

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)

Parcours universitaire

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

Parcours universitaire

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

Parcours universitaire

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

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

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 ...)

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 ...)

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

9 cours de 2h :

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

9 cours de 2h :

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

9 cours de 2h :

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

1. Acteurs des Humanités Numériques : Le CERES
2. Partie 1 : Traiter des données textuelles

Acteurs des Humanités Numériques : Le 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)

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

[images](#)



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 (STH)

06 juillet, 2023 • événements

[corpus](#) [accompagnements](#) [thèse](#)

CERES rencontre avec les doctorants et doctorants, épisode 1 : STH



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

OCR

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).

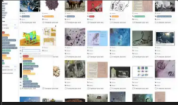


Outil CERES : Restweet

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...

09 juin, 2023 • articles



Tutoriel Europresse : de la requête à la collecte

09 juin, 2023 • articles



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

08 juin, 2023 • articles

Partie 1 : Traiter des données textuelles

Une application " numérique :

- des données
- des tâches

Une application " numérique :

- des données
- des tâches
- des méthodes

Une application "numérique" :

- des données
- des tâches
- des méthodes

Si Données = Texte,

Une application "numérique" :

- des données
- des tâches
- des méthodes

Si Données = Texte,

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

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

Concordance		Concordance Plot		File View		Clusters/N-Grams	
Word Types: 52093		Word Tokens: 1254248					
Rank	Freq	Word					
1	52285	de					
2	30488	la					
3	25690	et					
4	24956	le					
5	24477	à					
6	21909	l					
7	21672	il					
8	20456	les					
9	18701	un					
10	18624	d					
11	14587	des					
12	14212	en					
13	13657	une					
14	13086	que					
15	11760	est					
16	11133	qu					
17	10979	qui					
18	10787	je					
19	10383	dans					
20	9925	pas					

Search Term Words Case Regex H

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

Focus : Traitement Automatique des Langues

The screenshot displays the AntConc software interface. At the top, there are menu options: Concordance, Concordance Plot, File View, Clusters/N-Grams, Collocates, Word List, Keyword List. Below the menu, the title bar reads "Concordance Hits 128". The main window is divided into three columns: "Hit", "KWIC", and "File". The "Hit" column shows line numbers from 1 to 16. The "KWIC" column shows the search term "chance" highlighted in red in various contexts, such as "chance, Acceptes-tu ? Joseph ne répondit pas", "chance, ainsi, d'être envoyé en prison", "chance, allez, que je vous aie entendue appeler", "chance analogue à celle de certains jardiniers", "chance... assise à sa place préférée", "chance au positivisme : pour étrange qu'il fût", "chance aujourd'hui le Midi c'est bien joli mais", "chance auprès de cette sympathique tenancière", "chance aux chefs de l'état-major", "Chance, avec ce qu'elle implique d'inconsistance", "chance bien sûr Où irez-vous ? peut-être en France", "chance, c'est ce que tu veux. Regarde, je vais", "chance, c'est ce que tu veux... Regarde : je vois", "chance, c'est de parcourir tous les hôtels bordés", "chance, c'est un don du Ciel, c'est un", and "chance... — C'est vrai... sa voix est frêle, une". The "File" column lists the source files, including "C_1950-1972_Green_Moira.txt", "C_1951_Camus_Chute.txt", "L_1959_Sarraute_Planetarium.txt", "Y_1951_yourcenar_memoire.txt", "L_1976_Sarraute_Disent.txt", "C_1964_Sartre_Mots.txt", "L_1967_Simon_Histoire.txt", "Y_1977_yourcenar_archives-c.txt", "Y_1951_yourcenar_memoire.txt", "Y_1977_yourcenar_archives-c.txt", "L_1967_Simon_Histoire.txt", "L_1959_Sarraute_Planetarium.txt", "L_1959_Sarraute_Planetarium.txt", "C_1975_Tournier_Meteores.txt", "L_1976_Sarraute_Disent.txt", and "L_1976_Sarraute_Disent.txt". At the bottom, there are search controls: "Search Term" (chance), "Words" (checked), "Case" (unchecked), "Regex" (unchecked), "Advanced" (checked), "Search Window Size" (50), "Start" (button), "Stop" (button), "Sort" (button), "Show Every Nth Row" (1), "Kwic Sort" (checked), "Level 1 | 1R", "Level 2 | 2R", "Level 3 | 3R", and "Clone Results" (button).

Figure 2 – Un mot en contexte (Antconc)

Focus : Traitement Automatique des Langues

Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate
1	2	1	1	9.95233	utilisés
2	2	1	1	9.95233	raillés
3	2	1	1	9.95233	platras
4	2	1	1	9.95233	maquillée
5	2	1	1	9.95233	décrypter
6	2	1	1	9.95233	aimèrent
7	2	1	1	8.95233	sistere
8	2	0	2	8.95233	réintégré
9	2	1	1	8.95233	incorporée
10	2	2	0	8.95233	désagrément
11	2	1	1	8.36737	révision
12	2	1	1	8.36737	mortifié
13	2	2	0	8.36737	marbres
14	2	2	0	8.36737	imaginaient
15	2	1	1	8.36737	coutumier
16	1106	1177	24	8.34000	est

Search Term: Words Case Regex
être

Window Span: Same
From...: 5L To...: 5R

Min. Collocate Frequency:

Buttons: Start Stop Sort
Sort by: Invert Order

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

Focus : Traitement Automatique des Langues

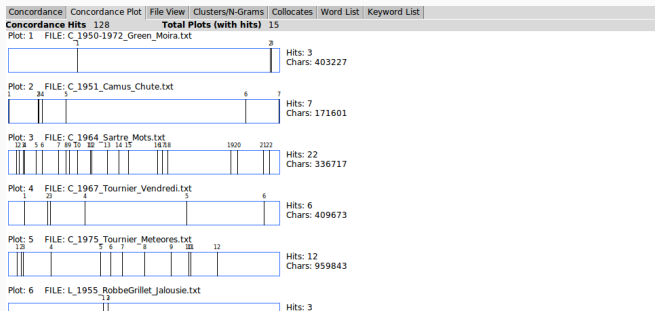


Figure 4 – Voir un mot au fil de différentes œuvres

Focus : Traitement Automatique des Langues

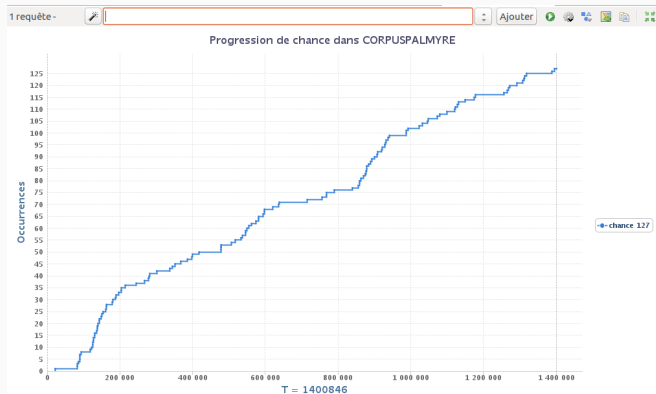


Figure 5 – La progression de la fréquence d'un mot dans le temps (TXM [Heiden et al., 2010])

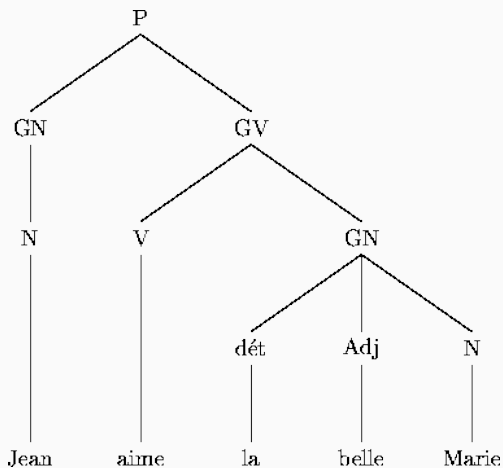


Figure 6 – Représentation Syntaxique d'une Phrase

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

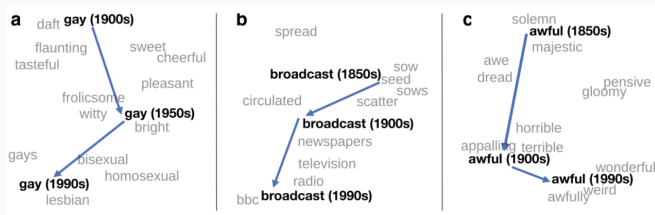


Figure 8 – Changement Sémantique (Dan Jurafsky [Hamilton et al., 2016])

Focus : Traitement Automatique des Langues



Figure 9 – Traduction et "Standardisation"

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

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

Le lien des sources à rechercher :

<https://tinyurl.com/HumaNumTabTD1>



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.



Lejeune, G. (2013).

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

PhD thesis, Université de Caen.