

Mesurer l'impact de Jean-Martin Charcot *via* l'extraction des termes médicaux : quelle approche adopter ?

Ljudmila PETKOVIC^{1,2,3,4}

`prenom.nom@sorbonne-universite.fr`

¹ Sorbonne Université, Faculté des Lettres, UFR Littératures françaises et comparée, ED III (ED019)

² Centre d'étude de la langue et des littératures françaises (CELLF), UMR 8599

³ Observatoire des textes, des idées et des corpus (ObTIC)

⁴ UFR Sociologie et Informatique pour les Sciences Humaines

Séminaire doctoral ObTIC

SCAI, salle 1

Paris, le 6 février 2025



1 Projet Charcot

2 Problématique

3 Méthodologie

4 Expériences

5 Conclusion

Valorisation des archives de Jean-Martin Charcot

Dans les petits papiers de Charcot : de l'expérimentation aux prémisses de la neurologie moderne¹

Thèse en cours

initiative OPUS^a

- dir. : Prof. D^r Glenn ROE
- co-enc. : D^r Motasem ALRAHABI

a. <https://institut-opus.sorbonne-universite.fr/node/478>



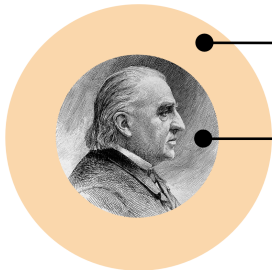
Fig. 1 – J.-M. Charcot (1825-1893),
Wikipédia.

- père de la neurologie moderne
- hystérie, « Parkinson », SLA...
- Freud, de la Tourette, Babinski...

1. <https://theses.fr/s382733>

- 1 Projet Charcot
- 2 Problématique**
- 3 Méthodologie
- 4 Expériences
- 5 Conclusion

Double objectif



Formaliser une approche numérique pour tracer l'évolution des concepts médicaux en général

Pister numériquement la circulation du discours médical de Charcot

Est-il possible de mesurer l'impact de Charcot sur son réseau scientifique* en s'appuyant sur les termes scientifiques qu'il a employés?

* élèves, collègues et successeurs

→ collaborateurs

Hypothèses

Le texte comme point d'entrée pour étudier les tendances de la circulation des idées à l'aide des caractéristiques structurelles spécifiques

(MILIA, 2023)

Certains **termes médicaux** dont Charcot a été l'inventeur (SLA) ou le transmetteur (*hystérie*) ont été repris de manière significative dans les écrits de son réseau scientifique.

Recherches quantitatives des circulations des savoirs

Tendances de circulation linguistique dans les fronts de recherche

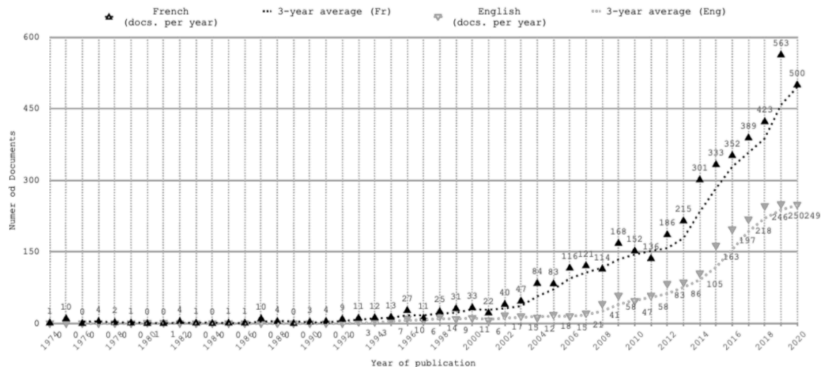


Fig. 2 – Nombre de documents représentant les intérêts thématiques en anglais et en français (MILIA, 2023).

Recherches quantitatives des circulations des savoirs

Tendances de circulation linguistique dans les fronts de recherche

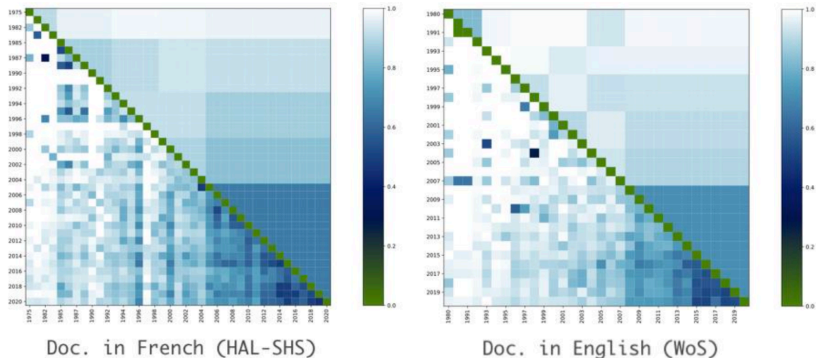


Fig. 3 – Détection de période à l'aide d'une corrélation significative des termes au fil des années dans deux sources et langues différentes (MILIA, 2023).

Recherches quantitatives des circulations des savoirs

- réception de la pensée scientifique de C. Bernard ([RIGUET, 2018](#))
- détection des réemplois textuels ([FEDCHENKO et al., 2024](#))
- Rankingdom – mesurer l'importance d'une entité ([SOULET, 2024](#))

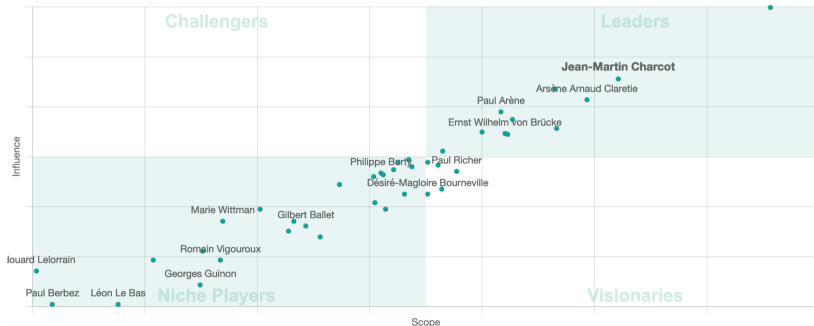


Fig. 4 – Positionnement de l'entité Charcot au sein de son domaine *via* l'analyse de quadrant.

- 1 Projet Charcot
- 2 Problématique
- 3 Méthodologie**
- 4 Expériences
- 5 Conclusion

Fonds Charcot²

SorbonNum

Bibliothèque de Sorbonne Université (BSU)

201 documents XML OCRisés (sans post-correction)

Corpus	Nb de docs	Nb de tokens
Charcot textes rédigés par Charcot	68	12 190 649 (38,12%)
Autres textes rédigés par les membres de son réseau scientifique	133	19 788 830 (61,88%)
Total	201	31 979 479 (100%)

Tab. 1 – Répartition du fonds Charcot selon les auteurs.

2. <https://patrimoine.sorbonne-universite.fr/collection/Fonds-Charcot>

Formalisation de l'approche

Extraction des concepts → problème d'extraction de la terminologie

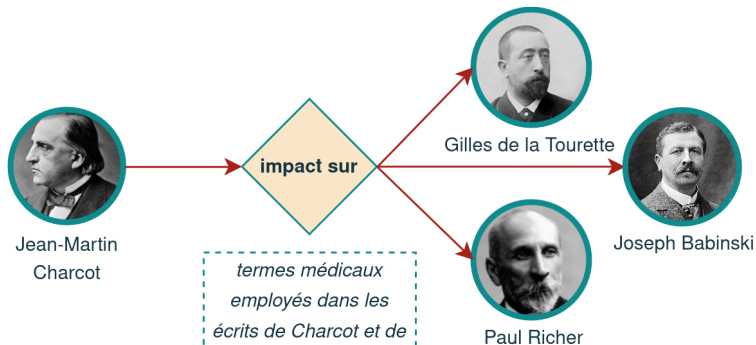


Fig. 5 – Opérationnalisation de l'impact de Charcot sur ses élèves.

Liste des concepts médicaux – vérité terrain

Extraction semi-automatique des termes en lien avec Charcot.

https://github.com/ljpetkovic/Charcot_circulations/tree/main/concepts

HYSTÉRIE (V. ÉPIDÉMIE, HÉMIANES-THÉSIE, HYPERESTHÉSIE OVARIENNE, ISCHURIE, SECOURS) ; — *épileptiforme*, 369 ; — *ovarienne*, 302 ; — *grave*, 306, 383 ; — *locale*, 320. — *infantile*, 451. — *locale traumatique*, 450.

HYSTÉRO-ÉPILEPSIE, 332, 367. — Signification de ce mot, 368 ; — à crises distinctes, 371. — Variétés de l' —, 370. — Nature de l' —, 373. — Température dans l' —,

```
<p>
<s>Hystérie (V. Epidémie, Hémianes-thésie, Hyperesthésie ovarienne,
</p>
<p>
<s>Hystéro-épilepsie, 332, 367. — Signification de ce mot, 368 ; —
</p>
```

Fig. 7 – Concepts médicaux, document XML.

Fig. 6 – Index des termes (CHARCOT, 1892).

<u>hystérie(s)?</u>
<u>hystérie(s)? épileptiforme(s)?</u>
<u>hystérie(s)? ovarienne(s)?</u>
<u>hystérie(s)? grave(s)?</u>
<u>hystérie(s)? locale(s)?</u>
<u>hystérie(s)? infantile(s)?</u>
<u>hystérie(s)? locale(s)? traumatique(s)?</u>
<u>hystéro-épilepsie(s)?</u>

Fig. 8 – Liste finale des concepts médicaux.

- 1 entre <s> et , - (regex)
- 2 sans termes génériques (*os*, *peau*)
- 3 prise en compte des sg. / pl. (regex)

- 1 Projet Charcot
- 2 Problématique
- 3 Méthodologie
- 4 Expériences**
- 5 Conclusion

Extraction de la terminologie : approche linguistique · TermSuite^{3, 4}

Termes uniques + motifs syntaxiques (6)

- système à base de règles + mesures statistiques
 - TF-IDF, spécificité, fréquences...
 - 261 termes en commun

Corpus Charcot				Corpus Autres		
Motif POS	Effectif	Fréq. relat. (%)	Exemple	Effectif	Fréq. relat. (%)	Exemple
N	261	52,10	<i>hystérie</i>	271	54,20	<i>somnambule</i>
A	151	30,14	<i>cérébral</i>	149	29,80	<i>hypnotique</i>
NA	73	14,57	<i>système nerveux</i>	73	14,60	<i>lame médullaire</i>
NPN	12	2,40	<i>cas de folie</i>	6	1,20	<i>scissure de sylvius</i>
NAA	3	0,60	<i>système nerveux central</i>	0	0,00	–
R	1	0,20	<i>[d']emblée</i>	1	0,20	<i>obliquement</i>
Total	501	100,00		500	100,00	

Tab. 2 – Répartition des parties du discours constituant les termes médicaux dans les deux corpus.

3. <https://termsuite.github.io/>

4. https://github.com/ljpetkovic/Charcot_TermSuite/tree/main

Répartition des motifs syntaxiques

- unigrammes de noms [N] et des adjectifs [A] : les plus fréquents
- trigrammes : les séquences les plus longues extraites
 - aucune occurrence du trigramme *sclérose latérale amyotrophique*
seule trace : adjectif [A] *latérale*
 - *idem* pour le quadrigramme *état de mal hystéro-épileptique*

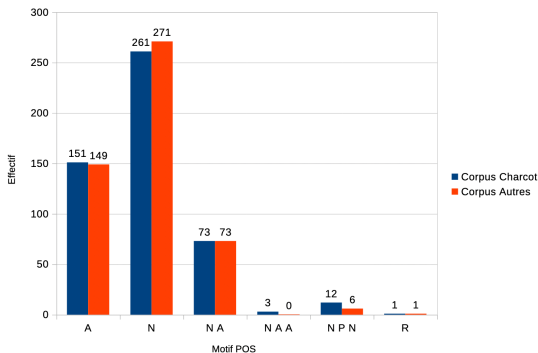


Fig. 9 – Analyse comparative des séquences syntaxiques.

TF-IDF

Premiers 12 termes communs, à partir du TF-IDF $\approx 0,15$

- prévalence des unigrammes

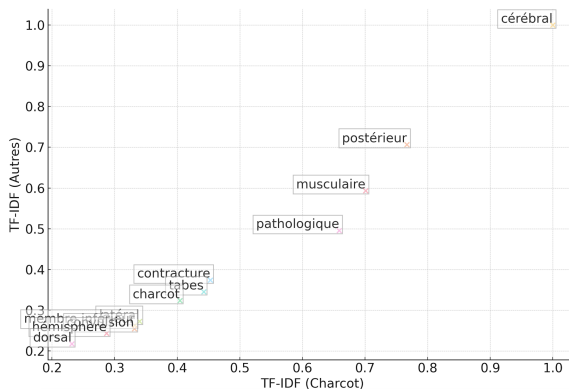


Fig. 10 – Termes communs avec leurs valeurs TF-IDF respectives.

- 1 Projet Charcot
- 2 Problématique
- 3 Méthodologie
- 4 Expériences
- 5 Conclusion**

Conclusion

Il est possible de quantifier la pertinence des termes scientifiques partagés entre Charcot et son réseau

⚠ Limitations de l'approche linguistique de l'extraction des termes
→ les termes plus pointus sont pénalisés

À faire :

- rendre compte du contexte d'énonciation des termes
- déterminer l'approche à adopter (plongements dynamiques?)
- trouver un corpus « externe » : prouver la pertinence des termes détectés pour la communauté scientifique

Références I



CHARCOT, J. M. (1892). *Œuvres complètes de J. M. Charcot. Leçons sur les maladies du système nerveux*. T. 1. <https://patrimoine.sorbonne-universite.fr/viewer/3468/?offset=1#page=2&viewer=picture&o=&n=0&q=>. Bureaux du progrès medical (voir p. 13).



FEDCHENKO, V., D. M. NICOLOSI et G. ROE (2024). À la recherche des réseaux intertextuels : défis de la recherche littéraire à grande échelle. In : *Humanités numériques* 9. <https://journals.openedition.org/revuehn/3940> (voir p. 9).



MILIA, M. (2023). Using Digital Text-Based Approaches to Study Knowledge Circulation. In : *Routledge Handbook of Academic Knowledge Circulation*. <https://www.taylorfrancis.com/chapters/edit/10.4324/9781003290650-52/using-digital-text-based-approaches-study-knowledge-circulation-mat%C3%ADas-milia>. Routledge, p. 556-574 (voir pp. 6-8).



RIGUET, M. (2018). L'impact de la physiologie dans la critique littéraire de la fin du XIX^{ème} siècle : l'exemple de Claude Bernard. In : *Epistémocritique : Littérature et savoirs*. <https://hal.science/hal-01903871> (voir p. 9).



SOULET, A. (2024). *Vers l'analyse à la demande des connaissances de Wikidata*. https://afia.asso.fr/wp-content/uploads/2024/05/SOULET_IA-et-HN-2024-05-03.pdf. Consulté le 6 février 2025 (voir p. 9).