Comparaison des approches pour l'extraction de phrases-clés

Ljudmila PETKOVIC[1,2,3,4]

prenom.nom@sorbonne-universite.fr

[1] Sorbonne Université, Faculté des Lettres, UFR Littératures françaises et comparée, ED III (ED019)
[2] Sorbonne Université, Centre d'étude de la langue et des littératures françaises (CELLF), UMR 8599
[3] Sorbonne Université, Observatoire des textes, des idées et des corpus (OBTIC)
[4] Sorbonne Université, UFR Sociologie et Informatique pour les Sciences Humaines

Séminaire doctoral OBTIC
SCAI, salle du Conseil
Paris, le 13 mars 2025

Termes inventés par Charcot : référence

paralysie agitante	**maladie de Parkinson**
ataxie locomotrice progressive	***tabes dorsalis***
arthropathies tabétiques	**arthropathie de Charcot**
trépidation épileptoïde du pied	**clonus**
sclérose en plaques disséminées	**sclérose multiple**
sclérose latérale amyotrophique	**maladie de Charcot / Lou Gehrig**
idée(s) fixe(s), maladie des tics	**syndrome de Tourette**
mouvements involontaires	**chorées**, **athétose**
incapacité d'être debout / de marcher	**astasie-abasie**
atrophie musculaire progressive	**maladie Charcot-Marie-Tooth**

(Walusinski, 2025 ; Camargo et al., 2023)

≠ termes transmis : **hystérie**
épilepsie
hypnose

Approches comparées

1. `TermSuite` (Cram et Daille, 2016)
   - linguistique, à base de règles → TD-IDF
2. **TF-IDF, BM25** (Robertson et Jones, 1976)
   - statistique
3. ***PatternRank*** (Schopf et al., 2022)
   - apprentissage profond
   - `keybert` + `keyphrase-vectorizers`
   - utilisation des étiquettes POS

Traitements effectués en local (1,2) et *via* la plateforme MeSU [a] (3).

a. https://sacado.sorbonne-universite.fr/fr/plateforme-mesu/

Critères de comparaison des approches

- prise en compte des synonymes des termes
  - ex. *paralysie agitante* → *maladie de Parkinson*
- recenser le score le plus élevé sur le terme ou sur son synonyme
- les méthodes classiques *vs.* celles de l'état de l'art

Le domaine potentiellement impactant : **syndrome de Tourette**

Terme	**TF-IDF** (*TermSuite*)	**TF-IDF**	**BM25**	*PatternRank*
maladie de Parkinson	0,05	0,0775	0,333	0,7936
ataxie locomotrice progressive	0,32	0,0386	0,4877	0,7431
arthropathies tabétiques	0,33	0,0934	0,4928	0,7506
trépidation épileptoïde du pied	0,0198	0,1227	0,2919	0.7597
sclérose en plaques disséminées	NA	0,178	0,8089	NA
tremblement	NA	0,1686	0,0362	0,7683
nystagmus	0,0243	0,1326	0,146	0,7474
embarras parole	NA	NA	0,0018	0,9347
sclérose latérale amyotrophique	NA	0,044	0,6586	NA
tics convulsifs	NA	0,1293	0,8385	**0,8331**
atrophie musculaire progressive	0,40	0.1118	0.3489	0,8053
aphasie	0,0587	0,2245	0,1334	0,7960
astasie-abasie	NA	0,0478	0,3565	0,7375
athétose	NA	0,2029	0,274	0,8068
chorées	NA	0,1336	0,0701	0,8047
hystérie	0,2724	0,3711	0,0442	0,8018
épilepsie	NA	0,164	0,0247	0,8199
hypnose	0,3543	1	0,2922	0,7738

Tab. 1 – Les scores de pertinence pour les termes de référence à partir du corpus « Autres ».

Analyse comparative des approches employées

- *PatternRank* valorise systématiquement les termes
- pas de consensus entre les métriques
  - l'écart le plus petit entre eux : *hypnose*

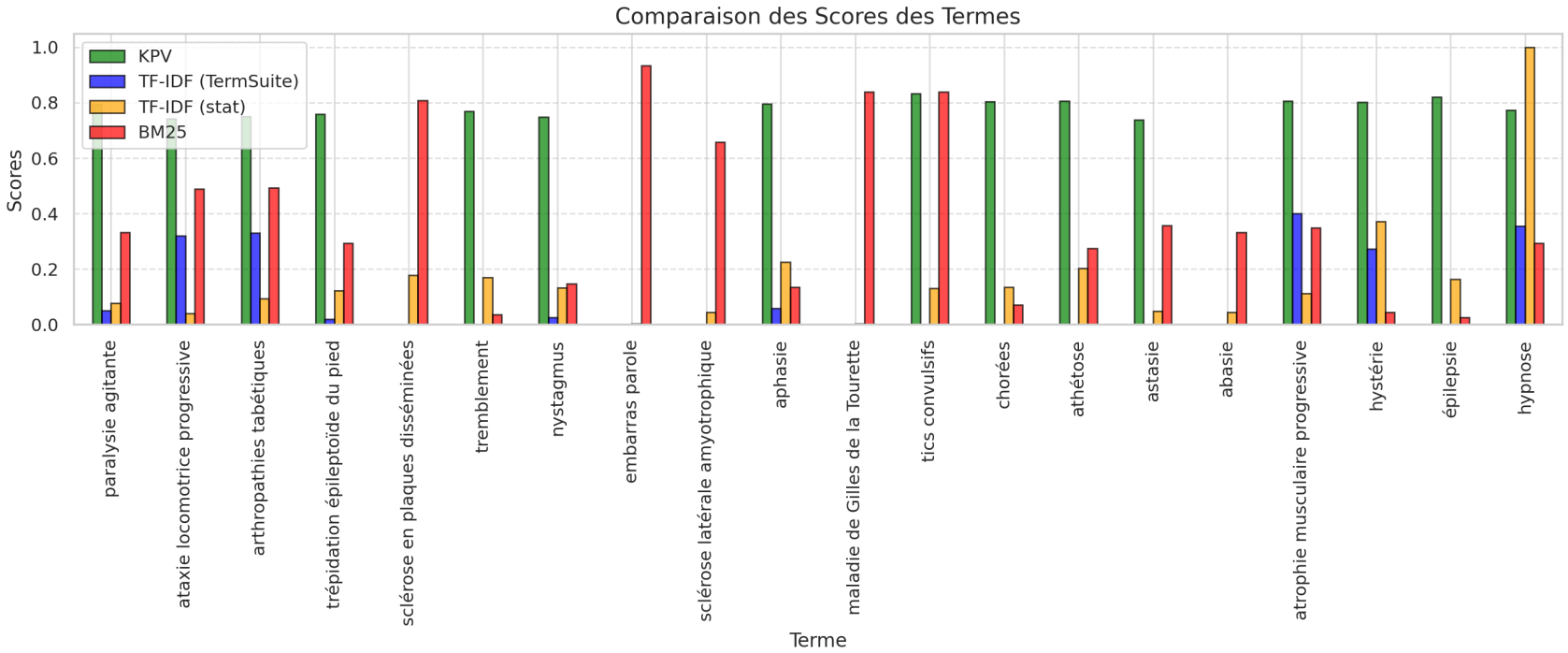

Fig. 1 – Visualisation des scores de pertinences pour chaque terme de référence

Chronologie d'une locution : indice de croissance de l'impact ?

- évolution de la fréquence des termes au sein des deux corpus
- convergence entre des termes : fin XIX^{e}, début XX^{e} s.
  - *ppm* : nombre d'occurrences par million de mots

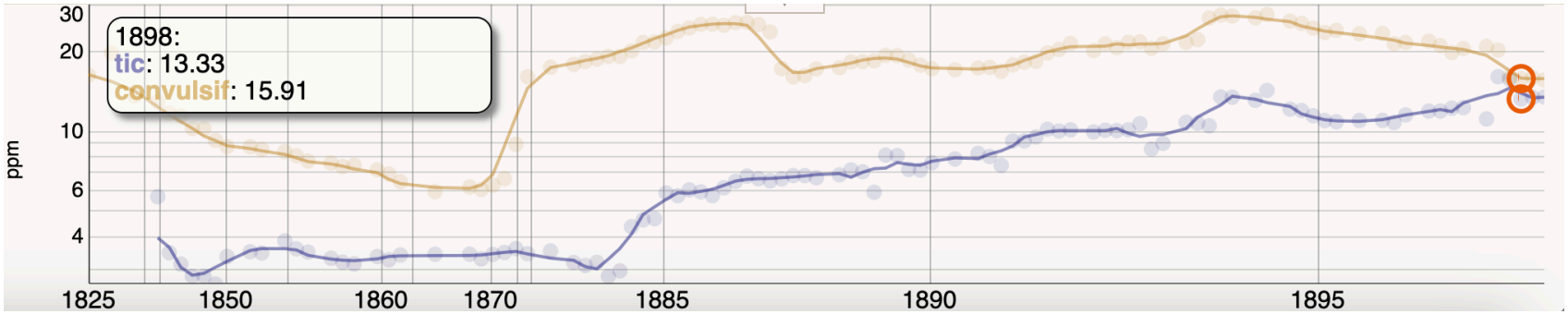

Fig. 2 – Chronologie de la fréquence du terme *tic convulsif*.

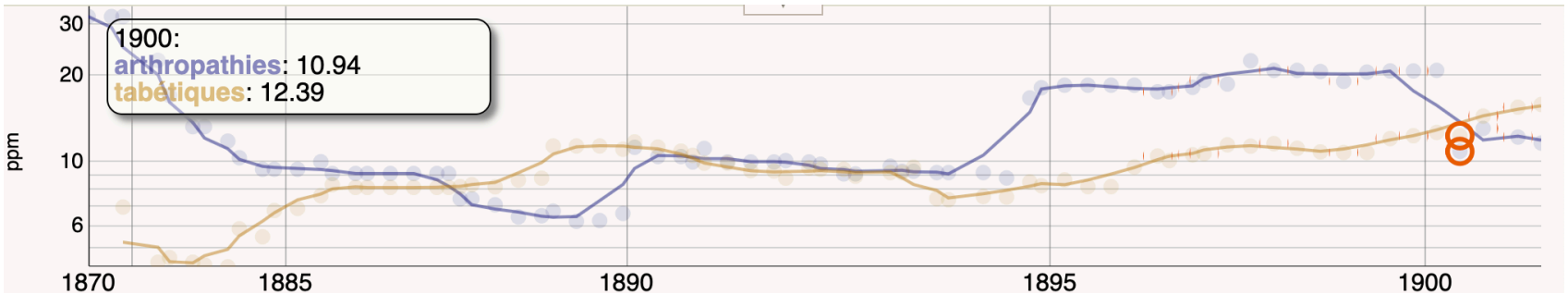

Fig. 3 – Chronologie de la fréquence du terme *arthropathies tabétiques*.

Conclusion

1. *PatternRank* : la méthode la plus fiable par rapport aux autres méthodes ?
   - capture la sémantique jusqu'aux pentagrammes
     - *méningite syphilitique hémorragique fibrineuse aiguë*
   - produit des scores de pertinence plus élevés
     - exception : scores BM25 (*SLA*, *embarras parole*) et TF-IDF (*hypnose*)
2. Les termes les plus impactants → syndrome de Tourette.
3. Absence des scores pour les termes comme *SEP* et SLA :
   - solution : chercher leurs symptomes ou leurs descriptions :
     - *amyotrophie spinale progressive*, *secousses nystagmiques*…

Références I

Camargo, C. H. F., L. Coutinho, Y. Correa Neto, E. Engelhardt, P. Maranhão Filho, O. Walusinski et H. A. G. Teive (2023). Jean-Martin Charcot : the polymath. In : *Arquivos de Neuro-psiquiatria* 81. https://www.thieme-connect.de/products/ejournals/pdf/10.1055/s-0043-1775984.pdf, p. 1098-1111 (*voir p. 3*).

Cram, D. et B. Daille (2016). Terminology Extraction with Term Variant Detection. In : *Proceedings of ACL-2016 system demonstrations*. https://aclanthology.org/P16-4003.pdf, p. 13-18 (*voir p. 5*).

Robertson, S. E. et K. S. Jones (1976). Relevance Weighting of Search Terms. In : *Journal of the American Society for Information science* 27.3. https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.4630270302?casa_token=TfyVkMGkDQsAAAAA:TCuXWzGHjo31RdxGR9jECRG2rZzqv0K3GOzHF7yAa2NfxtDFqxe-MmSHMC6e8OFiFxI4sLj2aW6OyDk, p. 129-146 (*voir p. 5*).

Schopf, T., S. Klimek et F. Matthes (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*. http://dx.doi.org/10.5220/0011546600003335. SCITEPRESS – Science et Technology Publications. DOI : 10.5220/0011546600003335. URL : http://dx.doi.org/10.5220/0011546600003335 (*voir p. 5*).

Références II

Walusinski, O. (2025). *Jean-Martin Charcot's Birth Bicentennia*. http://ishn.org/. Consulté le 13 mars 2025 (*voir p. 3*).