Introduction : Le numérique au service des Humanités . . . et vice-versa

Gaël Lejeune (gael.lejeune@sorbonne-universite.fr)

25 septembre 2025

UFR de Sociologie et d'Informatique pour les Sciences Humaines
STIH, Sorbonne Université
CERES (Expérimentations en Méthodes numériques pour les SHS)

Qui suis-je

Parcours universitaire

- L : Sciences du Langage (Caen)
- M : Traductologie (Caen, Corfou, Helsinki)

Qui suis-je

Parcours universitaire

- L : Sciences du Langage (Caen)
- M : Traductologie (Caen, Corfou, Helsinki)
- D : Informatique (Caen)

Qui suis-je

Parcours universitaire

- L : Sciences du Langage (Caen)
- M : Traductologie (Caen, Corfou, Helsinki)
- D : Informatique (Caen)
- puis Nantes, Paris XIII, Sorbonne Université

Qui suis-je

Parcours universitaire

- L : Sciences du Langage (Caen)
- M : Traductologie (Caen, Corfou, Helsinki)
- D : Informatique (Caen)
- puis Nantes, Paris XIII, Sorbonne Université

Spécialité : Traitement Automatique des Langues

Champs d'application

- Veille Epidémiologique et Extraction d'Info Multilingue

Qui suis-je

Parcours universitaire

- L : Sciences du Langage (Caen)
- M : Traductologie (Caen, Corfou, Helsinki)
- D : Informatique (Caen)
- puis Nantes, Paris XIII, Sorbonne Université

Spécialité : Traitement Automatique des Langues

Champs d'application

- Veille Epidémiologique et Extraction d'Info Multilingue
- Classification de Documents (dialectes, opinions ...)

Qui suis-je

Parcours universitaire

- L : Sciences du Langage (Caen)
- M : Traductologie (Caen, Corfou, Helsinki)
- D : Informatique (Caen)
- puis Nantes, Paris XIII, Sorbonne Université

Spécialité : Traitement Automatique des Langues

Champs d'application

- Veille Epidémiologique et Extraction d'Info Multilingue
- Classification de Documents (dialectes, opinions ...)
- Analyse Stylistique (auteurs, dates, genres ...)

Qui suis-je

Parcours universitaire

- L : Sciences du Langage (Caen)
- M : Traductologie (Caen, Corfou, Helsinki)
- D : Informatique (Caen)
- puis Nantes, Paris XIII, Sorbonne Université

Spécialité : Traitement Automatique des Langues

Champs d'application

- Veille Epidémiologique et Extraction d'Info Multilingue
- Classification de Documents (dialectes, opinions ...)
- Analyse Stylistique (auteurs, dates, genres ...)
- Analyse de Réseaux Soc. Numériques

Ce dont je vais parler ici

- Magie de l'outil automatique et pyrotechnie formelle

Ce dont je vais parler ici

- Magie de l'outil automatique et pyrotechnie formelle
- Les Humanités comme objet d'application du numérique

Ce dont je vais parler ici

- Magie de l'outil automatique et pyrotechnie formelle
- Les Humanités comme objet d'application du numérique
- . . . et comme objet de réflexion sur le numérique

Ce dont je vais parler ici

- Magie de l'outil automatique et pyrotechnie formelle
- Les Humanités comme objet d'application du numérique
- . . . et comme objet de réflexion sur le numérique
- Cadre : Centre d'Expérimentation des Méthodes Numériques en SHS

Domaines mobilisés

- Traitement Automatique des Langues (TAL)
- Vision par Ordinateur
- Fouille de Données

Organisation

9 cours de 2h :

- Du bla-bla . . . puis De la pratique

Organisation

9 cours de 2h :

- Du bla-bla . . . puis De la pratique
- De la réflexion

Organisation

9 cours de 2h :

- Du bla-bla . . . puis De la pratique
- De la réflexion et encore De la pratique

Outline

1. Acteurs des Humanités Numériques : Le CERES

2. Partie 1 : Traiter des données textuelles

Acteurs des Humanités Numériques : Le CERES

CERES

Centre d'Expérimentation pour les **Méthodes Numériques** en SHS

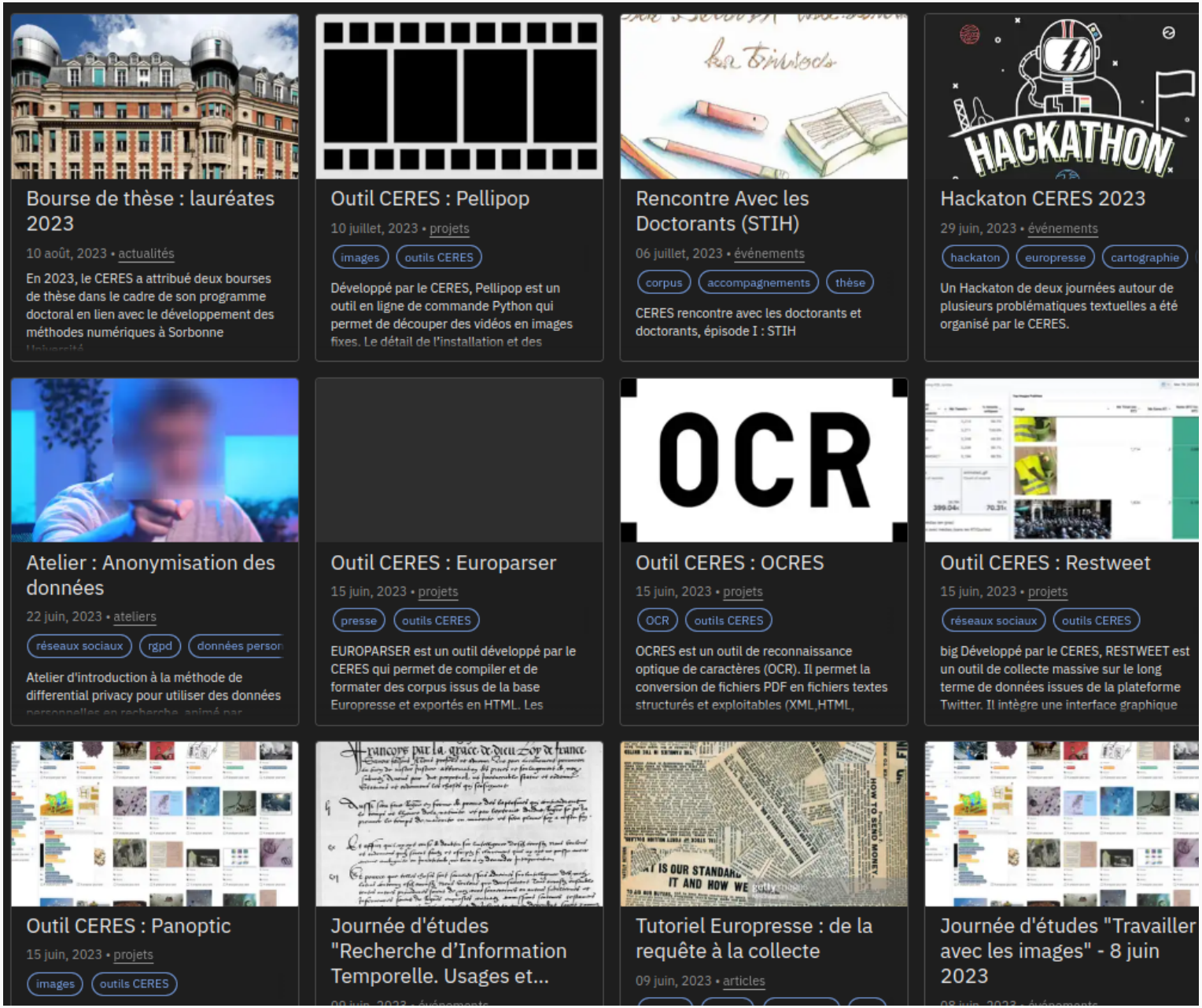
- Contexte : Faculté des Lettres de SU
- Pas ou peu d'ingénieurs de recherche
- Mais de nombreuses disciplines (sauf droit)

CERES

Centre d'Expérimentation pour les **Méthodes Numériques** en SHS

- Contexte : Faculté des Lettres de SU
- Pas ou peu d'ingénieurs de recherche
- Mais de nombreuses disciplines (sauf droit)
- Le numérique entre espérance et méfiance

CERES en un slide



Bourse de thèse : lauréates 2023

10 août, 2023 • actualités

En 2023, le CERES a attribué deux bourses de thèse dans le cadre de son programme doctoral en lien avec le développement des méthodes numériques à Sorbonne

Outil CERES : Pellipop

10 juillet, 2023 • projets

images | outils CERES

Développé par le CERES, Pellipop est un outil en ligne de commande Python qui permet de découper des vidéos en images fixes. Le détail de l'installation et des

Rencontre Avec les Doctorants (STIH)

06 juillet, 2023 • événements

corpus | accompagnements | thèse

CERES rencontre avec les doctorants et doctorants, épisode I : STIH

Hackaton CERES 2023

29 juin, 2023 • événements

hackaton | europresse | cartographie

Un Hackaton de deux journées autour de plusieurs problématiques textuelles a été organisé par le CERES.

Atelier : Anonymisation des données

22 juin, 2023 • ateliers

réseaux sociaux | rgpd | données perso

Atelier d'introduction à la méthode de differential privacy pour utiliser des données

Outil CERES : Europarser

15 juin, 2023 • projets

presse | outils CERES

EUROPARSER est un outil développé par le CERES qui permet de compiler et de formater des corpus issus de la base Europresse et exportés en HTML. Les

Outil CERES : OCRES

15 juin, 2023 • projets

OCR | outils CERES

OCRES est un outil de reconnaissance optique de caractères (OCR). Il permet la conversion de fichiers PDF en fichiers textes structurés et exploitables (XML,HTML,

Outil CERES : Restweet

15 juin, 2023 • projets

réseaux sociaux | outils CERES

big Développé par le CERES, RESTWEET est un outil de collecte massive sur le long terme de données issues de la plateforme Twitter. Il intègre une interface graphique

Outil CERES : Panoptic

15 juin, 2023 • projets

images | outils CERES

Journée d'études "Recherche d'Information Temporelle. Usages et...

Tutoriel Europresse : de la requête à la collecte

09 juin, 2023 • articles

Journée d'études "Travailler avec les images" - 8 juin 2023

Partie 1 : Traiter des données textuelles

Les termes du débat

Une application "numérique :

- des données
- des tâches

Les termes du débat

Une application "numérique :

- des données
- des tâches
- des méthodes

Les termes du débat

Une application "numérique :

- des données
- des tâches
- des méthodes

Si Données = Texte,

Les termes du débat

Une application "numérique :

- des données
- des tâches
- des méthodes

Si Données = Texte,

Et si Tâches = (Moteur de Recherche, Traduction, Génération . . .)

Les termes du débat

Une application "numérique :

- des données
- des tâches
- des méthodes

Si Données = Texte,

Et si Tâches = (Moteur de Recherche, Traduction, Génération . . .)

Alors la Méthode a à voir avec le Traitement Automatique des Langues

Focus : Traitement Automatique des Langues

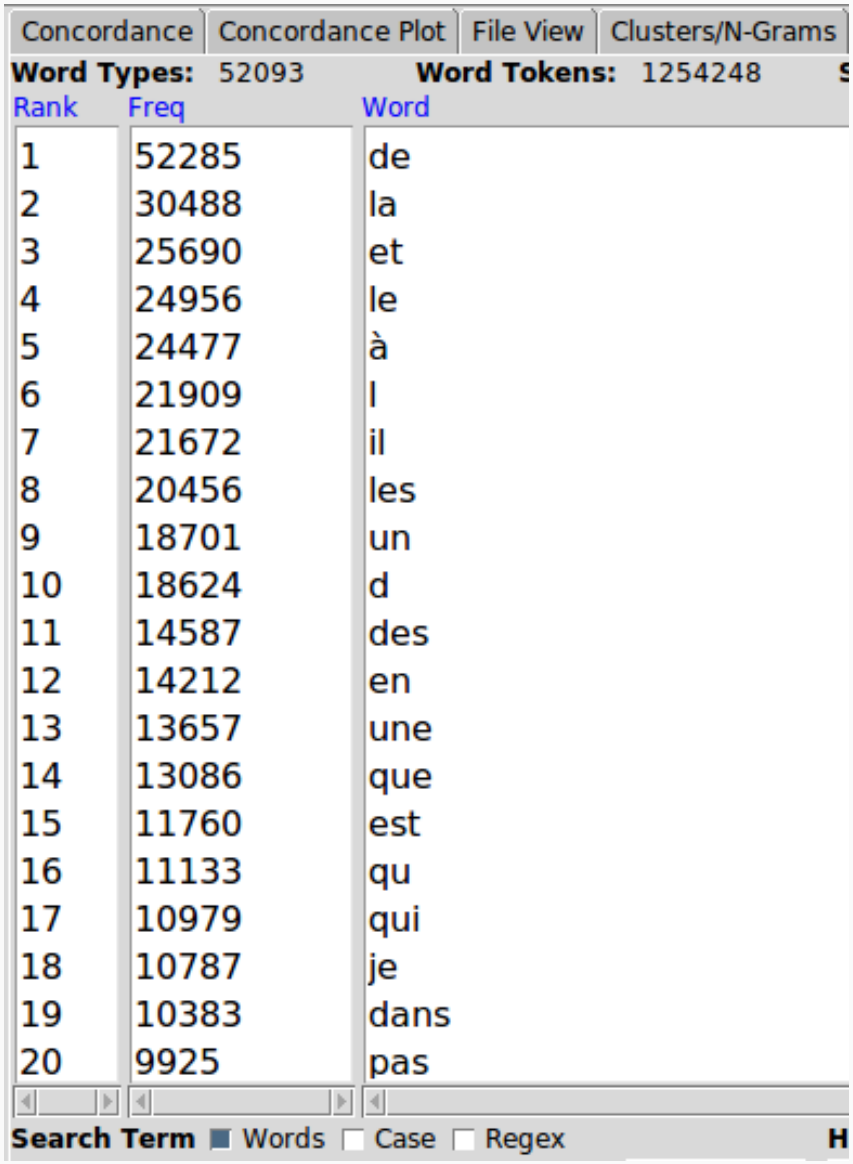


Figure 1 – Trouver les mots les plus fréquents (Antconc [Anthony, 2020])

Focus : Traitement Automatique des Langues

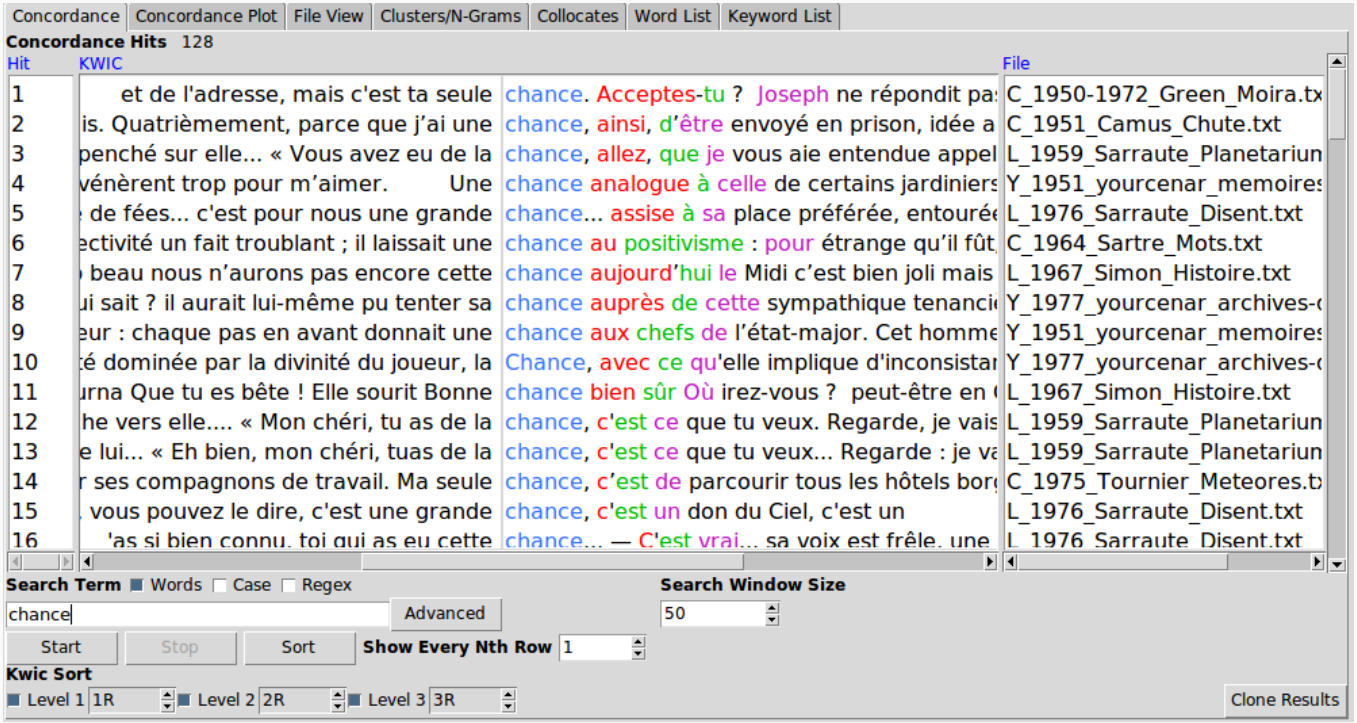


Figure 2 – Un mot en contexte (Antconc)

Focus : Traitement Automatique des Langues

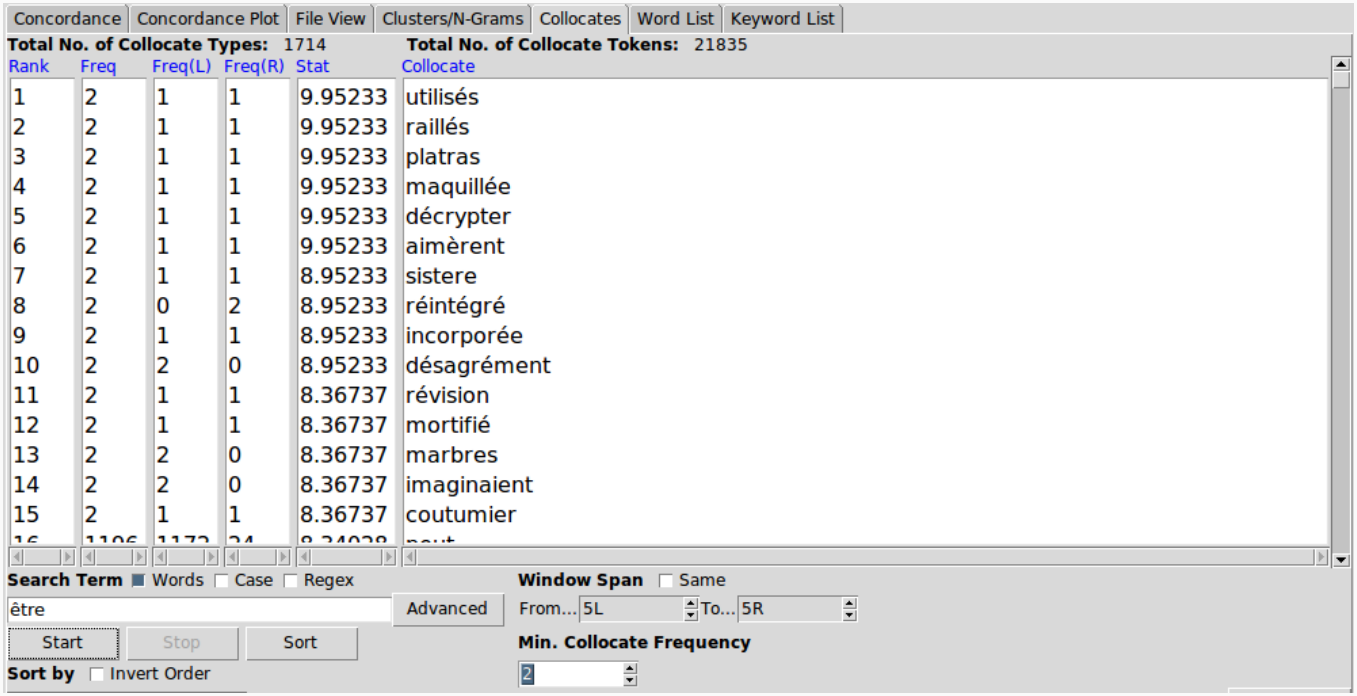


Figure 3 – Les co-occurrents d'un mot (Antconc)

Focus : Traitement Automatique des Langues

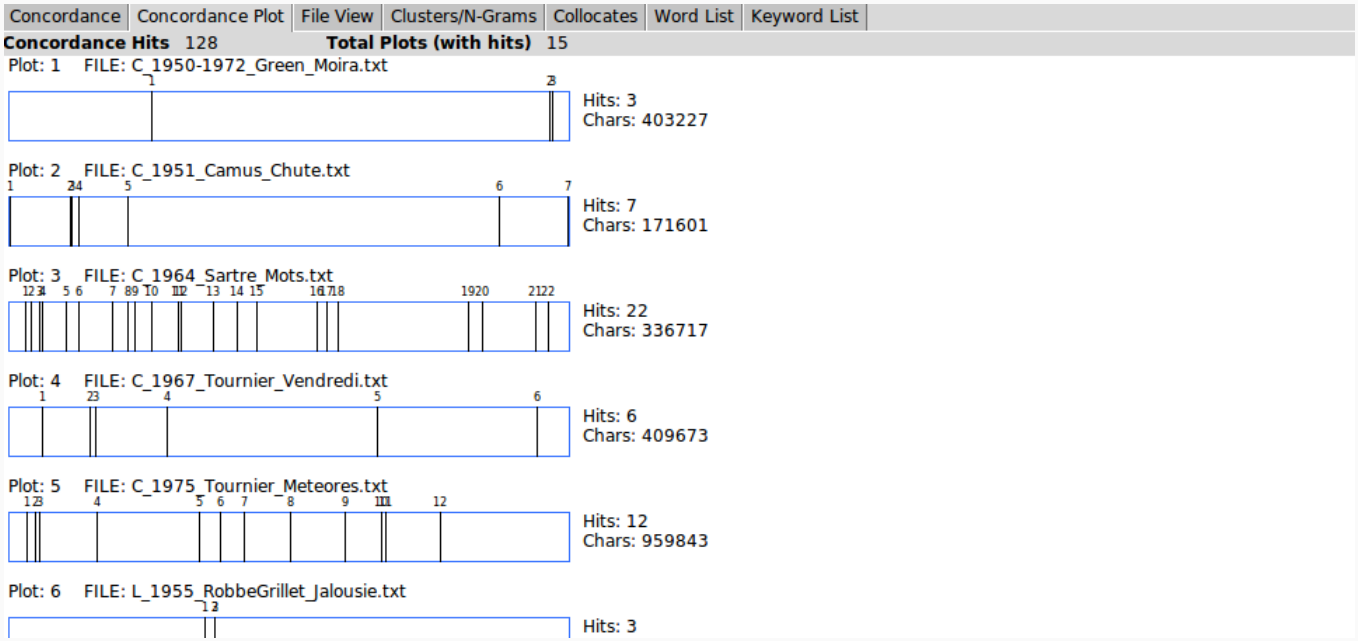


Figure 4 – Voir un mot au fil de différentes œuvres

Focus : Traitement Automatique des Langues

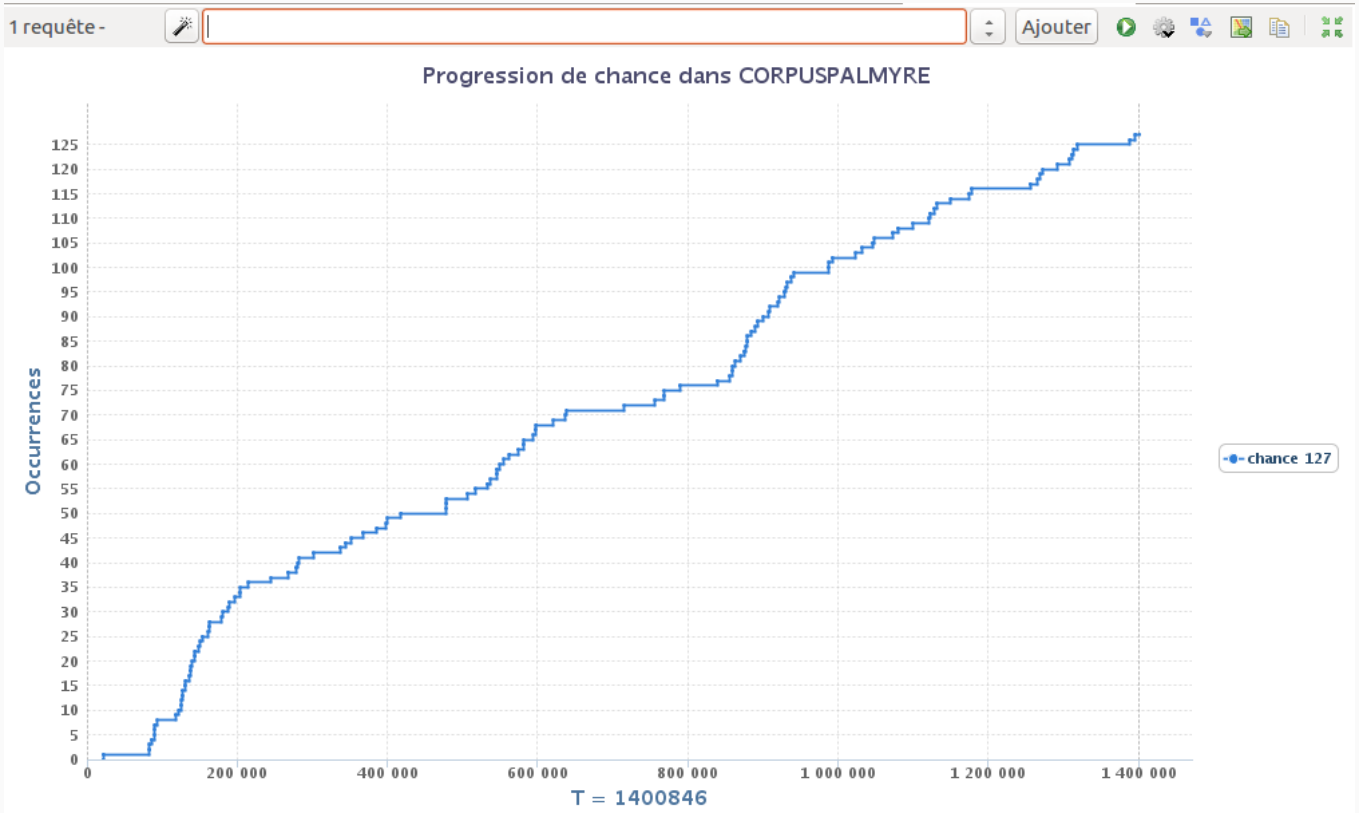


Figure 5 – La progression de la fréquence d'un mot dans le temps (TXM [Heiden et al., 2010])

Focus : Traitement Automatique des Langues

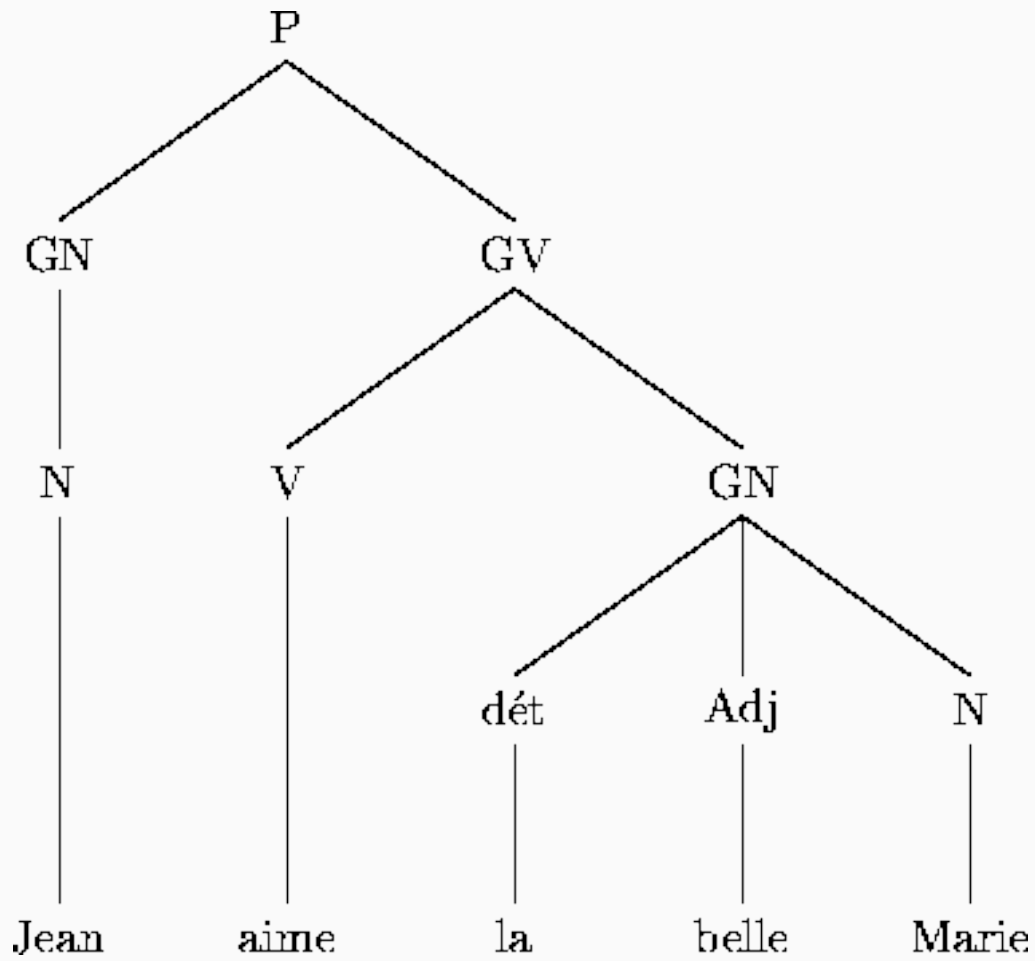


Figure 6 – Représentation Syntaxique d'une Phrase

Focus : Traitement Automatique des Langues

TITRE :Découverte d'un **nouveau virus** proche du **Sras**

Paragraphe 1/7 Un homme est hospitalisé et deux autres sont décédés après un séjour en Arabie saoudite. Les autorités sanitaires se veulent toutefois rassurantes car aucune contagion d'homme à homme n'a encore été détectée.
Paragraphe 2/7 Un **nouveau virus** appartenant à la famille du **Sras** (**syndrome respiratoire aigu sévère**), responsable de la mort de 800 personnes en 2002, a été identifié sur un Qatarien hospitalisé à Londres dans un état grave après avoir séjourné récemment en Arabie saoudite. C'est le troisième malade qui contracte ce **coronavirus** lors d'un séjour au Moyen-Orient, mais les autorités sanitaires se veulent rassurantes car aucune contamination d'homme à homme n'a pour l'instant été identifiée.
Paragraphe 3/7 C'est l'Organisation mondiale de la santé (OMS) qui a annoncé lundi la nouvelle via son système «d'alerte et de réponse globale». «Le patient est toujours en vie, mais d'après ce que nous savons, dans un état grave», a déclaré Gregory Hartl, porte-parole de l'OMS. L'homme de 49 ans souffre d'insuffisance respiratoire et rénale.
Paragraphe 4/7 Il existe un grand nombre de **coronavirus**. En général, ils provoquent des **rhumes** chez les humains. Mais une forme particulière de **coronavirus** à l'origine du **Sras** en 2002 avait tué 774 personnes dans le monde, dont 349 en Chine. Plus de 8000 personnes avaient été infectées.

Figure 7 – Extraction d'Information (Daniel [Lejeune, 2013])

Focus : Traitement Automatique des Langues

Figure 8 – Changement Sémantique (Dan Jurafsky [Hamilton et al., 2016])

Focus : Traitement Automatique des Langues

Figure 9 – Traduction et "Standardisation"

Conditions de réalisation de statistiques sur du texte

Conditions de réalisation de statistiques sur du texte

- Des données
- Représentatives
- Dans un format approprié
- En un mot : un **corpus** !

Conditions de réalisation de statistiques sur du texte

- Des données
- Représentatives
- Dans un format approprié
- En un mot : un **corpus** !

Le lien des sources à rechercher :
`https://tinyurl.com/HumaNumTabTD1`

References i

Anthony, L. (2020).

Antconc (software) version 3.5. 7 (windows).

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016).

Diachronic word embeddings reveal statistical laws of semantic change.

arXiv preprint arXiv :1605.09096.

Heiden, S., Magué, J.-P., and Pincemin, B. (2010).

Txm : Une plateforme logicielle open-source pour la textométrie-conception et développement.

In *JADT 2010*, volume 2, pages 1021–1032. Edizioni Universitarie di Lettere Economia Diritto.

References ii

Lejeune, G. (2013).

Veille épidémiologique multilingue : une approche parcimonieuse au grain caractère fondée sur le genre textuel.

PhD thesis, Université de Caen.