Observatoire
des humanités
numériques

Le TAL en pratique

Jean Barré, Thierry Poibeau

15 avril 2023

ENS-PSL

Table des matières

Le TAL : Toute une histoire

Le projet fou de Roberto Busa

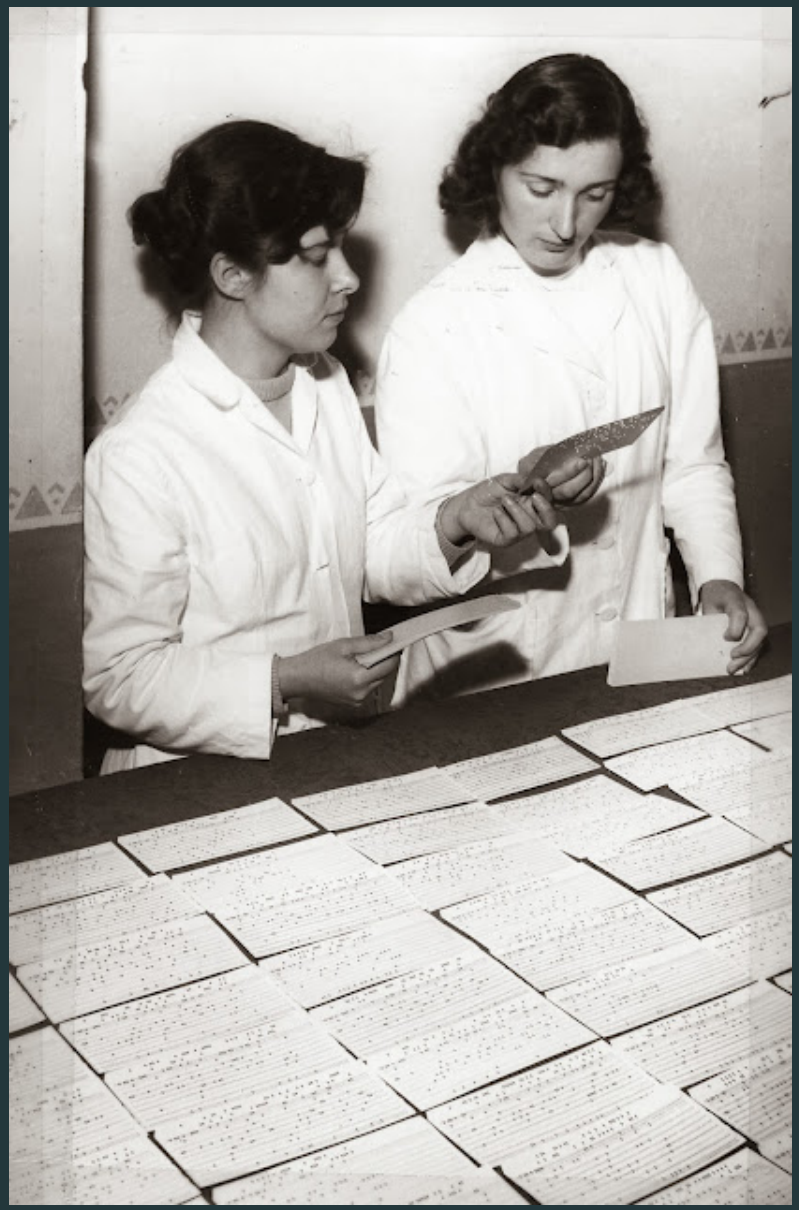
- Roberto Busa (1913-2011), prêtre jésuite italien spécialiste de Thomas d’Aquin
- L’objectif initial de Busa est philosophique et théologique
- Mais sa thèse est qu’on ne peut accéder à la pensée d’un·e auteurice que si on maîtrise sa façon d’employer le langage.
- Il lui faut donc étudier Thomas d’un point de vue philologique et linguistique.
- Il formule donc en 1946 le projet d’un grand concordancier des œuvres de Thomas
- https ://www.corpusthomisticum.org/it/index.age

Un premier traitement automatique de texte ?

- Réalisation immédiate : la tâche est trop vaste pour être faite sans assistance
- Il se met donc en quête de « machinerie » pour l'aider : « any gadget that might help »(Busa, 1980)
- Projet mené à partir de 1946 pendant 34 ans (+ de 30 personnes impliquées !).
- aide d'IBM pour le réaliser
- Transcription de 179 textes en forme lisible par des machines de l'époque (des cartes perforées !).
- Indexation de 10.632.980 mots, 1500 km de câble, 10.000h de calcul, 1.000.000 d'heures de travail humain.

L'index Thomisticus

Source : http ://melissaterras.blogspot.com/2013/10/for-ada-lovelace-day-father-busas.html



Introduction de l’informatique

- Comme outil pour embrasser une masse de données.
- Conduire l’analyse à un niveau jusque là inaccessible.
- Projet précurseur entre TAL et Humanités (Distant reading ?)

Le TAL aujourd'hui

Les tâches du TAL

- Récupération d'informations linguistiques (syntaxe, schéma de dépendances)
- Traduction automatique
- Classification automatique de texte (spam/non spam)
- Résumé de texte - Récupération de thèmes spécifiques
- Questions / Réponses - Chatbots
- Génération de texte

Le TAL aujourd'hui

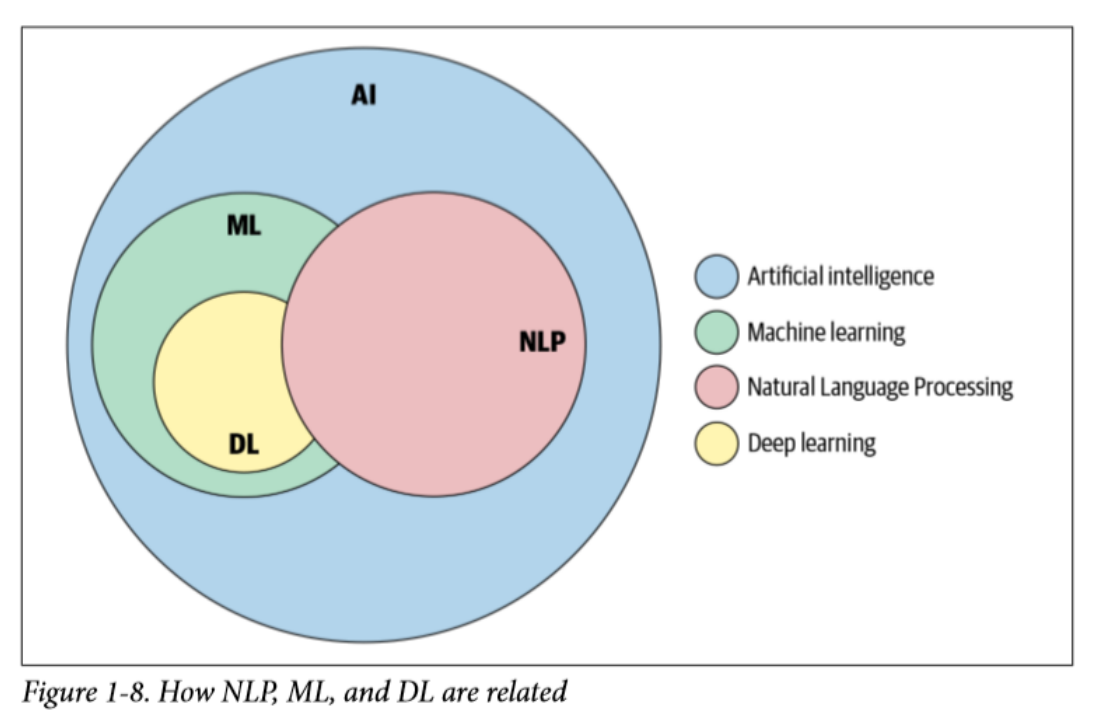


Figure 1-8. How NLP, ML, and DL are related

Figure 1 : Définition du champ. Source : [1]

La chaîne de traitement du TAL

La chaîne de traitement classique du TAL

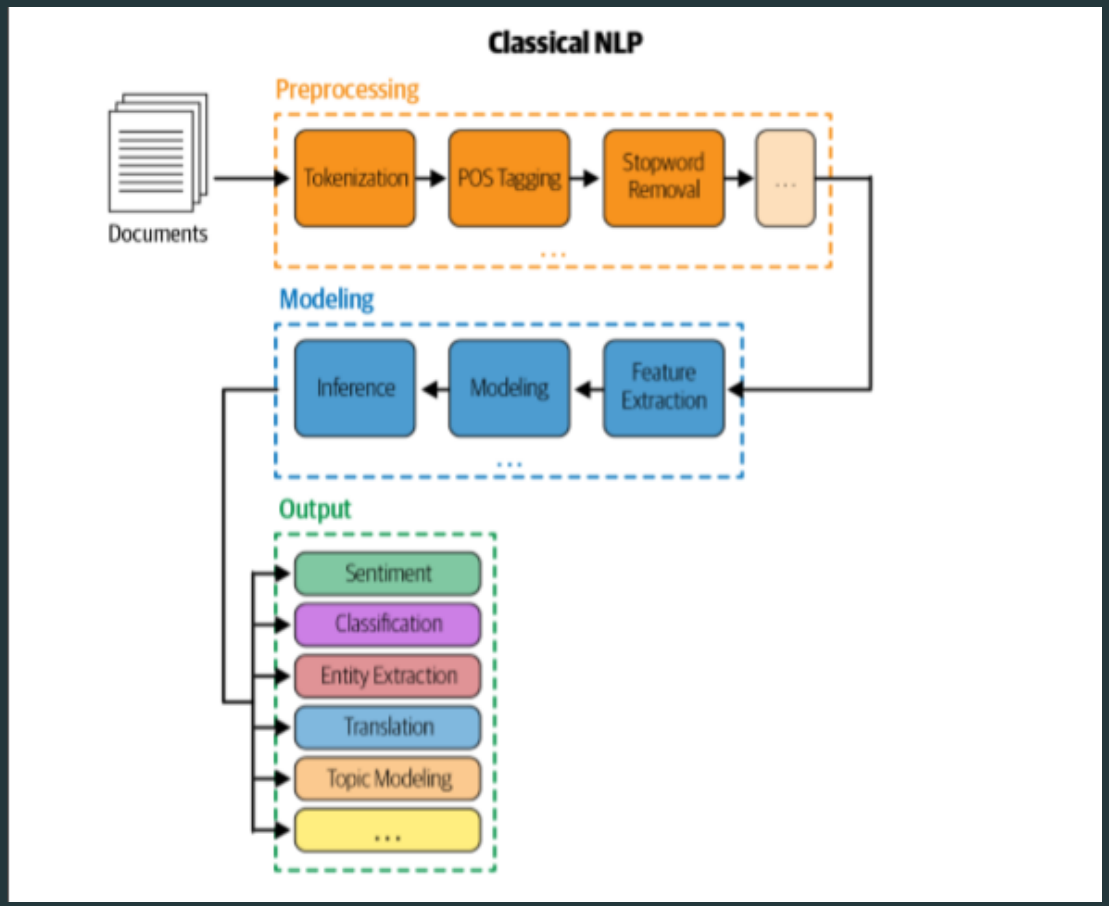


Figure 2 : Source : [1]

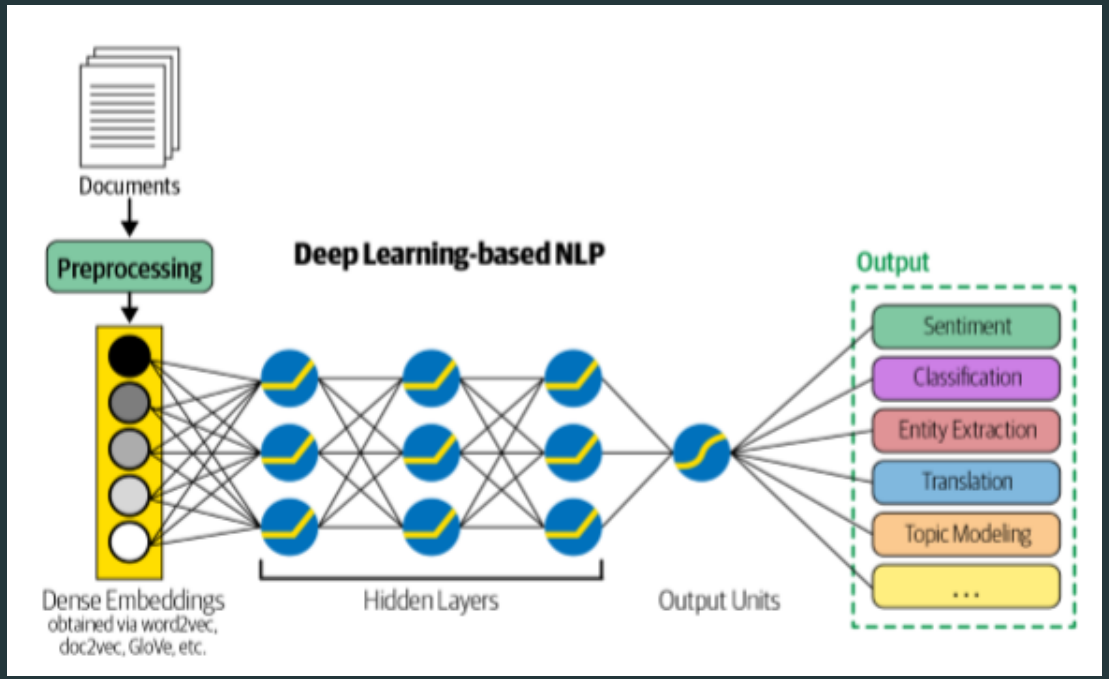


Figure 3 : Source : [1]

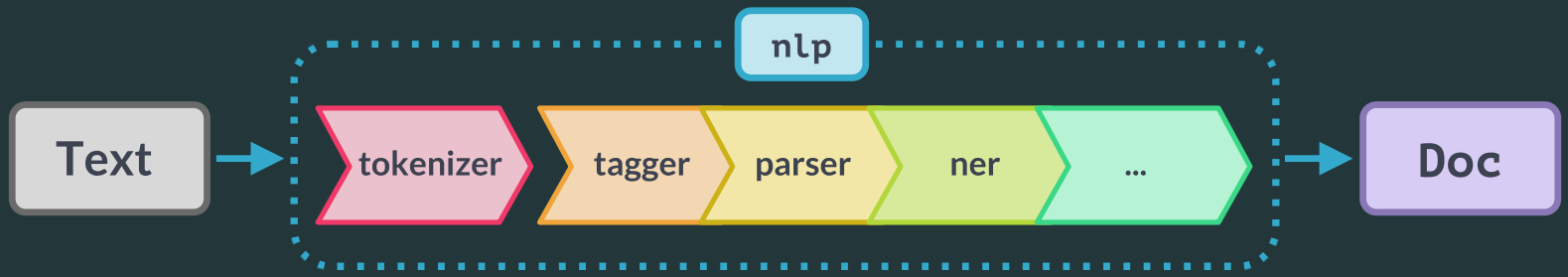


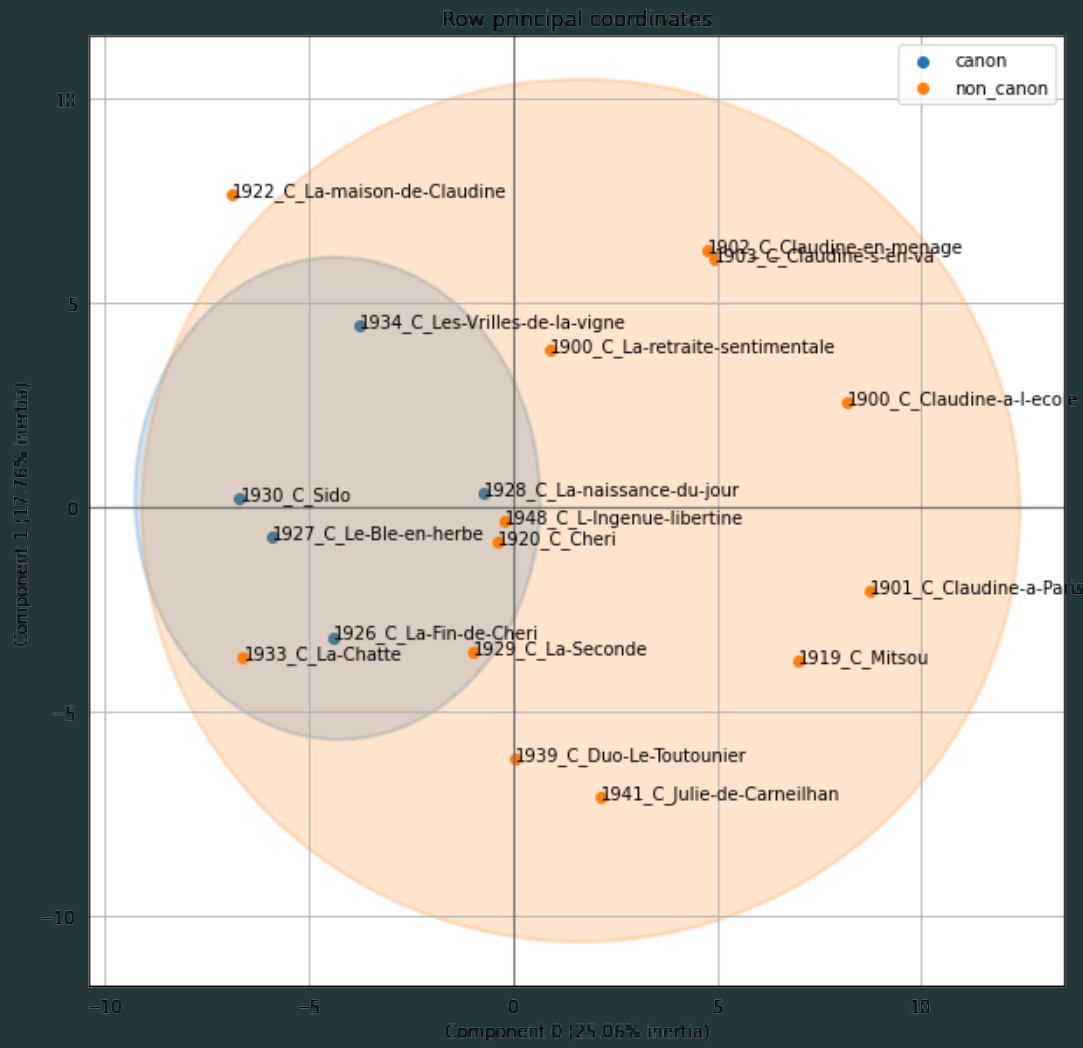
Figure 4 : Source : https ://spacy.io/usage/processing-pipelines/

Le TAL dans le contexte des Humanités Numériques - Exemple de la stylométrie

La stylométrie

Analyser des données textuelles

- dater, localiser des textes ;
- regrouper des textes selon des caractères stylistiques ;
- attribuer une pièce disputée entre plusieurs auteurs ;
- détecter des collaborations.



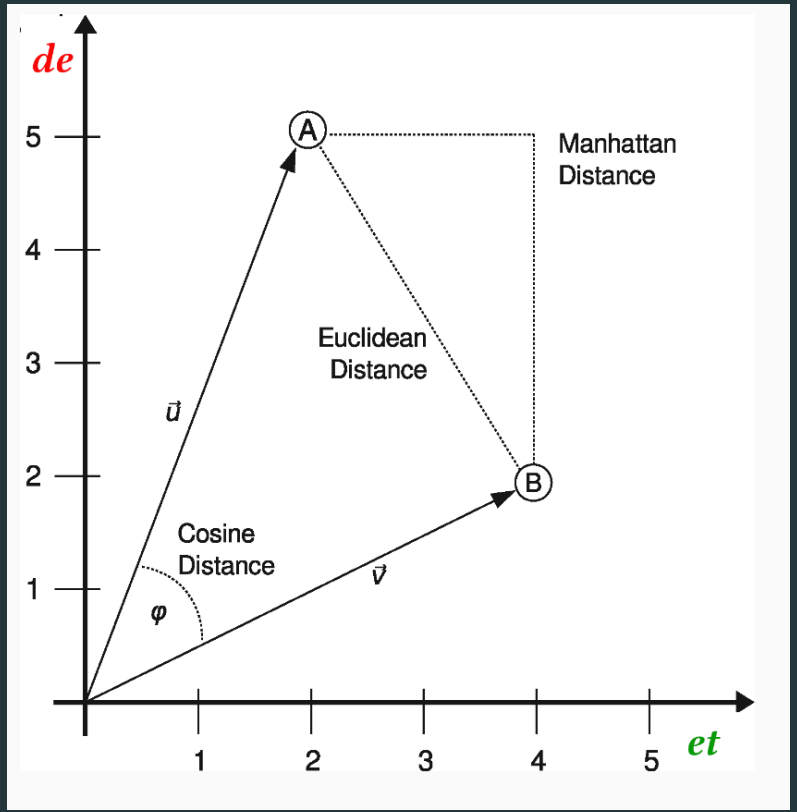
Mesure de distances entre des textes

Texte A

et de joie, il dit à son voisin de classe de gauche qu'il s'enthousiasmait de cette conférence et de cette journée.

Texte B

et de chaque côté de lui, il vit chiens et chats et poules et canards.



Mesurer le style ?

Wincenty Lutosławski - Principes de stylométrie (1890) [4]
Objectif : Classer chronologiquement les oeuvres de Platon

Postulat :
Chaque individu emploie une langue démontrant des propriétés particulières et mesurables, appelées stylome ou idiolecte.

Idiolecte :
Ensemble de traits linguistiques caractéristiques d'un individu

On peut détecter cela quantitativement - Importance des mots outils (mots les plus fréquents)
Propriétés inconscientes du style

Chaîne de traitement du TAL

- OCR / HTR
- Tokenisation
- Lemmatisation (lemme = partie canonique - entrée de dictionnaire)
- Étiquettage morpho-syntaxique (partie du discours - nom, verbe, pronom, substantif, ...)
- Sac de mots / Sac de séquence de mots - de POS

Molière a-t-il vraiment écrit ses pièces ? 1/2

Contexte & stylométrie

- Remise en cause de la paternité de certaines oeuvres de Molière.
- P. Corneille étant le vrai auteur ?
- Analyse du lexique, des lemmes, des rimes, de la morphosyntaxique sur un corpus de comédies.

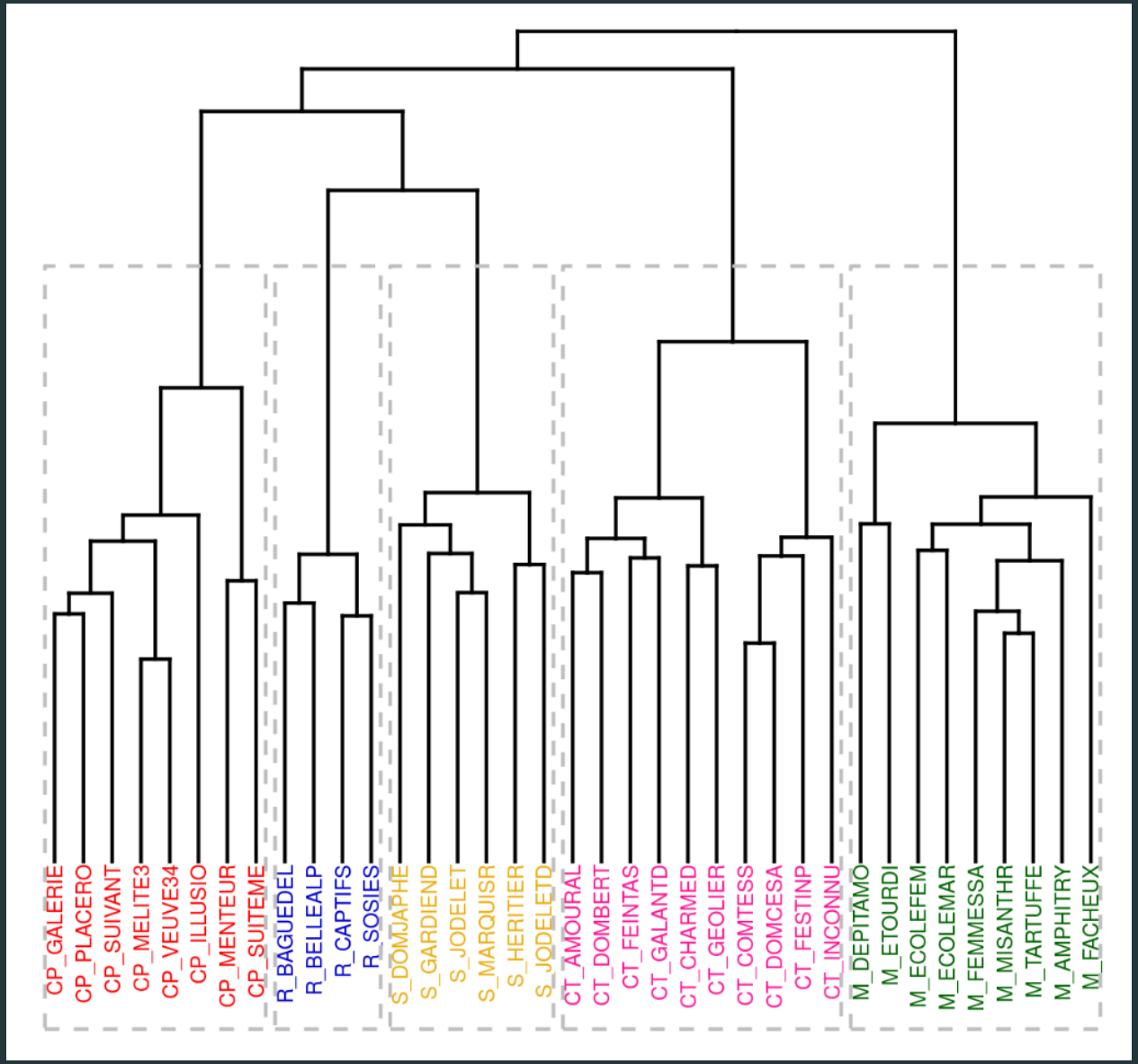


Figure 5 : Source : (Cafiero, 2019)[2]

Mesurer le signal d'une co-écriture

Psyché, Texte issue d'une collaboration :

- Molière a dressé le plan de la pièce. Il a rédigé le début de chaque acte
- Corneille s'est lui occupé de finir les actes
- Quinault - rédige les paroles du choeur (début et/ou fin des actes)
- Peut on détecter quantitativement le signal de cette collaboration ?

Mesurer le signal d'une co-écriture

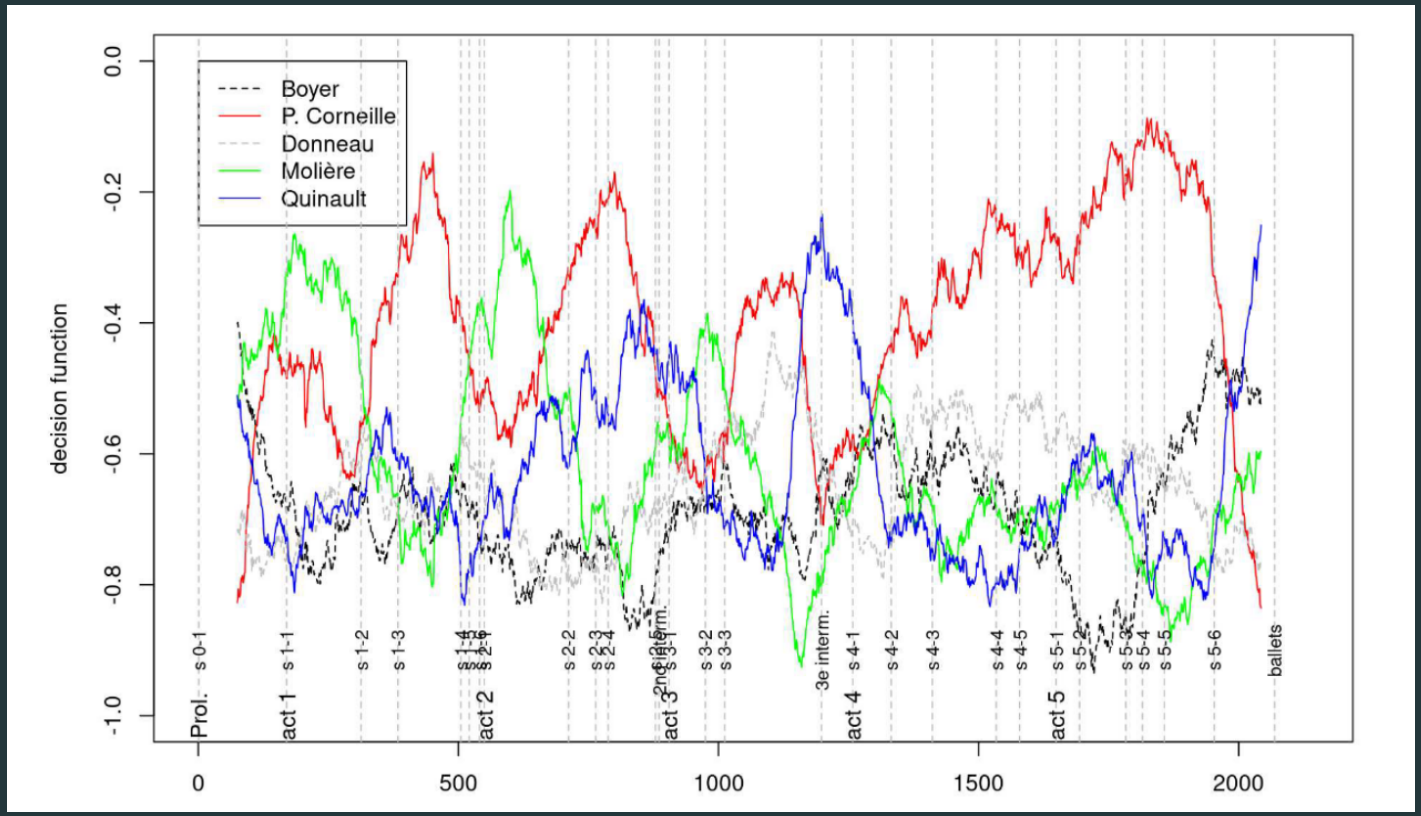


Figure 6 : Source : (Cafiero, 2021)[3]

Le TAL pour l'étude des textes en SHS ?

Gagner du temps et déléguer des tâches à l'ordinateur

- traiter des corpus très importants ou jusqu'à un niveau très fin, sans temps supplémentaire.
- réaliser des opérations répétitives difficilement envisageables à la main, en limitant le risque d'erreur humaine ;

Avoir une autre approche des mêmes données

- obtenir des réponses qu'on n'aurait pu obtenir par des moyens traditionnels.
- bénéficier de l'apport méthodologique d'autres champs scientifiques (biostatistiques, IA, TAL, etc.).

Ancrer son analyse dans les faits et leur mesure

- éviter un certain nombre d'écueils de l'analyse traditionnelle : surévaluation des phénomènes individuels, des individus aberrants, meilleure évaluation des tendances d'ensemble, etc.

Questions ?

Bibliographie indicative i

S. V. B, B. Majumder, A. Gupta, and H. Surana.
Practical natural language processing : a comprehensive guide to building real-world NLP systems.
O'Reilly Media, first edition edition.
OCLC : on1125266646.

F. Cafiero and J.-B. Camps.
Why molière most likely did write his plays.
5(11) :eaax5489.
Publisher : American Association for the Advancement of Science.

Bibliographie indicative ii

F. Cafiero and J.-B. Camps.
'psyché' as a rosetta stone ? assessing collaborative authorship in the french 17th century theatre.
In M. Ehrmann, F. Karsdorp, M. Wevers, T. L. Andrews, M. Burghardt, M. Kestemont, E. Manjavacas, M. Piotrowski, and J. van Zundert, editors, *Proceedings of the Conference on Computational Humanities Research, CHR2021, Amsterdam, The Netherlands, November 17-19, 2021*, volume 2989 of *CEUR Workshop Proceedings*, pages 377–391. CEUR-WS.org, 2021.

W. Lutoslawski.
Principes de stylométrie appliqués à la chronologie des œuvres de platon.
11(41) :61–81.