

Bases de TAL

Martina Barletta
Cours du 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 ?

05


Métriques

Pourquoi et comment
évaluer correctement ?

06

Exposé

sur les thématiques
du cours





01

Introduction

Qu'est-ce que c'est le TAL ?



Industries de la langue ?

Scannez le QR code à l'écran ou connectez vous avec le lien suivant au sondage : <https://www.menti.com/al13odojx7mn>



Industries de la langue

Traducteurs automatiques, dictionnaires informatisés, correcteurs automatiques

Prédiction de mots

Moteurs de recherche

Recommender systems

Classification automatique des informations

Recherche d'informations

Analyse d'opinion ou de sentiments

Génération automatique de textes, résumé automatique

Synthèse vocale, reconnaissance de la parole

Chatbots (service vocaux, assistants vocaux interactifs)

Robotique

Produits pour la communication augmentée

...

À la base de ces produits, on retrouve
le **TRAITEMENT AUTOMATIQUE DES LANGUES**

Traitement Automatique de la Langue

- Élaboration de programmes informatiques manipulant la forme langagière et capables, à travers la forme, de traiter le sens associé, pour résoudre des problèmes spécifiques
- Le but du TAL est de faire en sorte que les problèmes mettant en jeu des informations langagières deviennent calculables
- Problème central du TAL : **Traiter l'ambiguïté**





**Le langage est
redondant, implicite
et ambigu**



Le langage est ambigu, implicite

Ambiguïté : un même segment linguistique peut se prêter à deux interprétations mutuellement exclusives (Kerbrat-Orecchioni, 2005)

Exemple	Type d'ambiguïté
C'est un <u>vol</u> très risqué	Sémantique → Homonymie
Je <u>loue</u> pour l'année un appartement à Grenoble	Sémantique → Polysémie
Couvent (nom ou verbe)	Catégorielle
La petite porte le voile	Syntaxique catégorielle
Je vois un homme sur la colline avec un télescope	Syntaxique structurale

L'implicite dans le langage

ARTS

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

Le tableau trouvé en 2014 dans la région de Toulouse est estimé 120 millions d'euros. Les spécialistes ont trente mois pour approfondir l'expertise.

Connaissance du monde
Métonymie

Éducation et enseignement supérieur

Les dossiers chauds d'une rentrée scolaire inédite

Métaphore

Amériques

La Maison Blanche, un vieux rêve pour la vice-présidente

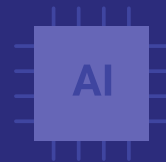
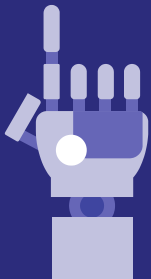
Connaissance du monde



02

Histoire du TAL

Du rapport ALPAC à ChatGPT



Traduction Automatique et Guerre Froide

1950 – USA – développement d'applications pour la traduction automatique

1954 – **Georgetown-IBM experiment**

Traduction automatique de 60 phrases russe → anglais

On préconisait la traduction automatique comme *problème déjà résolu* dans l'espace de 3/5 ans...

1966 – **Rapport ALPAC** – réduction des financements et investissements sur les ressources pour la traduction plutôt que sur la traduction automatique elle-même



“дух бодр, плоть же немощна”
“The spirit is willing, but the flesh is weak”

RU → EN → RU

“The vodka is strong, but the meat is rotten”
“водка хорошая, но мясо протухло”



(Exemple apocryphe non attesté)



“Out of mind, out of sight”

EN → RU → EN

“Invisible idiot”

cité par John Hutchins, Harper's Magazine, August 1962
"The whisky was invisible", or Persistent myths of MT"



Rapport ALPAC (I)

ALPAC

Automatic Language Processing Advisory Committee

Objectifs

Evaluer les progrès des derniers dix ans en linguistique computationnelle et en particulier en traduction automatique

Résultats du rapport

La recherche en TA n'était pas suffisante, il faut recentrer les recherches sur la linguistique computationnelle de base d'abord → réduction des financements en TA

Premier HIVER DE L'IA

Rapport ALPAC (II)

- Pas la peine d'investir plus dans la TA (marché relativement petit, traductions automatiques encore fortement erronées)
- S'orienter dans le développement d'outils de support à la traduction (dictionnaires numériques)

Passage de la Traduction Automatique
à la Traduction Assisté par l'Ordinateur

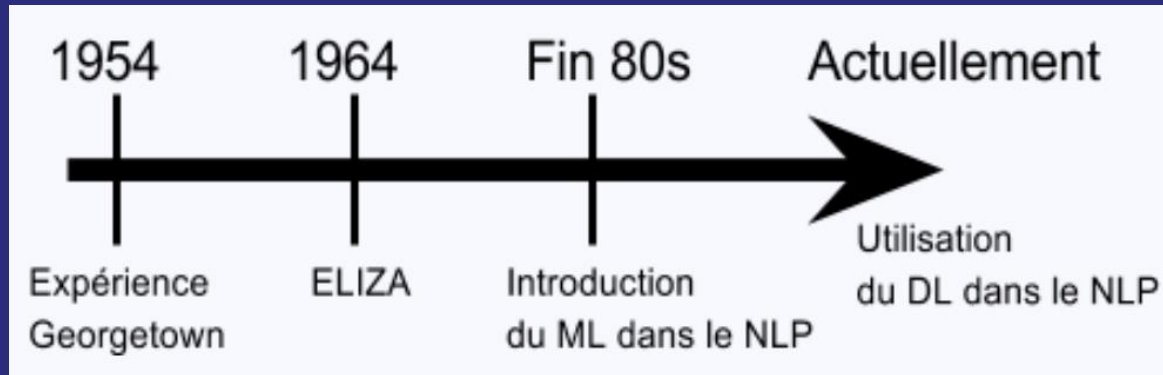
→ Naissance de la Linguistique Computationnelle
et de la Linguistique des Corpus

Périodisation

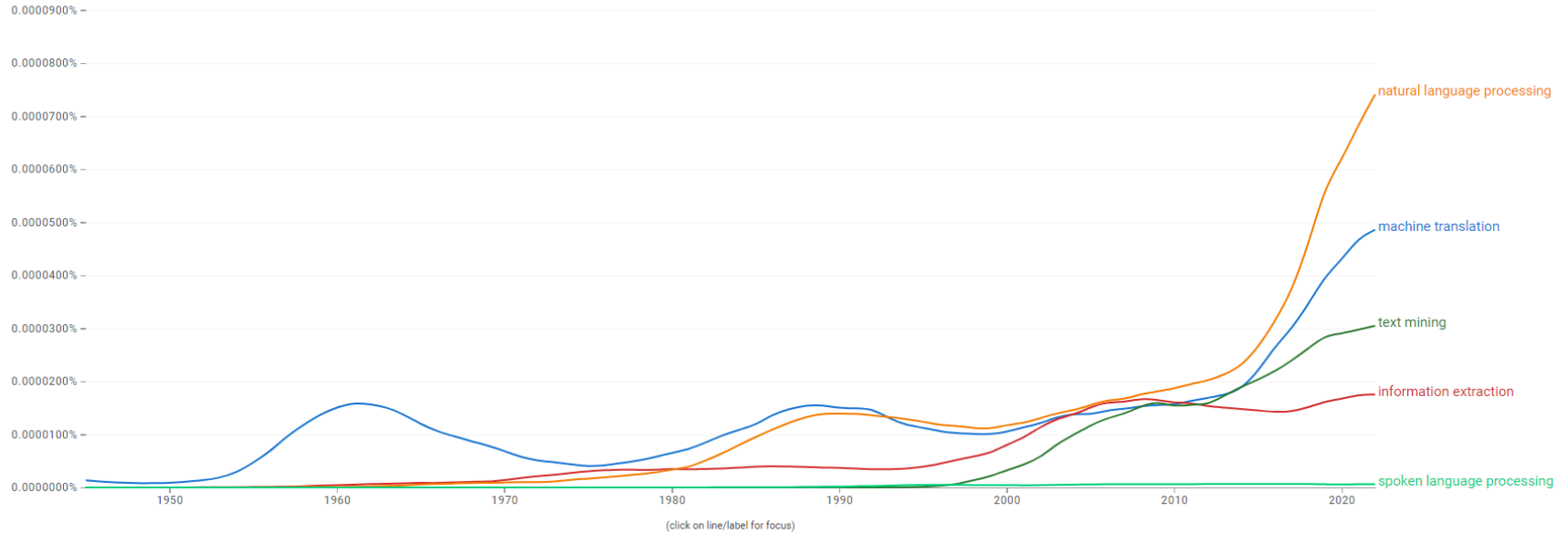
Période Symbolique (1950-1990)

Période Statistique (1990-2010)

Période Neuronale (2010-...)



Google Books Ngram Viewer



Période symbolique

1950-1992 ca

Symbolique → les connaissances du système sont représentées sous forme de **règles**, décrite à travers des **symboles** lisibles par l'humain (langage formel – Chomsky)

Applications :

SHRDLU (Winograd, 1968-1970)

ELIZA (Weizenbaum, 1964-1967)

Période symbolique

Test de Turing

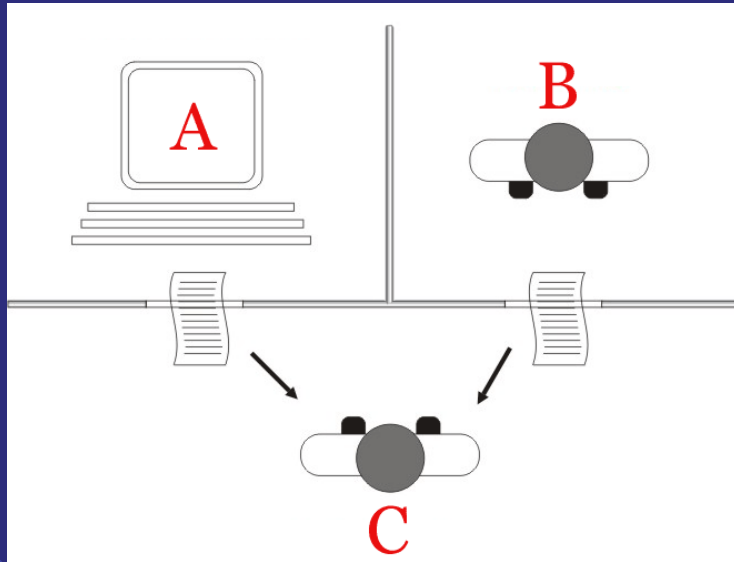


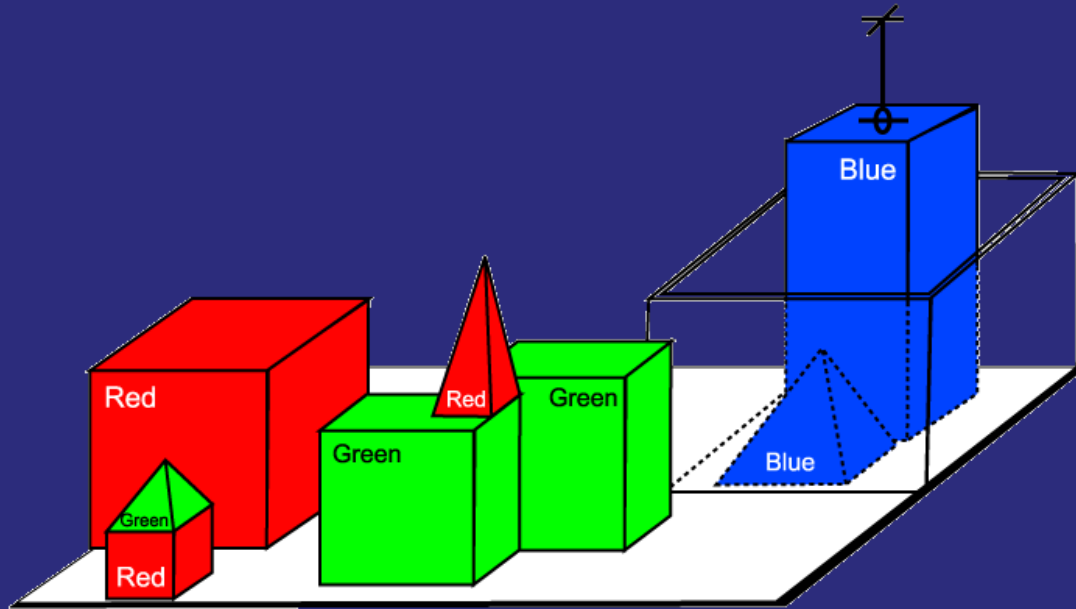
Illustration tirée de Wikipedia : « Test de Turing »

1950 - *Computing Machinery and Intelligence*

- Confrontation verbale à l'aveugle entre un examinateur humain (C) et une machine (A) ou un humain (B).
- Si l'examineur humain ne parviens pas à distinguer la machine d'un humain, la machine a passé le test.
- L'échange se fait à l'écrit.

Période symbolique

SHRDLU



Période symbolique

ELIZA

Psychothérapeute rogérien

Welcome to

```
EEEEEE LL      IIII ZZZZZZZ AAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LL      II      ZZZ  AAAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LLLLLL IIII ZZZZZZZ AA  AA
```

Eliza is a mock Rogerian psychotherapist.

The original program was described by Joseph Weizenbaum in 1966.

This implementation by Norbert Landsteiner 2005.

ELIZA: Please tell me what's been bothering you.

YOU: The weather is awful.

ELIZA: I'm not sure I understand you fully.

YOU: I hate the current weather.

ELIZA: You say you hate the current weather ?

YOU: Yes, that's what I said.

ELIZA: I understand.

YOU: Great.

ELIZA: Please go on.

YOU: I would say it's your turn now!

ELIZA: Why are you concerned over my turn now ?

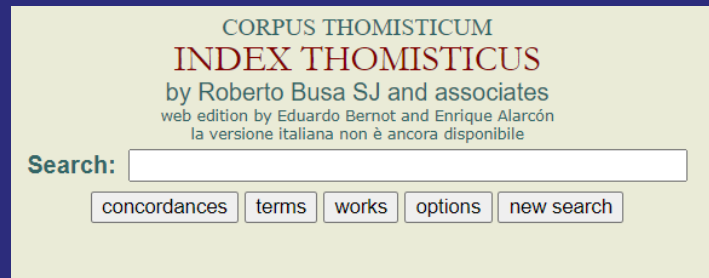
YOU: █

Période symbolique

Premiers expériences en Linguistique Computationnelle

Padre Roberto Busa et l'*Index Thomisticus*

- Exploiter l'ordinateur comme moyen de archivage et analyse des données textuels
- Réaliser une vérification intégrale et ponctuelle du lexique de St Thomas d'Aquin, mais les analyses qu'il voulait faire n'étaient pas disponibles dans les concordanciers papier disponibles à l'époque
- Décide de réaliser sa propre base de données
- 194 - Demande à Thomas Watson (IBM) de lui donner des calculateurs
- En 1980, s'achève son travail, et publie les 70 000 pages de *l'Index Thomisticus* (11 millions de mots!)



<https://www.corpusthomisticum.org/it/index.age>

Index Thomisticus

Lemma 06079

anima: anima animae

Type of lemma: (A) common word

Type of meaning: (NV) invisible thing

cases	c. freq.	places	pl. freq.		form	type	infl.	num.	gen.	case	deg.	voice	tense	mood	pers.	comp.	notes
12983	1,22%	6780	7,78%	<input checked="" type="checkbox"/>	anima	n.	irr. decl.	sing.	f.	nom.	pos.						<i>b</i>
8511	0,80%	4921	5,65%	<input type="checkbox"/>	animae	n.	irr. decl.	sing.	f.	gen.	pos.						<i>g</i>
60	0,01%	26	0,03%	<input type="checkbox"/>	anime	n.	irr. decl.	sing.	f.	gen.	pos.						<i>b</i>
1	0,00%	1	0,00%	<input type="checkbox"/>	animaeque	n.	irr. decl.	sing.	f.	gen.	pos.					-que	
4325	0,41%	2694	3,09%	<input type="checkbox"/>	animam	n.	irr. decl.	sing.	f.	acc.	pos.						
1	0,00%	1	0,00%	<input type="checkbox"/>	animamque	n.	irr. decl.	sing.	f.	acc.	pos.					-que	
609	0,06%	419	0,48%	<input type="checkbox"/>	animarum	n.	irr. decl.	pl.	f.	gen.	pos.						
349	0,03%	278	0,32%	<input type="checkbox"/>	animabus	n.	irr. decl.	pl.	f.	dat.	pos.						
743	0,07%	547	0,63%	<input type="checkbox"/>	animas	n.	irr. decl.	pl.	f.	acc.	pos.						<i>b</i>

(*b*) Not divided: Homographs belonging to other lemma entries have been assigned to this form, awaiting the analysis of all of its occurrences.

(*g*) Base-form of subsequent graphic variants. (The choice of this word as a base-form does not conform to scientific, but to technical criteria.)

(*b*) Secondary graphic variant.

General notice: Homographs within the same lemma entry have not yet been divided.



ATALA

Association pour le Traitement Automatique des Langues

- Fondée en **1959** !
- Aide à l'organisation de la conférence TALN et de sa session RECITAL
- Une des premières sociétés savantes au monde à s'occuper de TAL (reconnu par ACL)



Association for
Computational Linguistics

- Fondée en **1962**
- Organise la conférence ACL annuelle
- Sponsor de la revue Computational Linguistics

Période statistique

1993-2012 ca

- Disponibilité d'une grande quantité de données langagières
- Augmentation de la puissance de calcul
- Développement des corpus multilingues et de l'annotation de corpus

Introduction des algorithmes de Machine Learning

Applications :

- IBM Alignment Models (90s)

Période statistique

1993-2012 ca

2003 – le perceptron multi-niveau (Bengio et al.) obtient des meilleurs résultats du modèle word *n*-gram

2010 – RNN de Mikolov → Word2vec, word embeddings

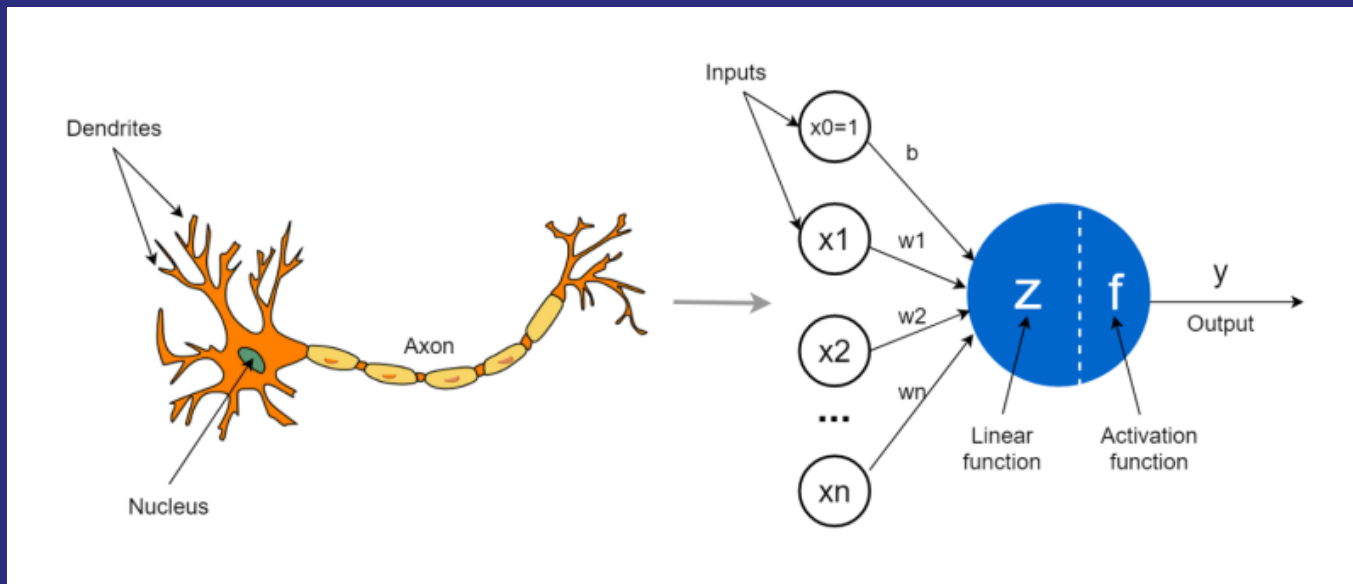
Mais aussi

• Treebanks et annotation de corpus

Période statistique

Le perceptron de Rosenblatt

- Classifieur linéaire



Période neuronale

2013-aujourd'hui

- Apprentissage Neuronale Profond (Deep Learning)
- Grande puissance de calcul disponible et très grandes quantités de données à disposition
- Différents approches (CNN, LSTM, Transformer...)
- Naissance des LLM (Large Language Models)

Période neuronale

2013-2017

2018-now

Période neuronale



ChatGPT

Ou l'explosion des LLM

- Chatbot et assistant virtuel
- Appartient à la famille des modèles GPT – Generative Pre-trained Transformer
- Fine-tuned pour la tâche de conversation à travers apprentissage supervisé et reinforcement learning à travers feedback humain

- Des travailleurs payés 2\$ au Kenya pour épurer le côté « toxique » de ChatGPT

“Despite the foundational role played by these data enrichment professionals, a growing body of research reveals the precarious working conditions these workers face,” says the Partnership on AI, a coalition of AI organizations to which OpenAI belongs. “This may be the result of efforts to hide AI’s dependence on this large labor force when celebrating the efficiency gains of technology. Out of sight is also out of mind.”

(Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic, Time, 18 Janvier 2023)

ChatQPT

Ou l'explosion des LLM

“(...) that for all its glamor, AI often relies on hidden human labor in the Global South that can often be damaging and exploitative. These invisible workers remain on the margins even as their work contributes to billion-dollar industries.”

(Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic, Time, 18 Janvier 2023)

Large Language Models

Modèles de langage

- La quantité des données d'entraînement
- La quantité d'hyperparamètres

Fig tirée de Minaee et al., 2024
Disponible sur arxiv

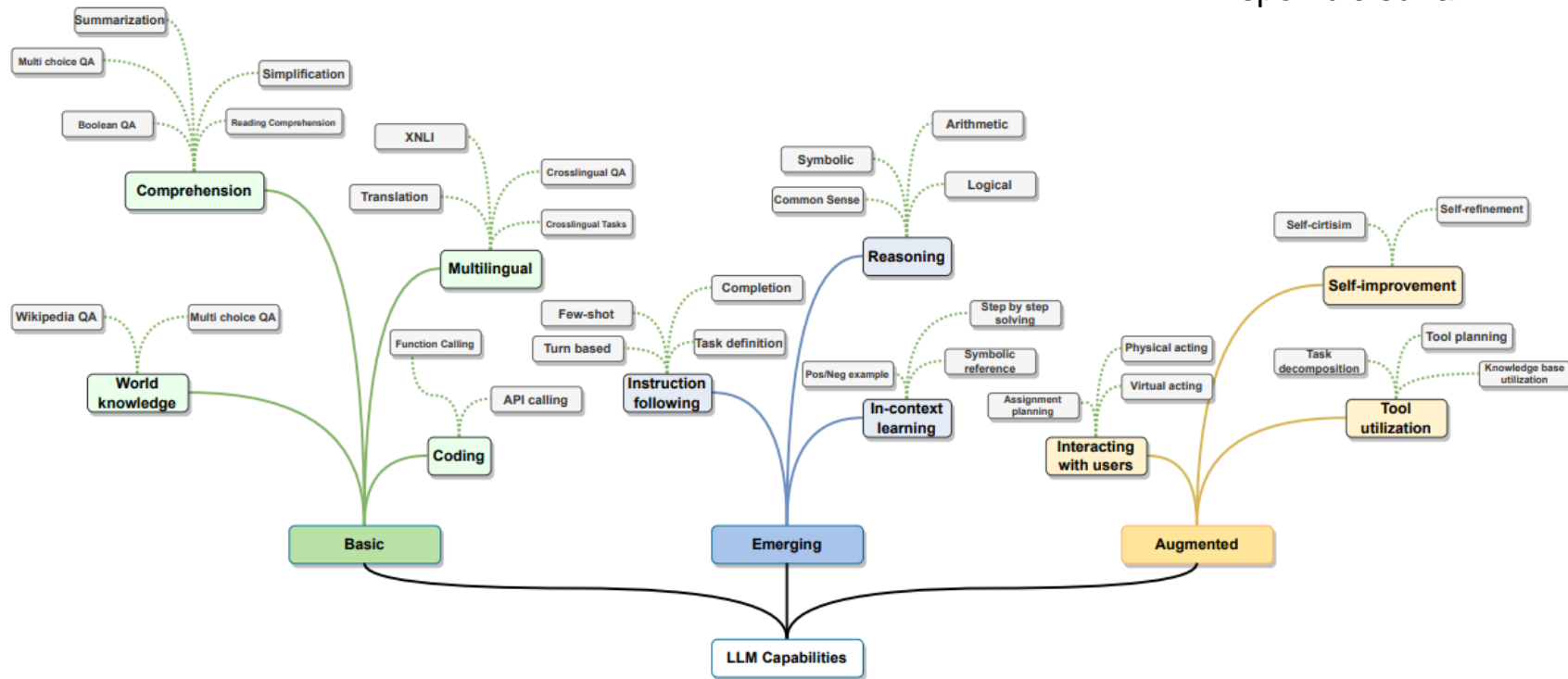


Fig. 1: LLM Capabilities.

Fig tirée de Minaee et al., 2024

Disponible sur arxiv

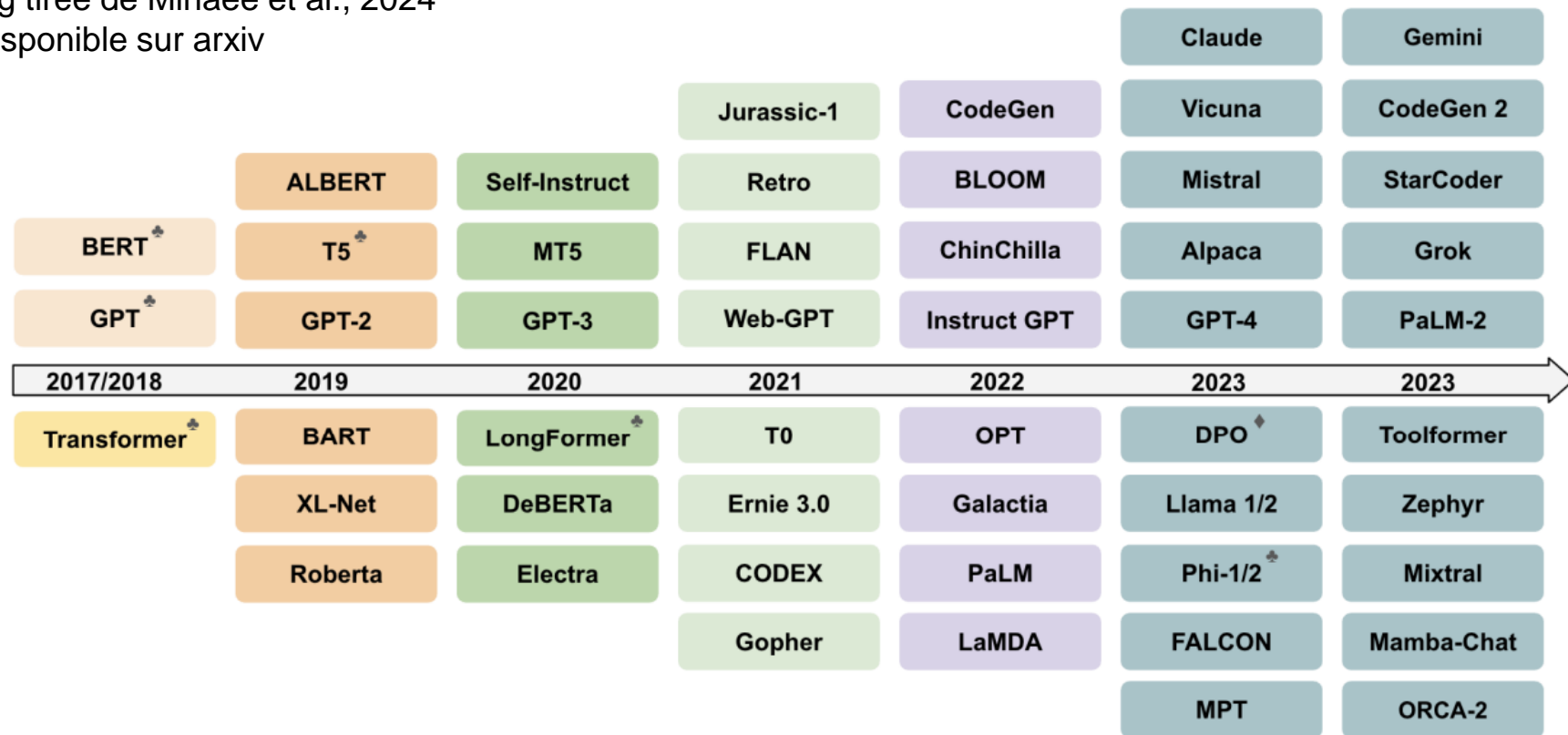


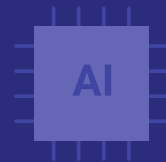
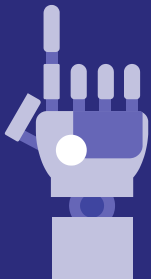
Fig. 24: Timeline of some of the most representative LLM frameworks (so far). In addition to large language models with our #parameters threshold, we included a few representative works, which pushed the limits of language models, and paved the way for their success (e.g. vanilla Transformer, BERT, GPT-1), as well as some small language models. ♠ shows entities that serve not only as models but also as approaches. ♦ shows only approaches.



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

Chaîne de traitement en TAL

Module

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



Le mot en TAL

- Passer d'une suite de caractères à une suite de **formes**
- Découpage en TOKEN
- Différentes manières de découper en tokens : mot, sous-mot (byte-pair encoding)
- La notion de token comprends toutes les suites de caractères qui ne sont pas exactement des mots

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。

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)

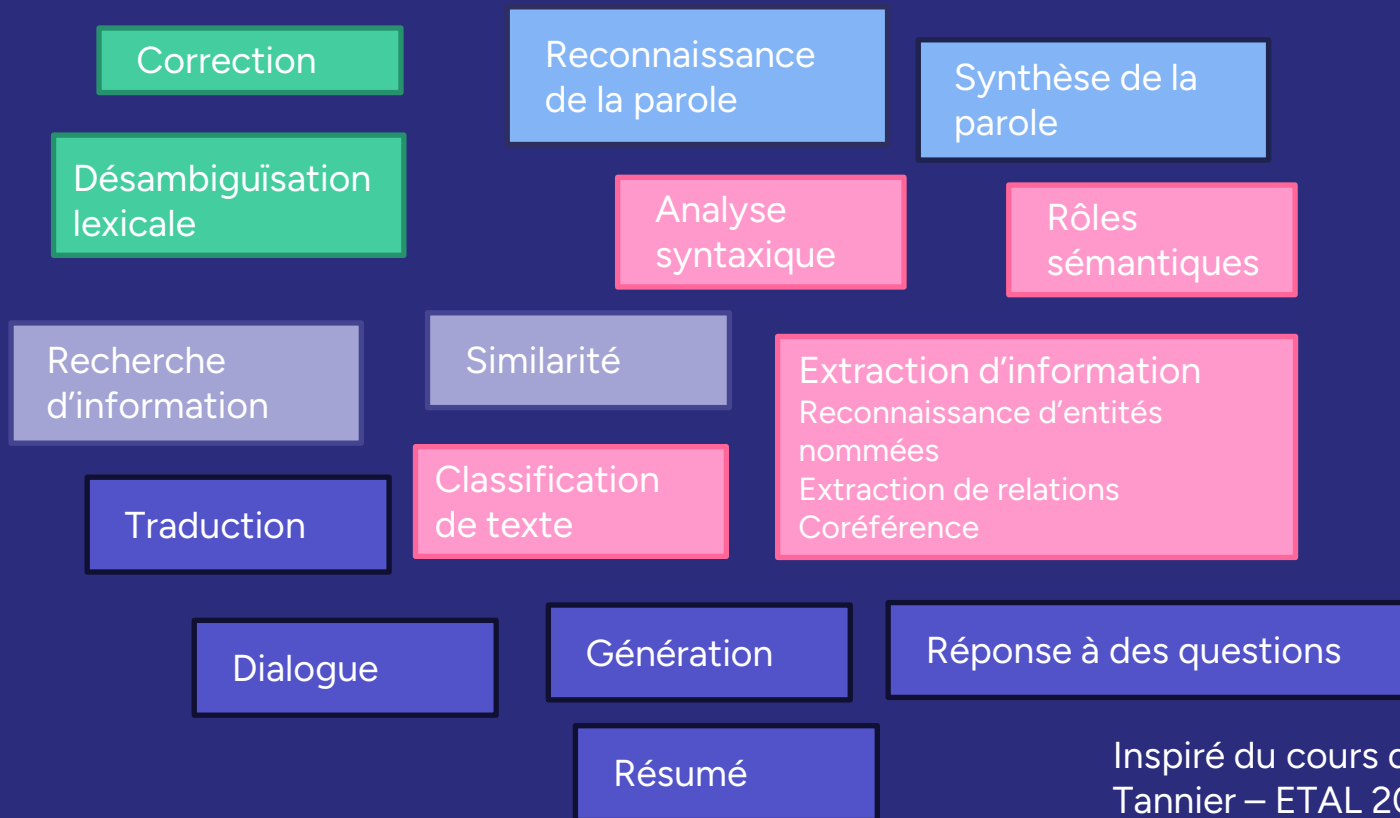
Annotation

- Associer aux tokens une ou plusieurs étiquettes, par exemple
 - Catégorie morphosyntaxique
 - Lemme
 - Traits morphosyntaxiques...

POS tagging (nom, verbe, adjectif, adverbe etc.)

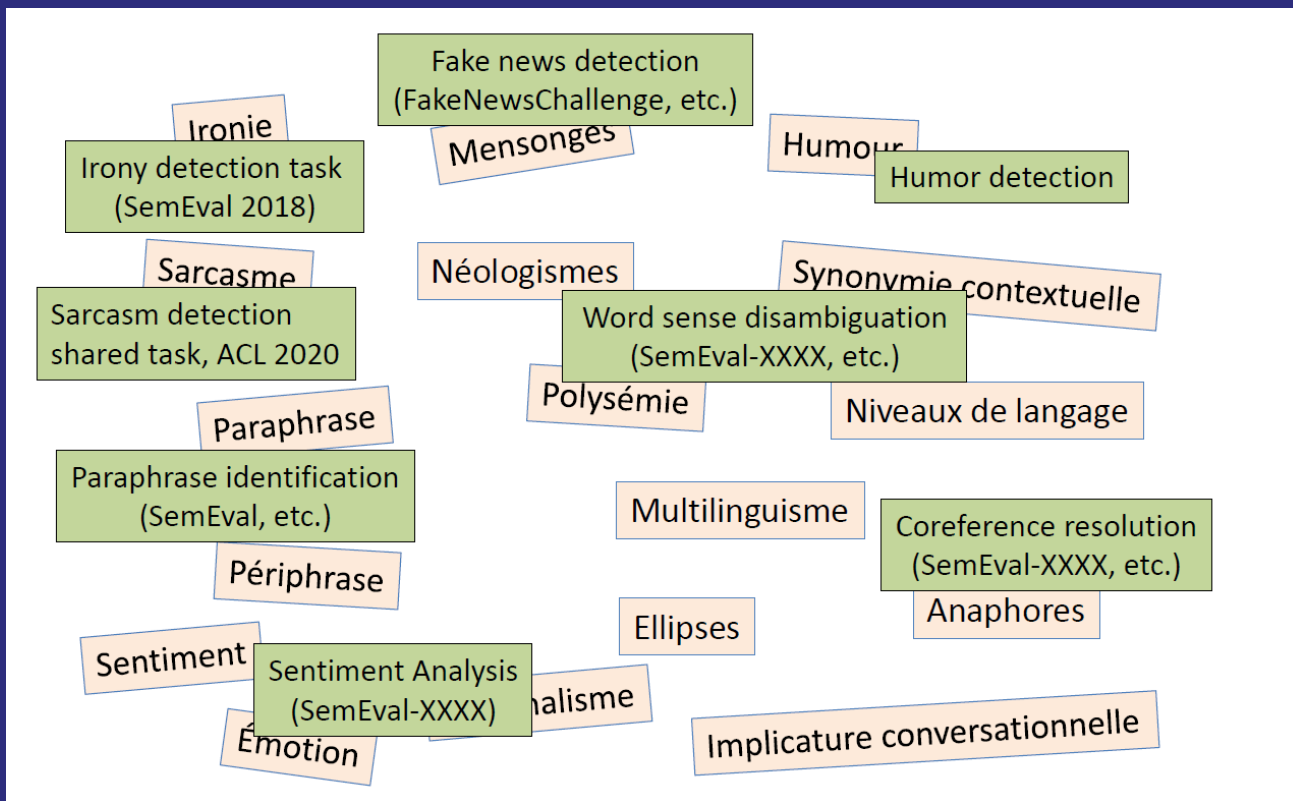
- Utile dans l'élimination des stop words
- Regroupement de termes complexes
- Manipuler des mots ambigus
- Construire une représentation syntaxique du texte

La notion de TASK en TAL



Inspiré du cours de X.
Tannier – ETAL 2021

Une shared task pour chaque tâche...



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

SemEval
CoNLL
EVALITA (Italie)
WMT
CLEF
DEFT fouille de texte (France)

BEA 2024 Shared Tasks @ [NAACL/BEA 2024](#)

[Automated Prediction of Item Difficulty and Item Response Time](#)
[Multilingual Lexical Simplification Pipeline](#)

BEA 2023 Shared Task @ [ACL/BEA 2023](#)

[Generating AI Teacher Responses in Educational Dialogues](#)

BEA 2019 Shared Task @ [ACL/BEA 2019](#)

[Grammatical Error Correction](#)

BEA 2018 Shared Tasks @ [NAACL/BEA 2018](#)

[Second Language Acquisition Modeling](#)
[Complex Word Identification](#)

BEA 2017 Shared Task @ [EMNLP/BEA 2017](#)

[Native Language Identification](#)

BEA 2016 Shared Task @ [NAACL/BEA 2016](#)

[Automated Evaluation of Scientific Writing](#)

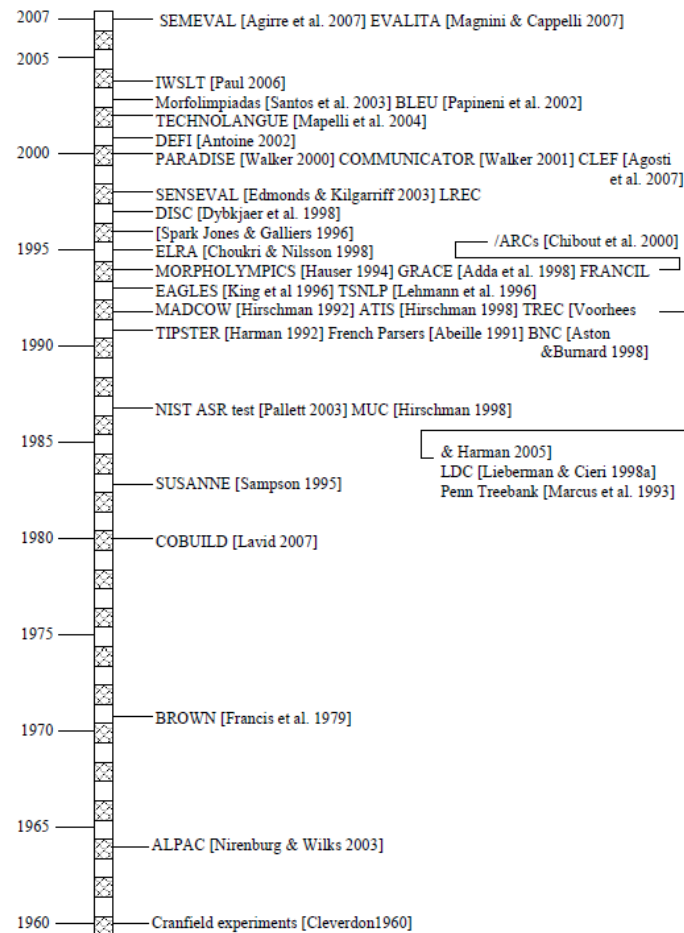


Figure 1. Salient events related to evaluation mentioned in this article (for evaluation campaign series, e.g. like TREC, only the first event is mentioned).

● Paroubek et al.,
2007

L'importance des campagnes d'évaluation...

- Permettent de faire avancer la recherche sur des thématiques spécifiques

Période statistique

Corpus

CoNLL – Computational Natural Language Learning
Shared Task → format(s) de donnée structuré(s) très répandu(s)
et utilisé(s)

1	They	they	PRON	PRP	Case=Nom Number=Plur	2	nsubj	2:nsubj 4:nsubj
2	buy	buy	VERB	VBP	Number=Plur Person=3 Tense=Pres	0	root	0:root
3	and	and	CCONJ	CC	—	4	cc	4:cc
4	sell	sell	VERB	VBP	Number=Plur Person=3 Tense=Pres	2	conj	0:root 2:conj
5	books	book	NOUN	NNS	Number=Plur	2	obj	2:obj 4:obj
6	.	.	PUNCT	.	—	2	punct	2:punct

Annotation de corpus

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

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

Focus sur la corréférence

Extraction d'information - **Corréférence**

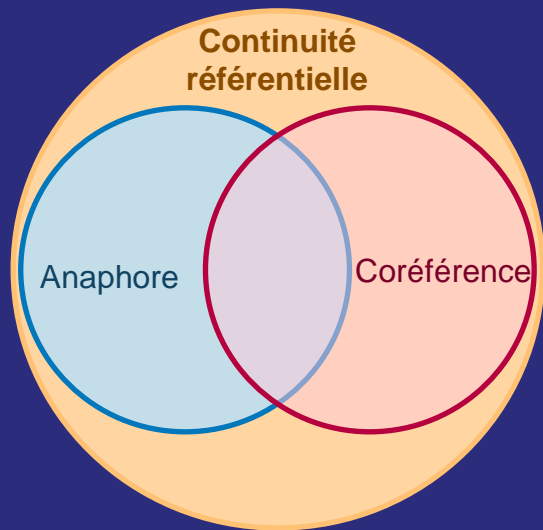
Une sorcière avait une maison noire et un chat noir. Il se cachait dans la maison pour ne pas qu'on le voie. Il ne sortait que la nuit. Un jour la sorcière en avait marre. Avec sa baguette magique elle l'a transformé en chat vert comme ça elle le voyait tout le temps.

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 » » (Schneider, 2019 p.11, Corblin, 1995)

Relation asymétrique

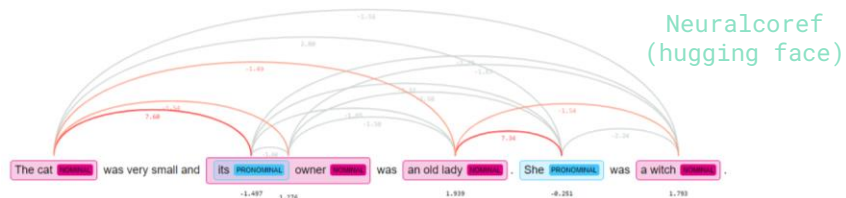
Coréférence

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

Relation symétrique

Utilité de la résolution de corréférence

Outils « off the shelf »



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

L. Grobol
(2020)

CI passing pypi v0.7.0 code style black

Detecting Coreferences for Oral French¹

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., Laidragin, F., Tellier, I., Lefeuvre, 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.

coreference resolution

 Scholar About 2,660 results (0.09 sec)

YEAR ▾

✕ Since 2021

Corpus annotés en coréférence

Dataset vs corpus