

Rencontres Minutes Département d'Informatique/STIH/CERES

STIH, CERES

2026-2027

UFR Sociologie-Informatique, Sorbonne Université

Un séminaire pour quoi ?

- Des rencontres régulières (1 fois/mois)
- Pas trop formelles

Un séminaire pour quoi ?

- Des rencontres régulières (1 fois/mois)
- Pas trop formelles
- Comprendre l'éco-système :
 - Enseignement : UFR et départements
 - Recherche : STIH, CERES, SCAI...

Un séminaire pour quoi ?

- Des rencontres régulières (1 fois/mois)
- Pas trop formelles
- Comprendre l'éco-système :
 - Enseignement : UFR et départements
 - Recherche : STIH, CERES, SCAI...
- Préparer le travail en équipe :
 - Connaître les collègues et partager des idées
 - Collaborer (en enseignement et en recherche)

Le Wiki STIH <http://stih-sorbonne-universite.fr/dokuwiki/doku.php?id=sem>

La "FAQ" UFR <https://gitlab.com/rundimeco/accueil-su/-/blob/master/guide>

1. Gaël Lejeune
2. Iglia-Nokolova Stoupak
3. Anaïs-Chambat
4. Dimitra-Niaouri
5. Francois-Olivier
6. Gianni-Gastaldi
7. Ljudmila-Petkovic
8. Rimane-Karam
9. Laurene Cave
10. Maxime Fily
11. Mélusine Caillard
12. Sophie Robert-Hayek
13. Marceau Hernandez

Gaël Lejeune

Qui suis-je ?

- Gaël Lejeune MCF HDR en informatique
- Thèse en 2013 : Veille Epidémiologique Multilingue

Mes centres d'intérêts :

Qui suis-je ?

- Gaël Lejeune MCF HDR en informatique
- Thèse en 2013 : Veille Epidémiologique Multilingue

Mes centres d'intérêts :

- Cuisine, Rugby et les Échecs

Qui suis-je ?

- Gaël Lejeune MCF HDR en informatique
- Thèse en 2013 : Veille Epidémiologique Multilingue

Mes centres d'intérêts :

- Cuisine, Rugby et les Échecs
- Extraction de Contenu/de structure (PDF, PNG, HTML...)
- Classification (polarité, émotion, dialectes, datation ...)
- Extraction d'Information et Reconnaissance d'Entités Nommées

Qui suis-je ?

- Gaël Lejeune MCF HDR en informatique
- Thèse en 2013 : Veille Epidémiologique Multilingue

Mes centres d'intérêts :

- Cuisine, Rugby et les Échecs
- Extraction de Contenu/de structure (PDF, PNG, HTML...)
- Classification (polarité, émotion, dialectes, datation ...)
- Extraction d'Information et Reconnaissance d'Entités Nommées

Dans un contexte de **variation** dans les données

- Multilinguisme : comment analyser n langues ?
- Hétérogénéité : comment traiter n états de textes ?
- Massification : comment travailler sur n To de textes ?

Projets de Recherche

Projets de Recherche

- Applicabilité de méthodes de TAL à des données spécialisées et bruitées (G. Roe et M. Alrahabi) **Ljudmila-Petkovic** (2021-)
- Simplification de textes et Diagnostic de complexité (E. Lacroix) **Iglika Nikolova Stouplak** (2023-)

Projets de Recherche

Projets de Recherche

- Applicabilité de méthodes de TAL à des données spécialisées et bruitées (G. Roe et M. Alrahabi) **Ljudmila-Petkovic** (2021-)
- Simplification de textes et Diagnostic de complexité (E. Lacroix) **Iglika Nikolova Stouplak** (2023-)
- COMMUNS DÉMOCRATIQUES & Make.org (J.P. Cointet, B.Piwowski, F. Yvon, Bernard Reber...), **Laurène Cave** (2024-)
- ANTONOMAZ (2018-) ANalyse auTOMatique et NumérisatiON des MAZarinades (K.Abiven) **X. Beckers** et **M. Hernandez** (2024-)
- IKOS Innovation Technologique dans le domaine du ferroviaire **Mélusine Caillard** (2025-)

Projets de Recherche

Projets de Recherche

- Applicabilité de méthodes de TAL à des données spécialisées et bruitées (G. Roe et M. Alrahabi) **Ljudmila-Petkovic** (2021-)
- Simplification de textes et Diagnostic de complexité (E. Lacroix) **Iglika Nikolova Stouplak** (2023-)
- COMMUNS DÉMOCRATIQUES & Make.org (J.P. Cointet, B.Piwowski, F. Yvon, Bernard Reber...), **Laurène Cave** (2024-)
- ANTONOMAZ (2018-) ANALyse auTOMatique et NumérisatiON des MAZarinades (K.Abiven) **X. Beckers** et **M. Hernandez** (2024-)
- IKOS Innovation Technologique dans le domaine du ferroviaire **Mélusine Caillard** (2025-)

Et le CERES : J. Bezançon, R. Karam, F. ALié/V. Julliard & Panoptic !

Iglike-Nokolova Stoupak

✉ email : iglika.nikolova.stoupak@gmail.com

✉ [page personnelle](#)

✉ chaîne YouTube [Monoglossia](#)

Iglika Nikolova-Stoupak

DOCTORANTE À L'ED CONCEPTS ET LANGAGES,
LABORATOIRE STIH (ANNÉE 4)

Directeurs de these : Gaël Lejeune, Eva Lacroix

Domaines d'intérêt et de recherche

- ❑ Langues à faibles ressources et langues historiques
- ❑ Éducation et apprentissage des langues étrangères

Sujet de thèse :

« Production de versions abrégées de textes littéraires : une approche multilingue »

Projets récents/en cours

(2021-2023) Traduction automatique entre le bulgare et le japonais,
Université de Kyoto

(2023-2024) CATS (Contextualisation Automatique de Termes de
Spécialité), laboratoire CENTAL, UCLouvain

(2024-) SHERBET (Stemmatologie pour la Transmission de la Bible
Hébraïque), LORIA, Université de Lorraine

Anaïs-Chambat



Quand la langue générale s'invite dans les dictionnaires de spécialité

Normes, autorités linguistiques et métalangage au XVIII^e siècle

10 septembre 2026

Séminaire de rentrée *Variation Linguistique et Linguistique Computationnelle*

Anaïs Chambat

ATER d'informatique pour les SHS
Sorbonne Université, STIH (UR 4509)

Doctorante en Sciences du langage
CY Cergy Paris Université, LT2D (UR 7518)

Le dictionnaire comme espace de construction du savoir

- Réceptacle d'une **représentation socioculturelle et lexiculturelle de la langue**, il est le reflet des réalités qu'elle désigne et de leur mise en mots (Galisson, 1988 et 1999)
- Quel que soit son programme linguistique, **sa temporalité n'est pas celle de la science**, mais bien « l'ensemble des connaissances antécédentes » (Auroux, 2006), celles d'aujourd'hui et de demain.

Le dictionnaire comme espace de construction du savoir

- Réceptacle d'une **représentation socioculturelle et lexiculturelle de la langue**, il est le reflet des réalités qu'elle désigne et de leur mise en mots (Galisson, 1988 et 1999)
- Quel que soit son programme linguistique, **sa temporalité n'est pas celle de la science**, mais bien « l'ensemble des connaissances antécédentes » (Auroux, 2006), celles d'aujourd'hui et de demain.

→ Comment appréhender cet « horizon de rétrospection » lexicographique ? (Auroux, 1986)

Un corpus privilégié : les dictionnaires réédités

→ Les dictionnaires réédités conservent une structure relativement stable, mais les modifications qu'ils recèlent peuvent être de différentes natures.

- **Discrètes** : reformulations légères, actualisation d'une référence, ajouts de synonymes ou de renvois, etc.
- **Massives** : évolutions sémantiques (changement de définition, d'exemple, d'auteur cité), glissements épistémologiques (passage d'un cadre théorique à un autre), modifications structurelles (ajout ou réécriture complète d'un article, déplacement de définitions), etc.

Un corpus privilégié : les dictionnaires réédités

→ Les dictionnaires réédités conservent une structure relativement stable, mais les modifications qu'ils recèlent peuvent être de différentes natures.

- **Discrètes** : reformulations légères, actualisation d'une référence, ajouts de synonymes ou de renvois, etc.
- **Massives** : évolutions sémantiques (changement de définition, d'exemple, d'auteur cité), glissements épistémologiques (passage d'un cadre théorique à un autre), modifications structurelles (ajout ou réécriture complète d'un article, déplacement de définitions), etc.

→ **Retracer** l'historicité des révisions éditoriales (modifications phrastiques) et typographiques successives (changements d'accentuation, de ponctuation ou encore de casse) dans une perspective diachronique.

Pour y répondre, nous avons créé COMPILEX !

→ Une interface de Comparaison Intertextuelle en LEXicographie

SECRET

1 Fiche HTML

← Retour à la recherche

● Basnage 1701 **1re éd.** — ● ColDeVillars 1753 **absent** — ● Lavoisien 1764 **absent** — ● Lavoisien 1771 **ajouté** — ● Lavoisien 1781 **modifié** —
● Lavoisien 1793 **dictionnaire** — ● Capuron 1806 **absent**

Comparaison inter-éditions

Basnage 1701

Fac-similés

XML sources

← secondines

secrétaire →

Lavoisien 1764, 1re éd.

— absent —

SECRET

Lavoisien 1771, 2e éd. p. 132

4 lignes - 46 tokens - 1 sens

ajouté

¶ 1

f. m. fecretum, arcanum, ce qu'on tient caché, qu'on ne veut pas découvrir pour quelques caufes particulieres. Voy. **arcane**. L'**Académie** dit que le c du mot de fecret fe prononce comme fi c'étoit un g; fegret.

SECRET

Lavoisien 1781, 3e éd. p. 132

Dice **98** Jacc **96** Cos **98** Seq **99**

4 lignes - 45 tokens - 1 sens

modifié

¶ 1

f. m. fecretum, arcanum, ce qu'on tient caché, qu'on ne veut pas découvrir pour quelques caufes particulieres. Voy. Voyez **ARCANE**. L'**Académie** dit que le c du mot de fecret fe prononce comme fi c'étoit un g; fegret.

SECRET

Lavoisien 1793, 4e éd. p. 518

Dice **100** Jacc **100** Cos **100** Seq **100**

4 lignes - 45 tokens - 1 sens

dictionnaire

¶ 1

f. m. fecretum, arcanum, ce qu'on tient caché, qu'on ne veut pas découvrir pour quelques caufes particulieres. Voyez **ARCANE**. L'**Académie** dit que le c du mot de fecret fe prononce comme fi c'étoit un g **g** **g**

Quand la langue générale s'invite dans les dictionnaires de spécialité...

4/5

Anaïs Chambat

- `anais.chambat@cyu.fr`
- `anais.chambat@sorbonne-universite.fr`
- `https://achambat.github.io/`
- `https://achambat-these.gitpages.huma-num.fr/compilex/index.html`

Merci de votre attention !

Dimitra-Niaouri

ARENAS

ANALYSIS OF AND RESPONSES
TO EXTREMIST NARRATIVES

Modèles DL/ML de nouvelle génération pour la détection des récits extrêmes dans le discours social en ligne

Dimitra Niaouri

Doctorante en 3e année · CY Cergy Paris Université
Laboratoires ETIS & AGORA · actuellement ATER à
Sorbonne Université

Encadrants : Michele Linardi, Julien Longhi & Nistor Grozavu

Idée centrale : passer de l'analyse de textes isolés à
une modélisation relationnelle et explicable des récits
extrémistes.



www.arenasproject.eu



This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101094731

Problématique

Des récits relationnels, pas seulement des textes isolés

Les récits extrémistes ne se réduisent pas à des propos toxiques isolés.
Ils organisent des groupes, des valeurs, des événements et des réponses proposées.

Contexte

Les crises récentes, les conflits géopolitiques, les tensions sociales et la désinformation ont accentué la polarisation des discours publics, notamment en ligne.

Limite actuelle

Les modèles se concentrent souvent sur des messages isolés, alors qu'un récit extrême peut se construire et se renforcer à travers plusieurs messages et leurs relations

Objectif

Développer des outils DL/ML pour caractériser, modéliser et détecter les EN dans différents corpus et contextes.

Travaux en cours

De la détection à la modélisation explicable

Mes travaux s'organisent autour de trois axes complémentaires

1. Détection du SUD

Évaluation comparative de modèles de langage pour le discours socialement inacceptable (SUD), à grande échelle et en classification multiclassée.

2. Apprentissage robuste au bruit

Analyse de l'influence des tokens et masquage ciblé pour améliorer la robustesse de la classification et détecter les annotations potentiellement bruitées.

3. Analyse par graphes

Graphes de messages pour classer les éléments narratifs et expliquer les relations influentes.

Objectif : détecter, caractériser et expliquer les récits extrémistes dans le discours social en ligne

Francois-Olivier

Modéliser nos représentations mentales

François Olivier

STIH, Sorbonne Université

Séminaire de rentrée de Variation Linguistique et Linguistique Computationnelle
10 septembre 2026

Structures mentales sous-jacentes au langage

Quelle est la structure sous-jacente à ces phrases ?

- John est dans la cuisine
- Mary est en France
- Être hors de danger
- I'm in trouble
- Sortir du coma
- To be in love
- $2 \in \mathbb{N}$

Modéliser des schémas spatio-temporels qualitatifs

Le langage formel

- Logique intuitionniste du premier ordre sans quantificateurs.
- Non-monotone (**négation par défaut**).
- **Prédicats spatiaux**.
- Opérateurs temporels (I , $\mathbb{F}$, $\square$,...)

Pour la phrase « Bill a quitté le pays ».

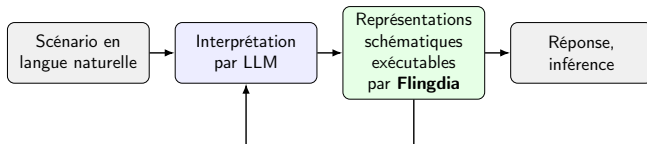
- $\square(I \rightarrow \text{in}(\text{bill}, \text{pays}))$
- $\square(\text{move}(\text{bill}) \vee \neg \text{move}(\text{bill}))$
- $\square(\mathbb{F} \rightarrow \neg \text{in}(\text{bill}, \text{pays}))$

Flingdia : l'implémentation

- point, rect, ligne.
- inside, bigger, above,...
- clingo (ASP) + métaprogrammation.
- solutions + diagrammes.
- github.com/fcs-olivier/flingdia

```
point(bill).    rect(pays).  
  
&in(bill,pays) :- &initial.  
{ &move(bill) }.  
:- &final, &in(bill,pays).
```

Embodied-LM : ancrer le raisonnement des modèles de langue

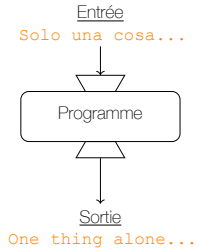


F. Olivier et Z. Bouraoui. *Towards a Neurosymbolic Reasoning System Grounded in Schematic Representations*.
NeSy 2025, PMLR 284 :420–438.

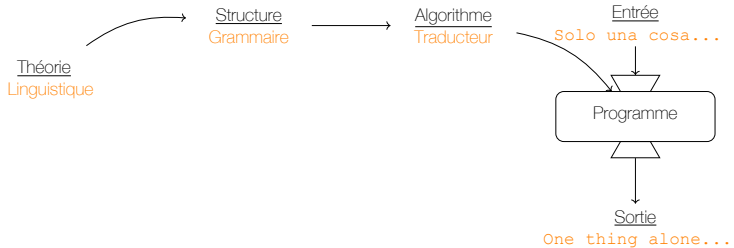
proceedings.mlr.press/v284/olivier25a.html

Gianni-Gastaldi

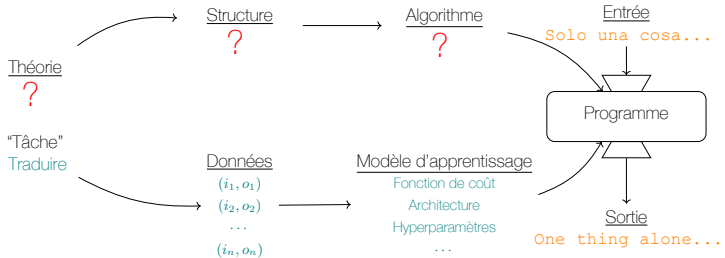
Motivation: La structure implicite des données



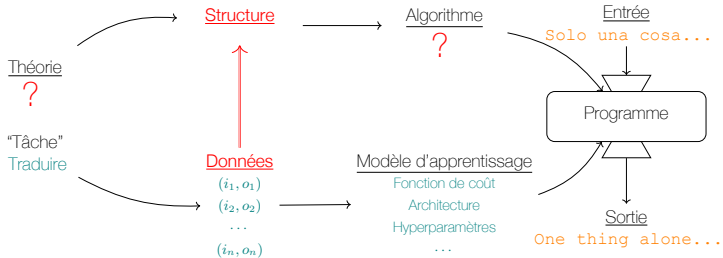
Motivation: La structure implicite des données



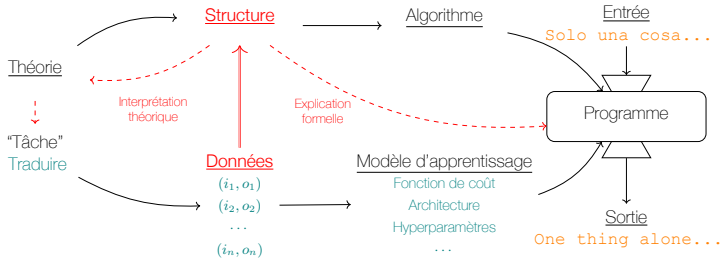
Motivation: La structure implicite des données



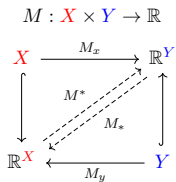
Motivation: La structure implicite des données



Motivation: La structure implicite des données

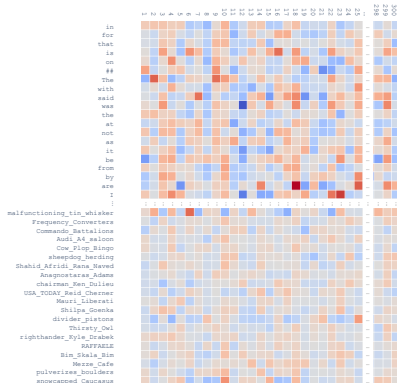


Opérateur “distributionnel” et points fixes

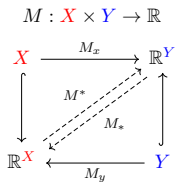


$$\begin{aligned}
 M_* M^* : \mathbb{R}^X &\rightarrow \mathbb{R}^X \\
 M^* M_* : \mathbb{R}^Y &\rightarrow \mathbb{R}^Y
 \end{aligned}$$

$$\begin{aligned}
 \{u_i \in \mathbb{R}^X \mid M_* M^* u_i &= \lambda_i u_i\} \\
 \{v_i \in \mathbb{R}^Y \mid M^* M_* v_i &= \lambda_i v_i\}
 \end{aligned}$$



Opérateur “distributionnel” et points fixes

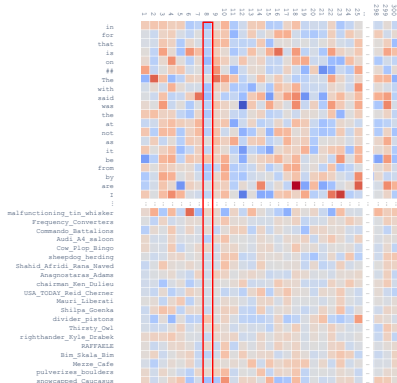


$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$

$$M^* M_* : \mathbb{R}^Y \rightarrow \mathbb{R}^Y$$

$$\{u_i \in \mathbb{R}^X \mid M_* M^* u_i = \lambda_i u_i\}$$

$$\{v_i \in \mathbb{R}^Y \mid M^* M_* v_i = \lambda_i v_i\}$$



Opérateur “distributionnel” et points fixes

$$\begin{array}{ccc}
 & M : X \times Y \rightarrow \mathbb{R} & \\
 \textcolor{red}{X} & \xrightarrow{M_x} & \mathbb{R}^{\textcolor{blue}{Y}} \\
 \downarrow & \swarrow M^* & \uparrow \\
 \mathbb{R}^{\textcolor{red}{X}} & \xleftarrow{M_y} & Y
 \end{array}$$

Diagram illustrating the operator M and its adjoint M^* between spaces X and Y .

$$\left. \begin{array}{l} M_* M^* : \mathbb{R}^{\textcolor{red}{X}} \rightarrow \mathbb{R}^{\textcolor{red}{X}} \\ M^* M_* : \mathbb{R}^{\textcolor{blue}{Y}} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}} \end{array} \right\} \text{opérateur distributionnel}$$

$$\left. \begin{array}{l} \{u_i \in \mathbb{R}^{\textcolor{red}{X}} \mid M_* M^* u_i = \lambda_i u_i\} \\ \{v_i \in \mathbb{R}^{\textcolor{blue}{Y}} \mid M^* M_* v_i = \lambda_i v_i\} \end{array} \right\} \begin{array}{l} \text{points} \\ \text{fixes} \end{array}$$

$$\begin{array}{ccc}
 & \mathcal{C}^{\text{op}} \times \mathcal{D} \rightarrow \text{Set} & \\
 \textcolor{brown}{C} & \xrightarrow{\mathcal{M}_c} & (\text{Set}^{\textcolor{green}{D}})^{\text{op}} \\
 \downarrow \textcolor{brown}{Y}_{\text{oneda}} & \swarrow \mathcal{M}^* & \uparrow \textcolor{green}{Y}_{\text{oneda}} \\
 \text{Set}^{\textcolor{brown}{C}^{\text{op}}} & \xleftarrow{\mathcal{M}_d} & \textcolor{green}{D}
 \end{array}$$

Diagram illustrating the operator $\mathcal{M}$ and its adjoint $\mathcal{M}^*$ between categories $\mathcal{C}$ and $\mathcal{D}$.

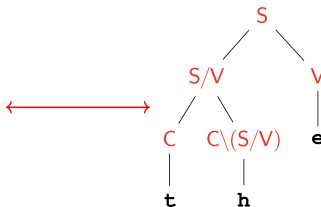
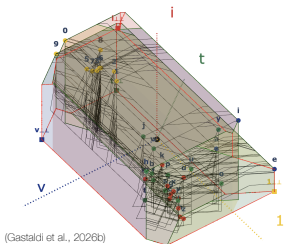
$$\left\{ \begin{array}{l} \mathcal{M}_* \mathcal{M}^* : \text{Set}^{\textcolor{brown}{C}^{\text{op}}} \rightarrow \text{Set}^{\textcolor{brown}{C}^{\text{op}}} \\ \mathcal{M}^* \mathcal{M}_* : (\text{Set}^{\textcolor{green}{D}})^{\text{op}} \rightarrow (\text{Set}^{\textcolor{green}{D}})^{\text{op}} \end{array} \right.$$

$$\left\{ \begin{array}{l} \{f \in \text{Set}^{\textcolor{brown}{C}^{\text{op}}} \mid \mathcal{M}_* \mathcal{M}^*(f) \cong f\} \\ \{g \in (\text{Set}^{\textcolor{green}{D}})^{\text{op}} \mid \mathcal{M}^* \mathcal{M}_*(g) \cong g\} \end{array} \right.$$

Profoncteur et structure du noyau

$$\begin{array}{ccccc}
 e_i & & s_i & & \text{mesure} \\
 \downarrow & & \downarrow & & \downarrow \\
 \mathbf{C}^{\text{op}} \times \mathbf{D} & \xrightarrow{\quad} & \bar{\mathbb{R}} & & \\
 & \Downarrow & & & \\
 \mathcal{M}^* : \bar{\mathbb{R}}^{\mathbf{C}^{\text{op}}} & \rightleftharpoons & (\bar{\mathbb{R}}^{\mathbf{D}})^{\text{op}} : \mathcal{M}_*
 \end{array}$$

Profoncteur et structure du noyau



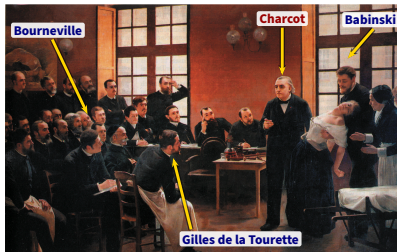
Gastaldi and Pellissier, 2021
(Gastaldi et al., 2026a)

$$\begin{aligned}
 & \mathbf{C}^{\text{op}} \times \mathbf{D} \rightarrow \bar{\mathbb{R}} \\
 & \quad \Downarrow \\
 & \mathcal{M}^* : \bar{\mathbb{R}}^{\mathbf{C}^{\text{op}}} \Leftrightarrow (\bar{\mathbb{R}}^{\mathbf{D}})^{\text{op}} : \mathcal{M}_*
 \end{aligned}$$

Red curved arrows connect the 3D diagram to the left side of the equation and the syntactic tree to the right side.

Ljudmila-Petkovic

Objectiver l'influence d'un savant dans un fonds bruité



Une leçon clinique à la Salpêtrière, A. Brouillet, 1887.

Ljudmila PETKOVIĆ

doctorante, ED 433 · CELLF · ObTIC

dir. G. Lejeune, co-enc. G. Roe et M. Alrahabi

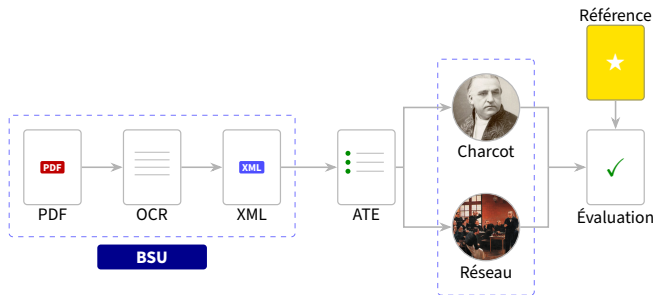
*Dans les petits papiers de Charcot : de
l'expérimentation aux prémisses de la
neurologie moderne*

2021-2026, thèse en voie d'achèvement

- Charcot fonde la neurologie française et forme une école
- ses concepts sont repris et reformulés par son réseau scientifique

Ce transfert laisse-t-il une trace mesurable dans les textes ?

Du PDF océrisé au terme évalué

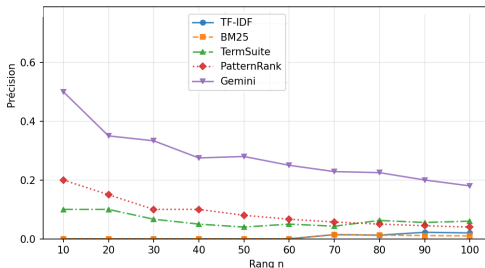


Corpus : 160 docs (C. 68, R. 92), OCR sans post-correction, ≈ 24 M de tokens

Extraction : TF-IDF · BM25 · TermSuite · PatternRank · Gemini

Évaluation : sur une liste de 33 termes de référence

Un LLM vs. les mesures classiques



Précision au rang n , corpus Charcot.

- Gemini surpasse les autres mesures à tous les rangs
- TF-IDF et BM25 restent à zéro sur leurs 60 premiers rangs

Pistes : *clustering* non supervisé des termes; corpus enrichi...

Rimane-Karam

L'épopée numérique de Baybars.

Outils et méthodes pour explorer la variation linguistique dans un corpus de littérature populaire en moyen arabe damascène

Rimane Karam

Sorbonne Université, CERES, Orient & Méditerranée, LiPoL

Sous la direction de : Julien Dufour, Iyas Hassan et Gaël Lejeune

Laurene Cave

Maxime Fily

Mélusine Caillard

Sophie Robert-Hayek

Marceau Hernandez
