Mining of the Heritage Collections in Medical Domain

Ljudmila PETKOVIC[1,2,3], Motasem ALRAHABI[1,3], Glenn ROE[1,2,3]

firstname.lastname@sorbonne-universite.fr

[1]Sorbonne University | Faculty of Arts, Languages, Literature and Humanities
[2]Center for French Language and Literature Studies (CELLF), UMR 8599
[3]Observatory of Texts, Ideas and Corpora (OBTIC)

Summer School on Digital Humanities
September 13-14, 2023 · UNAM, Mexico City (Mexico)

Overview

"Napoleon of Neurosis" or "Paganini of Hysteria" (Marmion 2015)

Charcot's Portrait (Wikipedia).

Jean-Martin Charcot (1825-1893)

- father of modern neurology, forefather of psychoanalysis

1868	first diagnosed multiple sclerosis
1869	first diagnosed amyotrophic lateral sclerosis
1870	hysteria: *neurological* pathology, *both* sexes
1887-88	hypnosis as a method of investigating hysteria
	states: lethargy, catalepsy, somnambulisme
	Tuesday clinical lessons, Salpêtrière Hospital, Paris
1872	coined the term "Parkinson's disease"

(Gomes and Engelhardt 2013 ; White 1997)

Charcot's Impact on his Scientific and Artistic Network

Disciples – Salpêtrière school

Sigmund Freud (1856-1939)	psychoanalytic theory
Gilles de la Tourette (1857-1932)	Tourette's syndrome
Joseph Babinski (1857-1904)	Babinski sign
Pierre Janet (1859-1947)	dissociation theory

Writers (Koehler 2013)

Émile Zola (1840–1902)	*Lourdes*
Leo Tolstoy (1828–1910)	*The Kreutzer Sonata*
Leopoldo Alas Clarín (1852–1901)	*La Regenta*
Luigi Capuana (1839–1915)	*Giacinta*

At the Junction of Digital Humanities and the History of Science

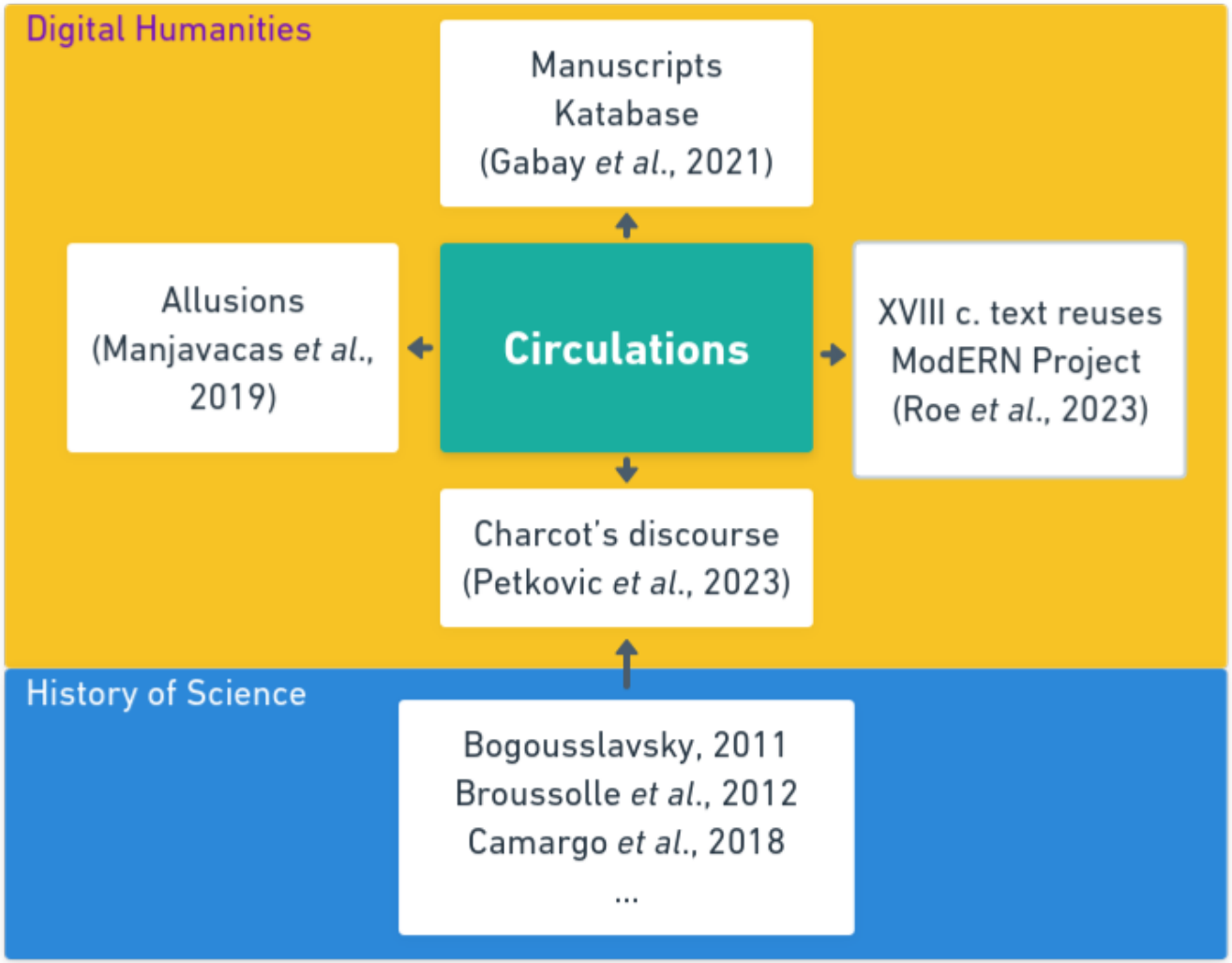

Figure 1: Digital Studies of the Concept of Circulations.

Research Question

How to measure the degree of intertextuality between Charcot and his scientific and/or artistic network through the prism of Digital Humanities?

Charcot Collection

SorbonNum[1] – The Library of Sorbonne University

201 OCRed documents (without post-correction)

- "Charcot" : texts written by Charcot
- "Others" : texts written by a person from his scientific network

Corpus	# of docs	# of tokens
"Charcot"	68	12 190 649 (38,12%)
"Others"	133	19 788 830 (61,88%)
Total	**201**	**31 979 479** (100%)

Table 1: Corpus Distribution based on the Charcot Collection[2].

[1] https://patrimoine.sorbonne-universite.fr/

[2] https://patrimoine.sorbonne-universite.fr/collection/Fonds-Charcot

Measuring the Degree of Intertextuality

Computationally measuring the impact of Charcot on his network
→ uni-directional intertextuality

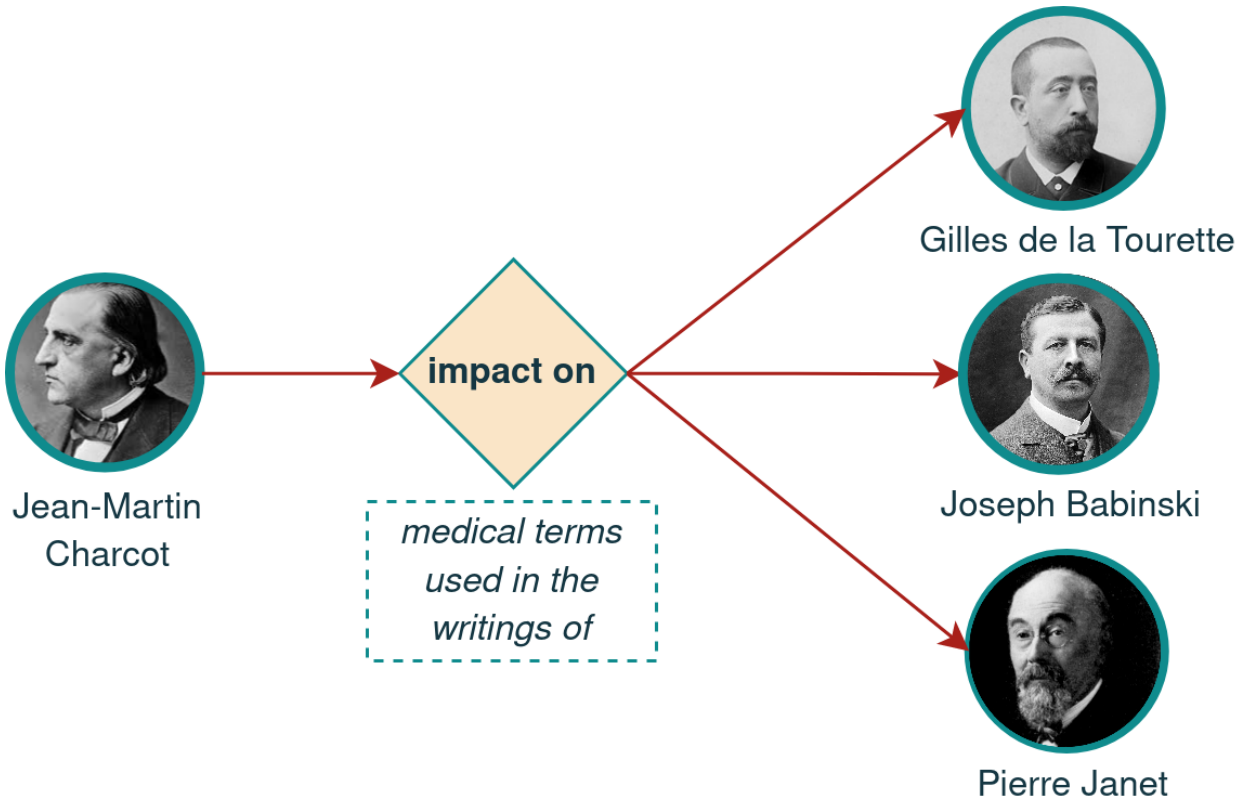

Figure 2: Operationalization of Charcot's Impact on his Disciples.

First Analysis

OBVIE[3]

- search engine allowing advanced corpus search (XML-TEI)
- identification of the most important nouns of each corpus
  - raw frequencies, Jaccard, Dice, PPMI, χ^2, G-test measures
- identification of similar texts in order of relevance based on terms in common

[3] https://obtic.huma-num.fr/obvie/

OBVIE – Charcot Corpus[4]

⚠impossible to quantify the relevance of multi-word expressions

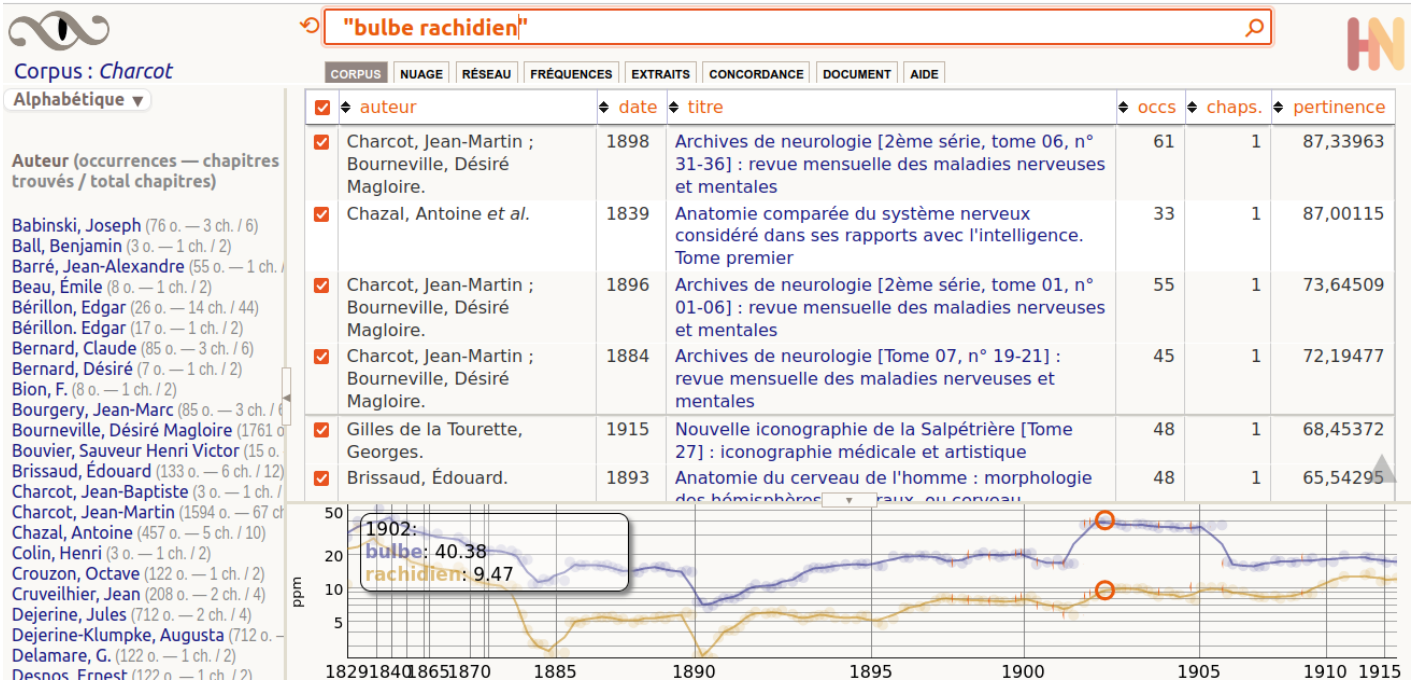

Figure 3: Distribution of Token Frequencies with the Timeline for those Constituting the Expression *bulbe rachidien* (from the "Charcot" and "Others" Corpora).

[4] https://obtic.huma-num.fr/obvie/charcot/?view=corpus

Second Analysis

TextPair[5]

- alignment of similar text sequences in the two corpora
- generates a list of similar passages for each text
- overlapping word sequences (word trigrams)
- compare these results with those of sequences in other texts

[5] https://artfl-project.uchicago.edu/text-pair

Second Analysis – TextPair[6]

⚠retrieving an important number of results – filtering required

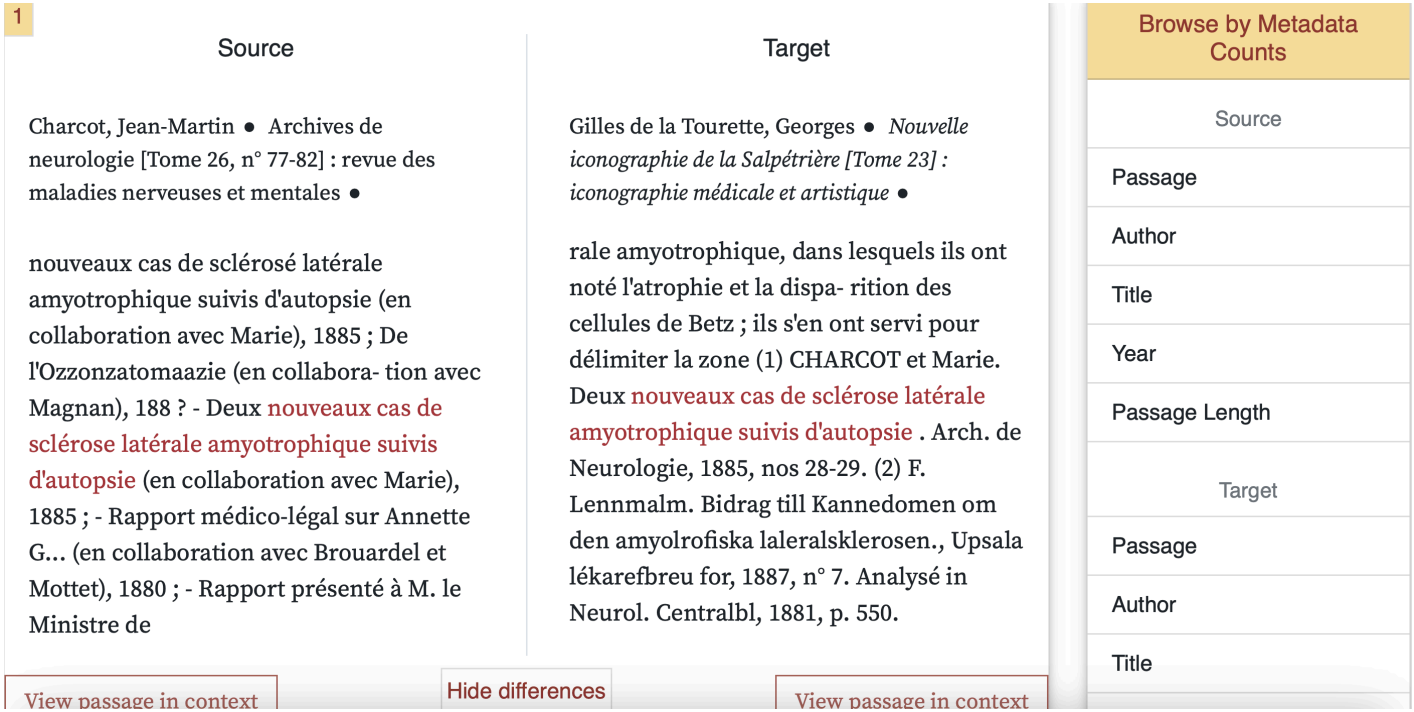

Figure 4: Alignment and Comparing the Charcot's Texts with those of Georges Gilles de la Tourette (the Only Result) by Launching the Query *sclérose latérale amyotrophique*.

[6] https://anomander.uchicago.edu/text-pair/charcot2autres/

Our Approach

Identification of medical terms in the two corpora based on the weight of their appearance

Weighting measures		
TF-IDF	BM25	BERT
• evaluates the importance of a term contained in a document relating to a bigger corpus • rewards the frequency of terms and penalizes the frequency of documents	• TF-IDF's improvement • handles long documents and term saturation issues	• pre-trained model on large corpora (unsupervised learning, Transformers architecture) • learns words and sentence representations (capturing context + semantics)

List of Medical Terms

Extraction of terms or expressions popularized by Charcot (*hystérie*, *sclérose latérale* etc.)

- index of an edition of the complete works of Charcot[7]
- without the generic terms (*os*, *cerveau*, *etc.*)
- taking into account the sg. and pl. forms (regex)

[7] Charcot 1892

Intensification of the Charcot's Vocabulary in the "Others" Corpus

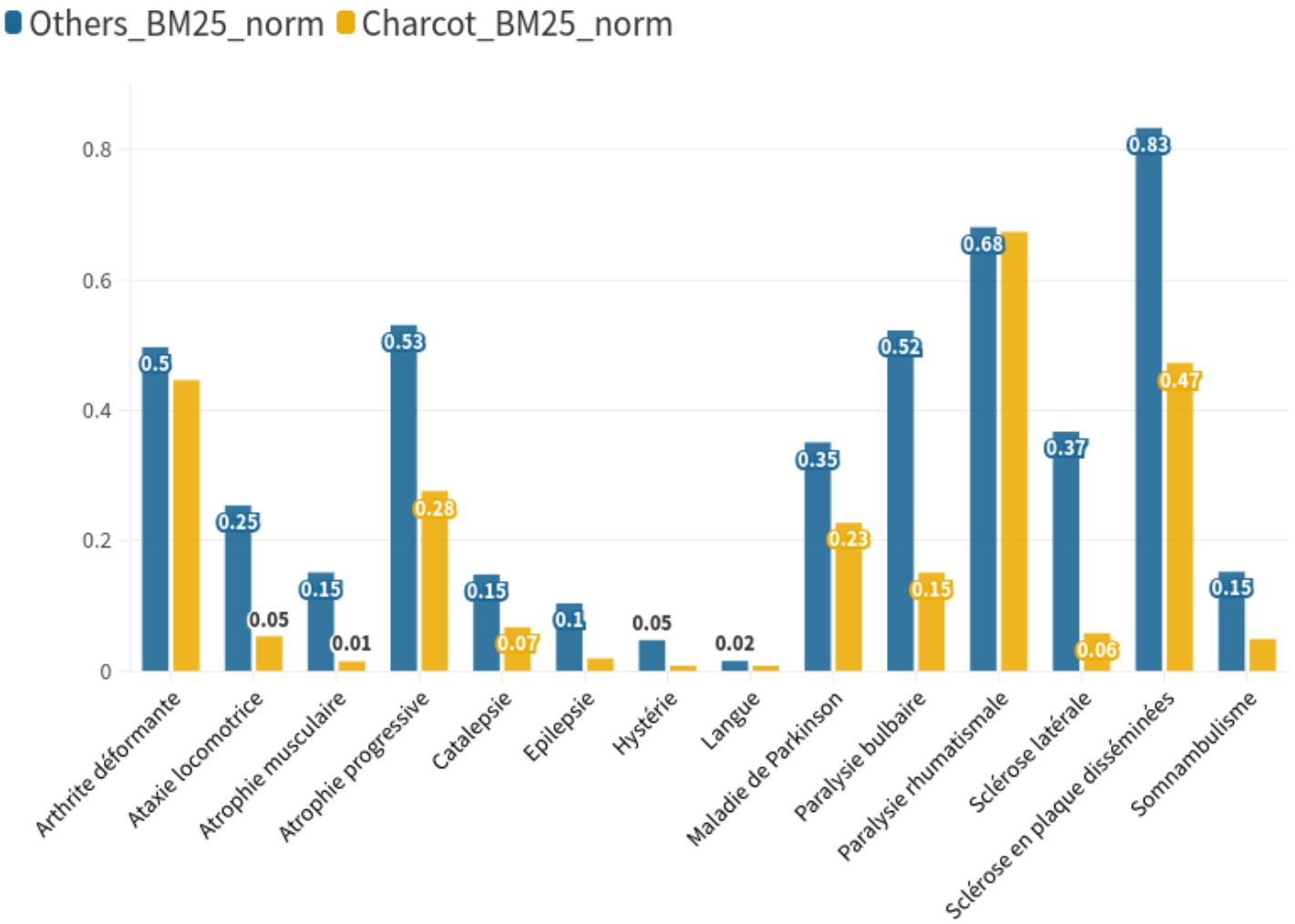

Figure 5: Terms Relevance in the Two corpora (BM25).

Experiments with BERT

Vaswani et al. 2017

- word embeddings and attention mechanism
- `bert-base-multilingual-cased` model

"Charcot" Corpus	"Others" Corpus
diplopie (0,92)	préambule (0,47)
myélite partielle (0,91)	délire (0,47)
état de mal épileptique (0,91)	miracle (0,47)
paralysie labio-glosso-laryngée (0,91)	cicatrices vicieuses (0,46)
PATHOLOGIES	ABSTRACT NOTIONS

Calculation of Terms Relevance – "Charcot" Corpus

Term	"Charcot" Corpus			
	Frequency	TF-IDF	BM25	BERT
Arthrite déformante	30	0,16	**0,45**	0,80
Ataxie locomotrice	559	0,35	0,05	0,83
Atrophie musculaire	1105	0,20	0,02	0,84
Atrophie progressive	40	0,14	0,27	0,72
Catalepsie	681	0,54	0,07	0,88
Épilepsie	414	0,09	0,02	0,78
Hystérie	5775	0,51	0,01	0,74
Langue	2695	0,24	0,01	0,72
Maladie de Parkinson	75	0,21	0,23	0,81
Paralysie bulbaire	149	0,27	0,15	0,89
Paralysie rhumatismale	8	0,07	**0,67**	0,86
Sclérose latérale	445	0,30	0,06	0,88
Sclérose en plaque disséminées	45	0,25	**0,47**	0,87
Somnambulisme	847	0,49	0,05	0,89

Table 2: Calculation of Terms Relevance According to the TF-IDF, BM25 and BERT Measures in the "Charcot" Corpus.

Calculation of Terms Relevance – "Others" Corpus

Terme	"Others" Corpus			
	Fréquence	TF-IDF	BM25	BERT
Arthrite déformante	24	0,02	**0,50**	0,40
Ataxie locomotrice	169	0,08	0,25	0,39
Atrophie musculaire	1465	0,43	0,15	0,42
Atrophie progressive	22	0,02	0,53	0,39
Catalepsie	975	0,28	0,15	0,39
Épilepsie	577	0,12	0,10	0,41
Hystérie	4934	0,45	0,05	0,41
Langue	3591	0,11	0,02	0,41
Maladie de Parkinson	130	0,09	0,35	0,37
Paralysie bulbaire	93	0,09	0,52	0,40
Paralysie rhumatismale	14	0,02	**0,68**	0,44
Sclérose latérale	127	0,09	0,37	0,41
Sclérose en plaque disséminées	12	0,02	**0,83**	0,40
Somnambulisme	3410	1	0,15	0,43

Table 3: Calculation of Terms Relevance According to the TF-IDF, BM25 and BERT Measures in the "Others" Corpus.

Towards a (more) distant reading of the Charcot corpus

First explorations of the Charcot corpus

- advanced search and text alignment → computer-assisted text analysis (OBVIE, TextPair)
- need to measure the impact of Charcot on his network *via* the main medical concepts of his work → distant reading

A Novel Approach

- quantification of the relevance of polylexical concepts in the corpora, according to three different weighting metrics
- identification of lexical phenomena thanks to visualizations (validation of specialists of Charcot's work required)

Perspectives

Future research

1. Charcot vs. Others : initiator or transmitter of certain terms?
2. semantic analysis of passages containing these concepts → enunciative modalities
   - opinions, agreements, disagreements, definitions, etc.
3. OCR post-correction (deep learning) and evaluation of its impact on downstream tasks
4. dynamic topic modeling[8] in order to trace the diachronic evolution of Charcot's terms

[8] *cf.* Blei and Lafferty 2006.

Data and scripts

GitHub repo :
https://github.com/ljpetkovic/Charcot_circulations

Acknowledgments

Many thanks to Valentina Fedchenko (research engineer of the ObTIC project team) and Simon Gabay (assistant professor of the Chair of Digital Humanities at University of Geneva, Switzerland) for their valuable advice.

References I


Blei, David M. and John D. Lafferty (2006). "Dynamic Topic Models". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, pp. 113–120. ISBN: 1595933832. DOI: 10.1145/1143844.1143859.


Bogousslavsky, Julien (Oct. 2010). "Hysteria after Charcot: Back to the Future". In: *Following Charcot: A Forgotten History of Neurology and Psychiatry*. S.Karger AG. ISBN: 978-3-8055-9556-8. DOI: 10.1159/000321783.


Camargo, Carlos Henrique et al. (Oct. 2018). "Jean-Martin Charcot's influence on Career of Sigmund Freud, and the Influence of this Meeting for the Brazilian Medicine". In: *Revista Brasileira de Neurologia* 54, pp. 40–46. URL: https://docs.bvsalud.org/biblioref/2018/07/907032/revista542v4-artigo6.pdf.

References II


Charcot, Jean-Martin (1892). *Œuvres complètes de J.-M. Charcot : Leçons sur les maladies du système nerveux*. Vol. 1. Paris: Bureaux du Progrès médical. URL: https://patrimoine.sorbonne-universite.fr/viewer/3468/?offset=1#page=2&viewer=picture&o=&n=0&q=.


Gabay, Simon et al. (2021). "Katabase: À la recherche des manuscrits vendus". In: *Humanistica 2021*. URL: https://hal.science/hal-03066108/document.


Gomes, Marleide da Mota and Eliasz Engelhardt (2013). "Jean-Martin Charcot, father of modern neurology: an homage 120 years after his death". In: *Arquivos de neuro-psiquiatria* 71. https://doi.org/10.1590/0004-282X20130128, pp. 815–817.

References III


Joyeux-Prunel, Béatrice (2019). "Visual Contagions, the Art Historian, and the Digital Strategies to Work on Them". In: *Artl@s Bulletin* 8.3, p. 8. URL: https://docs.lib.purdue.edu/artlas/vol8/iss3/8/.


Koehler, Peter J. (2013). "Charcot, La Salpêtrière, and Hysteria as Represented in European Literature". In: *Progress in Brain Research* 206, pp. 93–122. DOI: 10.1016/B978-0-444-63364-4.00023-5.


Manjavacas, Enrique, Brian Long, and Mike Kestemont (June 2019). "On the Feasibility of Automated Detection of Allusive Text Reuse". In: *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Minneapolis, USA: Association for Computational Linguistics, pp. 104–114. DOI: 10.18653/v1/W19-2514.

References IV


Marmion, Jean-François (2015). *Freud et la psychanalyse*. Sciences Humaines. URL: https://www.cairn.info/freud-et-la-psychanalyse--9782361063542-page-22.htm.


Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *CoRR* abs/1706.03762. arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.


White, Michel B (1997). "Jean-Martin Charcot's contributions to the interface between neurology and psychiatry". In: *Canadian journal of neurological sciences* 24.3. https://doi.org/10.1017/S0317167100021909, pp. 254–260.