Remerciements

...

Résumé

Ce projet de thèse propose une étude interdisciplinaire centrée sur la valorisation du fonds patrimonial de Jean-Martin Charcot, fondateur de la neurologie moderne au XIX^{e} siècle en France, au prisme des humanités numériques (HN) et du traitement automatique des langues (TAL). Plus concrètement, cette recherche se concentre sur l'exploration de la circulation des savoirs, à travers les reprises du discours scientifique de Charcot sous forme de concepts médicaux dans les écrits d'autres scientifiques. Le présent mémoire vise également à approfondir les recherches issues du travail de Petkovic *et al.* (2023) s'inscrivant dans l'optique de l'exploration quantitative de ce type de circulation.

Au-delà du cas de Charcot, ce travail vise à établir un protocole permettant d'appréhender la circulation de concepts de manière automatisée. ======= Ce projet de thèse propose une étude interdisciplinaire centrée sur la valorisation du fonds patrimonial de Jean-Martin Charcot, fondateur de la neurologie moderne au XIX^{e} siècle en France, au prisme des humanités numériques (HN) et du traitement automatique des langues (TAL). Plus concrètement, cette recherche se concentre sur l'exploration de la circulation des savoirs, à travers les reprises du discours scientifique de Charcot sous forme de concepts médicaux dans les écrits d'autres scientifiques. Le présent mémoire vise également à approfondir les recherches issues du travail de Petkovic *et al.* (2023) s'inscrivant dans l'optique de l'exploration quantitative de ce type de circulation. Au-delà du cas de Charcot, ce travail vise à établir un protocole permettant d'appréhender la circulation de concepts de manière automatisée.

Mots-clés : Jean-Martin Charcot ; humanités numériques ; traitement automatique des langues ; champs lexicaux.

Table des matières

Chapitre 1 Introduction

L'intérêt pour ce parcours doctoral s'ancre dans les expériences de l'autrice en valorisation numérique de ressources textuelles variées, menées lors du master en « Informatique pour les sciences humaines » [1] à l'université de Belgrade et du certificat de spécialisation en linguistique [2] à Genève. Ces travaux ont porté sur des corpus tels que les paroles de chansons rock d'ex-Yougoslavie (Petkovic, 2019), les manuscrits de M^me^ de Sévigné (Gabay *et al.*, 2020, 2021) [3], ou encore les catalogues d'expositions [4]. À ces projets s'ajoute le stage effectué au sein de l'équipe-projet Observatoire des Textes, des Idées et des Corpus (ObTIC) [5] de Sorbonne Université entre le 29 mars 2021 et le 31 juillet 2021, sous l'encadrement de Prof. D^r^ Glenn Roe et D^r^ Motasem Alrahabi, ingénieur de recherche. Dans le cadre du projet de la Très Grande Bibliothèque (TGB) [6], l'enjeu de ce stage a été d'exploiter le corpus constitué d'un grand volume de documents XML-TEI en français, océrisés [7] (transcrits automatiquement) et non corrigés, issus des collections Gallica de la Bibliothèque nationale de France (BnF). À l'issue du stage, le projet de recherche doctoral a été mis en place lors de la campagne d'attribution de contrats doctoraux 2021 par l'institut Observatoire des patrimoines de l'Alliance Sorbonne Université (OPUS) [8]. Ce projet a pour objectif de répondre aux enjeux portés par l'Institut, orientés vers la valorisation du fonds Charcot, tout en s'inscrivant dans les dynamiques de la communauté des HN, dont les travaux portent sur les objets patrimoniaux sous toutes leurs formes : (im)matériels, culturels ou naturels. Au carrefour de l'histoire des sciences et de la linguistique computationnelle, ce projet fait également partie des travaux de l'axe #3 qui sont menés par l'équipe-projet ObTIC et qui sont tournés vers le dialogue constant entre les recherches qualitatives et quantitatives.

La motivation à poursuivre ce travail de recherche découle d'une part de l'abondance

1. http://arhiva.rect.bg.ac.rs/en/education/interdisciplinary/computing.php
2. https://www.unige.ch/lettres/linguistique/program/postgrade
3. Projet *Katabase* : https://katabase.huma-num.fr/
4. Projet *Virtual Contagions* : https://www.unige.ch/visualcontagions/
5. https://obtic.sorbonne-universite.fr/presentation/.
6. http://obvil.lip6.fr/tgb
7. Forme francisée dérivée de l'abréviation anglaise OCR pour *optical character recognition*, soit « reconnaissance optique de caractères ».
8. https://institut-opus.sorbonne-universite.fr/node/478

des études consacrées à la valorisation des archives patrimoniales et aux circulations culturelles (voir section 6). Elle s'appuie également sur la confusion que suscite le terme *circulation des savoirs*, souvent méconnu des chercheur·se·x·s ne travaillant pas dans le domaine du TAL ou des HN. Cette barrière disciplinaire est apparue de manière particulièrement tangible lors des échanges de l'autrice avec les spécialistes de Charcot en médecine, histoire des sciences et critique littéraire. C'est pourquoi cette thèse entend constituer un pas vers la démystification des méthodes quantitatives, mises au service des recherches en histoire des sciences et plus largement en sciences humaines et sociales (ci-après « SHS »).

1 Objectifs

Ce projet de thèse propose une étude interdisciplinaire centrée sur la valorisation du fonds patrimonial de Jean-Martin Charcot, fondateur de la neurologie moderne au XIX^{e} siècle en France, au prisme des humanités numériques (HN) et du traitement automatique des langues (TAL). Plus concrètement, cette recherche se concentre sur l'exploration de la circulation des savoirs, à travers les reprises du discours scientifique de Charcot sous forme de concepts médicaux dans les écrits d'autres scientifiques. Le présent mémoire vise également à approfondir les recherches issues du travail de Petkovic *et al.* (2023) s'inscrivant dans l'optique de l'exploration quantitative de ce type de circulation. Dans le cadre de cette thèse, nous nous intéressons à la circulation à travers l'analyse de la genèse et de la migration du discours médical – en pathologie anatomique, neurologie et psychologie – dans les écrits co-rédigés par Charcot ainsi que dans ceux de ses disciples, collaborateurs et successeurs constituant son « réseau scientifique ». Si l'importance des contributions scientifiques de Charcot est un sujet largement étudié du point de vue de l'histoire des neurosciences (Bogousslavsky, 2011 ; Broussolle *et al.*, 2012 ; Camargo *et al.*, 2024, parmi de nombreux autres travaux), cet aspect reste inexploré dans une perspective quantitative. Cela n'est pas très étonnant, étant donné que l'étude des textes dans les archives Internet et des données en ligne dans un contexte de circulation des connaissances en général reste un domaine en cours de développement (Milia, 2023). À ce titre, nous nous tâchons à mesurer informatiquement l'impact des travaux de Charcot sur son réseau scientifique. Cette mesure se fonde sur l'analyse des concepts-clés en matière de son discours scientifique, et plus particulièrement sur l'opérationnalisation du terme « influence », définie ici comme une intertextualité unidirectionnelle, allant des écrits de Charcot (ci-après corpus « Charcot ») vers ceux de son réseau scientifique (ci-après corpus « Autres »). Il s'agit donc *in fine* d'aborder computationnellement la question des circulations, non pas des artefacts matériels comme les manuscrits (Gabay *et al.*, 2021) et les images (Joyeux-Prunel, 2019), mais des phénomènes textuels complexes (Manjavacas *et al.*, 2019) ayant une dimension théorique forte.

Dans le cadre de l'analyse numérique de l'impact scientifique de Charcot, nous étudions *in fine* la circulation de ses théories et des concepts médicaux dont il était inventeur (p. ex. *SLA*) et transmetteur (p. ex. *hystérie*) [9]. Cette démarche nous oblige de :

1. formaliser en premier lieu la définition du terme *concept scientifique*, identifiable dans un corpus numérique, tout en prenant en compte les difficultés inhérentes à la définition d'un concept *per se*, ainsi qu'à celle de ses termes apparentés : *idée*, *terme*, *mot* ou *mot-clé* (parties 7 et 8) ;
2. comprendre, conceptualiser et opérationnaliser « comment des concepts, des théories ou des méthodes circulent, s'échangent, s'empruntent, se transfèrent et se transforment dans le passage d'une discipline à une autre », questionnement partagé avec LANDAIS (2014, p. 331) (partie 8).

Le pré-requis pour analyser ce type de circulation est de formaliser les concepts scientifiques identifiables dans notre corpus d'étude. Cela est intrinsèquement lié au double objectif de cette thèse, car nous souhaitons :

- formaliser une approche numérique pour tracer l'évolution des concepts médicaux en général ;
- en prenant comme cas d'étude les archives de Charcot, pister numériquement la circulation de son discours médical dans la communauté scientifique de son époque.

C'est donc de cette problématique que découle la question de savoir s'il est possible de mesurer l'impact de Charcot sur l'histoire des neurosciences à travers les termes scientifiques qu'il a employés et qui ont été repris par son réseau scientifique. Dans la lignée de pensée de MILIA (2023), nous considérons un texte comme point d'entrée pour étudier les tendances de la circulation des idées à l'aide des caractéristiques structurelles spécifiques. En l'occurrence, nous avançons l'hypothèse que certains termes médicaux dont Charcot a été l'inventeur (*sclérose latérale amyotrophique – SLA*) ou le transmetteur (*hystérie*) ont été repris de manière significative dans les écrits de son réseau. Ces termes se déclinent sous forme des unigrammes (mots uniques), mais aussi des expressions à mots multiples (angl. *multi-word expressions*), plus précisément des collocations, « associations conventionnelles de mots, arbitraires et récurrentes, dont les éléments ne sont pas nécessairement contigus et dont la signification est largement transparente » (NERIMA *ET AL.*, 2006, p. 96). La complexité syntaxique de ces termes, qui constituent le champ notionnel médical et potentiellement des savoirs en circulation, peut être résumée ainsi :

2 La complexité du terme « circulation des savoirs »

Les savoirs participent à un double mouvement d'héritage et de transmission. En effet, leur circulation sur le temps long reflète ces dynamiques de transmission, es-

9. Comme déjà expliqué dans la partie 4, Charcot n'a pas inventé ce terme, mais en réinterprété le sens.

Partie(s) du discours	Exemple
NOM	*hystérie*
NOM + ADJECTIF	*ataxie locomotrice*
NOM + ADJECTIF + ADJECTIF	*sclérose latérale amyotrophique*
NOM + PRÉPOSITION + NOM + ADJECTIF	*état de mal hystéro-épileptique*

TABLEAU 1.1 – Exemples des concepts scientifiques avec leurs parties du discours.

sentielles à la formation de courants de pensée, ainsi qu'à l'affirmation d'une identité construite autour d'un savoir partagé (ADELL, 2011, p. 251).

======= L'intérêt de poursuivre ce parcours doctoral puise ses racines dans les participations de l'autrice de ce mémoire aux travaux portant sur l'exploration numérique des ressources textuelles aussi diverses que les paroles de chansons rock d'ex-Yougoslavie (PETKOVIC, 2019), les manuscrits de M^me^ de Sévigné (GABAY *ET AL.*, 2020, 2021) ou les catalogues d'expositions dans le cadre du projet *Virtual Contagions* [10]. À ces projets s'ajoute également le stage effectué au sein de l'équipe-projet Observatoire des Textes, des Idées et des Corpus (OBTIC) [11] de Sorbonne Université entre le 29 mars 2021 et le 31 juillet 2021, sous l'encadrement de Prof. D^r^ Glenn Roe et D^r^ Motasem Alrahabi, ingénieur de recherche. Dans le cadre du projet de la Très Grande Bibliothèque (TGB), l'enjeu de ce stage a été d'exploiter le corpus constitué d'un grand volume de documents XML-TEI en français, océrisés [12] (transcrits automatiquement) et non corrigés, issus des collections Gallica de la Bibliothèque nationale de France (BNF). À l'issue du stage, le projet de recherche doctoral a été mis en place lors de la campagne d'attribution de contrats doctoraux 2021 par l'institut Observatoire des patrimoines de l'Alliance Sorbonne Université (OPUS) [13]. Ce projet a pour objectif de répondre aux enjeux portés par l'Institut, orientés vers la valorisation du fonds Charcot, tout en s'inscrivant dans les dynamiques de la communauté des HN, dont les travaux portent sur les objets patrimoniaux sous toutes leurs formes : (im)matériels, culturels ou naturels. Au carrefour de l'histoire des sciences et de la linguistique computationnelle, ce projet fait également partie des travaux de l'axe #3 qui sont menés par l'équipe-projet OBTIC et qui sont tournés vers le dialogue constant entre les recherches qualitatives et quantitatives.

Outre l'intérêt porté aux projets en HN et en TAL, la motivation à poursuivre ce travail de recherche découle également de l'abondance des études consacrées à la valorisation des archives patrimoniales et aux circulations culturelles (voir section 6). Néanmoins, l'intention de mener à bien ce projet se renforce face à la confusion que suscite le terme *circulation des savoirs*, souvent méconnu des chercheur·se·s·x ne travaillant pas

10. https://www.unige.ch/visualcontagions/.
11. https://obtic.sorbonne-universite.fr/presentation/.
12. Forme francisée dérivée de l'abréviation anglaise OCR pour *optical character recognition*, soit « reconnaissance optique de caractères ».
13. https://institut-opus.sorbonne-universite.fr/node/478

dans le domaine du TAL ou des HN. Cette barrière disciplinaire est apparue de manière particulièrement tangible lors des échanges de l'autrice avec les spécialistes de Charcot en médecine, histoire des sciences et critique littéraire. C'est pourquoi cette thèse entend constituer un pas vers la démystification des méthodes quantitatives, mises au service des recherches en histoire des sciences et plus largement en sciences humaines et sociales (ci-après « SHS »).

3 Problématique

Les savoirs participent à un double mouvement d'héritage et de transmission. En effet, leur circulation sur le temps long reflète ces dynamiques de transmission, essentielles à la formation de courants de pensée, ainsi qu'à l'affirmation d'une identité construite autour d'un savoir partagé (Adell, 2011, p. 251). Dans le cadre de cette thèse, nous nous intéressons à la circulation à travers l'analyse de la genèse et de la migration du discours médical – en pathologie anatomique, neurologie et psychologie – dans les écrits co-rédigés par Charcot ainsi que dans ceux de ses disciples, collaborateurs et successeurs constituant son « réseau scientifique ». Si l'importance des contributions scientifiques de Charcot est un sujet largement étudié du point de vue de l'histoire des neurosciences (Bogousslavsky, 2011 ; Broussolle *et al.*, 2012 ; Camargo *et al.*, 2024, parmi de nombreux autres travaux), cet aspect reste inexploré dans une perspective quantitative. Cela n'est pas très étonnant, étant donné que l'étude des textes dans les archives Internet et des données en ligne dans un contexte de circulation des connaissances en général reste un domaine en cours de développement (Milia, 2023). À ce titre, nous nous tâchons à mesurer informatiquement l'impact des travaux de Charcot sur son réseau scientifique. Cette mesure se fonde sur l'analyse des concepts-clés en matière de son discours scientifique, et plus particulièrement sur l'opérationnalisation du terme « influence », définie ici comme une intertextualité uni-directionnelle, allant des écrits de Charcot (ci-après corpus « Charcot ») vers ceux de son réseau scientifique (ci-après corpus « Autres »). Il s'agit donc *in fine* d'aborder computationnellement la question des circulations, non pas des artefacts matériels comme les manuscrits (Gabay *et al.*, 2021) et les images (Joyeux-Prunel, 2019), mais des phénomènes textuels complexes (Manjavacas *et al.*, 2019) ayant une dimension théorique forte.

»»»> Stashed changes De nombreux·ses chercheur·se·s·x partagent le point de vue selon lequel la notion de circulation des savoirs constitue un champ de recherche vaste, ainsi qu'un nouveau paradigme de la connaissance depuis le début du XXIe siècle et l'avènement du Web 2.0 (Landais, 2014 ; Quet, 2014). Cette phase de l'évolution du Web se caractérisait notamment par la transformation majeure de l'Internet en vue du développement des réseaux sociaux, des blogs et des sites participatifs, tout en permettant aux utilisateur·trice·s·x de créer, partager et interagir avec du contenu Web. Nous traversons actuellement l'ère du Web 3.0, né dans les années 2010 et appelé également « Web séman-

tique », qui permet de lier et structurer l'information afin d'en extraire la connaissance (ANDRADE 2013, p. 107). Néanmoins, en se référant à la circulation des savoirs, LANDAIS (2014, p. 331) remarque que ce phénomène connaît une croissance importante grâce aux outils de la numérisation de la production scientifique et de l'édition numérique des ouvrages.

Le terme en question reste toutefois assez complexe en raison de visions différentes sur la façon de le définir. Afin d'éclairer cette problématique, QUET (2014, pp. 221–222) souligne trois aspects suivants :

1. **Éléments de la circulation**. Qu'est-ce qui circule ?
   - individus (savants, techniciens, traducteurs, etc.) ;
   - objets matériels (instruments scientifiques, ouvrages etc.) :
   - constructions symboliques (théories, concepts etc.).
2. **Conceptions de la circulation et méthodes de son analyse** ;
   - définition de la circulation comme « traduction », « diffusion », « accès » ou « succès » ;
   - critères méthodologiques possibles pour étudier la circulation p. ex. d'une théorie :
   – circulations géographiques des principaux concepteurs qu'on lui reconnaît ;
   – circulations et lectures des textes produits par leurs concepteurs ;
   – usages et applications analogiques qui en sont faits dans d'autres domaines.
   - enjeux d'articulation de ces différents niveaux d'observation du point de vue méthodologique et de celui de la production du texte de recherche, dans le cas des croisements de ces niveaux.
3. **Conceptions analytiques et normatives des savoirs**
   - affaiblissement des catégories des « savoirs profanes » et « savoirs scientifiques », ainsi que de l'opposition entre eux ;
   - revalorisation des savoirs implicites et de la dimension pratique des connaissances ;
   - glorification de la circulation comme porteuse de valeurs *a priori* positives : confrontation à l'autre, hybridation, production de nouveauté, etc.

Le pré-requis pour analyser ce type de circulation est de formaliser les concepts scientifiques identifiables dans notre corpus d'étude. Cela est intrinsèquement lié au double objectif de cette thèse, car nous souhaitons :

- formaliser une approche numérique pour tracer l'évolution des concepts médicaux en général ;
- en prenant comme cas d'étude les archives de Charcot, pister numériquement la circulation de son discours médical dans la communauté scientifique de son époque.

C'est donc de cette problématique que découle la question de savoir s'il est possible de mesurer l'impact de Charcot sur l'histoire des neurosciences à travers les termes scientifiques qu'il a employés et qui ont été repris par son réseau scientifique. Dans la lignée de pensée de MILIA (2023), nous considérons un texte comme point d'entrée pour étudier les tendances de la circulation des idées à l'aide des caractéristiques structurelles

spécifiques. En l'occurrence, nous avançons l'hypothèse que certains termes médicaux dont Charcot a été l'inventeur (*sclérose latérale amyotrophique – SLA*) ou le transmetteur (*hystérie*) ont été repris de manière significative dans les écrits de son réseau. Ces termes se déclinent sous forme des unigrammes (mots uniques), mais aussi des expressions à mots multiples (angl. *multi-word expressions*), plus précisément des collocations, « associations conventionnelles de mots, arbitraires et récurrentes, dont les éléments ne sont pas nécessairement contigus et dont la signification est largement transparente » (NERIMA *ET AL.*, 2006, p. 96). La complexité syntaxique de ces termes, qui constituent le champ notionnel médical et potentiellement des savoirs en circulation, peut être résumée ainsi :

Partie(s) du discours	Exemple
NOM	*hystérie*
NOM + ADJECTIF	*ataxie locomotrice*
NOM + ADJECTIF + ADJECTIF	*sclérose latérale amyotrophique*
NOM + PRÉPOSITION + NOM + ADJECTIF	*état de mal hystéro-épileptique*

TABLEAU 1.2 – Exemples des concepts scientifiques avec leurs parties du discours.

4 Jean-Martin Charcot : un médecin à l'aube de la neurologie moderne

« Les vraies révolutions sont lentes et elles ne sont jamais sanglantes. »

— ANOUILH (1956)

La science progresse en corrigeant constamment les erreurs, c'est-à-dire que les erreurs précèdent nécessairement l'établissement de la connaissance scientifique. Bien que ce processus de correction des erreurs puisse être observé de manière diachronique, il est de nature circulaire. En outre, si une doctrine devient obsolète avec le temps et l'avènement des technologies avancées permettant de recueillir de nouvelles preuves, une doctrine actuellement en vigueur deviendra tout de même à son tour obsolète à un moment. L'un des exemples le plus connu de l'obsolescence scientifique est sans doute le passage du modèle géocentrique de l'univers, défendu par Aristote et Ptolémée (selon lesquels la Terre est immobile au centre de l'Univers), à la conception héliocentrique de Nicolas Copernic, qui affirmait que la Terre tournait autour du Soleil.

Un tel cycle des observations empiriques peut être bouleversé, selon BACHELARD (1934, p. 26), par la « rupture et non pas continuité entre l'observation et l'expérimentation ». Autrement dit, la rupture épistémologique survient lors d'un renversement fondamental dans la façon d'établir une connaissance dans un domaine particulier. De fait,

ce phénomène caractérise une « révolution scientifique » (KOYRÉ, 1957, p. 2), terme apparenté avec celui du « changement de paradigme », introduit par KUHN (1962, p. 66). D'après ce dernier, les *paradigmes* désignent les « découvertes scientifiques universellement reconnues qui, pour un temps, fournissent à une communauté de chercheur·euse·x·s des problèmes types et des solutions ». En plus des révolutions en matière de changement de théories scientifiques entières, l'histoire de la pensée scientifique connaît également les ruptures au niveau de la construction des concepts scientifiques. Ces concepts sont nés de ce palimpseste des processus permanents de successions et de rectifications des idées (ASTOLFI *ET AL.*, 2008, p. 24).

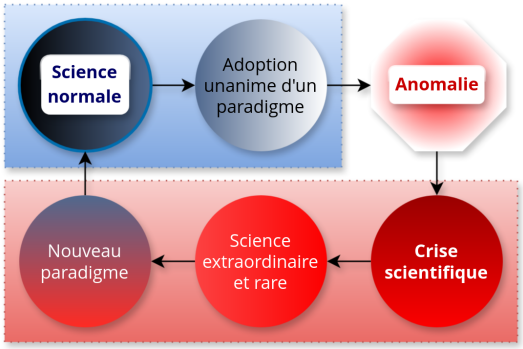

FIGURE 1.1 – Conception kuhnienne du progrès scientifique, adaptée de AMIRI (2012).

Dans cette optique, la structure des révolutions scientifiques désigne un modèle épistémique constitué des épisodes non cumulatifs du développement scientifique (Figure 1.1), marqués par des passages radicaux d'un paradigme à un autre. Le nouveau paradigme ne désigne donc pas une extension de l'ancien paradigme ; au contraire, ce dernier est entièrement ou partiellement remplacé par un nouveau paradigme incompatible avec le précédent.

Cela est bel et bien un signe de l'*émergence* d'une nouvelle théorie ou découverte, tout en prouvant que le développement historique des théories est fondamentalement discontinu. Dans un esprit similaire, BACHELARD (1970, p. 72) souligne :

> « Il ne saurait y avoir de vérité *première*. Il n'y a que des erreurs *premières*. On ne doit donc pas hésiter à inscrire à l'actif du sujet son expérience essentiellement malheureuse. La première et la plus essentielle fonction de l'activité du sujet est de se tromper. Plus complexe sera son erreur, plus riche sera son expérience. L'expérience est très précisément le souvenir des erreurs rectifiées. L'être pur est l'être détrompé. »

Un exemple du changement de paradigme est l'évolution du terme *hystérie*, introduit par Hippocrate dans l'Antiquité au V^e^ s. av. J.-C., qui expliquait cette maladie par un déplacement de l'utérus dans le corps féminin [14]. Au Moyen Âge, surtout à partir du XIII^e^ s., les *hystériques* étaient considérées par l'Église comme possédées par le diable et, par conséquent, chassées, torturées ou soumises aux exorcismes dans une perspective religieuse (TASCA *ET AL.*, 2012, p. 113). Néanmoins, certains scientifiques de la Renaissance commencent progressivement à s'éloigner de l'étiologie démonologique de cette

14. Ce terme est issu du mot grec ὑστέρα, par le latin *hystera*, « matrice ». Par dérivation, le terme « hystérique » se référait à une personne « (femme) malade de l'utérus », selon REY (1998, p. 1767).

maladie ; un cas notable est celui du médecin Charles Le Pois (1563-1633), qui fut le premier à désigner le cerveau, et plus précisément, le *sensorium commune* [15], comme siège de la maladie hystérique en 1618 [16], en associant l'hystérie autant aux hommes qu'aux femmes (WRIGHT, 1980, p. 235).

Pour mieux comprendre l'importance de ce changement de pensée radical, il convient également de souligner que notre compréhension actuelle du système nerveux central est basée sur les premières descriptions faites de manière rigoureuse par Constanzo Varolio (1543-1575) au XVI^e^ s. (TUBBS *ET AL.*, 2008, p. 734) [17]. À l'époque des Lumières en Angleterre (fin XVII^e^ – début XVIII^e^ s.), Thomas Willis (1621-1675), créateur du terme *neurologia* [18] en 1664 (MONTEIRO *ET AL.*, 2021, p. 2), maintint et développa cette conception en caractérisant cette maladie comme principalement convulsive en raison des explosions des « esprits animaux » dans le cerveau (WILLIS, 1681, p. 1). Enfin, l'histoire de la neurologie trouve son ancrage à la fin du XIX^e^ siècle dans les travaux de Jean-Martin Charcot (1825-1893). Ce n'est qu'à cette période que la maladie en question a été systématiquement traitée comme un trouble neurologique (TASCA *ET AL.*, 2012, p. 114).

Figure emblématique et directeur de l'illustre École de la Salpêtrière (basée à l'actuelle hôpital de la Pitié-Salpêtrière à Paris), Charcot a laissé une trace indélébile dans le domaine de la neurologie. Il est essentiellement connu pour ses études portant sur les troubles névrotiques, notamment l'hystérie. Selon lui, l'hystérie découle d'une dégénérescence héréditaire du système nerveux, en montrant qu'elle est en fait plus fréquente chez les hommes que chez les femmes (TASCA *ET AL.*, 2012, p. 114). Charcot a été reconnu pour ses travaux de recherche sur l'hypnose qu'il a utilisée afin d'induire l'état modifié de conscience d'un sujet, permettant l'analyse des symptômes hystériques et leur traitement. Son nom est également associé aux descriptions de nombreuses pathologies connues aujourd'hui, comme la *maladie de Parkinson*, la *sclérose en plaques disséminées*, abbr. *SEP* (ou *sclérose multiple*), la *sclérose latérale amyotrophique*, abbr. *SLA* (soit la *maladie de Charcot*, ou *maladie Lou-Gehrig*) etc [19].

Ces explorations des abîmes de l'esprit humain lui ont valu de nombreuses appellations : à part avoir été globalement considéré comme le père de la neurologie française et moderne (TEIVE *ET AL.* 2022, p. 761 ; BROUSSOLLE *ET AL.* 2012, p. 301), d'autres noms symboliques lui ont été associés, notamment « Napoléon des névroses », « Paganini de

15. ce que KANT (1863, p. 452). appelle plus tard « siège commun de la sensibilité » pour désigner l'ensemble des perceptions.

16. LE POIS (1618, p. 101) a noté que les symptômes communément appelés hystériques se référaient à l'épilepsie, mais qu'il était prouvé que l'épilepsie elle-même était une maladie *idiopathique* (existant par elle-même, sans lien avec une autre maladie) de la tête, et non pas provoquée par les troubles de l'utérus ou des intestins.

17. Il s'agit de l'identification et de la description de la structure cérébrale agissant comme un relai entre le cerveau et le cervelet, appelée *pont* (lat. *pons*) par VAROLIO (1573), soit *pont de Varole* (lat. *pons Varolii*), en l'honneur du célèbre anatomiste, qui fut le premier à examiner le cerveau de sa base vers le haut.

18. Terme présent dans WILLIS (1664).

19. Pour un aperçu détaillé des contributions majeures de Charcot dans le domaine de la médecine, voir CAMARGO *ET AL.* (2024, p. 1102).

l'hystérie » (MIRBEAU & MICHEL 1995, p. 124), ou même « César de la Faculté » (CAMARGO *ET AL.*, 2024, p. 1109). Dans la même lignée de pensée, l'École de la Salpêtrière était caractérisée comme la « Mecque de la neurologie » grâce aux activités de Charcot (TEIVE *ET AL.* 2014, p. 637 ; GOETZ 2017, p. 628 ; CAMARGO *ET AL.* 2024, p. 1100). En outre, de nombreuses références à Charcot et des descriptions d'attaques hystériques figurent non seulement dans la littérature médicale, mais aussi dans des romans naturalistes français et européens, notamment en Pays-Bas, Russie, pays scandinaves, Espagne, Italie et Allemagne (KOEHLER, 2013).

Charcot a créé un véritable réseau scientifique autour de soi grâce à ses idées novatrices qui ont eu un grand retentissement parmi ses collaborateurs, élèves et savants polymathes. Parmi eux, nous ne nommons que quelques figures majeures souvent citées dans la littérature (GOMES & ENGELHARDT 2013, p. 816 ; BOGOUSSLAVSKY 2014, p. 55 ; CAMARGO *ET AL.* 2024, p. 1100), notamment :

- Paul Richer (1849-1933), anatomiste, neurologue et sculpteur, qui a résumé les premières études de Charcot sur l'hystérie dans ses *Études cliniques sur l'hystéro-épilepsie ou grande hystérie* ;
- Georges Gilles de la Tourette (1857-1904), psychiatre et neurologue, qui a décrit les symptômes de la *maladie des tics*, renommée *syndrôme de Tourette* en son hommage par Charcot lui-même ;
- Pierre Janet (1839-1916), philosophe, neurologue et psychiatre, concepteur des termes *dissociation* et *sous-conscient* ;
- Désiré Magloire Bourneville (1840-1909), homme politique et neurologue, qui a publié le premier tome de l'ouvrage monumental l'*Iconographie photographique de la Salpêtrière*, dédiée à l'hystérie, sous l'égide de Charcot ;
- Joseph Babinski (1857-1932), neurologue et neurobiologiste, concepteur du terme *pithiatisme*, qui a découvert le réflexe cutané plantaire, appelé également *signe de Babinski*.

L'impact colossal de Charcot sur sa propre discipline se reflète aussi dans le changement d'intérêt radical du célèbre psychanalyste Sigmund Freud (1856-1939), caractérisé par son passage de la neurologie générale à l'hystérie, à l'hypnose et à d'autres troubles psychologiques. En effet, son séjour dans le service de Charcot à Paris en 1885-1886 a donné lieu au développement de la théorie psychanalytique (CAMARGO *ET AL.*, 2018, p. 41). Les concepts modernes du trouble de stress post-traumatique et des troubles somatoformes en psychopathologie du traumatisme puisent également les racines dans l'œuvre de Charcot (WHITE 1997, p. 254). Néanmoins, certains scientifiques ont fortement contesté le raisonnement scientifique de Charcot, comme le neurologue Hippolyte Bernheim (1840-1919) avec l'École de Nancy pendant les années 1880-1890. Cette polémique porte sur la nature de l'hypnose qui, pour Charcot, représentait un état pathologique propre aux hystériques, et non pas un état de sommeil obtenu par suggestion qui est susceptible d'applications thérapeutiques (et donc, applicable à pratiquement

n'importe qui), comme le soutenait BERNHEIM (1891, pp. 130–131).

«««< Updated upstream

5 Structure du mémoire

À METTRE À JOURCe mémoire est structuré en cinq parties principales : après l'introduction, nous esquissons l'évolution des théories scientifiques dans une perspective épistémologique, en prenant comme cas d'étude les contributions majeures de Charcot (chapitre **??**). Ensuite, nous proposons une revue de la littérature portant sur les modalités des circulations des objets patrimoniaux du point de vue numérique (chapitre **??**). Le chapitre 2 donne un aperçu de la constitution du corpus de recherche. Le chapitre 1 présente les premières tentatives de l'analyse computationnelle de l'impact de Charcot sur son réseau scientifique, ainsi que les limites de ces approches, en proposant une nouvelle méthode pour la quantification de la pertinence des expressions polylexicales. Enfin, le chapitre 3 propose une conclusion et des pistes pour des recherches futures.

======= »»»> Stashed changes

6 Circulations des savoirs et valorisation des archives : état de l'art

Incontestablement, l'époque actuelle est profondément marquée par le « déluge des données », phénomène représentatif de la quatrième paradigme de la science, selon Jim Gray (HEY *ET AL.*, 2009, p. 30). Par conséquent, les projets numériques sont aujourd'hui « pilotés par les données » [20] et ceux qui sont centrés sur les explorations des circulations culturelles au prisme du numérique se concrétisent à grande échelle. Sont fortement axés sur cette thématique :

1. certaines chaires universitaires, notamment celle des Humanités numériques à l'université de Genève (JOYEUX-PRUNEL & GABAY, 2022) [21] ;
2. de divers évènements scientifiques, comme la journée d'étude « Circulation des écrits littéraires de la première modernité et humanités numériques » [22], les colloques Humanistica 2023 [23], ACFAS 2023 [24] etc. ;
3. des numéros de certaines revues, par exemple « Circulation des discours dans les récits complotistes », dont les articles portent sur les thématiques aussi diverses

20. Traduction du terme *data-driven* introduit par JOHNS (1991), issu de l'expression *data-driven learning*.
21. *Cf.* les projets de la chaire : https://www.unige.ch/lettres/humanites-numeriques/recherche/projets-de-la-chaire.
22. https://www.fabula.org/actualites/86846/circulation-des-ecrits-litteraires-de-la-premiere-modernite-et-humanites-numeriques.html
23. https://humanistica2023.sciencesconf.org/
24. https://www.crihn.org/nouvelles/2022/12/11/colloque-de-la-transformation-des-sciences-humaines-par-les-humanites-numeriques-acfas-2023/

que les circulations textuelles internationales du discours complotiste des « Illuminati » (CHAUDET, 2022), « conspirationniste » sur Twitter (GIRY & NOUVEL, 2022) ou antiféministe en ligne (MORIN & MÉSANGEAU, 2022).

La question de recherche sous-tendant ce mémoire s'approche tangentiellement des travaux de RIGUET (2018) et de ROE *ET AL.* (2023). Le premier travail porte sur la réception de la pensée scientifique du physiologiste français Claude Bernard dans la critique littéraire, illustrée par l'alignement des textes de Bernard avec des ouvrages de critique littéraire. Le second article porte sur la détection de réemplois textuels à grande échelle et l'analyse de réseaux pour identifier les « influenceurs » dans les ouvrages français du siècle des Lumières.

Pour ce qui est des projets individuels, la question de l'estimation de l'importance d'une entité issue d'un domaine ontologique occupe une place centrale dans le travail de SOULET (2024), ce qui a résulté dans le développement de l'outil de représentation des connaissances Rankingdom [25]. L'un de ses aspects concerne les déclinaisons de la notion d'importance d'une entité résultant aux métriques correspondantes, comme présenté dans le tableau 1.3 : ces métriques sont calculées pour l'entité `Jean-Martin Charcot`.

Métrique	*Définition*	*Exemple*
PORTÉE (POPULARITÉ)	nombre d'assertions décrivant une entité	Charcot est décrit par 546 assertions.
INFLUENCE	nombre d'entités liées à une entité	191 entités liées à Charcot.
À PROPOS	nombre d'entités impactées (œuvres originales, évènements. . .)	56 entités à propos de Charcot.
INDEX *A*	nombre maximum d'entités impactées *a* ayant le comptage « à propos »	13 entités impactées par Charcot ont le comptage « à propos » supérieur à 13.
IMPACT	somme de tous les comptages « à propos » de toutes les entités impactées	L'impact de Charcot est 826.

TABLEAU 1.3 – Aperçu des métriques Rankingdom pour quantifier l'importance de l'entité `Jean-Martin Charcot`.

De plus, des calculs effectués à partir de la portée et de l'influence de Charcot permettent de générer un graphique de « quadrant magique de Gartner » (angl. *Gartner Magic Quadrant*) [26]. Cette représentation sur la figure 1.2 met en valeur quatre types d'entités :

- **acteurs de niche** : entités avec une portée et une influence modestes (p. ex. Pierre Marie) ;
- ***challengers*** : entités ayant une certaine reconnaissance et une influence considérable, mais qui sont de taille mineure, more concentrées, avec une portée plus petite

25. https://rankingdom.org/about.
26. Le nom provient de la société américaine de conseil Gartner qui « publie chaque année les résultats de ses analyses dans plus de 100 secteurs technologiques » (GUEMAS, 2024).

(Charles-Joseph Bouchard) ;

- **visionnaires** : entités avec une grande portée, dont l'influence reste néanmoins limitée et qui recevront plus de reconnaissance ultérieurement (Paul Richer) ;
- ***leaders*** : entités les plus importantes, avec une grande portée, connues à grande échelle et dans plusieurs domaines, tout en étant reconnues comme ayant une grande influence (Charcot).

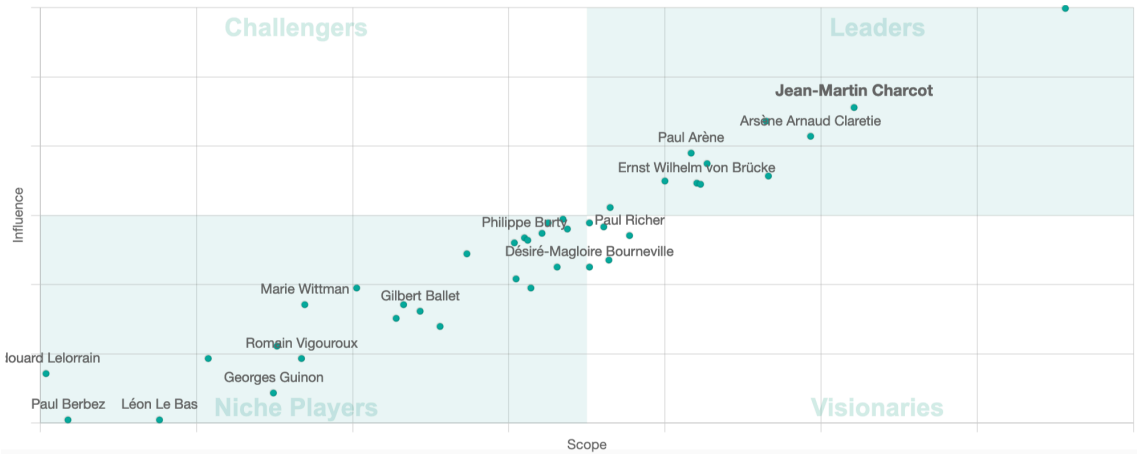

FIGURE 1.2 – Positionnement de l'entité `Jean-Martin Charcot` au sein de son domaine et comparaison avec les entités les plus similaires à lui *via* une analyse de quadrant de l'outil Rankingdom [27].

L'impact de Charcot peut également être visualisé à l'aide de Rankingdom à travers le graphique qui apporte une dimension temporelle (figure 1.3). Il s'agit notamment de la cumulation temporelle de son impact, où l'on peut observer qu'il s'étend sur la période 1856–1994.

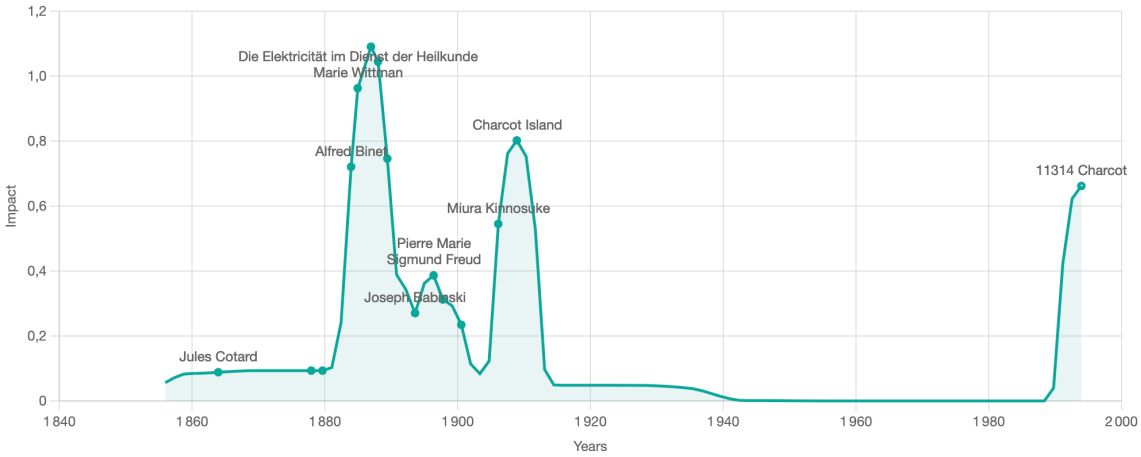

FIGURE 1.3 – Analyse temporelle de l'impact de l'entité `Jean-Martin Charcot` à l'aide de l'outil Rankingdom [28].

27. https://www.rankingdom.org/entity/Q20710?search=jean-martin+charcot. Le domaine dans lequel Charcot figure est relativement large, y compris les figures du domaine médical (p. ex. Bourneville), mais aussi littéraire (Arsène Arnaud Claretie).

28. https://www.rankingdom.org/entity/Q20710?search=jean-martin+charcot.

Enfin, il est également possible de lister les entités impactées, comprenant les personnes (p. ex. Sigmund Freud), les notions médicales (SEP) ou bien les entités géographiques afférentes (Île Charcot).

Type	Titre	

SotA, cadre théorique de la thèse

https://docs.google.com/document/d/1eoW3mDiHYB9vrPtG-5pdPuaUAAUJpWDa/edit#heading=h.gjdgxs

Étant donné l'importance des travaux de Charcot et ses contributions dans le domaine de la neurologie et au-delà, nous souhaitons explorer la notion de la circulation des savoirs au prisme du numérique à travers son impact. Avant d'aborder la question d'opérationnalisation de son impact, nous tenons d'abord à décortiquer les mécanismes à l'origine des circulations des savoirs à grande échelle, ainsi que de définir la notion d'un « concept » pouvant véhiculer les informations importantes concernant les circulations en question.

7 Comment les mots deviennent-ils des concepts ?

Afin de pouvoir analyser les concepts médicaux liés à Charcot, il est important de déterminer de quelle manière un mot ou un groupe de mots devient un concept général ou scientifique. Les termes *idée*, *concept*, *terme*, *mot* et *mot-clé* figurent parmi des notions fondamentales dans les disciplines aussi théoriques (linguistique générale, épistémologie ou philosophie) que numériques ou celles ayant un aspect appliqué, comme p. ex. traitement automatique des langues (TAL) et humanités numériques. Malgré leur présence répandue dans les domaines cités, ainsi que leur utilisation devenue quasi banale dans le langage courant, ces notions demeurent sans définition fixe et universellement acceptée en raison de la disparité des contextes dans lesquels elles sont utilisées. En plus, elles sont interdépendantes et la frontière entre eux est floue.

Concernant la notion du concept, quelques remarques philosophiques de LECOURT (1999, p. 261-263) méritent d'être soulignées ici. Premièrement, l'invention de l'entité du concept remonte à l'ère d'Aristote, qui l'a caractérisé comme une abstraction, un mode de connaissance médiat et général, et comme mode de classification entre le genre et l'espèce (*intension* et *extension*, respectivement). L'intension du concept de chat est sa définition : « animal à quatre pattes de la famille des félins », tandis que son extension est un chat concret : le chat tigré, mon chat etc. Deuxièmement, un concept décrit un sujet, il est définissable et représente un résultat de l'abstraction du donné [29] empirique qui forme une de ses extensions. Cette notion n'est pas à confondre avec celle de l'*idée*,

29. Le concept de « donné » est utilisé en philosophie pour désigner « ce qui est immédiatement présent à l'esprit avant que celui-ci n'y applique ses procédés d'élaboration », http://stella.atilf.fr/.

qui représente elle-même l'objet de connaissance et la condition même du concept, distinction faite de manière systématique chez Kant. Finalement, au-delà des définitions du concept présentées ci-dessus du point de vue phénoménologique à travers l'intension et l'extension, la notion du concept peut également être comprise comme un élément d'un jugement qui peut être une loi scientifique. En d'autres mots, la conception d'un concept inclut non seulement les descriptions d'un sujet en utilisant les prédicats à une place (a.), mais s'étend aussi aux relations *n*-aires (b.) ou même à celles entre des concepts plus abstraits qui impliquent des propriétés allant au-delà des simples prédicats (c.). Cette théorie plus « inférentielle » est à l'origine des concepts scientifiques, dont l'illustration nous retrouvons dans les exemples suivants :

(a.) « le chat est roux » : *le chat* est un sujet (concept) est et *être roux* est un prédicat ;

(b.) « le chat voit un chien » : le sujet *le chat* forme une relation binaire avec un objet *un chien* à l'aide du prédicat *voir* ;

(c.) « Dans un *triangle rectangle*, le *carré* de la *longueur* de l'*hypoténuse* est *égal* à la *somme* des *carrés* des *longueurs* des deux *côtés* de l'*angle droit* » : les concepts mathématiques sont typographiés en italique.

Nous juxtaposons ce point de vue aux réflexions sociohistoriques de STENGERS (1987) qui rendent compte des particularités des concepts scientifiques. D'après elle, l'attribut *scientifique* est associé à leur objectivité et leur puissance explicative, or il n'implique pour autant pas une neutralité d'avis qui est considérée néfaste pour les recherches scientifiques et en même temps fictive. L'autrice renforce cette idée en prétendant que le concept scientifique est forcément controversé, puisqu'il est sujet aux discussions, aux polémiques et aux consensus, ce qui impose une prise de position. Le concept scientifique a des rôles particuliers dans les opérations régissant un champ scientifique, notamment sa singularité, son pouvoir d'extension et d'organisation effective des phénomènes, en s'opposant ainsi à la simple présentation des idées de la part de son·sa émetteur·trice, tout en comprenant un aspect polémique (STENGERS, 1987, pp. 10-11).

À ces traits s'ajoute celui que la même autrice appelle « la propagation épidémique » (p. 16), où les domaines « infectés » par un concept scientifique peuvent être autonomes et devenir une source de nouvelle propagation. Cela est illustré sur l'exemple du concept « programme » en biologie (matériel génétique et sa fonction) qui a migré vers le domaine de l'informatique (opération d'un ordinateur). Les concepts sont donc capables de voyager d'une science à l'autre, ce qui a inspiré la métaphore des « concepts nomades », marqués par leur circulation spatio-temporelle et linguistique. Outre la nature itinérante des concepts scientifiques qui contribue à l'interdisciplinarité et à la production des savoirs nouveaux, STENGERS (1987, pp. 21-23) se réfère aux opérations de la « capture » de la scientificité par ces concepts et du « durcissement » conséquent des sciences. À savoir, certains concepts atteignent le degré de maturité après s'être avéré être adéquats et pertinents dans les démarches scientifiques dont ils « capturent » la scientificité, permettant ainsi que le statut des sciences se solidifie ou « durcisse ». La capture

implique la définition, mais aussi la redéfinition d'une notion par les spécialistes d'une science. Les points de vue de STENGERS (1987) relèvent de la théorie constructiviste du savoir scientifique, selon laquelle la science est une « construction » collective issue du contexte socio-historique (p. ex. interaction entre les scientifiques, les institutions etc.), et non pas d'une accumulation neutre et objective de faits.

Cette approche est complémentaire à l'histoire des concepts (allem. *Begriffsgeschichte*), dans laquelle les significations des concepts en général sont considérées d'être les dérivés d'un contexte sociopolitique. Plus précisément, cette transformation d'un ou plusieurs mots en un concept survient lorsque cette construction linguistique comprend toute la gamme des significations dérivées d'un tel contexte (KOSELLECK & RICHTER, 2011, p. 19). À titre d'exemple, le concept d'un *état* ne peut être interprété qu'à travers ses différents constituants, dont *souveraineté territoriale, législation, fiscalité*, parmi maints d'autres. L'histoire des concepts concerne principalement les manifestations de conflits sociopolitiques particuliers qui doivent être compris dans leur contexte approprié, où p. ex. les mots comme *liberté* ou *démocratie* portent la connotation polémique dont le sens ne peut être précisé qu'à travers leurs antithèses (*esclavage* et *dictature*, respectivement). Les concepts sont donc les concentrations par défaut ambiguës d'une multitude de contenus sémantiques, uniquement interprétables et indéfinissables, par contraste avec des significations des mots qui peuvent être définies de manière exacte (KOSELLECK & RICHTER, 2011, p. 20). De plus, les concepts comme *histoire* ou *progrès* sont caractérisés comme « collectifs singuliers » qui marquent un passage du domain concret d'un individu (plusieurs *histoires* et *progrès* individuels) au domain abstrait et général du collectif social (une *histoire* ou un *progrès* général ou collectif). Ce phénomène linguistique, ainsi que la création des concepts comme *industrie, usine, classe moyenne* etc., reflète un changement de paradigme dans l'organisation sociale survenu lors des révolutions politiques et industrielles (HOBSBAWM, 2010, p. 1). La période charnière concernée par ce phénomène est nommée *Sattelzeit* [30], entre 1750 et 1850, durant laquelle les concepts historiques deviennent abstraits, singularisés, respatialisés et retemporalisés (KOSELLECK & RICHTER, 2011, pp. 34-35). Cela traduit le lien fort entre l'histoire du langage et l'histoire des idées.

Ces considérations sont appliquables à d'autres « concepts nomades » en sciences humaines et sociales (ci-après SHS), comme *travail*, *intelligencija*, *Ancien Régime*, *avant-garde*, *Occident* etc. qui font partie du *Dictionnaire des concepts nomades en sciences humaines* (CHRISTIN, 2011). Plusieurs questionnements ont été soulevés par GHERMANI (2011, p. 117) eu égard de leur émergence, notamment pour déterminer à quel moment un concept devient une entrée dans un dictionnaire des SHS : « *Pourquoi un concept fait-il son entrée dans un dictionnaire ? Au terme de quel processus ? À l'inverse, comment cette percée lexicale est-elle parfois impossible ou refusée ?* ». Contrairement aux processus de la propagation et de la capture qui permettaient à un concept d'obtenir le statut de scientificité, l'autrice sou-

30. Trad. allem. « époque de selle ».

ligne les pratiques scientifiques conduisant aux rétractations et aux masquages de sens des concepts en SHS, p. ex. dans le cas du terme « confession [religieuse] », dont le sens varie en fonction de l'historiographie dans laquelle il figure (GHERMANI, 2011, p. 117). Enfin, BAL (2002, p. 34) va plus loin en excluant la « diffusion » et en mettant en avant la « propagation » comme le critère discriminatoire de la nature itinérante des concepts.

Pour résumer la complexité de la définition des concepts du point de vue de leur histoire, nous citons ici BAL (2002, p. 51), selon laquelle les concepts sont :

- datés, et donc marqués par une évolution ;
- les mots : archaïsmes et néologismes relevant des mécanismes étymologiques qui leur donnent une dimension philosophique ;
- syntaxiques au sein d'une langue ;
- en évolution constante ;
- créés, et non pas donnés *a priori.*

Concernant plus précisément le concept scientifique, l'épistémologie en esquisse les traits suivants, comme souligné par RUMELHARD (1986) et cité dans ASTOLFI *ET AL.* (2008, p. 25) :

- le concept scientifique possède une dénomination et une définition, avec le sens le plus univoque possible, *a contrario* du concept linguistique, en principe équivoque et polysémique ;
- fonction opératoire : le concept scientifique est un outil intellectuel, un instrument théorique permettant d'interpréter des phénomènes ;
- fonction d'opérateur, caractérisé par son degré de formalisation et par les interconnexions avec les techniques scientifiques ;
- une extension, une compréhension, un domaine et des limites de validités en lien étroit avec sa définition fixée ;
- le concept scientifique peut être compris comme un nœud dans un réseau de relations organisé, au sein duquel il dialogue avec d'autres concepts et théories scientifiques.

8 Repérage des termes scientifiques dans un corpus numérique

Si nous nous limitons aux théories abordées jusqu'à maintenant, nous pouvons considérer que les concepts médicaux de Charcot ont eu le rôle des vecteurs de la crise conceptuelle, ce qui représentait une forme de *Sattelzeit* dans le domaine de la médecine. Autrement dit, ces concepts ont été détournés de leurs sens initiaux ayant une apparence formelle neutre (descriptions des pathologies), vers ceux exerçant un certain impact sur la communauté scientifique que nous souhaitons mesurer informatiquement. Néanmoins, l'analyse numérique des concepts n'est pas une tâche triviale non plus, car tous les logiciels ne traitent pas des textes de la même manière. D'après SILBERZTEIN (2022, p. 2), les

logiciels comme TXM [31], Sketch Engine [32] ou IRaMuTeQ [33] traitent les documents comme des *séquences de formes graphiques* (dans notre cas, les séquences « hystérie » et « arthrite déformante » seraient composées d'une et de deux formes graphiques, respectivement). Ces formes sont définies comme les séquences contiguës de caractères alphabétiques délimités par des non-lettres ou les délimiteurs, qui peuvent être considérées comme des informations potentiellement pertinentes pour une étude. D'autres logiciels, comme NooJ [34], peuvent traiter ces séquences comme les *unités linguistiques atomiques*, quel que soit le nombre de formes graphiques (SILBERZTEIN, 2022, pp. 2-3). Ainsi, l'unité linguistique atomique « hystérie » serait recensée dans un dictionnaire des entrées lexicales simples (DELAS), tandis que « arthrite déformante » ferait partie du dictionnaire des entrées lexicales composées (DELAC) [35].

Afin d'extraire automatiquement les concepts scientifiques, nous les opérationnalisons comme des *termes* scientifiques. On en trouve une analogie proche dans la distinction terminologique relevée par SAUSSURE *ET AL.* (1915, pp. 74-75) entre un *signifié* (p. ex. le concept d'un arbre dans notre système cognitif) et un *signifiant* (mot, parole, pictograme désignant un arbre) qui consitue un *signe* (référent, un arbre réel). Les termes sont des expressions textuelles et unités sémantiques qui désignent des concepts dans un domaine d'expertise spécifique. Par conséquent, la tâche d'extraction des concepts peut donc être formalisée comme un problème d'extraction de la terminologie (angl. *automatic text extraction* – *ATE*), dont les enjeux appartiennent au domaine de l'extraction d'information (angl. *information retrieval*), et plus largement, à celui du TAL. L'ATE a pour objectif de faciliter l'identification manuelle des termes à partir de corpus spécifiques à un domaine en fournissant une liste de termes candidats (TRAN *ET AL.*, 2023, p. 1). Jusqu'à maintenant, trois grandes méthodes d'extraction de la terminologie ont été recensées dans la littérature : linguistique, statistique et la méthode basée sur les apprentissages machine et profond (angl. *machine learning* et *deep learning*, respectivement). À FAIRE

- TermoStat [36] (DROUIN, 2003) : termes simples vs. complexes nominaux ; à base de règles ; limite de corpus : 30 Mo + connexion échouée ou phénomène de bottleneck ; extraction des POS
- extraction terminologique `TermSuite` [37] (CRAM & DAILLE, 2016) : scalable, `TreeTagger`
- approche linguistique : besoin d'expert du domaine, analyse syntaxique, POS tagging qui a ses limitations, ne peut pas mesurer la pertinence du terme
- approche statistique, pas besoin d'expert du domaine, mesure de pertinence : *termhood* et *unithood* (KAGEURA & UMINO, 1996, pp. 6-7) vs. fréquence
- approche apprentissage machine / profond : (TRAN *ET AL.*, 2023)

31. https://txm.gitpages.huma-num.fr/textometrie/
32. https://www.sketchengine.eu/
33. http://www.iramuteq.org/
34. https://nooj.univ-fcomte.fr/
35. Ce principe est repris lors du développement du logiciel Unitex https://unitexgramlab.org/fr.
36. https ://termostat.ling.umontreal.ca/doc_termostat/doc_termostat.html
37. https://termsuite.github.io/

- Pour nous, concept scientifique est opérationnalisé comme un terme scientifique.

Comment définir les concepts scientifiques du point de vue du TAL / analyse du corpus ? concepts, termes et mots-clés

Dans le domaine du traitement automatique des langues (TAL), le terme « concept » peut s'apparenter à celui des « entités nommées », comme en témoignent les recherches sur l'extraction automatique de la terminologie biomédicale (Jolly *et al.*, 2024 ; Navarro *et al.*, 2023). Un concept d'un domaine de connaissance peut faire partie d'un thésaurus, liste organisée de termes contrôlés et normalisés, auquel cas le concept est appelé « descripteur ». (Rennesson *et al.*, 2020, p. 16).

Un exemple de ce phénomène est le terme mot, qui véhicule une réalité particulière appartenant à chaque langue (Mounin 1968, p. 65).

Nous n'entendons pas le terme concept dans le sens de Saussure,.... signe = concept (signifié) + image acoustique (signifiant)

Même si l'on reprend la description de Saussure qui considère le mot comme « une image acoustique associé à un concept », nous nous heurtons ensuite au problème de la définition du terme *concept*. Le structuralisme linguistique de Bloomfield souligne ce point, en ajoutant que les linguistes ne sont pas outillés pour démêler complètement ce réseau complexe. Ce structuraliste poursuit en disant que le langage peut en effet être perçu comme une abstraction construite à partir de nos connaissances sur celui-ci, mais qu'il faut « décrire d'abord le fonctionnement de cet instrument de communication » et expliquer comment nous (dé)construisons les énoncés en tant que locuteurs ou auditeurs (Mounin 1968, pp. 94-95).

- ok, et c'est quoi le concept en linguistique (de Saussure) et en analyse du discours
- nous différencions des concepts des « figements linguistiques » (Bezançon & Lejeune, 2023)

Dans le souci de différencier ces notions à travers les disciplines citées, nous présentons ci-dessous quelques-uns de leurs traits discriminatoires qui ne prétendent être ni exhaustifs ni limitatifs :

	Philosophie Épistémologie	Linguistique	TAL
Idée	objet de connaissance (Lecourt, 1999, p. 261)		
Concept	représentation de l'objet de connaissance (Lecourt, 1999, p. 261)		
Signifié	(Astolfi *et al.*, 2008, p. 27)		
Signifiant	mode de représentation des signifiés (Astolfi *et al.*, 2008, p. 27)		
Terme			
Mot			
Mot-clé			

- **Proposer de formaliser la définition du concept (identifiables dans un corpus), mots clés ? Embeddings ? —>**
- nous nous appuyons sur une approximation d'un tel concept, car la tâche d'automatisation et d'implémentation dans l'optique computationnelle enlève forcément quelques traits de concepts abordés dans ce chapitre

Comment suis-je arrivée à ma méthodologie ?

Première partie

Cadre théorique et état de l'art

Chapitre 2 Pister la circulation du discours médical au prisme du numérique

1 Constitution du corpus Charcot

Le fonds patrimonial de Jean-Martin Charcot est conservé à la Bibliothèque de Neurosciences Jean-Martin Charcot par la Bibliothèque numérique patrimoniale de Sorbonne Université (BSU) [1]. Ce fonds regroupe des ouvrages suivants :

- fonds historique Charcot (bibliothèque personnelle de Charcot) : ouvrages, périodiques, collection de thèses et de tirés à part, manuscrits, observations, collection neurologique couvrant la seconde partie du XIX^{e} siècle, fonds bibliophilique ancien ;
- collections de la bibliothèque des Internes de la Salpêtrière : ouvrages, périodiques, thèses en neurologie et psychiatrie pour la période 1800-1950 ;
- donations en ouvrages du docteur Achille Souques.

Dans un souci de préservation d'ouvrages originaux et de valorisation de collections ayant un caractère iconographique notable, une partie de ce fonds a été numérisée. Ces archives numérisées sont disponibles sur le portail numérique SorbonNum [2], porte d'entrée unique vers les collections scientifiques patrimoniales et numériques de Sorbonne Université, ainsi que sur Gallica, bibliothèque numérique de la Bibliothèque nationale de France (BnF) [3].

Le fonds numérisé a été décrit et divisé par la BSU en quatre grandes typologies de documents :

1. Fonds iconographique
   - **Album des internes :** Album des promotions annuelles d'internes, photographiées et classées par établissements de l'Assistance Publique, entre 1860 et 1963 ;

1. https://www.sorbonne-universite.fr/bu/decouvrir-nos-bibliotheques/la-bibliotheque-charcot.
2. anc. Jubilothèque, https://patrimoine.sorbonne-universite.fr/collection/Fonds-Charcot
3. https://gallica.bnf.fr/services/engine/search/sru?operation=searchRetrieve&version=1.2&query=%28gallica%20all%20%22Charcot%2C%20Jean-Martin%22%29&lang=fr&suggest=0.

- **Photographies sur les aliénés de Bicêtre par Désiré Magloire Bourneville** : deux albums présentant les photographies des « petits enfants anormaux » hospitalisés à Bicêtre dans le service du docteur Bourneville, collaborateur de Charcot.

2. Leçons et manuscrits des leçons de Charcot
   - **Manuscrits des leçons et observations de Charcot (1825-1893)** : leçons orales de Charcot, rédigées intégralement de sa main et annotées ;
   - **Leçons de Charcot** : numérisation des volumes de l'*Œuvre Complète* de Charcot consacrés au système nerveux et à l'enseignement clinique, comme par exemple les célèbres leçons du Mardi, sur l'hystérie notamment.

3. Périodiques
   - ***Les Recherches cliniques et thérapeutiques sur l'épilepsie, l'hystérie et l'idiotie (1872-1903)*** de Bourneville. Y est retracée toute l'activité du Service des Enfants Idiots, à la Salpêtrière puis à Bicêtre, par le biais des compte-rendu illustrés de photographies et rédigés par Bourneville ;
   - ***Revue de l'Hypnotisme (1887-1910)*** : périodique consacré à l'hypnotisme que Charcot a réhabilité, publiant les principaux articles théoriques sur cette discipline ;
   - ***Journal du magnétisme (1845-1861)*** : la collection reflète les recherches sur le magnétisme, renouvelées au milieu du XIX^{e} siècle ;
   - ***Revue photographique des hôpitaux de Paris (1869-1872)***. Première revue exposant les applications de la photographie à la médecine, notamment la médecine hospitalière, à travers les études menées à l'Hôpital Saint Louis, et à la Salpêtrière ;
   - ***Iconographie Photographique de la Salpêtrière (1875-1879)***. La collection présente les observations de patientes examinées à la Salpêtrière, accompagnées de photographies d'Albert Londe, présentant les divers stades de la crise d'hystérie ;
   - ***Nouvelle Iconographie de la Salpêtrière (1888-1918)***. La revue est fondée sous la direction de Charcot par Paul Richer, Gilles de la Tourette et Albert Londe, directeur du service photographique. Elle réunit la collection de clichés constituée à la Salpêtrière a pour but la représentation objective des pathologies observées. Elle prend la relève de l'*Iconographie Photographique de la Salpêtrière*. Les articles sont illustrés de photographies, de dessins et de lithographies ;
   - ***Archives de neurologie (1880-1907)***. Sous-titrée « Revue trimestrielle des maladies nerveuses et mentales », les Archives de neurologie sont publiées sous la direction de Charcot par Bourneville. La revue édite, groupe, catégorise et compare la masse des travaux de pathologie nerveuse. Les *Archives de neurologie* sont devenues bisannuelles en 1881.

4. Ouvrages de la bibliothèque de Charcot
   - **Collection d'atlas d'anatomie et de pathologie du système nerveux**, publiés durant le XIX^{e} siècle. L'iconographie de ces ouvrages est remarquable, à commencer par l'*Atlas de Vicq d'Azyr*, médecin du roi Louis XVI ;
   - **Traités**. Cette collection regroupe à la fois des traités sélectionnés dans la biblio-

thèque de Charcot (comme l'*Opera omnia*... de Thomas Willis, 1682, comportant des gravures), des atlas et des textes significatifs des successeurs de Charcot, issus de la bibliothèque des Internes de la Salpêtrière (par exemple l'*Anatomie des centres nerveux* des Déjerine).

Répartition des œuvres sur les années : chronologiquement, cf. le graphique généré au SCAI

Le corpus de travail est constitué de 201 documents OCRisés (sans post-correction), gracieusement fournis au format XML par la BSU. Nous avons procédé, dans un premier temps, à une restructuration des textes en XML-TEI [4] à l'aide de l'outil TEINTE [5], afin de permettre la fouille avancée du corpus Charcot à travers des outils développés au sein de l'équipe-projet OBTIC. D'une part, le moteur de recherche OBVIE [6] permet de repérer des textes similaires par ordre de pertinence à partir des termes en commun. D'autre part, l'algorithme TEXTPAIR génère une liste de passages similaires, c'est-à-dire les séquences de mots qui se chevauchent (n-grammes de mots) pour chaque texte, en comparant ensuite ces résultats avec ceux de séquences dans d'autres textes [7].

Afin de mesurer l'impact de Charcot sur son entourage et d'analyser la circulation de concepts véhiculés dans le corpus, nous avons commencé par séparer les documents rédigés par Charcot de ceux rédigés par ses co-auteurs (p. ex. Bourneville) ou les auteurs thématiquement proches de lui (p. ex. de la Tourette). Nous avons obtenu respectivement 68 (corpus « Charcot ») et 133 (corpus « Autres ») documents, comme présenté dans le tableau 2.1.

Corpus	Nombre de documents	Nombre de tokens	Mémoire
Charcot textes rédigés par Charcot	68	12 190 649 (38,12 %)	79,1 Mo
Autres textes rédigés par les membres de son réseau scientifique	133	19 788 830 (61,88 %)	127,2 Mo
TOTAL	**201**	**31 979 479** (100 %)	**206,3 Mo**

TABLEAU 2.1 – Répartition du fonds Charcot selon les auteurs.

Les deux corpus issus du fonds Charcot sont librement disponibles et interrogeables sur les deux plateformes OBVIE [8] et TEXTPAIR [9].

4. Originalement, ces fichiers ne contenaient que les balises `<doc>`, `<id_doc>` et `<pages>`.
5. https://github.com/OBVIL/teinte_obtic
6. https://obtic.huma-num.fr/obvie/. Pour d'amples informations sur le fonctionnement de cet outil, *cf.* ALRAHABI (2022).
7. https://artfl-project.uchicago.edu/text-pair.
8. https://obtic.huma-num.fr/obvie/charcot/?view=corpus
9. https://anomander.uchicago.edu/

métadonnées sur les deux corpus...

2 Exploration du corpus Charcot : OBVIE et TEXTPAIR

Une première exploration du corpus Charcot à travers l'application OBVIE nous a permis d'identifier les substantifs les plus importants de chaque corpus en utilisant les fréquences brutes ou des méthodes plus fines comme TF-IDF, BM25 (détaillées dans la partie 4), $\chi 2$ ou le TEST GAMMA. Cependant, l'application ne permet pas de quantifier la pertinence des expressions polylexicales, soit les n-grammes de mots, très fréquentes dans les deux corpus et dont la décomposition entraînerait une perte d'information (p. ex. le terme polysémique « bulbe » qui a une valeur spécifique dans l'expression figée *bulbe rachidien*). En observant la figure 2.1, nous constatons que l'abscisse donne l'information sur les dates de publication des ouvrages compris dans les corpus, alors que l'ordonnée indique le nombre d'occurrences par million de mots, soit *parties par million* (*ppm*) [10].

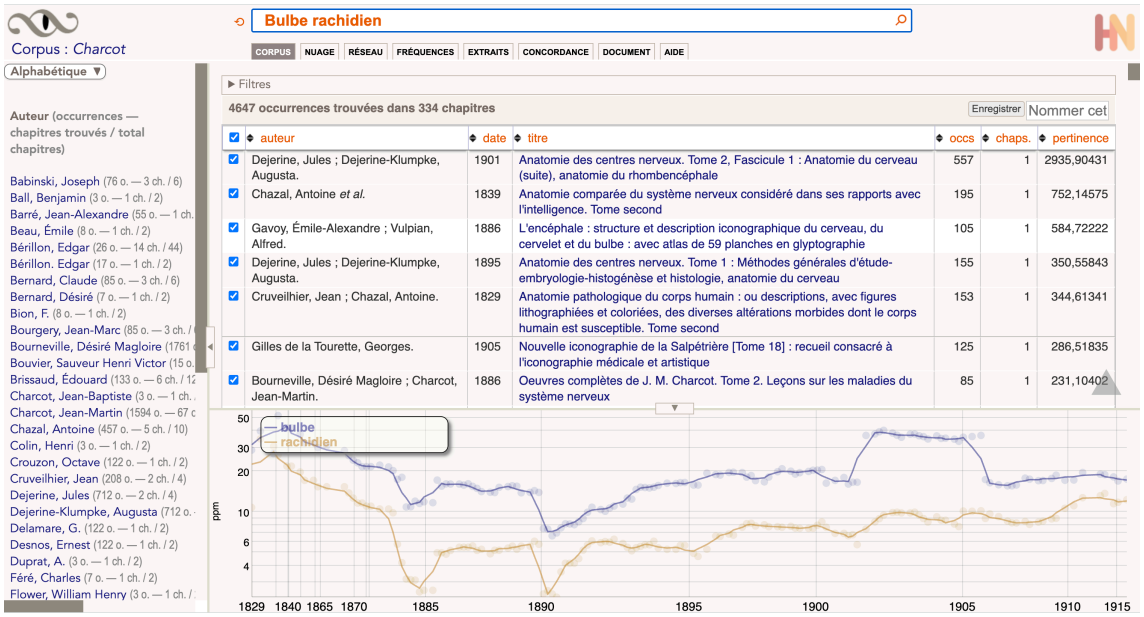

FIGURE 2.1 – Distribution des fréquences des tokens avec la frise chronologique pour ceux constituant l'expression « bulbe rachidien » (issus du corpus « Charcot » et du corpus « Autres ») dans le logiciel OBVIE.

Concernant l'alignement des séquences similaires aux deux corpus, TEXTPAIR nous a permis, par une lecture attentive, de faire des comparaisons entre les textes et de rechercher des termes au sein des passages similaires, malgré le nombre de résultats assez conséquent (*cf.* la figure 2.2). En raison de sa capacité de détecter les passages similaires, notamment les citations directes, les plagiats ou les réemplois, ce logiciel, ainsi qu'un autre logiciel de détection de plagiat, peuvent nous servir de *baseline* pour comparer leurs résultats avec ceux proposés dans la partie 4.

10. *Cf.* le guide d'utilisation d'OBVIE détaillé : https://obtic.huma-num.fr/obvie//static/aide.html.

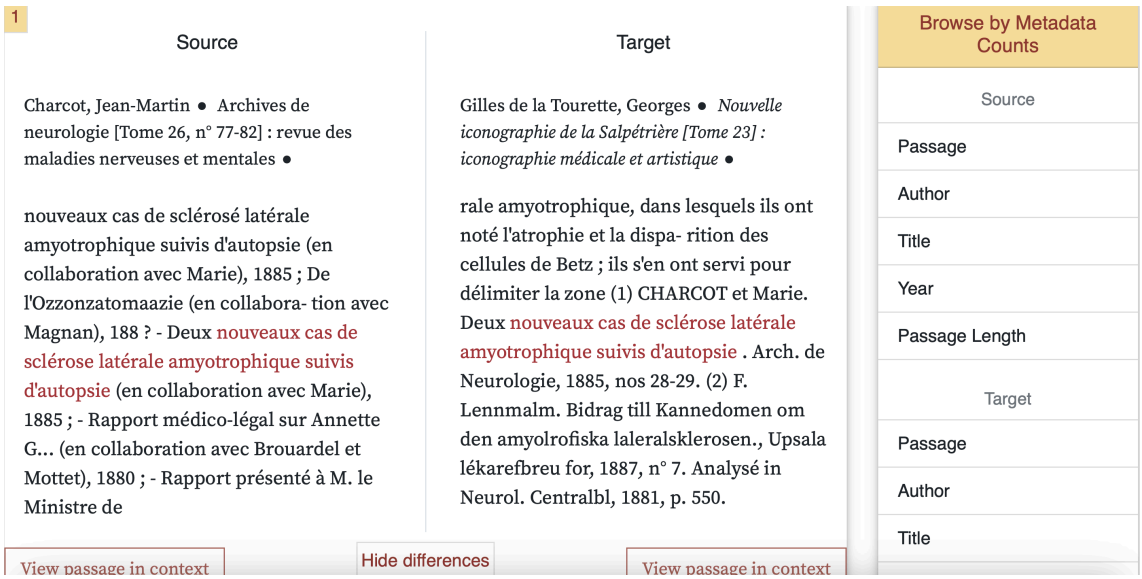

FIGURE 2.2 – Alignement et comparaison d'un texte de Charcot à celui de Georges Gilles de la Tourette (le seul résultat) en lançant la requête *sclérose latérale amyotrophique*.

3 Extraction de la terminologie : approche linguistique

Dans le cadre de l'approche linguistique de l'extraction terminologique, nous avons tenté d'utiliser l'outil `TermoStat`. Bien que le traitement d'un échantillon minuscule des corpus (1-2 documents) ait pu se terminer avec succès, le passage à l'échelle de l'intégralité des corpus n'a pas été possible en raison des limitations citées dans la section 8. Même en respectant la limite du corpus de 30 Mo, le traitement a été extrêmement chronophage, sans pour autant générer aucun résultat après plusieurs jours de calcul. En revanche, l'utilisation de l'outil `TermSuite` s'est avéré comme un moyen alternatif bien plus efficace pour générer les résultats souhaités, étant donné que le traitement de chaque corpus a duré environ une vingtaine de minutes [11]. Nous disposons de deux tableaux issus de l'extraction des termes uniques correspondant aux deux corpus, ainsi que de leurs diverses caractéristiques et mesures statistiques (motifs syntaxiques des parties du discours, fréquences brute et documentaire, TF-IDF, spécificité. . .). Concernant les motifs syntaxiques, nous en avons extrait 6 types, marqués par leurs étiquettes `TreeTagger` comme illustré dans le tableau 2.2.

Le tableau 2.3 montre les motifs syntaxiques extraits, avec leurs fréquences absolues (nombres d'occurrences extraites), leurs fréquences relatives (pourcentages de toutes les étiquettes extraites), avec des exemples des termes représentatifs correspondant à chaque motif. Nous soulignons que 5 motifs sont communs aux deux corpus, hormis celui de `[N A A]`, extrait uniquement à partir du corpus « Charcot ». Toutefois, il est intéressant de noter qu'aucune occurrence du terme très fréquent *sclérose latérale amyotrophique*, ainsi que de ses éléments constitutifs *sclérose* et *amyotrophique*, n'a été extraite ; on n'en retrouve les traces que dans l'adjectif `[A]` extrait *latérale*. Les trigrammes sont

11. Les résultats sont disponibles dans le dépôt GitHub https://github.com/ljpetkovic/Charcot_TermSuite

Étiquette	Signification
A	`adjectif`
N	`nom`
N A	`nom + adjectif`
N A A	`nom + adjectif + adjectif`
N P N	`nom + préposition + nom`
R	`adverbe`

TABLEAU 2.2 – Étiquettes `TreeTagger` extraites avec `TermSuite`, accompagnées de leurs significations.

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

TABLEAU 2.3 – Répartition des parties du discours constituant les termes médicaux dans les corpus « Charcot » et « Autres ».

les séquences les plus longues extraites, notamment 6 occurrences de `[N P N]` (corpus « Autres »), 12 de `[N P N]` et 3 de `[N A A]` (les deux dernières dans le corpus « Charcot »). Les séquences plus longues sont très pertinentes pour l'extraction de la terminologie précise (p. ex. *sclérose* est terme moins précis que *sclérose latérale amyotrophique*). Dans la figure 2.3, nous exposons les répartitions des motifs syntaxiques constituant les termes médicaux extraits en termes de leurs effectifs dans les corpus « Charcot » et « Autres ». Nous nous apercevons que les motifs les plus fréquents sont les unigrammes (mots individuels) de noms `[N]` et les adjectifs `[A]`, alors que les bigrammes et les trigrammes sont présents dans une moindre mesure. Cela laisse à penser que `TermSuite` ne parvient pas à extraire les termes médicaux sous forme de quadrigrammes (séquence de quatre mots consécutifs, p. ex. *sclérose en plaques disséminées*) ou des séquences plus longues (*état de mal hystéro-épileptique*) qui sont bel et bien mentionnées dans les deux corpus. En plus, cette expérience confirme les limites de l'approche linguistique de l'extraction des termes scientifiques à base de règles, notamment à l'aide des expressions régulières et des automates à états finis. En effet, il est bien connu que leur construction est une tâche fastidieuse, restrictive et non maintenable sur le long terme, surtout en cas d'un grand ensemble de termes.

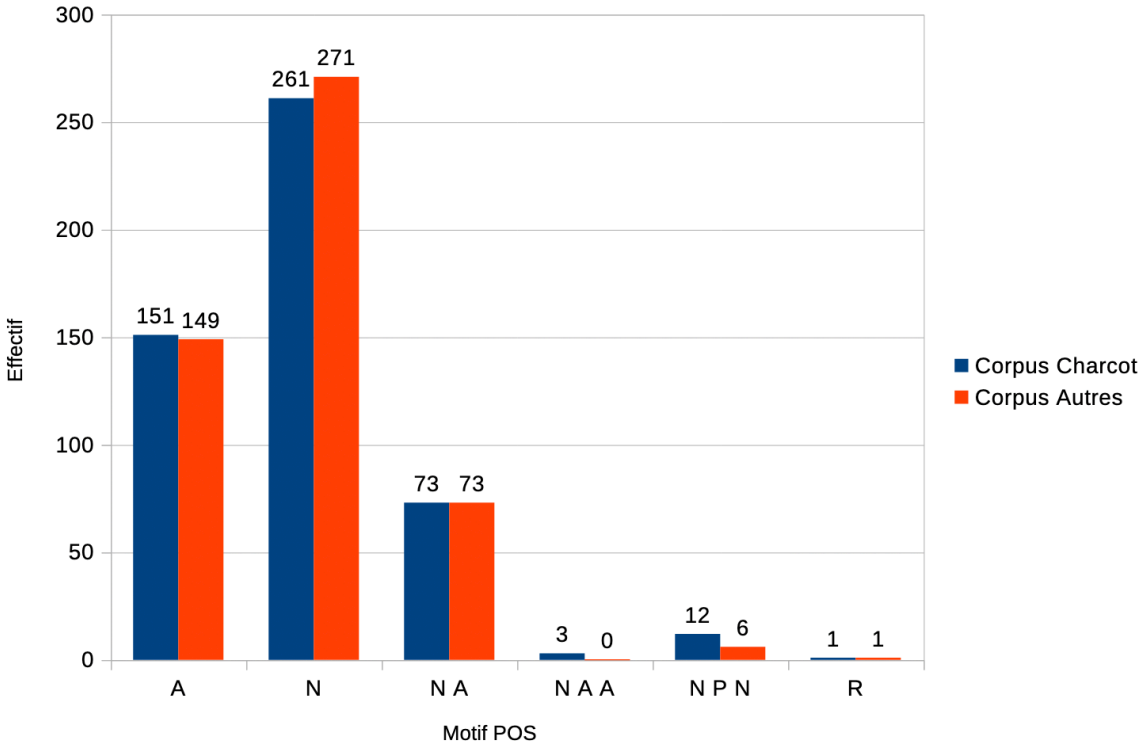

FIGURE 2.3 – Analyse comparative des séquences syntaxiques constituant les termes scientifiques dans le corpus « Charcot » et « Autres ».

Toutefois, `TermSuite` donne la possibilité d'analyser la pertinence des termes à travers les mesures statistiques. Pour ce qui est de celle de TF-IDF, elle mesure la pertinence d'un terme en calculant sa fréquence brute moins l'inverse de sa fréquence documentaire (le nombre de documents contenant ce terme indique sa dispersion). L'intuition derrière cette mesure se résumerait ainsi : si un terme apparaît souvent dans quelques documents, il s'agit d'un terme spécialisé. En revanche, si un terme apparaît souvent dans beaucoup de documents, il sera moins pertinent (un exemple classique de ce phénomène sont les mots vides, p. ex. *par exemple*). Nous remarquons que les bigrammes comme, p. ex. *paralysie générale* $(0,53)$ et *lame médullaire* $(0,51)$ n'ont été sous-valorisés dans aucun corpus par rapport aux unigrammes selon cette mesure, puisqu'ils figurent dans les parties supérieures des listes des termes extraites. Cela ne s'applique pas aux trigrammes comme *troubles de sensibilité* $(0,10)$ et *pli de passage* $(0,14)$. Enfin, nous avons également récupéré les résultats de la spécificité des termes extraits. Cette mesure se réfère au ratio d'étrangeté (angl. *weirdness ratio*) (KHURSHID *ET AL.*, 2000), soit à la « termicité » des termes dans le corpus par rapport au langage général. Elle exprime le degré de relation d'un terme avec un domaine spécifique. Par exemple, le terme *atrophie* a une termicité plus élevée $(4,44)$ que *maladie nerveuses* $(3,15)$ dans le corpus « Charcot » ; le même rapport est observé pour les termes *atrophie* $(4,38)$ et *antécédent personnel* $(3,17)$ dans le corpus « Autres ».

4 Extraction des phrases-clés : méthodes statistiques

Afin de surmonter les limites rencontrées avec ces deux outils, nous avons proposé une nouvelle méthode pour identifier des concepts dans les deux corpus en nous basant sur le poids de leur apparition, calculé selon trois différentes mesures de pondération [12] :

- TF-IDF (Robertson & Jones, 1976) est une méthode qui permet d'évaluer l'importance d'un terme contenu dans un document relativement à un corpus plus large en récompensant la fréquence des termes, sans tenir compte des variations de longueur du document ;
- BM25 est une fonction de classement qui classe un ensemble de documents en fonction des termes de requête apparaissant dans chaque document, quelle que soit l'interrelation entre les termes de requête au sein d'un document (par exemple, leur proximité relative). Il s'agit d'une tentative d'amélioration de TF-IDF, notamment pour prendre en compte divers facteurs tels que la longueur du document et les problèmes engendrés par la possible saturation des termes (Robertson *et al.*, 2009, p. 355) ;
- BERT (Devlin *et al.*, 2019) est un modèle pré-entraîné qui utilise l'apprentissage non-supervisé sur de grandes quantités de données textuelles pour apprendre des représentations de mots et de phrases, et comprendre le contexte et la sémantique. Il est basé sur l'architecture des *transformeurs*, qui est un type de grand modèle de langue utilisé pour le TAL.

La liste des concepts retenus pour l'étude est composée de termes ou expressions popularisés par Charcot, comme *hystérie*, *sclérose latérale* etc. (Camargo *et al.*, 2024, p. 1102) [13]. Pour chaque entrée, nous avons pris en compte les formes du singulier et du pluriel obtenues grâce à des expressions régulières. La liste est produite de façon supervisée et provient du croisement entre la liste des termes obtenus avec OBVIE et l'index d'une édition des œuvres complètes de (Charcot, 1892, pp. 493–507), dont nous avons retiré les termes génériques (*os*, *cerveau*, etc.).

Comme nous pouvons l'observer sur la figure 2.4, la mesure BM25 révèle une intensification du lexique de Charcot dans le corpus « Autres ». Plus précisément, tous les termes évalués sont identifiés comme plus signifiants dans le discours des « Autres » que dans celui de Charcot, les scores étant plus élevés pour 14 termes (sur 14 évalués) utilisés par le réseau de Charcot. D'ailleurs, d'après le tableau **??** (en annexe), c'est la seule mesure dont les valeurs témoignent clairement d'un lexique partagé entre Charcot et ses successeurs et collaborateurs, *a contrario* des deux autres mesures, où le rapport en question est inversé (la grande majorité des termes étant plus pertinente dans le discours de Charcot, et son impact étant donc moins accentué). Concrètement, les termes les plus pertinents semblent être *sclérose en plaque disséminées* (score 0,83), *paralysie rhumatismale* (0,68), *atrophie progressive* (0,53) et *arthrite déformante* (0,50).

12. Le code est disponible en ligne : https://github.com/ljpetkovic/Charcot_circulations.
13. *Cf.* la liste exhaustive des termes et des expressions popularisés par Charcot en annexe.

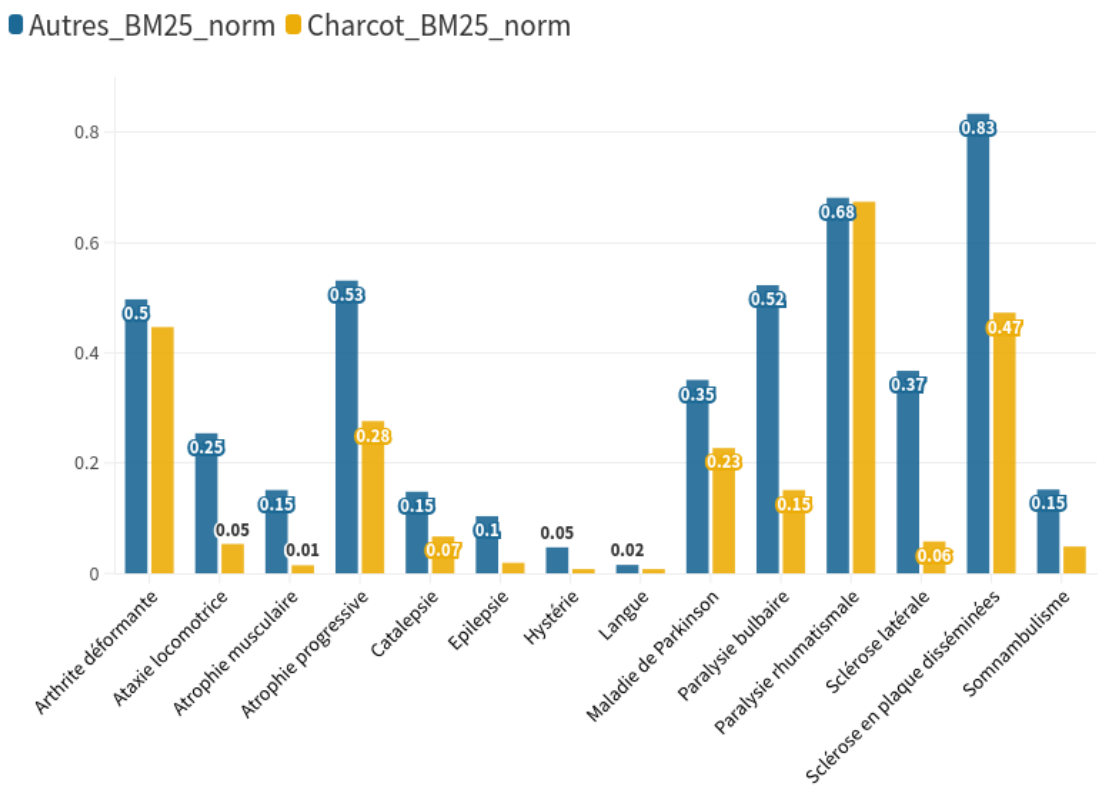

FIGURE 2.4 – Visualisation de pertinence des concepts dans les deux corpus suivant la métrique BM25. Les valeurs des concepts associées au corpus « Autres » sont représentées en bleu, alors que celles du corpus « Charcot » en jaune.

D'autre part, nous avons utilisé BERT pour mesurer le poids des termes dans les deux corpus. Bien que ce type de modèle ne fournisse pas directement de poids pour les mots, nous pourrions cependant en extraire des informations utiles pour estimer l'importance ou le poids des mots dans les textes. Différentes approches sont généralement utilisées pour obtenir une représentation de l'importance des mots, en exploitant les informations des plongements lexicaux et des mécanismes d'attention (VASWANI *ET AL.*, 2023). Pour ce travail en cours, nous avons utilisé le modèle `bert-base-multilingual-cased`. Les premiers résultats obtenus se trouvent dans le tableau **??** et restent à améliorer. Cependant, nous avons observé que les termes les plus pertinents pour le discours de Charcot étaient ceux qui désignent les noms des différentes pathologies (*diplopie*, *myélite partielle*, *état de mal épileptique*, *paralysie labio-glosso-laryngée* etc.), contrairement à d'autres notions plus abstraites (*vicieuses*, *délire*, *miracle*) qui sont prédominantes dans le corpus « Autres » (termes non renseignés dans le tableau en question). La présence de ce dernier type de notion n'est pas étonnant, étant donné que Charcot aborde la question des guérisons miraculeuses dans ses recherches [14].

14. Voir notamment son œuvre *La foi qui guérit* (CHARCOT, 1897).

5 Extraction des phrases-clés : méthode à base d'apprentissage profond

En complément de la méthode du calcul de pertinence des termes médicaux fournis de manière supervisée (partie 4), nous exposons ici des résultats de l'approche non-supervisée pour extraire des mots/phrases-clés pertinents à partir de nos deux corpus [15]. L'objectif de cette approche est de détecter les termes communs entre les deux corpus et de montrer la répartition des termes les plus pertinents dans le réseau de Charcot. Deux algorithmes librement disponibles sont présentés ici pour illustrer cette dernière approche : `keybert` (GROOTENDORST *ET AL.*, 2023) [16] et `keyphrase-vectorizers` [17]. Lors du passage à l'échelle avec la quantité de données considérable (voir le tableau 2.1), nous avons fait face un manque de puissance de calcul des ordinateurs locaux. Pour faciliter l'extraction des phrases-clés, nous avons obtenu l'accès à la plateforme technologique MESU et un accompagnement technique grâce à l'unité de service SACADO (Service d'Aide au Calcul et à l'Analyse de Données) [18] de Sorbonne Université.

5.1 Librairie `keybert`

Cette librairie Python permet d'exploiter les plongements de mots (angl. *word embeddings*) du type BERT pour générer des mots/phrases-clés les plus similaires à un document. La figure 2.5 illustre la chaîne de traitement appliquée à nos deux corpus :

1. les corpus « Charcot » et « Autres » sont utilisés comme les données d'entrée au format `.txt` ;
2. les documents d'entrée ont été tokenisés en phrases-clés candidates avec la fonction `CountVectorizer` ;
3. les plongements des documents et de leurs phrases-clés candidates ont été générés par le modèle de langue `sentence-transformers` ;
4. la similarité cosinus a été calculée entre les documents d'entrée et les phrases-clés candidates, où celles avec les scores les plus élevés sont extraites.

Une première tentative de génération des phrases-clés les plus pertinentes dans les deux corpus n'a produit que deux termes : ARTICULATION DE [*sic*] ÉPAULE et PARALYSIE FACIALE PÉRIPHÉRIQUE. Par ailleurs, en observant les 15 phrases-clés les plus pertinentes dans le corpus « Autres » (figure 2.6), nous constatons un manque de diversification des résultats et des phrases-clés qui se ressemblent (*la sensibilité tactile*, *sensibilité*

15. *Cf.* le dépôt GitHub https://github.com/ljpetkovic/Charcot_KeyBERT_Keyphrase-Vectorizers/.
16. https://maartengr.github.io/KeyBERT/
17. https://pypi.org/project/keyphrase-vectorizers/
18. https://sacado.sorbonne-universite.fr/fr/.
19. Illustration reprise de https://maartengr.github.io/KeyBERT/guides/quickstart.html#installation.

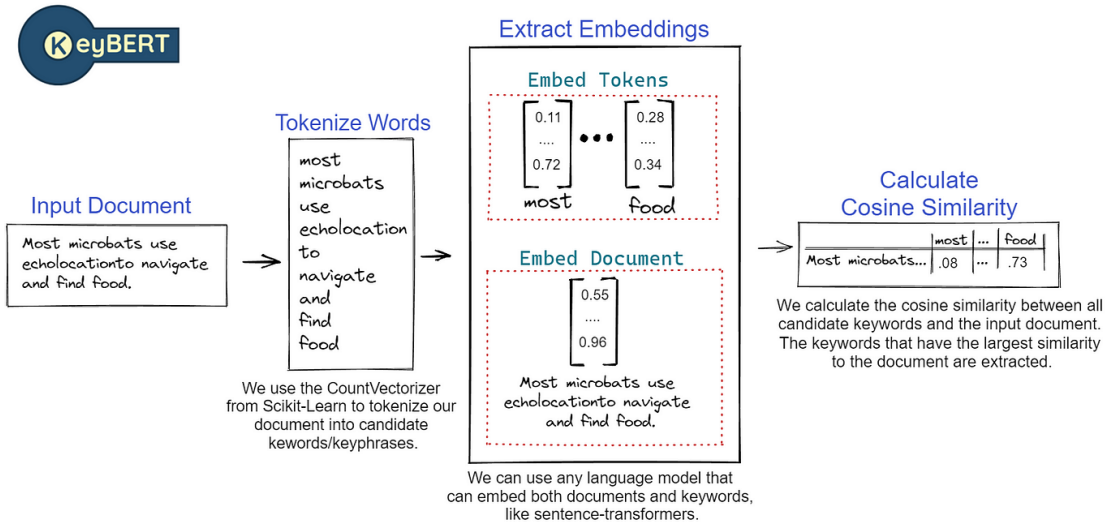

FIGURE 2.5 – *Pipeline* de la librairie `keybert` [19].

tactile au, *la sensibilité tend* etc.) [20]. Un autre problème observé était la non-grammaticalité des phrases-clés extraites (*sémi lunaire segment*, *prière le malade* etc.), ce qui nous a incités à tester une approche plus fine, décrite dans la partie 5.2.

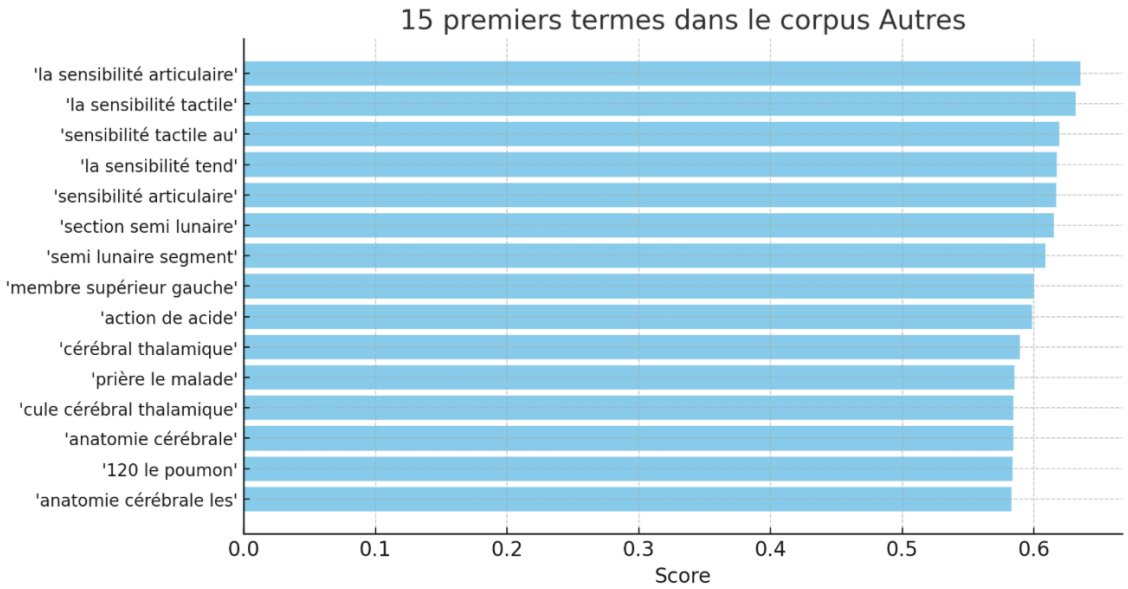

FIGURE 2.6 – Répartition des 15 termes les plus pertinents dans le corpus « Autres » selon `keybert`.

5.2 Approche *PatternRank*

Cette approche exploite la librairie `keyphrase-vectorizers` qui offre la possibilité d'extraire les phrases-clés pertinentes et spécifiques à l'aide des balises de parties de discours. Cela nous a paru comme une piste intéressante, étant donné que les termes

20. Pour assurer que les phrases-clés ne se ressemblent pas, il faut utiliser le paramètre `use_mmr` et spécifier sa valeur entre 0 et 1.

médicaux (surtout ceux plus pointus) que l'on souhaitait extraire étaient généralement des n-grammes constitués des substantifs, suivis d'un ou plusieurs adjectifs (p. ex. *sclérose latérale amyotrophique*). Voici les étapes de la chaîne de traitement de l'approche *PatternRank* (figure 2.7) :

1. les corpus « Charcot » et « Autres » sont utilisés comme les données d'entrée au format `.txt` ;
2. les tokens ont été extraits et étiquetés avec les balises de partie du discours et les expressions régulières `<N.*>+<ADJ.*>*` (sans utiliser le paramètre `use_mmr`) ;
3. les tokens ont été sélectionnés selon les balises de partie de discours souhaitées et gardés comme les phrases-clés candidates ;
4. les plongements des documents et de leurs phrases-clés candidates ont été générés par le modèle de langue (en l'occurrence `flair` [21]) ;
5. les similarités cosinus ont été calculées entre ces deux types de plongements, et les phrases-clés candidates ont été triées par ordre décroissant ;
6. les phrases-clés les plus représentatives ont été extraites.

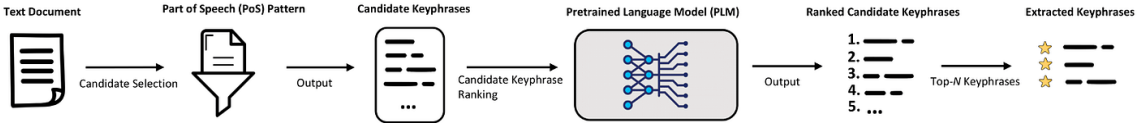

FIGURE 2.7 – *Workflow* de la méthode *PatternRank* (SCHOPF *ET AL.*, 2022, p. 2).

La figure 2.8 nous informe sur les 15 termes les plus pertinents et fréquents, extraits avec la librairie `keyphrase-vectorizers`, que l'on retrouve dans les deux corpus. Malgré certains tokens tronqués, très probablement en raison d'un OCR imparfait (*ments* → *mouvements*, *decins* → *médecins* etc.), nous observons une diversification des résultats. Après cela, il reste la question de mieux comprendre le rôle des phrases-clés extraites dans les écrits de l'entourage de Charcot et/ou si elles sont vraiment significatives ou pas.

21. https://github.com/flairNLP/flair

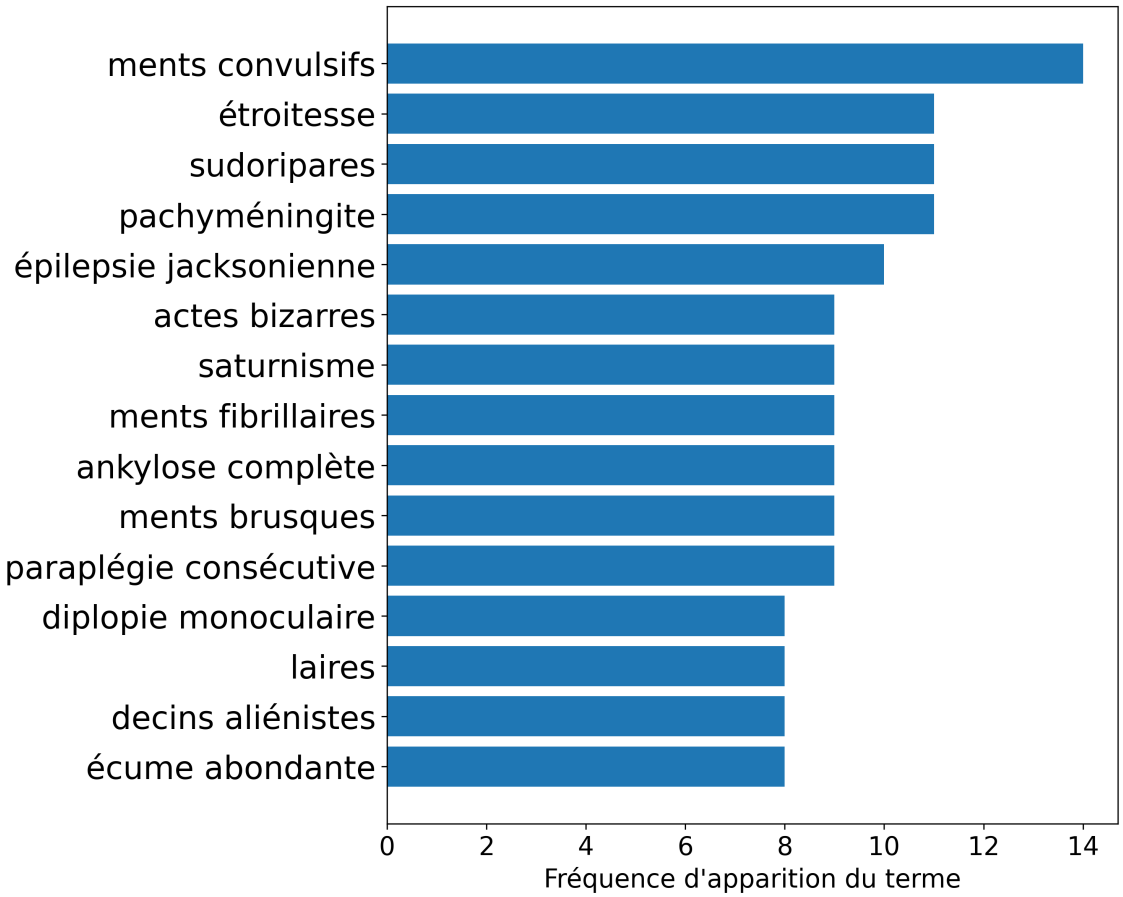

FIGURE 2.8 – Les 15 termes les plus fréquents partagés par les deux corpus selon `keyphrase-vectorizers`.

Chapitre 3 Conclusion

1 Contributions et perspectives

Intro / Rappel Contexte
Nous avons donc pu en tirer la problématique suivante :
Discussion et perspectives

Bibliographie

Adell, N. (2011). Chapitre 6 – La circulation des savoirs. *Collection U*, (pp. 251–292). https://shs.cairn.info/anthropologie-des-savoirs--9782200270360-page-251?lang=fr&utm_source=chatgpt.com. (pages 4, 5)

Alrahabi, M. (2021). Ariane : dispositif de fouille et de lecture synthétique de textes. In *DigitAl Humanities and cuLtural herItAge : data and knowledge management and analysis (Atelier Dahlia)*. https://hal.science/hal-03167271.

Alrahabi, M. (2022). Obvie : interface web pour la fouille et la comparaison de textes. In *Atelier DigitAl Humanities and cuLtural herItAge : data and knowledge management and analysis durant la conférence francophone sur l'Extraction et la Gestion des Connaissances (egc2022)*. https://hal.science/hal-03543362/. (page 25)

Amiri, V. V. (24 novembre 2012). T. S. Kuhn. *Histo Philo Sciences*. https://histoirephilosciences.wordpress.com/depuis-le-20eme-siecles/une-nouvelle-epistemologie/t-s-kuhn/. (page 8)

Andrade, P. (2013). Sociologie sémantico-logique des ruines : pour une herméneutique hybride de la ruine du web 2.0 au web 3.0. *Sociétés*, 120(2), 105–119. https://doi.org/10.3917/soc.120.0105. (page 6)

Anouilh, J. (1956). *Pauvre Bitos ou le dîner de têtes*. Gallimard, coll. « Folio », n° 301. https://archive.org/details/anouilh-pauvre-bitos-ou-le-diner-de-tetes-1979. (page 7)

Astolfi, J.-P., Darot, É., Ginsburger-Vogel, Y., & Toussaint, J. (2008). Chapitre 2. concept, conceptualisation. *Pratiques pédagogiques*, 2, 23–33. https://shs.cairn.info/mots-cles-de-la-didactique-des-sciences--9782804157166-page-23?lang=fr. (pages 8, 17, and 19)

Bachelard, G. (1934). *La formation de l'esprit scientifique : contribution à une psychanalyse de la connaissance*. Vrin. https://gastonbachelard.org/wp-content/uploads/2015/07/formation_esprit.pdf. (page 7)

Bachelard, G. (1970). *Idéalisme discursif*. Vrin, présentation de Georges Canguilhem : Paris. https://www.academia.edu/27217437/BACHELARD_Gaston_%C3%89tudes_Vrin_1970_. (page 8)

Bal, M. (2002). *Travelling Concepts in the Humanities : A Rough Guide*. University of Toronto Press. https://s3.amazonaws.com/arena-attachments/89974/705194c45c063480ed0bb3af6fdd2dfc.pdf. (page 17)

Bernheim, H. (1891). *De la suggestion et de ses applications à la thérapeutique*. Paris : Octave Doin. https://gallica.bnf.fr/ark:/12148/bpt6k97805169. (page 11)

Bezançon, J. & Lejeune, G. (2023). Reconnaissance de défigements dans des tweets en français par des mesures de similarité sur des alignements textuels. In C. Servan & A. Vilnat (Eds.), *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs* (pp. 56–67). Paris, France : ATALA. https://aclanthology.org/2023.jeptalnrecital-long.5. (page 19)

Bogousslavsky, J. (2011). *Following Charcot : A Forgotten History of Neurology and Psychiatry*, volume 29. Karger Medical and Scientific Publishers. https://nah.sen.es/en/issues/lastest-issues/135-journals/volume-2/issue-2/270-the-mysteries-of-hysteria. (pages 2, 5)

Bogousslavsky, J. (2014). The Mysteries of Hysteria. *Neurosciences and History*, 2(2), 54–73. https://nah.sen.es/vmfiles/abstract/NAHV2N2201454_73EN.pdf. (page 10)

Broussolle, E., Poirier, J., Clarac, F., & Barbara, J.-G. (2012). Figures and institutions of the neurological sciences in Paris from 1800 to 1950. Part III : Neurology. *Revue Neurologique*, 168(4), 301–320. https://doi.org/10.1016/j.neurol.2011.10.006. (pages 2, 5, and 9)

Camargo, C. H. F., Coutinho, L., Correa Neto, Y., Engelhardt, E., Maranhão Filho, P., Walusinski, O., & Teive, H. A. G. (2024). Jean-Martin Charcot : the polymath. *Arquivos de Neuro-psiquiatria*, 81, 1098–1111. https://www.thieme-connect.de/products/ejournals/pdf/10.1055/s-0043-1775984.pdf. (pages 2, 5, 9, 10, and 30)

Camargo, C. H. F., Marques, P. T., de Oliveira, L. P., Germinian, F. M., de Paola, L., & Teive, H. A. G. (2018). Jean-Martin Charcot's Influence on Career of Sigmund Freud, and the Influence of this Meeting for the Brazilian Medicine. *Revista Brasileira de Neurologia*, 54(2). https://docs.bvsalud.org/biblioref/2018/07/907032/revista542v4-artigo6.pdf. (page 10)

Charcot, J. M. (1892). *Œuvres complètes de J. M. Charcot. Leçons sur les maladies du système nerveux.*, volume 1. Bureaux du progrès medical. https:

//patrimoine.sorbonne-universite.fr/viewer/3468/?offset=1#page=2&viewer=picture&o=&n=0&q=. (page 30)

Charcot, J.-M. (1897). *La foi qui guérit*. F. Alcan (Paris). https://gallica.bnf.fr/ark:/12148/bpt6k68008w. (page 31)

Chaudet, C. (2022). Les « Illuminati » du pamphlet au roman : circulations d'un discours complotiste à grande échelle depuis le tournant du XIXe siècle. *Mots. Les langages du politique*, (pp. 19–36). https://www.cairn.info/revue-mots-2022-3-page-19.htm. (page 12)

Christin, O. (2011). *Dictionnaire des concepts nomades en sciences humaines*. Métailié. https://editions-metailie.com/livre/dictionnaire-des-concepts-nomades-en-sciences-humaines/. (page 16)

Cram, D. & Daille, B. (2016). Terminology Extraction with Term Variant Detection. In *Proceedings of ACL-2016 system demonstrations* (pp. 13–18). https://aclanthology.org/P16-4003.pdf. (page 18)

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert : Pre-training of Deep Bidirectional Transformers for Language Understanding. (page 30)

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 9(1), 99–115. https://doi.org/10.1075/term.9.1.06dro. (page 18)

Gabay, S., Du Noyer, L. R., Levenson, M. G., Petkovic, L., & Bartz, A. (2020). Quantifying the Unknown : How many manuscripts of the marquise de Sévigné still exist ? In *Digital Humanities DH2020*. https://hal.science/hal-02898929. (pages 1, 4)

Gabay, S., Petkovic, L., Bartz, A., Levenson, M. G., & Du Noyer, L. R. (2021). Katabase : À la recherche des manuscrits vendus. In *Humanistica 2021* (pp. 1–7). https://hal.science/hal-03066108. (pages 1, 2, 4, and 5)

Ghermani, N. (2011). Confessions. In O. Christin (Ed.), *Dictionnaire des concepts nomades en sciences humaines* (pp. 117–133). Métailié. https://www.academia.edu/5335160/_Confession_. (pages 16, 17)

Giry, J. & Nouvel, D. (2022). Étudier les discours « conspirationnistes » et leur circulation sur Twitter : Les théories du complot comme objets du traitement automatique du langage et de l'analyse des données textuelles. *Mots. Les langages du politique*, (pp. 37–55). https://www.cairn.info/revue-mots-2022-3-page-37.htm. (page 12)

Goetz, C. (2017). Charcot : Past and present. *Revue Neurologique*, 173(10), 628–636. https://doi.org/10.1016/j.neurol.2017.04.004. (page 10)

Gomes, M. d. M. & Engelhardt, E. (2013). Jean-Martin Charcot, father of modern neurology : an homage 120 years after his death. *Arquivos de Neuro-Psiquiatria*, 71, 815–817. https://doi.org/10.1590/0004-282X20130128. (page 10)

Grootendorst, M., Mishra, A., Matsak, A., OysterMax, Govil, P., Ogura, Y., Warmerdam, V. D., & yusuke1997 (2023). Maartengr/keybert : v0.8. https://doi.org/10.5281/zenodo.8388690. (page 32)

Guemas, G. (12 février 2024). Qu'est-ce que le Gartner Magic Quadrant ? https://tool-advisor.fr/blog/gartner-magic-quadrant/htm. Tool Advisor, consulté le 24 mars 2025. (page 12)

Hey, T., Tansley, S., & Tolle, K. M. (2009). Jim Gray on eScience : A Transformed Scientific Method. In T. Hey, S. Tansley, & K. M. Tolle (Eds.), *The Fourth Paradigm*. Microsoft Research. http://languagelog.ldc.upenn.edu/myl/JimGrayOnE-Science.pdf. (page 11)

Hobsbawm, E. (2010). *The Age of Revolution : 1789-1848*. Hachette UK. https://files.libcom.org/files/Eric%20Hobsbawm%20-%20Age%20Of%20Revolution%201789%20-1848.pdf. (page 16)

Johns, T. F. (1991). Should You be Persuaded. Two Samples of Data-Driven Learning Materials. https://api.semanticscholar.org/CorpusID:53988458. (page 11)

Jolly, A., Pandey, V., Singh, I., & Sharma, N. (2024). Exploring Biomedical Named Entity Recognition via SciSpacy and BioBERT models. *The Open Biomedical Engineering Journal*, 18(1). https://doi.org/10.2174/0118741207289680240510045617. (page 19)

Joyeux-Prunel, B. (2019). Visual Contagions, the Art Historian, and the Digital Strategies to Work on Them. *Artl@s Bulletin*, 8(3), 128–144. https://docs.lib.purdue.edu/artlas/vol8/iss3/8/. (pages 2, 5)

Joyeux-Prunel, B. & Gabay, S. (2022). Circulations des savoirs, de la recherche à l'enseignement. *Arabesques*. https://doi.org/10.35562/arabesques.2847. (page 11)

Kageura, K. & Umino, B. (1996). Methods of Automatic Term Recognition : A Review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2), 259–289. https://doi.org/10.1075/term.3.2.03kag. (page 18)

Kant, É. (1863). *Anthropologie d'un point de vue pragmatique* (trad. J. Tissot). Librairie Ladrange (originalement publié en 1798). https://fr.wikisource.org/wiki/Page:Kant_-_Anthropologie.djvu/452. (page 9)

Khurshid, A., Gillman, L., & Tostevin, L. (2000). Weirdness indexing for logical document extrapolation and retrieval. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*. https://trec.nist.gov/pubs/trec8/papers/surrey2.pdf. (page 29)

Koehler, P. J. (2013). Chapter 6 – Charcot, La Salpêtrière, and Hysteria as Represented in European Literature. In S. Finger, F. Boller, & A. Stiles (Eds.), *Literature, Neurology, and Neuroscience : Neurological and Psychiatric Disorders*, volume 206 of *Progress in Brain Research* (pp. 93–122). Elsevier. https://www.sciencedirect.com/science/article/pii/B9780444633644000235. (page 10)

Koselleck, R. & Richter, M. (2011). Introduction and Prefaces to the Geschichtliche Grundbegriffe : (Basic Concepts in History : A Historical Dictionary of Political and Social Language in Germany). *Contributions to the History of Concepts*, 6(1), 1–37. https://www.berghahnjournals.com/view/journals/contributions/6/1/choc060102.xml. (page 16)

Koyré, A. (1957). *From the Closed World to the Infinite Universe*, volume 1. Baltimore, Johns Hopkins Press. https://archive.org/details/fromclosedworldt0000koyr/page/2/mode/2up?q=%22revolution%22. (page 8)

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press. https://www.lri.fr/~mbl/Stanford/CS477/papers/Kuhn-SSR-2ndEd.pdf. (page 8)

Landais, É. (2014). « Frédéric Darbellay, éd., *La circulation des savoirs. Interdisciplinarité, concepts nomades, analogies, métaphores* » : Berne, P. Lang, 2012, 245 pages. *Questions de communication*, 26, 331–333. https://doi.org/10.4000/questionsdecommunication.9367. (pages 3, 5, and 6)

Le Pois, C. (1618). *Selectiorum observationum et consiliorum de praetervisis hactenus morbis affectibusque praeter naturum, ab aqua seu serosa colluvie et diluvie ortis, liber singularis.* Authore Carolo Pisone, Ponte ad Monticulum, apud Carolum Mercatorem. https://archive.org/details/BIUSante_05814/page/n3/mode/2up. (page 9)

Lecourt, D., Ed. (1999). *Dictionnaire d'histoire et philosophie des sciences*. Puf. https://www.librairiedalloz.fr/livre/9782130544999-dictionnaire-d-histoire-et-philosophie-des-sciences-4e-edition-dominique-lecourt/. (pages 14, 19)

Manjavacas, E., Long, B., & Kestemont, M. (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* (pp. 104–114). Minneapolis, USA : Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-2514. (pages 2, 5)

Milia, M. (2023). Using Digital Text-Based Approaches to Study Knowledge Circulation. In *Routledge Handbook of Academic Knowledge Circulation* (pp. 556–574). Routledge. https://www.taylorfrancis.com/chapters/edit/10.

4324/9781003290650-52/using-digital-text-based-approaches-study-knowledge-circulation-mat%C3%ADas-milia. (pages 2, 3, 5, and 6)

Mirbeau, O. & Michel, P. (1995). *Chroniques du diable*, volume 555. Presses Univ. Franche-Comté. https://mirbeau.asso.fr/darticlesfrancais/Marquer-Mirbeau%20et%20Charcot.pdfen. (page 10)

Monteiro, F., Nardi, A., & Gomes, M. (2021). The 400th anniversary of the birth of Thomas Willis (1621-1675) : an invaluable contributor to neuroscience. *Revista Brasileira de Psiquiatria*, 44. https://doi.org/10.1590/1516-4446-2021-2159. (page 9)

Morin, C. & Mésangeau, J. (2022). Les discours complotistes de l'antiféminisme en ligne. *Mots. Les langages du politique*, (pp. 57–78). https://shs.cairn.info/revue-mots-2022-3-page-57?lang=fr. (page 12)

Mounin, G. (1968). *Clefs pour la linguistique*. Collection Clefs. Seghers. https://books.google.fr/books?id=7SgDAAAAMAAJ. (page 19)

Navarro, D. F., Ijaz, K., Rezazadegan, D., Rahimi-Ardabili, H., Dras, M., Coiera, E., & Berkovsky, S. (2023). Clinical named entity recognition and relation extraction using natural language processing of medical free text : A systematic review. *International Journal of Medical Informatics*, 177, 105122. https://doi.org/10.1016/j.ijmedinf.2023.105122. (page 19)

Nemickienė, Ž. (2011). "Concept" in Modern Linguistics : the Component of the Concept "Good". *Filologija*, 16, 26–36. https://core.ac.uk/outputs/62656539?source=oai.

Nerima, L., Seretan, V., & Wehrli, E. (2006). Le problème des collocations en TAL. *Nouveaux cahiers de linguistique française*, 27, 95–115. https://access.archive-ouverte.unige.ch/access/metadata/fc3fad28-5b90-42ec-bea5-0c6d54cb5452/download. (pages 3, 7)

Petkovic, L. (2019). Creation and Analysis of the Yugoslav Rock Song Lyrics Corpus from 1967 to 2003. *INFOtheca : Journal of Information and Library Science*, 19(1), 5–29. https://doi.org/10.18485/infotheca.2019.19.1.1. (pages 1, 4)

Petkovic, L., Alrahabi, M., & Glenn, R. (2022). Impact de la correction automatique de l'ocr/htr sur la reconnaissance d'entités nommées dans un corpus bruité. *Journal of Information Sciences*, 21(2), 42–57. https://doi.org/10.34874/IMIST.PRSM/jis-v21i2.36599.

Petkovic, L., Alrahabi, M., & Roe, G. (2023). Circulation du discours médical de jean-martin charcot. In *Humanistica 2023*. https://hal.science/HUMANISTICA-2023/hal-04107099v1. (pages 3, 2)

Quet, M. (2014). « Frédéric Darbellay, *La circulation des savoirs. Interdisciplinarité, concepts nomades, analogies, métaphores* ». *Revue d'anthropologie des connaissances*, 8(8-1). https://doi.org/10.3917/rac.022.0221. (pages 5, 6)

Rennesson, M., Georget, M., Paillard, C., Perrin, O., Pigeotte, H., & Tête, C. (2020). Le thésaurus, un vocabulaire contrôlé pour parler le même langage. *Médecine Palliative*, 19(1), 15–23. Documentation et pratiques documentaires en soins palliatifs. Coordonné par Caroline Tête. (page 19)

Rey, A. (1998). *Dictionnaire historique de la langue française. Tome 2.* Le Robert. https://www.plouffe.fr/simon/Dictionnaires/Le%20Robert%20Dictionnaire%20Historique%202a.pdf. (page 8)

Riffaterre, M. (1980). La trace de l'intertexte. *Pensée (La) Paris*, (215), 4–18. https://api.semanticscholar.org/CorpusID:170902390.

Riguet, M. (2018). L'impact de la physiologie dans la critique littéraire de la fin du XIX^{ème} siècle : l'exemple de Claude Bernard. *Epistémocritique : Littérature et savoirs*. https://hal.science/hal-01903871. (page 12)

Robertson, S., Zaragoza, H., *et al.* (2009). The probabilistic relevance framework : Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4), 333–389. https://dx.doi.org/10.1561/1500000019. (page 30)

Robertson, S. E. & Jones, K. S. (1976). Relevance Weighting of Search Terms. *Journal of the American Society for Information science*, 27(3), 129–146. https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.4630270302?casa_token=TfyVkMGkDQsAAAAA:TCuXWzGHjo31RdxGR9jECRG2rZzqv0K3GOzHF7yAa2NfxtDFqxe-MmSHMC6e8OFiFxI4sLj2aW6OyDk. (page 30)

Roe, G., Fedchenko, V., & Nicolosi, D. M. (2023). Enlightenment Inflluencers : Networks of Text Reuse in 18th-century France. In *Digital Humanities 2023* (pp. 296–299). https://doi.org/10.5281/zenodo.8107964. (page 12)

Rumelhard, G. (1986). *La génétique et ses représentations dans l'enseignement*. Berne : Peter Lang. https://shs.cairn.info/mots-cles-de-la-didactique-des-sciences--9782804157166-page-23?lang=fr. (page 17)

Saussure, F. d., Bally, C., Sechehaye, A., & Riedlinger, A. (1915). *Cours de linguistique générale / Ferdinand de Saussure; publié par Charles Bailly et Albert Sechehaye avec la collaboration de Albert Riedlinger*. Grande bibliothèque Payot. Genève : Payot. https://www.arbredor.com/collections/etudes-et-essais/77-cours-de-linguistique-generale. (page 18)

Schopf, T., Klimek, S., & Matthes, F. (2022). PatternRank : Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase Extraction. In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management* : SCITEPRESS – Science and Technology Publications. http://dx.doi.org/10.5220/0011546600003335. (page 34)

Silberztein, M. (2022). Linguistic Resources for Corpus Processing : the ATISHS project. In *JADT2022 International Conference on Statistical Analysis of Textual Data*. https://hal.science/hal-03854939/. (pages 17, 18)

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 24 mars 2025. (page 12)

Stengers, I. (1987). *D'une science à l'autre. : Des concepts nomades*. Seuil. https://archive.org/details/dunesciencealaut0000unse. (pages 15, 16)

Tasca, C., Rapetti, M., Carta, M. G., & Fadda, B. (2012). Women And Hysteria In The History Of Mental Health. *Clinical Practice & Epidemiology in Mental Health : CP & EMH*, 8, 110–119. https://doi.org/10.2174/1745017901208010110. (pages 8, 9)

Teive, H. A. G., Coutinho, L., Camargo, C. H. F., Munhoz, R. P., & Walusinski, O. (2022). Thomas Willis' legacy on the 400th anniversary of his birth. *Arquivos de Neuro-Psiquiatria*, 80, 759–762. https://doi.org/10.1055/s-0042-1755278. (page 9)

Teive, H. A. G., Germiniani, F., Munhoz, R. P., & Paola, L. d. (2014). 126 hysterical years - the contribution of Charcot. *Arquivos de Neuro-Psiquiatria*, 72, 636–639. https://doi.org/10.1590/0004-282x20140068.PMID:25098481. (page 10)

Tran, H. T. H., Martinc, M., Caporusso, J., Doucet, A., & Pollak, S. (2023). The Recent Advances in Automatic Term Extraction : A Survey. *arXiv preprint arXiv :2301.06767*. https://arxiv.org/pdf/2301.06767. (page 18)

Tubbs, R. S., Loukas, M., Shoja, M. M., Apaydin, N., Ardalan, M. R., Shokouhi, G., & Oakes, W. J. (2008). Costanzo Varolio (Constantius Varolius 1543–1575) and the Pons Varolli. *Neurosurgery*, 62(3), 734–737. https://doi.org/10.1227/01.neu.0000317323.63859.2a. (page 9)

Varet, V. (2023). Les nouvelles modalités numériques : *blockchain*, Web 3.0, NFT, métavers... *Legipresse*, 68(HS1), 59–70. https://doi.org/10.3917/legip.hs68.0059.

Varolio, C. (1573). *De nervis opticis nonnullisq : aliis praeter communem opinionem in humano capite obseruatis*. Patavii : apud P. et A. Meiettos fratres. https://gallica.bnf.fr/ark:/12148/bpt6k325486q. (page 9)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need. https://arxiv.org/abs/1706.03762. (page 31)

White, M. B. (1997). Jean-Martin Charcot's Contributions to the Interface Between Neurology and Psychiatry. *Canadian Journal of Neurological Sciences*, 24(3), 254–260. https://doi.org/10.1017/S0317167100021909. (page 10)

Willis, T. (1664). *Cerebri anatome : cui accessit nervorum descriptio et usus*. Londini : Typis Ja. Flesher, impensis Jo. Martyn & Ja. Allestry, apud insigne Campanæ in Cœmeterio, D. Pauli. https://books.google.fr/books/?id=L2xEAAAAcAAJ&pg=PP9#v=onepage&q&f=false. (page 9)

Willis, T. (1681). *An Essay of the Pathology of the Brain and Nervous Stock in which Convulsive Diseases are Treated of*. London : Printed by J. B. for T. Dring. https://quod.lib.umich.edu/e/eebo/A66496.0001.001?rgn=main;view=fulltext. (page 9)

Wright, J. P. (1980). Hysteria and Mechanical Man. *Journal of the History of Ideas*, 41(2), 233–247. https://doi.org/10.2307/2709458. (page 9)