Cross-Lingual Abstractive Keyphrase Generation for Historical Newspapers

Simon Clematide[1], **Jenifer Meyer**[1], **Juri Opitz**[1], **Maud Ehrmann**[2], **Kaspar Beelen**[3]
[1]Department of Computational Linguistics, University of Zurich (UZH), Switzerland,
[2]Digital Humanities Laboratories, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[3]School of Advanced Study, University of London, United Kingdom
<firstname>.<lastname>@{uzh.ch,epfl.ch,sas.ac.uk}

Abstract

We investigate large language models (LLMs) for cross-lingual abstractive keyphrase generation from historical newspapers. The task consists of producing a small set of English keyphrases for articles written in German, French, and Luxembourgish, combining translation, abstraction, and normalization. We conduct a human-centered pilot study comparing model outputs using human selections, LLM-as-judge assessments, and inter-annotator agreement analysis, followed by a medium-scale application to multilingual data from the Impresso corpus. Results show that LLM-generated keyphrases can support semantic enrichment and exploratory analysis of historical collections, while highlighting the subjective and methodologically challenging nature of keyphrase evaluation.

1. Introduction

Digitized historical newspapers provide unprecedented access to past public discourse, but their effective exploration remains challenging due to optical character recognition (OCR) noise, multilinguality, diachronic language variation, and sparse metadata (Bingham, 2010). Semantic enrichment techniques are therefore essential to support retrieval, browsing, and large-scale exploratory analysis (Düring et al., 2021).

Keyphrase-based indexing provides a compact abstraction of document content and follows a long archival tradition of concept-based indexing and thesaurus-driven cataloguing. Automatic keyphrase generation has therefore become essential for large digitized collections. However, most existing methods are designed for contemporary, monolingual, and relatively clean text, limiting their applicability to noisy historical data and low-resource languages (Chiron et al., 2017; van Strien et al., 2020; Ehrmann et al., 2023b).

Large newspaper digitization initiatives further highlight the need for semantic enrichment (Ridge et al., 2019; Neudecker, 2022; Ehrmann et al., 2023a). While full-text search enables lexical lookup, exploratory research often requires higher-level thematic access across time periods, languages, and publication contexts (Gaillard, 2022; Bunout et al., 2023).

Conceptual keyphrases provide a compact abstraction layer that can be aggregated, compared, and visualized across collections. In multilingual settings, English keyphrases additionally function as a pivot representation enabling cross-lingual comparison. They can therefore complement search, faceted navigation, and other semantic access methods, and may support exploratory interfaces by suggesting concepts users did not initially query (Whitelaw, 2015; Düring et al., 2024).

Recent instruction-following LLMs offer a promising alternative (Ouyang et al., 2022). These models can jointly perform translation, abstraction, and normalization, making them suitable for generating conceptual keyphrases across languages and time periods. We evaluate their use for cross-lingual conceptual keyphrasing of historical newspapers and assess their suitability for semantic enrichment of multilingual collections.

Task Definition We study cross-lingual abstractive keyphrase generation for historical newspapers, primarily drawn from Swiss and Luxembourgish collections. Given an article written in a source language (German, French, or Luxembourgish), the goal is to generate a small set (typically three to five) of *conceptual keyphrases in English* that summarize its main topics.

This task differs from classical extractive keyphrase extraction in three ways. First, the keyphrases need not occur verbatim in the source text but may be semantically implied. Second, they are produced in a target language (English), introducing an explicit cross-lingual abstraction step. Third, the objective is conceptual coverage and topical diversity rather than surface-form fidelity. We use the term *abstractive* to denote keyphrases that need not occur verbatim in the source article. Instead, the model is asked to infer higher-level conceptual descriptors, translate them into English, and normalize across OCR noise, historical spelling, and multilingual variation. In this sense, the task combines summarization, translation, and semantic abstraction rather than surface-form extraction alone.

Named entities (persons, locations, organizations, events) are deliberately excluded from the

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 218–223

> You are an archivist for historical newspaper articles who indexes historical newspaper articles with conceptual keyphrases in English. Given a JSON object containing metadata and a historical newspaper article, please index with an adequate number (between 3 and 5) of most relevant keywords in English in JSON format. Do not create keywords consisting of names for persons, locations, or events. In addition, add one summary sentence in English.

Figure 1: System prompt for keyphrase abstraction.

keyphrase inventory. This design choice reflects the availability of a dedicated named entity recognition and linking pipeline within the broader semantic enrichment workflow. By separating conceptual keyphrasing from entity-centric annotation, the approach avoids redundant effort and allows keyphrases to focus on higher-level thematic dimensions.

The keyphrases function as lightweight semantic descriptors for indexing, clustering, and exploratory analysis rather than as gold-standard annotations.

2. Related Work

Prior work on keyphrase extraction and generation typically assumes contemporary monolingual data. Historical texts pose additional challenges, including OCR noise and linguistic variation, and reliable gold standards are difficult to obtain (Piotrowski, 2012; McGillivray and Tóth, 2020).

Recent studies address multilingual keyphrase generation, for example through retrieval-augmented models (Gao et al., 2022) or multilingual datasets such as EUROPA (Salaün et al., 2024). However, cross-lingual abstractive keyphrase generation for noisy historical texts remains underexplored.

Recent work has begun to examine large language models for zero-shot keyphrase extraction and generation more systematically. Mohan et al. (2025) investigate instruction-tuned LLMs for zero-shot keyphrase generation and show that increasingly specialized instructions do not consistently improve results, whereas multi-sample generation with aggregation can yield clear gains. In a complementary setting, Kang and Shin (2025) study zero-shot keyphrase extraction and show that prompt design matters substantially: simple prompts can already be competitive, task-relevant role prompting often helps, and combining direct extraction with candidate-based selection can further improve performance. Together, these studies suggest that LLM-based keyphrase methods are highly sensitive to prompting and decoding choices, while also confirming their promise as flexible zero-shot alternatives to more traditional extraction and generation pipelines.

Given these challenges, we evaluate systems using an exploratory, human-centered framework that prioritizes adequacy, consistency, and downstream usefulness.

	DE1	DE2	FR1	FR2	LB1	LB2
Claude-3.5-Sonnet	3	5	4	5	4	4
DeepSeek-V3	3	5	5	5	5	5
GPT-3.5 Turbo	5	5	5	4	5	5
GPT-4o mini	3	5	3	5	5	5
# Unique KP	12	12	14	18	15	16

Table 1: Keyphrase counts by model and article

	DeepSeek-V3	GPT-3.5 Turbo	GPT-4o mini
Claude 3.5 Sonnet	10.4%	5.9%	4.1%
DeepSeek-V3	—	9.6%	12.5%
GPT-3.5 Turbo	—	—	8.3%

Table 2: Pairwise overlap (%) between models.

3. Pilot Study: Human and Model Judgments

To implement this evaluation strategy, we conducted a pilot study comparing human and model judgments.

Data and Models We selected a pilot sample (N=6) of historical newspaper articles from the *Impresso* corpus,[1] with two articles each in German, French, and Luxembourgish, covering different periods. Several LLMs (Claude 3.5 Sonnet, DeepSeek-V3, GPT-3.5 Turbo, GPT-4o mini) were prompted, using provider-default decoding settings, to generate English keyphrases using the instruction shown in Figure 1. The user prompt provided the article text together with two metadata fields, namely the newspaper title and the publication date.

Inter-Model Agreement Keyphrase counts are similar across models (Table 3), reflecting adherence to the prompt constraint. However, pairwise overlap scores remain low (Table 2), suggesting substantial variation in model interpretations of document content. Applying a stemmer before computing overlap does not substantially alter the results.

Low pairwise overlap should not be interpreted straightforwardly as system failure. For conceptual keyphrasing, especially in a cross-lingual historical setting, multiple non-overlapping but semantically plausible outputs may adequately represent the same article. At the same time, this variability limits the usefulness of strict lexical overlap as a primary evaluation criterion and cautions against strong claims of comparative superiority.

[1] https://impresso-project.ch

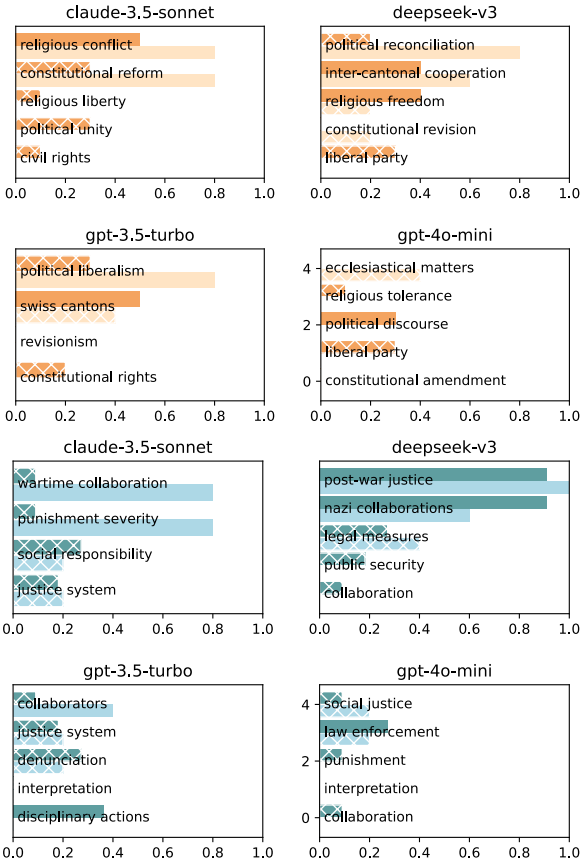

Figure 2: Human and GPT-4o preferences over keyphrases by four models for two representative articles: a French article (1873) on constitutional reform debates in Switzerland (top) and a Luxembourgish article on post-war justice and collaboration (bottom). Bar length indicates the proportion of selections (0–1). Dark bars represent human selections, light bars GPT-4o selections, and hatched bars denote candidates outside the top-five consensus.

Human Annotation Given the low overlap between model outputs, we conducted a human evaluation. Eleven annotators selected up to five keyphrases per article, aiming for broad coverage while avoiding synonyms.

ChatGPT Annotation Although human annotation provides a qualitative reference, it does not scale to larger datasets. We therefore evaluate LLM-as-judge as an approximation of human preference. Using GPT-4o, we replicate the human selection task by asking the model to choose the best-fitting keyphrases among the generated candidates. The procedure is repeated five times with different random seeds to assess output stability.

Consensus Reference Since each example receives multiple human and LLM annotations, we construct a consensus reference that maximizes agreement. Figure 2 illustrates this process for two representative articles, showing how human and

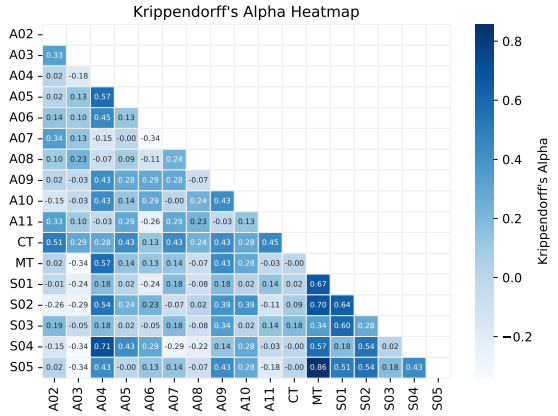

Figure 3: Inter-annotator agreement for French keyphrases from two articles. Human annotators are labeled A*dd*, GPT-4o preference runs with different random seeds are labeled S*dd*, the human consensus reference is CT, and the GPT-4o consensus reference is MT. Darker cells indicate higher agreement.

GPT-4o selection frequencies reveal both shared preferences and divergent judgments across models. Despite variation in wording and level of abstraction, both examples exhibit a coherent thematic core (e.g., constitutional reform in the French case and post-war justice and collaboration in the Luxembourgish case), supporting the interpretation that differences reflect emphasis rather than topic mismatch. For each article, we manually select five diverse keyphrases from the human annotations and five from the GPT-4o runs to retain representative and complementary choices.

Results Figure 3 shows agreement patterns for the French sample. Agreement between annotators and models is moderate, reflecting variation in wording and topical emphasis rather than obvious errors. The heatmap nevertheless reveals non-random structure: some annotators and model-based references cluster more closely than others, and the repeated GPT-4o selection runs appear more consistent with each other than with the human consensus. This pattern suggests that LLM-based judgments capture a coherent preference signal, but not simply the same one as human annotators.

Annotators selected on average 42.7% of model keyphrases for German articles, compared with 28.1% for French and 28.4% for Luxembourgish. Across languages, DeepSeek-V3 receives the highest overall selection rates, indicating closer alignment with human judgments. Annotators generally favor multi-word expressions, and agreement remains limited even among humans, underscoring the subjective nature of keyphrase evaluation. These findings highlight the limitations of strict lex-

German	French	Luxembourgish
	20th Century	
workers' rights	international relations	satire
labor movement	cultural events	humor
economic policy	labor movement	tradition
labor unions	workers' rights	cultural heritage
labor rights	economic crisis	patriotism
	21st Century	
stock market	television programming	linguistics
television programming	cultural events	cultural identity
investment	stock market	theater
financial data	documentary films	grammar
market trends	football	phonetics

Table 3: Top five generated keyphrases in 20th- and 21st-century newspapers by language.

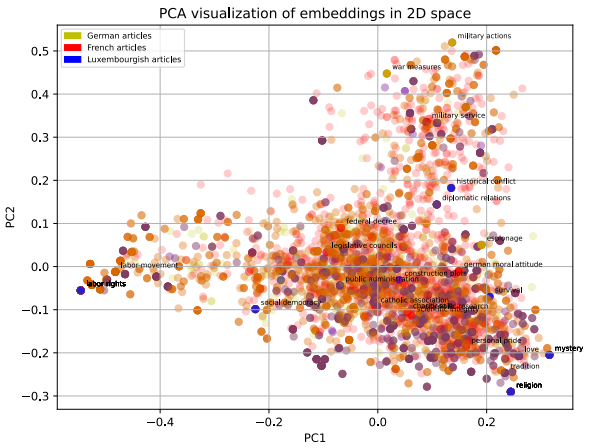

Figure 4: PCA projection of article and keyphrase embeddings for newspapers from the 1910s. Points represent semantic embeddings; labeled items indicate example keyphrases. Clusters correspond to thematic groupings such as war, legislation, and religion.

ical overlap measures and motivate embedding-based similarity approaches. Given the exploratory nature and limited size of the pilot, we do not treat the comparison as a definitive model ranking. We selected DeepSeek-V3 for the medium-scale application because, in this pilot, it aligned best overall with annotator preferences and was substantially cheaper for larger-scale processing.

4. Medium-Scale Application

We apply DeepSeek-V3 to 3,870 German, 7,272 French, and 512 Luxembourgish newspaper articles (18th–21st centuries) to analyze topical and diachronic patterns. Articles were sampled based on length (5,000–25,000 characters), temporal balance (maximum three articles per year and newspaper), and OCR quality, estimated as the proportion of word types recognized by dictionaries.

Cross-Linguistic Thematic Profiles Table 3 shows clear cross-linguistic thematic differences. German newspapers are dominated by labor and economic terminology, French newspapers combine international and cultural topics, and Luxembourgish newspapers emphasize cultural and identity-related themes. These contrasts suggest that generated keyphrases capture language-specific discourse profiles rather than reflecting uniform model bias.

To examine whether generated keyphrases reflect historically meaningful structure, we analyze their distribution across periods. Table 4 reveals systematic diachronic variation across periods. Early periods are dominated by legal, administrative, and diplomatic terminology, reflecting the informational function of early newspapers. The late nineteenth and early twentieth centuries show increased prominence of public health, labor, and economic topics, followed by interwar emphasis on crisis, employment, and war-related themes. Postwar decades introduce media, institutional, and cultural vocabulary, indicating diversification of newspaper content. These shifts broadly align with known historical developments, suggesting that generated keyphrases capture meaningful semantic signals rather than surface lexical artifacts.

Embedding texts and keyphrases To assess whether generated keyphrases support large-scale semantic exploration, we project article and keyphrase embeddings into a shared space using a multilingual text embedder. Figure 4 shows a two-dimensional projection of embeddings for articles and generated keyphrases from the 1910s. The projection reveals coherent thematic structure: items related to military activity cluster together, while administrative and religious topics occupy distinct regions. This indicates that the generated keyphrases capture meaningful semantic distinctions and can support exploratory analysis of historical collections.

Interpretive Implications Taken together, the frequency patterns, period distributions, and embedding projections suggest that generated keyphrases capture not only topical content but also structural properties of historical discourse. The consistency of thematic clustering across languages and periods indicates that the model is sensitive to underlying semantic regularities rather than merely reproducing superficial lexical associations.

Methodological Perspective These findings support the use of LLM-based keyphrase generation as a lightweight semantic enrichment layer for large historical corpora. Instead of replacing traditional indexing or annotation, such automatically generated abstractions can complement existing metadata by enabling cross-lingual comparison, thematic exploration, and corpus-level analysis with minimal manual effort.

1789–1848	1849–1875	1876–1914	1918–1939	1945–1989
legal proceedings	diplomatic relations	international relations	international relations	cultural events
public notices	international relations	public health	economic crisis	international relations
political unrest	public opinion	public opinion	labor movement	sports competition
legal notices	federal council	federal council	workers' rights	radio programming
property auction	public administration	diplomatic relations	public administration	social justice
diplomatic relations	political conflict	theater	unemployment	labor rights
commerce	government policy	public safety	economic policy	television programming
public auctions	military conflict	workers' rights	public works	professional training
public administration	political unrest	legal proceedings	World War I	democracy
real estate	railway construction	crime	diplomatic relations	collective bargaining

Table 4: Top ten keyphrases per historical period (all languages combined). The distribution shows a shift from legal–administrative topics in early periods to industrial and labor themes in the early twentieth century, and later to media, cultural, and institutional domains.

Comparative Patterns and Diachronic Trends Across languages, three broad trends emerge. First, labor- and union-related terminology prominent in the twentieth century declines in the twenty-first century. Second, German and French newspapers show increased salience of media, sports, and financial topics, indicating diversification and financialization of discourse. Third, Luxembourgish shifts toward metalinguistic and policy-oriented themes, suggesting growing linguistic reflexivity.

Overall, keyphrase distributions point to a transition from labor-centered and structurally political discourse to more diversified and media-oriented thematic landscapes, while cross-linguistic distinctions remain clearly visible. Detailed frequency lists are omitted due to space constraints.

Limitations Several limitations should be noted. First, the corpus is opportunistic rather than systematically curated: article selection depends on digitization availability within the *Impresso* infrastructure and is therefore not culturally or historiographically representative. Observed thematic distributions may thus reflect collection bias as much as historical reality. Second, keyphrase quality was evaluated through human preference and agreement rather than against a gold standard, limiting comparability with benchmark-style evaluations. Third, diachronic language variation remains an important consideration for this task. Although historical forms and discourse conventions do not always align neatly with present-day English conceptual labels, our examples suggest that cross-lingual normalization still captures broader thematic structure well enough to support exploratory analysis. Accordingly, the findings should be interpreted as exploratory rather than definitive.

5. Conclusion and Future Work

We presented an exploratory study of cross-lingual abstractive keyphrase generation for historical newspapers using instruction-following LLMs. A human-centered pilot evaluation shows moderate agreement but consistent preference patterns, with DeepSeek-V3 aligning most closely with human judgments. A medium-scale application demonstrates that generated keyphrases capture coherent cross-linguistic and diachronic structure and support semantic clustering of historical content.

Overall, the findings suggest that LLM-based keyphrase generation can provide a practical abstraction layer for multilingual historical collections, enabling indexing, comparison, and exploratory analysis beyond lexical search. At the same time, corpus bias, evaluation subjectivity, and cross-lingual normalization effects limit generalizability and call for further systematic investigation.

Future work will examine robustness across more balanced and culturally diverse corpora, analyze cost–performance trade-offs in greater detail, and investigate embedding-based normalization and clustering techniques to improve vocabulary consistency. Large-scale deployment within digital humanities infrastructures remains a promising next step.

Acknowledgments

This work has been supported by the Swiss National Science Foundation (grant No. CRSII5_213585) and by the Luxembourg National Research Fund (No. 17498891).

6. Bibliographical References

Adrian Bingham. 2010. The digitization of newspaper archives: Opportunities and challenges for historians. *Twentieth Century British History*, 21(2):225–231.

Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, editors. 2023. *Digitized Newspapers*

- *A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology*. Studies in Digital History and Hermeneutics. De Gruyter Oldenbourg, Berlin, Germany.

Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*, JCDL '17, pages 249–252, Piscataway, NJ, USA. IEEE, IEEE Press.

Marten Düring, Estelle Bunout, and Daniele Guido. 2024. Transparent generosity. Introducing the impresso interface for the exploration of semantically enriched historical newspapers. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, pages 35–55.

Marten Düring, Roman Kalyakin, Estelle Bunout, and Daniele Guido. 2021. Impresso inspect and compare. visual comparison of semantically enriched historical newspaper articles. *Information*, 12(9):348.

Maud Ehrmann, Marten Düring, Clemens Neudecker, and Antoine Doucet. 2023a. Computational Approaches to Digitised Historical Newspapers (Dagstuhl Seminar 22292). *Dagstuhl Reports*, 12(7):112–179.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023b. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):27:1–27:47.

Claire-Lise Gaillard. 2022. Feuilleter La Presse Ancienne Par Giga Octets. In Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, editors, *Digitised Newspapers – A New Eldorado for Historians? Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitization*. De Gruyter, Berlin, Germany.

Yifan Gao, Qingyu Yin, Zheng Li, Rui Meng, Tong Zhao, Bing Yin, Irwin King, and Michael Lyu. 2022. Retrieval-augmented multilingual keyphrase generation with retriever-generator iterative training. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1233–1246, Seattle, United States. Association for Computational Linguistics.

Byungha Kang and Youhyun Shin. 2025. Empirical study of zero-shot keyphrase extraction with large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3670–3686, Abu Dhabi, UAE. Association for Computational Linguistics.

Barbara McGillivray and Gábor Mihály Tóth. 2020. *Applying language technology in humanities research: Design, application, and the underlying logic*. Springer Nature.

Jayanth Mohan, Jishnu Ray Chowdhury, Tomas Malik, and Cornelia Caragea. 2025. Zero-shot keyphrase generation: Investigating specialized instructions and multi-sample aggregation on large language models.

Clemens Neudecker. 2022. Cultural Heritage as Data: Digital Curation and Artificial Intelligence in Libraries. In *Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022)*, volume 3234 of *CEUR Workshop Proceedings*, Berlin, Germany. CEUR.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Michael Piotrowski. 2012. *Natural language processing for historical texts*. Morgan & Claypool Publishers.

Mia Ridge, Giovanni Colavizza, Laurel Brake, Maud Ehrmann, Jean-Phillipe Moreux, and Andrew Prescott. 2019. The Past, Present and Future of Digital Scholarship with Newspaper Collections. In *DH 2019 Book of Abstracts*, page 9, Utrecht.

Olivier Salaün, Frédéric Piedboeuf, Guillaume Le Berre, David Alfonso-Hermelo, and Philippe Langlais. 2024. EUROPA: A legal multilingual keyphrase generation dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12718–12736, Bangkok, Thailand. Association for Computational Linguistics.

Daniel van Strien, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pages 484–496, Valletta, Malta. SCITEPRESS - Science and Technology Publications.

Mitchell Whitelaw. 2015. Generous Interfaces for Digital Cultural Collections. *Digital Humanities Quarterly*, 9(1).