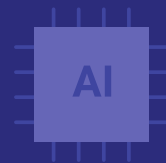
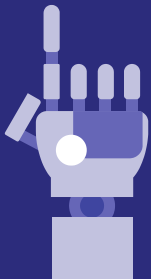




# Bases de TAL

Master 1 Sciences du Langage  
Parcours Industries de la Langue – Linguistic Data Sciences

Martina Barletta  
3-6 septembre 2024





# Table des contenus

**01**

## **Introduction**

au Traitement  
Automatique des Langues

**02**

## **Histoire du TAL**

Du rapport ALPAC à  
ChatGPT

**03**

## **Tâches**

Et campagnes  
d'évaluation

**04**

## **Evaluation**

Comment on évalue des  
systèmes TAL et  
l'annotation des données ?

**05**


## **Métriques**

Pourquoi et comment  
évaluer correctement ?

**06**

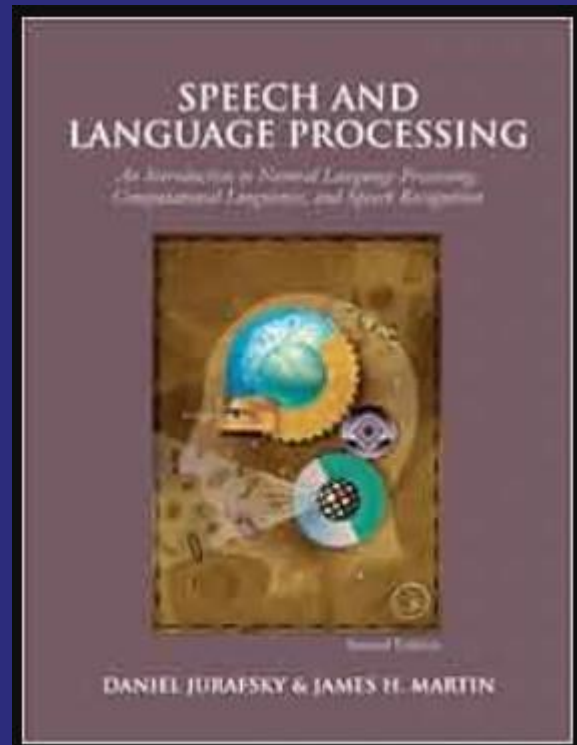
## **Exposé**

sur les thématiques  
du cours



# Speech and Language Processing

Dan Jurafsky et James H. Martin  
Full version available online!

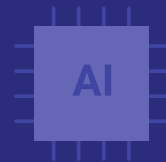
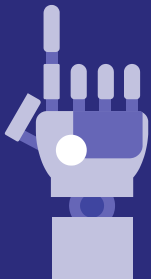




03

# Tâches

Et chaîne de traitement



# Réduire la langue à des problèmes calculables

Se soucier d'un problème à la fois  
pour ne pas s'occuper de la  
langue comme système



Le problème général est  
divisé  
en plusieurs sous-  
problèmes plus petits

**Théorie de la calculabilité** →  
définir les limites de ce qui peut  
être calculé à travers un algo  $\neq$   
**Théorie de la complexité** → définir  
quelle est l'efficacité d'un  
algorithme

# Chaîne de traitement en TAL

## Module

Manipulations de l'objet texte  
pour identifier et/ou étiqueter  
des objets linguistiques  
à différents niveaux textuels



# Comment tokeniser un texte ?

En juillet 1799, les soldats de l'armée de Napoléon découvraient près de la ville de Rosette, sur le delta du Nil, une pierre qui allait devenir l'une des plus célèbres de l'Antiquité.

Cette pierre, datant de 196 av. J.-C., relatait les honneurs rendus au roi Ptolémée V par les temples d'Égypte sous forme d'un "texte parallèle" en deux langues (le grec et l'égyptien) et trois écritures (les textes égyptiens étant écrits à la fois en hiéroglyphes et en démotique).

Son étude permit à Jean-François Champollion d'apporter en 1822 la clé du déchiffrement de l'écriture hiéroglyphique, découverte qui eut un retentissement considérable car elle mettait fin aux nombreuses controverses et mythes qui avaient entouré cette écriture.

# Comment tokeniser un texte ?

1. Tokenisez la première phrase du texte à la main
2. Téléchargez le texte veronis.txt sur github
3. Utilisez le fichier comme fichier d'entrée sur <https://corliapi.ortolang.fr/stanza/> (option écrit)
4. Téléchargez le fichier CoNLL-U et comparez-le avec votre tokenisation



# Le mot en TAL (plutôt, token)

- Passer d'une suite de caractères à une suite de **formes** → découpage en **TOKEN**
  - Séparés par des espaces ? **Pomme de terre**
  - Par des signes de ponctuation ? Chauve-souris, aujourd'hui, I'm en anglais
  - Constitués ou qui commencent par de signes de ponctuation ? ☺, #nplrules
- Certains tokens ne sont pas forcément des mots
  - Chiffres, dates, heures...
  - Acronymes
  - LA PONCTUATION !!! (*Let's eat, grandma* VS *Let's eat grandma*)
- Pas toutes les langues utilisent des espaces pour découper leurs mots (japonais)
- Dans la langue parlée, les disfluences ou les fillers, sont-ils des tokens ?

# Plusieurs types de tokenisation

"Don't you love 😊 Transformers? We sure do."

- Sur les espaces

["Don't", "you", "love", "😊", "Transformers?", "We", "sure", "do."]

- Token = mot (au sens linguistique) et ponctuation

["Don", "'", "t", "you", "love", "😊", "Transformers", "?", "We", "sure", "do", "."]

- Espaces, ponctuation et tokenisation à base de règles

["Do", "n't", "you", "love", "😊", "Transformers", "?", "We", "sure", "do", "."]

- Autres algorithmes (Byte-Pair Encoding)

[https://huggingface.co/docs/transformers/tokenizer\\_summary](https://huggingface.co/docs/transformers/tokenizer_summary)

# Lemmatisation

- Obtenir la forme canonique ou lemme d'un mot à partir d'une forme donnée
  - Verbe – forme à l'infinitif (sans flexion)
    - Il court → courir
  - Pour un nom, adjectif, article, ... - forme au masculin singulier
    - Cheval, chevaux → cheval
- La lemmatisation demande des ressources et un traitement linguistique (couteuse)
- Elle permet d'agréger des variantes flexionnelles et non pas des mots ayant la même racine

## Stemming (racinisation)

- Obtenir la racine d'un mot, commune à toutes les variantes morphologiques d'un mot à travers la suppression des flexions et des suffixes
- Elle est généralement à base de règles, rapide et dépend de la langue
- Demande moins de ressources que la lemmatisation (vocabulaire plus petit)

# Annotation

Segmenter un texte en plusieurs sous-unités et associer une étiquette aux unités qui nous intéressent traiter

Annoter tous les tokens d'un texte et associer une ou plusieurs étiquettes à chaque token

POS-tagging

Délimiter des tokens ou suite de tokens dans les textes et leur associer des étiquettes

Entités nommées  
Coréférence

# Annotation (I)

- L'annotation servait à fournir des informations pour le développement et mise à l'épreuve des théories en linguistique, ou, comment on l'appelle aujourd'hui, à la linguistique des corpus (...) ces ressources servent aujourd'hui à la linguistique mais aussi au traitement automatique des langues (...)

*(Ide, Handbook of linguistic annotation, 2017)*

# Annotation (II)

- Associer à chaque token une ou plusieurs étiquettes
- Délimiter un token ou ensemble de tokens et associer une étiquette

# Annotation (III)

- Associer à chaque token une ou plusieurs étiquettes

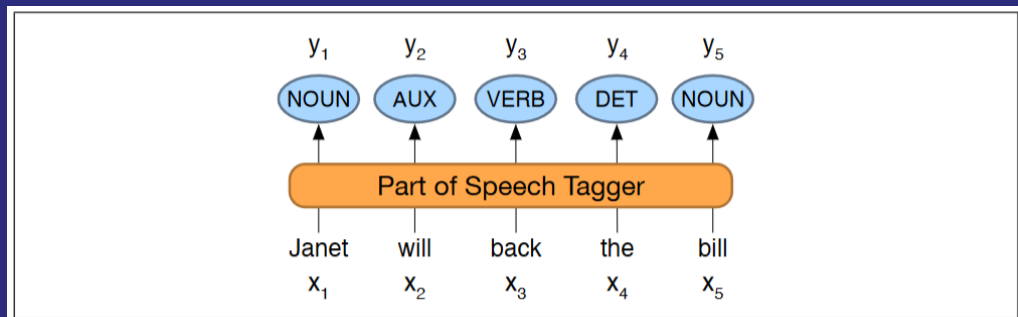
“Such tasks in which we assign, to each word  $x_i$  in an input word sequence, a label  $y_i$ , so that the output sequence  $Y$  has the same length as the input sequence  $X$  are called **sequence labeling tasks**.” (Jurafsky et Martin, 2008)

- Étiquetage morphosyntaxique
  - Lemmatisation
  - Traits morphosyntaxiques
- La **précision** des algorithmes d’annotation morphosyntaxique est actuellement très haute (autour du 97%)

# POS tagging

Associer à un token sa catégorie morphosyntaxique  
(nom, verbe, adjectif, adverbe etc.)

- Utile dans l'élimination des stop words
- Regroupement de termes complexes
- Désambiguïsation
- Construire une représentation syntaxique du texte





# (Stop Words)

- Mots vides en français (prépositions, articles, pronoms...)
- Sa distribution est uniforme sur les textes du corpus – sa fréquence est similaires dans tous les textes d'un corpus → mots qui ne sont pas discriminants dans le significat du texte
- Pour certaines tâches, on peut les retirer des textes pour faciliter le traitement
- BONUS : allez regarder la loi de Zipf !

# Exo POS tagging

Le chat mange la souris

La pomme de terre n'était pas connue  
en Amérique avant 1596.

La petite porte le voile

Tagset : DET, NOUN, ADJ, VERB, PRON...  
(formalisme UD)

Quelles versions de la dernière phrase sont  
possibles ?

Open class words	Closed class words	Other
<a href="#">ADJ</a>	<a href="#">ADP</a>	<a href="#">PUNCT</a>
<a href="#">ADV</a>	<a href="#">AUX</a>	<a href="#">SYM</a>
<a href="#">INTJ</a>	<a href="#">CCONJ</a>	<a href="#">X</a>
<a href="#">NOUN</a>	<a href="#">DET</a>	
<a href="#">PROPN</a>	<a href="#">NUM</a>	
<a href="#">VERB</a>	<a href="#">PART</a>	
	<a href="#">PRON</a>	
	<a href="#">SCONJ</a>	

Le – DET  
chat – NOUN  
mange – VERB  
la – DET  
souris – NOUN  
. – PUNCT

La – DET  
*pomme de terre* – NOUN ?  
    *pomme* – NOUN  
    *de* – ADP  
    *terre* – NOUN

n' – ADV  
était – AUX  
pas – ADV  
connue – VERB  
en – ADP  
Amérique – PROPN  
avant – ADP  
1596 – NUM  
. – PUNCT

Open class words	Closed class words	Other
<u>ADJ</u>	<u>ADP</u>	<u>PUNCT</u>
<u>ADV</u>	<u>AUX</u>	<u>SYM</u>
<u>INTJ</u>	<u>CCONJ</u>	<u>X</u>
<u>NOUN</u>	<u>DET</u>	
<u>PROPN</u>	<u>NUM</u>	
<u>VERB</u>	<u>PART</u>	
	<u>PRON</u>	
	<u>SCONJ</u>	

La - DET  
petite - ADJ  
porte - NOUN  
le - PRON  
voile - VERB



La - DET  
petite - NOUN  
porte - VERB  
le - DET  
voile - NOUN



# Annotation (II)

- Délimiter un token ou ensemble de tokens et associer une étiquette

Le phénomène qu'on veut annoter est « dilué » dans le texte, il faut d'abord le délimiter puis associer l'étiquette nécessaire

- Entités nommées
- Coréférence
- Sentiment analysis
- Hate speech
- Toxic content

# Entités nommées

**Les noms propres (PROPN) sont habituellement des unités polylexicales (Multi Word Phrases)**

*Named entity – entité qui peut être indiquée à travers un nom propre*  
Personne, lieu, organisation...

Ex. Marie Curie, Léon Marchand, New York, Université Grenoble Alpes

NER – named entity recognition → trouver des empanes de texte qui représentent des noms propres et étiqueter le type d'entité représenté

# Exo entités nommées

TAGSET : PER, LOC, ORG, TIME, MONEY

Citing high fuel prices, United Airlines said Friday it has increased fares by \$6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately matched the move, spokesman

Tim Wagner said. United, a unit of UAL Corp., said the increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

## Solution

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **\$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].



# Entités nommées

## AstroERC

**Extension du corpus TDAC (Time-Domain Astrophysic Corpus)**  
**Alkaan et al., 2024**

Annotation des entités nommées, la coréférence et les relations  
sémantiques dans un corpus d'astrophysique composé de 300  
rapports d'observation en anglais

# AstroERC

We discovered PS19did on MJD 58666.31 = 2019-07-02.31 at w=19.9 +/- 0.1 mag.

[...] The new transient source is in the galaxy UGC 11003.

[...] A spectrum was obtained of the possible supernova with the 1.82-m Plaskett telescope.

[...] Adopting the host galaxy redshift z=0.03566 (NED) yields an expansion velocity...

[...] Followup observations of this intrinsically faint transient are encouraged.

We discovered PS19did on MJD 58666.31 = 2019-07-02.31 at w=19.9 +/- 0.1 mag.

[...] The new transient source is in the galaxy UGC 11003.


[...] A spectrum was obtained of the possible supernova with the 1.82-m Plaskett telescope.




[...] Adopting the host galaxy redshift z=0.03566 (NED) yields an expansion velocity...






[...] Followup observations of this intrinsically faint transient are encouraged.

FIGURE 1 – Extrait d'un rapport d'observation. A gauche, un exemple d'annotation en entités nommées uniquement, et à droite, l'annotation des entités nommées avec en plus l'annotation des mentions de coréférences et des relations sémantiques entre les objets célestes (mentions de type `CelestialObject`) et leurs propriétés physiques.













# Pour quel but ?














 Astro-COLIBRI












































Personalize 

Status: logged out Infos: ✓ v2.16.0

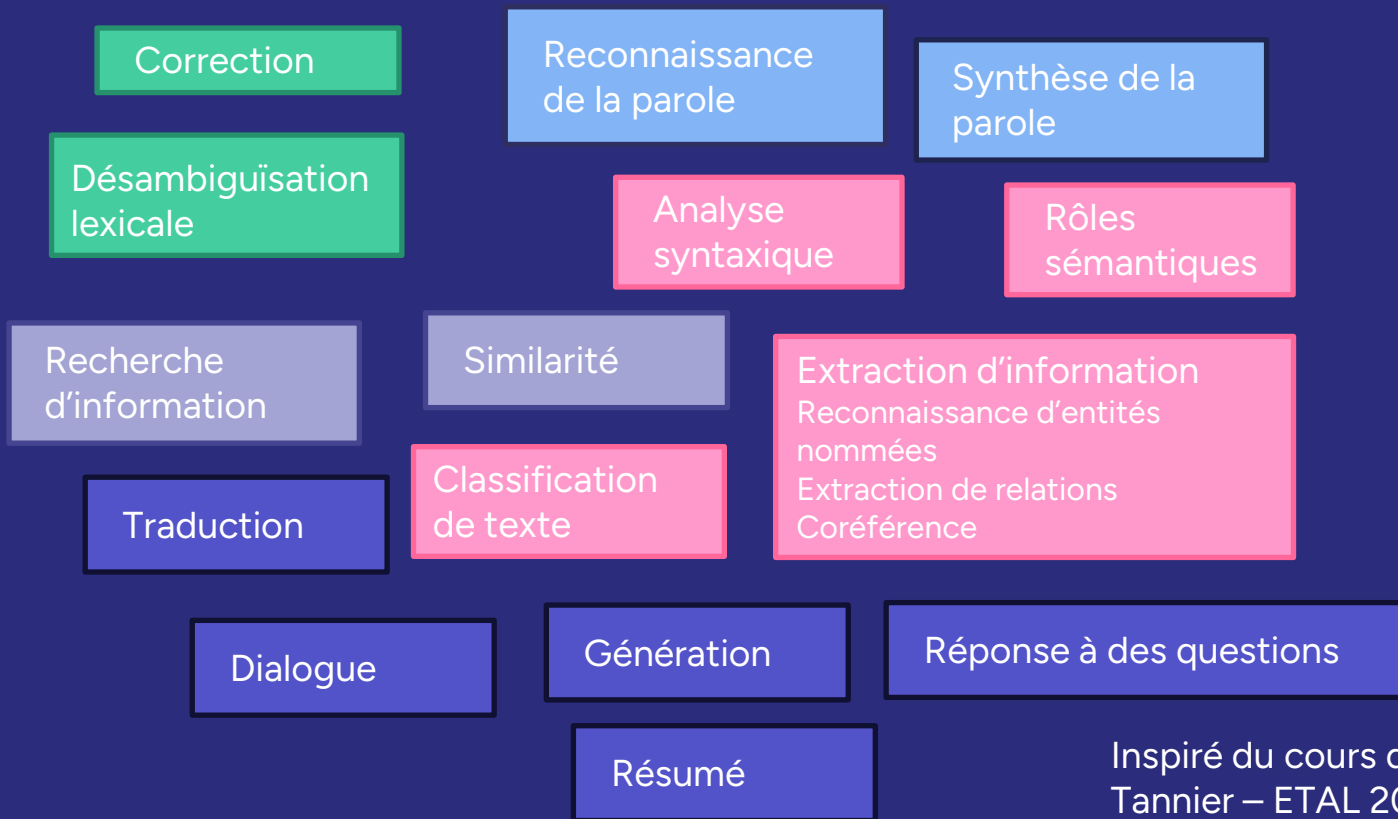
Observatories  Swift  SVOM  Fermi  HAWC  IceCube  AMON  Integral  GECAM  FlaapLUC  LVC  Catalogs  Other

Event type  FRB  Unclassified OT  Classified OT  SN  GRB  burst  neutrino  nuem  GW  4FGL  TeV CAT  SGR/AXP  IceCat General

 2024-08-20 

# La notion de TASK en TAL



Inspiré du cours de X.  
Tannier – ETAL 2021

# Shared task

- Campagnes qui rassemblent chercheurs et industriels
- Trouver une solution à un problème commun en utilisant le même jeu de données et les mêmes métriques d'évaluation (framework pas toujours défini)

Chaque équipe propose un système, entraîné (habituellement) sur un jeu de données commun (gold standard annotation), évalué sur un jeu de données commun (blind dataset parfois) et selon les mêmes métriques.

SemEval – International Workshop on Semantic Evaluation

CoNLL – **SIGNLL conference on Computational Natural Language Learning**

FakeNewsChallenge

EVALITA (Italie)

DEFT fouille de texte (France)

# EVALITA 2023

- Affect

- EMit – Categorical Emotion Detection in Italian Social Media (O. Araque, S. Frenda, D. Nozza, V. Patti, R. Sprugnoli)
- EmotivITA – Dimensional and Multi-dimensional emotion analysis (G. Gafà, F. Cutugno, M. Venuti)

- Authorship Analysis

- PoliticiT – Political Ideology Detection in Italian Texts (D. Russo, S.M. Jiménez-Zafra, J.A. García-Díaz, T. Caselli, M. Guerini, L.A. Ureña-López, R. Valencia-García)
- GeoLingIt – Geolocation of Linguistic Variation in Italy (A. Ramponi, C. Casula)
- LangLearn – Language Learning Development (C. Alzetta, D. Brunato, F. Dell’Orletta, A. Miaschi, K. Sagae, C.H. Sánchez-Gutiérrez, G. Venturi)

# EVALITA 2023

- Computational Ethics

- HaSpeede 3 – Political and Religious Hate Speech Detection (M. Lai, F. Celli, A. Ramponi, S. Tonelli, C. Bosco, V. Patti)
- HODI – Homotransphobia Detection in Italian (D. Nozza, G. Damo, A.T. Cignarella, T. Caselli, V. Patti)
- MULTI-Fake-DetectiVE – MULTImodal Fake News Detection and VERification (A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L.C. Passaro)
- ACTI – Automatic Conspiracy Theory Identification (G. Russo, N. Stoehr, M. Horta Ribeiro)

- New Challenges in Long Standing Tasks

- NERMuD -Named-Entities Recognition on Multi-Domain Documents (T. Paccosi, A. Palmero Aprosio)
- CLinkaRT – Linking a Lab Result to its Test Event in the Clinical Domain (B. Magnini, B. Altuña, A. Lavelli, M. Speranza, R. Zanolì)
- WiC-ITA – Word-in-Context task for Italian (P. Cassotti, L. Siciliani, L. Passaro, M. Gatto, P. Basile)
- DisCoTEX – Assessing DIScourse COherence in Italian TEXTs (D. Brunato, D. Colla, F. Dell'Orletta, I. Dini, D.P. Radicioni, A.A. Ravelli)

# DEFT 2024

Défi Fouille de Textes@TALN 2024

- Réponse automatique à des questionnaires à choix multiples issus d'annales d'examens de pharmacie
- Corpus (d'évaluation ?) utilisé : FrenchMedMCQA
- FrenchMedQA – 3105 question fermées composés de :
  - Un identifiant
  - La question
  - Cinq options
  - L'ensemble des réponse(s) correcte(s)



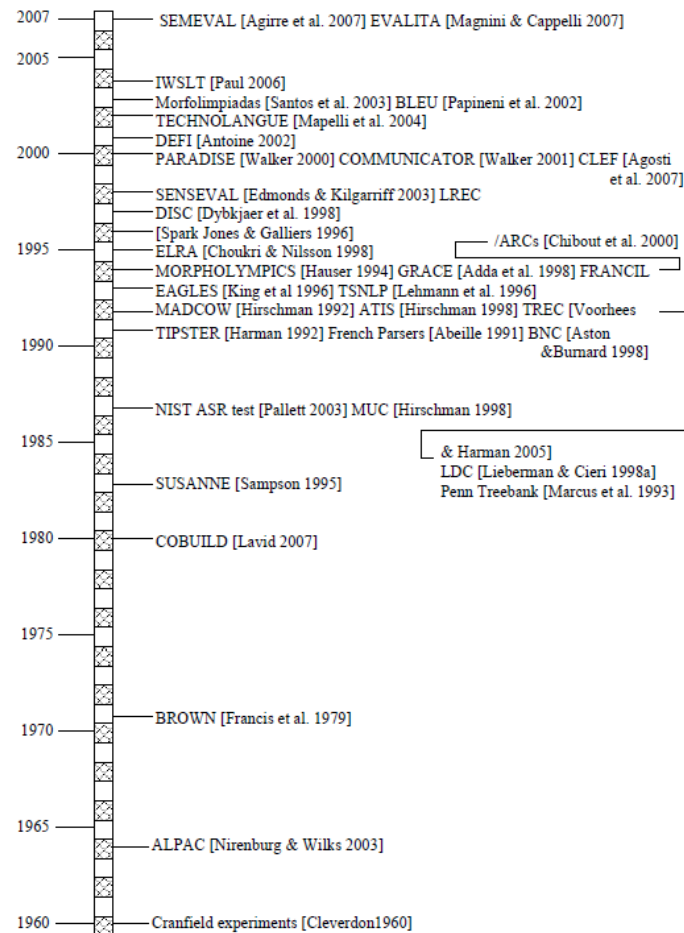
# DEFT 2024

Défi Fouille de Textes@TALN 2024

- Deux tâches :
  - Tâche principale – identifier l'ensemble des réponses correctes parmi les cinq proposées pour une question donnée. Moins de 3 milliards de paramètres
  - Tâche annexe : pareil mais pas de limites sur la taille des modèles
- Métriques :
  - Exact Match Ratio (taux des réponses parfaitement justes)
  - Hamming Score (taux des réponses juste parmi l'ensemble des réponses et référence)

# L'importance des campagnes d'évaluation...

- Permettent de faire avancer la recherche sur des thématiques spécifiques à travers la compétition
- Problèmes éthiques, entre autres le manque de ressource adéquates pour certaines équipes (ordinateurs assez puissants)
- Les gagnants parfois ne publient pas de manière claire (secretiveness)
- Manque de description des résultats négatifs (ne permet pas d'avancer)
- Certaines équipes se retirent des compétitions si pas bien placés pour ne pas nuire à leur réputation – surtout dans l'industrie
- Plutôt que trouver une vraie solution, on s'occupe d'adapter notre système à un data set existant et à le mesurer avec la métrique prévue



**Figure 1.** Salient events related to evaluation mentioned in this article (for evaluation campaign series, e.g. like TREC, only the first event is mentioned).

● Paroubek et al.,  
2007

# Focus sur la corréférence

## Extraction d'information - **Corréférence**

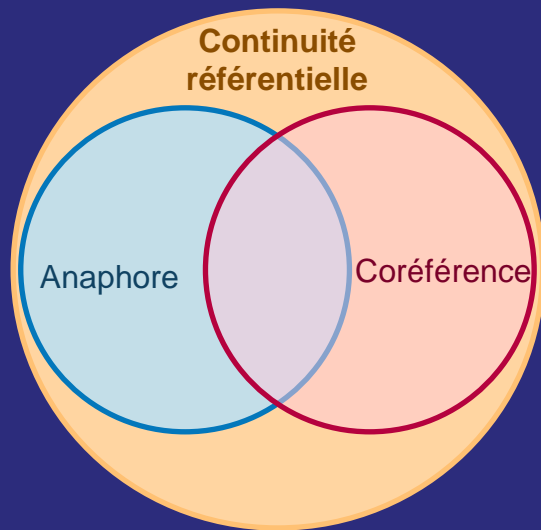
Une sorcière avait une maison noire et un chat noir. Il se cachait dans la maison pour ne pas qu'on le voie. Il ne sortait que la nuit. Un jour la sorcière en avait marre. Avec sa baguette magique elle l'a transformé en chat vert comme ça elle le voyait tout le temps.

*Texte normalisé tiré du corpus Scoledit, CE2, élève 207*

Une sorcière > la sorcière > sa > elle > elle

un chat noir > Il > le > Il > l' > chat vert > le

# Anaphore et coréférence



## *Anaphore*

- (Corblin, 1995 ; Poesio, 2016)
- « suppose la mise en relation d'une expression non autonome du point de vue de la référence et d'une expression référentielle susceptible de la « saturer » (Schnedeker, 2019 p.11, Corblin, 1995)

*Relation asymétrique*

## *Coréférence*

- « forme d'identité référentielle entre les référents évoqués » (Schnedeker, 2019, p. 13)

*Relation symétrique*

# Utilité de la résolution de corréférence

- Systèmes de dialogue
- Réponse aux questions
- Traduction automatique
- Text summarization (de quoi on parle dans un texte)?
- Études sur le développement de l'écriture à l'école primaire

# Outils « off the shelf »

Neuralcoref  
(hugging face)

The cat was very small and its owner was an old lady. She was a witch.

Mémoire de Master 2  
Linguistique Informatique Traduction, option Informatique

## ODACR :

un Outil de Détection Automatique  
des Chaînes de Référence  
à base de règles linguistiques

rédigé par Bruno OBERLÉ  
sous la direction de Mme TODIRASCU

README.md

## DeCOFre

passing pypi v0.7.0 code style black

Detecting Coreferences for Oral French<sup>1</sup>.

This was developed for application on spoken French as part of my PhD thesis, it is relatively easy to apply it to other languages and genres, though.

## Croc

### Coreference Resolver for Oral Corpora

The documentation is in the [pdf file](#) (in french).

Please cite:  
Désoyer, A., Landragin, F., Tellier, I., Lefevre, A. & Antoine, J.-Y.  
(2014) "Les coréférences à l'oral : une expérience  
d'apprentissage automatique sur le corpus ANCOR.",  
Traitement Automatique des Langues (TAL) 55(2),  
<http://www.atala.org/-Volume-55->, 2014, pp. 97-121.

reference resolution

Scholar About 2,660 results (0.09 sec)

Since 2021

# Corpus annotés en coréférence

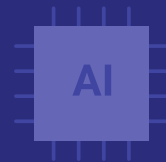
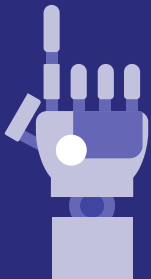


# CoNLL-2011

Modeling unrestricted coreference in OntoNotes

04

# Évaluation



05

# Métriques

