Les *embeddings*, nouvel outil d'une histoire numérique des grands corpus de presse ?

Arthur Michelet[1], and Martin Grandjean[1]

[1] Université de Lausanne

Abstract

This research focuses on the development of a procedure for analyzing very large corpora of news articles. It discusses the cross-analysis of archival sources and news articles using text embeddings within a notebook designed for historians and students.

Mots-clés: Histoire, histoire économique, histoire des médias, relations publiques, traitement automatique des langues, embeddings, humanités numériques, humanités computationnelles, visualisation de données
Keywords: History, economic history, media history, public relations, natural language processing, embeddings, digital humanities, computational humanities, data visualization

1 Introduction

En raison de leur nature massive et hautement standardisée, et plus récemment en raison d'un intense effort de numérisation, les archives de presse constituent des sources d'intérêt majeur pour l'histoire numérique, un véritable « Eldorado » à explorer [2]. Au cours de la dernière décennie, l'intérêt pour ces corpus a donné lieu à des projets interdisciplinaires à grande échelle, tels que Numapresse (http ://www.numapresse.org), ViralTexts (https ://viraltexts.org), NewsEye (https ://www.newseye.eu/) ou Impresso (https ://impresso-project.ch/). Il s'agit là d'un terrain particulièrement fertile pour l'innovation, qu'il s'agisse du traitement et de l'enrichissement de ces données, de leur analyse à l'aide de toute une gamme d'outils de lecture distante, ou de leur mise à disposition du grand public via des interfaces en ligne [4].

Nous développons, questionnons et critiquons une approche de *distant reading* [17] basée sur les *text embeddings*, une méthode qui consiste à calculer la proximité sémantique des articles de presse dans un espace multidimensionnel. Cette approche vise à développer un pipeline d'analyse qui consiste à calculer la proximité de milliers d'articles de la presse suisse (francophones et germanophones principalement) et de la comparer à des sources historiques externes pour discuter de la pertinence des clusters détectés, dans notre cas des documents permettant d'étudier l'activité de communication des banques suisses au XXe siècle. Au-delà d'un outil d'analyse destiné uniquement aux membres du projet, le *notebook* ainsi créé [13] a pour vocation de rejoindre un *datalab* public à disposition des chercheuses et chercheurs en histoire numérique des médias et être utilisé dans le cadre de plusieurs enseignements de Master dans les universités des porteurs du projet. Lier une étude de cas à ce questionnement épistémologique et pédagogique nous permet d'éviter les écueils d'une approche trop théorique et de fonder ces réflexions sur des problématiques historiques concrètes.

Arthur Michelet, and Martin Grandjean. "Les *embeddings*, nouvel outil d'une histoire numérique des grands corpus de presse?." *Actes de la Conférence Humanistica*, éd. par Serena Crespi, Simon Gabay, Martin Grandjean, Ariane Pinche, Marie Puren et Léa Saint-Raymond. Vol. 4. Anthology of Computers et the Humanities. 2026, 175–181. https://doi.org/10.63744/R2LLVRsBeNuI.

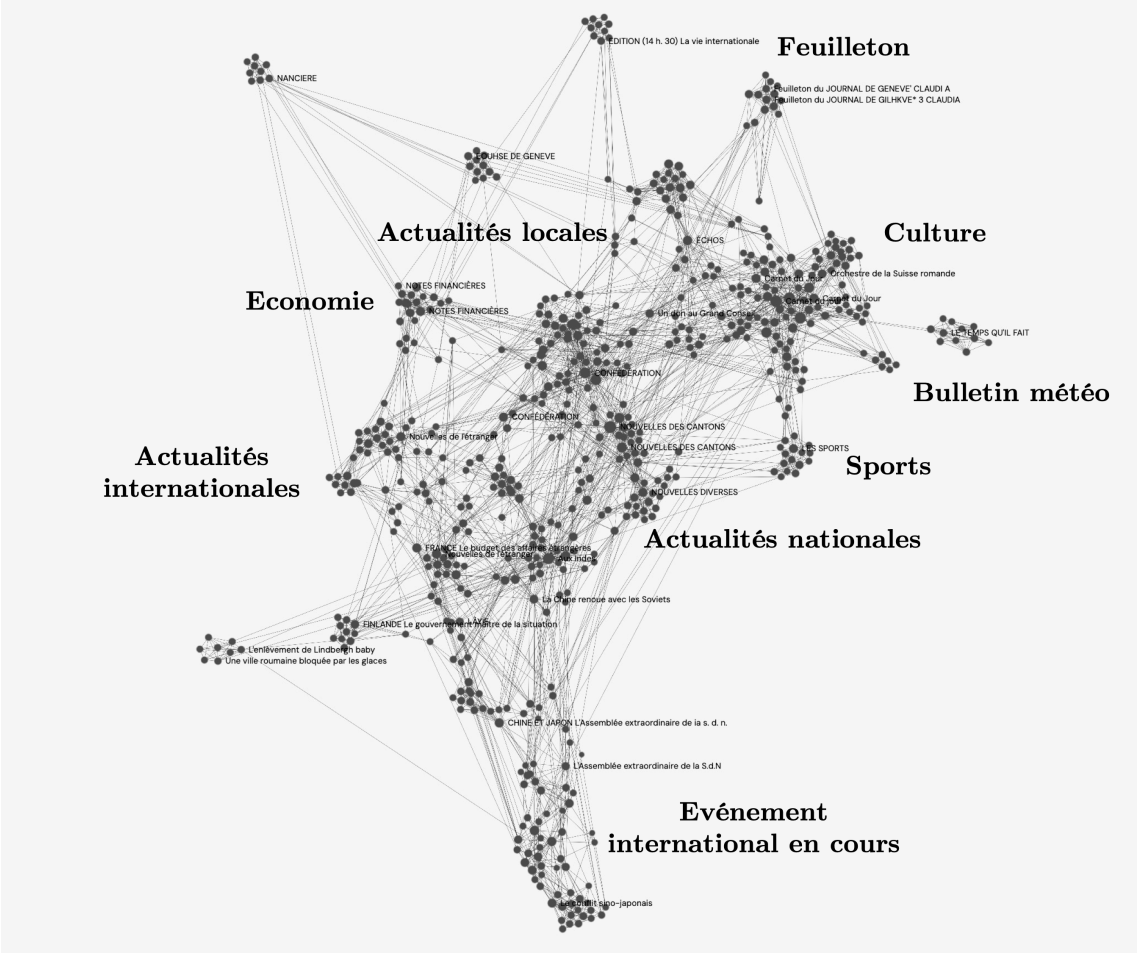


FIGURE 1 – UMAP (sous la forme d'un *7-nearest neighbors network*) de tous les articles du Journal de Genève du 1er au 10 mars 1932. Chaque nœud est un article, deux nœuds sont connectés s'ils font partie des 7 textes les plus proches l'un de l'autre. Exemple produit avec un *notebook* dédié à l'analyse de réseaux de textes [9] et visualisé avec Gephi Lite [6].

2 Les *text embeddings* au service d'une analyse des corpus de presse

Récente addition à la boîte à outils du *Natural Language Processing* (NLP), les *embeddings* (ou "plongement sémantique") complètent des approches d'analyse de grands corpus de presse déjà existantes comme le *topic model* [14] en permettant de mesurer la distance entre plusieurs textes, promettant de "cartographier la signification ou le contenu des phrases et documents sous forme de représentations vectorielles" [15]. Une réduction de la dimensionnalité de cette distance, mesure de similarité sémantique entre ces textes qui s'exprime dans un espace de vecteurs à plusieurs centaines voire milliers de dimensions, permet de produire une représentation graphique synthétique qui aplatit la complexité sémantique du corpus choisi sous la forme d'une carte (UMAP) de clusters de textes.

Au vu de la nature sérielle des grands corpus de presse numérisés, il semble qu'une telle méthode permet de s'inscrire efficacement dans une démarche de *distant reading*. Les cartes produites permettent en effet une première évaluation de la nature du corpus, les clusters étant susceptibles de permettre une classification thématique des différents articles, soit pour comprendre la couverture médiatique globale de l'actualité par la presse pendant une période donnée (Figure 1), soit, à plus petite échelle, pour appréhender les différents discours à l'intérieur d'une actualité particulière.

L'appropriation de telles méthodes de la linguistique computationnelle par les sciences historiques pose une série de questions :

— Dans quels cas ce type d'approche est-elle justifiée ?
— Correspond-elle à ce que d'autres méthodes de classification proposent déjà pour l'exploration de grands corpus de textes ? Dans ce cas, est-ce que la redondance permet de déterminer des classes/clusters robustes ou au contraire rend-elle cette approche superflue ? Si ce n'est pas le cas, est-ce que les différences montrent quelque chose de qualitativement probant, des cas d'usages privilégiés pour les unes ou les autres, ou cela met-il au contraire en évidence un manque de fiabilité de cette nouvelle méthode ?
— Apporte-t-elle des résultats plus probants lorsque le corpus de presse est très vaste et généraliste (auquel cas elle sert à comprendre une sorte de "panorama" global de la couverture médiatique) ou quand les articles ont été soigneusement sélectionnés autour d'un thème ou d'une période resserrée (auquel cas il s'agit avant tout de tracer les différentes manières de couvrir un événement) ?
— Dans quelles conditions est-il possible de faire une analyse conjointe d'un corpus d'articles de presse et d'un corpus d'autres documents historiques ? La réduction de la dimensionnalité permet-elle valablement de comparer les représentations vectorielles de ces deux types de textes ensemble ? Cette comparaison permet-elle d'enrichir l'étude de grands corpus médiatiques ou ne fait-elle que de montrer des évidences ?

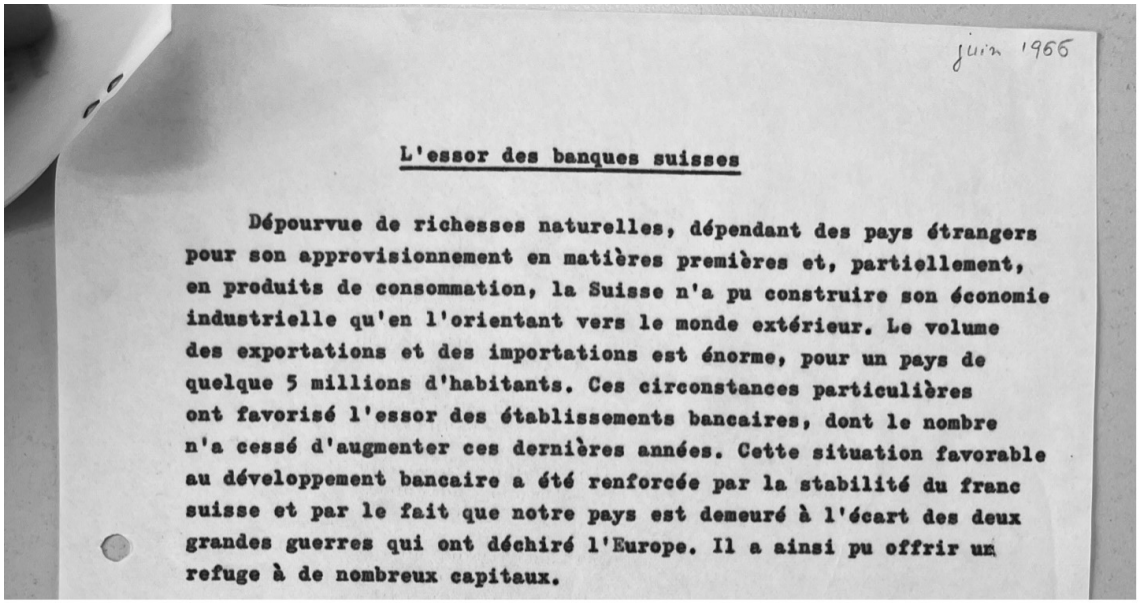
3 Sources et données : actualité bancaire et relations publiques des banques suisses

Afin d'explorer la dimension comparative des *embdeddings*, notre recherche utilise deux corpus distincts qui permettent d'aborder une étude de cas sur l'histoire de la communication des banques suisses. Tout d'abord, elle repose sur les collections de presse numérisées rendues accessibles par le projet Impresso sur sa plateforme (https ://impresso-project.ch/app/) et *via* son API (https ://impresso-project.ch/datalab/). S'ajoutent ensuite des sources tirées de recherches dans les archives de l'Association suisse des banquiers (ASB), conservées aux Archives économiques suisses (Bâle).

L'actualité bancaire offre un objet d'étude particulièrement adapté à l'exploration de grands corpus de presse par *embeddings*, ainsi qu'à la comparaison, par la mesure de similarité d'*embeddings*, de ces corpus et de sources externes. Les banques sont en effet présentes dans la presse quotidiennement et sous des formes qui vont bien au-delà des articles : leur activité est couverte par des dépêches financières, diverses informations tabulaires, de la publicité, etc. L'actualité bancaire se retrouve donc sur un large spectre de la surface éditoriale d'un journal.

En outre, dès la fin des années 1940, les banques cherchent, en Suisse, à influencer collectivement le flux d'information les concernant [12]. S'inscrivant dans une dynamique également observée au Royaume-Uni [16], l'Association suisse des banquiers — organisation patronale et faîtière [8] — établit une commission de relations publiques en 1947 [11]. Celle-ci vise, d'une part, à faire de la "publicité éditoriale", c'est-à-dire à impulser une couverture médiatique favorable aux banques en cultivant des liens personnels, parfois proches, avec les journalistes, et, d'autre part, à produire directement du contenu médiatique souvent non signé (articles de journaux et de revues, émissions de radio, court-métrages, etc.) ou du contenu publicitaire (encarts, spots TV, etc.). Si l'on peut documenter, par un travail d'archives, la mise en place institutionnelle et la fabrication matérielle des relations publiques de l'ASB, la question de l'efficacité de ce travail demeure, comme souvent dans l'étude des "pratiques d'influence" [3] ou de la "communication persuasive organisée" [1], difficile à aborder et à évaluer.

Dès lors que les *embeddings* permettent de mesurer la similarité sémantique entre différents textes, on peut se demander s'ils pourraient être mobilisés pour étudier la correspondance entre un contenu de type relations publiques et une couverture médiatique. Une approche *data-driven* de


juin 1966

L'essor des banques suisses

Dépourvue de richesses naturelles, dépendant des pays étrangers pour son approvisionnement en matières premières et, partiellement, en produits de consommation, la Suisse n'a pu construire son économie industrielle qu'en l'orientant vers le monde extérieur. Le volume des exportations et des importations est énorme, pour un pays de quelque 5 millions d'habitants. Ces circonstances particulières ont favorisé l'essor des établissements bancaires, dont le nombre n'a cessé d'augmenter ces dernières années. Cette situation favorable au développement bancaire a été renforcée par la stabilité du franc suisse et par le fait que notre pays est demeuré à l'écart des deux grandes guerres qui ont déchiré l'Europe. Il a ainsi pu offrir un refuge à de nombreux capitaux.

FIGURE 2 – Extrait d'un texte rédigé par le président de la commission de relations publiques de l'ASB (Hans Bauer), destiné à être publié dans la revue britannique Statist. Ce texte encapsule un certain type de discours sur les banques suisses typique des relations publiques de l'ASB que l'on retrouve généralement dans tout contenu médiatique que cette association produit. Source : Archives économiques suisses (Bâle), PA 650, G 3-0-1160, NF_0408, 1964/08-1967/05, Hans Bauer (Société de Banque Suisse), "L'essor des banques suisses", juin 1966.

ce type permet-elle d'explorer conjointement les structures sémantiques de l'actualité bancaire et des relations publiques collectives des banques suisses ? Les *embeddings* peuvent-ils répondre à la question de savoir si l'ASB parvient ou non à influencer à une échelle significative la couverture des banques dans la presse suisse ?

4 Développer un *notebook* d'analyse comparée

Notre démarche s'inscrit dans une volonté de développer des outils d'analyse accessibles et reproductibles. En cela, notre contribution prolonge nos réflexions sur le développement de *notebooks* publics, amenés à être utilisés par des chercheur.euse.s non spécialistes des méthodes d'humanités numériques. Le projet Impresso, dont l'objectif premier était la constitution d'une base de données d'archives de presse, a ainsi lancé à l'automne 2025 un datalab (https ://impresso-project.ch/datalab/). Comportant un comité éditorial, des règles, une évaluation externe ainsi qu'une citabilité, ce site vise à rendre accessible des *notebooks* utilisant les modèles développés au sein du projet tels que la reconnaissance d'agence de presse ou l'analyse de cooccurrence d'entités nommées. Dans cette optique, nous avons fait de l'analyse comparative de deux corpus un *notebook* voué à être rendu publiquement disponible sur le *datalab* d'Impresso.

Le *notebook* [13] a deux objectifs principaux : d'une part, *embed* les textes présents dans deux corpus distincts, c'est-à-dire assigner un vecteur à haute dimension à chaque texte, et, de l'autre, produire une visualisation unique regroupant l'ensemble de ces données. Il se découpe ainsi plus précisément en trois étapes distinctes : l'*embedding* des textes, la réduction des vecteurs à trois dimensions, et la production d'une représentation visuelle pour l'exploration des données. La première phase repose sur les modèles développés par Impresso. Le projet propose une bibliothèque python accessible que nous avons utilisée en passant par son API. Pour l'*embedding* de textes, Impresso utilise le modèle *multilingual General Text Embeddings* (mGTE), ainsi qu'une de ses

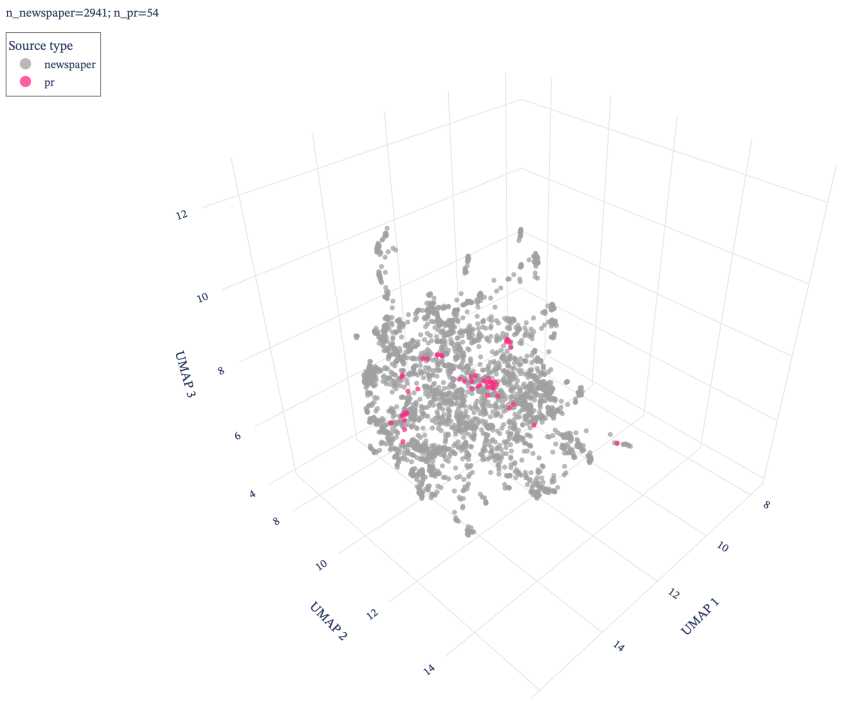


FIGURE 3 – Exemple d'output du *notebook* [13] : comparaison des *embeddings* de deux corpus. En gris, 2941 articles de presse suisse (français et allemand) publiés entre le 01.01.1950 et le 31.12.1980 mentionnant "Association suisse des banquiers" ou "Schweizerische Bankiervereinigung". Les articles sont tirés de la plateforme Impresso (voir la requête complète ici). En rose, 54 textes (articles, articles payés, communiqués de presse et discours officiels) tirés des archives de l'ASB (Archives économiques suisses, Bâle) et de Leo Schürmann, membre de sa commission de relations publiques (Archives fédérales suisses, Berne), rédigés entre 1950 et 1980.

variantes adaptées aux problèmes d'OCR [10]. Le mGTE est très performant tant sur une langue que sur plusieurs [5], ce qui en fait un modèle idéal pour un corpus multilingue comme le nôtre, qui comprend des textes en allemand, en français et en anglais. En second lieu, nous avons utilisé UMAP pour la compression des vecteurs en trois dimensions. UMAP est en effet très efficace pour réaliser ce genre d'opération en préservant à la fois les relations de voisinage locales et la structure globale des données. Enfin, Plotly a été utilisé pour créer une visualisation interactive, en trois dimensions rendant visible les deux collections (Figure 3).

Sachant que le *clustering* des vecteurs indique la proximité sémantique des textes, l'analyse semble tout d'abord mettre en évidence les structures thématiques de chacun des corpus. En second lieu, elle apparaît révéler les aires de correspondance entre ceux-ci. En théorie, elle démontre en effet quels pans du corpus de presse se chevauchent avec le corpus de relations publiques, mettant du même coup en évidence ce qui, de la couverture médiatique générale, n'est pas visé par les relations publiques et, inversement, ce qui des relations publiques ne se retrouve pas dans la presse. Plusieurs questions restent cependant ouvertes :

— Comment interpréter les chevauchements des deux corpus ? Indiquent-ils une simple proximité sémantique fondée sur l'occurrence d'un certain nombre de mots-clés communs ou permet-elle d'approcher plus en profondeur le traitement (cadrage) des différents sujets ?
— Comment prendre en compte l'évolution historique des deux corpus de sources dans la production de visualisations UMAP et la comparaison des vecteurs ? Quelle doit être la granularité temporelle minimum des données historiques utilisées ? Comment comparer non plus uniquement des corpus de données différents, mais des versions de ces corpus à

différentes périodes (p. ex., par tranche de cinq ou dix ans) ?

— Comment tester la robustesse des résultats en allant au-delà du visuel ? Peut-on recourir à des méthodes d'analyse de réseaux (modularité ou autre détection de communautés dans le *k-nearest neighbors network*), de classification thématique automatisée par LLM ou de mesure statistique de proximité des vecteurs afin de mesurer très concrètement et plus précisément la correspondance entre les textes des corpus ? Peut-on, par exemple, associer des mots-clefs à chaque cluster ? Comment intégrer de telles méthodes de vérification dans le pipeline d'analyse ?

5 Conclusion

Parce qu'elle repose sur une expérience de recherche historique qualitative, un processus empirique de développement et des questionnements pédagogiques, cette proposition de communication est autant une contribution aux méthodes computationnelles qu'une réflexion sur l'intégration de celles-ci dans la démarche exploratoire des historien.ne.s. Les expériences menées ici permettent un premier état des lieux des possibilités, qui, si elles sont prometteuses, posent un certain nombre de questions assez classiques des humanités numériques, ou de toute rencontre d'une discipline avec les méthodes quantitatives : aussi rigoureux et reproductibles qu'ils soient, ces résultats se prêtent-ils à une interprétation historique stable ? Autrement dit, s'agit-il d'une méthode purement exploratoire pour les spécialistes de ces corpus ou est-ce que les *text embeddings* sont destinés à se généraliser et prendre une place dans la boîte à outil des historien.ne.s computationnel.le.s [7] ?

Références

[1] Bakir, Vian, Herring, Eric, Miller, David et Robinson, Piers. « Organized Persuasive Communication : A New Conceptual Framework for Research on Public Relations, Propaganda and Promotional Culture ». In : *Critical Sociology 45*, no. 3 (2019), p. 311-328. DOI : 10.1177/0896920518764586.

[2] Bunout, Estelle, Ehrmann, Maud et Clavert, Frédéric. *Digitised Newspapers – A New Eldorado for Historians ?: Reflections on Tools, Methods and Epistemology*. Berlin : De Gruyter Oldenbourg, 2023. DOI : 10.1515/9783110729214.

[3] Cohen, Yves. « Une école de liberté historiographique ». In : *Critique 843–844*, no. 8-9 (2017), p. 700-711. DOI : 10.3917/criti.843.0700.

[4] Ehrmann, Maud, Düring, Marten, Neudecker, Clemens et Doucet, Antoine. « Computational Approaches to Digitised Historical Newspapers (Dagstuhl Seminar 22292) ». In : *Dagstuhl Reports 12*, no. 7 (2023), p. 112-179. DOI : 10.4230/DagRep.12.7.112.

[5] Enevoldsen, Kenneth et al. « MMTEB : Massive Multilingual Text Embedding Benchmark ». In : *International Conference on Representation Learning 2025* (2025), p. 101715-101771. DOI : 10.48550/arXiv.2502.13595.

[6] Girard, Paul, Jacomy, Alexis, Simard, Benoît et Jacomy, Mathieu. « Gephi Lite : A Lighter Web Based Version of Gephi ». In : *Digital Humanities 2025*. Lisbonne, Portugal : Alliance of Digital Humanities Organizations, 2025.

[7] Grandjean, Martin. « The Digital Historian's Craft ». In : *Handbook of Digital and Computational Research Methods*, sous la dir. d'Anders Koed Madsen et Anders Kristian Munk. Cheltenham : Edward Elgar Publishing, 2026, p. 32-47. DOI : 10.4337/9781802208993.00009.

[8] Guex, Sébastien et Mazbouri, Malik. « De l'association des représentants de la banque en Suisse (1912) à l'Association Suisse des Banquiers (1919). Genèse et fonctions de l'organisation faîtière du secteur bancaire suisse ». In : *Genèse des organisations patronales en Europe (19e–20e siècles)*, sous la dir. de Danièle Fraboulet et Pierre Vernus. Rennes : Presses universitaires de Rennes, 2012, p. 205-225.

[9] Jacomy, Mathieu. « Text List to Semantic Network (Embedding) ». 2025. URL : https://jacomyma.github.io/mapping-controversies/nb/.

[10] Michail, Andrianos, Opitz, Juri, Wang, Yining, Meister, Robin, Sennrich, Rico et Clematide, Simon. « Cheap Character Noise for OCR-Robust Multilingual Embeddings ». In : *Findings of the Association for Computational Linguistics : ACL 2025*, sous la dir. de Wanxiang Che, Joyce Nabende, Ekaterina Shutova et Mohammad Taher Pilehvar. Vienne : Association for Computational Linguistics, 2025, p. 11705-11716. DOI : 10.18653/v1/2025.findings-acl.609.

[11] Michelet, Arthur. « Construire l'information bancaire : La publicité de l'Association Suisse des Banquiers et Les médias de masse (1948–1975) ». In : *L'influence et ses pratiques*, sous la dir. d'Yves Cohen, Irene Di Jorio et Hugo Souza de Cursi. Lille : Presses universitaires du Septentrion, 2026, sous presse.

[12] Michelet, Arthur. « The Transmedia Publicity of Swiss Banks : Public Relations Strategies and Mass Media Coverage (1960–1977) ». In : *Transmedia History : Circulations, Reconfigurations and New Methodologies*. Lausanne, Suisse, mars 2025, p. 67-71. DOI : 10.5281/zenodo.15046947.

[13] Michelet, Arthur et Grandjean, Martin. « Comparing Corpora Using Embeddings ». Zenodo. 2026. DOI : 10.5281/zenodo.19006513.

[14] Murugaraj, Keerthana, Lamsiyah, Salima, During, Marten et Theobald, Martin. « Mining the Past : A Comparative Study of Classical and Neural Topic Models on Historical Newspaper Archives ». In : *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, sous la dir. de Mika Hämäläinen, Emily Öhman, Yuri Bizzoni, So Miyagawa et Khalid Alnajjar. Albuquerque : Association for Computational Linguistics, 2025, p. 452-463. DOI : 10.18653/v1/2025.nlp4dh-1.39.

[15] Opitz, Juri, Möller, Lucas, Michail, Andrianos, Padó, Sebastian et Clematide, Simon. « Interpretable Text Embeddings and Text Similarity Explanation : A Survey ». 2025. DOI : 10.48550/arXiv.2502.14862.

[16] Reveley, James et Singleton, John. « Clearing the Cupboard : The Role of Public Relations in London Clearing Banks' Collective Legitimacy-Seeking, 1950–1980 ». In : *Enterprise & Society 15*, no. 3 (sept. 2014), p. 472-498. DOI : 10.1093/es/khu033.

[17] Underwood, Ted. « A Genealogy of Distant Reading ». In : *Digital Humanities Quarterly 11*, no. 2 (2017). DOI : 10.63744/ktwb2atwsbep.