

Health, vitality, and trust: embedding-based cross-lingual retrieval of medicinal advertisements in European and colonial press

Vy Cao¹ 

¹ Centre for Contemporary and Digital History (C²DH), Department of Digital History and Historiography, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Abstract

This paper presents a case study of medicinal advertisements published in the early twentieth century across European and colonial press. It offers a critical reflection on the retrieval and evaluation of multilingual corpora within large-scale media archives, with particular attention to their validity and relevance for historical research. The study develops a reproducible workflow for embedding-based retrieval, focusing on keyword-level queries as a first stage of cross-lingual corpus construction. Retrieved materials are evaluated through a two-step protocol combining cosine similarity ranking and expert close reading, enabling a comparative assessment of semantic precision and discursive coherence across selected keywords. The paper argues that embeddings provide a powerful alternative to string-based search in multilingual and orthographically unstable corpora, while demonstrating that quantitative cluster stability is a necessary but insufficient condition for meaningful retrieval. This finding has particular implications for languages such as Vietnamese, where compound terms are prone to lexical and semantic drift in cross-lingual vector spaces.

Keywords: medicinal advertising, European and colonial press, cross-lingual embedding-based retrieval, dataset curation, qualitative evaluation

1 Introduction

The growing availability of digitized newspapers raises new methodological challenges for multilingual corpus construction, particularly when working across linguistically and historically heterogeneous collections [1; 7]. To address these challenges, this study proposes a cross-lingual retrieval pipeline based on embeddings within a shared vector space. Developed within the *Impresso* project¹, this experiment offers a critical reflection on integrating embeddings into large-scale and multilingual archives. It participates in a broader research project examining the construction and circulation of medicinal advertisements in European and colonial press, specifically those promoting “health” and “vitality”.

2 Printing and modernity in South-East Asia

Since the eighteenth century, Christian missionaries played an increasingly important role in the transmission of new printing techniques in Southeast Asia, strengthening maritime routes to China and consolidating relations with Thailand, Singapore, Malaysia, and the Indochinese peninsula [2].

Vy Cao. “Health, vitality, and trust: embedding-based cross-lingual retrieval of medicinal advertisements in European and colonial press.” In: *Computational Cultural Science Workshop 2026*, ed. by Suzanne Mpouli, Jean-Baptiste Camps, Florian Cafiero, and Chahan Vidal-Gorène. Vol. 6. Anthology of Computers and the Humanities. 2026, 15–22. <https://doi.org/10.63744/KgzC1tjzIESz>.

© 2026 by the authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

¹ Impresso mediates European digitized collections from GLAM partners such as the National Library of France (BnF), the National Library of Luxembourg (BNL), and the Swiss Federal Archives (SFA). As of 2025, it has gathered a total of 134 titles, and nearly 53 million content items and 4 million images. It is funded by the Swiss National Science Foundation and the Luxembourg National Research Fund.

The production of typefaces in regional languages was largely dominated by Protestant missionaries from the United States and the United Kingdom, before type-making practices spread to local artisans across Asia. By the mid-nineteenth century, printing houses and type foundries had been established in Burma, India, Malaysia, Thailand, and Singapore².

In 1862, the Colonial printing house was officially established in Saigon, the capital of French Cochinchina. Prior to this, significant printing activities in Vietnamese had been conducted outside the territory. Thailand served as a refuge for Vietnamese Catholic communities, fostering diasporic and missionary connections that catalyzed early Vietnamese-language printing initiatives abroad³. The presence of French missionary networks in China also contributed to the circulation of works in Vietnamese.

At the turn of the twentieth century, French Cochinchina experienced a profound transformation of its publishing landscape [2; 3]. The gradual decline of the Colonial printing house in 1904 opened the field to private entrepreneurs, while a new regulatory framework granted printers greater autonomy. This shift generated a dual dynamic in which French publishers consolidated their activities, while Vietnamese-run workshops revitalized local cultural production. In this evolving landscape, advertisements for Vietnamese medicinal products began to flourish. Leading pharmacies commissioned writers, Fine Arts graduates, and traditional artisans to craft their visual identity, leveraging print culture as a space for commercial visibility and discursive legitimation.

3 Research design

This paper asks whether semantic similarity-based retrieval can identify European press materials meaningfully comparable to Vietnamese medicinal advertisements, using heuristic categories of trust as a shared analytical frame. The research design unfolds in two complementary stages. First, keywords and textual segments in Vietnamese are extracted from a digitized catalogue of Vietnamese medicinal products. This curated corpus will serve as a training set for a classification model operationalizing “trust” across four dimensions, to be fully developed in subsequent work:

- **Institutional trust**, indexed by references to professional institutions and figures, as well as markers of credentialing (e.g., hospitals, doctors, diplomas);
- **Procedural trust**, signaled by prescriptive and evidentiary elements such as instructions, dosage, prices, and appeals to proof;
- **Interpersonal trust**, captured through relational formats including letters and testimonials;
- **Affective trust**, articulated through discourses mobilizing folklore, family, nature, and other repertoires of care.

Second, the annotated corpus is embedded into the *Impresso* semantic vector space using the multilingual GTE (General Text Embeddings) model [4]. The choice of GTE is determined by the existing project infrastructure, which this experiment aims to critically evaluate as a retrieval tool for historical and multilingual materials [5].

This paper focuses on the first stage of the pipeline, interrogating whether a reliable corpus can be curated at the keyword level — the most controlled and interpretable entry point, before any extension to longer textual units. By documenting this early stage of corpus formation, the paper contributes to methodological transparency and the reuse of data and workflows across research contexts [6].

² In Calcutta, the Serampore Baptist Mission established its own printing press as early as 1800.

³ In 1838, Jean-Baptiste Pallegoix founded the Assumption printing house in Bangkok, publishing works in both Latin and romanized Vietnamese.

4 Embedding-based cross-lingual curation: workflow and data

This section presents the workflow underlying the experiment, from the integration of external Vietnamese materials into a shared embedding space to the evaluation of retrieved dataset⁴.

4.1 A Vietnamese pharmaceutical catalogue as source data

Data are extracted from the first Sino-Vietnamese catalogue published in Cholon in 1925, by one of the most prosperous pharmacies called Nhị Thiên Đường⁵. This catalogue, *Vệ sanh chỉ nam* (A guide to hygiene), contains more than 300 pages on medicinal products and remedies. Their composition made use of diverse narrative techniques and devices to attract new literate Vietnamese readers, and promote their business in traditional medicine⁶.



Figure 1: “Phụ khoa kim phụng hườn” (Women’s remedy), *Vệ sanh chỉ nam* (A guide to hygiene), Saigon, Imprimerie de l’Union Nguyễn Văn Của, 1925, pp. 32-33. Source: Gallica.

A significant portion of the Vietnamese vocabulary in these catalogues resists automatic translation. Regional and temporal spelling variations reflect the fluid standardization of romanized Vietnamese in the early twentieth century, highlighting a critical limitation of string-based keyword searches. Embedding-based search presents a compelling alternative to capture these semantic nuances.

An initial set of Vietnamese keywords was manually extracted from the catalogue and aligned with English glosses and semantic categories. Table 1 presents this curated list, which serves as the basis for subsequent embedding and retrieval procedures across multilingual corpora.

⁴ This experiment will be published as a Jupyter Notebook, offering a transparent and documented workflow for external evaluation and reuse.

⁵ In 2021, the BnF digitized more than 10,000 printed materials from the Indochinese collection, gathered through the colonial legal deposit system operating in Indochina between 1922 and the early 1940s. This initiative exposed new challenges including the encoding of Vietnamese diacritics, metadata normalization, and the limitations of OCR for historical documents in low-resourced languages.

⁶ Over a third of the 145 Sino-Vietnamese catalogues in the Gallica collection weave literary content directly into the pharmaceutical listings.

Keywords (VI)	Keywords (EN)	Word Categories
bệnh	illness, disease	generic medical vocabulary
cứu	to cure, to treat	generic medical vocabulary
lương y	traditional physician	generic medical vocabulary
người bệnh	patient, sick person	generic medical vocabulary
thầy thuốc	healer, practitioner	generic medical vocabulary
vệ sinh	hygiene, sanitation	generic medical vocabulary
y viện	medical institution, clinic	generic medical vocabulary
đau	pain	physiological condition
lành mạnh	healed, healthy	physiological condition
thai	pregnancy	physiological condition
gân xương	muscles and bones, joints	anatomical reference
huyết	blood	anatomical reference
hườn	medicinal pill (often honey-based)	medicinal product
tán	powder medicine	medicinal product
thuốc men	medicines, remedies	medicinal product
á phiện	opium	ingredient for remedies
đồng tiện	child's urine	ingredient for remedies
lá sương phủ	frost-covered leaves	ingredient for remedies
bụng đau	abdominal pain	disease or symptom
đầu nóng	fever	disease or symptom
đờm	phlegm	disease or symptom
ghẻ chốc	scabies, skin infection	disease or symptom
hạch	swollen lymph nodes, glands	disease or symptom
ho	cough	disease or symptom
ỉa mửa	diarrhea, vomiting	disease or symptom
lậu	gonorrhea	disease or symptom
nhiễm gió	catching a cold	disease or symptom
phát nhiệt	fever	disease or symptom
phong đàm	phlegm-related disorder	disease or symptom
thổ huyết	vomiting blood	disease or symptom
trĩ	hemorrhoids	disease or symptom
ung thư u nhọt	tumors, sores, abscesses	disease or symptom

Table 1: Vietnamese keywords and their English equivalents with semantic categories.

This selection functions as a heuristic entry point into the corpus, capturing recurrent terms related to medicinal products, symptoms, and substances.

4.2 Retrieval evaluation: from cosine similarity to close reading

The evaluation focuses on the thirty-two Vietnamese keywords listed above. For each keyword, ten documents are retrieved and organized in an internal ranking from 0 to 9, yielding a total of 320 retrieved documents. Cosine similarity scores provide a first quantitative indicator of retrieval quality. The ten highest-performing keywords yield scores between 0.714 and 0.760, while the five lowest-ranked score between 0.638 and 0.655, forming a distinct lower band⁷.

⁷ No hard threshold is applied, as the goal is to observe the full gradient of retrieval quality rather than to filter results in advance.

The protocol adopts a qualitative, expert-driven⁸ approach designed to probe the semantic behavior of the model across different types of terms [6]. Expert evaluation proceeds in two complementary steps. First, the ten keywords with the highest cosine similarity score for their top retrieved document, and the ten retrieved documents with the highest individual scores across the entire dataset of 320 results, are surveyed comparatively [8]. Since each keyword generates ten retrieved documents ranked from 0 to 9, a single high-performing keyword can occupy multiple positions in the top retrieved documents list. This parallel observation identifies whether best-performing keywords also generate the most semantically proximate documents.

This comparison reveals a high degree of overlap. The six highest-ranked keywords — *bụng đau* (abdominal pain), *thuốc men* (medicines, remedies), *đau* (pain), *trĩ*, *huyết* (blood), and *lành mạnh* (healed, healthy)— appear in both lists with identical scores, confirming that the best-performing keywords also generate the most semantically meaningful documents. The main divergence between the two lists concerns *thuốc men* (medicines, remedies), which occupies three positions in the top 10 retrieved documents. This suggests that *thuốc men* produces the most coherent cluster in the vector space. On the contrary, *thầy thuốc* (healer, practitioner), *nhiễm gió* (catching a cold), and *cứu* (to cure, to treat) appear in the top keywords but not in the top documents. This comparison illustrates the complementary value of the two perspectives. The top documents reveal cluster density, while the top keywords provide a more balanced view of each term’s relative performance across the full keyword set (Table 2).

Top retrieved documents			Top keywords		
Score	VI	EN	Score	VI	EN
0.760	bụng đau	abdominal pain	0.760	bụng đau	abdominal pain
0.741	thuốc men	medicines, remedies	0.741	thuốc men	medicines, remedies
0.732	đau	pain	0.732	đau	pain
0.731	thuốc men	medicines, remedies	0.730	trĩ	hemorrhoids
0.730	thuốc men	medicines, remedies	0.727	huyết	blood
0.730	trĩ	hemorrhoids	0.726	lành mạnh	healed, healthy
0.727	huyết	blood	0.726	đầu nóng	fever
0.726	bụng đau	abdominal pain	0.725	thầy thuốc	healer, practitioner
0.726	lành mạnh	healed, healthy	0.720	nhiễm gió	catching a cold
0.726	đầu nóng	fever	0.714	cứu	to cure, to treat

Table 2: Top retrieved documents and top keywords by cosine similarity score.

Second, for each of the ten highest-ranked keywords, the three top documents retrieved within their internal ranking are subjected to detailed close reading. This internal variation is treated as an analytical signal, indicating the degree to which a keyword maps onto a coherent discursive field in the European press corpus. Each retrieved document is evaluated along three qualitative dimensions:

- **Semantic match**, defined as the degree to which the retrieved content corresponds to the meaning of the query term.
- **Discourse type**, which identifies the genre of the document (e.g., advertisement, medical advice).
- **Cluster stability**, defined as the degree of semantic coherence among the top-ranked results for a given query.

⁸ Close reading was conducted by the author, with full reading proficiency in Vietnamese.

A preliminary survey of cosine similarity scores reveals meaningful variation in cluster stability (Table 3). Six keywords produce stable clusters: *thuốc men* (medicines, remedies), *huyết* (blood), *thầy thuốc* (healer, practitioner), *nhễm gió* (catching a cold), *cứu* (to cure, to treat), and *đau* (pain). Four keywords produce unstable clusters: *trĩ* (hemorrhoids), *đầu nóng* (fever), *bụng đau* (abdominal pain), and *lành mạnh* (healed, healthy), with *trĩ* representing the most extreme case with a drop of 0.063 between the first and third retrieved documents. Notably, cluster stability does not correlate directly with the score of the highest-scoring retrieved document: *bụng đau* yields the highest first result across all keywords yet produces an unstable cluster, while *cứu* and *thầy thuốc*, with more modest scores, generate the most coherent retrievals.

Keyword (VI)	Keyword (EN)	Cosine Similarity		
		Rank 0	Rank 1	Rank 2
Stable clusters				
thuốc men	medicines, remedies	0.741	0.731	0.730
huyết	blood	0.727	0.725	0.710
thầy thuốc	healer, practitioner	0.725	0.725	0.720
nhễm gió	catching a cold	0.720	0.707	0.703
cứu	to cure, to treat	0.714	0.711	0.707
đau	pain	0.732	0.714	0.709
Unstable clusters				
trĩ	hemorrhoids	0.730	0.667	0.667
đầu nóng	fever	0.726	0.683	0.677
bụng đau	abdominal pain	0.760	0.726	0.706
lành mạnh	healed, healthy	0.726	0.692	0.688

Table 3: Internal distribution of cosine similarity scores for the top three retrieved documents of the ten highest-scoring keywords, grouped by cluster stability.

The close reading evaluation maps the conditions of success and failure of embedding-based retrieval with analytical precision. Semantic match is strongest for terms designating specific yet popular medical vocabularies — *bụng đau* (abdominal pain), *huyết* (blood), *thầy thuốc* (healer) — where retrieved documents converge on coherent and discursively diverse clusters. It weakens for polysemous terms, such as *đau* (pain), *lành mạnh* (healthy), and breaks down for compound terms whose components activate divergent semantic fields in the target languages. In this case, *nhễm gió* (catching a cold) retrieves meteorological rather than clinical content, and *cứu* (to cure) consistently retrieves rescue narratives rather than medical treatment. The case of *trĩ* (hemorrhoids) illustrates the outer limit of the model’s reach — a clinically precise term with low lexical frequency in generalist press produces both quantitative and qualitative instability.

The distribution of discourse types across keywords constitutes another interesting result. Medicinal advertisements surface predominantly for terms designating concrete symptoms or remedies - *bụng đau* (abdominal pain), *thuốc men* (medicines, remedies), while medical history articles and narratives dominate for terms with broader physiological resonance such as *huyết* (blood) and *đau* (pain). Literary texts emerge specifically for terms carrying relational or moral dimensions, such as *thầy thuốc* (healer, practitioner) and *cứu* (to cure, to treat). This genre distribution suggests that the specificity and cultural aspect of a query term shapes the discursive field it activates in the vector space. It has direct implications for the subsequent analysis of trust configurations in medicinal advertisements across European and colonial press.

The close reading analysis also complicates the preliminary classification on cluster stability.

Two quantitatively stable clusters — *nhiễm gió* (catching a cold) and *cứu* (to cure) — conceal distinct qualitative failures that only close reading can expose. These failures are particularly revealing in a cross-lingual context involving Vietnamese, a language in which meaning is built through the compounding of monosyllabic morphemes. Lexical drift, as observed with *nhiễm gió*⁹, occurs when the model captures one semantically dominant component of a compound term, producing meteorological rather than medical retrievals. Semantic shift, as observed with *cứu*, occurs when the model activates a legitimate but contextually inappropriate sense of a polysemous term — here “to rescue” rather than “to cure”, because that sense is more densely represented in the target corpus. In both cases, the stability of cosine similarity scores creates a false impression of coherence that only close reading can expose. Embedding models tend to decompose Vietnamese compounds and privilege the most lexically frequent component in the target corpus — a form of compositional interference that string-based search would partially avoid by treating the compound as an invariant unit, reinforcing the case for combining both retrieval strategies [5; 8].

5 Conclusion and future perspectives

This paper has introduced a reproducible pipeline for curating cross-lingual historical datasets within an embedded vector space. By embedding thirty-two Vietnamese keywords from a 1925 pharmaceutical catalogue into the *Impresso* multilingual GTE vector space, the experiment tested whether semantic proximity could support cross-lingual corpus construction as a prerequisite for analyzing configurations of trust in European and colonial medicinal advertising. This research argues that embeddings are most productive when treated as heuristic instruments whose outputs require qualitative evaluation. The combination of cosine similarity ranking and close reading constitutes the methodological core of the proposed pipeline.

The two-step evaluation produced a nuanced set of results. First-rank cosine similarity is a reliable but partial indicator of retrieval quality, and quantitative cluster stability is a necessary but insufficient condition for discursive coherence. Stable clusters may conceal lexical or semantic shift, while unstable clusters may still yield historically meaningful results. The results also invite reflection on keyword selection. Terms designating common medical conditions consistently outperform polysemous or clinically precise terms, as they map onto well-represented discursive fields in the European press corpus. Compound terms in Vietnamese pose additional cross-lingual challenges due to compositional interference in the vector space.

The next phase will extend the pipeline in two directions. First, the experiment will vary the length and discourse type of query units to evaluate how textual granularity affects retrieval precision and thematic coherence. Second, as the corpus scales beyond feasible expert close reading, LLM-assisted evaluation will be explored as a complementary assessment layer, while expert judgment is reserved for ambiguous and historically sensitive cases. The reliability and limitations of LLM-assisted evaluation in this multilingual and historical context will themselves constitute an object of critical inquiry.

References

- [1] Bingham, Adrian. “The Digitization of Newspaper Archives: Opportunities and Challenges for Historians”. In: *Twentieth Century British History* 21, no. 2 (2010), pp. 225–231. DOI: 10.1093/tcbh/hwq007. URL: <https://academic.oup.com/tcbh/article/21/2/225/1703826>.

⁹ The term *gió* means “wind”.

- [2] Cao, Vy. “Between the Sacred and the Secular: Publishing, Books, and Everyday Life in Colonial Cochinchina”. In: *Vietnam Over the Long Twentieth Century*, ed. by Liam C. Kelly and Gerard Sages. Global Vietnam: Across Time, Space and Community. Singapore: Springer, 2024, pp. 85–100. DOI: 10.1007/978-981-97-3611-9_5. URL: https://link.springer.com/chapter/10.1007/978-981-97-3611-9_5.
- [3] Cao, Vy. “Des presses coloniales aux ateliers vietnamiens. Production et marché du livre en Cochinchine française (1900–1945)”. In: *Histoire et civilisation du livre. Revue internationale* 21 (2025), pp. 135–163. DOI: 10.47421/HCL_21_135-163. URL: https://revues.droz.org/HCL/article/view/HCL_21_135-163.
- [4] Gomez-Perez, Jose Manuel, Denaux, Ronald, and Garcia-Silva, Andres. “Understanding Word Embeddings and Language Models”. In: *A Practical Guide to Hybrid Natural Language Processing: Combining Neural Models and Knowledge Graphs for NLP*, ed. by Jose Manuel Gomez-Perez, Ronald Denaux, and Andres Garcia-Silva. Cham, Switzerland: Springer International Publishing, 2020, pp. 17–31. ISBN: 978-3-030-44830-1. DOI: 10.1007/978-3-030-44830-1_3. URL: https://link.springer.com/chapter/10.1007/978-3-030-44830-1_3.
- [5] Michail, Andrianos, Clematide, Simon, and Sennrich, Rico. “Examining Multilingual Embedding Models Cross-Lingually Through LLM-Generated Adversarial Examples”. 2025. arXiv: 2502.08638 [cs.CL]. URL: <https://arxiv.org/abs/2502.08638>.
- [6] Nguyen, Cindy Anh and Rojas, Alejandro Alvarado. “Exploratory Computation in Digital Humanities: A Qualitative Evaluation Framework”. In: *Journal of Open Humanities Data* 12, no. 50 (2026), pp. 1–13. DOI: 10.5334/johd.500. URL: <https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.500>.
- [7] Oberbichler, Sarah, Boroş, Emanuela, Doucet, Anne, Marjanen, Joonas, Pfanzelter, Elisabeth, Rautiainen, Jaakko, and Tolonen, Mikko. “Integrated Interdisciplinary Workflows for Research on Historical Newspapers: Perspectives from Humanities Scholars, Computer Scientists, and Librarians”. In: *Journal of the Association for Information Science and Technology* 73, no. 2 (2022), pp. 225–239. DOI: 10.1002/asi.24565. URL: <https://onlinelibrary.wiley.com/doi/10.1002/asi.24565>.
- [8] Ruder, Sebastian, Vulić, Ivan, and Søgaard, Anders. “A Survey of Cross-lingual Word Embedding Models”. In: *Journal of Artificial Intelligence Research* 65 (2019), pp. 569–631. DOI: 10.1613/jair.1.11640. URL: <https://www.jair.org/index.php/jair/article/view/11640>.