



université
PARIS-SACLAY

inria



CentraleSupélec

Introduction au TAL

ETAL 2025

1er septembre 2025

Laboratoire Interdisciplinaire des Sciences du Numérique – LISN, CNRS
sophie.rosset@lisn.fr

1 Définition(s)

2 La langue et le TAL

3 TAL : tâches et méthodes

4 Le TAL à l'ère des LLMs

5 Conclusion

6 EN et évaluation

7 Références et sources variées

Ce que ne sera pas ce cours

- Une présentation détaillée de méthodes
- Une présentation détaillée des tâches de TALP

Ce que ce sera ce cours

- Une (tentative de) explication de ce qu'est le TAL, pourquoi c'est particulier ...
- Une présentation (rapide et très haut niveau) des méthodes générales
- Une (tentative de) réflexion générale sur le TAL à l'ère des LLM

Plan de la présentation

1 Définition(s)

- Bref historique
- Le TAL

2 La langue et le TAL

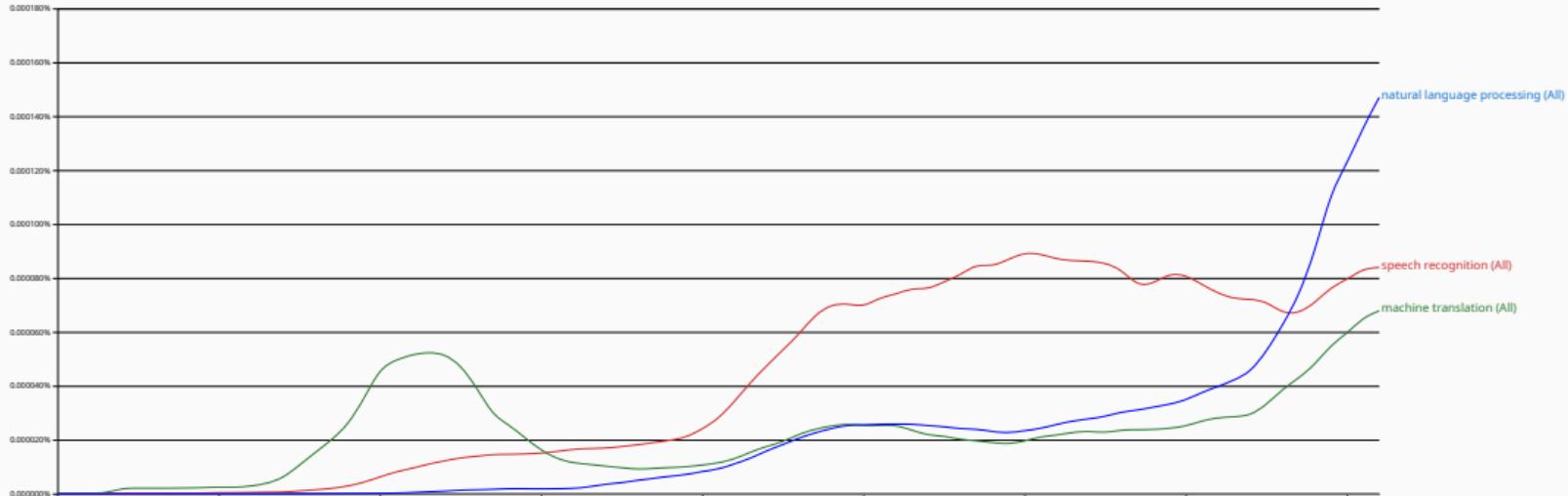
3 TAL : tâches et méthodes

4 Le TAL à l'ère des LLMs

5 Conclusion

6 EN et évaluation

De quoi on parle quand ?



<https://books.google.com/ngrams/>

- traitement automatique de la langue = TA et langue
- langue vs langage : fr, en etc. = pour communiquer ; langage = pour faire qqcvh (langage informatique par ex)
- traitement automatique : programme ?
- TAL : comment un ordinateur va se dépatouiller avec la langue

- premiers ordinateurs → traiter des choses qui ont de la langue dedans
- → traiter de la communication avec les humains
- à travers le temps, on a précisé, formalisé ce qu'on voulait faire avec : traduction, reconnaissance automatique de la parole, dialogue, compréhension, recherche d'informations...

Les premiers pas : traduction automatique (années 40-70)

- Contexte : Après la Seconde Guerre mondiale, besoins en traduction rapide (notamment pendant la Guerre froide) et en cryptanalyse
- Pionnier : Warren Weaver [[Wiever, 1949](#)]
The attached memorandum on translation from one language to another, and on the possibility of contributing to this process by the use of modern computing devices of very high speed, capacity, and logical flexibility, has been written with one hope only [...]
- 1954 : Georgetown–IBM experiment, 60 phrases soigneusement choisies, de russe vers anglais → d'ici 3 à 5 ans, la traduction automatique sera un problème résolu...
- 1966, rapport ALPAC : Après 10 années de recherche financées, les attentes n'ont pas été atteintes... → baisse drastique des financements
- 1959 : création de l'ATALA (Association pour l'étude et le développement de la Traduction Automatique et de la Linguistique Appliquée) !!!

Les premiers pas : reconnaissance de la parole (années 40-70)

- 1952 : Bell Laboratories, 1952 système “Audrey” : reconnaissance des chiffres (une seule voix) [Davis et al., 1952]
- 1962 : IBM shoebox, William C. Dersch (voir <https://tinyurl.com/4j8uvep5>)
 - commande vocale : “Seven plus three plus six plus nine plus five. Subtotal.” → it worked !

The limitation of this fine instrument, the human voice, became apparent when modern man confronted the machine. Now machine, however large and complex, must be instructed. And since they do not understand human speech, men have invented many ingenious control devices. Good ones, if sometimes less convenient than spoken commands.

→ nécessaire de faire des systèmes de reconnaissance de la parole (incluant la compréhension)
- 1969 : Bell Labs suspend ses financements suite à une critique par J. R. Pierce (Directeur Exécutif, Recherche) de la recherche sur la reconnaissance automatique de la parole. Les financements ont repris à l'arrivée à ce poste de J. L. Flanagan (1971 ?)...

- Vouloir des systèmes qui comprennent la langue, la traduisent, la reconnaissent est aussi vieux (voire plus) que les ordinateurs
- Il y a un lien avec l'IA dont l'une des composantes est la compréhension de la langue
- Un rêve aussi dans la SF et autres
 - Les assistants vocaux omniscients : HAL 9000 (2001 : L'Odyssée de l'espace, 1968), Mother (Alien, 1979), LCARS/Computer (Star Trek, dès 1966)
 - Les androïdes et robots à conversation naturelle : C-3PO (Star Wars, 1977), R2-D2 (Star Wars, même s'il parle avec des bips !), Data (Star Trek : The Next Generation)
 - Les traducteurs universels : Universal Translator (Star Trek), Babel Fish (Le Guide du voyageur galactique, Douglas Adams, 1979), Tardis (Doctor Who),
 - Les intelligences artificielles incarnées : Samantha (Her, 2013), Jarvis/F.R.I.D.A.Y. (Marvel Cinematic Universe), GLaDOS (Portal, 2007)
 - Les systèmes de reconnaissance et de contrôle vocal : KITT (K 2000, 1982), EVA (Wall-E, 2008)
 - et bien d'autres

- problème de l'ambiguïté
- la langue et la logique
- traduction et théorie de l'information
- Langues and invariants

- problème de l'ambiguïté
 - il suffit de prendre un contexte - pas besoin qu'il soit long car finalement peu de mots sont ambigüs (hmmm)
- la langue et la logique
- traduction et théorie de l'information
- Langues and invariants

- problème de l'ambiguïté
- la langue et la logique
 - ordinateur construit avec des boucles régénératives d'un certain caractère formel est capable de déduire toute conclusion légitime à partir d'un ensemble fini de prémisses McCulloch and Pitts [1988]
 - et que la langue (écrite) est "une expression de caractère logique"
→ le problème de la traduction automatique peut être résolu de manière formelle (hmmm)
- traduction et théorie de l'information
- Langues and invariants

- problème de l'ambiguïté
- la langue et la logique
- traduction et théorie de l'information
 - Il est possible d'utiliser la théorie de l'information et la cryptographie (puisque cela concerne les propriétés statistiques fondamentales de la communication)
 - → Wiener n'a pas été convaincu

"I frankly am afraid the boundaries of words in different languages are too vague and the emotional and international connotations are too extensive to make any quasi mechanical translation scheme very hopeful."
- Langues and invariants

- problème de l'ambiguïté
- la langue et la logique
- traduction et théorie de l'information
- Langues and invariants

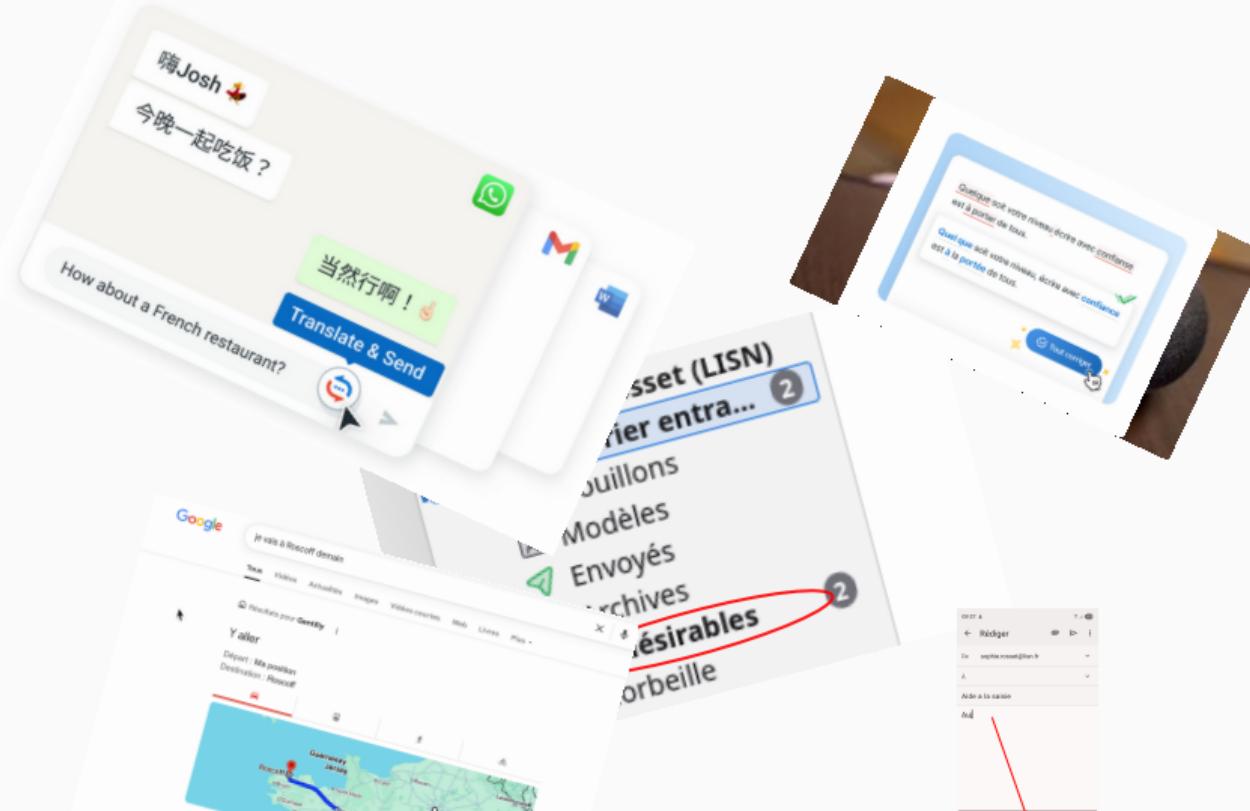
Thus may it be true that the way to translate from Chinese to Arabic, or from Russian to Portuguese, is not to attempt the direct route, shouting from tower to tower. Perhaps the way is to descend, from each language, down to the common base of human communication - the real but as yet undiscovered universal language - and then re-emerge by whatever particular route is convenient.

- problème de l'ambiguïté
 - la langue et la logique
 - traduction et théorie de l'information
 - Langues and invariants
- si Wiever avait une vision peut-être un peu simpliste des problèmes posés, il n'en reste pas moins qu'il avait compris pas mal d'entre eux... (allez lire [Translation !](#))

Définition(s)

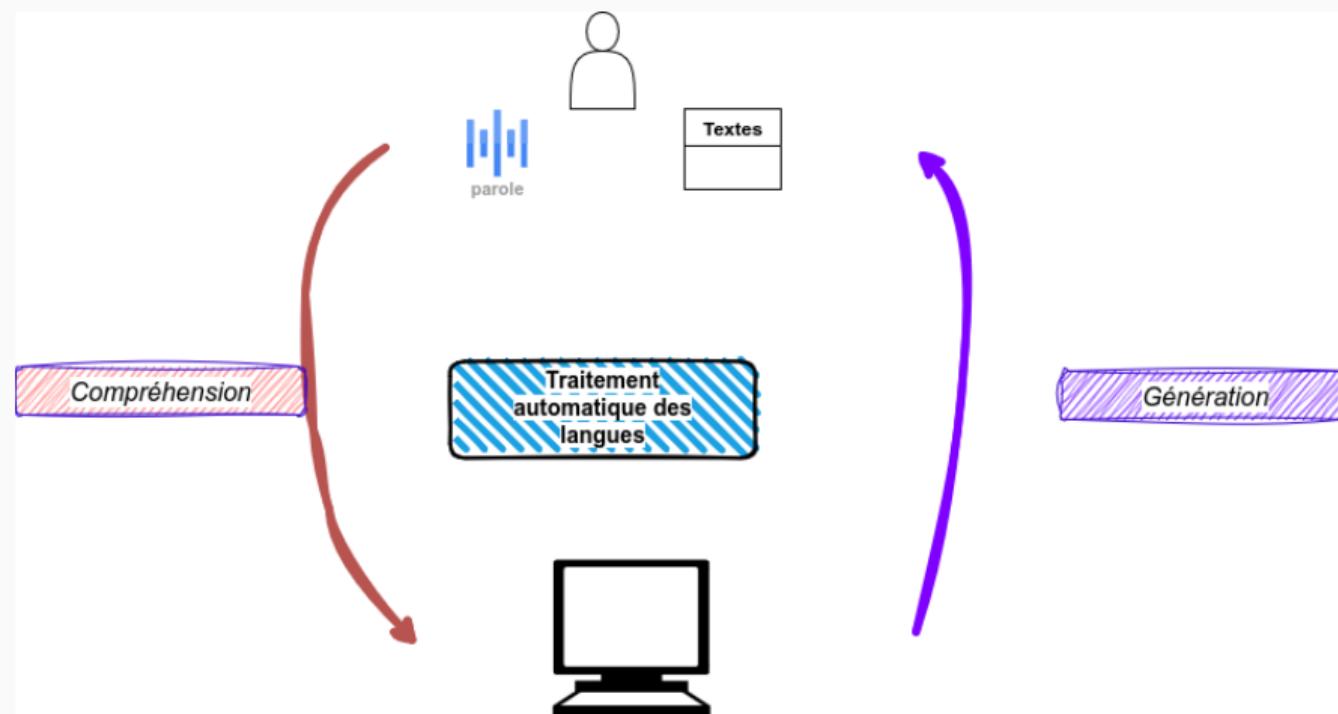
Le TAL

Le TAL est présent partout



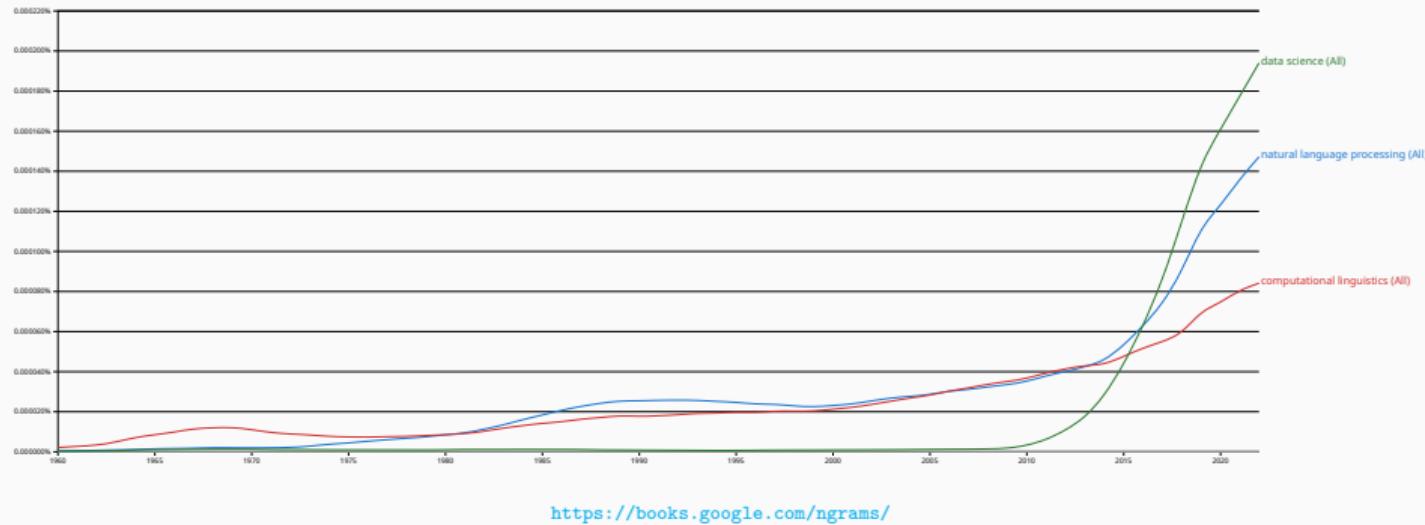
Le TAL est

Un ensemble d'outils qu'on cherche à développer pour traiter la langue comme nous le faisons...

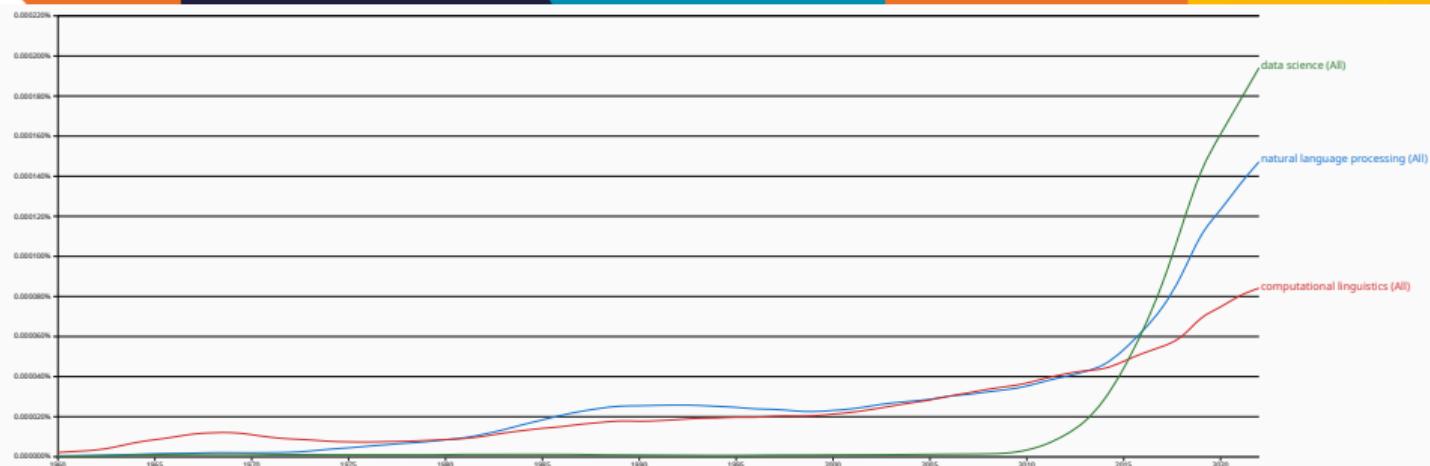


Une définition large

- Le traitement automatique de la langue et de la parole (TALP) est une branche de l'intelligence artificielle (IA) qui vise à permettre aux ordinateurs de comprendre, analyser, manipuler et générer le langage humain de manière utile.
- → proposer et implémenter des méthodes permettant à un ordinateur/système/etc. de manipuler la langue, la communication humaine



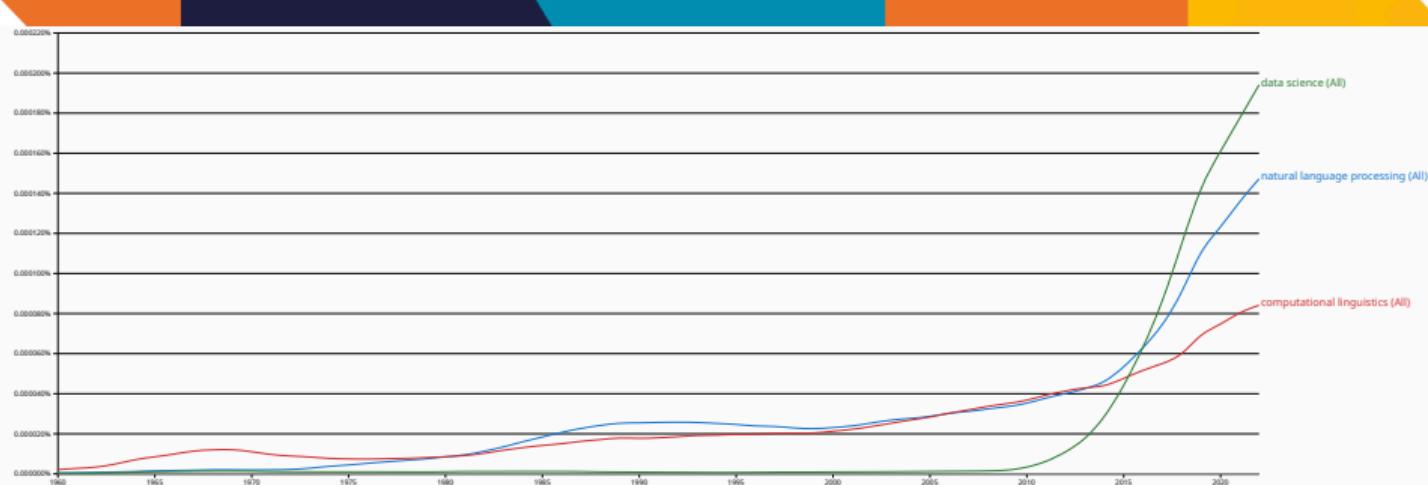
TAL vs X



<https://books.google.com/ngrams/>

- TAL vs linguistique computationnelle
 - LC : comprendre et modéliser la langue (très présent au début de la TA)
 - TAL : proposer des méthodes et développer des systèmes informatiques capables de manipuler la langue
 - les deux peuvent (et doivent ?) se nourrir l'une de l'autre (voir plus loin LLM).

TAL vs X



<https://books.google.com/ngrams/>

- TAL vs sciences des données
 - SD : extraire de la connaissance à partir de données, quelle que soit leur nature (numériques, textuelles, images, etc.)
 - TAL : comprendre, interpréter, générer le langage humain à l'aide d'ordinateurs/systèmes
 - il y a un lien entre les deux même si la nature des données diffèrent, enrichissement mutuel possible

Plan de la présentation

1 Définition(s)

2 La langue et le TAL

- Données linguistiques et distributions
- La langue n'est pas un langage comme les autres

3 TAL : tâches et méthodes

4 Le TAL à l'ère des LLMs

5 Conclusion

6 EN et évaluation

La langue et ses particularités

- données linguistiques et distribution
- langue \neq langage
- langue : multicanale, multimodale

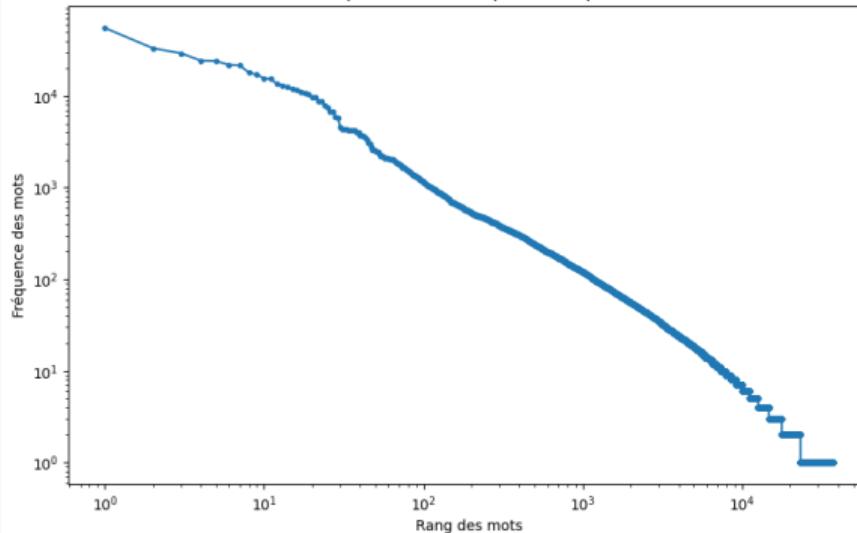
Optimisée pour la communication entre humains

La langue et le TAL

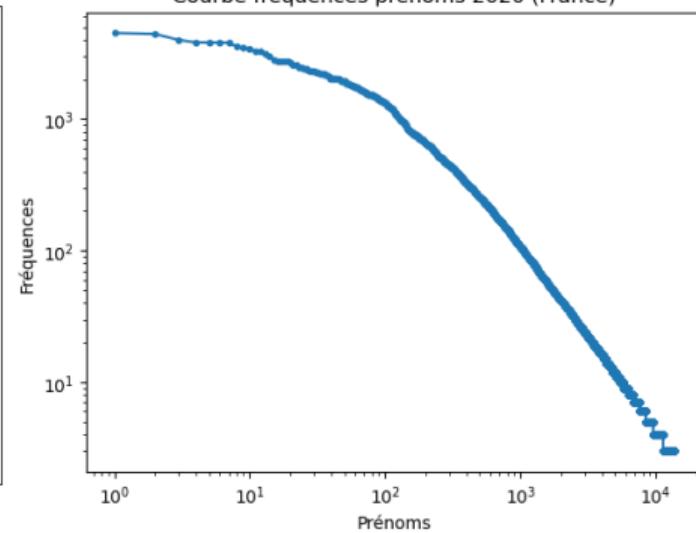
Données linguistiques et distributions

Distributions

Courbe fréquences de mots pour le corpus ESTER2

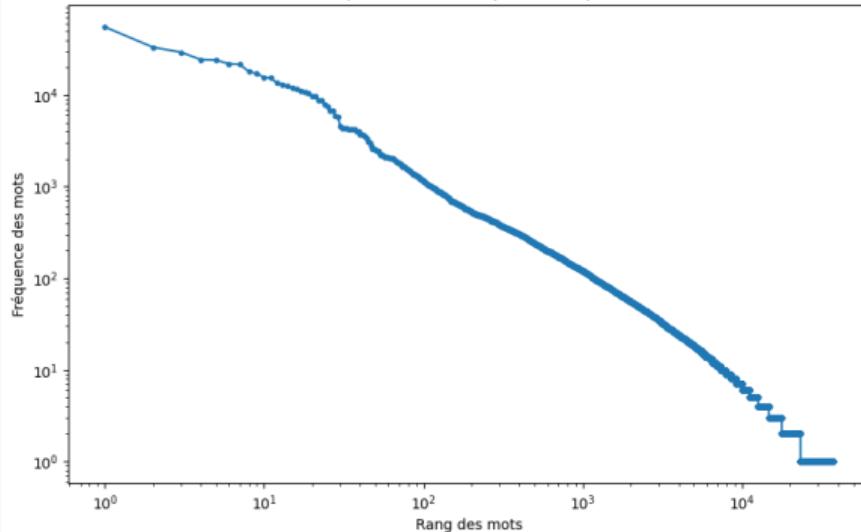


Courbe fréquences prénoms 2020 (France)

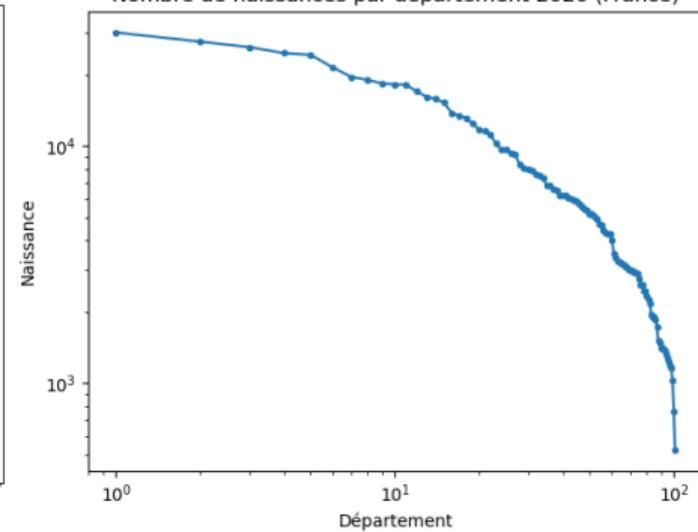


Distributions

Courbe fréquences de mots pour le corpus ESTER2



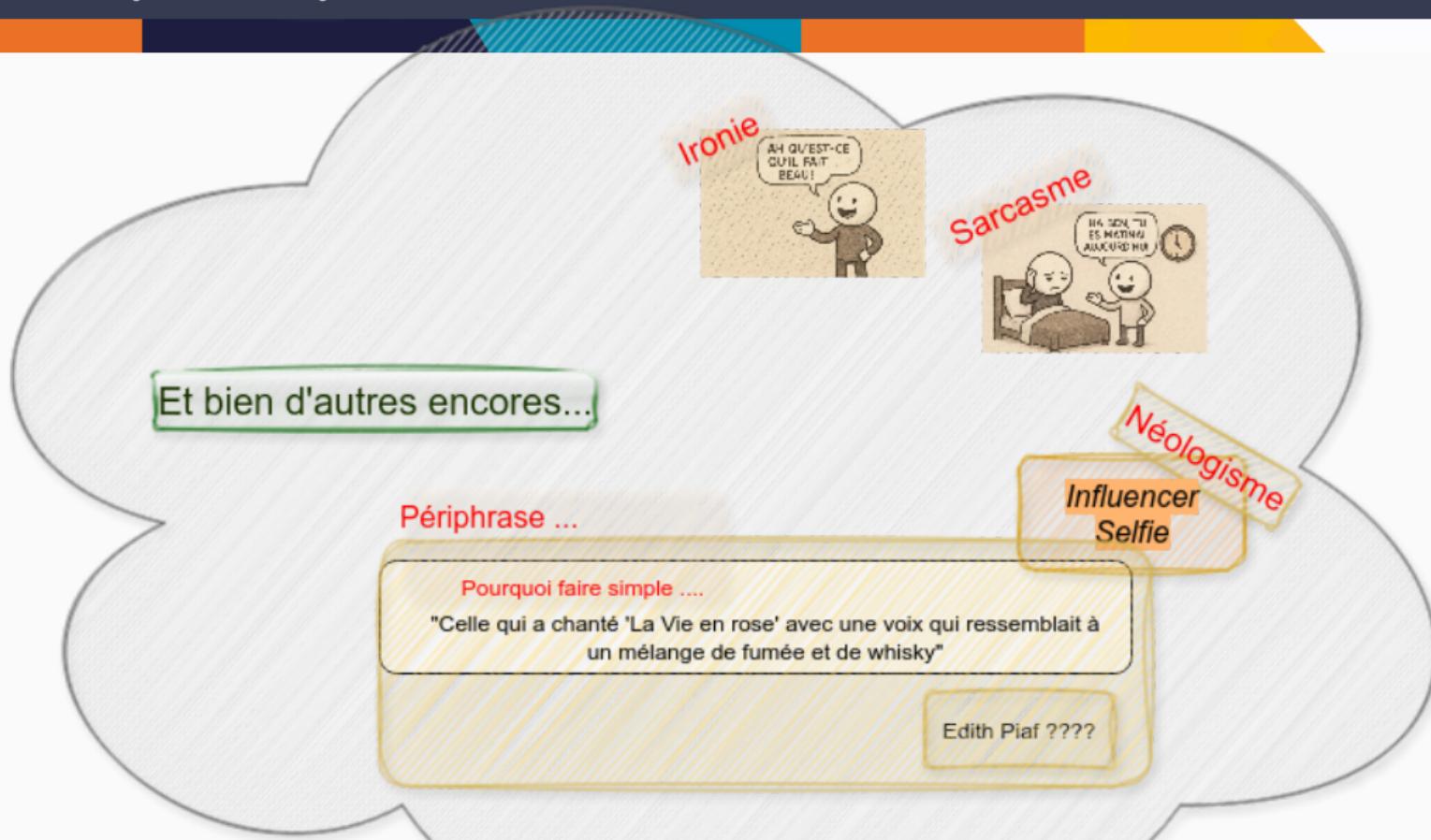
Nombre de naissances par département 2020 (France)



La langue et le TAL

La langue n'est pas un langage comme les autres

Langue est optimisée pour la communication entre humains



Homonymie

J'ai pris un avocat
la belle porte le voile

homophonie

Cinq capucins, sains de corps et d'esprit, portaient sur leur sein le seing du saint père.

Métonymie

Le gouvernement a cédé à la rue
Il a bu un Bordeaux/un verre/une bouteille

Des graphies différentes

Kadhafi Khadafi Qaddafi Gaddafi Kadafi Khaddafi Qadhafi Gadafi Kaddafi Khadafy Qadafi
Gadafy Kadaffi Khadaffi Qaddaffi Gaddafy

...

La langue est utilisée/créée par des humains qui évoluent

Amis lecteurs, qui ce livre lisez, Despouillez vous de toute affection ; Et, le lisant, ne vous scandalisez : Il ne contient mal ne infection. Vray est qu'icy peu de perfection Vous apprendrez, si non en cas de rire ; Aultre argument ne peut mon cuer elire, Voyant le dueil qui vous mine et consomme : Mieulx est de ris que de larmes escrire, Pour ce que rire est le propre de l'homme.

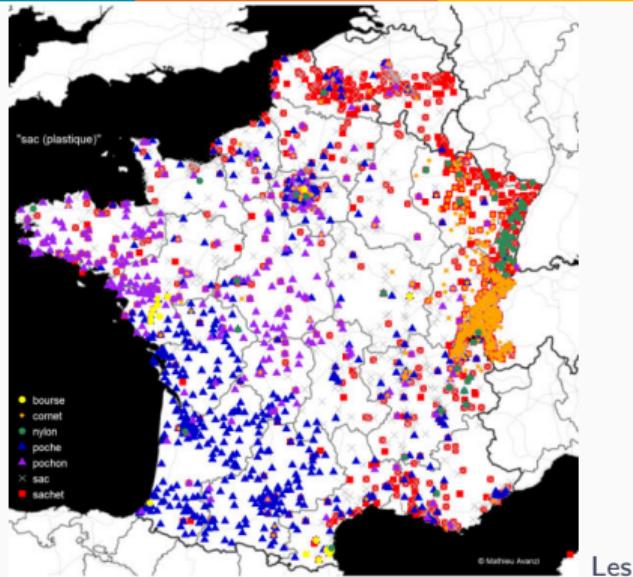
Amis lecteurs, qui lisez ce livre, Dépouillez-vous de toute passion. Et, en le lisant, ne soyez pas scandalisés. Il ne contient ni mal ni corruption ; Il est vrai qu'ici vous ne trouverez Guère de perfection, sauf si on se met à rire ; Autre sujet mon cuer ne peut choisir À la vue du chagrin qui vous mine et consume. Il vaut mieux traiter du rire que des larmes, Parce que rire est le propre de l'homme.

La vie très horrificque du grand Gargantua père de Pantagruel (1534), Rabelais

La langue permet d'exprimer de différentes façons...

Je tiens à exprimer mon désaccord sur ce point.
À ce sujet, j'ai une opinion divergente et je
tiens à le faire savoir. Je n'approuve pas. Je ne
suis pas d'accord.

Cc Tom koi29 ? Coucou Tom quoi de neuf ?
Eske Ta capT ou pa ? Est-ce que tu as capté
ou pas ? Est-ce que tu as compris ou pas ?



dénominations du "sac (plastique)" en français régional, d'après les résultats de l'enquête Euro-1. (d'après l'article de Mathieu Avanzi, "Cornet, poche, pochon, sac, sachet ou autre ?" du 31 août 2016)

La langue permet d'exprimer de différentes façons...

Elle est multimodale, multicanale ...

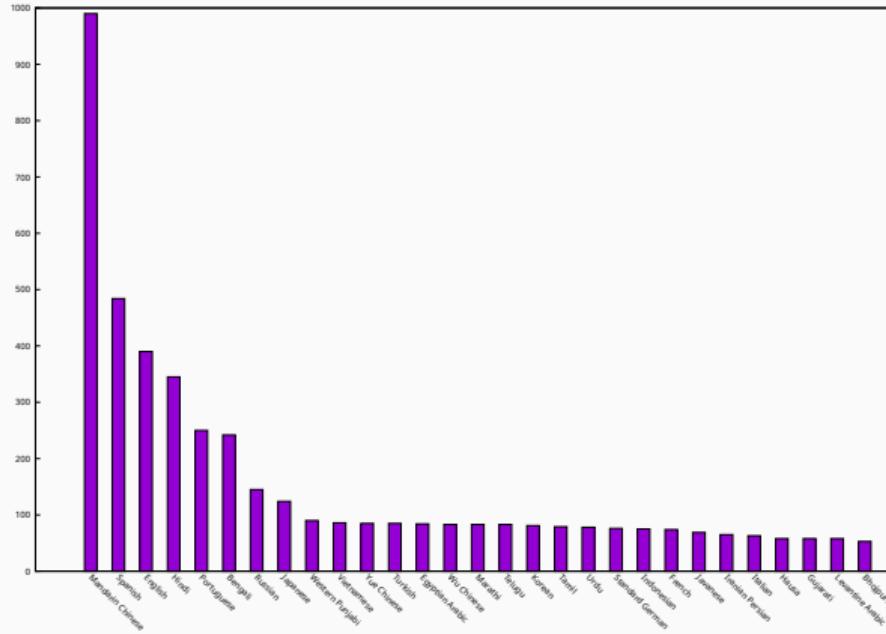


Vidéos de la séance du 9 juillet 2025



Allocution de E. Macron - 5 mars 2025

La langue est ... multiple



- env. 7000 langues parlées dans le monde
- une 20aine parlées par 95% de la population.

Wikipedia et Langue et Liste des langues par nombre de spkr

Ces langues peuvent se mélanger (code switching) :

This morning I hantar my baby tu dekat babysitter tu lah (This morning I took my baby to the babysitter) - anglais, malais

Cette fête a l'air fun, let's go !

Hay que mopear en la kitchen (je vais passer la serpillère dans la cuisine, avec *mopear* qui est une création), - anglais, espagnol

Holnap morning shiftet csinálok'. (Je fais le shift du matin demain.) - anglais, hongrois

En bref, la langue est



un système évolutif de signes linguistiques, vocaux, graphiques et/ou gestuels, qui permet la communication entre les individus. (Wikipedia)

un système optimisé par et pour les humains

...

Plan de la présentation

1 Définition(s)

2 La langue et le TAL

3 TAL : tâches et méthodes

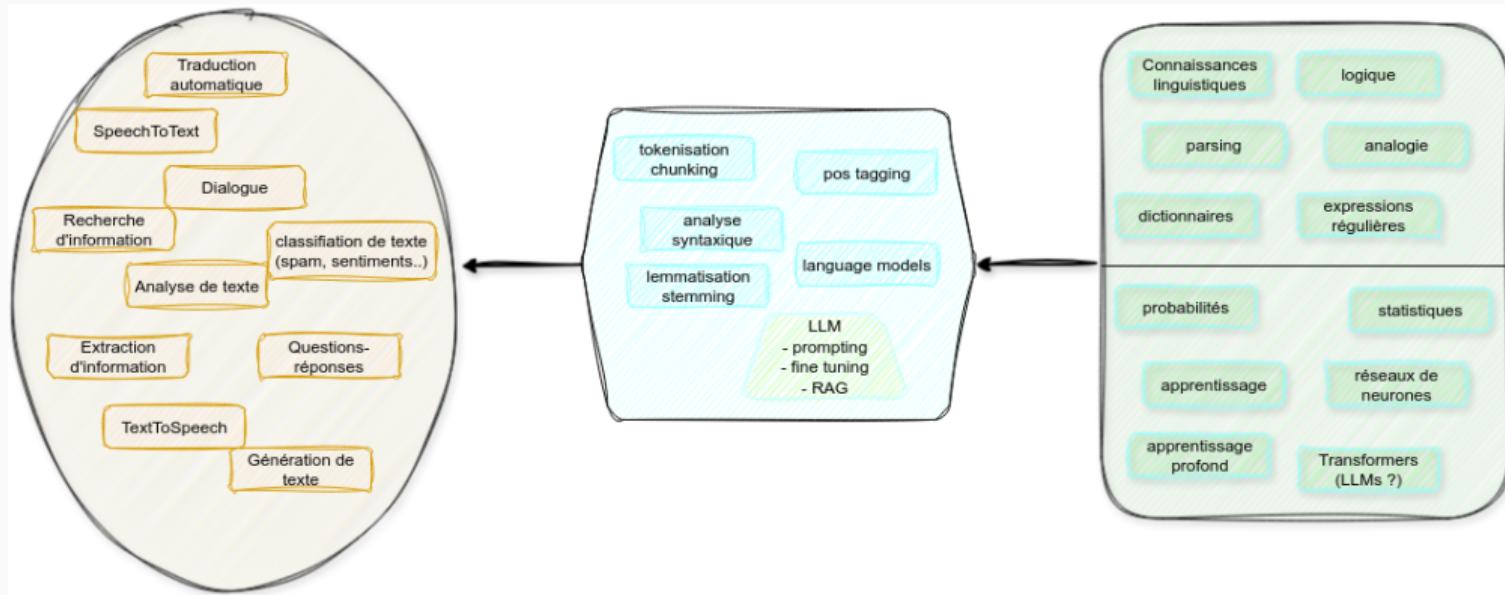
- Un monde ... pas réel... ni idéal
- Construire un système de TAL (ère pré-LLM)
- Exemples avec les entités nommées

4 Le TAL à l'ère des LLMs

5 Conclusion

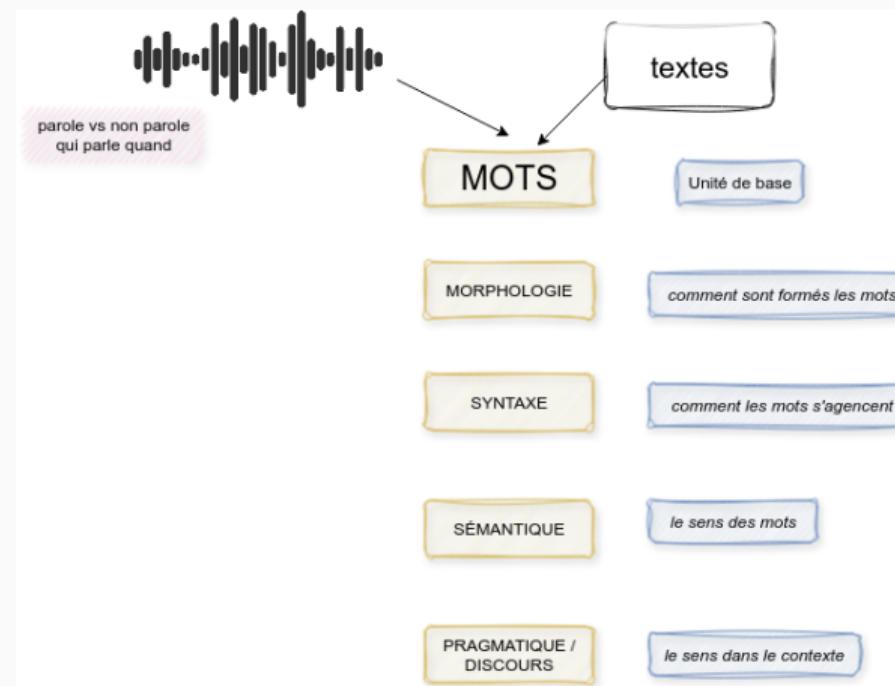
TAL : tâches et méthodes/techniques

Des tâches, des techniques, des méthodes...



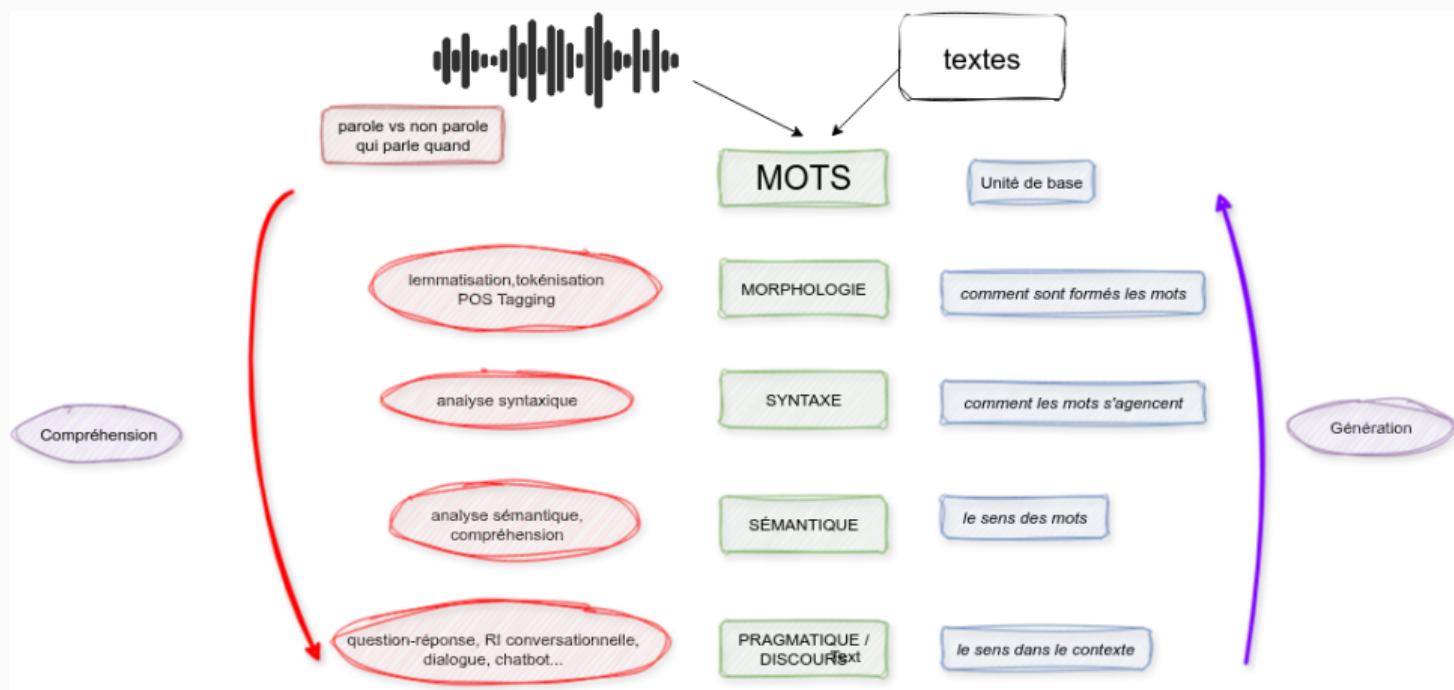
Comment on fait (dans un monde ... pas réel)

La langue est un système, structuré et organisé... on pourrait s'appuyer dessus :



Comment on fait (dans un monde ... pas réel)

La langue est un système, structuré et organisé... on pourrait s'appuyer dessus :



Un mot, vous avez dit un mot ?

La terre est bleue comme une orange

Un mot, vous avez dit un mot ?

La terre est bleue comme une orange

Un mot, vous avez dit un mot ?

Mais

U.S.A ?

Avons-nous ?

L'arbre ?

2020-2021 ?

EDF ?

s'il-vous-plaît ?

<https://etal2025.sciencesconf.org/> ?

#JPP ?

Un mot, vous avez dit un mot ?

- attarsinnaanngorpoq : il lui devint possible de partir

Un mot, vous avez dit un mot ?

- attarsinnaanngorpoq : il lui devint possible de partir
- attar·sinnaa·nngor·poq : attaq (partir), sinnaa (pouvoir), nngoq (devenir), pu (indicatif) + q (3e pers.) - groenlandais

Un mot, vous avez dit un mot ?

- attarsinnaanngorpoq : il lui devint possible de partir
- Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz : loi sur le transfert des obligations de surveillance de l'étiquetage de la viande bovine, 1999

Un mot, vous avez dit un mot ?

- attarsinnaanngorpoq : il lui devint possible de partir
 - Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz : [loi sur le transfert des obligations de surveillance de l'étiquetage de la viande bovine](#), 1999
 - Rindfleisch·etikettierungs·überwachungs·aufgaben·übertragungs·gesetz : Rind (boeuf), Fleisch (viande), Etikettierung (étiquetage), Überwachung (surveillance), Aufgabe (tâche), Übertragung (transfert), Gesetz (loi) - allemand
- ⇒ au-delà des problèmes, la segmentation en mot est utile, simple et explicable

Beaucoup, beaucoup de mots → réductions de la taille du vocabulaire ?

- Lemmatisation
 - forme canonique du mot (infifitif pour un verbe, par exemple mangera, mangeons → manger)
 - demande des ressources, des règles, des lexiques → dépendante d'une langue
 - → de 39902 à 29015 (corpus ESTER2)
- Racinisation (Stemming)
- Suppression des mots outils (stop words)
- Subwords (tokens)

Beaucoup, beaucoup de mots → réductions de la taille du vocabulaire ?

- Lemmatisation
- Racinisation (Stemming)
 - racine du mot : on supprime les flexions, préfixes etc.
 - algorithme : Porter, Carry, Paice/Husk
 - → de 39902 à 22442 (corpus ESTER2)
- Suppression des mots outils (stop words)
- Subwords (tokens)

Beaucoup, beaucoup de mots → réductions de la taille du vocabulaire ?

- Lemmatisation
- Racinisation (Stemming)
- Suppression des mots outils (stop words)

- Ce sont les mots courts les plus fréquents

pos	mots	fréquence
1	de	55221
2	la	33199
3	le	29263
4	l	24310
...	...	
47	france	2796

- quel seuil ?
- Subwords (tokens)

Beaucoup, beaucoup de mots → réductions de la taille du vocabulaire ?

- Lemmatisation
- Racinisation (Stemming)
- Suppression des mots outils (stop words)
- Subwords (tokens)
 - Byte-Pair Encoding (BPE), WordPiece, UnigramLM et SentencePiece
 - granularité plus fine que le stemming ou la lemmatisation
 - permet de gérer les mots inconnus (si le modèle en est capable)
 - peut ajouter de la complexité et du bruit dans le modèle

⇒ la tokenisation permet de gérer les mots inconnus, de représenter n'importe quelle séquence de caractère, n'est pas linguistiquement fondée, d'où perte d'explicabilité

TAL : tâches et méthodes

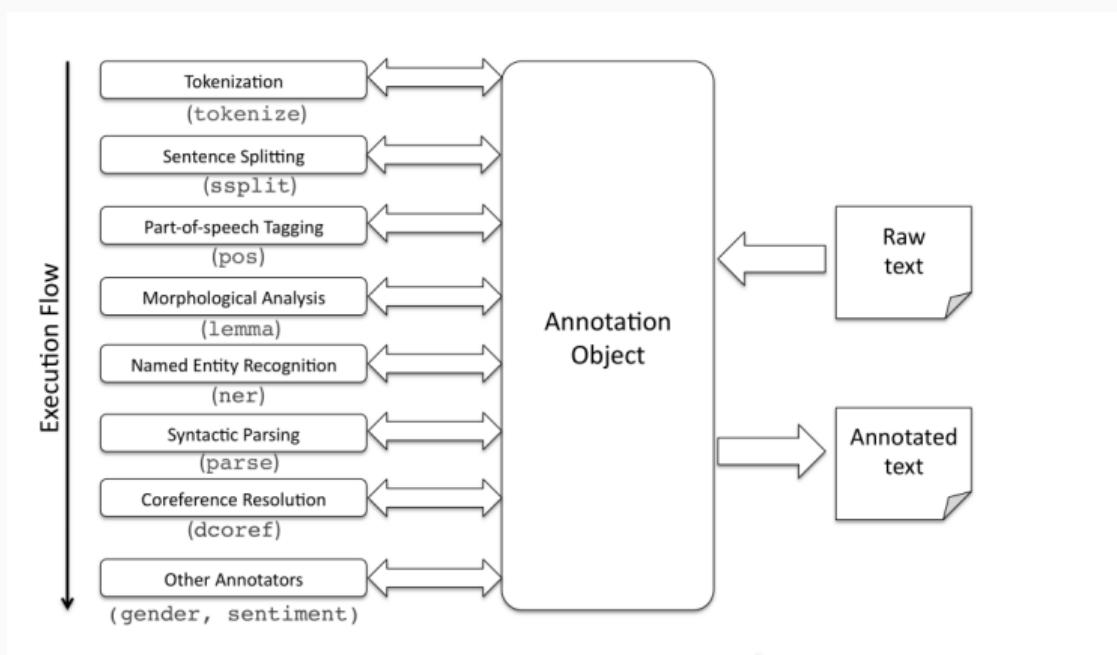
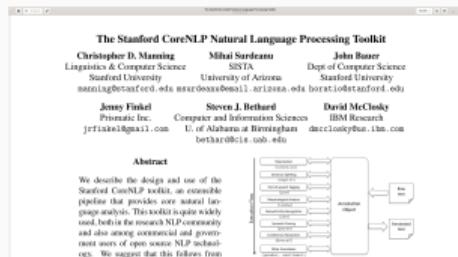
Construire un système de TAL (ère pré-LLM)

Construire un système de TAL

- Ce n'est pas UN programme
- C'est un ensemble de "briques"

Construire un système de TAL

[Manning et al., 2014]

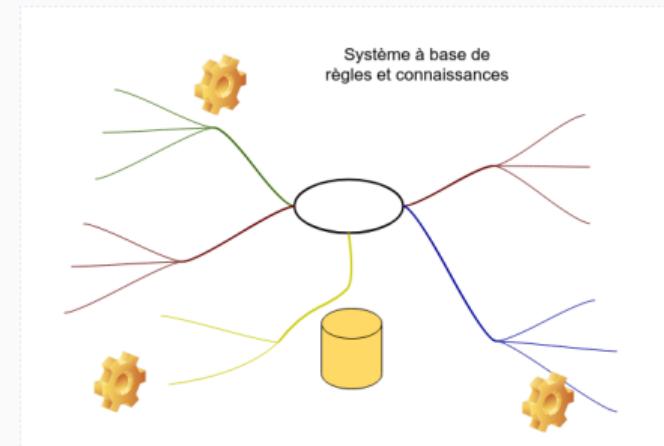


Comment on fait ?



Système à base de règles

- définir la tâche (manuel, guide)
- annoter test
- faire des listes de mots clés, d'expression régulière etc.
- nettoyer les données
- appliquer sur test
- évaluer

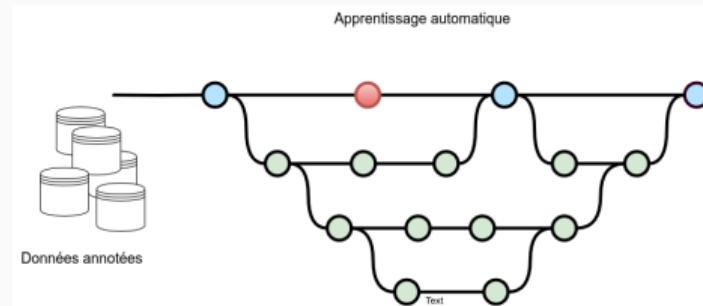


Comment on fait ?



Système avec apprentissage

- définir la tâche (manuel, guide)
- nettoyer les données
- annoter test, trn, dev
- choisir un algorithme d'apprentissage (NB, SVM, MaxENT, CRF, etc.)
- apprendre le modèle sur trn, bidouiller sur dev
- appliquer sur test
- évaluer

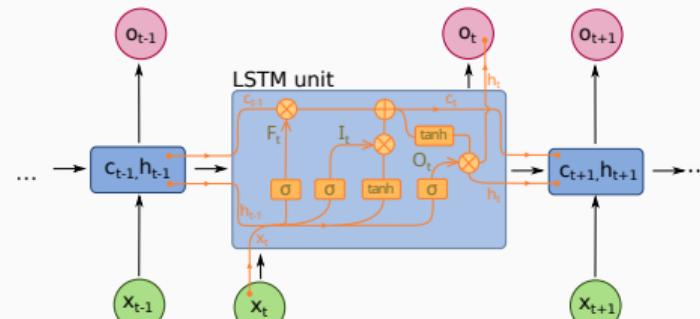


Comment on fait ?



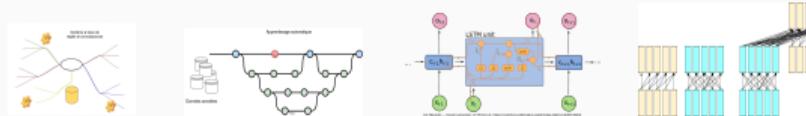
Système avec apprentissage profond

- choisir une architecture
- annoter test, trn, dev
- nettoyer les données
- apprendre un modèle sur le trn, bidouiller sur le dev
- appliquer sur test
- évaluer



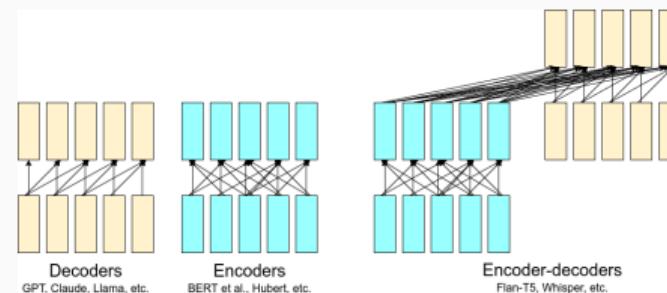
Per fdeolche — Travail personnel, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=60148410>

Comment on fait ?



Système avec LLM

- choisir un LLM
- choisir un prompt ???
- appliquer sur test
- évaluer



⇒ données annotées pour apprentissage, test...

Quelle que soit la tâche et l'approche choisie, besoin de données annotées

Implique de :

- définir la tâche (le modèle d'annotation)
- rédiger un guide d'annotation
- (faire) annoter des données

⇒ métriques d'évaluation

Quelle que soit la tâche et l'approche choisie, besoin de métriques d'évaluation

Implique de :

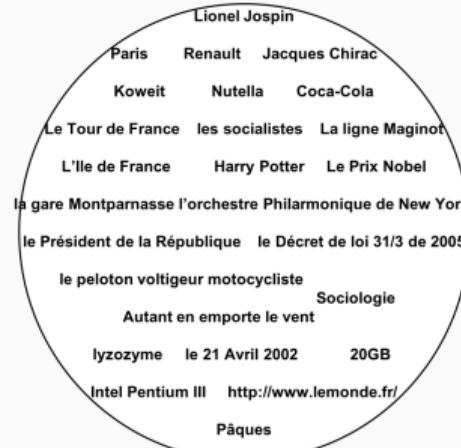
- définir ce qu'on veut évaluer
- définir une métrique

TAL : tâches et méthodes

Exemples avec les entités nommées

Les EN dans le monde : le problème de la catégorisation

- Le choix des catégories



- La détermination de ce qu'elles recouvrent

Catégorie PERSONNE :

Lionel Jospin	les Démocrates	Bison Futé
les Windsors	les Talibans	le Prince Charmant
la famille Kennedy	Zorro	l'épouse Chirac
les frères Cohen	St Nicolas	...

→ catégorisation instable

Les EN dans le texte : le problème de l'annotation

- Combinaisons de syntagmes : une ou plusieurs entités ?

Les Banques centrales américaine et européenne ont décidé...

Bill et Hillary Clinton

l'Université de Corte

- Un syntagme : quelles frontières ?

la candidate Ségolène Royal, Professeur Paolucci

George W. Bush Jr., La Mecque, l'Abbé Pierre

- Une entité : quelle unité lexicale ?

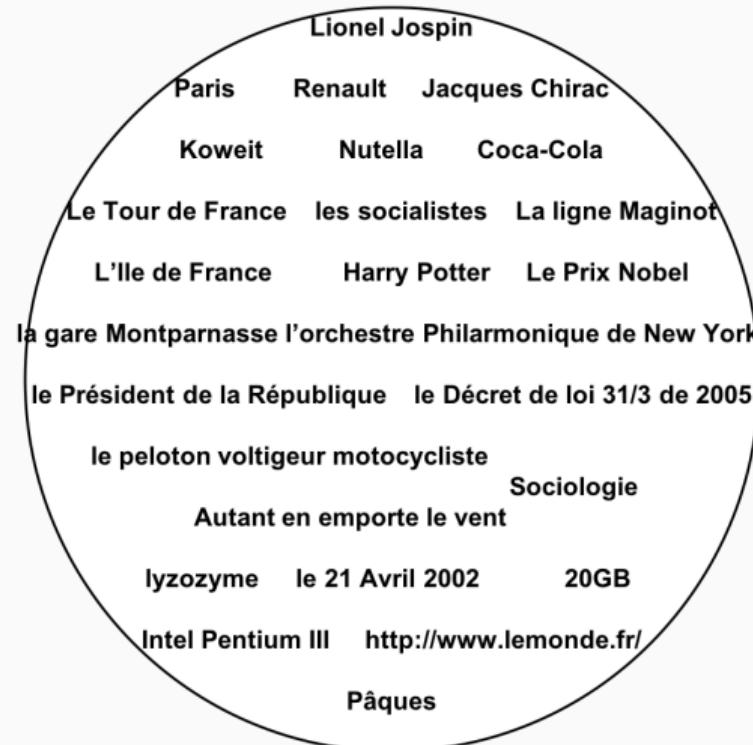
Jacques Chirac, Monsieur Chirac, le Président Jacques Chirac,

le Président français, le Président de la République française, Chichi

→ caractérisation imprécise, diversité des mentions

Le “matériau” de départ

Unités lexicales considérées
comme des entités nommées



Le “matériau” de départ



noms propres



Ce que l'on peut hésiter à qualifier
de nom propre



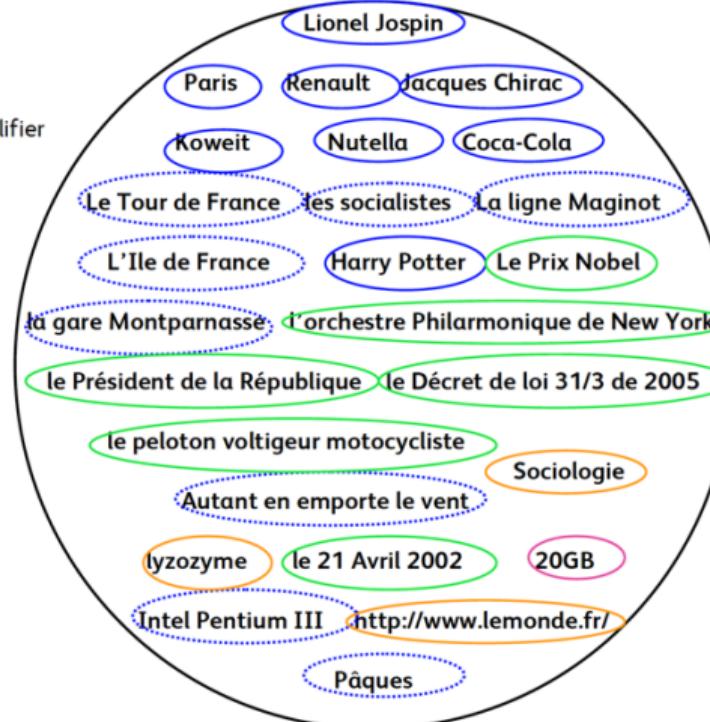
descriptions définies



expressions numériques



autre



Considération des aspects liés au TAL

Etant donné un **modèle applicatif** et un **corpus**, on appelle entité nommée toute expression linguistique qui **réfère** à une **entité unique** du modèle de manière **autonome dans le corpus**.

Illustration

Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.

Laguna	le président de la République en 2005		
	Jacques Chirac	Napoléon III	
	le président	je	30°
		l'Empereur des Français	2028hPa
Ivan	le président de la République en 2007		
	l'ouragan	Louise Colet	l'été 2004

Application : générique « typique »

Modèle : Personnes, Lieux, Organisations

Corpus : journalistique français de 1998 à 2008

Illustration

Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.

Laguna	le président de la République en 2005		
	Jacques Chirac	Napoléon III	
le président	je		30°
	l'Empereur des Français		2028hPa
Ivan	le président de la République en 2007		
	l'ouragan	Louise Colet	l'été 2004

Application : étude sur le climat

Modèle : températures, mesures atmosphérique, ouragan, dates, périodes, ...

Corpus : totalité des observations météorologiques sur une période données

De la linguistique au TAL, spécification d'un cadre théorique pour les EN :

- perspective linguistique : non réductibles à une catégorie mais caractérisables par un comportement référentiel
- perspective TAL : existent relativement à un modèle applicatif précis

→ Pas d'entité nommée « en soi », seulement des critères linguistiques et un modèle.

Conséquences

- point de vue général : explication de l'hétérogénéité et de la variabilité de l'ensemble 'entités nommées'
- point de vue pratique : critères de décision pour annoter
- point de vue méthodologique : besoin impératif d'expliciter le modèle

- Une typologie (ou *tagset*) est une **formalisation descriptive** des catégories d'EN à prendre en compte :
 - quoi reconnaître (cibler des éléments appartenant à des catégories spécifiques)
 - comment le représenter (pour un élément, choisir une catégorie parmi d'autres)
- De **nombreuses variations** en fonction des domaines et des applications
 - différences de catégories
 - différences de structure
 - différences sur la définition de ce que recouvrent les catégories

→ des années 80 aux années 2010, une multitude de typologies (plus de 20 inventoriées en 2016), et donc de corpus...

→ Guide d'annotation QUAERO [[Rosset et al., 2011](#)], 86 p.!!!

Comparaison de typologies par l'exemple

MUC d'après le Bureau du recensement des LOC[Etats-Unis] , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE[2011] .

ACE d'après le ORG[Bureau du recensement des Etats-Unis] , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE[2011] .

ESTER d'après le ORG[Bureau du recensement des LOC[Etats-Unis]] , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE[2011] .

QUA d'après le ORG [name [Bureau du recensement] des
LOC [name[Etats-Unis]]] , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE[year[2011]] .

1. **reconnaissance** : détecter, repérer des entités nommées dans les flux textuels (on pose les frontières dans le texte)
2. **classification** : catégoriser les éléments reconnus selon des catégories sémantiques pré-définies (on affecte un type)
3. **désambiguïsation/liaison** : lier les mentions d'entités à une référence unique (on lie à une référence)
4. **extraction de relation** : découvrir des relations entre entités (*father-of, born-in, alma mater*)

Construire des systèmes logiciels qui effectuent ces tâches de manière automatique.

Exigences :

- **qualité** : ne pas faire trop d'erreurs
- **exhaustivité** : ne pas manquer trop d'entités
- **robustesse** : ne pas échouer face à des cas non canoniques

En pratique :

- difficile de répondre à ces 3 exigences simultanément
- recherche du **meilleur compromis** en fonction des ressources et de l'application.

La **représentation des textes** comme séquences de mots donne 2 niveaux de granularité :

- les **caractères**, qui forment un mot
- les **mots**, qui composent une séquence (un texte)

Les **indices** peuvent être caractérisés au niveau :

- des caractères : **indices morphologiques**
- des mots eux-mêmes : **indices lexicaux**
- de la séquence de mots : **indices contextuels**

⇒ très utiles et utilisés même si pas toujours fiables (aussi avec TAL à base de règles qu'avec ML)

Détection des EN : approches symboliques (les règles)

- **Objectif** : insertion de balises dans les textes indiquant où se trouvent les ENs
- **Principe** : conception de *règles* formant un *grammaire locale*
- De nombreuses **boîtes à outils** :
 - GATE
 - LingPipe
 - NooJ
 - OpenNLP
 - OpenCalais
 - Unitex
 - WMatch

Possibilité d'avoir des prétraitements :

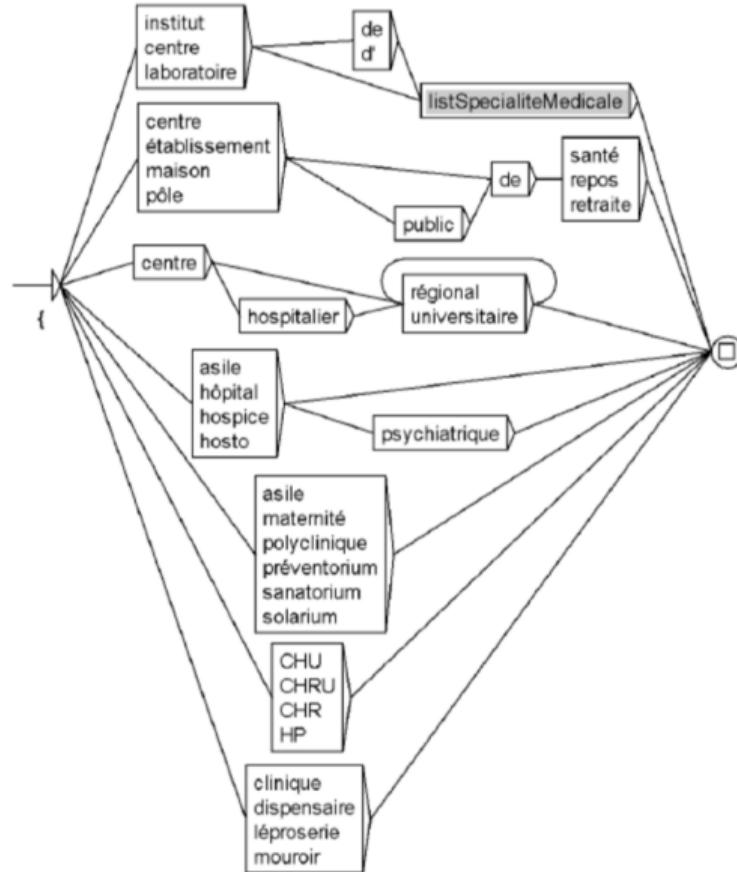
segmentation en mots, en phrases, étiquetage morphosyntaxique.

→ indices supplémentaires fort utiles,
mais qui impactent les performances si bruités.

- des éléments sont indiqués au sein des nœuds
- les noeuds sont agencés de manière à reconnaître des expressions linguistiques
- les transitions sont réalisées par présence d'indices (morphologiques, lexicaux, internes ou externes)
- plusieurs transitions sont réalisées par juxtaposition de nœuds
- l'automate ne reconnaît une expression linguistique que s'il existe un chemin depuis le nœud initial (à gauche) jusqu'au nœud final (à droite).

Objectif : contraindre correctement l'automate, afin qu'il reconnaisse toutes les expressions linguistiques souhaitées, et aucune autre.

Automates



Basculement vers les approches statistiques

Au début des années 2000, grâce à la mise à disposition de jeux de données volumineux.

Mais les approches symboliques sont toujours présentes :

- combinées avec des méthodes statistiques
- prédominent pour les langues ou les typologies sans corpus de données suffisants
- gardent l'avantage pour le contrôle et de l'ingénierie : plus compréhensibles, modulables, possibilités de réglages fins.
- majoritaires dans le milieu industriel.

Objectif : déterminer les paramètres d'un modèle à partir de données, d'où le terme *apprentissage*

Ces paramètres et ce modèle sont ensuite utilisés pour prendre les décisions les plus probables (ou vraisemblables) sur de nouvelles données à traiter.

Il s'agit, simultanément, de spécifier le modèle et de généraliser les données.

A partir des années 1960, émergent les modèles connexionnistes (perceptron, réseaux de neurones), qui mettent en relation des propriétés sur les objets modélisés.

Puis les modèles markoviens (modèles de Markov à états cachés), qui simulent des processus stochastiques.

→ remise en cause du principe déterministe des automates et la manière dont sont élaborés les systèmes.

Objectif : tenir compte de la vraisemblance d'**étiquettes contiguës**

François Hollande

- *Hollande* : Lieu ou Personne ?
- *François*, annoté comme Personne, peut conditionner l'annotation du mot *Hollande*

Option : modèles génératifs comme les modèles de Markov à états cachés.

Calcul des probabilités inversé : déterminer, pour une suite d'étiquettes, la probabilité qu'elle génère un texte donné.

$$P(M_1, M_2 \dots M_n | E_1, E_2 \dots E_n) = \prod_{i=1}^n P(M_i | E_i) * P(E_i | E_{i-1})$$

Soit le produit des probabilités de génération $P(M_i | E_i)$ et de transition $P(E_i | E_{i-1})$.

Modèles à décisions contextuelles (HMM)

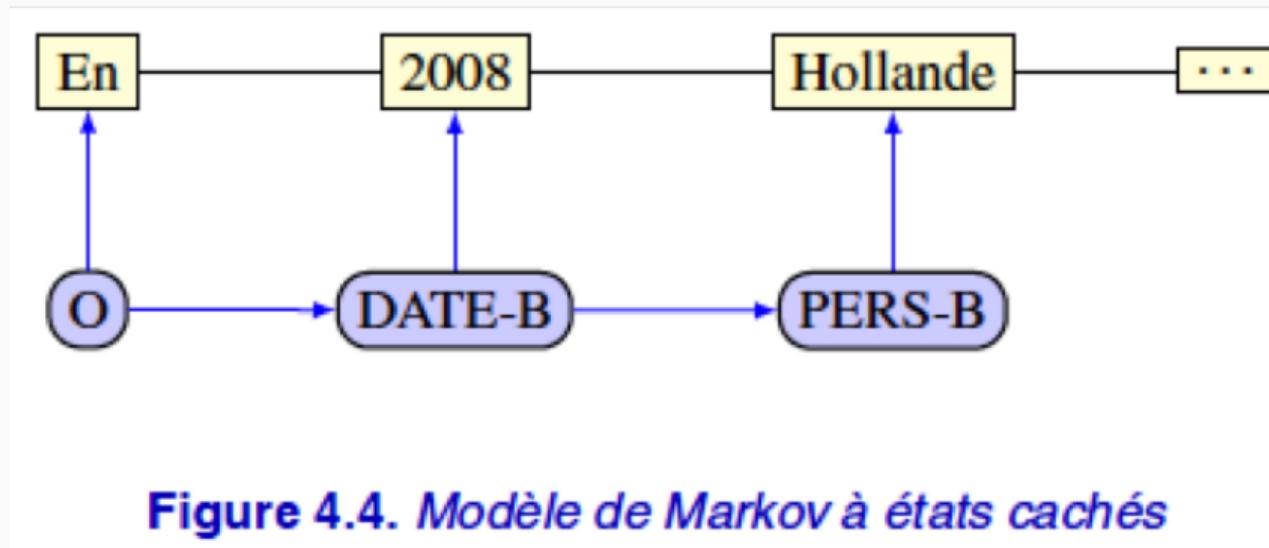


Figure 4.4. Modèle de Markov à états cachés

Décisions non indépendantes : la solution la plus vraisemblable est choisie en fonction des étiquettes préalablement choisies.

Modèles utilisant des indices multiples (softmax, MaxEnt)

Objectif : considérer plus d'indices que les mots, i.e. prendre en compte la morphologie, les indices lexicaux, le contexte, etc.

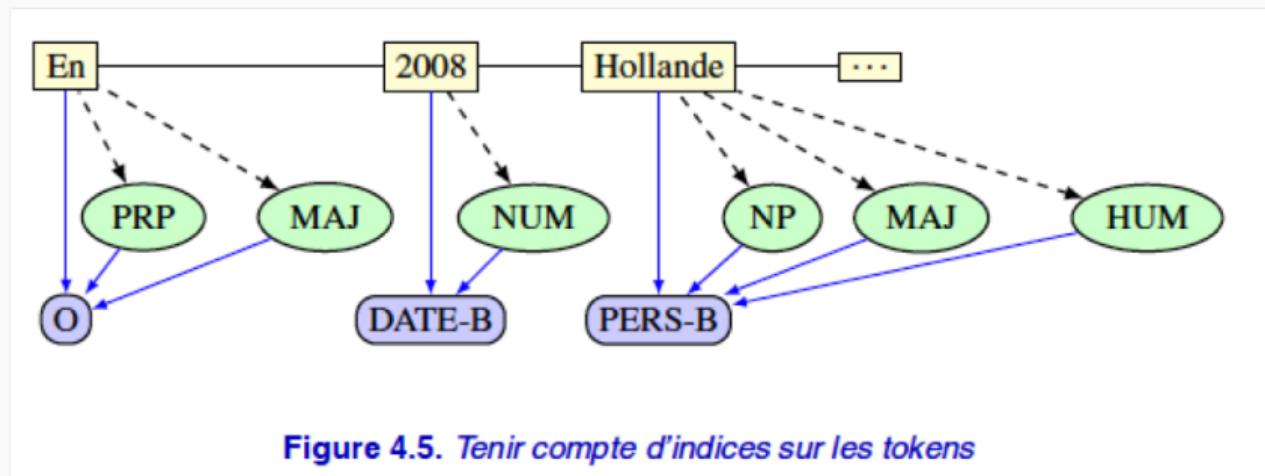


Figure 4.5. Tenir compte d'indices sur les tokens

Champs markoviens conditionnels (CRF)

Les CRF (*Conditional Random Fields* ou champs markoviens conditionnels) combinent les deux aspects précédents :

- **tenir compte du contexte** pour prendre des décisions
(une décision sur un mot influence la décision pour le mot suivant)
- **tenir compte de multiples indices**
(analyses en prétraitements)

Modèle qui obtient de très bonnes performances pour la reconnaissance d'EN.

Ref : Lafferty, MacCallum, Pereira, 2001.

Champs markoviens conditionnels (CRF)

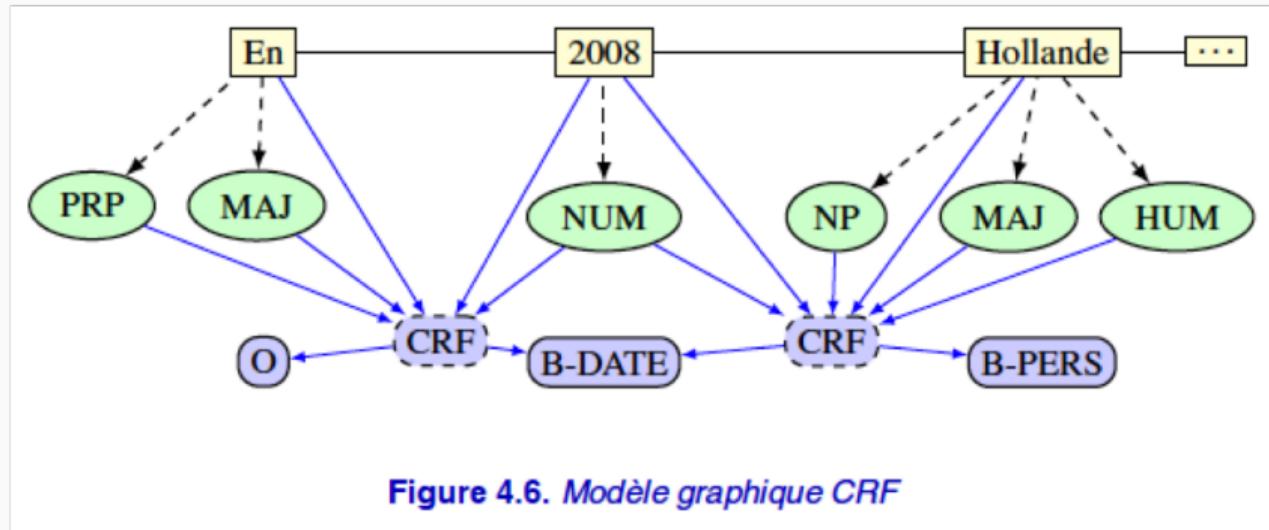


Figure 4.6. Modèle graphique CRF

$$G(e, m, f_1 \dots f_k) = \exp \left(\sum_{p=1}^k \alpha_{ep} * f_p \right)$$

Détection des EN et réseaux de neurones profonds

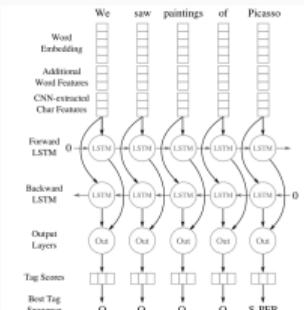


Figure 1: The (unrolled) BLSTM for tagging named entities. Multiple tables look up word-level feature vectors. The CNN (Figure 2) extracts a fixed length feature vector from character-level features. For each word, these vectors are concatenated and fed to the BLSTM network and then to the output layers (Figure 3).

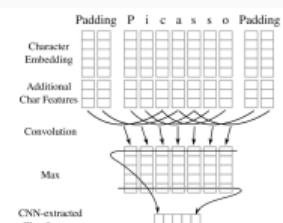


Figure 2: The convolutional neural network extracts character features from each word. The character embedding and (optionally) the character type feature vector are computed through lookup tables. Then, they are concatenated and passed into the CNN.

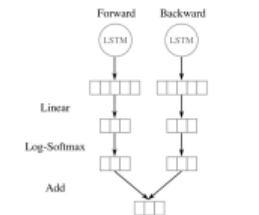


Figure 3: The output layers ("Out" in Figure 1) decode output into a score for each tag category.

Hyper-parameter	CoNLL-2003 (Round 2)		OntoNotes 5.0 (Round 1)	
	Final	Range	Final	Range
Convolution width	3	[3, 7]	3	[3, 9]
CNN output size	53	[15, 84]	20	[15, 100]
LSTM state size	275	[100, 500]	200	[100, 400] ¹⁰
LSTM layers	1	[1, 4]	2	[2, 4]
Learning rate	0.0105	[10^{-3} , $10^{-1.8}$]	0.0008	[$10^{-3.5}$, $10^{-1.5}$]
Epochs ¹¹	80	-	18	-
Dropout ¹²	0.68	[0.25, 0.75]	0.63	[0, 1]
Mini-batch size	9	- ¹³	9	[5, 14]

Table 3: Hyper-parameter search space and final values used for all experiments

Architecture et paramètres [Chiu and Nichols, 2016]

Plan de la présentation

1 Définition(s)

2 La langue et le TAL

3 TAL : tâches et méthodes

4 Le TAL à l'ère des LLMs

- Les LLMs en bref
- LLM et TAL
- Impact des LLM

5 Conclusion

Le TAL à l'ère des LLMs

Les LLMs en bref

- Attention dans un encoder-decoder : 2014 [Bahdanau et al., 2015]
- Transformer : 2017 [Vaswani et al., 2017]
- Learning with instructions : 2022 [Ouyang et al., 2022]

- Attention dans un encoder-decoder : 2014 [Bahdanau et al., 2015]

Published as a conference paper at ICLR 2015

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TR

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

The most important distinguishing feature of this approach from the basic encoder–decoder is that it does not attempt to encode a whole input sentence into a single fixed-length vector. Instead, it encodes the input sentence into a sequence of vectors and chooses a subset of these vectors adaptively while decoding the translation. This frees a neural translation model from having to squash all the information of a source sentence, regardless of its length, into a fixed-length vector. We show this allows a model to cope better with long sentences.

LLM : points de repère

- Attention dans un encoder-decoder : 2014 [Bahdanau et al., 2015]
- Transformer : 2017 [Vaswani et al., 2017]

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukasz.kaiser@google.com

Ilia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

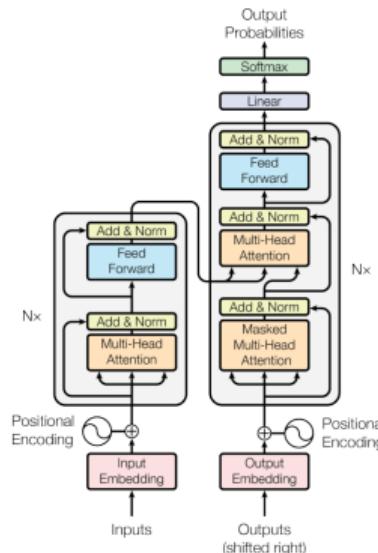


Figure 1: The Transformer - model architecture.

- Attention dans un encoder-decoder : 2014 [Bahdanau et al., 2015]
- Transformer : 2017 [Vaswani et al., 2017]
- Learning with instructions : 2022 [Ouyang et al., 2022]

Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*

Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray

John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens

Amanda Askell[†] Peter Welinder Paul Christiano^{*†}

Jan Leike* Ryan Lowe*

OpenAI

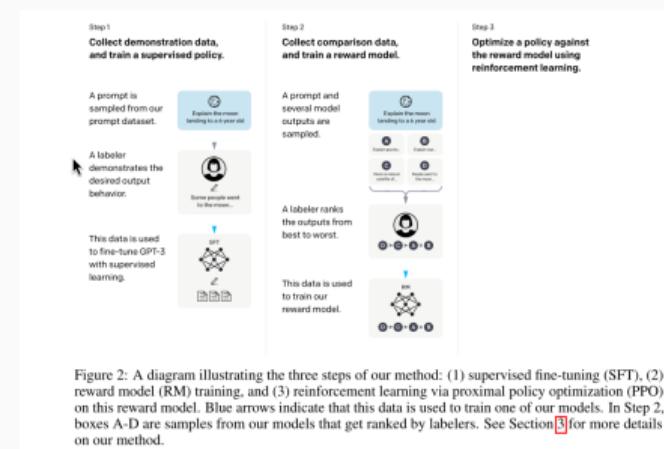


Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers. See Section 3 for more details on our method.

LLM en deux mots

- un modèle de langue
- entraîné sur de très très très grandes quantité de données
 - common crawl, colossal clean crawled corpus (C4, [Gao et al., 2020])...

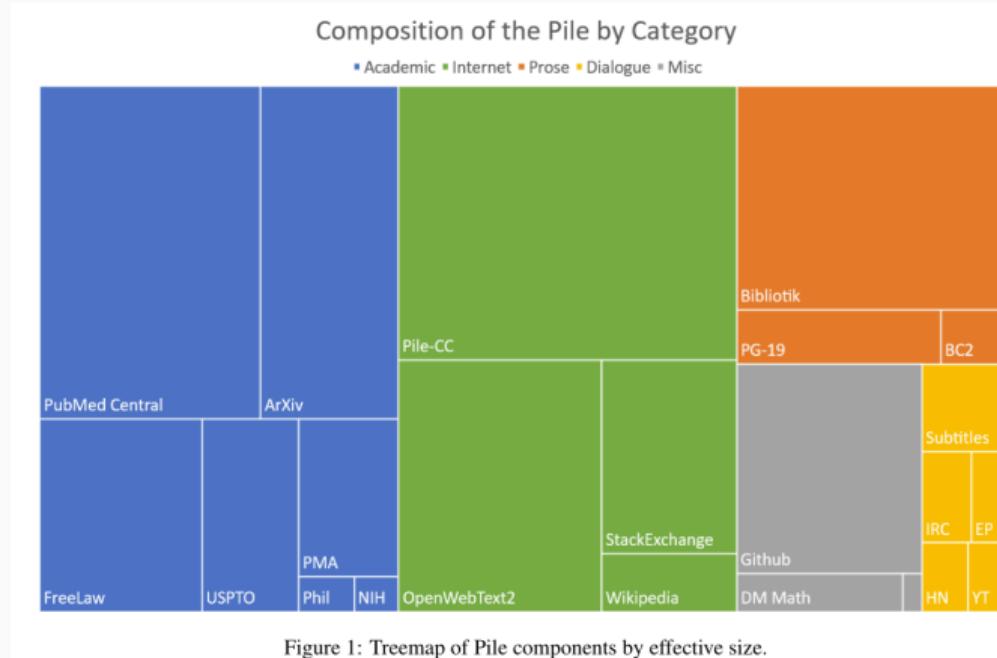
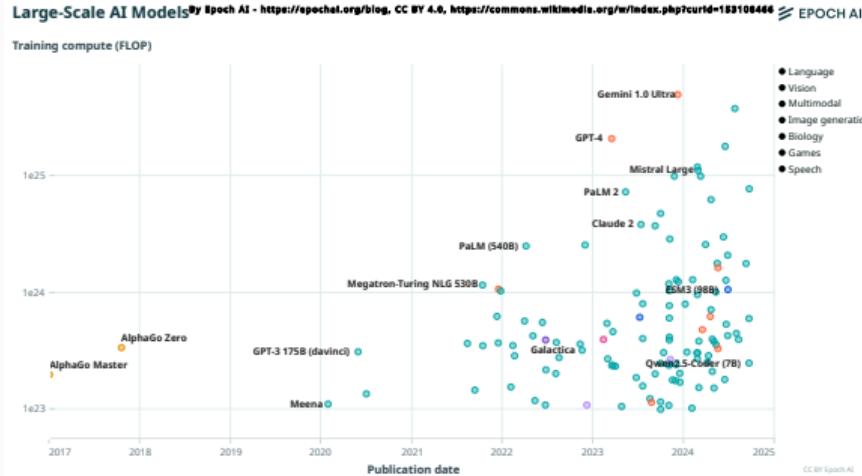


Figure 1: Treemap of Pile components by effective size.

LLM en deux mots

- un modèle de langue
- entraîné sur de très très très grandes quantité de données
- de plus en plus gros \Rightarrow Large Language Model (also see [LLM timeline](#))



- un modèle de langue
- entraîné sur de très très très grandes quantité de données
- pour générer (du texte)
 - entraîné à prédire des mots (par exemple en apprenant à prédire le mot suivant (ou autre))
- apprennent un tas de choses sur la langue (et plus !)
- peuvent être fine-tunés sur des données spécifiques
- peuvent être utilisés en mode zero ou few-shot notamment via du *prompting*

- un modèle de langue
- entraîné sur de très très très grandes quantité de données ⇒ Large Language Model
- pour générer (du texte)
- apprenent un tas de choses sur la langue (et plus !)
- peuvent être fine-tunés sur des données spécifiques
- peuvent être utilisés en mode zero ou few-shot notamment via du *prompting*
 - tout ou presque peut être exprimé en langue naturelle, y compris les instructions donnés au système pour accomplir des tâches NLP (*prompting*)

⇒ beaucoup de tâche peuvent se réécrire en tâche de génération !

Le TAL à l'ère des LLMs

LLM et TAL

BERT RedisCOVERS THE CLASSICAL NLP PIPELINE

Ian Tenney¹ Dipanjan Das¹ Ellie Pavlick^{1,2}

¹Google Research ²Brown University

{iftenney, dipanjand, epavlick}@google.com

ordering emerges. We find that while this traditional pipeline order holds in the aggregate, on individual examples the network can resolve out-of-order, using high-level information like predicate-argument relations to help disambiguate low-level decisions like part-of-speech. This provides new evidence corroborating that deep language models can represent the types of syntactic and semantic abstractions traditionally believed necessary for language processing, and moreover that they can model complex interactions between different levels of hierarchical information.

BERT RedisCOVERS THE CLASSICAL NLP PIPELINE

Ian Tenney¹ Dipanjan Das¹ Ellie Pavlick^{1,2}

¹Google Research ²Brown University

{iftenney, dipanjand, epavlick}@google.com

6 Conclusion

Did BERT rediscover an NLP pipeline? Not in a naïve, architectural sense. GridLoc reveals a structure in BERT that is more intricate than a flowchart of a pipeline could accurately portray, and yet it does seem to be linguistically founded. We find that probing results regarding BERT layers are unstable, diverging across sentence input, random seeds and the early iterations of training. The distribution of linguistically motivated task features along token positions, on the other hand, is relatively more stable. Moreover, GridLoc’s results on tree depth provide preliminary evidence of POSs being used to conduct novel but linguistically generalizable inference concerning a derivative syntactic phenomenon.

Does BERT Rediscover a Classical NLP Pipeline?

Jingcheng Niu

Wenjie Lu

Gerald Penn

University of Toronto

Vector Institute

{niu, luwenjie, gpenn}@cs.toronto.edu

GPT-NER: Named Entity Recognition via Large Language Models

[Wang et al., 2025a]

Shuhe Wang[†], Xiaofei Sun[†], Xiaoya Li[‡], Rongbin Ouyang[†]
Fei Wu[†], Tianwei Zhang[†], Jiwei Li[†], Guoyin Wang[†]

I am an excellent linguist. The task is to label location entities in the given sentence. Below are some examples

Task Description

Input: Only France and Britain backed Fischler's proposal. **Example 1**

Output: Only @@France## and @@Britain## backed Fischler's proposal.

Input: Germany imported 47,600 sheep from Britain last year, nearly half of total imports. **Example 2**

Output: @@Germany## imported 47,600 sheep from @@Britain## last year, nearly half of total imports.

Few-shot Demonstrations

Input: It brought in 4275 tonnes of British mutton. some 10 percent of overall imports. **Example 3**

Output: It brought in 4275 tonnes of British mutton. some 10 percent of overall imports.

Input: China says Taiwan spoils atmosphere for talks.

Output: @@China## says @@Taiwan## spoils atmosphere for talks.

Input Sentence

Figure 1: The example of the prompt of GPT-NER. Suppose that we need to recognize location entities for the given sentence: *China says Taiwan spoils atmosphere for talks*. The prompt consists of three parts: (1) **Task Description**: It's surrounded by a red rectangle, and instructs the GPT-3 model that the current task is to recognize **Location** entities using linguistic knowledge. (2) **Few-shot Demonstrations**: It's surrounded by a yellow rectangle giving the GPT-3 model few-shot examples for reference. (3) **Input Sentence**: It's surrounded by a blue rectangle indicating the input sentence, and the output of the GPT-3 model is colored green.

English CoNLL2003 (FULL)			
Model	Precision	Recall	F1
<i>Baselines (Supervised Model)</i>			
BERT-Tagger (Devlin et al., 2018)	-	-	92.8
BERT-MRC (Li et al., 2019a)	92.33	94.61	93.04
GNN-SL (Wang et al., 2022)	93.02	93.40	93.2
ACE+document-context (Wang et al., 2020)	-	-	94.6 (SOTA)
<i>GPT-NER</i>			
GPT-3 + random retrieval	77.04	68.69	72.62
GPT-3 + sentence-level embedding	81.04	88.00	84.36
GPT-3 + entity-level embedding	88.54	91.4	89.97
<i>Self-verification (zero-shot)</i>			
+ GPT-3 + random retrieval	77.13	69.23	73.18
+ GPT-3 + sentence-level embedding	83.31	88.11	85.71
+ GPT-3 + entity-level embedding	89.47	91.77	90.62
<i>Self-verification (few-shot)</i>			
+ GPT-3 + random retrieval	77.50	69.38	73.44
+ GPT-3 + sentence-level embedding	83.73	88.07	85.9
+ GPT-3 + entity-level embedding	89.76	92.06	90.91

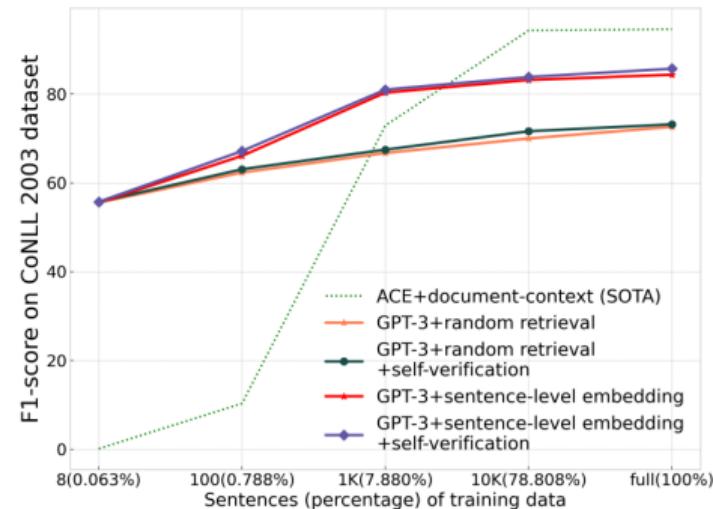
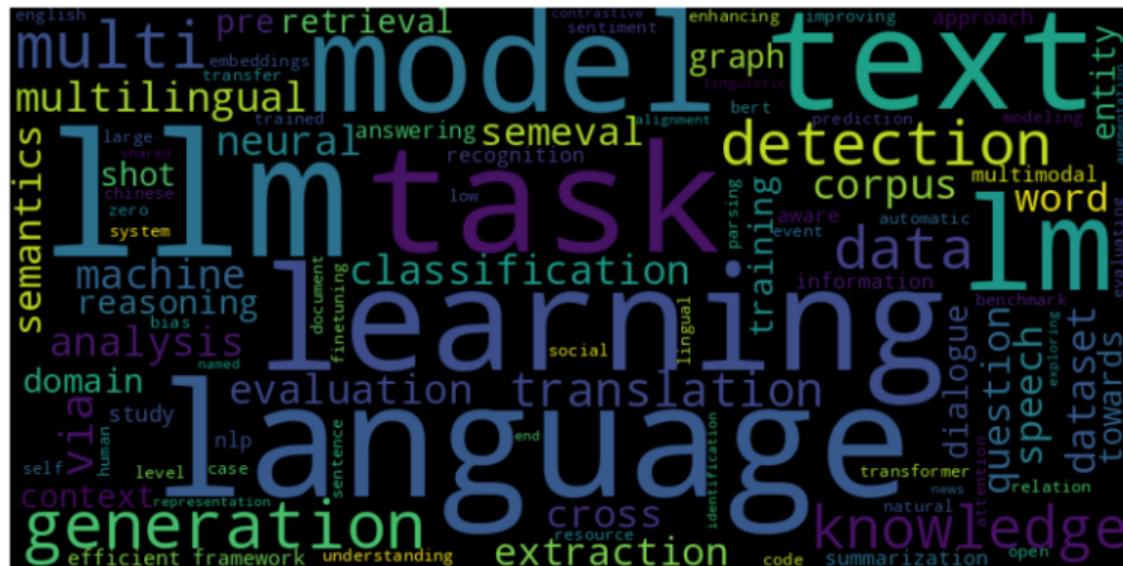
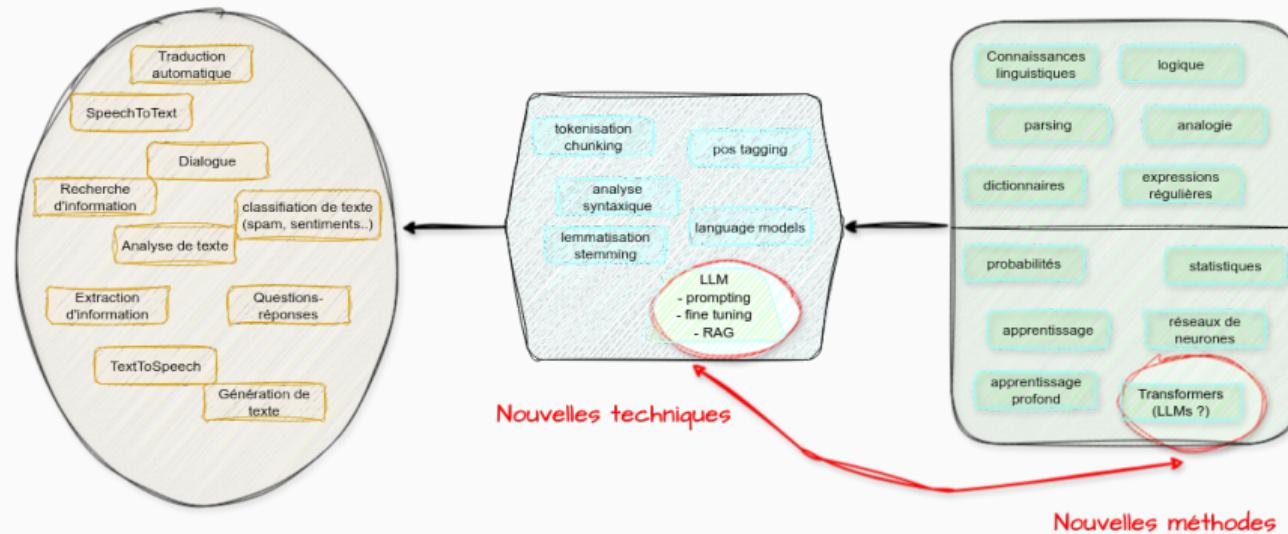


Figure 3: Low-resource comparisons on CoNLL2003 dataset.



Le TAL vise à proposer et développer des méthodes pour permettre aux ordinateurs de comprendre, analyser, manipuler et générer la langue de manière utile



Différentes familles de modèles pour différentes familles de tâches

- Encoders
 - BERT et famille BERT
 - entraînés à prédire un mot étant donné le contexte autour
 - habituellement fine-tunés
 - utilisation avec ajout de connaissance (RAG par ex)
 - tâche de classification, d'extraction d'information
- Decoders
 - GPT, Claude, Llama...
 - entraînés à prédire les mots de gauche à droite
 - tâches de génération, résumé, chatbot, ...
 -
- Encoder-decoders
 - Whisper, Flan-T5
 - entraînés à mapper une séquence sur une autre

Le TAL à l'ère des LLMs

Impact des LLM

- Modèles génératifs et à usage général
- Cassent la linguistique structurale
- On les "programme" en langue nat
- On passe de modèles spécifiques qu'on tente de généraliser → modèles généraux (*General Purpose AI, GPAI*) qu'on tente de spécifier
 - adaptation : finetuning total (huge cost !), partiel (vive les PEFT), instruct tuning
 - autre : prompting

- Entraînés sur des données pas toujours propres ou correctes
 - "privacy"
 - droits d'auteurs
 - consentement
- Fonctionnement des LLM
 - risque d'hallucinations
 - explicabilité
 - coût carbone (Anne-Laure Ligozat, vendredi matin)

Plan de la présentation

1 Définition(s)

2 La langue et le TAL

3 TAL : tâches et méthodes

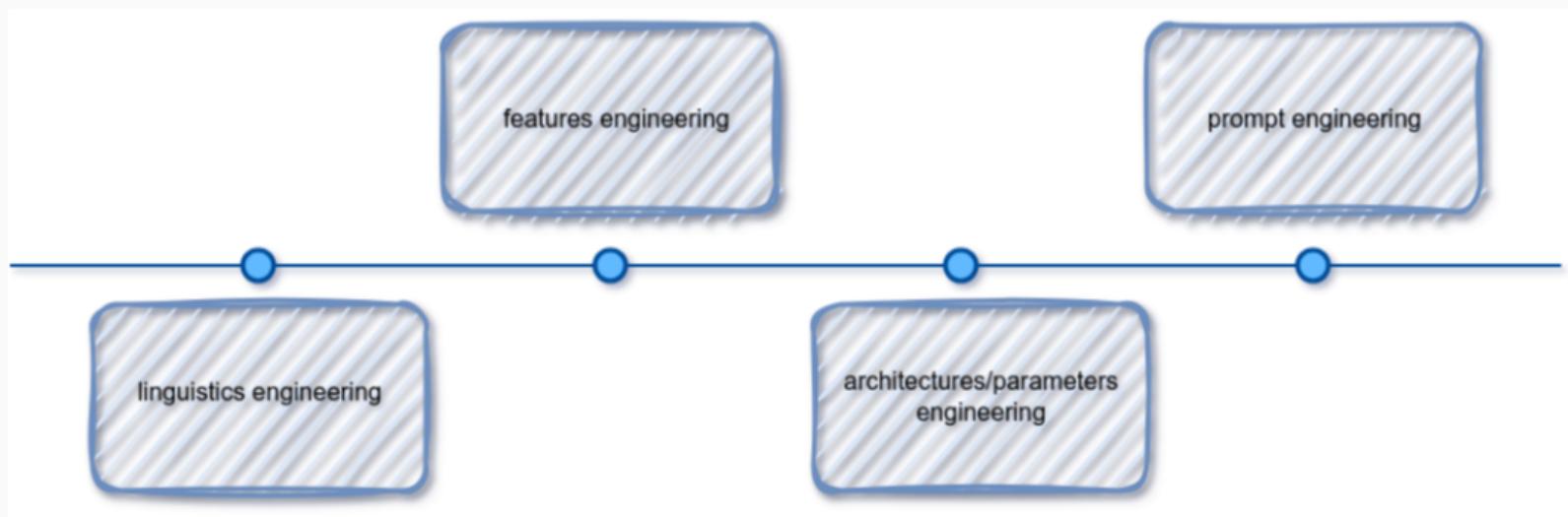
4 Le TAL à l'ère des LLMs

5 Conclusion

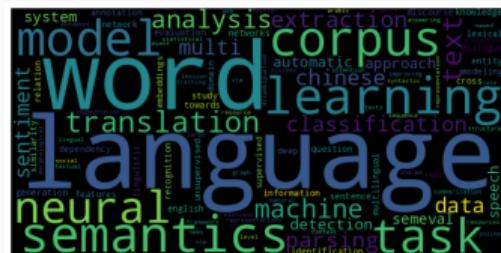
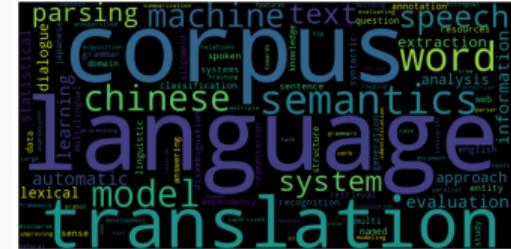
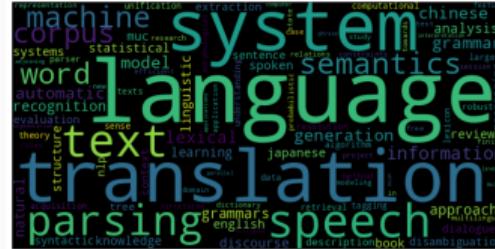
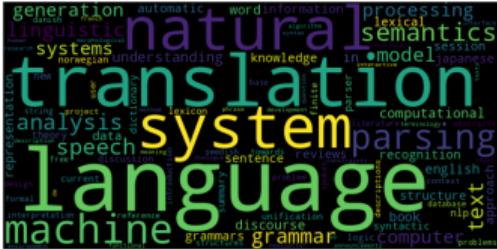
6 EN et évaluation

7 Références et sources variées

Evolution du TAL



Evolution du TAL



Plan de la présentation

1 Définition(s)

2 La langue et le TAL

3 TAL : tâches et méthodes

4 Le TAL à l'ère des LLMs

5 Conclusion

6 EN et évaluation

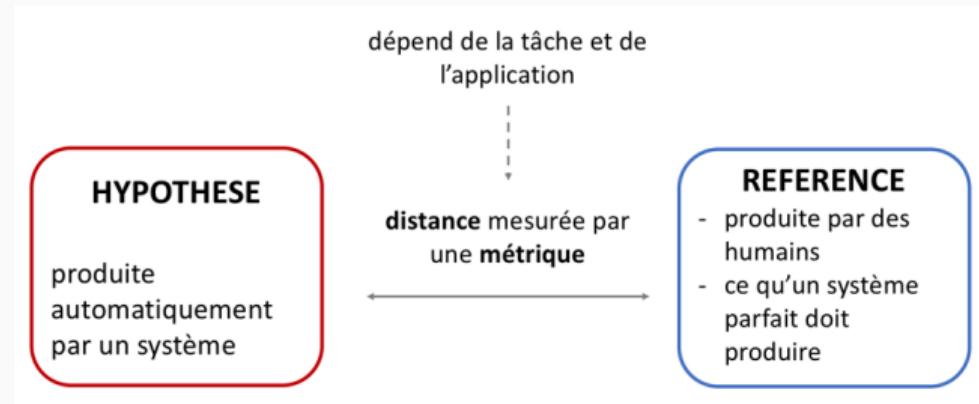
■ Introduction

■ Les mesures classiques

EN et évaluation

- Première formalisation de la procédure d'évaluation : MUC3 ?
- Motivation : avoir des éléments de comparaison stables et effectifs entre hypothèses et références

Protocole d'évaluation



Objectif : mesurer à quel point le système trouve les “bonnes réponses”

Quelle “bonnes réponse” ?

- traduction ou le résumé automatique : bonne réponses multiples
- EN : on peut supposer une seule et unique “bonne réponse”

Avantages

- **Transparence** : “règles du jeu” connues par tous
- **Coût** : réduit par rapport à une évaluation manuelle pour chaque hypothèse des systèmes ;
- **Reproductibilité** : réutilisation au delà des campagnes permettant une comparaison des résultats dans la production scientifique

1. Une **métrique** mesurant la distance entre une référence et une hypothèse ;
2. Un **algorithme d'alignement** de la référence et de l'hypothèse.
3. Un **algorithme de projection** des entités annotées sur la transcription manuelle de référence vers la transcription automatique

Précision

Ratio entre le nombre de **réponses correctes** et toutes les **réponses données** par un système

$$P = \frac{C}{C + S + I} \quad (1)$$

- C : nombre d'objets **corrects** dans l'hypothèse ;
- I : nombre d'**insertions** par le système ;
- S : nombre de **substitutions** par le système (entités mal typées).
- soit $C + S + I$: nombre total d'objets dans l'hypothèse.

Rappel

Ratio entre le nombre de **réponses correctes** et le nombre des **réponses attendues** (i.e. présentes dans la référence)

$$R = \frac{C}{C + S + D} \quad (2)$$

- D : nombre total d'**omission** (*deletions*) opérées par le système (entités non détectées) ;
- $C + S + D$: nombre total d'objets dans la référence.

Exemple 1

REF: <pers> Bertrand Delanoë </pers> a été élu maire de <loc> Paris </loc>

HYP1 : <pers> Bertrand Delanoë </pers> a été élu <pers> maire </pers> de <loc> Paris </loc>

- Precision = $\frac{2}{3} = 0,67$
- Rappel = $\frac{2}{2} = 1$

→ ici HYP1 produit du **bruit**

Exemple 2

REF: <pers> Bertrand Delanoë </pers> a été élu maire de <loc> Paris </loc>

HYP2: <pers> Bertrand Delanoë </pers> a été élu maire de Paris

- Precision = 1
- Rappel = $\frac{1}{2} = 0.5$

→ HYP2 produit du **silence**

- La **précision** tient compte des **insertions** et **substitutions**
- Le **rappel** tient compte des **omissions**

Comment combiner les 2 en une seule mesure ?

F-mesure, définie comme la **moyenne harmonique entre Précision et Rappel** :

$$F = (1 + \beta^2) \times \frac{P \times R}{\beta^2 P + R} \quad (3)$$

Où β est un **poids** permettant d'ajuster l'importance de P ou R
(si 1, égale importance).

Exemples

REF: <pers> Bertrand Delanoë </pers> a été élu maire de <loc> Paris </loc>

HYP1 : <pers> Bertrand Delanoë </pers> a été élu <pers> maire </pers> de <loc> Paris </loc>

HYP2: <pers> Bertrand Delanoë </pers> a été élu maire de Paris

$$F(HYP1) = (1 + 1^2) \times \frac{0,67 \times 1}{1^2 \times 0,67 + 1} = 0,80 \quad (4)$$

$$F(HYP2) = (1 + 1^2) \times \frac{1 \times 0,5}{1^2 \times 1 + 0,5} = 0,67 \quad (5)$$

Inconvénients des mesures classiques

- Fusionner P et R minimise le poids des erreurs d'insertion et d'omission par rapport aux erreurs de substitution, quel que soit β ?
- Avec les typologies fines et complexes, besoin d'une métrique différenciant les erreurs.

Différents types d'erreur

REF: the <pers.ind> president of Ford </pers.ind>

HYP1 : the <pers.ind> president </pers.ind> of Ford

→ erreur de frontière

HYP2 : the <pers.coll> president of Ford </pers.coll>

→ erreur de sous-type

HYP3 : the <pers.coll> president </pers.coll> of Ford

→ erreur de sous-type et de frontière.

ERR, Error Per Response

- définie lors de MUC5 ?
- inspirée du taux d'erreurs mots (WER pour *Word Error Rate*) en RAP ?
- mesure des erreurs : plus le taux est bas, mieux c'est.

$$ERR = \frac{S + D + I}{C + S + D + I} \quad (6)$$

REF: <pers> Bertrand Delanoë </pers> a été élu maire de <loc> Paris </loc>

HYP1: <pers> Bertrand Delanoë </pers> a été élu <pers> maire </pers> de <loc> Paris </loc>

HYP2: <pers> Bertrand Delanoë </pers> a été élu maire de Paris

$$ERR(HYP1) = \frac{0 + 0 + 1}{2 + 0 + 0 + 1} = \frac{1}{3}$$

$$ERR(HYP2) = \frac{0 + 1 + 0}{1 + 0 + 1 + 0} = \frac{1}{3}$$

Le poids des insertions est moins important que celui des substitutions et des omissions?.

Une augmentation de I provoque une augmentation de ERR moins importante qu'une augmentation de $S + D$.

$$ERR = \frac{S + D + I}{N + I} \quad (7)$$

Avec N = nombre d'entités dans la référence.

Pour $N = 100$, $S + D = 10$, $I = 10$, on a :

$$ERR = \frac{10 + 10}{100 + 10} = \frac{20}{110}$$

Si on augmente $S + D$ de 10 :

$$ERR = \frac{20 + 10}{100 + 10} = \frac{30}{110} = 0,27$$

Si on augmente I de 10 :

$$ERR = \frac{10 + 20}{100 + 20} = \frac{30}{120} = 0,25$$

De plus, avoir I dans le dénominateur rend les résultats non comparables.

SER : Slor Error Rate

- proposée par ?
- identique au WER utilisé en RAP
- utilisée lors de ACE, ESTER-2, QUAERO et ETAPE
- suppression du nombre d'insertion (I) du dénominateur :

$$SER = \frac{S + D + I}{C + D + S} = \frac{S + D + I}{R} \quad (8)$$

où R = nombre total d'entités de la référence.

Possibilité d'affiner l'importance relative des erreurs :

$$SER = \frac{\alpha_1 S_t + \alpha_2 S_f + \beta D + \gamma I}{R} \quad (9)$$

- S_t et S_f : nombre total de substitution de type et de frontières ;
- D et I : nombre total d'omissions et insertions ;
- α_1 α_2 β et γ : poids affectées à chaque catégorie d'erreur.

- représentation en “slot” des hypothèses et de la référence
 - slot= un segment de texte avec des frontières (début/fin) et un type
- structure plate qui ne peut pas traiter les entités imbriquées

ETER : Entity Tree Error Rate

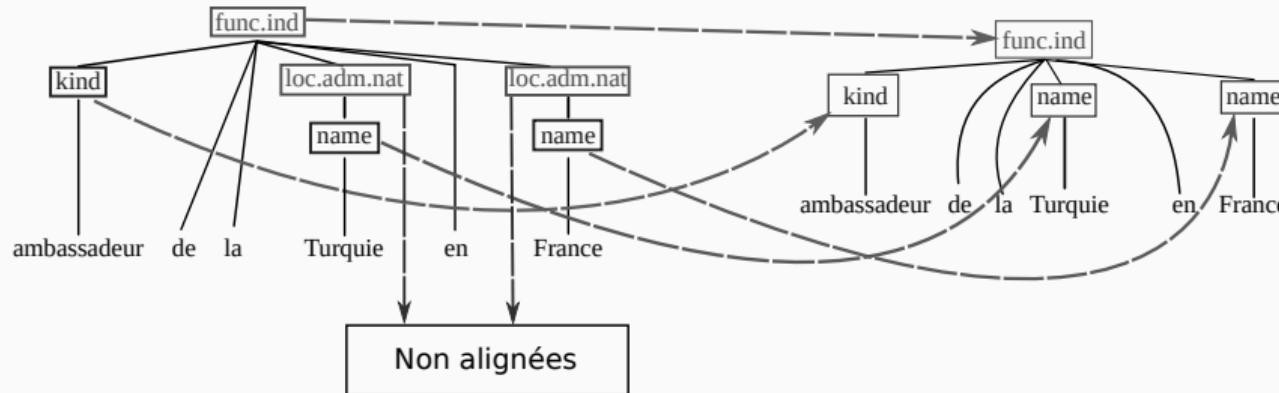
Basée sur une comparaison des arbres d'entités ?

$$ETER = \frac{I + D + \sum_{(e_r, e_h)} E(e_r, e_h)}{N_E} \quad (10)$$

- I : nombre total d'insertions d'arbre-entité ;
- D : nombre total d'omission d'arbre-entité ;
- (e_r, e_h) : paires d'entités-arbres référence/hypothèse associées à l'issue de l'alignement ;
- $E(r, h)$: erreur calculée pour chaque paire d'entité-arbre (e_r, e_h) (peut être zéro) ;
- N_E : nombre d'entité-arbre dans la référence.

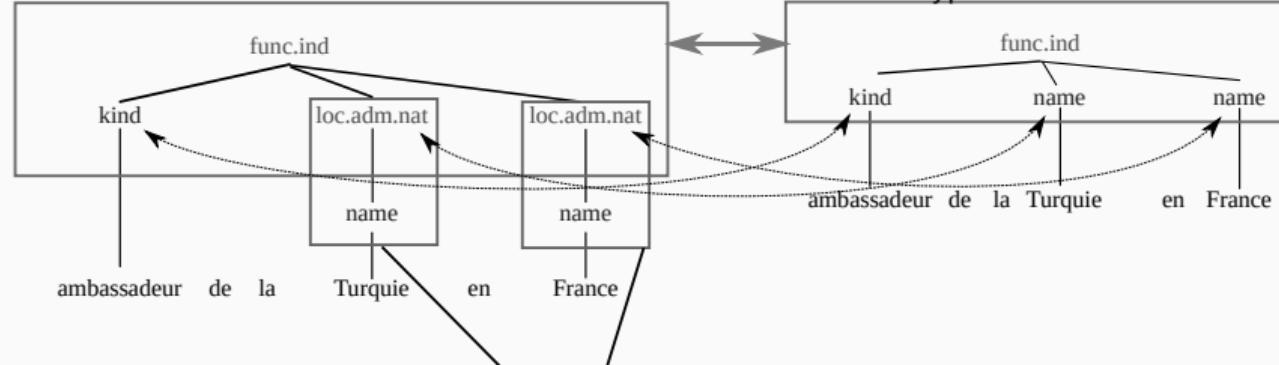
Référence

Hypothèse



Référence

Hypothèse



$$ETER = \frac{I + D + \sum_{(e_r, e_h)} E(e_r, e_h)}{N_E} \quad (11)$$

Le calcul d'erreur pour les paires d'entité-arbre $E(r, h)$ a 2 parties

- erreur de détection et de classification de l'entité
- erreur de décomposition E_c

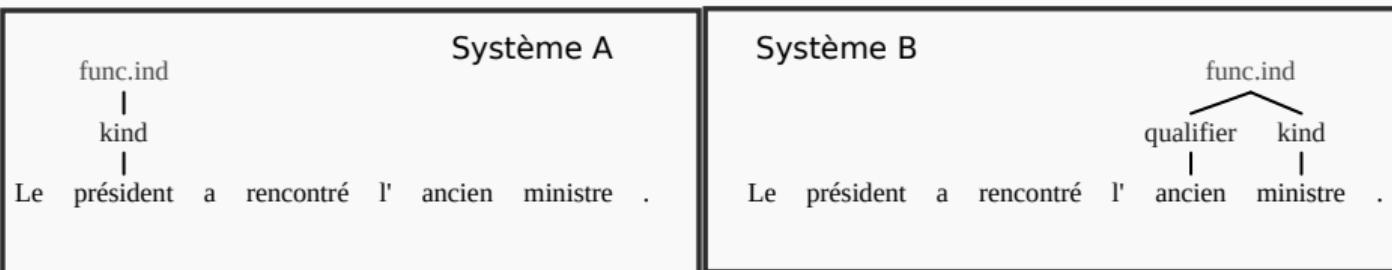
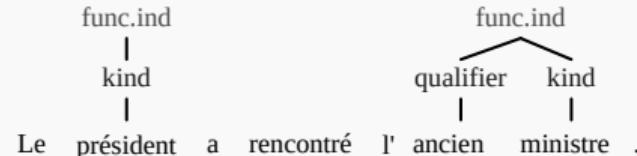
$$E(r, h) = (1 - \alpha)E_T(e_r, e_h) + \alpha E_c(e_r, e_h), \alpha \in [0..1] \quad (12)$$

- $E_T(e_r, e_h)$: erreur de classification, dépend de la distance entre (e_r, e_h) ;
- $E_c(e_r, e_h)$: erreur de décomposition, dépend de la distance entre les constituants des entités-arbres (e_r, e_h) ;
- α paramètre fixant le poids relatif de la décomposition par rapport à la classification.

→ $E_c(e_r, e_h)$ se rapproche d'un SER local

ETER : exemple

Référence



- **Système A** : 3 omissions, 0 insertion, 0 substitution, 2 slots corrects
→ $3/5$ soit $SER = 60\%$
- **Système B** : 2 omissions, 0 insertion, 0 substitution, 3 slots corrects.
→ $2/5$ soit $SER = 40\%$

Or ces deux systèmes ont omis une entité chacun et devrait avoir un score équivalent. Avec ETER,

- Assez peu de corpus et de campagnes d'évaluation
 - en France : ESTER 1 et 2, ETAPE (+ QUAERO), REPERE (pour les personnes, multimodal)
 - à l'international : campagne ACE (2000-2008)
- Difficile de comparer REN sur textes et REN sur parole car on ne dispose pas de corpus et campagnes comparables (types de données + typologies)
- Résultats nettement différents entre transcriptions manuelles et transcriptions automatiques

Pour comparer simplement, utilisée par ? :

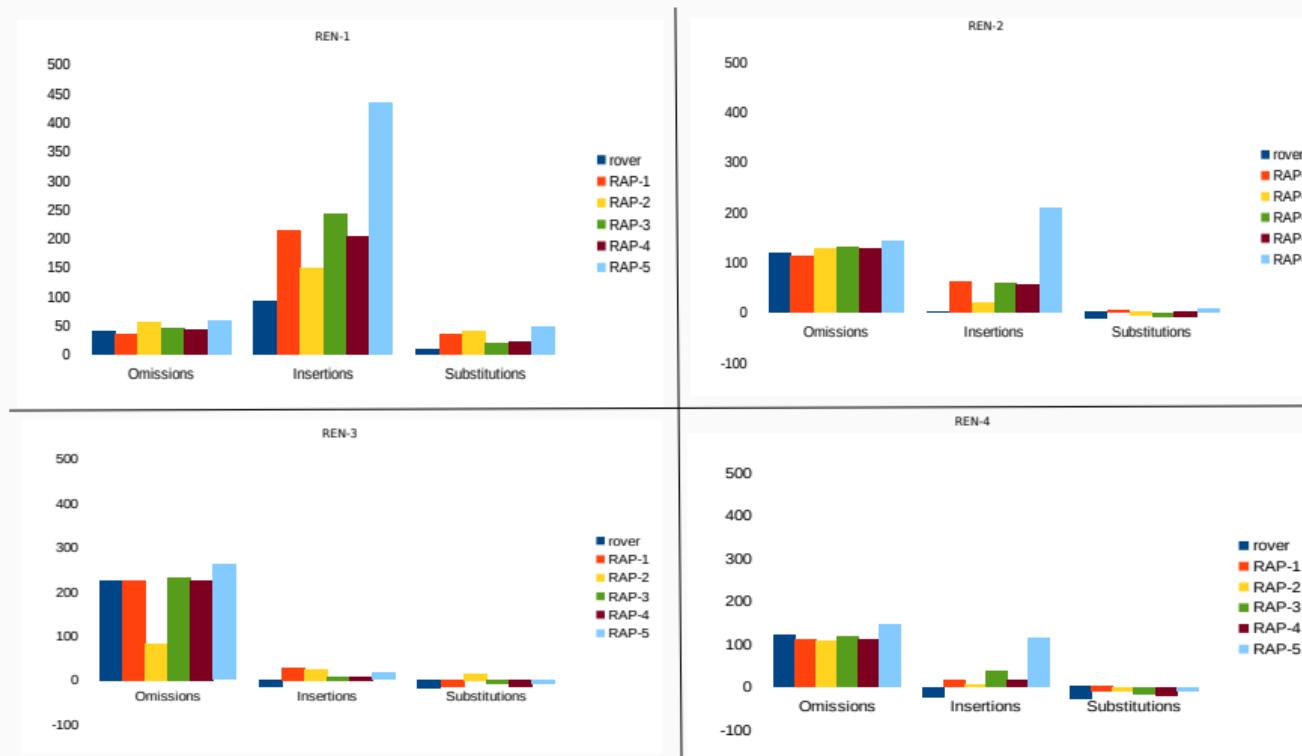
$$PAE(e) = 100 * \frac{NB_A(e) - NB_M(e)}{NB_M(e)} \quad (13)$$

Avec :

- e une erreur de REN de type omission, insertion ou substitutions ;
- NB_A le nombre des erreurs de type e sur les transcriptions automatiques ;
- NB_M le nombre des erreurs de type e sur les transcriptions manuelles.

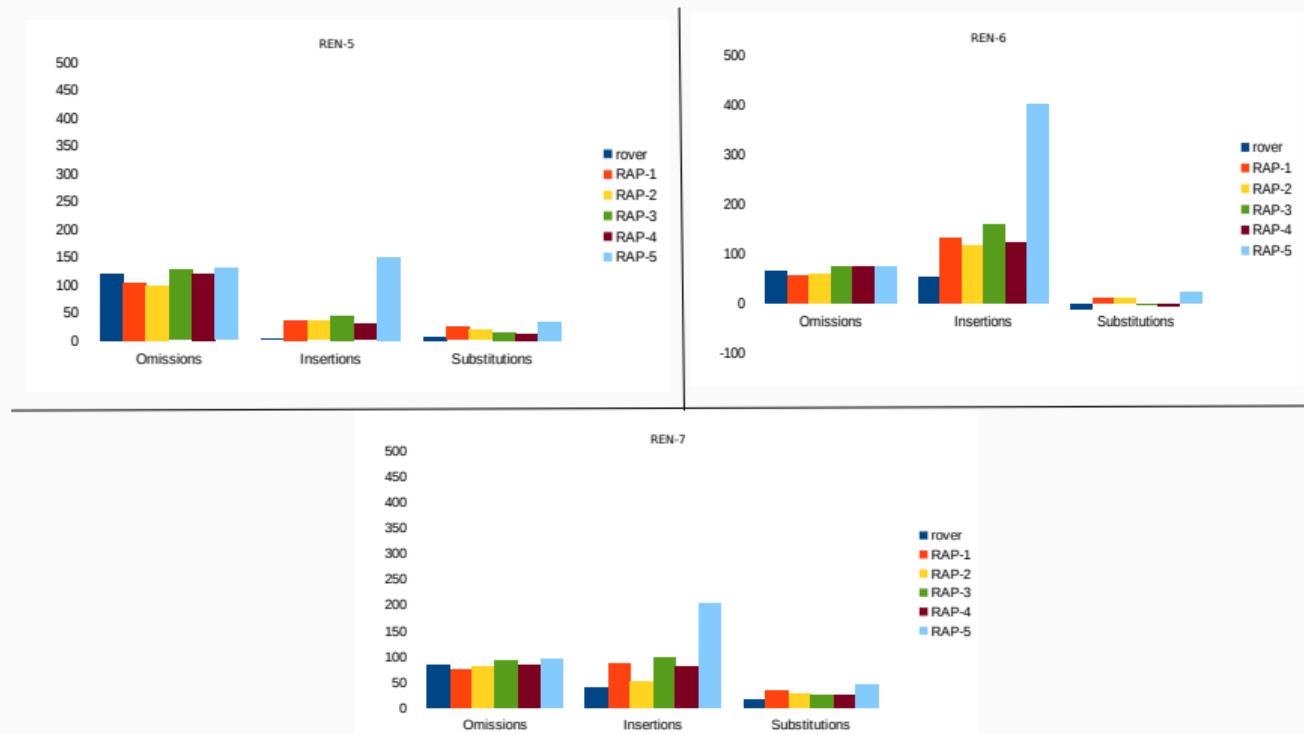
REN sur transcriptions automatiques et manuelles

PAE ETAPE-1



REN sur transcriptions automatiques et manuelles

PAE ETAPE-2



Quelques constats

1. sur ETAPE plus un système ASR insère de mots, plus un système REN insère d'entités (pas observées sur les données QUAERO)
2. impact fort des erreurs (notamment omissions et insertions) sur la mention de l'entité
3. impact non nul des erreurs des mots qui introduisent une EN

Point 2 et 3 en lien direct avec la façon dont les systèmes REN sont développés.

Plan de la présentation

1 Définition(s)

2 La langue et le TAL

3 TAL : tâches et méthodes

4 Le TAL à l'ère des LLMs

5 Conclusion

6 EN et évaluation

7 Références et sources variées

Liens divers : informations diverses

- Cours de Xavier Tannier
- Bling : Blog de linguistique illustré
- la collection Pangloss, archive ouverte de langues en danger ou peu documentées
- Histoire de l'ATALA
- L'ordinateur qui parle - Joseph Mariani
- The shoebox à IBM
- CMU Advance Natural Language Processing, Spring 2025
- Natural Language Processing with Deep Learning - Stanford / Spring 2024
- Wikipedia, histoire de Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz
- Conférence invitée de H. Ney, ETAL 2023
- Lectures in natural language processing, Stuttgart
- Introduction aux IAG, Vincent Guige, 2024

- Can language models synthesize scientific literature?, Université de Washington
- OpenHands: An Open Platform for AI Software Developers as Generalist Agents[Wang et al., 2025b]

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. URL <https://arxiv.org/abs/1409.0473>.
- Jason P.C. Chiu and Eric Nichols. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4 :357–370, 2016. doi : 10.1162/tacl_a_00104. URL <https://aclanthology.org/Q16-1026/>.
- K. Davis, R. Biddulph, and S. Balashek. Automatic recognition of spoken digits. *Journal of the Acoustical Society of America*, 24 :637–642, 1952. URL <https://rauterberg.employee.id.tue.nl/presentations/bell-labs.pdf>.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile : An 800gb dataset of diverse text for language modeling, 2020. URL <https://arxiv.org/abs/2101.00027>.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In Kalina Bontcheva and Jingbo Zhu, editors, *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi : 10.3115/v1/P14-5010. URL <https://aclanthology.org/P14-5010/>.

Warren S. McCulloch and Walter Pitts. (1943) warren s. mcculloch and walter pitts, "a logical calculus of the ideas immanent in nervous activity," bulletin of mathematical biophysics 5 : 115-133. In *Neurocomputing, Volume 1 : Foundations of Research*. The MIT Press, 04 1988. ISBN 9780262267137. doi : 10.7551/mitpress/4943.003.0004. URL <https://doi.org/10.7551/mitpress/4943.003.0004>.

- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL <https://arxiv.org/abs/2203.02155>.
- Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum. Entités nommées structurées : guide d'annotation quaero. Technical report, 2011. autres.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fdb053c1c4a845aa-Paper.pdf.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. GPT-NER : Named entity recognition via large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics : NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi : 10.18653/v1/2025.findings-naacl.239. URL

<https://aclanthology.org/2025.findings-naacl.239/>.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F. Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, Hoang H. Tran, Fuqiang Li, Ren Ma, Mingzhang Zheng, Bill Qian, Yanjun Shao, Niklas Muennighoff, Yizhe Zhang, Binyuan Hui, Junyang Lin, Robert Brennan, Hao Peng, Heng Ji, and Graham Neubig. Openhands : An open platform for AI software developers as generalist agents. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL <https://openreview.net/forum?id=0Jd3ayDDoF>.

Warren Wiever. Translation, 1949. URL
<https://aclanthology.org/1952.earlymt-1.1.pdf>.