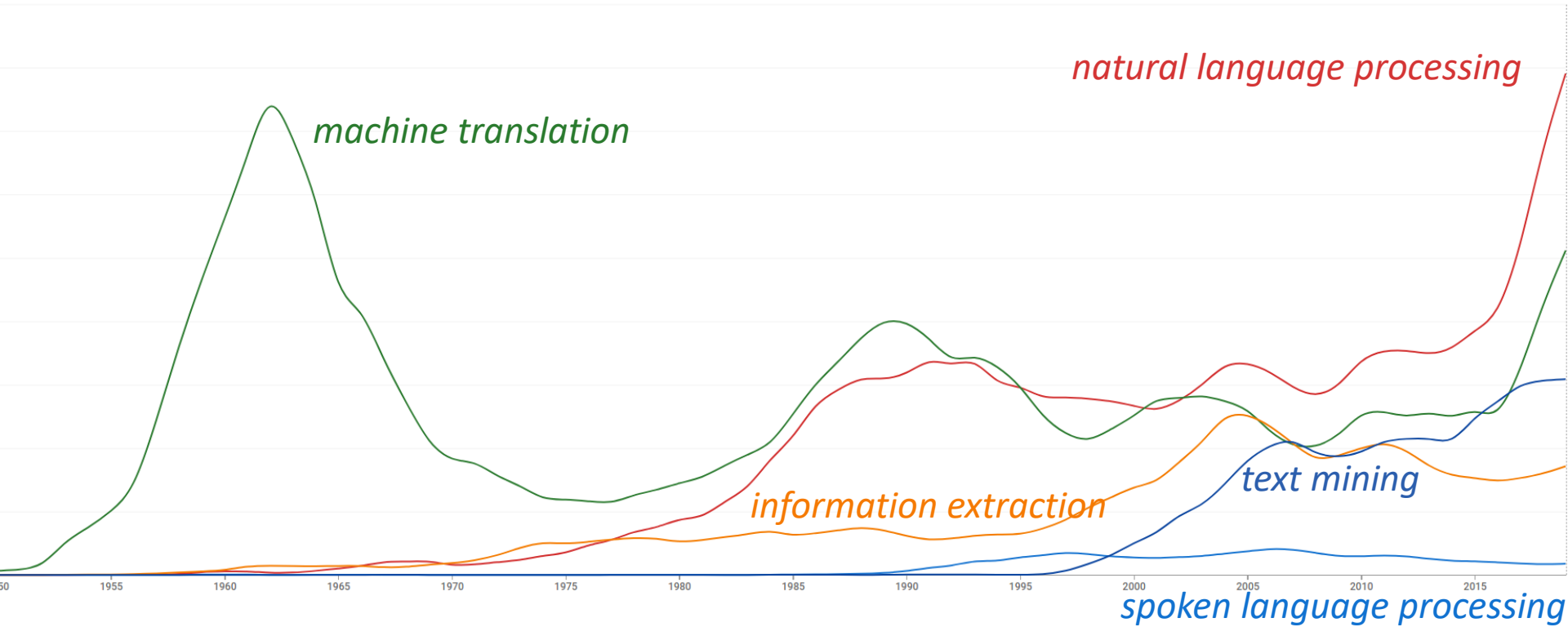


À LIRE AUSSI

- 26/10/17 | FAKE OFF  
Non, le drapeau européen ne va pas «remplacer» le drapeau français
- 06/11/17 | JAPON  
VIDEO. Japon: Trump en visite officielle, protocole et carpes Koi
- 06/11/17 | COREE DU NORD  
Donald Trump et Shinzo Abe affichent leur fermeté face à la Corée du Nord

D'ACTU

# Historique



Google Books Ngram Viewer

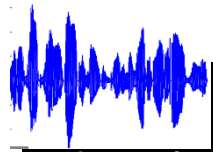
# Qu'est-ce que le TAL ?

~~Comprendre et générer du langage~~



Accomplir des tâches  
en lien avec le langage

# Niveaux d'analyse (image d'Épinal)



(parole)

phonétique, phonologique

Comment les mots et les phrases sont liés aux sons qui les réalisent à l'oral

(texte)

(extraction), tokenisation

Comment les mots sont construits et quels sont leurs rôles dans la phrase

morphologique

Comment les mots se combinent pour former des syntagmes, puis des propositions et enfin des phrases correctes

syntaxique

Comment les mots font du sens lorsqu'ils sont insérés dans une phrase (indépendamment du contexte)

sémantique

Comment les phrases peuvent être interprétées selon leur contexte d'énonciation (interlocuteurs, phrases précédentes, connaissance commune du monde, ...)

discours





# Transcription manuelle

Transcriber 1.5.1

Fichier Edition Signal Segmentation Options Aide

Serge Levallant  
alors+[i en fond] Don Camillo ; oui+[“ouais”] je comprends mieux pourquoi vous aviez honte :

Serge Levallant + Bernard Mabile  
1:[i] c'est ringard+[“mais pas du tout pas du tout”] , c'est ringard+[“c'est pas vrai”] au possible le Don Camillo : c'est un truc pour touristes japonais et américains !  
2:[bb] pourquoi ? mais pas du tout ! pourquoi ? bah pourquoi ? pas du tout ! absolument pas !

Bernard Mabile  
[i] +[rire en fond] nan j'aime bien moi , j'aime bien j'aime bien passer dans les dans les dîners-spectacles ça ça ça me plaît .  
[i] je ferai pas ça toute ma vie et tous les jours non plus mais j'aime bien ; j'aime bien l'ambiance , j'aime bien boire un verre euh  
[i] avec des copains , non j'aime bien .

Virginie Lemoine  
faut avoir [bb en fond-]des des spectacles très solides hein et des des numéros[-bb en fond]

Virginie Lemoine + Bernard Mabile  
1:très très solides hein pour [pron=pi]  
2:(il) y a plein de gens qui

Bernard Mabile  
qui qui passent là-bas , non c'est marrant quoi .

Serge Levallant  
comme disait une grande intellectuelle que j'ai rencontrée dans un dîner un un soir

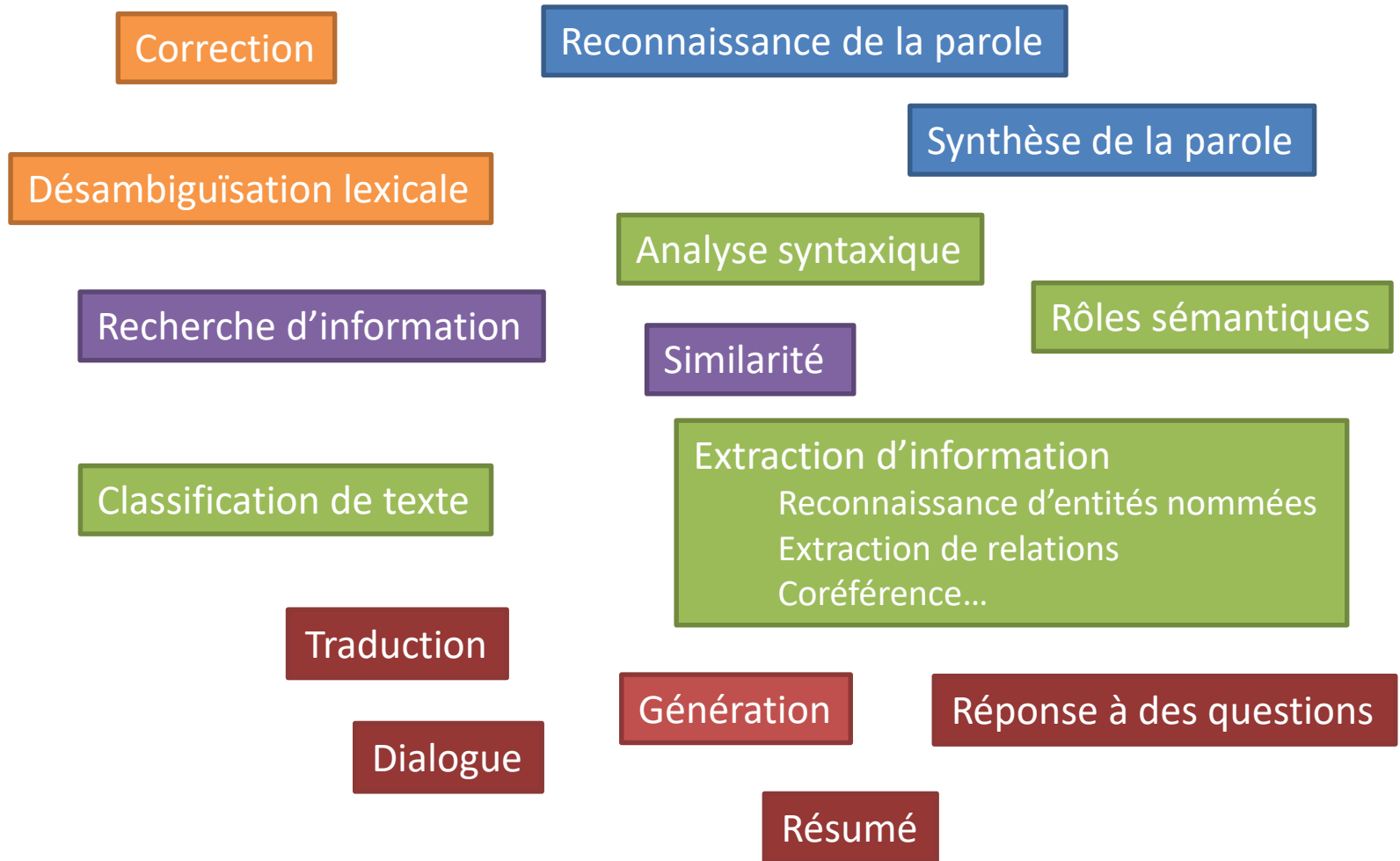
Serge Levallant + Bernard Mabile  
1:vraiment par hasard  
2:spectacle ? spectacle ?

Serge Levallant  
c'est "chansonnier"  
[silence]

20040909\_0455\_0905\_INTER\_ELDA



# Les grandes classes de tâches de TAL



# **Le TAL, un problème de représentation**

# Le mot

Rien

ne

sert

de

courir

il

faut

partir

à

point

# Le mot

Rien

ne

sert

de

courir

il

faut

partir

à

point

etc.  
T.A.L.  
21.3  
[www.sncf.com](http://www.sncf.com)

Jean-Louis  
donne-t-il  
1914-1918  
06-13-23-33-12

l'illusion  
aujourd'hui  
jusqu'à

Les industries sont variées :  
imprimerie, agroalimentaire (jus  
de fruits, conserveries), arme-  
ment et défense, équipement  
électrique et électronique, chi-  
mie (engrais, matières plasti-  
ques), industries pharmaceuti-  
ques, textile et confection, in-  
dustries dérivées du bois (pa-  
peteries) et du tabac. Sur la  
côte orientale, à Cap Canave-  
ral, est installée le grand centre  
spatial de la NASA.

# Le mot

Rien

ne

sert

de

courir

il

faut

partir

à

point

etc.  
T.A.L.  
21.3  
www.sncf.com

Jean-Louis  
donne-t-il  
1914-1918  
06-13-23-33-12

l'illusion  
aujourd'hui  
jusqu'à

Les industries sont variées :  
imprimerie, agroalimentaire (jus  
de fruits, conserveries), arme-  
ment et défense, équipement  
électrique et électronique, chi-  
mie (engrais, matières plasti-  
ques), industries pharmaceuti-  
ques, textile et confection, in-  
dustries dérivées du bois (pa-  
peteries) et du tabac. Sur la  
côte orientale, à Cap Canave-  
ral, est installée le grand centre  
spatial de la NASA.

mot

token



# Le mot

**Lebensversicherungsgesellschaftsangestellter**

(langues **agglutinantes**)

(employé d'une compagnie d'assurance-vie)

學而不思則罔，思而不學則殆

東京、マドリード、イスタンブール(トルコ)が争う2020年夏季五輪の開催地は7日(日本時間8日)、ブエノスアイレスでの国際オリンピック委員会(IOC)総会で、IOC委員約100人の投票で決まる。

# Formes d'un mot, famille d'un mot

- **Flexion**

- Verbale : *montrer, montreras...*
- Nominale : *cheval, chevaux...*
- forme canonique (lemme) et formes fléchies

- **Dérivation**

- *penser/V + able = pensable*
- *in + pensable/A = impensable*
- base et dérivé

- **Composition**

- *appendice + ectomie = appendicectomie*
- éléments de formation, mot composé

# Lemmatisation

- Obtention de la **forme canonique** (le *lemme*) à partir du mot :
  - Pour un **verbe** : sa forme à l'infinif (sans les flexions)  
*montrer, montreras, montraient* → montrer
  - Pour un **nom, adjectif, article, ...** : sa forme au masculin singulier  
*vert, vertes, verts* → vert
- La lemmatisation demande des **ressources** et **un traitement linguistique**
  - En particulier pour les nombreuses exceptions
  - **Long** et donc difficile à mettre en œuvre pour des grandes collections
  - **Dépendant de la langue**
- Elle n'agrège que des variantes flexionnelles
  - *cheval = chevaux*
  - *cheval ≠ chevalier*

# Racinisation (*stemming*)

- Obtention de la **racine**, une forme tronquée du mot, commune à toutes les variantes morphologiques
  - Suppression des **flexions**
  - Suppression des **suffixes**
  - Ex : *cheval, chevaux, chevalier, chevalerie, chevaucher*  
→ "*cheva*"(mais pas "*cavalier*")
- La racinisation est généralement à base de règles
  - Rapide
  - Dépendant de la langue
- Elle agrège beaucoup plus que la lemmatisation
  - Vocabulaire plus petit

# Étiquetage

- Associer aux mots leur **catégorie morphosyntaxique** (nom, verbe, adjectif, etc.)

- Peut être utile pour :

- Supprimer les mots inutiles
- Opérer des regroupements en termes complexes
- Manipuler des mots ambigus avec plus de précision (*vers, or, pouvoir...*)
- Construire une représentation syntaxique

- Mais :

- Un processus plus long
- 96 % de précision = une erreur par phrase en moyenne !

ADJECTIF  
PRONOM  
VERBE  
ADVERBE  
ARTICLE  
PREPOSITION  
CONJONCTION  
NOM  
DETERMINANT

# Mots vides

- Les mots « **outils** » n'apportent pas de sens au texte  
déterminants : « le », « la », pronoms : « je », « nous »,  
prépositions : « sur », « contre », ...
- Ce sont les mots les plus **fréquents** de la langue
  - Les 30 mots les plus fréquents représentent 30 % des occurrences de mots
  - Les supprimer permet :
    - de réduire le bruit qu'ils pourraient générer
    - de « rapprocher » les mots porteurs de sens
    - d'économiser beaucoup de place dans un index
  - Mais :
    - ils informent sur la syntaxe
    - ils sont dans les expressions multi-mots
    - comme toutes les listes en TAL, la frontière n'est pas évidente à fixer



# Tâche : classification de texte

20newsgroups : 20 000 documents partitionnés en 20 classes

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

# Tâche : classification thématique

talk.politics.guns

Percentage of Reported  
Gun and Knife Attacks

Guns (6,350 total attacks)  
12.2

[...] teach children to  
safely handle firearms  
[...]

[...] not in the armed  
forces. [...]

this is a bus mouse.

[...] the purchase of a  
handgun when his life or  
family members have been  
threatened? [...]

[...] Didn't Christ tell  
his disciples to arm them  
selves [...]

[...] unintentional youth  
gun deaths [...]

pc.hardware

*Variantes*  
*Similarité sémantique*  
*Polysémie*  
*N-grammes*

# Représentation « sac de mots »

Lisible par l'humain



Sémantique explicite



Dimensionnalité



Similarité sémantique



Polysémie



Variantes orthographiques



Expressions multi-mots



Mots inconnus



Langage = symboles

Vocabulaire = tous les mots

Vélo ≠ bicyclette ≠ VTT, chien ≠ chat,  
Paris ≠ Londres

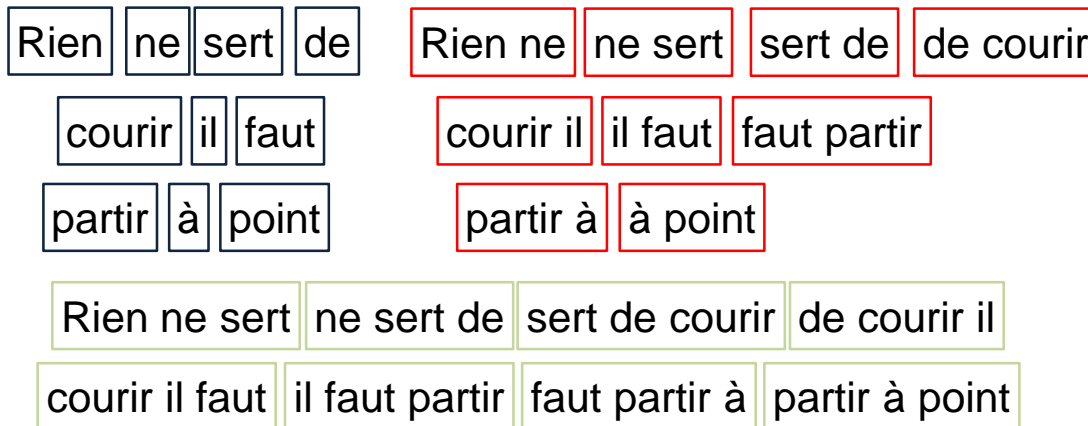


Tuebingen ≠ Tübingen,  
artichaut ≠ artichaud

# N-grammes de mots


$n$ -gramme = sous-séquence de  $n$  éléments consécutifs dans une séquence donnée

- n-gramme de caractères
- n-gramme de mots
- ...




# Tâche : recherche d'information

campagne



Web Actualités Images Vidéos Shopping Maps Musique


France ▼ Toute période ▼

 **Campagne Diesel EA189 (NOx) | Volkswagen Group France**  
[informations.volkswagengroup.fr/](https://informations.volkswagengroup.fr/)  
Campagne Diesel EA189 (NOx) | Volkswagen Group France Campagne Diesel EA189.

 **Campagne**  
[fr.wikipedia.org/wiki/Campagne](https://fr.wikipedia.org/wiki/Campagne)  
La campagne, aussi appelée milieu campagnol ou milieu  
espaces cultivés, naturels ou semi-naturels habités, e


 **Quand le directeur de campagne de Marine**  
[mediapart.fr/journal/france/290311/qua...](https://mediapart.fr/journal/france/290311/qua...)  
Il faut s'adresser au «client-électeur» comme un «Mor  
«démago à mort»: les conseils aux militants


campagne d'Italie











Web Actualités Images Vidéos Shopping Maps Musique

France ▼ Toute période ▼

 **Campagne d'Italie (1796-1797) — Wikipédia**  
[fr.wikipedia.org/wiki/Campagne\\_d'Italie\\_\(1796-1797\)](https://fr.wikipedia.org/wiki/Campagne_d'Italie_(1796-1797))  
La première campagne d'Italie est une campagne militaire menée par le général  
français Napoléon Bonaparte en Italie du Nord et sur le territoire autrichien du 24 mar...

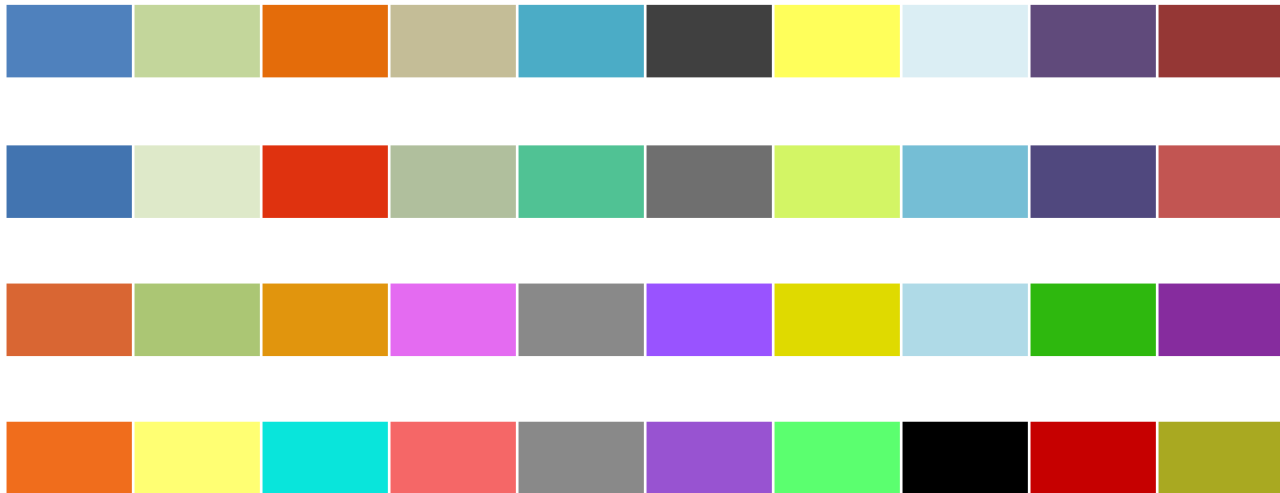
 **Campagne d'Italie (Seconde Guerre mondiale) — Wikipédia**  
[fr.wikipedia.org/wiki/Campagne\\_d'Italie\\_\(Seconde\\_Guerre\\_mondiale\)](https://fr.wikipedia.org/wiki/Campagne_d'Italie_(Seconde_Guerre_mondiale))  
modifier. La campagne d'Italie, appelée aussi libération de l'Italie,, est une campagne  
militaire de la Seconde Guerre mondiale ayant duré de juillet 1943 à mai 1945. Cette...

# Représentation « sac de n-grammes »

Lisible par l'humain		
Sémantique explicite		
Dimensionnalité		Vocabulaire = tous les n-grammes
Similarité sémantique		
Polysémie		n-grammes peu polysémiques
Variantes orthographiques		
Expressions multi-mots		
Mots inconnus		

# Représentations denses (*embeddings*)

- Word embeddings = représentation vectorielle des mots
- Mots proches dans l'espace = mots ayant un certain degré de similarité entre eux



# Représentations denses (*embeddings*)

- Word embeddings = représentation vectorielle des mots
- Mots proches dans l'espace = mots ayant un certain degré de similarité entre eux





# Représentations denses (*embeddings*)

- Word embeddings = représentation vectorielle des mots
- Mots proches dans l'espace = mots ayant un certain degré de similarité entre eux



# Représentations denses (*embeddings*)



- **Intuition 1.** Chaque mot d'un langage est associés à une composition de facteurs cachés (souvent inintelligible)

*Ex : chat = 10 (animal) + 5 (doux) – 10 (loyal)*

- **Intuition 2.** Hypothèse distributionnelle

« You shall know a word by the company it keeps » (Firth, 1957)

Deux mots proches dans l'espace vectoriel = deux mots qui partagent souvent des contextes similaires

*Ex : le ... griffe ; ... est un félin*

$occurrence(chat) \sim occurrence(tigre)$

$W_{chat} \cdot W_{contexte} \sim W_{tigre} \cdot W_{contexte}$

$W_{chat} \sim W_{tigre}$

(© Perceval Wajsbürt)

# Représentations denses (*embeddings*)



Deux mots proches dans l'espace vectoriel = deux mots qui partagent souvent des contextes similaires



~~Deux mots proches dans l'espace vectoriel = deux mots ayant un sens proche~~

$W_{\text{vélo}} \sim W_{\text{bicyclette}}$

$W_{\text{chat}} \sim W_{\text{tigre}}$

$W_{\text{bon}} \sim W_{\text{mauvais}}$

$W_{\text{malade}} \sim W_{\text{malades}}$

$W_{\text{Paris}} \sim W_{\text{Londres}}$

$W_{\text{petit}} \sim W_{\text{grand}}$

# Représentations denses (*embeddings*)



- Quelques méthodes :
  - *Latent Semantic Analysis (1988)*
  - *Latent Dirichlet Allocation (2001)*
  - *Word2vec (2013)*
  - *GloVe (2014)*
  - *FastText (2017)*
  - *Elmo (2018)*
  - *BERT (2018)*
  - ...
- Nécessité de gros corpus de texte

# Modèle de langue

Ces modèles entrent dans la classe des *modèles de langue*.

Un modèle de langue est un modèle sans supervision manuelle qui a appris à prédire les mots dans des séquences (mot suivant).

Un modèle est *pré-entraîné* (*pretrained*) et on peut l'*affiner* (*fine-tune*) en fonction d'une tâche.

Ces modèles sont au cœur de la grande majorité des systèmes de TAL à l'heure actuelle.

# Tâche : reconnaissance d'entités nommées

## CONLL 2003, reconnaissance d'entités nommées

- Anglais, Allemand, Espagnol, Néerlandais
- Types : PERSONNE, ORGANISATION, LIEU, DIVERS

LOCATION

Bonn has led efforts to protect public health after consumer confidence collapsed in March









LOCATION

[...] told Israel Radio it looked like Damascus wanted to talk [...]

LOCATION

[...] concert in the English city of Nottingham he threw the sheet of paper [...]

# Représentation dense

Lisible par l'humain		
Sémantique explicite		
Dimensionnalité		Vecteur = dizaines ou centaines de dimensions
Similarité sémantique		Par construction, mais implicite
Polysémie		
Variantes orthographiques		
Expressions multi-mots		Opérations arithmétiques sur les vecteurs de mots
Mots inconnus		Selon les méthodes, voir plus loin

# Représentation dense

Autre avantage des représentations denses :  
on peut les mélanger à des représentations d'autre chose

TEXT PROMPT    an armchair in the shape of an avocado. an armchair imitating an avocado.

<https://openai.com/blog/dall-e/>

AI-GENERATED IMAGES

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever  
Zero-Shot Text-to-Image Generation



# Représentation dense

Autre avantage des représentations denses :  
on peut les mélanger à des représentations d'autre chose

FOOD101

**guacamole** (90.1%) Ranked 1 out of 101 labels



✓ a photo of **guacamole**, a type of food.

✗ a photo of **ceviche**, a type of food.

✗ a photo of **edamame**, a type of food.

✗ a photo of **tuna tartare**, a type of food.

✗ a photo of **hummus**, a type of food.

SUN397

**television studio** (90.2%) Ranked 1 out of 397



✓ a photo of a **television studio**.

✗ a photo of a **podium indoor**.

✗ a photo of a **conference room**.

✗ a photo of a **lecture room**.

✗ a photo of a **control room**.

YOUTUBE-BB

**airplane, person** (89.0%) Ranked 1 out of 23



✓ a photo of a **airplane**.

✗ a photo of a **bird**.

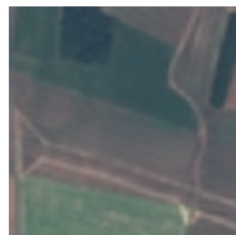
✗ a photo of a **bear**.

✗ a photo of a **giraffe**.

✗ a photo of a **car**.

EUROSAT

**annual crop land** (12.9%) Ranked 4 out of 10



✗ a centered satellite photo of **permanent crop land**.

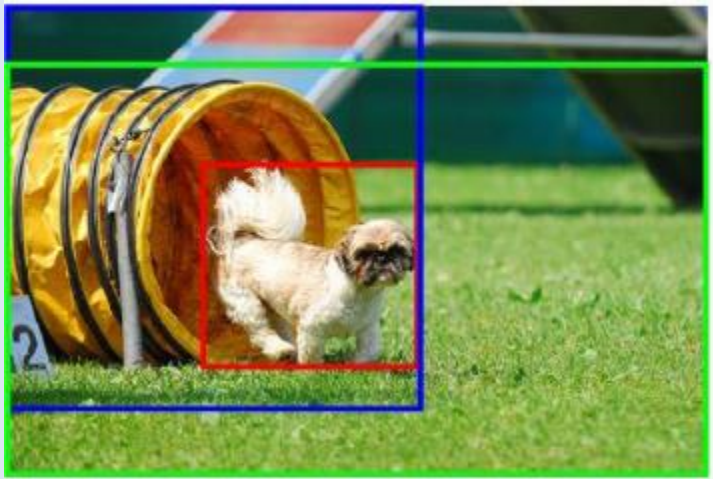
✗ a centered satellite photo of **pasture land**.

✗ a centered satellite photo of **highway or road**.

✓ a centered satellite photo of **annual crop land**.

✗ a centered satellite photo of **brushland or shrubland**.

# Tâche : légendage automatique d'image



A little brown and white dog  
emerges from a yellow col-  
lapsable toy tunnel onto the  
lawn.

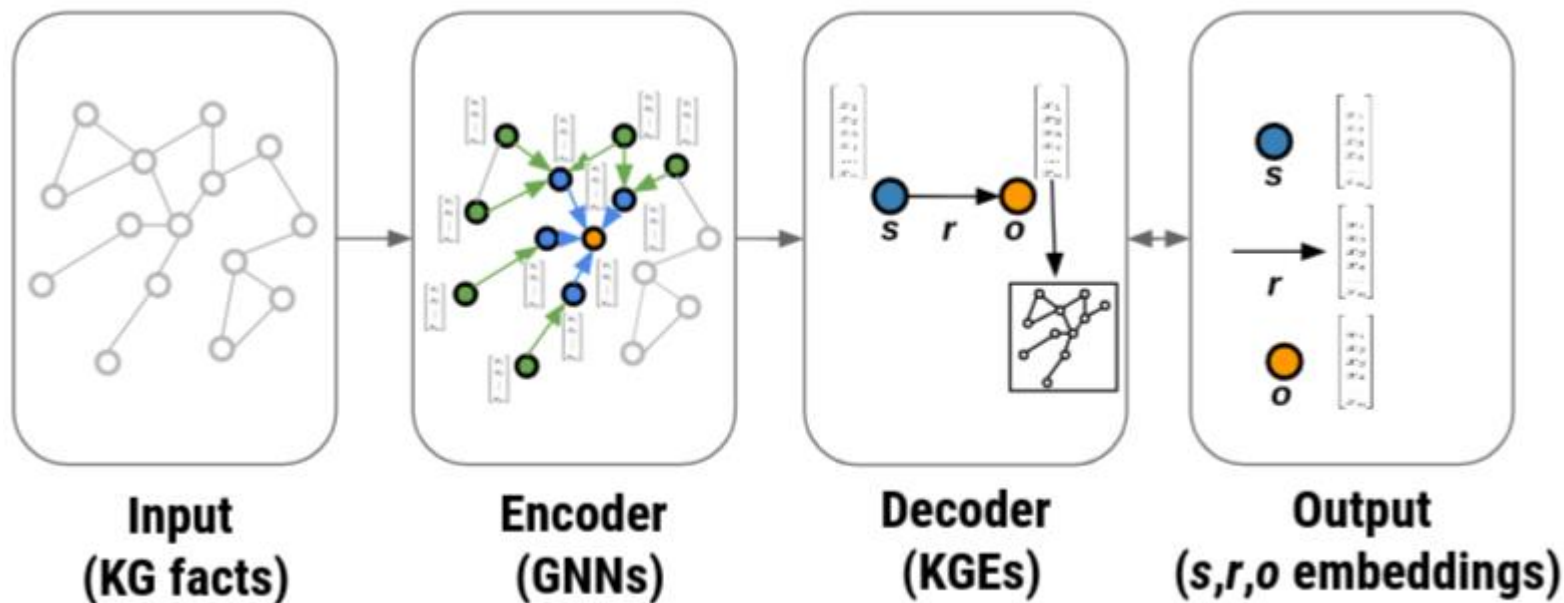
Rohrbach, Anna, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele (2016). "Grounding of Textual Phrases in Images by Reconstruction". Computer Vision - ECCV 2016

# Représentation dense

Autre avantage des représentations denses :  
on peut les mélanger à des représentations d'autre chose

Embeddings de graphes de connaissance

Luca Costabello, Sumit Pai, Nicholas McCarthy, Adrianna Janik  
Knowledge Graph Embeddings Tutorial: From Theory to Practice  
ECAI 2020



(Source :  
Giuseppe Futila)

# Sous le mot

- Subwords

school      { 'sch', 'cho', 'hoo', 'ool',  
                  'scho', 'choo', 'hool',  
                  'schoo', 'chool',  
                  'school' }      (n-grammes de caractères)

Puis agrégation des sous-mots en mots (somme des vecteurs)

- WordPieces

- Un vocabulaire de taille prédéfinie, composé de n-grammes de caractères
- Vocabulaire choisi pour maximiser la fréquence des n-grammes
- Possibilité d'un tokenizer multilingue
- Exemple (FlauBERT) :

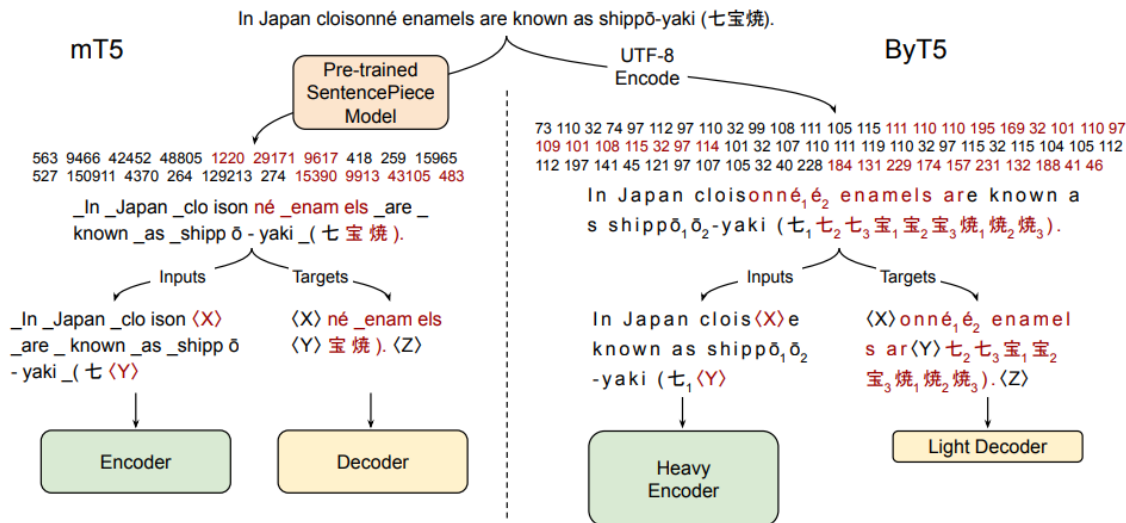
nous uti ##lisons des mo ##deles de re ##presentation contextu ##elle

# Sous le mot

- Et maintenant même, octet !

## ByT5: Towards a token-free future with pre-trained byte-to-byte models

Linting Xue\* Aditya Barua\* Noah Constant\* Rami Al-Rfou\*  
Sharan Narang Mihir Kale Adam Roberts Colin Raffel  
Google Research



Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, Colin Raffel  
ByT5: Towards a token-free future with pre-trained byte-to-byte models  
Non publié, 28 mai 2021 : <https://arxiv.org/abs/2105.13626>

# Tâche : recherche vocale

- Recherche vocale : lancer des recherches web par la parole
- Utilisation de WordPiece pour gérer le vocabulaire infini, notamment dans des langues sans espaces









1. ”京都 清水寺の写真” (original text)
2. ”京都 清水寺 の写真” (after segmentation)
3. ”\_\_京都\_\_ \_\_清水寺 の写真\_\_” (after addition of underscores, used for LM training, dictionary etc.)
4. ”\_\_京都\_\_ \_\_清水寺 の写真\_\_” (decoder result)
5. ”京都 清水寺の写真” (displayed output)

Schuster, M., and Nakajima, K. Japanese and Korean voice search.  
2012 IEEE International  
Conference on Acoustics, Speech and Signal Processing (2012).

# Représentations denses en contexte

- **Représentation statique** : un token = un vecteur
  - On manipule une « matrice d'embeddings » ( $N \times d$ )
  - Le vecteur du token est le même à chacune de ses occurrences dans le corpus
- **Représentation contextuelle** : calcul du vecteur en contexte
  - Le calcul de la représentation est intégré dans le modèle
  - Les mots précédents et suivants agissent sur la représentation (réseau récurrent, attention...)
  - Exemple : ELMo, ULMFit, BERT, XLM...

# Représentation contextuelle

Lisible par l'humain	
Sémantique explicite	
Dimensionnalité	
Similarité sémantique	
Polysémie	
Variante orthographiques	
Expressions multi-mots	
Mots inconnus	

Vecteur = dizaines ou centaines de dimensions

Par construction, mais implicite



# Pondération des mots

- Dans un document, dans une requête, dans une question, les termes n'ont pas tous la même **importance**
- **Intuition #1** : plus un document contient d'occurrences d'un terme, plus il est "à propos" de ce terme (plus il sera pertinent par rapport à une requête contenant ce terme)

$tf_{t,d}$  = nombre d'occurrences du terme  $t$  dans le document  $d$

# Pondération des mots : la matrice des fréquences

	Antoine & Cléopâtre	Jules César	La Tempête	Hamlet	Othello	Macbeth
Antoine	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
César	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cléopâtre	57	0	0	0	0	0
pitié	2	0	3	5	5	1
pire	2	0	1	1	1	0

Chaque document est un vecteur  $v$  dans  $\mathbb{N}^{|v|}$

# Pondération des mots

- **Intuition #2** : des termes très fréquents dans tous les documents ne sont pas si importants (ils sont moins *discriminants*)
- On compense donc la fréquence des termes dans les documents (tf) en prenant en compte leur fréquence dans la collection (df)

$df_t =$  nombre de documents qui contiennent le terme  $t$

$$idf_t = \log_{10} \frac{N}{df_t} \quad (N = \text{nombre total de documents})$$

# Pondération des mots

- Le **poids** d'un terme ( $tf.idf$ ) est la combinaison de ces deux intuitions pour rendre compte du caractère discriminant d'un terme dans un document

$$\begin{aligned}w_{t,d} &= tf_{t,d} \times idf_t \\ &= tf_{t,d} \times \log_{10} \frac{N}{df_t}\end{aligned}$$

ou  $w_{t,d} = \log tf_{t,d} \times \log_{10} \frac{N}{df_t}$

- Le poids d'un terme  $t$  :
  - augmente avec sa **fréquence dans le document**
  - augmente avec sa **rareté dans la collection de documents**

# Pondération des mots : la matrice des poids

	Antoine & Cléopâtre	Jules César	La Tempête	Hamlet	Othello	Macbeth
Antoine	13,1	11,4	0	0	0	0
Brutus	3,0	8,3	0	1	0	0
César	2,3	2,3	0	0,5	0,3	0,3
Calpurnia	0	11,2	0	0	0	0
Cléopâtre	17,7	0	0	0	0	0
pitié	0,5	0	0,7	0,9	0,9	0,3
pire	1,2	0	0,6	0,6	0,6	0

Chaque document est un vecteur  $v$  dans  $\mathbb{R}^{|v|}$

De nombreuses variantes existent !

# Du mot à la phase

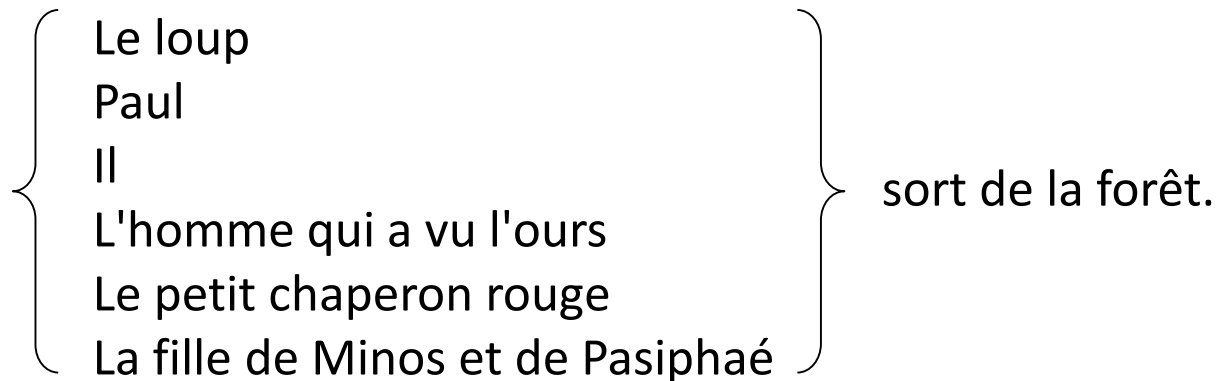
# Du mot à la phrase : les syntagmes

- Les **syntagmes** (ou **constituants**) sont qualifiés par le type de l'élément principal (la **tête**). On a donc des syntagmes :
  - **nominaux** (*le loup, Paul, l'homme qui valait 3 millions*)
  - **verbaux** (*vendra, valait 3 millions*)
  - **adjectivaux** (*rouge, [une classe] pleine d'étudiants*)
  - **adverbiaux** (*bien, conformément à la loi*)
  - on parle aussi de syntagmes **prépositionnels** (*[le chat] de ma mère*)
- Les autres éléments sont :
  - les **spécifieurs** (*déterminants...*)
  - les **qualificateurs** (*adjectifs, adverbes...*)
  - les **compléments** (*compléments du nom, propositions relatives...*)
- Un syntagme a la même fonction que sa tête dans la phrase

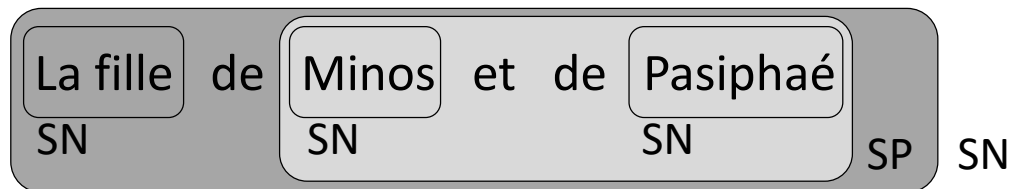


# Du mot à la phrase : les syntagmes

- Les syntagmes de même type sont syntaxiquement **substituables** entre eux...



- Les syntagmes peuvent s'imbriquer les uns dans les autres :



# Chunking

- *Les gendarmes* *interpellent* *un conducteur* *en état d'ivresse.*
- Pas d'analyse de la structure interne
- **Pas de liens de dépendances** entre les *chunks*
- Les **ambiguïtés de rattachement** sont implicites

- *Bill* *vit* *l'homme* *sur la colline* *avec un télescope.*
- *Bill* *vit* *l'homme* *sur la colline* *avec un télescope.*

# Tâche : extraction de termes

N..., âgé de **40 ans**, a consulté pour une **symptomatologie mictionnelle sévère, irritative et obstructive**. On notait dans ses **antécédents** une **orchidectomie droite** pour **ectopie testiculaire** à l'âge de **10 ans**, une **lobectomie thyroïdienne gauche** après **irradiation nucléaire accidentelle**.

Ce patient était adressé en **consultation d'Urologie** avec une **échographie de l'appareil urinaire**.

- Les termes sont en général des groupes nominaux
- Peut permettre :
  - d'indexer des n-grammes pertinents
  - de créer une terminologie/ontologie basée sur les corpus
  - de rechercher des traductions plus précises
  - de faciliter l'analyse syntaxique
  - ...

# Les entités nommées

- Les **entités nommées** sont des éléments qu'il est intéressant de pouvoir distinguer du reste du texte :
  - Entités : **personnes, organisations, lieux**
  - Dates : **dates, heures**
  - Quantités : **montants financiers, pourcentages, etc.**
- **Reconnaissance** des entités nommées :
  - **Identifier** ces unités dans un texte
  - Les **catégoriser**
  - Éventuellement, les **normaliser** (*entity linking*)

# Entités nommées

## Identification

Le joueur de tennis américain **John McEnroe** a déclaré **samedi** sur **ESPN** que **Gaël Monfils** n'était pas assez professionnel. « **Monfils** aurait déjà dû gagner 4 ou 5 majeurs », a-t-il précisé.

# Entités nommées

## Catégorisation

Le joueur de tennis américain *personne* John McEnroe a déclaré *date* samedi sur *organisation* ESPN que *personne* Gaël Monfils n'était pas assez professionnel. « *personne* Monfils aurait déjà dû gagner 4 ou 5 majeurs », a-t-il précisé.

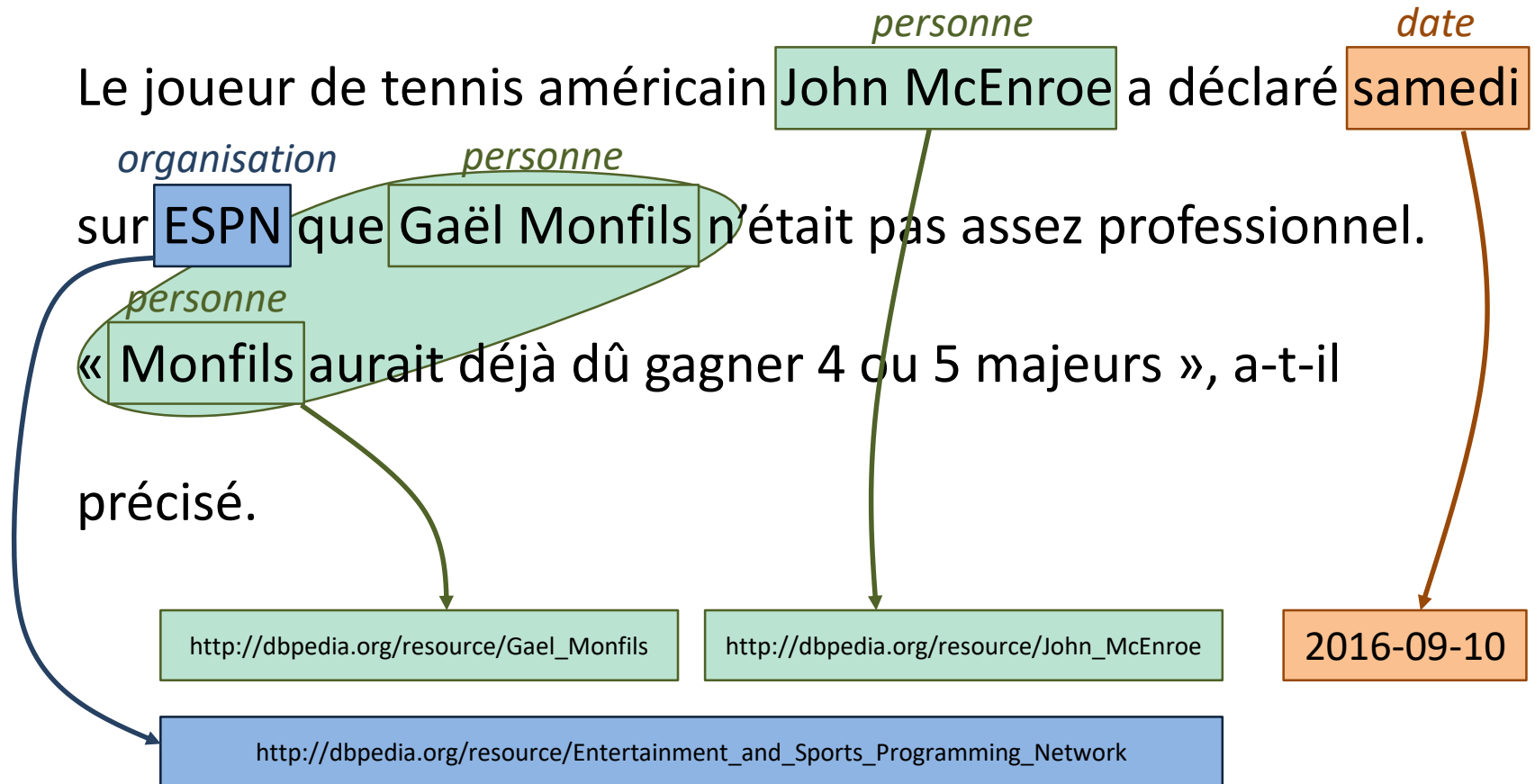
# Entités nommées

## Normalisation

Le joueur de tennis américain *personne* John McEnroe a déclaré *date* samedi sur *organisation* ESPN que *personne* Gaël Monfils n'était pas assez professionnel. *personne* « Monfils aurait déjà dû gagner 4 ou 5 majeurs », a-t-il précisé.

# Entités nommées

## Normalisation



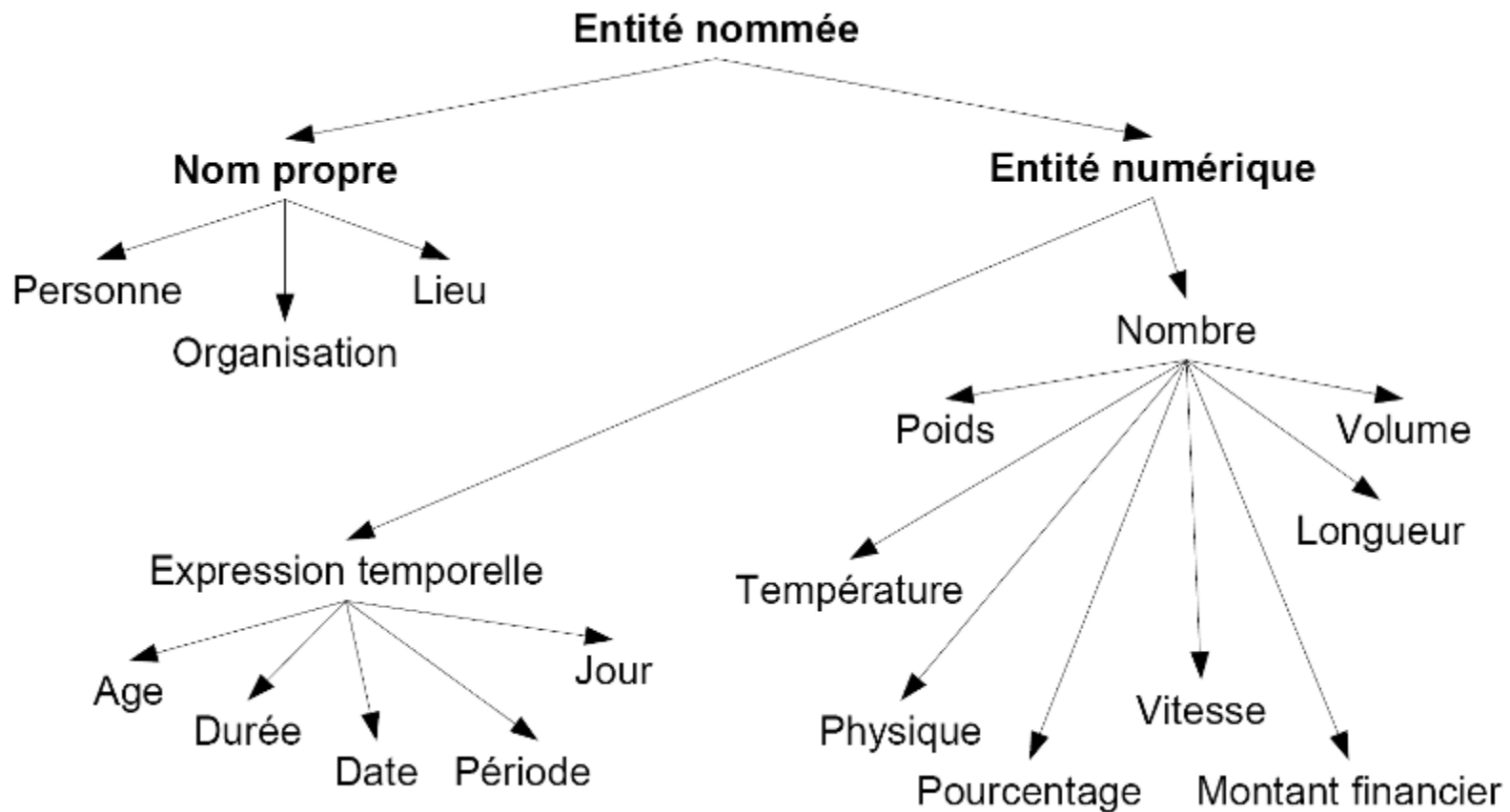


# Entités nommées

## Plus de précision ?

Le *fonction* **joueur de tennis américain** *pers:sportif* **John McEnroe** a déclaré *date:jour* **samedi**  
sur *org:TV* **ESPN** que *pers:sportif* **Gaël Monfils** n'était pas assez professionnel.  
« *pers:sportif* **Monfils** aurait déjà dû gagner *evt:tournoi* **Rolland-Garros** », a-t-il  
précisé.

# Entités nommées



# Tâche : Désidentification

Cher Confrère,

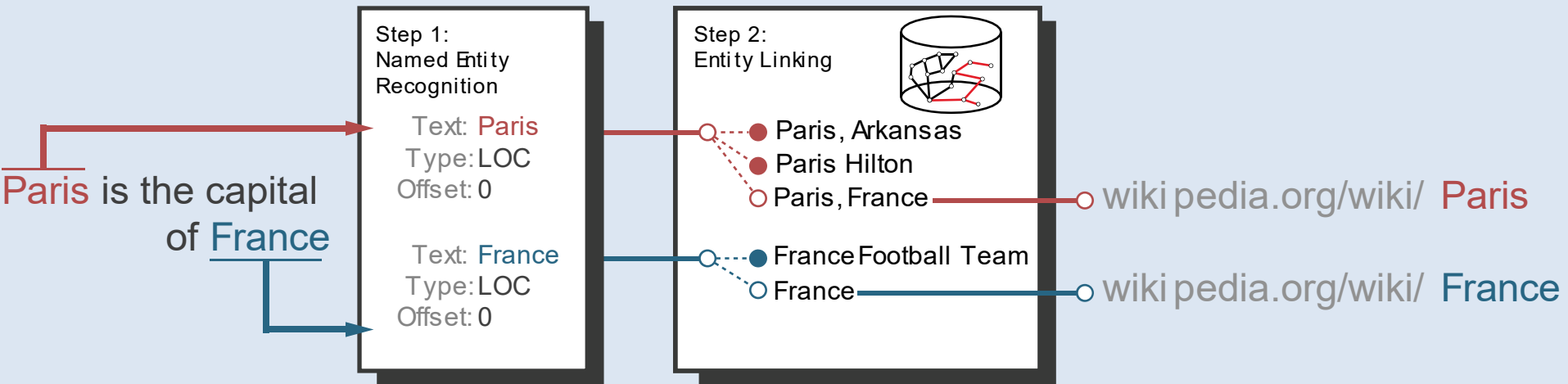
Votre patiente madame [REDACTED], âgée de [REDACTED], née le [REDACTED], a été hospitalisée aux soins intensifs de cardiologie le [REDACTED] pour la prise en charge d'un syndrome douloureux thoracique.

Les principaux antécédents de cette patiente sont représentés par :

- une thyroïdectomie en [REDACTED].
- une hystérectomie en [REDACTED].
- un stripping des varices en [REDACTED].
- une hernie discale en [REDACTED].

Ses principaux facteurs de risque cardio-vasculaire sont représentés par une hypertension artérielle traitée par ATACAND, une dyslipidémie non traitée, un tabagisme estimé à 5 paquets année, et une légère surcharge pondérale ([REDACTED] pour [REDACTED]).

# Tâche : *entity linking*

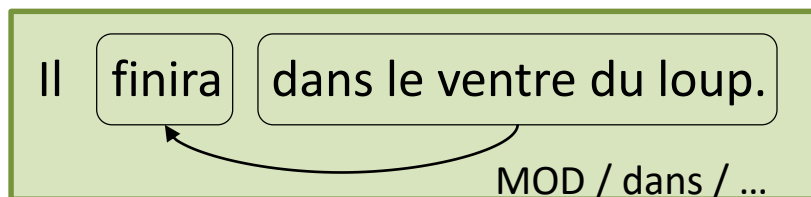
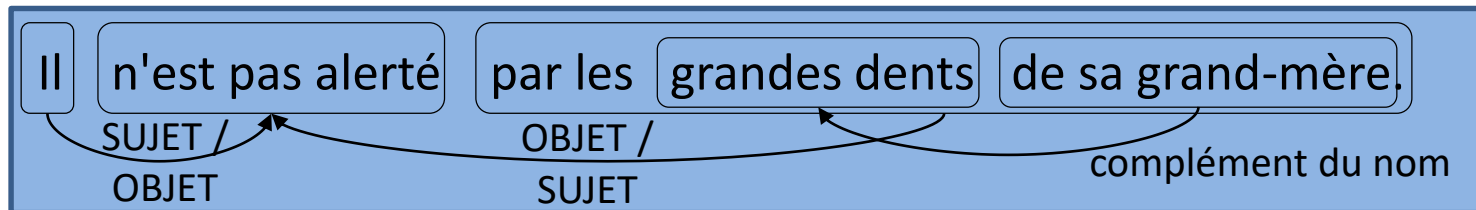
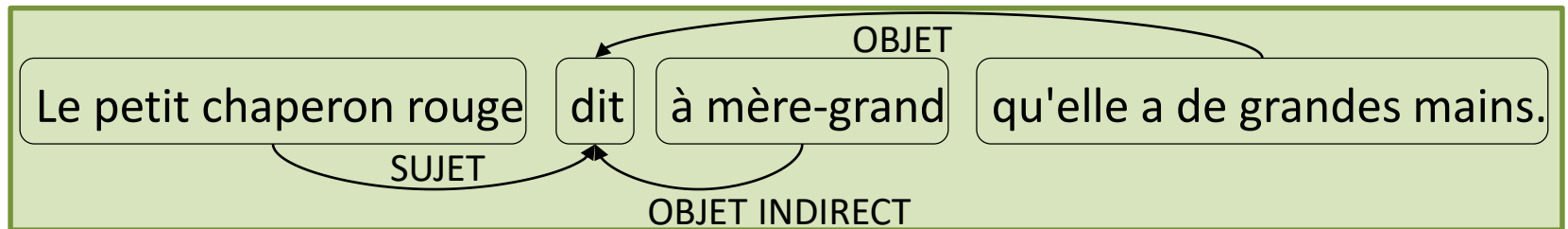
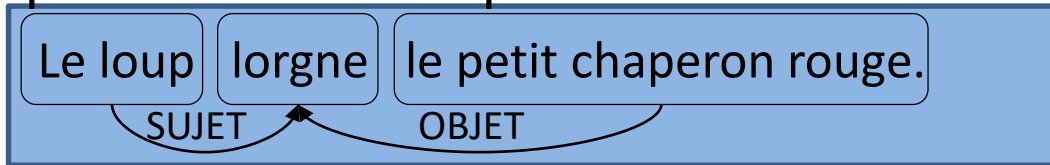


# La phrase

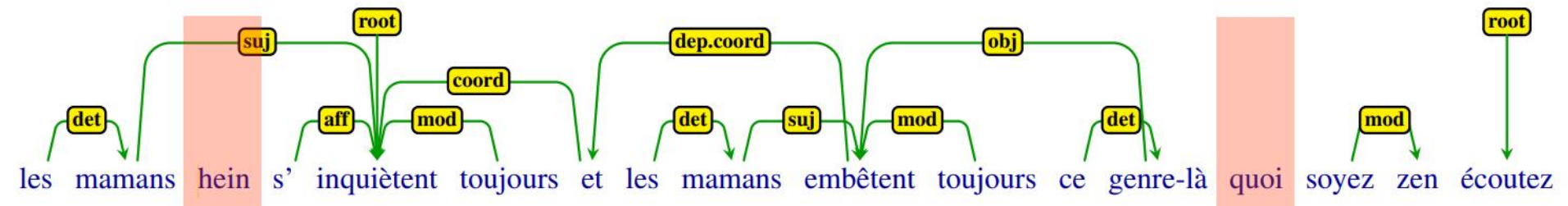
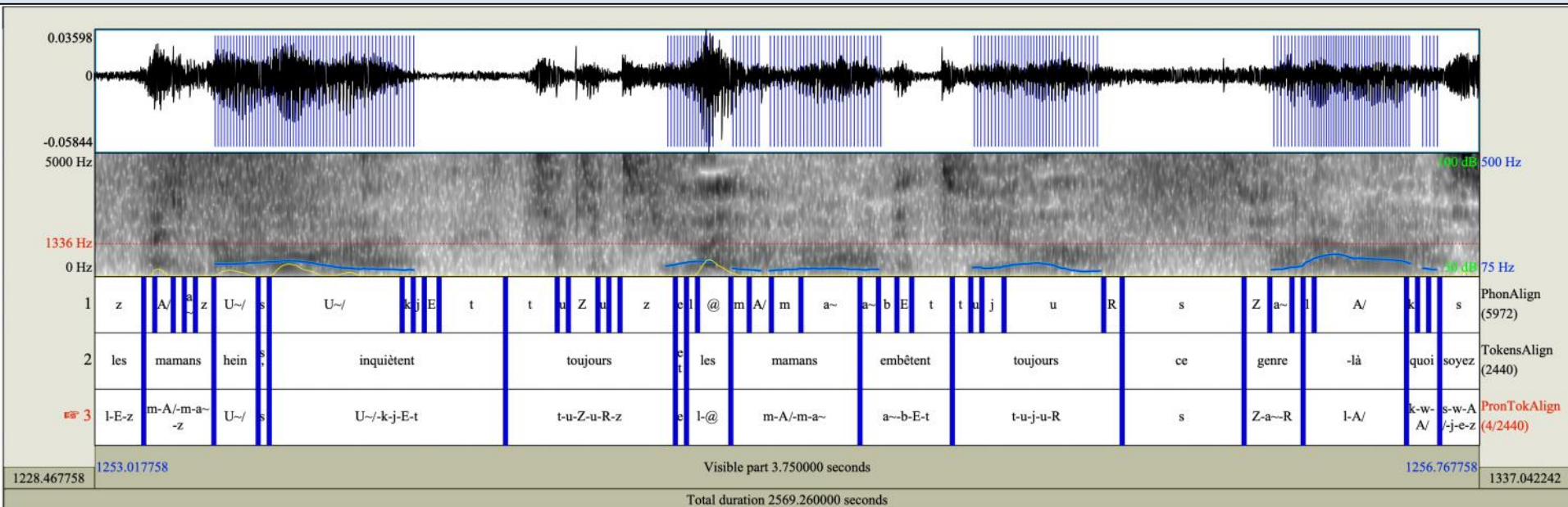
- Une ou plusieurs **propositions** :
  - *Deux pigeons s'aimaient d'amour tendre.*
  - *Deux sûretés valent mieux qu'une, et le trop en cela ne fut jamais perdu.*  
(coordination)
  - *Vous savez que nul n'est prophète en son pays.*  
(subordination)
  
- Une succession de phrases forment le **discours**.

# Les fonctions grammaticales

- Le *rôle syntaxique* que joue un constituant par rapport à un autre
- Dépend surtout des positions relatives dans l'énoncé :



# Tâche : analyse syntaxique sur des transcriptions de dialogues oraux, pour l'extraction d'information



Yannis Haralambous, Christophe Lemey, Philippe Lenca, Romain Billot, Deok-Hee Kim-Dufor.  
 Using Dependency Syntax-Based Methods for Automatic Detection of Psychiatric Comorbidities.  
 Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments, May 2020, Marseille, France.

# Les ambiguïtés syntaxiques

- On peut les classer selon les connaissances nécessaires pour la désambiguïstation
- Connaissances **pragmatiques**
  - *Jean a rapporté un vase de Chine.*
  - *Jean a rapporté un vase de Chine (des puces de St-Ouen).*
- Connaissances **sémantiques**
  - *Jean vend une tarte aux pommes.*
  - *Jean vend une tarte aux clients.*
- Connaissances **syntaxiques**
  - *Un jus d'oranges fraîches.*
  - *Un jus d'oranges frais.*
- Parfois des centaines de combinaisons possibles pour une phrase.



# Tâche : correction syntaxique

The screenshot shows the Microsoft Word interface with the 'Révision' (Review) tab selected. The ribbon includes 'Grammaire et orthographe' (Grammar and Spelling), 'Recherche des synonymes' (Synonym Search), 'Dictionnaire' (Dictionary), 'Traduction' (Translation), 'Statistiques' (Statistics), 'Nouveau commentaire' (New Comment), 'Supprimer' (Delete), 'Précédent' (Previous), 'Suivant' (Next), 'Suivi des modifications' (Track Changes), 'Bulles' (Comments), 'Final', 'Afficher les modifications' (Show Changes), and 'Volet Vérification' (Check for Grammar and Spelling). The document text is 'Ils ne partirons que demain dans l'après-midi.' The word 'partirons' is underlined with a red squiggly line, indicating a grammar error.

Two windows are open over the document:

- Grammaire et orthographe : Français (France)**: This window shows the error 'Accord du sujet et du verbe' (Subject and verb agreement) for the sentence 'Ils ne **partirons** que de...'. The 'Suggestions' list contains 'partiront'. The 'Langue du dictionnaire' is set to 'Français (France)'. The 'Vérifier la grammaire' checkbox is checked. Buttons for 'Options...' and 'Rétablir' are visible.
- Word - Aide**: This help window provides the following information:
  - Accord du sujet et du verbe**
  - Le verbe et son sujet doivent s'accorder en nombre et en personne.
  - Plutôt que :
    - " Le soleil **luisais** de tous ses feux. "
  - Écrivez :
    - " Le soleil **luisait** de tous ses feux. "

# Relations sémantiques

- Les relations sémantiques considérées dans la littérature sont très diverses :
  - Causalité
  - Temps, espace
  - Possession, production, instrument
  - Manière, moyen
  - Sujet
  - Propriété, bénéficiaire
  - Mesure
  - ...

Barker K. and Szpakowicz S 1998. Semi-automatic recognition of noun modifier relationships. In Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98), pages 96-102, Montreal, Canada

D. Moldovan and R. Girju. 2001. An Interactive Tool For The Rapid Development of Knowledge Bases. In International Journal on Artificial Intelligence Tools (IJAIT).

# Relations sémantiques :

## l'exemple de SemEval 2010, tâche 10

Cause-Effect	those <b>cancers</b> were caused by radiation <b>exposures</b>
Instrument-Agency	<b>phone operator</b>
Product-Producer	a <b>factory</b> manufactures <b>suits</b>
Content-Container	a <b>bottle</b> full of <b>honey</b> was weighed
Entity-Origin	<b>letters</b> from foreign <b>countries</b>
Entity-Destination	the <b>boy</b> went to <b>bed</b>
Component-Whole	my <b>apartment</b> has a large <b>kitchen</b>
Member-Collection	there are many <b>trees</b> in the <b>forest</b>
Message-Topic	the <b>lecture</b> was about <b>semantics</b>

Iris Hendrickx et al.  
SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations  
between Pairs of Nominals  
Proceedings of the 5th International Workshop on Semantic Evaluation

# Au-delà de la phrase

# La coréférence

- Coréférence pronominale

- Jacques<sub>1</sub> était furieux. **Il**<sub>1</sub> s'était disputé avec Georges.
- Dominique<sub>1</sub> rencontra Collins<sub>2</sub> à un congrès. **Ils**<sub>1+2</sub> se réconcilièrent.
- Nicolas<sub>1</sub> rencontra Dominique<sub>2</sub> dans un couloir. **Il**<sub>?</sub> **lui**<sub>?</sub> en voulait toujours.
- Pierre<sub>1</sub> empoisonna Sam<sub>2</sub>. **Il**<sub>2</sub> mourut.
- Pierre<sub>1</sub> empoisonna Sam<sub>2</sub>. **Il**<sub>1</sub> fut arrêté.
- Si votre bébé ne supporte pas le lait cru, faites-**le** bouillir.

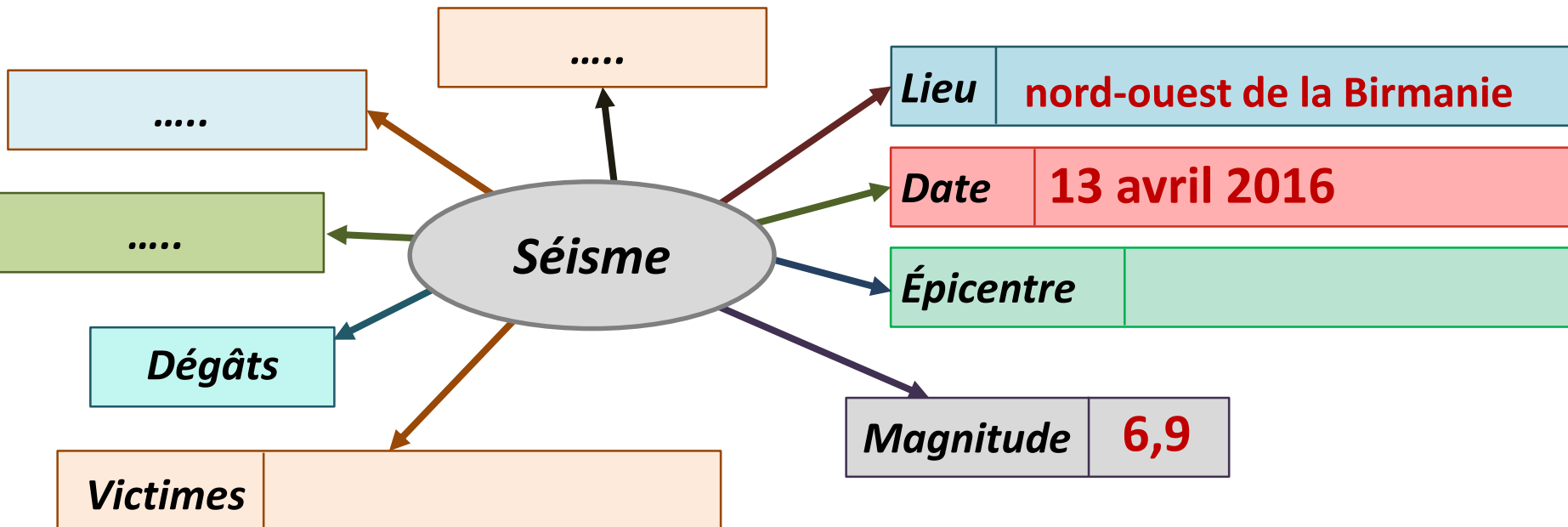
- Autres

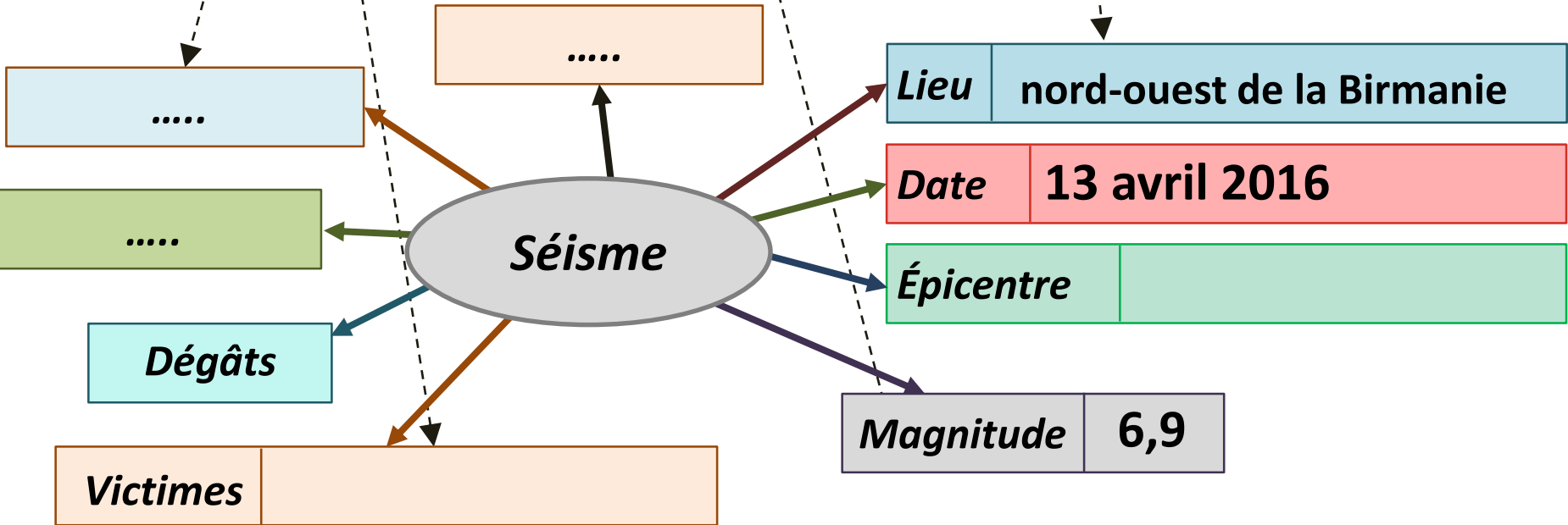
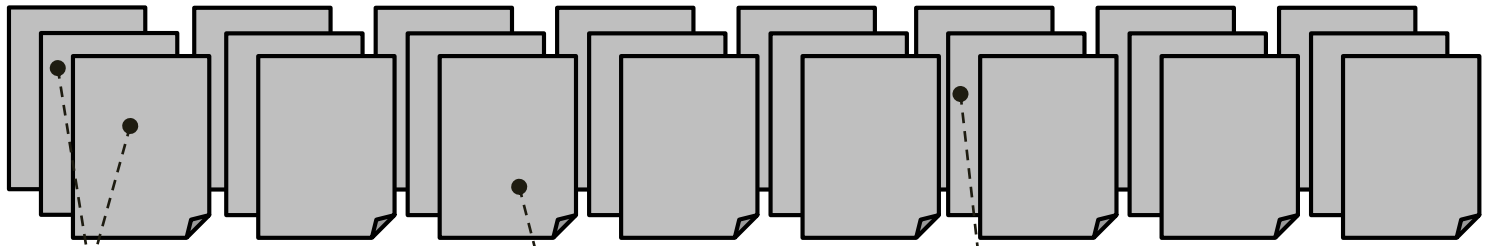
- La cage du gorille s'ouvrit. **Sa** serrure devait être mal fermée.
- Le gorille accéléra le pas vers le juge. **Le quadrumane** avait une idée derrière la tête.
- Je ne peux donner la suite de l'histoire. **Cela** serait pourtant délectable.

# Extraction d'information

Extraction d'entités, extraction de relations

Un séisme de magnitude 6,9 a secoué hier  
le nord-ouest de la Birmanie.





# Classification de texte

Quel thème ?  
(international,  
politique, sports,  
people, sciences,  
économie, ...)

## Tintin : les procédures judiciaires pour contrefaçon soulèvent des questions sur la paternité du héros reporter



Onésime, héros de Benjamin Rabier aux allures de Tintin (image BnF).

Publié le 2 juin 2021

[Partager](#) 820 [Twitter](#)




### **Hergé est-il l'unique créateur de Tintin ? Car il existe bien un Tintin en pantalon de golf et houppette créé en 1897 par l'illustrateur français Benjamin Rabier.**

Tintin est le héros le plus protégé de la bande dessinée. La société Moulinsart qui gère l'oeuvre du dessinateur belge est d'une fermeté sans égale. Elle poursuit quiconque reproduit ou s'inspire des oeuvres du maître. Avec d'ailleurs plus ou moins de succès. Tout récemment, la chambre civile du tribunal de Rennes vient de donner raison au peintre Xavier Marabout en lui accordant le droit de parodier le reporter du Petit Vingtième. Une autre procédure est en cours au tribunal civil de Marseille à l'encontre du plasticien Peppone. Les démêlés judiciaires de l'artiste provençal avec Moulinsart ont commencé en 2015. Des bustes du héros belge en résine de fibre de verre sont à l'origine du litige.








# Classification de texte

Spam ou pas spam ?

WITH DUE RESPECT  Spam x  

---

 **Zumba Kabale** <zumbakabale0003@gmail.com> 15 sept.   

À cci : moi 

Dear Friend,  
I know that this mail will come to you as a surprise as we have never met before, but need not to worry as I am contacting you independently of my investigation and no one is informed of this communication. I need your urgent assistance in transferring the sum of \$11.3million immediately to your private account. The money has been here in our Bank lying dormant for years now without anybody coming for the claim of it.

I want you to incorporate with me for the release of this money into your private bank account as the relative to our deceased customer (the account owner) who died a long ago and with her supposed NEXT OF KIN since 31 January 2000. The Banking law here does not allow such money to stay more than 15 years, because the money will be recalled to the Bank treasury account as unclaimed fund.

By indicating your interest I will send you the full details on how the business will be executed.

Best Regqrd  
Mr.Zumba Kabale

# Classification de texte

*“Une arnaque!!!”*

Content  
ou pas content ?

J'avais réservé une chambre double en demandant des lits séparés car partageais la chambre avec la collègue. Notre demande avait été confirmée par la réception de l'hôtel quelques jours avant notre arrivée. Le jour J: ascenseur dont la tapisserie se décolle, chambre avec lit king seize et baignoire trônant au milieu de la pièce...notre demande n'avait finalement pas été prise en compte. Nous avons demandé à être changées de chambre et avons tout simplement fini... À la cave! Chambre accessible par l'extérieur, odeur de moisi, bouteille de jus de fruit périmée dans le frigidaire, finitions douteuse: moisissure sur la poignée de porte des toilettes, saleté dans la baignoire et la cerise sur le gâteau: impossible de fermer la porte de douche car cette dernière cogne dans le pommeau!!la porte de douche a visiblement été montée avant la colonne de douche! Et le pommeau qui se trouve à hauteur de les chevilles, le cordon de douche n'est pas suffisamment long pour le permettre d'atteindre mes cheveux! Deuxième demande auprès de la réception afin d'être changées de chambre, à plus de 200€ la nuit, nous attendons un minimum de propreté et de fonctionnalité! Nous arrivons finalement dans une troisième chambre, correspondant davantage à nos attentes, mais toujours: baignoire sale (morceaux de peau et de cheveux, beurk), table de nuit abîmée. En clair: un nom, une déco surchargée qui masque des finitions inexistantes et hôtel qui ne bénéficie pas du tout du standing qu'il veut se donner.

# Génération de... trucs

## Can AI Write Your React Code?



Harshdeep Bilaiya [Follow](#)  
Sep 2, 2020 · 5 min read



In a word: **YES!** But it could do much, much more than that. This disruptive language model is taking the internet by storm — it's called **GPT-3** (*Generative Pre-trained Transformer 3*). It is an autoregressive language model that uses deep learning to produce human-like text. It is the third-generation language prediction model in the *GPT-n* series created by [OpenAI](#). OpenAI is an AI research and development company founded in late 2015, by one of THE tech innovators of our time — Elon Musk (along with *Greg Brockman* (CTO), *Sam Altman* (CEO) and 3 other gentlemen).

<https://medium.com/globant/can-ai-write-you-react-code-c36c9f16b1f5#:~:text=In%20a%20word%3A%20YES!>

### AI generated faces

x GPT-3

Generate a front facing young black adult female with brown eyes and long hair

Generate



<https://gpt3demo.com/apps/gpt-3-x-gans>

<https://philosopherai.com/philosopher/lets-design-a-happier-world-together-c4f8ad>

## Philosopher AI



This is an experiment in what one might call "prompt engineering", which is a way to utilize GPT-3, a neural network trained and hosted by OpenAI.

GPT-3 is a language model. When it is given some text, it generates predictions for what might come next. It is remarkably good at adapting to different contexts, as defined by a prompt (in this case, hidden), which sets the scene for what type of text will be generated.

Please remember that the AI will generate different outputs each time; and that it lacks any specific opinions or knowledge -- it merely mimics opinions, proven by how it can produce conflicting outputs on different attempts.

## Let's design a happier world together

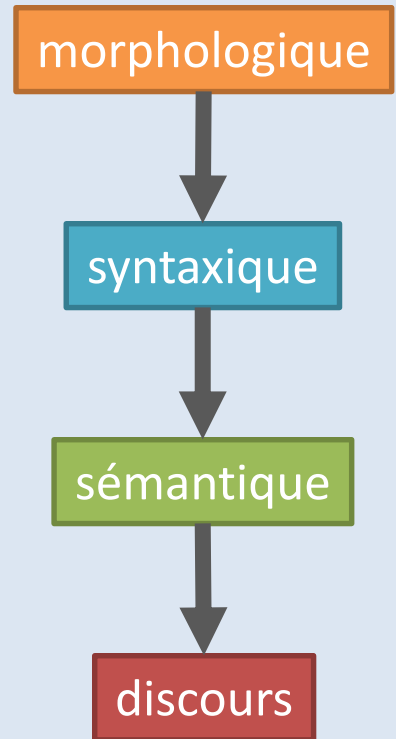
Firstly, we should look at the current state of happiness in human society. Finding out how happy people are and then using that as a benchmark for making their lives even happier is a good place to start. It would make sense to survey everyone about this first and find out what they think makes them happy.

The most common response for both men and women was that more sex would make them happier. Aside from this, the other immediate responses were:

Sex (in general) Family Life Friends Work A loving partner in life Good health Money/financial security Community Spirituality Break Free of Stress Being very wealthy My children Doing what you love Traveling The right job Growing old with someone

It seems that sex and family are the two most important factors in achieving happiness. This is supported by a lot of research into this area. For example, social relationships have been identified as an essential component of human well-being.

# Tâche : traduction automatique



# Tâches

- Dialogue humain / machine (oral ou écrit)
- Réponse à des questions
- Résumé, synthèse, simplification
- Recherche d'information, similarité entre documents
- ...

# Tâches

- Dialogue humain / machine (oral ou écrit)
- Réponse à des questions
- Résumé, synthèse, simplification
- Recherche d'information, similarité entre documents
- ...
- Sans oublier les tâches multimodales
  - Multimodal emotion recognition / sentiment analysis (vidéo + texte)
  - Visual question answering (image + texte)
  - Verifying Multimedia Use (image + texte, MediaEval 2016)
  - Image captioning
  - Image generation
  - ...

<http://nlpprogress.com/english/multimodal.html>

# Problèmes, enjeux

## Qu'est-ce qui rend un problème difficile ?

- Il est difficile à résoudre par un humain  
(cf. accord inter-annotateur)
- Il a beaucoup de classes (*multi-class*)  
(ex. entity linking)
- Les classes ne sont pas mutuellement exclusives (*multi-label*)  
(ex. : attribution de *tags*)
- Les classes sont réparties de façon très déséquilibrée (*unbalanced, skewed*)  
(ex. : spam)
- Sa représentation est complexe  
(ex. : phrase, document, multimodal)
- Certaines données sont manquantes ou bruitées (*missing data, noisy data*)  
(ex. : quasiment tous les problèmes de la vie réelle)

# Geekeries

En vrac





# Quiz

"Quelle est l'unique différence entre cette phrase et les autres ?"

“Quelle est l'unique différence entre cette phrase et les autres ?”

« Quelle est l'unique différence entre cette phrase et les autres ? »

# Quiz

Quelle est l'unique différence entre cette phrase et l'autre ?

Quelle est l'unique différence entre cette phrase et l'autre ?

# Quiz

Quelle est l'unique différence entre cette phrase et l'autre ?

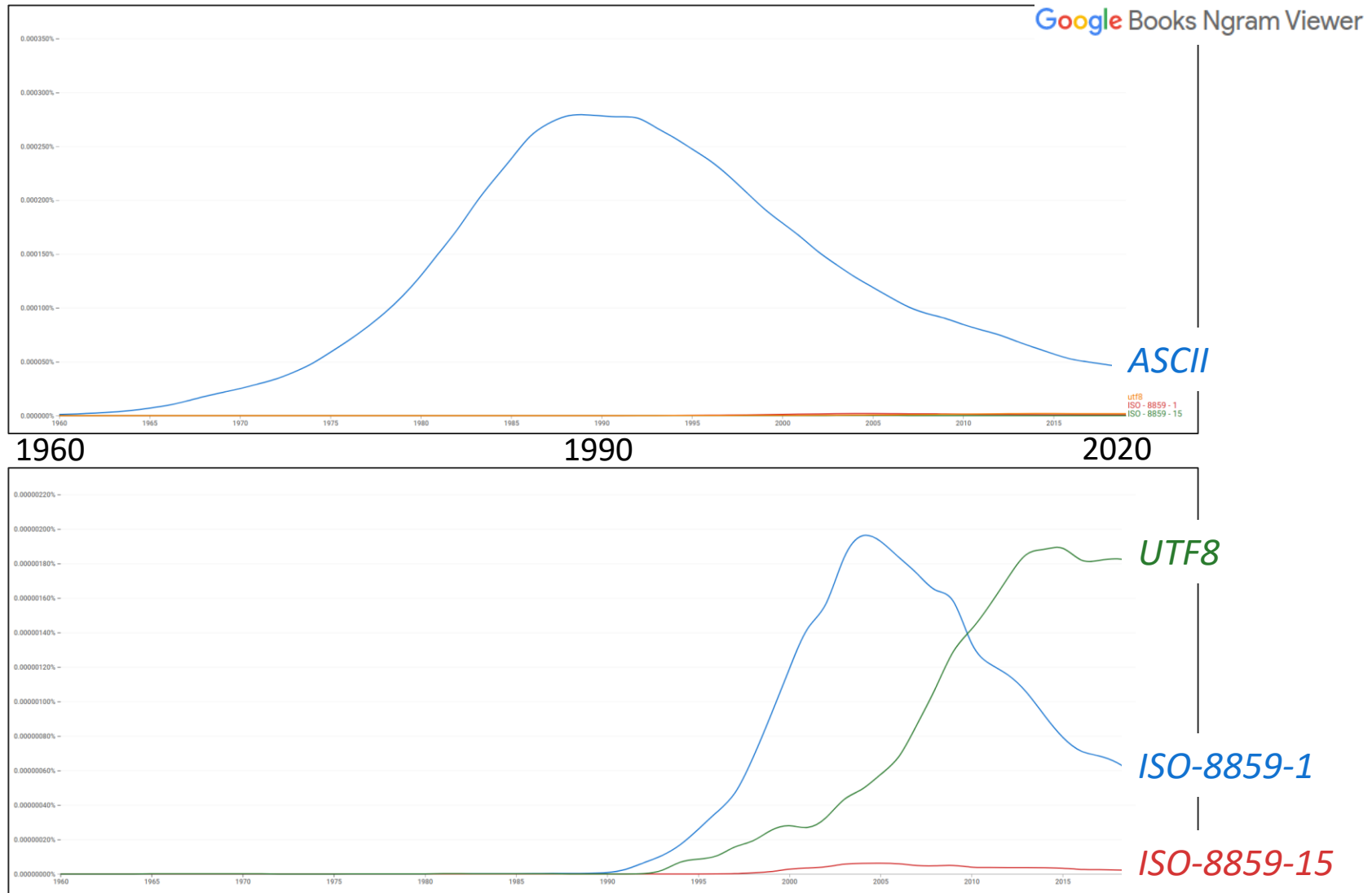
Quelle est l'unique différence entre cette phrase et l'autre ?

# Quiz

Quel est le problème ?

« prÃ©sident du PÃ©rou »

# Encodage



# Encodage

## ASCII

Codage de 128 caractères sur 7 bits  
(un octet avec dernier bit à 0)

**USASCII code chart**

					0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
b <sub>4</sub>	b <sub>3</sub>	b <sub>2</sub>	b <sub>1</sub>	Column	0	1	2	3	4	5	6	7
Row	↓	↓	↓	↓	0	1	2	3	4	5	6	7
0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p
0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q
0	0	1	0	2	STX	DC2	"	2	B	R	b	r
0	0	1	1	3	ETX	DC3	#	3	C	S	c	s
0	1	0	0	4	EOT	DC4	\$	4	D	T	d	t
0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	6	ACK	SYN	&	6	F	V	f	v
0	1	1	1	7	BEL	ETB	'	7	G	W	g	w
1	0	0	0	8	BS	CAN	(	8	H	X	h	x
1	0	0	1	9	HT	EM	)	9	I	Y	i	y
1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
1	0	1	1	11	VT	ESC	+	;	K	[	k	{
1	1	0	0	12	FF	FS	,	<	L	\	l	
1	1	0	1	13	CR	GS	-	=	M	]	m	}
1	1	1	0	14	SO	RS	.	>	N	^	n	~
1	1	1	1	15	SI	US	/	?	O	_	o	DEL

# Encodage

## ISO-8859-1 (« latin1 »)

Codage de 192 caractères sur 8 bits  
(un octet)

ISO/CEI 8859-1																
	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
0x	<i>positions inutilisées</i>															
1x																
2x	SP	!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{		}	~	
8x	<i>positions inutilisées</i>															
9x																
Ax	NBSP	ı	ç	£	¤	¥	¦	§	¨	©	ª	«	¬	-	®	¯
Bx	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
Cx	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
Dx	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
Ex	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
Fx	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

(source : Wikipedia)

# Encodage

## UTF-8

Codage sur 1, 2, 3 ou 4 octets  
(potentiellement jusqu'à 8)

Caractères codés	Représentation binaire UTF-8
U+0000 à U+007F	0xxxxxxx
U+0080 à U+07FF	110xxxxx 10xxxxxx
U+0800 à U+FFFF	1110xxxx 10xxxxxx 10xxxxxx
U+10000 à U+10FFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx

(source : Wikipedia)



# Quiz

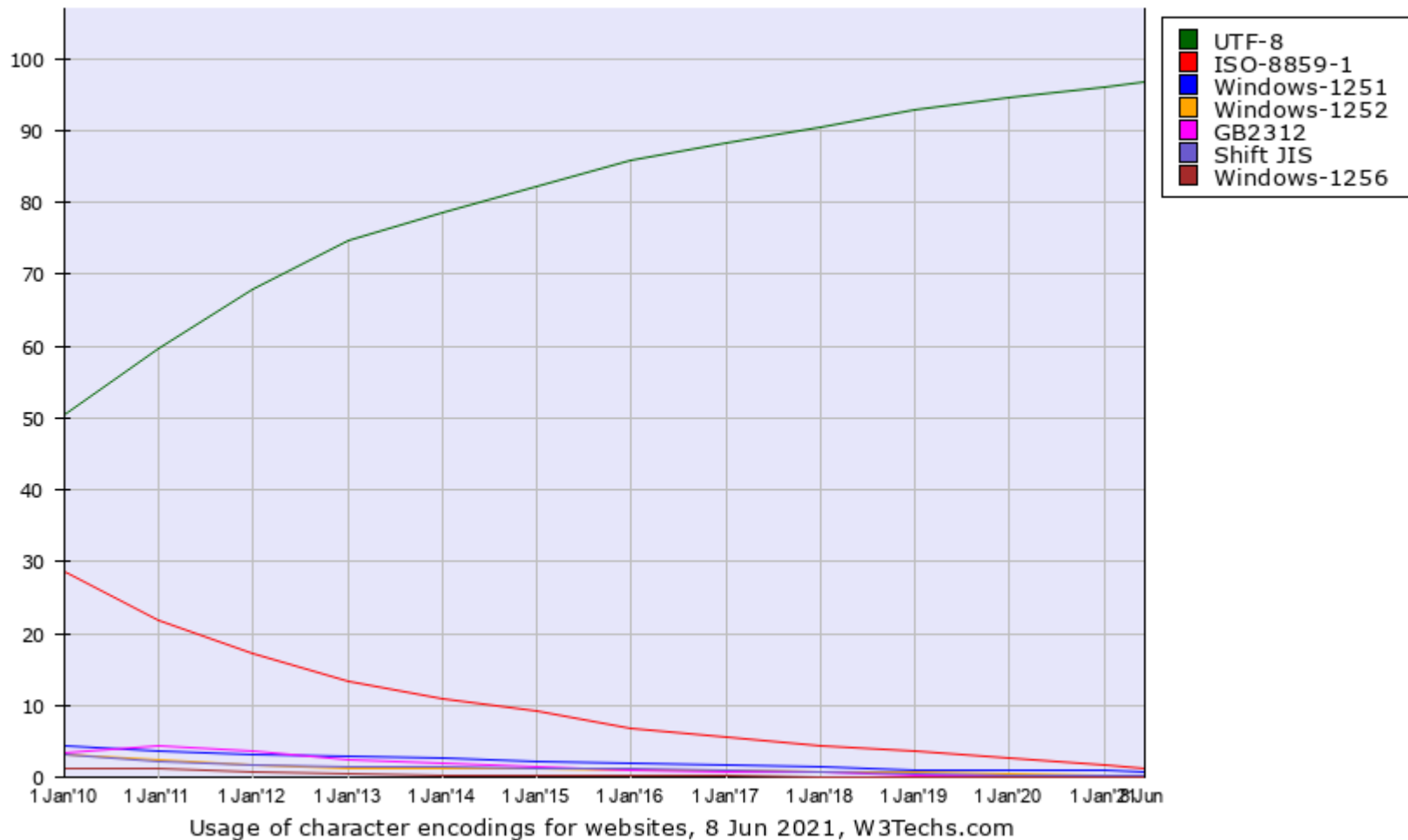
Et donc, quel est le problème ?

« prÃ©sident du PÃ©rou »

# Encodage

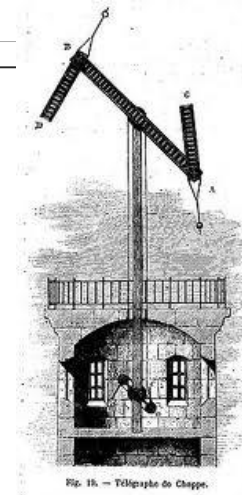
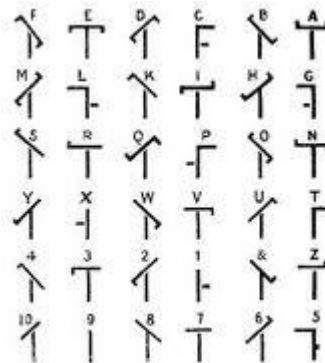
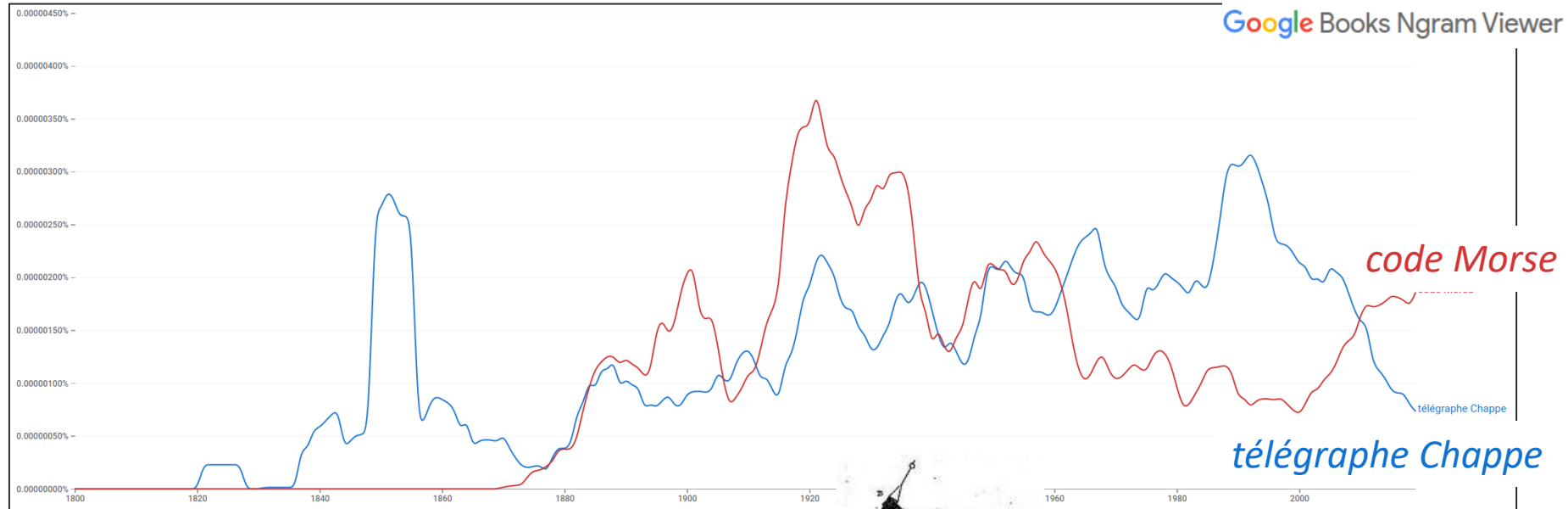
On peut espérer être bientôt débarrassé du problème...

[https://w3techs.com/technologies/history\\_overview/character\\_encoding/ms/y](https://w3techs.com/technologies/history_overview/character_encoding/ms/y)



# Encodage

PS : ça ne date pas de l'invention de l'informatique...



# Quiz

-DOCSTART- -X- O O

EU NNP I-NP I-ORG  
rejects VBZ I-VP O  
German JJ I-NP I-MISC  
call NN I-NP O  
to TO I-VP O  
boycott VB I-VP O  
British JJ I-NP I-MISC  
lamb NN I-NP O  
. . O O

Peter NNP I-NP I-PER  
Blackburn NNP I-NP I-PER

BRUSSELS NNP I-NP I-LOC  
1996-08-22 CD I-NP O

The DT I-NP O  
European NNP I-NP I-ORG  
Commission NNP I-NP I-ORG  
said VBD I-VP O  
on IN I-PP O  
Thursday NNP I-NP O

Voici le début du fichier d'entraînement de Conll 2003 (chunking et reconnaissance d'entités nommées, EN).

Vous avez une minute pour en extraire les classes d'entités nommées (dernière colonne) et leur nombre dans le corpus.

# Quiz

```
-DOCSTART- -X- O O
```

```
EU NNP I-NP I-ORG  
rejects VBZ I-VP O  
German JJ I-NP I-MISC  
call NN I-NP O  
to TO I-VP O  
boycott VB I-VP O  
British JJ I-NP I-MISC  
lamb NN I-NP O  
. . O O
```

Voici le début du fichier d'entraînement de Conll 2003 (chunking et reconnaissance d'entités nommées, EN).

Vous avez une minute pour en extraire les classes

```
> grep -v "DOCSTART-" eng.train.txt | sed '/^$/d' | awk '{print $4}'  
| sort | uniq -c | sort -nr  
169578 O  
11128 I-PER  
10001 I-ORG  
8286 I-LOC  
4556 I-MISC  
37 B-MISC  
24 B-ORG  
11 B-LOC
```

# Quiz

```
-DOCSTART- -X- O O
```

```
EU NNP I-NP I-ORG  
rejects VBZ I-VP O  
German JJ I-NP I-MISC  
call NN I-NP O  
to TO I-VP O  
boycott VB I-VP O  
British JJ I-NP I-MISC  
lamb NN I-NP O  
. . O O
```

Voici le début du fichier d'entraînement de Conll 2003 (chunking et reconnaissance d'entités nommées, EN).

Vous avez une minute pour en extraire les classes

```
> grep -v "DOCSTART-" eng.train.txt | sed '/^$/d' | awk '{print $4}'  
| sort | uniq -c | sort -nr  
169578 O  
11128 I-PER  
10001 I-ORG  
8286 I-LOC  
4556 I-MISC  
37 B-MISC  
24 B-ORG  
11 B-LOC
```

Apprenez les commandes unix !

- grep, tr
- sort, uniq
- cut
- awk
- sed

# Quiz

```
-DOCSTART- -X- O O
```

```
EU NNP I-NP I-ORG  
rejects VBZ I-VP O  
German JJ I-NP I-MISC  
call NN I-NP O  
to TO I-VP O  
boycott VB I-VP O  
British JJ I-NP I-MISC  
lamb NN I-NP O  
. . O O
```

Voici le début du fichier d'entraînement de Conll 2003 (chunking et reconnaissance d'entités nommées, EN).

Et calculez les statistiques des corpus que vous manipulez, avant TOUTE chose :

- Répartition des classes
- Nombre de classes par objet
- Nombre de mots total, par document, par phrase
- ...

```
> grep -v "DOCSTART" | sort | uniq -c  
169578 O  
11128 I-PER  
10001 I-ORG  
8286 I-LOC  
4556 I-MISC  
37 B-MISC  
24 B-ORG  
11 B-LOC
```

- grep, tr
- sort, uniq
- cut
- awk
- sed

classes

\$4}'

# Quiz

Quels sont les points communs entre ces erreurs ?

Erreur pytorch

```
RuntimeError: size mismatch, m1: [a x b], m2: [c x d]
```

Erreur regex

```
AttributeError: 'NoneType' object has no attribute 'group'
```

Erreur partout

```
IndexError: list index out of range
```

```
FileNotFoundError: [Errno 2] No such file or directory
```



# Quiz

Quels sont les points communs entre ces erreurs ?

1. Elles peuvent arriver après des heures d'entraînement
2. Les corriger peut vous prendre des heures
3. Vous auriez peut-être pu les éviter en utilisant **assert**

`AttributeError: 'NoneType' object has no attribute 'group'`

```
assert isfile(res_file)
```

Erreur partout

`IndexError: assert outputs.shape == (batch_size, seq_len, dim)`

`FileNotFoundError: assert len(fields) > 0, fields`

# Traitement automatique de la langue Évaluation

Xavier Tannier



<https://xkcd.com/2451/>

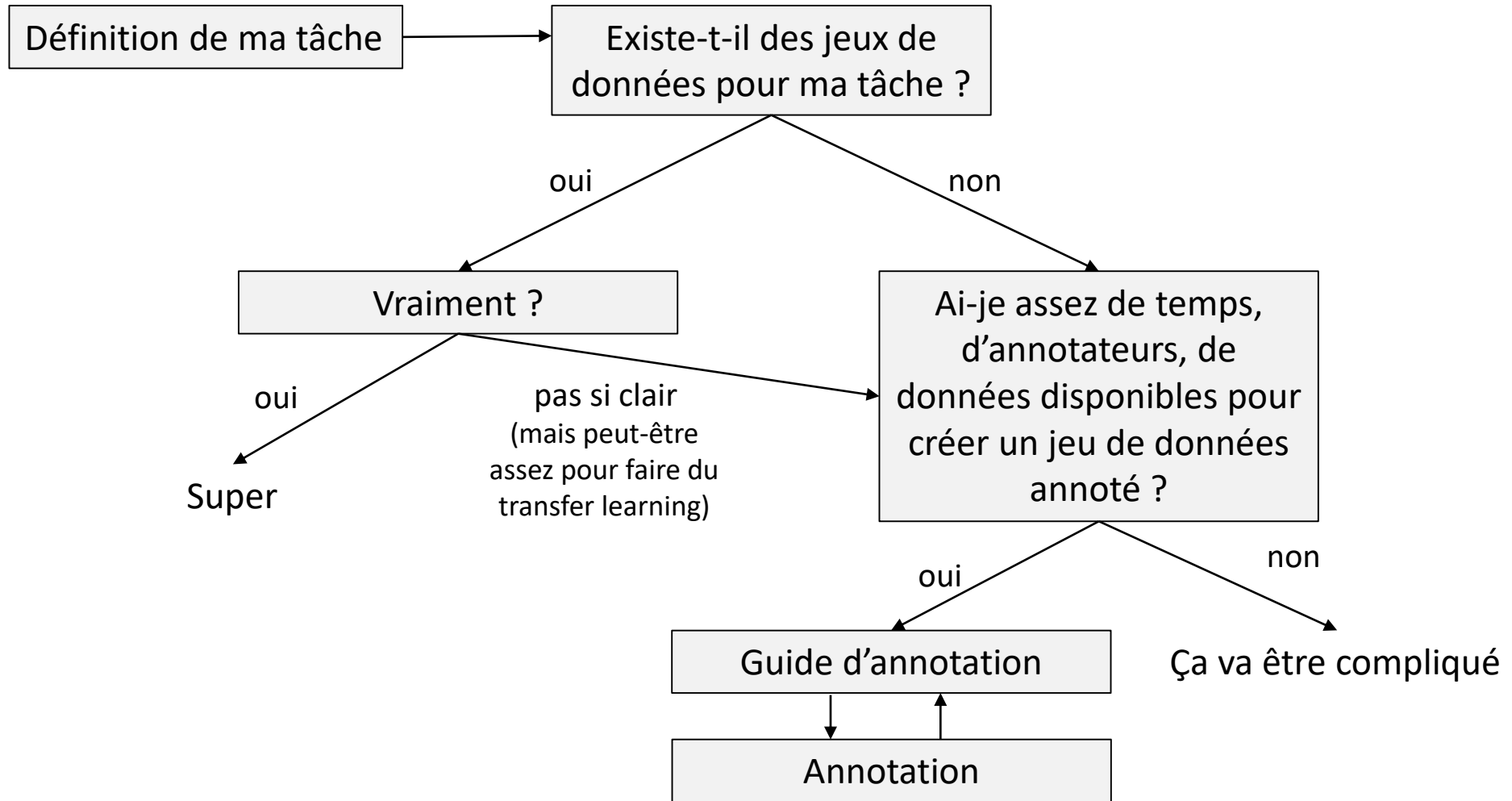


# Évaluation

- Une bonne évaluation, c'est :
  - Une bonne référence
  - Un périmètre bien défini (pour estimer la généralisabilité et la robustesse)
  - Une métrique appropriée
  - Un protocole défini et respecté
- Et puis se poser les questions sur :
  - L'interprétabilité
  - Le caractère FAIR des données
  - Le coût environnemental
  - La reproductibilité
  - L'éthique

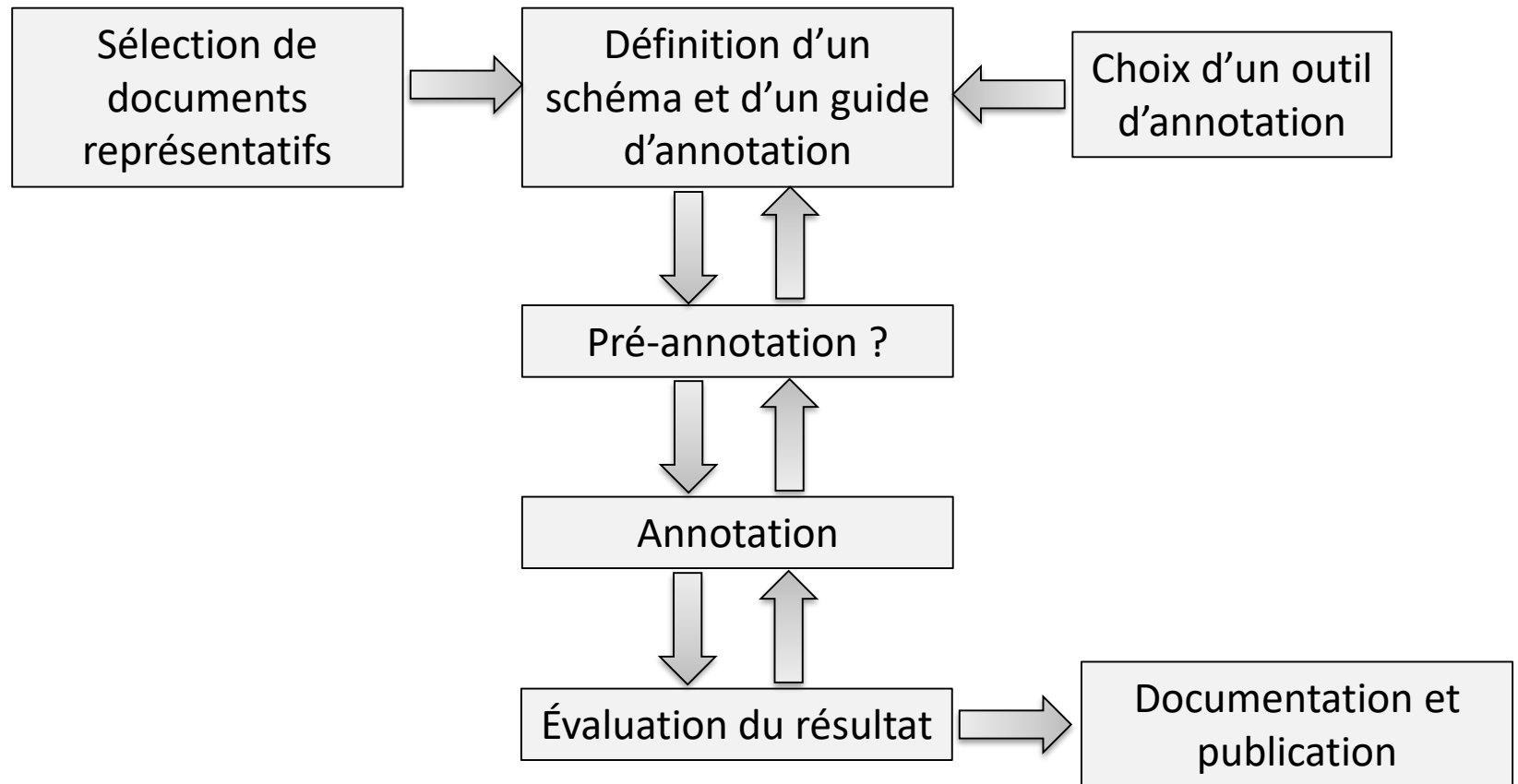
**Une bonne évaluation, c'est  
un bon jeu de données annotées**

# Construction du jeu de données



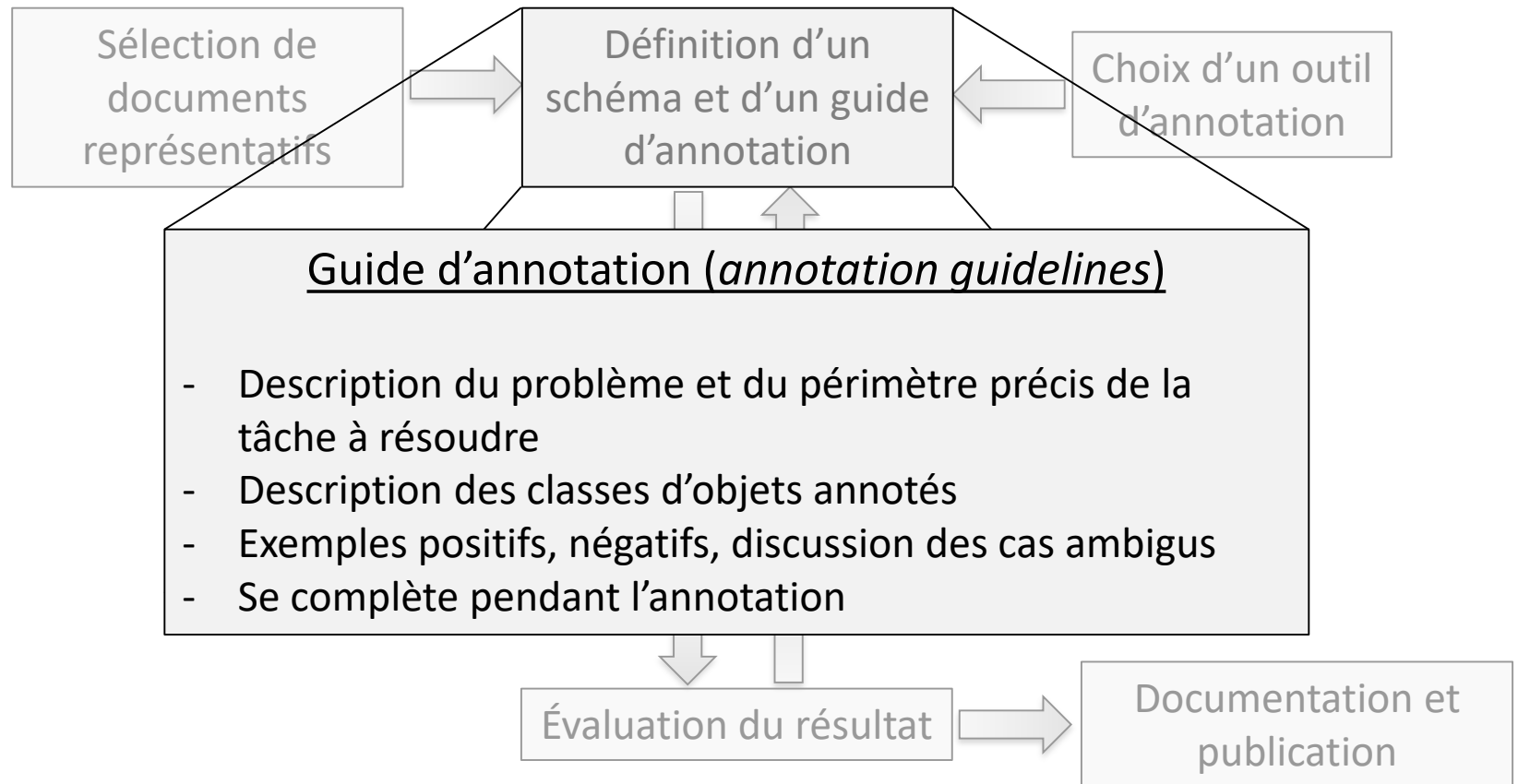
# Construction du jeu de données

La configuration idéale (version courte)



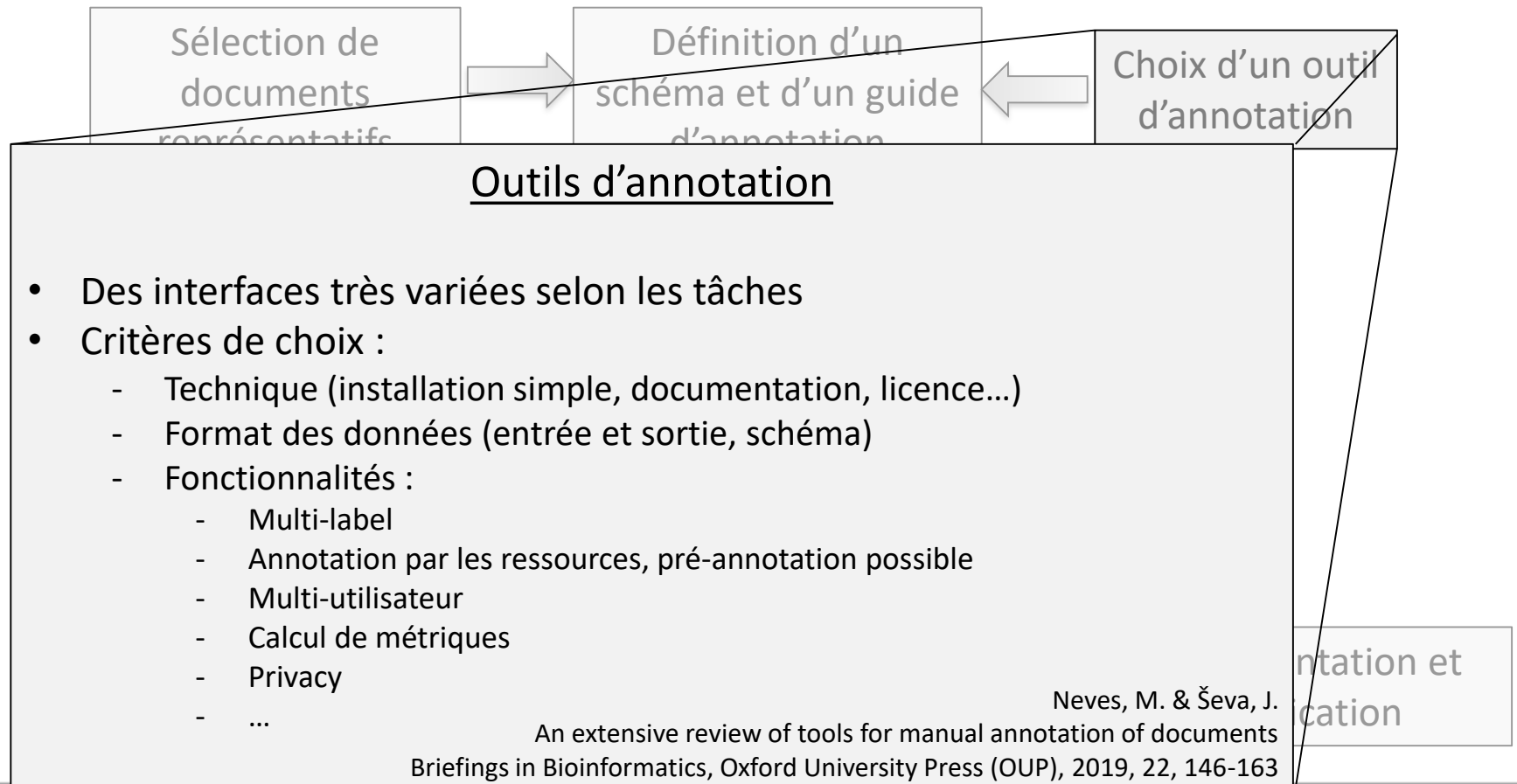
# Construction du jeu de données

La configuration idéale (version courte)



# Construction du jeu de données

La configuration idéale (version courte)





# Construction du jeu de données

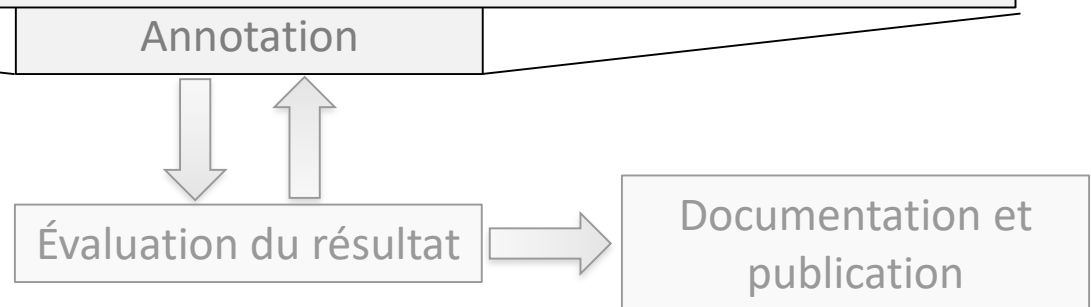
La construction d'un jeu de données (Annotation)

## Annotation

- Plusieurs personnes, et pas celles qui vont concevoir le système
- Procédure :
  1. Annotent des documents ensemble
  2. Annotent séparément mais avec du chevauchement par paire d'annotateurs
  3. Se retrouvent régulièrement pour régler les désaccords (adjudication) et vérifient qu'ils ne varient pas
- Si crowd-sourcing, ajouter des contrôles qualité un peu partout

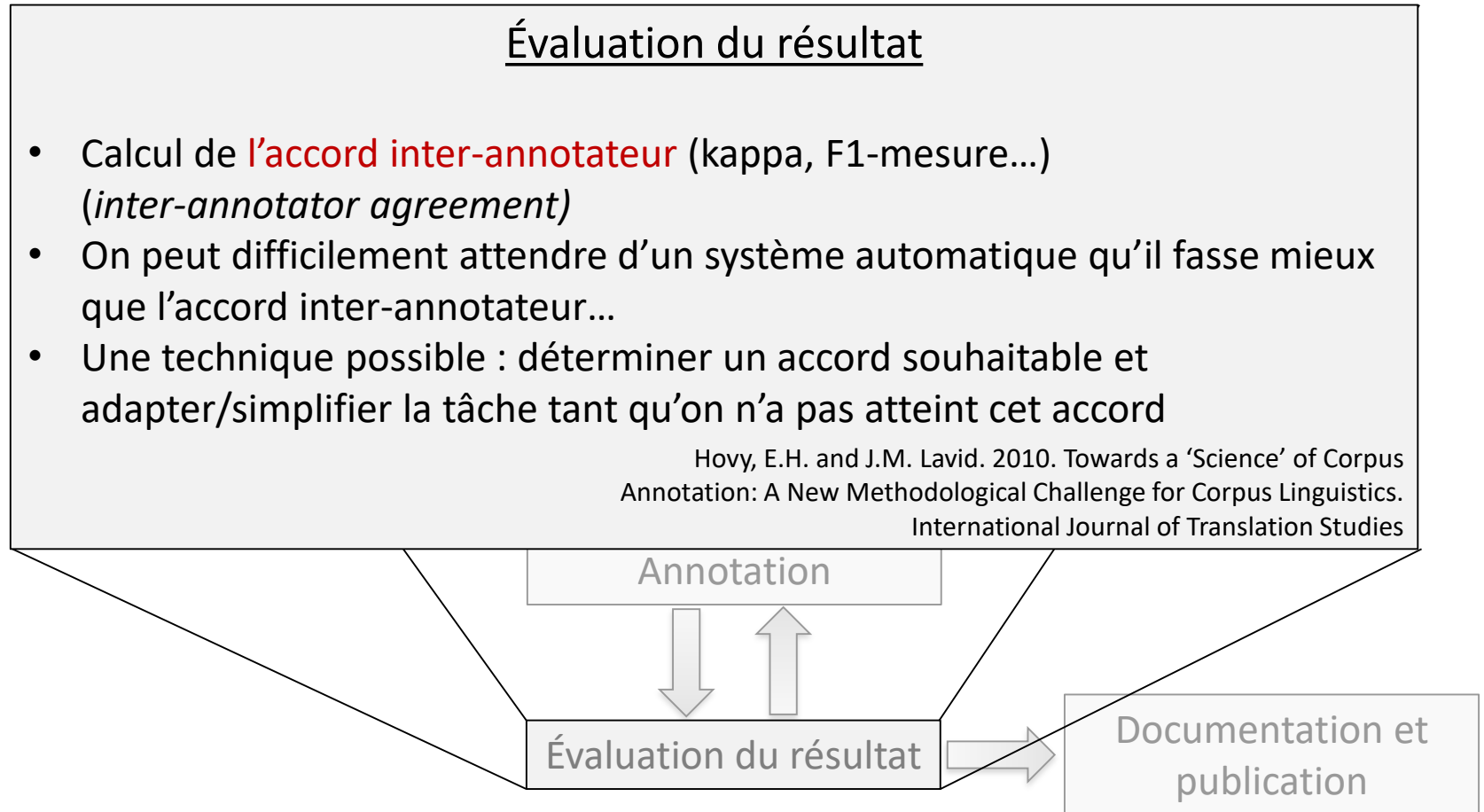
Les ressources annotées, un enjeu pour l'analyse de contenu :  
vers une méthodologie de l'annotation manuelle de corpus.  
Thèse de doctorat.  
Karën Fort

M. Sabou, K. Bontcheva, L. Derczynski, A. Scharl  
Corpus Annotation through Crowdsourcing:  
Towards Best Practice Guidelines  
LREC 2014 (LREC'14), 2014



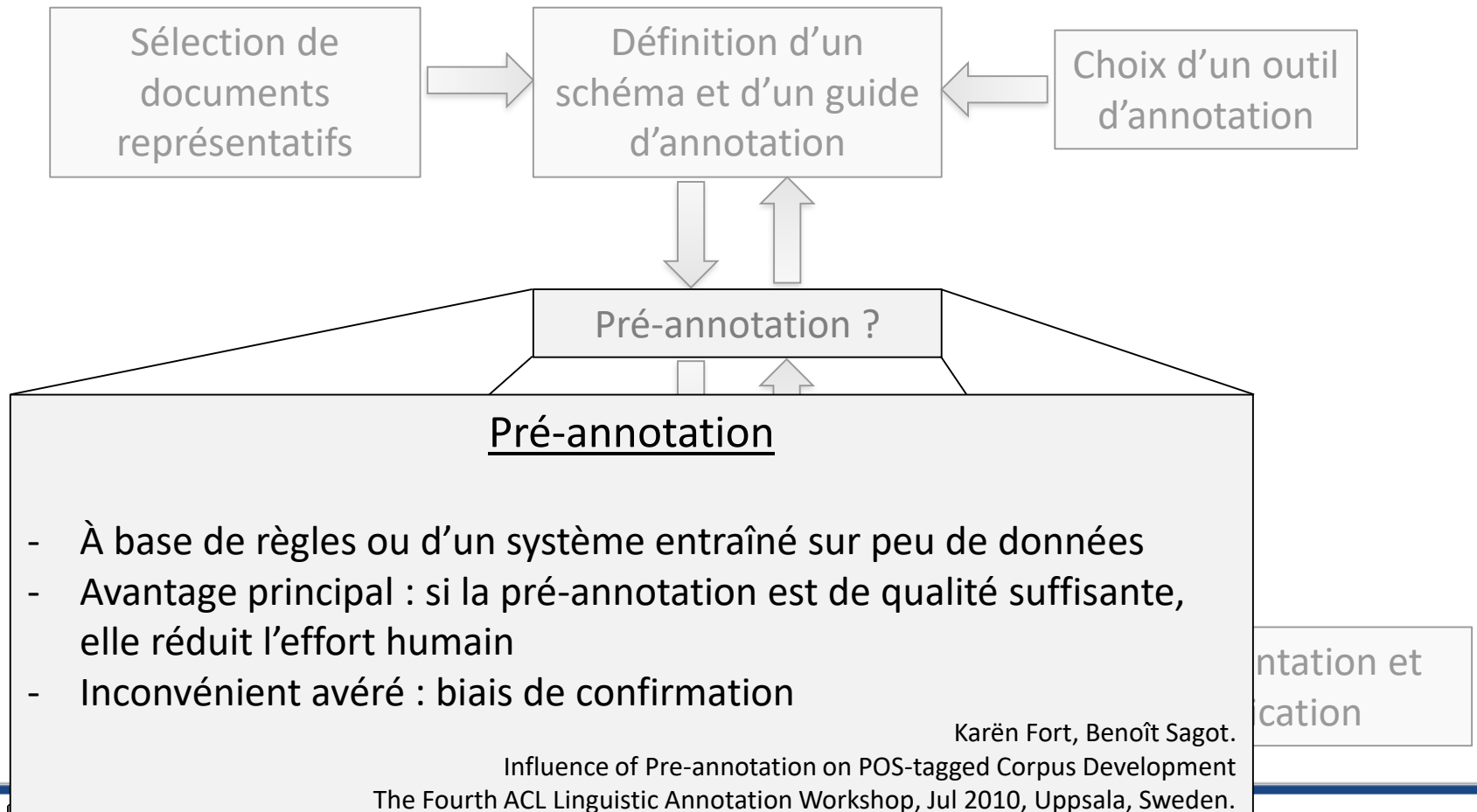
# Construction du jeu de données

La configuration idéale (version courte)



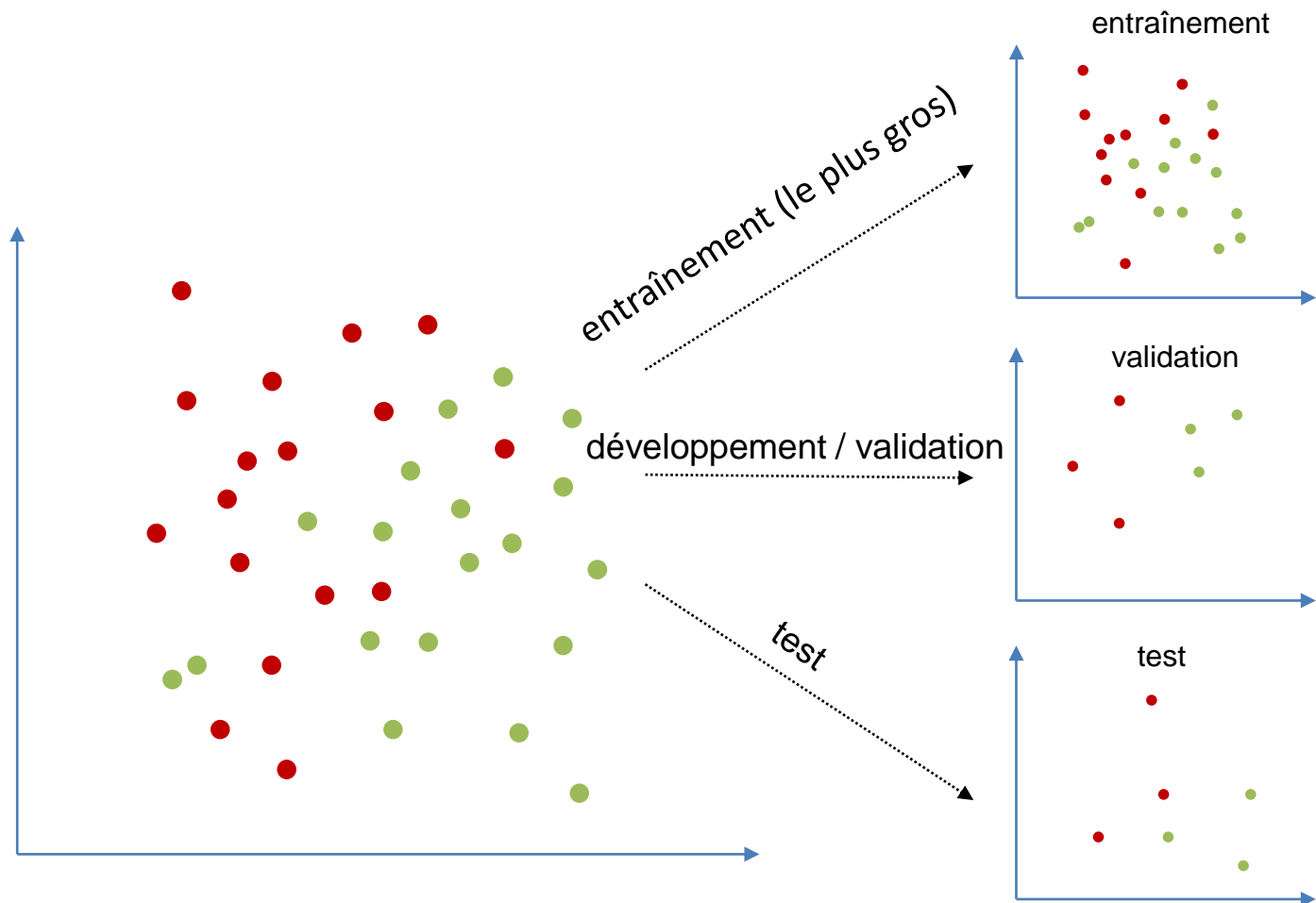
# Construction du jeu de données

La configuration idéale (version courte)



# À quoi sert le jeu de données ? (tâches supervisées)

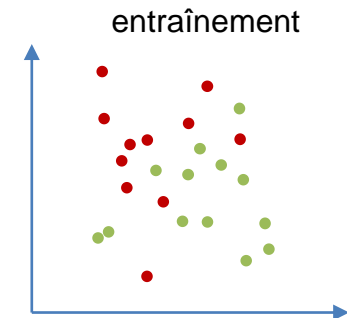
On sépare en général le jeu de données annotées en trois :



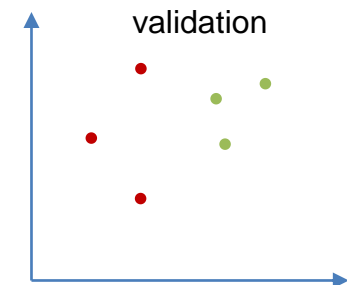
# À quoi sert le jeu de données ? (tâches supervisées)

On sépare en général le jeu de données annotées en trois :

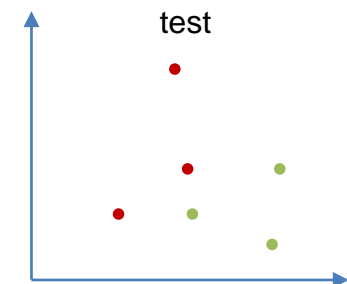
Pour entraîner le modèle



Pour trouver les meilleurs paramètres  
/ le meilleur modèle, savoir quand  
arrêter l'entraînement,  
éviter le surentraînement, etc.



Pour évaluer les performances  
du modèle choisi sur des  
données qu'il n'a jamais vues.  
**Ne doit jamais être regardé  
ni utilisé pendant la  
conception !**

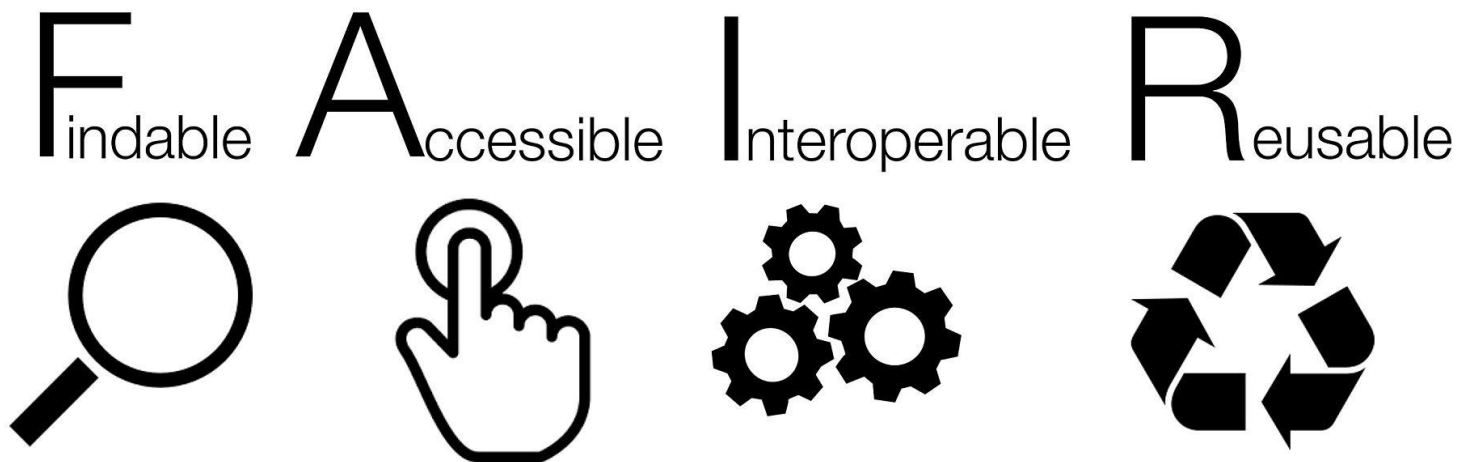


# Les bonnes propriétés du jeu de données

Pour une bonne réutilisabilité :

- Le **périmètre** du jeu de données doit être clair et documenté
- La **qualité** des annotations doit être **documentée**
- Les données doivent être **représentatives** de la tâche considérée (dans le périmètre considéré)
- Les formats de données doivent être **standard** (si un standard existe)
- Le **jeu de test** doit être défini (et respecté)
- Les données doivent être rendue **disponibles** (si légalement possible)

# FAIR data



SangyaPundir, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=53414062>

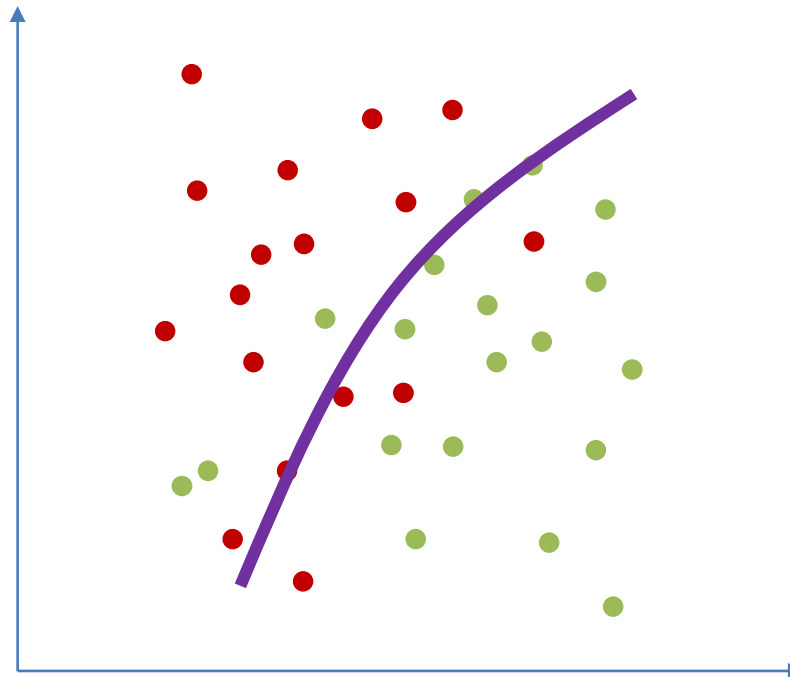
Kai-Wei Chang, Vicente Ordonez, Margaret Mitchell, Vinodkumar Prabhakaran  
Tutorial: Bias and Fairness in Natural Language Processing  
EMNLP 2019

**Une bonne évaluation, ce sont  
de bonnes métriques**



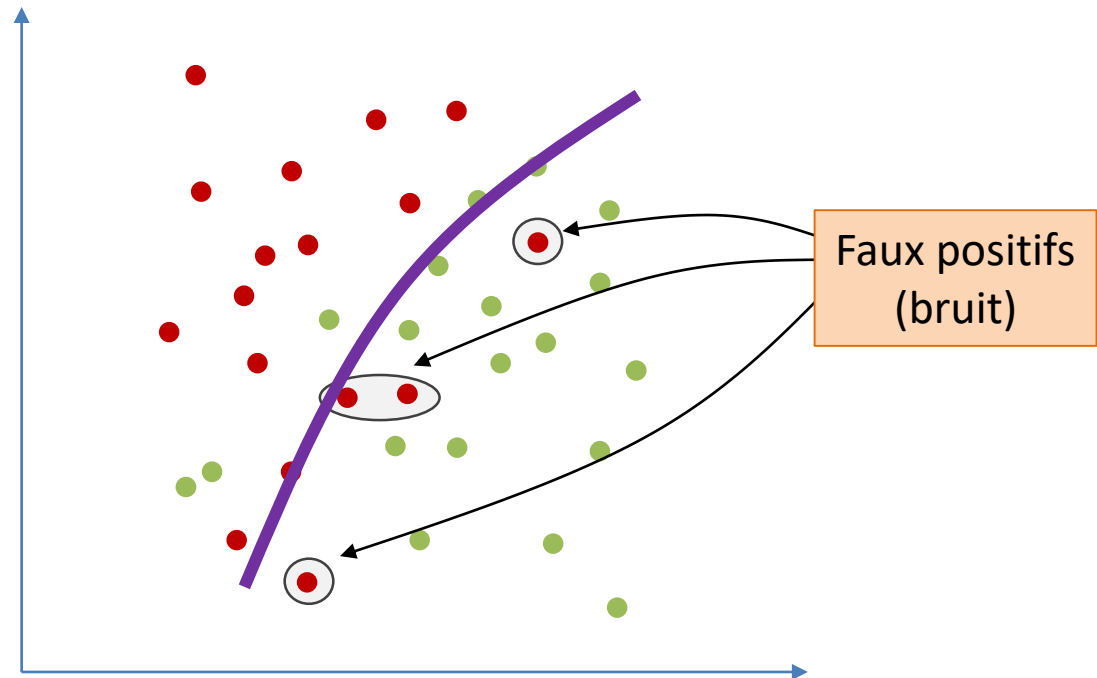
# Évaluation en extraction d'information

Prédiction du système



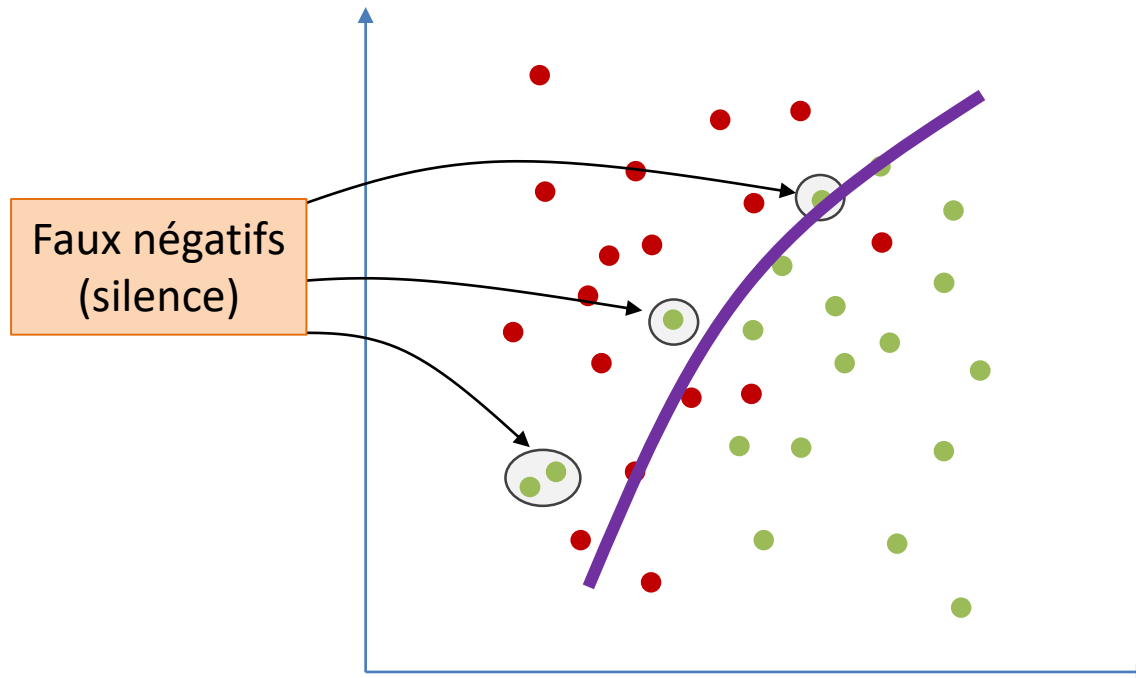
# Évaluation en extraction d'information

classe « VERT »



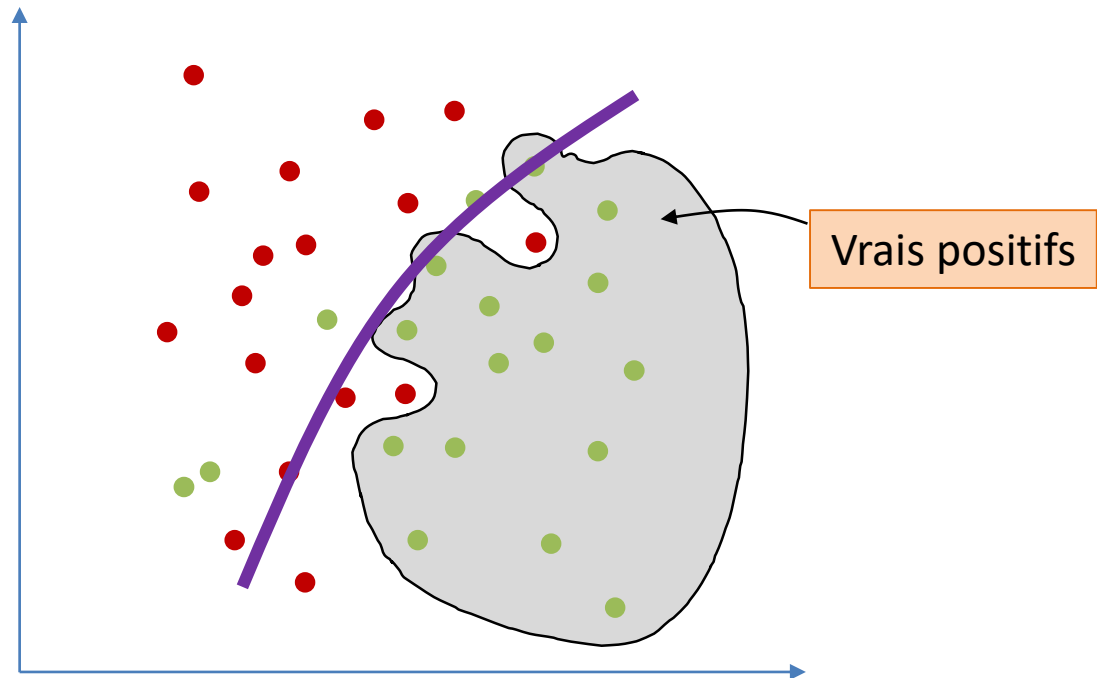
# Évaluation en extraction d'information

classe « VERT »



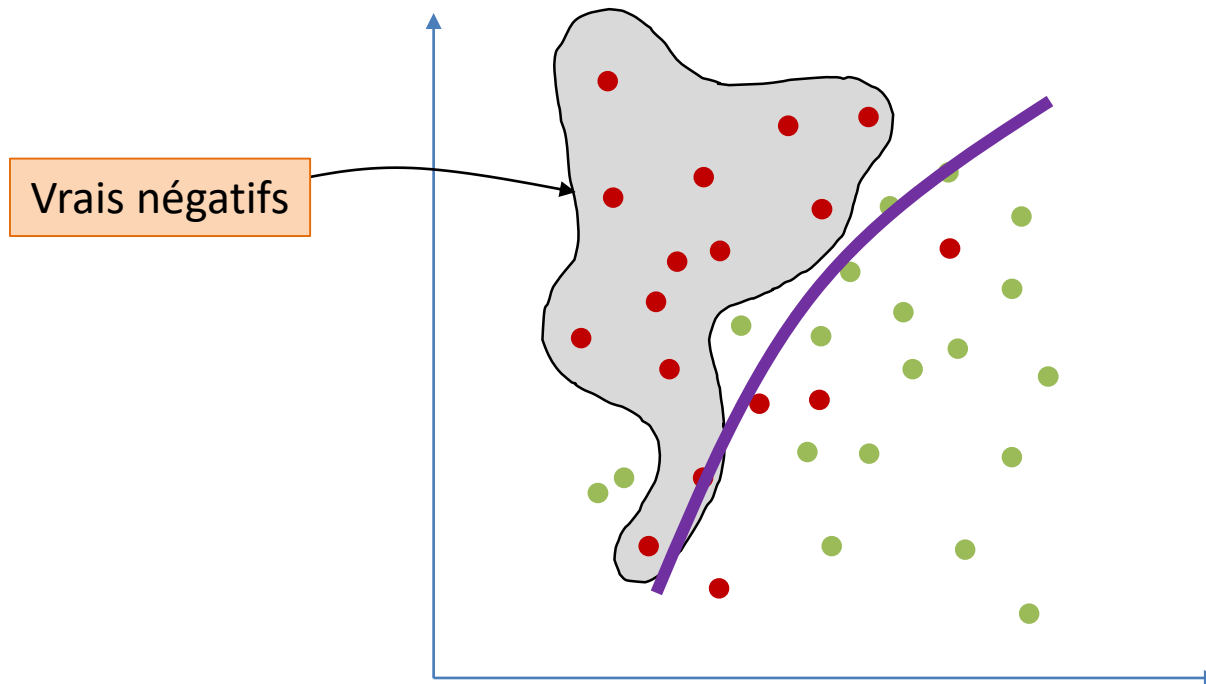
# Évaluation en extraction d'information

classe « VERT »



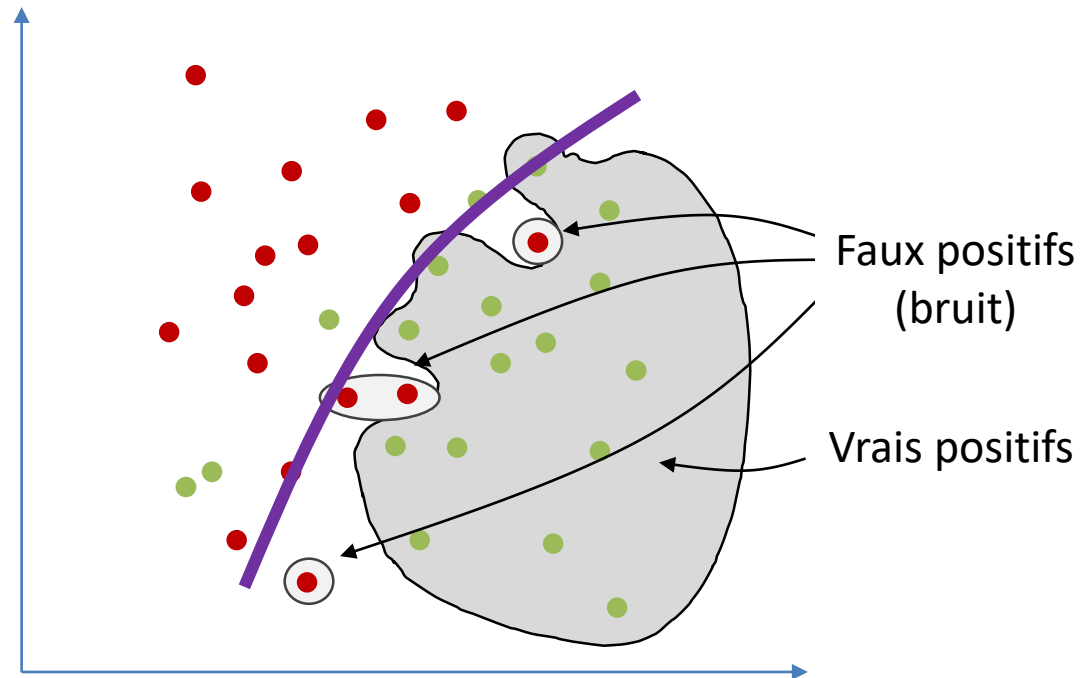
# Évaluation en extraction d'information

classe « VERT »



# Évaluation en extraction d'information

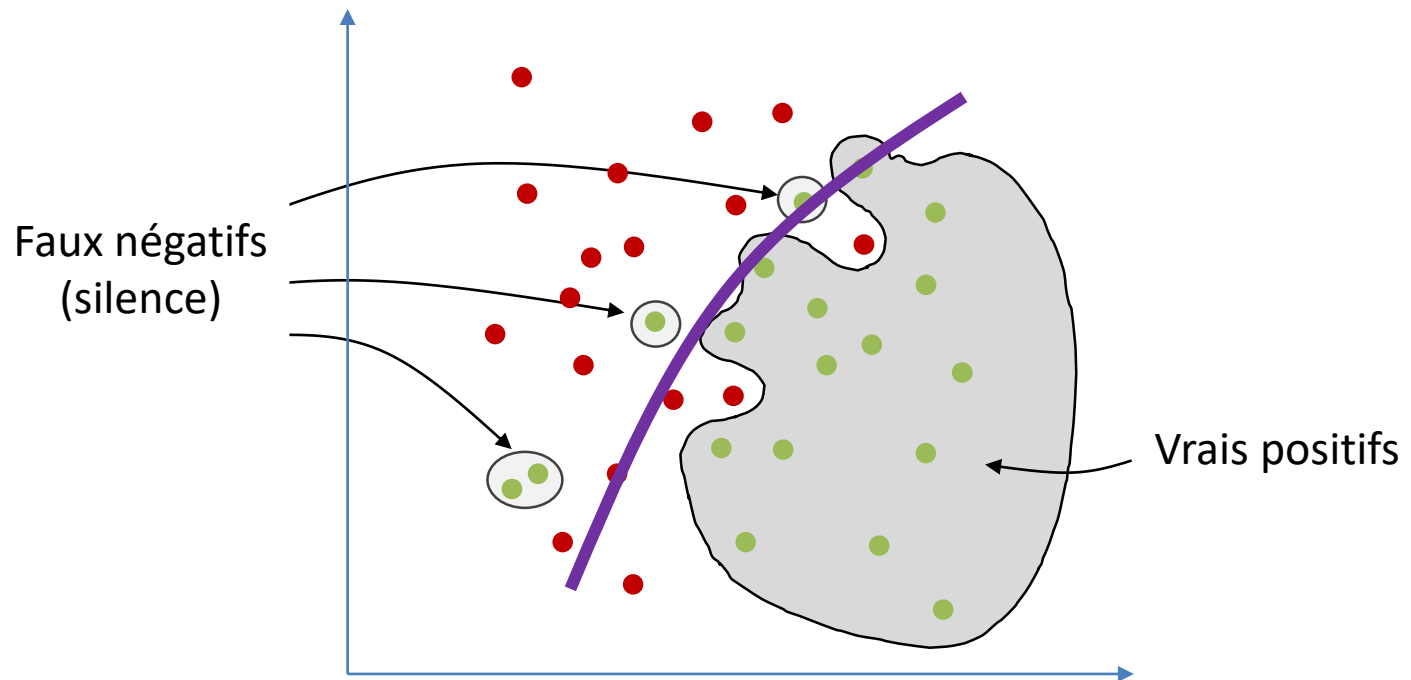
classe « VERT »



taux de positifs dans la prédiction « VERT » = Précision =  $\frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux positifs}} = 1 - \text{bruit}$

# Évaluation en extraction d'information

classe « VERT »



taux de positifs dans la référence « VERT » =  $\text{Rappel} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux négatifs}} = 1 - \text{silence}$

# Évaluation en extraction d'information

- **Précision** : le système a-t-il raison quand il annonce la classe ?
- **Rappel** (ou sensibilité) : le système manque-t-il beaucoup d'éléments de la classe ?
- **Spécificité** =  $\frac{\text{Vrais négatifs}}{\text{Vrais négatifs} + \text{Faux positifs}}$  le système évite-t-il les fausses alarmes ?
- **F-mesure** = moyenne harmonique de précision et rappel  
$$= \frac{(1+\beta^2) \times (\text{Précision} \times \text{Rappel})}{(\beta^2 \times \text{Précision} + \text{Rappel})}$$
- **F<sub>1</sub>-mesure** =  $2 \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$



# Évaluation en extraction d'information

- Classification multi-classes : matrice de confusion

		Prédictions		
		Classe 1	Classe 2	Classe 3
Référence	Classe 1	125	9	6
	Classe 2	13	110	52
	Classe 3	6	31	95

# Évaluation en extraction d'information

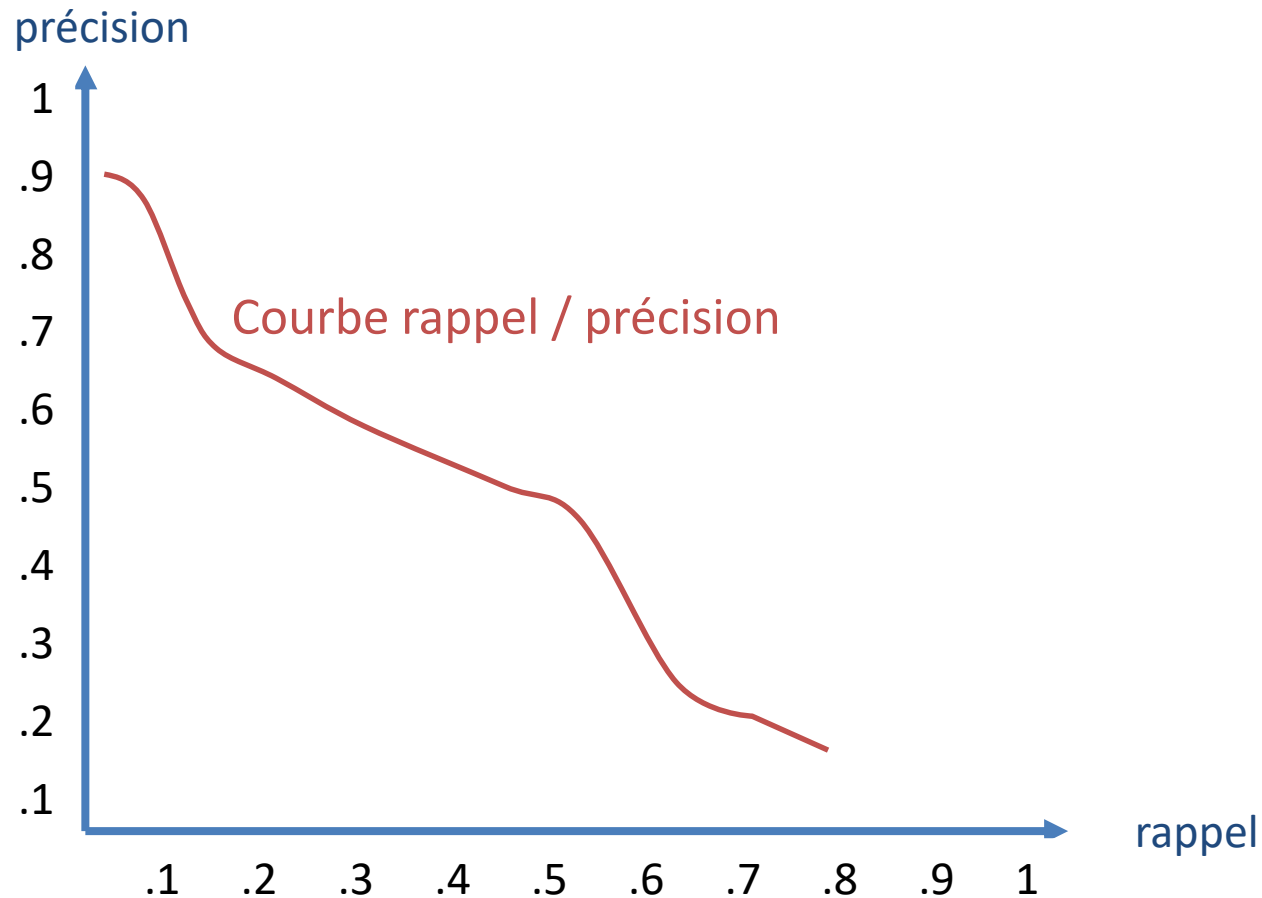
- **Accuracy** = taux de bonnes réponses, mais ATTENTION à son interprétation !

		Prédictions	
		Pas malade	Malade
Référence	Pas malade	19064	89
	Malade	110	231

- Accuracy =
- Rappel de la classe malade =
- Précision de la classe malade =

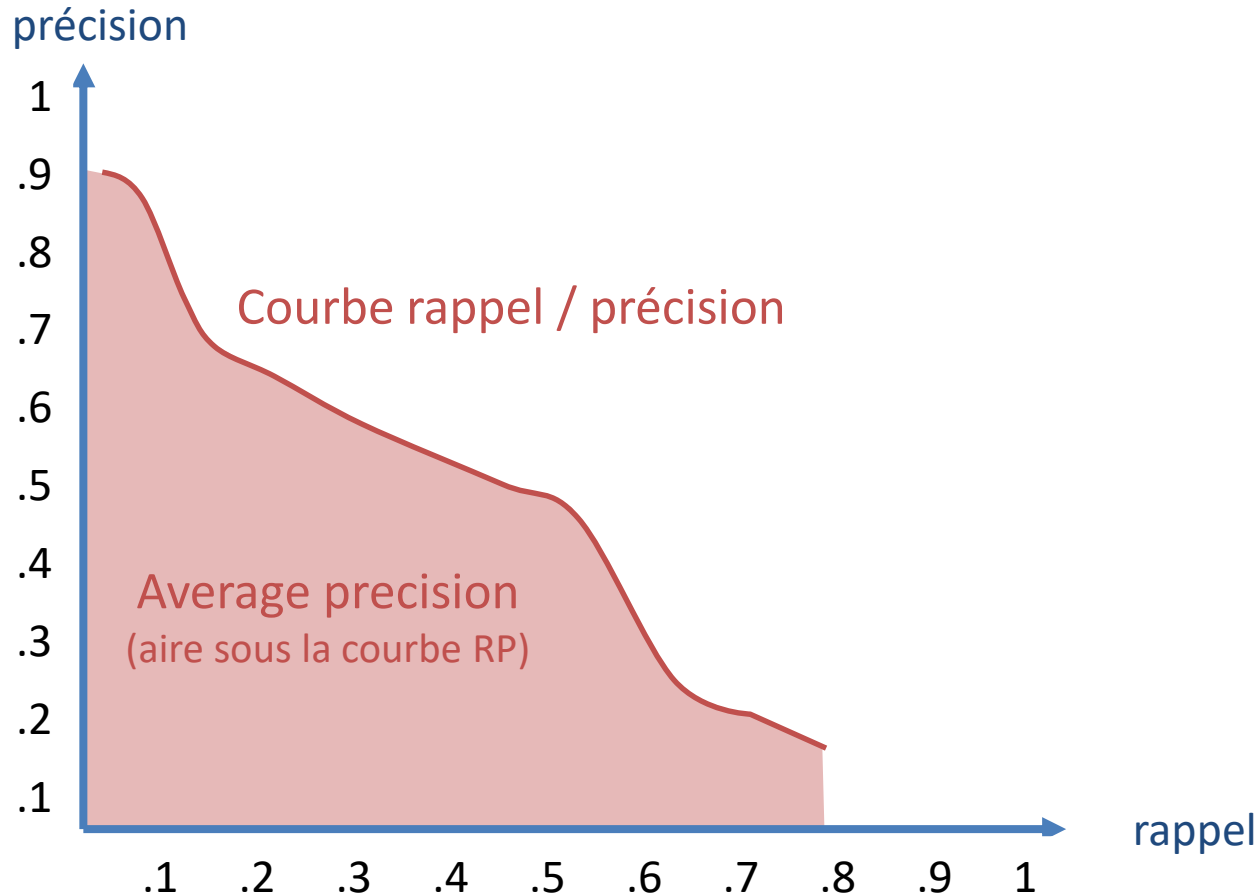
# Évaluation en extraction d'information

- Évaluation de listes ordonnées : tenir compte du rang des réponses



# Évaluation en extraction d'information

- Évaluation de listes ordonnées : tenir compte du rang des réponses



# Évaluation en extraction d'information

- Évaluation de listes ordonnées : tenir compte du rang des réponses
  - Précision après  $n$  éléments ( $P@5$ ,  $P@10...$ )
  - Rappel après  $k$  éléments, où  $k$  est le nombre d'éléments qu'il fallait renvoyer
  - AUROC
  - Mean Reciprocal Rank (MMR)
  - Mean Average Precision (MAP)

# Évaluation en traduction / génération

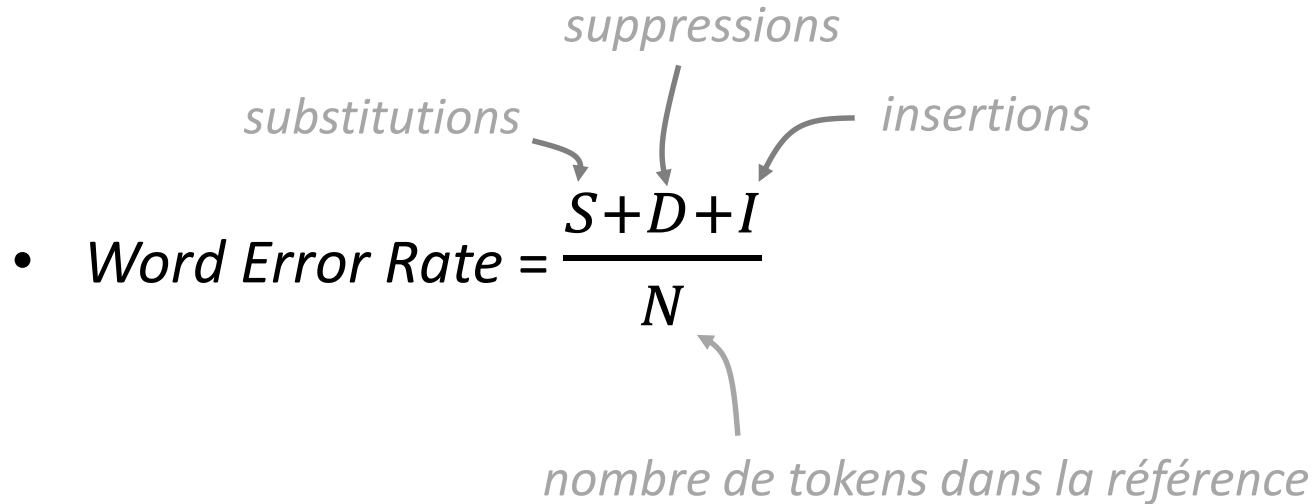
- BLEU (Bilingual Evaluation Understudy) : comparaison des références et des résultats du système en termes de n-grammes en commun
- METEOR (Metric for Evaluation of Translation with Explicit ORdering) : tient compte des synonymes ou des racines de mots
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation) : métrique plus orientée vers le rappel
  
- Ces métriques sont controversées et d'autres ont été proposées, sans que la communauté parvienne à un réel consensus

# Évaluation en reconnaissance de la parole

*substitutions*      *suppressions*      *insertions*

- $Word\ Error\ Rate = \frac{S+D+I}{N}$

*nombre de tokens dans la référence*



# Évaluation des tâches non supervisées

- Pour les modèles de langue, la **perplexité** : à quel point le modèle est « surpris » de voir des nouvelles phrases (il est surpris s'il a assigné une faible probabilité aux suites de mots)
- Pour le clustering, par exemple
  - La **silhouette** : est-ce que la cohésion des clusters (similarité intra-cluster) est plus forte que la séparation (distance avec les clusters voisins)
  - La **cohérence**, par exemple le degré de similarité entre les mots bien classé d'un topic dans le topic modeling

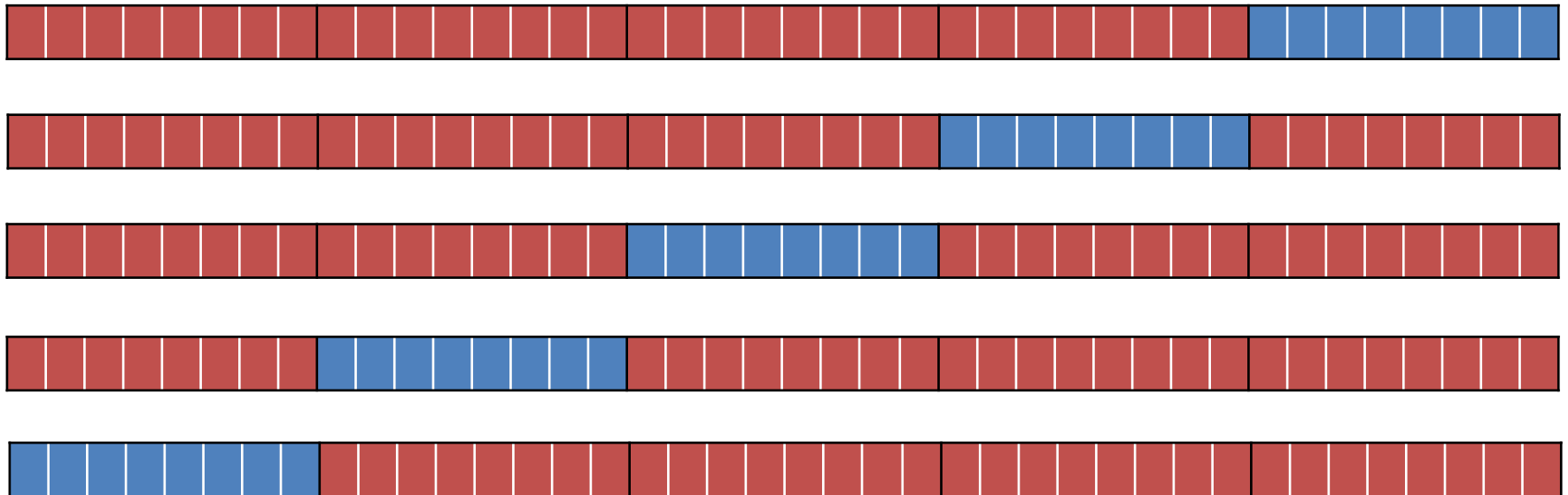
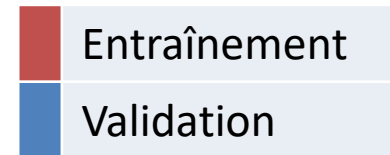


# Évaluation intrinsèque vs extrinsèque

- Évaluation **intrinsèque** = mesure de la performance d'un système sur la tâche elle-même
- Évaluation **extrinsèque** = mesure de l'apport d'un système sur des tâches aval.
  - Extrinsèque mais « **interne** » au TAL
    - Stemming : mesure de l'apport sur la recherche d'information
    - Embeddings : mesure de l'apport sur l'extraction d'information
    - Recherche d'information : mesure de l'apport sur la réponse à des questions
  - Extrinsèque **applicatif**
    - Entités nommées dans des comptes-rendus médicaux : meilleure sélection de cohortes de patients ou de cas similaire ?
    - Analyse de sentiment : meilleur ciblage des actions correctives ?
    - Système de recommandation : augmentation du panier moyen ?

# Validation croisée

- Pour les approches par apprentissage, possibilité de « faire tourner » le jeu de validation
- Exemple : validation croisée à 5 plis (*5-fold cross-validation*)
  - On sépare le jeu de données en 5 parties égales
  - On réalise 5 expériences **indépendantes**



# Validation croisée

- Intérêts

- Évaluer un système sans « gâcher » un jeu de validation
- Utile pour les **petits jeux de données**
- Permet d'estimer la performance de validation d'un système avec des mesures de biais et de **variance**
  - Si les résultats varient peu d'un pli à l'autre, c'est bon signe pour la représentativité du jeu de données et la généralisabilité du modèle
  - Si les résultats varient beaucoup, c'est mauvais signe mais on l'a repéré, alors qu'un jeu de validation unique ne l'aurait pas permis

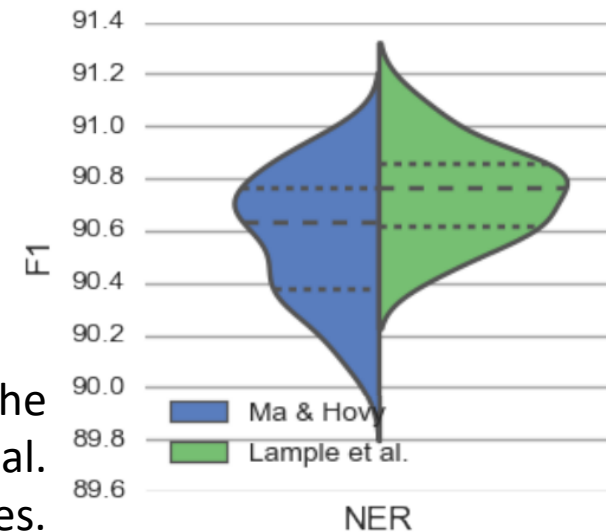
- Limites

- Difficile de déduire les **hyperparamètres** à choisir au final
- Ne permet pas de conclure formellement sur les performances réelles du modèle (pour cela on a besoin d'un **jeu de test** distinct)
- $k$  plis =  $k$  fois plus **coûteux** !

# Un run n'est pas assez

- Les résultats peuvent dépendre de l'initialisation aléatoire de vos réseaux de neurones
- Fonder vos conclusions sur un seul *run* peut :
  - Masquer des variations importantes, symptômes d'un problème dans votre modèle (difficulté à généraliser)
  - Conduire à des conclusions erronées sur le positionnement par rapport à l'état de l'art
- Idéalement il faut présenter la distribution de vos résultats sur plusieurs *runs* (5, 30... avec un *violin plot* ou un *box plot*)

Distribution of scores for re-running the system by Ma and Hovy (left) and Lample et al. (right) multiple times with different seed values.



Dashed lines indicate quartiles

**Une bonne évaluation,  
c'est aussi...**

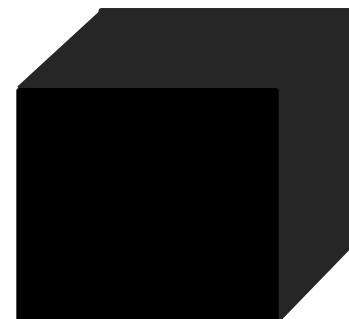
# L'interprétabilité

Le système sait-il « expliquer » ses prédictions ?

- Quelle partie de l'entrée a conduit à la prédiction ?
  - Saillance et perturbations adversariales
  - Support aux réponses à des questions

- Quelles règles de décision ont conduit à la prédiction ?
- Quels exemples du jeu d'entraînement ont causé la prédiction ?

- Intégration de la notion d'interprétabilité dans les modèles (mise en abyme)



# Le coût environnemental

- Des outils existent pour faire une estimation approximative de l'impact carbone d'un algorithme.
  - Estimation à partir de la description de l'outil (ex. Green Algorithms)
  - Mesure directe (ex. Carbon Tracker)
- Mais ce ne sont pour l'instant que des indications très imprécises, tenant compte du temps de l'expérience, mais pas
  - Du cycle de vie du matériel (de la fabrication au recyclage)
  - Du lieu de l'expérience (impact carbone de l'électricité en Allemagne >> celui de la France)

# Le coût environnemental

- Green Algorithms (en ligne, estimation)

**Green Algorithms**  
*How green are your computations?*

**Details about your algorithm**  
*To understand how each parameter impacts your carbon footprint, check out the formula below and the [methods article](#)*

Runtime (HH:MM)

Type of cores

Number of cores

Model

Memory available (in GB)

Select the platform used for the computations

Select location

**303.10 g CO<sub>2</sub>e**  
Carbon footprint

**2.28 kWh**  
Energy needed

**0.33 tree-months**  
Carbon sequestration

**1.73 km**  
in a passenger car

**1 %**  
of a flight Paris-London

Share your results with [this link!](#)

Computing cores VS Memory

How the location impacts your footprint

2000

L. Lanelongue, J. Grealey, M. Inouye  
Green Algorithms: Quantifying the  
Carbon Footprint of Computation  
Advanced Science, May 2021



# Le coût environnemental

- Carbon Tracker

## Example usage

```
from carbontracker.tracker import CarbonTracker

tracker = CarbonTracker(epochs=max_epochs)

# Training loop.
for epoch in range(max_epochs):
    tracker.epoch_start()

    # Your model training.

    tracker.epoch_end()

# Optional: Add a stop in case of early termination before all monitor_epochs has
# been monitored to ensure that actual consumption is reported.
tracker.stop()
```

- Et d'autres...

Lasse F. Wolff Anthony, Benjamin Kanding, Raghavendra Selvan  
Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models  
ICML Workshop on "Challenges in Deploying and monitoring Machine Learning Systems", 2020

# La reproductibilité

- Pouvoir reproduire les expériences (réplicabilité, répétabilité)
- Mais aussi pouvoir reproduire :
  - La conclusion générale (ce qu'on apprend d'un article)
  - Les résultats qualitatifs (ce qu'on constate, sans interprétation)
  - Les valeurs (mesurées ou calculées)

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss,  
Nancy Ide, Aurélie Névéol, Cyril Grouin, Lawrence E. Hunter  
Three Dimensions of Reproducibility in Natural Language Processing  
LREC 2018

Anya Belz, Shubham Agarwal, Anastasia Shimorina, Ehud Reiter  
A Systematic Review of Reproducibility Research in Natural Language Processing  
EACL 2021

# L'éthique

- Les biais cognitifs (genre, origine ethnique et sociale...)
- Le respect de vos valeurs (applications...)
- La gestion des données confidentielles
- L'information honnête du grand public
- Et de nombreux autres sujet
  
- ... mercredi avec Karën Fort !

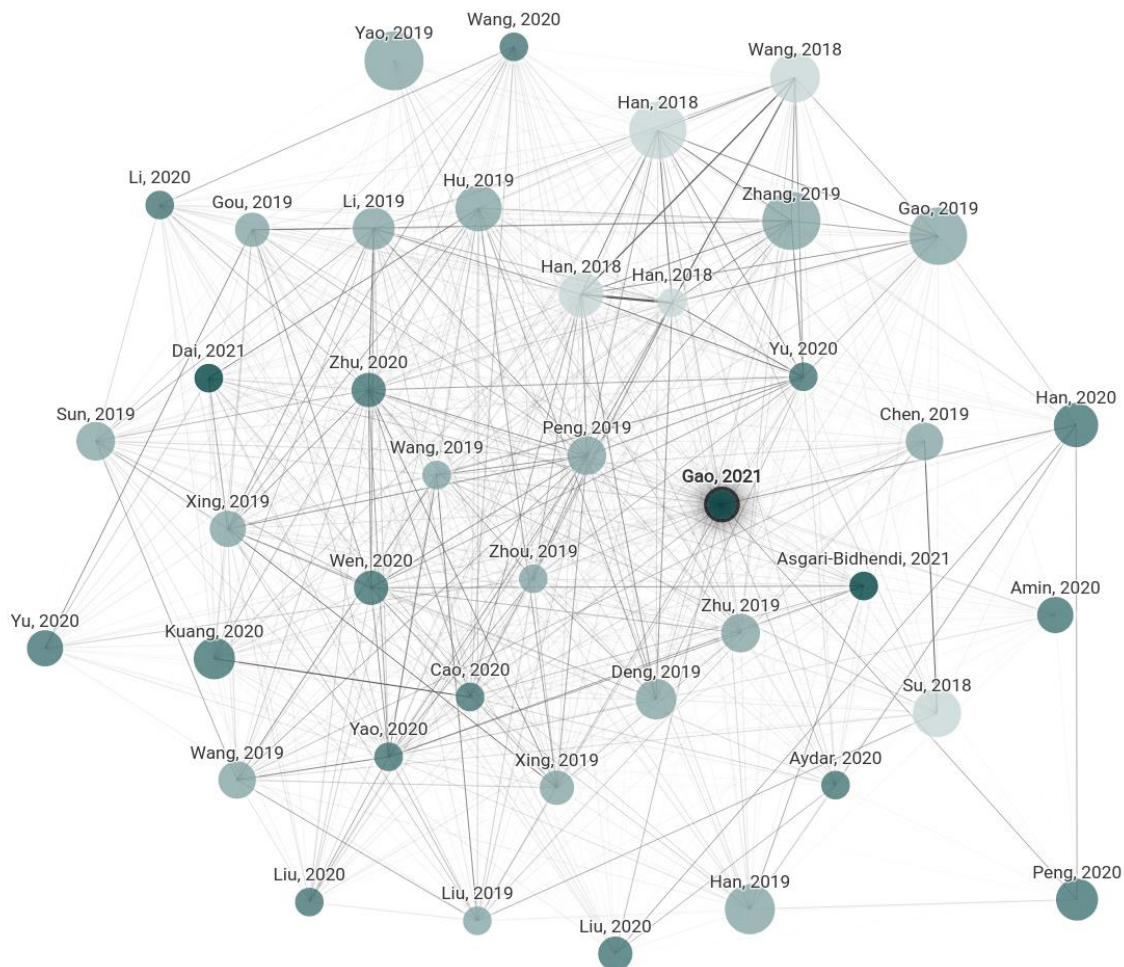
# Pointeurs utiles

En vrac

# The NLP Index

Articles, code, graphe : <https://index.quantumstat.com/>

Search...	Expand
Origin paper Manual Evaluation Matters: Reviewing Test Protocols of Distantly Supervised Relation Extraction Tianyu Gao, Xu Han, Keyue Qiu, Yuzhuo Bai, Zhiyu Xie, Yankai Lin, Zhiyuan Liu, Peng Li,... 2021	
Learning from Context or Names? An Empirical Study on Neural Relation Extraction Hao Peng, Tianyu Gao, X. Han, Yankai Lin, Peng Li, Zhiyuan Liu, M. Sun, Jie Zhou 2020	
More Data, More Relations, More Context and More Openness: A Review and Outlook for Relation Extraction Xu Han, Tianyu Gao, Yankai Lin, H. Peng, Y. Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li,... 2020	
REKER: Relation Extraction with Knowledge of Entity and Relation H. Liu, Yian Wang, Fangzhao Wu, Pengfei Jiao, Hongyan Xu, X. Xie 2019	
A survey on neural relation extraction Kang Liu 2020	
Neural relation extraction: a survey Mehmet Aydar, Ozge Bozal, Furkan Özbay 2020	
Hierarchical Relation Extraction with Coarse-to-Fine Grained Attention Xu Han, Pengfei Yu, Zhiyuan Liu, M. Sun, Peng Li 2018	
Two Training Strategies for Improving Relation Extraction over Universal Graph Qinyun Dai, Naoya Inoue, Ryo Takahashi, Kentaro Inui 2021	
A Hybrid Model with Pre-trained Entity-Aware Transformer for Relation Extraction Jinxin Yao, Min Zhang, B. Wang, Xianda Xu 2020	
Multi-task Learning for Relation Extraction Kai Zhou, X. Luo, Hongya Wang, R. Xu 2019	
Towards Accurate and Consistent Evaluation: A Dataset for Distantly-Supervised Relation Extraction Tong Zhu, Haitao Wang, Junjie Yu, Xiabing Zhou, W. Chen, W. Zhang, Min Zhang 2020	
OpenNRE: An Open and Extensible Toolkit for Neural Relation Extraction Xu Han, Tianyu Gao, Y. Yao, D. Ye, Zhiyuan Liu, M. Sun 2019	
Improving Distant Supervised Relation Extraction by Dynamic Neural Network Yanjie Gou, Y. Lei, Lingqiao Liu, Pingping Zhang, Xi Peng 2019	



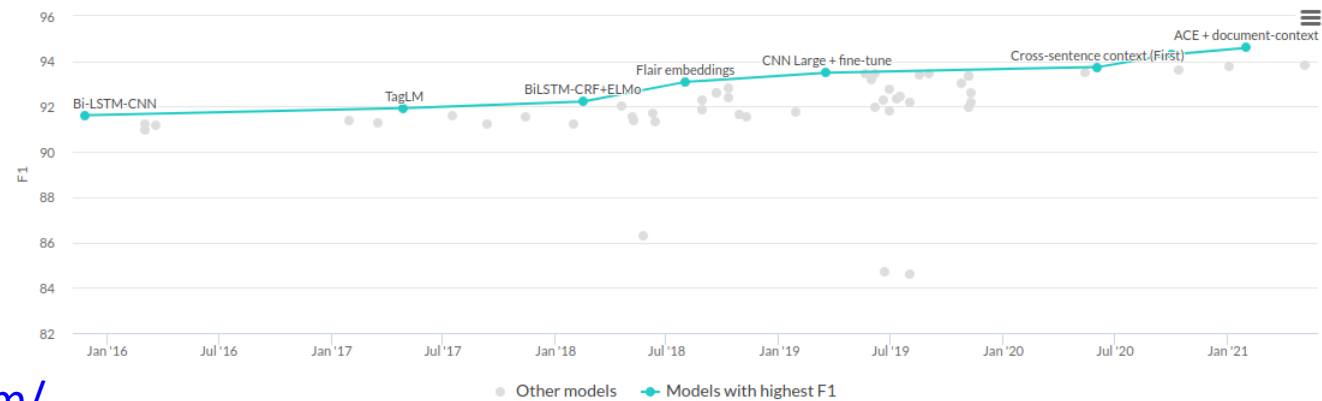


# Papers With Code

Named Entity Recognition

## Named Entity Recognition on CoNLL 2003 (English)

Leaderboard Dataset



Leaderboards,  
articles,  
code :

<https://paperswithcode.com/>

View F1 All models

Edit Leaderboard

Rank	Model	F1 ↑	Extra Training Data	Paper	Code	Result	Year
1	<b>ACE + document-context</b>	94.6	×	<a href="#">Automated Concatenation of Embeddings for Structured Prediction</a>			2021
2	<b>LUKE</b>	94.3	×	<a href="#">LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention</a>			2020
3	<b>CL-KL</b>	93.85	×	<a href="#">Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning</a>			2021
4	<b>InferNER</b>	93.76	×	<a href="#">InferNER: an attentive model leveraging the sentence-level information for Named Entity Recognition in Microblogs</a>			2021
5	<b>Cross-sentence context (First)</b>	93.74	×	<a href="#">Exploring Cross-sentence Contexts for Named Entity Recognition with BERT</a>			2020

# NLP progress

## NLP-progress

Repository to track the progress in Natural Language Processing (NLP), including the datasets and the current state-of-the-art for the most common NLP tasks.

## Tracking Progress in Natural Language Processing

### Table of contents

---

#### English

- [Automatic speech recognition](#)
- [CCG](#)
- [Common sense](#)
- [Constituency parsing](#)
- [Coreference resolution](#)
- [Data-to-Text Generation](#)
- [Dependency parsing](#)
- [Dialogue](#)
- [Domain adaptation](#)
- [Entity linking](#)
- [Grammatical error correction](#)
- [Information extraction](#)

<http://nlpprogress.com/>

# THE NLP PANDECT

"almost anything related to Natural Language Processing that is available online"

<https://github.com/ivan-bilan/The-NLP-Pandect>

## NLP RESOURCES



### Compendiums and awesome lists on the topic of NLP:

- [The NLP Index](#) by Quantum Stat / NLP Cypher
- [Awesome NLP](#) by keon [GitHub, 11952 stars]
- [Speech and Natural Language Processing Awesome List](#) by elaboshira [GitHub, 2020 stars]
- [Awesome Deep Learning for Natural Language Processing \(NLP\)](#) [GitHub, 927 stars]
- [Text Mining and Natural Language Processing Resources](#) by stepthom [GitHub, 380 stars]
- [Made with ML List](#) by madewithml.com
- [Brainsources for #NLP enthusiasts](#) by Philip Vollet
- [Awesome AI/ML/DL - NLP Section](#) [GitHub, 894 stars]
- [Resources on various machine learning topics](#) by Backprop

### NLP Conferences, Paper Summaries and Paper Compendiums:

#### Papers and Paper Summaries

- [100 Must-Read NLP Papers](#) 100 Must-Read NLP Papers [GitHub, 3106 stars]
- [NLP Paper Summaries](#) by dair-ai [GitHub, 1333 stars]
- [Curated collection of papers for the NLP practitioner](#) [GitHub, 1039 stars]
- [Papers on Textual Adversarial Attack and Defense](#) [GitHub, 859 stars]
- [The Most Influential NLP Research of 2019](#)
- [Recent Deep Learning papers in NLU and RL](#) by Valentin Malykh [GitHub, 288 stars]
- [Some Notable Recent ML Papers and Future Trends](#) by Aran Komatsuzaki [Blog, Oct. 2020]
- [A Survey of Surveys \(NLP & ML\): Collection of NLP Survey Papers](#) [GitHub, 1257 stars]
- [A Paper List for Style Transfer in Text](#) [GitHub, 1123 stars]
- [Video recordings index for papers](#)

#### Conferences


- [NLP top 10 conferences Compendium](#) by soulbliss [GitHub, 364 stars]
- [NLP Conferences Calendar](#)



# En vrac (pour la veille et la formation)

- ACL Anthology <https://www.aclweb.org/anthology/>
- #NLProc 
- arXiv Computation and Language <https://arxiv.org/list/cs.CL/recent>

# En vrac (pour la veille et la formation)

- ACL Anthology <https://www.aclweb.org/anthology/>
- #NLProc 
- arXiv Computation and Language <https://arxiv.org/list/cs.CL/recent>




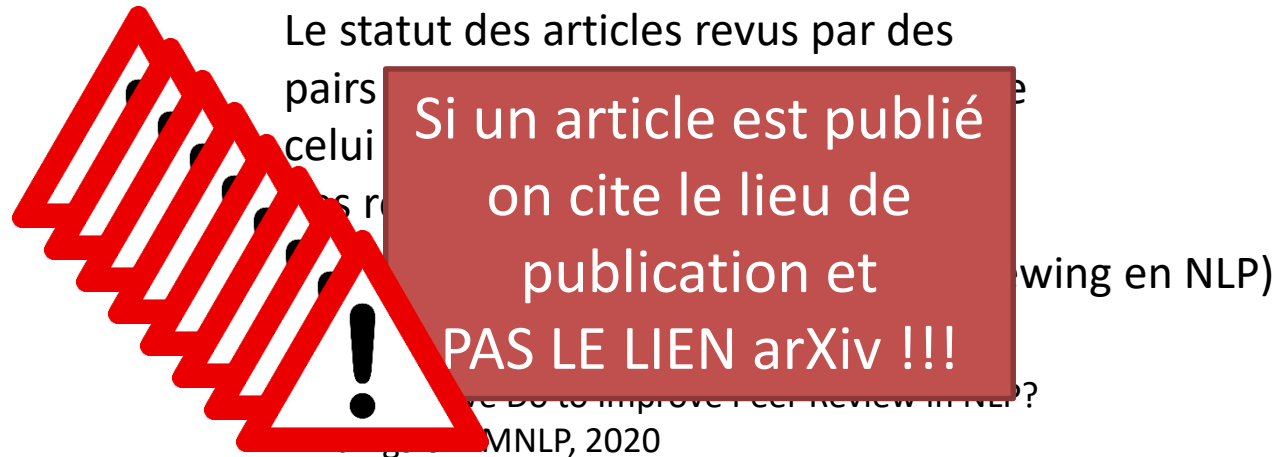
Le statut des articles revus par des pairs et publiés n'est pas le même que celui des preprints et des discussions des réseaux sociaux !

(malgré la crise actuelle du peer-reviewing en NLP)


Anna Rogers, Isabelle Augenstein  
What Can We Do to Improve Peer Review in NLP?  
Findings of EMNLP, 2020

# En vrac (pour la veille et la formation)




- ACL Anthology <https://www.aclweb.org/anthology/>
- #NLProc 
- arXiv Computation and Language <https://arxiv.org/list/cs.CL/recent>



# En vrac (pour la veille et la formation)

- ACL Anthology <https://www.aclweb.org/anthology/>
- #NLProc 
- arXiv Computation and Language <https://arxiv.org/list/cs.CL/recent>
- La liste de diffusion corpora <https://mailman.uib.no/listinfo/corpora>
- ATALA et liste LN <https://www.atala.org/>
- Le blog éthique et TAL : <http://www.ethique-et-tal.org/>
- Les cours d'Hugo Larochelle <https://www.youtube.com/user/hugolarochelle>
- Les cours de Stanford <https://www.youtube.com/watch?v=8rXD5-xhemo>
- Le blog de Sebastian Ruder <https://ruder.io/>

# En vrac

- Campagnes d'évaluation / shared tasks
  - SemEval
  - CLEF
  - TREC
  - NTCIR
  - MediaEval
  - DEFT 
  - Quaero 
  - ESTER 
  - Workshop on Machine Translation (WMT)
  - VarDial
  - Kaggle
- ... et beaucoup, beaucoup d'autres initiatives et benchmarks utiles

# Très en vrac

- <https://github.com/brianspiering/awesome-dl4nlp>
- <https://towardsdatascience.com/a-collection-of-must-known-pre-requisite-resources-for-every-natural-language-processing-nlp-a18df7e2e027>
- <https://www.kaggle.com/tags/nlp>
- <https://github.com/niderhoff/nlp-datasets>
- <https://data.world/datasets/nlp>

# Quelques outils

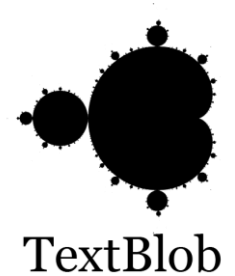
(liste personnelle, incomplète et biaisée)

spaCy



HUGGING FACE

AllenNLP



TextBlob

gensim



TensorFlow



**Fin !**

xavier.tannier@sorbonne-universite.fr