



*This list includes all the publications which were created in the context of the NewsEye project.*

# Book Chapters

1. Marjanen, J. (2021). [National Sentiment: Nation Building and Emotional Language in Nineteenth-Century Finland](#). In V. Kivimäki, S. Suodenjoki, & T. Vahtikari (Eds.), *Lived Nation as the History of Experiences and Emotions in Finland, 1800-2000* (pp. 61–83). Springer International Publishing. [https://doi.org/10.1007/978-3-030-69882-9\\_3](https://doi.org/10.1007/978-3-030-69882-9_3)
2. Oberbichler S. & Pfanzer E., [Tracing Discourses in Digital Newspaper Collections: A Contribution to Digital Hermeneutics while Investigating 'Return Migration' in Historical Press Coverage](#), in: *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 Oldenbourg, 2022.
3. Gasteiner M. & Enderlin A., [Crossing or Intersecting the Emperor's Desk with digitized Newspaper Data: Entity-source-networks in the late Habsburg Empire](#), in: *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 Oldenbourg, 2022.

# Journal Articles

1. Hengchen, S., Ros, R., Marjanen, J., & Tolonen, M. (2021). A data-driven approach to studying changing vocabularies in historical newspaper collections. *Digital Scholarship in the Humanities*, 36(Supplement\_2), ii109–ii126. <https://doi.org/10.1093/llc/fqab032>
2. Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Boros, E., Hamdi,

- A., Doucet, A., Sidere, N., & Coustaty, M. (2021). MELHISSA: A multilingual entity linking architecture for historical press articles. *International Journal on Digital Libraries*. <https://doi.org/10.1007/s00799-021-00319-6>
3. Linhares Pontes, E., Huet, S., Torres Moreno, J. M., Silva, T. G. da, & Linhares, A. C. (2020). A Multilingual Study of Multi-Sentence Compression using Word Vertex-Labeled Graphs and Integer Linear Programming. *Computación y Sistemas*, 24(2). <https://doi.org/10.13053/cys-24-2-3335>
  4. Marjanen, J., Kurunmäki, J., Pivovarov, L., & Zosa, E. (2020). The expansion of isms, 1820-1917: Data-driven analysis of political language in digitized newspaper collections. *Journal of Data Mining & Digital Humanities, HistInformatics*, 6159. <https://doi.org/10.46298/jdmdh.6159>
  5. Marjanen, J., Vaara, V., Kanner, A., Roivainen, H., Mäkelä, E., Lahti, L., & Tolonen, M. (2019). A National Public Sphere? Analyzing the Language, Location, and Form of Newspapers in Finland, 1771–1917. *Journal of European Periodical Studies*, 4(1), 54–77. <https://doi.org/10.21825/jeps.v4i1.10483>
  6. Nguyen, T. T. H., Jatowt, A., Coustaty, M., & Doucet, A. (2021). Survey of Post-OCR Processing Approaches. *ACM Computing Surveys*, 54(6), 1–37. <https://doi.org/10.1145/3453476>
  7. Oberbichler, S., Boroş, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H., & Tolonen, M. (2022). Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. *Journal of the Association for Information Science and Technology*, 73(2), 225–239. <https://doi.org/10.1002/asi.24565>
  8. Oberbichler, S., Hechl, S., & Pfanzelter, E. (2020). Als eine andere Epidemie die Welt in Atem hielt: Die Spanische Grippe 1918/19 in der österreichischen Presse. *Tiroler Chronist - Fachblatt von Und Für Chronisten in Nord-, Süd- Und Osttirol*, 154, 15–22.
  9. Oberbichler, S., & Pfanzelter, E. (n.d.). Topic-specific corpus building: A step towards a representative newspaper corpus on the topic of return migration using text mining methods. *Journal of Digital History Journal*, jd001. <https://journalofdigitalhistory.org/en/article/4yxHGiqXYRbX>
  10. Pfanzelter, E., Oberbichler, S., Marjanen, J., Langlais, P.-C., & Hechl, S. (2021). Digital interfaces of historical newspapers: Opportunities, restrictions and recommendations. *Journal of Data Mining & Digital Humanities, HistInformatics*(HistInformatics), 6121. <https://doi.org/10.46298/jdmdh.6121>

---

## Conference Papers

1. Alhalaseh, R., Munezero, M., Leinonen, M., Leppänen, L., Avikainen, J., & Toivonen, H. (2018). Towards Data-Driven Generation of Visualizations for Automatically Generated News Articles. *Proceedings of the 22nd International Academic Mindtrek Conference*, 100–109. <https://doi.org/10.1145/3275116.3275131>
2. Bernard, Guillaume, Suire, Cyrille, Faucher, Cyril, & Doucet, Antoine. (2022a). A Comprehensive Extraction of Relevant Real-World-Event

- Qualifiers for Semantic Search Engines. *Linking Theory and Practice of Digital Libraries*, 12866, 153–164. <https://doi.org/10.5281/ZENODO.5900757>
3. Bernard, Guillaume, Suire, Cyrille, Faucher, Cyril, & Doucet, Antoine. (2022b). Event Related Document Retrieval with Multilingual Real World Event Representation. *Proceedings of the 20th International Semantic Web Conference*, 2980. <https://doi.org/10.5281/ZENODO.5900742>
  4. Boros, E., Hamdi, A., Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Sidere, N., & Doucet, A. (2020). Alleviating Digitization Errors in Named Entity Recognition for Historical Documents. *Proceedings of the 24th Conference on Computational Natural Language Learning*, 431–441. <https://doi.org/10.18653/v1/2020.conll-1.35>
  5. Boros, E., Hamdi, A., Pontes, E. L., Cabrera-Diego, L. A., Moreno, J. G., Sidère, N., & Doucet, A. (2021, April 15). *Atténuer les erreurs de numérisation dans la reconnaissance d'entités nommées pour les documents historiques*. Conférence en Recherche d'Informations et Applications - CORIA 2021, French Information Retrieval Conference, Grenoble, France. [https://doi.org/10.24348/CORIA.2021.MINI\\_24](https://doi.org/10.24348/CORIA.2021.MINI_24)
  6. Boros, E., Moreno, J. G., & Doucet, A. (2021a, December 14). Exploring Entities in Event Detection as Question Answering. *Proceedings of the 44th European Conference on Information Retrieval*. 44th European Conference on Information Retrieval, Stavanger, Norway.
  7. Boros, E., Moreno, J. G., & Doucet, A. (2021b). Event Detection with Entity Markers. In D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, & F. Sebastiani (Eds.), *Proceedings of the 43rd European Conference on Information Retrieval (ECIR) (ECIR 2021)* (Vol. 12657, pp. 233–240). Springer International Publishing. [https://doi.org/10.1007/978-3-030-72240-1\\_20](https://doi.org/10.1007/978-3-030-72240-1_20)
  8. Boros, Emanuela, & Doucet, Antoine. (2021, January 1). *Transformer-based Methods for Recognizing Ultra Fine-grained Entities (RUFES)*. TAC 2020 Workshop, Online. <https://doi.org/10.5281/ZENODO.4555778>
  9. Boros, Emanuela, Linhares Pontes, Elvys, Cabrera-Diego, Luis Adrián, Hamdi, Ahmed, Moreno, Jose G., Sidère, Nicolas, & Doucet, Antoine. (2020). Robust Named Entity Recognition and Linking on Historical Multilingual Documents. *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, 2696, 1–17. <https://doi.org/10.5281/ZENODO.4059652>
  10. Cabrera-Diego, Luis Adrián, Moreno, Jose G., & Doucet, Antoine. (2021a, April 12). Simple ways to improve NER in every language using markup. *Proceedings of the 2nd International Workshop on Cross-Lingual Event-Centric Open Analytics Co-Located with the 30th The Web Conference (WWW 2021)*. 2nd International Workshop on Cross-lingual Event-centric Open Analytics co-located with the 30th The Web Conference (WWW 2021), Ljubljana, Slovenia. <https://doi.org/10.5281/ZENODO.4680998>
  11. Cabrera-Diego, Luis Adrián, Moreno, Jose G., & Doucet, Antoine. (2021b). Using a Frustratingly Easy Domain and Tagset Adaptation for Creating Slavic Named Entity Recognition Systems. *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, 98–104. <https://doi.org/10.5281/ZENODO.4730478>

12. Doucet, Antoine, Gasteiner, Martin, Granroth-Wilding, Mark, Kaiser, Max, Kaukonen, Minna, Labahn, Roger, Moreux, Jean-Philippe, Muehlberger, Guenter, Pfanzelter, Eva, Therenty, Marie-Eve, Toivonen, Hannu, & Tolonen, Mikko. (2020, July 23). *NewsEye: A digital investigator for historical newspapers*. 15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Ottawa, Canada. <https://doi.org/10.5281/ZENODO.3895269>
13. Duong, Quan, Hämäläinen, Mika, & Hengchen, Simon. (2020). An Unsupervised method for OCR Post-Correction and Spelling Normalisation for Finnish. *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 240–248. <https://doi.org/10.5281/ZENODO.4242890>
14. Duong, Quan, Pivovarov, Lidia, & Zosa, Elaine. (2021). Benchmarks for Unsupervised Discourse Change Detection. *Proceedings of the 6th International Workshop on Computational History (HistoInformatics 2021)*, 2981. <https://doi.org/10.5281/ZENODO.5548067>
15. Frossard, E., Coustaty, M., Doucet, A., Jatowt, A., & Hengchen, S. (2020). Dataset for Temporal Analysis of English-French Cognates. *Proceedings of the 12th Language Resources and Evaluation Conference*, 855–859. <https://doi.org/10.5281/ZENODO.3693651>
16. Gutehrlé, Nicolas, Harlamov, Oleg, Karimi, Farimah, Wei, Haoyu, Jean-Caurant, Axel, & Pivovarov, Lidia. (2021, September 30). SpaceWars: A Web Interface for Exploring the Spatio-temporal Dimensions of WWI Newspaper Reporting. *Proceedings of the 6th International Workshop on Computational History (HistoInformatics 2021)*. 6th International Workshop on Computational History, Aachen, Germany. <https://doi.org/10.5281/ZENODO.5566462>
17. Hamdi, A., Carel, E., Joseph, A., Coustaty, M., & Doucet, A. (2021). Information Extraction from Invoices. In J. Lladós, D. Lopresti, & S. Uchida (Eds.), *Proceedings of the 16th International Conference on Document Analysis and Recognition (ICDAR 2021)* (Vol. 12822, pp. 699–714). Springer International Publishing. [https://doi.org/10.1007/978-3-030-86331-9\\_45](https://doi.org/10.1007/978-3-030-86331-9_45)
18. Hamdi, A., Jean-Caurant, A., Sidere, N., Coustaty, M., & Doucet, A. (2019). An Analysis of the Performance of Named Entity Recognition over OCRed Documents. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 333–334. <https://doi.org/10.1109/JCDL.2019.00057>
19. Hamdi, A., Jean-Caurant, A., Sidère, N., Coustaty, M., & Doucet, A. (2020). Assessing and Minimizing the Impact of OCR Quality on Named Entity Recognition. In M. Hall, T. Merčun, T. Risse, & F. Duchateau (Eds.), *Proceedings of the 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020* (Vol. 12246, pp. 87–101). Springer International Publishing. [https://doi.org/10.1007/978-3-030-54956-5\\_7](https://doi.org/10.1007/978-3-030-54956-5_7)
20. Hamdi, A., Linhares Pontes, E., Boros, E., Nguyen, T. T. H., Hackl, G., Moreno, J. G., & Doucet, A. (2021). A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2328–2334. <https://doi.org/10.1145/3404835.3463255>

21. Huynh, V.-N., Hamdi, A., & Doucet, A. (2020). When to Use OCR Post-correction for Named Entity Recognition? In E. Ishita, N. L. S. Pang, & L. Zhou (Eds.), *Proceedings of the 14. International Conference on Data Analytics in Logistics (ICDAL 2020)* (Vol. 12504, pp. 33–42). Springer International Publishing. [https://doi.org/10.1007/978-3-030-64452-9\\_3](https://doi.org/10.1007/978-3-030-64452-9_3)
22. Kettunen, K., & La Mela, M. (2020). Digging Deeper into the Finnish Parliamentary Protocols – Using a Lexical Semantic Tagger for Studying Meaning Change of Everyman’s Rights (allemensrätten). *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, 63–80. <https://doi.org/10.5281/ZENODO.3676371>
23. Klaus, Barbara. (2020). Can Umlauts Ruin Your Research in Digitized Newspaper Collections? A NewsEye Case Study on “The Dark Sides of War” (1914–1918). *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, 2612, 267–274. <https://doi.org/10.5281/ZENODO.4686731>
24. Kutuzov, A., & Pivovarova, L. (2021). Three-part diachronic semantic change dataset for Russian. *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, 7–13. <https://doi.org/10.18653/v1/2021.lchange-1.2>
25. Leppänen, Leo, & Toivonen, Hannu. (2021). A Baseline Document Planning Method for Automated Journalism. *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 101–111. <https://doi.org/10.5281/ZENODO.5562428>
26. Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Boros, E., Hamdi, A., Sidère, N., Coustaty, M., & Doucet, A. (2020). Entity Linking for Historical Documents: Challenges and Solutions. In E. Ishita, N. L. S. Pang, & L. Zhou (Eds.), *Proceedings of the 22nd International Conference on Asia-Pacific Digital Libraries (ICADL 2020)* (Vol. 12504, pp. 215–231). Springer International Publishing. [https://doi.org/10.1007/978-3-030-64452-9\\_19](https://doi.org/10.1007/978-3-030-64452-9_19)
27. Linhares Pontes, E., Hamdi, A., Sidere, N., & Doucet, A. (2019). Impact of OCR Quality on Named Entity Linking. In A. Jatowt, A. Maeda, & S. Y. Syn (Eds.), *Digital Libraries at the Crossroads of Digital Information for the Future* (Vol. 11853, pp. 102–115). Springer International Publishing. [https://doi.org/10.1007/978-3-030-34058-2\\_11](https://doi.org/10.1007/978-3-030-34058-2_11)
28. Linhares Pontes, E., Moreno, J. G., & Doucet, A. (2020). Linking Named Entities across Languages using Multilingual Word Embeddings. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 329–332. <https://doi.org/10.1145/3383583.3398597>
29. Marjanen, Jani, Pivovarova, Lidia, Zosa, Elaine, & Kurunmäki, Jussi. (2019). Clustering Ideological Terms in Historical Newspaper Data with Diachronic Word Embeddings. *HistoInformatics 2019: International Workshop on Computational History 2019: The 5th International Workshop on Computational History (HistoInformatics 2019): Co-Located with the 23rd International Conference on Theory and Practice of Digital Libraries (TPDL 2019)*, 21–29. <https://doi.org/10.5281/ZENODO.3689466>
30. Marjanen, Jani, Zosa, Elaine, Hengchen, Simon, Pivovarova, Lidia, & Tolonen, Mikko. (2021). Topic modelling discourse dynamics in historical newspapers. *Proceedings of the 5th Conference Digital Humanities in the*

*Nordic Countries* (DHN 2020), 63–77. <https://doi.org/10.5281/ZENODO.5648114>

31. Martinc, M., Montariol, S., Zosa, E., & Pivovarov, L. (2020). Capturing Evolution in Word Usage: Just Add More Clusters? *Companion Proceedings of the Web Conference 2020*, 343–349. <https://doi.org/10.1145/3366424.3382186>
32. Michael, J., Labahn, R., Gruning, T., & Zollner, J. (2019). Evaluating Sequence-to-Sequence Models for Handwritten Text Recognition. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1286–1293. <https://doi.org/10.1109/ICDAR.2019.00208>
33. Michael, J., Weidemann, M., Laasch, B., & Labahn, R. (2021). ICPR 2020 Competition on Text Block Segmentation on a NewsEye Dataset. In A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, & R. Vezzani (Eds.), *Pattern Recognition. ICPR International Workshops and Challenges* (Vol. 12668, pp. 405–418). Springer International Publishing. [https://doi.org/10.1007/978-3-030-68793-9\\_30](https://doi.org/10.1007/978-3-030-68793-9_30)
34. Montariol, S., Martinc, M., & Pivovarov, L. (2021). Scalable and Interpretable Semantic Change Detection. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4642–4652. <https://doi.org/10.18653/v1/2021.naacl-main.369>
35. Mutuvi, S., Boros, E., Doucet, A., Jatowt, A., Lejeune, G., & Odeo, M. (2020). Multilingual Epidemiological Text Classification: A Comparative Study. *Proceedings of the 28th International Conference on Computational Linguistics*, 6172–6183. <https://doi.org/10.18653/v1/2020.coling-main.543>
36. Mutuvi, S., Boros, E., Doucet, A., Lejeune, G., Jatowt, A., & Odeo, M. (2021a, April 15). *Étude comparative de méthodes de classification multilingue appliquées à l'épidémiologie*. COnférence en Recherche d'Informations et Applications - CORIA 2021, French Information Retrieval Conference, Grenoble, France. <https://doi.org/10.5281/ZENODO.4734471>
37. Mutuvi, S., Boros, E., Doucet, A., Lejeune, G., Jatowt, A., & Odeo, M. (2021b). Multilingual Epidemic Event Extraction. In H.-R. Ke, C. S. Lee, & K. Sugiyama (Eds.), *Proceedings of the 23rd International Conference on Asian Digital Libraries* (Vol. 13133, pp. 139–156). Springer International Publishing. [https://doi.org/10.1007/978-3-030-91669-5\\_12](https://doi.org/10.1007/978-3-030-91669-5_12)
38. Mutuvi, S., Boros, E., Doucet, A., Lejeune, G., Jatowt, A., & Odeo, M. (2021c). Token-Level Multilingual Epidemic Dataset for Event Extraction. In G. Berget, M. M. Hall, D. Brenn, & S. Kumpulainen (Eds.), *Proceedings of the 25th International Conference on Theory and Practice of Digital Libraries* (Vol. 12866, pp. 55–59). Springer International Publishing. [https://doi.org/10.1007/978-3-030-86324-1\\_6](https://doi.org/10.1007/978-3-030-86324-1_6)
39. Mutuvi, S., Doucet, A., Lejeune, G., & Odeo, M. (2020). A Dataset for Multilingual Epidemiological Event Extraction. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 4139–4144. <https://doi.org/10.5281/ZENODO.3693647>
40. Mutuvi, S., Doucet, A., Odeo, M., & Jatowt, A. (2018). Evaluating the Impact of OCR Errors on Topic Modeling. In M. Dobрева, A. Hinze, & M. Žumer (Eds.), *Proceedings of the 20th International Conference on Asia-Pacific Digital Libraries* (Vol. 11279, pp. 3–14). Springer International Publishing.

41. Nguyen, N. K., Boroş, E., Lejeune, G., & Doucet, A. (2020). Impact Analysis of Document Digitization on Event Extraction. *Proceedings of the 4th Workshop on Natural Language for Artificial Intelligence (NL4AI 2020)*, 2735, 17–28. <https://doi.org/10.5281/ZENODO.4734268>
42. Nguyen, N. K., Boros, E., Lejeune, G., Doucet, A., & Delahaut, T. (2021). L3i\_LBPAM at the FinSim-2 task: Learning Financial Semantic Similarities with Siamese Transformers. *Companion Proceedings of the Web Conference 2021*, 302–306. <https://doi.org/10.1145/3442442.3451384>
43. Nguyen, T. T. H., Jatowt, A., Nguyen, N.-V., Coustaty, M., & Doucet, A. (2020). Neural Machine Translation with BERT for Post-OCR Error Detection and Correction. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 333–336. <https://doi.org/10.1145/3383583.3398605>
44. Nguyen, T.-T.-H., Coustaty, M., Doucet, A., Jatowt, A., & Nguyen, N.-V. (2018). Adaptive Edit-Distance and Regression Approach for Post-OCR Text Correction. In M. Dobрева, A. Hinze, & M. Žumer (Eds.), *Proceedings of the 20th International Conference on Asia-Pacific Digital Libraries* (Vol. 11279, pp. 278–289). Springer International Publishing. [https://doi.org/10.1007/978-3-030-04257-8\\_29](https://doi.org/10.1007/978-3-030-04257-8_29)
45. Nguyen, T.-T.-H., Jatowt, A., Coustaty, M., Nguyen, N.-V., & Doucet, A. (2019a). Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 29–38. <https://doi.org/10.1109/JCDL.2019.00015>
46. Nguyen, T.-T.-H., Jatowt, A., Coustaty, M., Nguyen, N.-V., & Doucet, A. (2019b). Post-OCR Error Detection by Generating Plausible Candidates. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 876–881. <https://doi.org/10.1109/ICDAR.2019.00145>
47. Piskorski, Jakub, Babych, Bogdan, Kancheva, Zara, Kanishcheva, Olga, Lebedeva, Maria, Marcinczuk, Michał, Nakov, Preslav, Osenova, Petya, Pivovarova, Lidia, Pollak, Senja, Pribá, Pavel, Radev, Ivaylo, Robnik-Šikonja, Marko, Starko, Vasyl, Steinberger, Josef, & Yangarber, Roman. (2021). Slav-NER: The 3rd Cross-lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic languages. *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, 122–133. <https://doi.org/10.5281/ZENODO.4635585>
48. Pivovarova, L., Jean-Caurant, A., Avikainen, J., Alnajjar, K., Granroth-Wilding, M., Leppänen, L., Zosa, E., & Toivonen, H. (2020). Personal Research Assistant for Online Exploration of Historical News. In J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, & F. Martins (Eds.), *Proceedings of the 42nd European Conference on IR Research* (Vol. 12036, pp. 481–485). Springer International Publishing. [https://doi.org/10.1007/978-3-030-45442-5\\_62](https://doi.org/10.1007/978-3-030-45442-5_62)
49. Pivovarova, Lidia, & Zosa, Elaine. (2022, January 25). Visual Topic Modelling for NewsImage Task at MediaEval 2021. *Working Notes Proceedings of the MediaEval 2021 Workshop*. MediaEval 2021 Workshop. <https://doi.org/10.5281/ZENODO.5900719>
50. Rautiainen, J. (2019, July 15). *Opening Digitized Newspapers for Different User Groups—Successes and Challenges*. IFLA World Library and

- Information Congress 2019, Athens, Greece. <https://doi.org/10.5281/ZENODO.3403158>
51. Rigaud, C., Doucet, A., Coustaty, M., & Moreux, J.-P. (2019). ICDAR 2019 Competition on Post-OCR Text Correction. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1588–1593. <https://doi.org/10.1109/ICDAR.2019.00255>
  52. Ros, R., & Oberbichler, S. (2020). The Helsinki Digital Humanities Hackathon: Two Perspectives on Multidisciplinary Historical Newspapers Research in a Hackathon Context. *Proceedings of the Twin Talks 2 and 3 Workshops at DHN 2020 and DH 2020*, 66–74. <https://doi.org/10.5281/ZENODO.3689228>
  53. Sumikawa, Y., Jatowt, A., Doucet, A., & Moreux, J.-P. (2019). Large Scale Analysis of Semantic and Temporal Aspects in Cultural Heritage Collection's Search. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 77–86. <https://doi.org/10.1109/JCDL.2019.00021>
  54. University of Helsinki, Finland, Pivovarova, L., Marjanen, J., & Zosa, E. (2019). Word Clustering for Historical Newspapers Analysis. *Proceedings of the Workshop on Language Technology for Digital Historical Archives - with a Special Focus on Central-, (South-)Eastern Europe, Middle East and North Africa*, 3–10. [https://doi.org/10.26615/978-954-452-059-5\\_002](https://doi.org/10.26615/978-954-452-059-5_002)
  55. Zosa, E., Granroth- Wilding, M., & Pivovarova, L. (2020). A Comparison of Unsupervised Methods for Ad hoc Cross-Lingual Document Retrieval. *Proceedings of the Workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, 32–37. <https://doi.org/10.5281/ZENODO.3751036>
  56. Zosa, E., & Granroth-Wilding, M. (2019). Multilingual Dynamic Topic Model. *Proceedings - Natural Language Processing in a Deep Learning World*, 1388–1396. [https://doi.org/10.26615/978-954-452-056-4\\_159](https://doi.org/10.26615/978-954-452-056-4_159)
  57. Zosa, Elaine, Hengchen, Simon, Marjanen, Jani, Pivovarova, Lidia, & Tolonen, Mikko. (2020, March 1). *Disappearing Discourses: Avoiding anachronisms and teleology with data-driven methods in studying digital newspaper collections*. DHN 2020. <https://doi.org/10.5281/ZENODO.3631613>
  58. Zosa, Elaine, Mutuvi, Stephen, Granroth-Wilding, Mark, & Doucet, Antoine. (2022, January 25). Evaluating the Robustness of Embedding-Based Topic Models to OCR Noise. *Proceedings of the 23rd International Conference on Asian Digital Libraries 2021*. 23rd International Conference on Asian Digital Libraries 2021, Online. <https://doi.org/10.5281/ZENODO.5900730>
  59. Zosa, Elaine, Shekhar, Ravi, Karan, Mladen, & Purver, Matthew. (2021). Not All Comments are Equal: Insights into Comment Moderation from a Topic-Aware Model. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 1652–1662. <https://doi.org/10.5281/ZENODO.5562465>
  60. Boros, Emanuela; Khoa Nguyen, Nhu; Lejeune, Gaël; Coustaty, Mickael; Doucet, Antoine (2021). Transformer-based Methods with #Entities for Detecting Emergency Events on Social Media. TREC Incident Streams 2021. <https://zenodo.org/record/6334513#.YjyMfIXMKU>
  61. Cabrera-Diego, Luis Adrián; Boros, Emanuela; Doucet, Antoine (2021). Elastic Embedded Background Linking for News Articles with Keywords,

- Entities and Events. TREC Incident Streams 2021. [zenodo.org/record/6334523](https://zenodo.org/record/6334523)
62. Zosa, Elaine, Pivovarova, Lidia, Boggia, Michele & Ivanova, Sardana (2022). Multilingual Topic Labelling of News Topics using Ontological Mapping. 44th European Conference on Information Retrieval (ECIR 2022). <https://zenodo.org/record/6334491#.YiX86Gdv7z8>
63. Emanuela Boros; Carlos-Emiliano Gonzalez-Gallardo; Jose G. Moreno; Antoine Doucet (2022). L3i at SemEval-2022 Task 11: Straightforward Additional Context for Multilingual Named Entity Recognition. <https://zenodo.org/record/6369947#.YjyN1IXMKUk>
64. Elaine Zosa; Emanuela Boros; Boshko Koloski; Lidia Pivovarova (2022). EMBEDDIA at SemEval-2022 Task 8: Investigating Sentence, Image, and Knowledge Graph Representations for Multilingual News Article Similarity. [zenodo.org/record/6369944](https://zenodo.org/record/6369944)
- 

## Datasets

1. Bernard, G., Doucet, A., Faucher, C., & Suire, C. (2021). *Event representation on Wikidata and Wikipedia with, and without the analysis of vernacular languages* (Version 1) [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.4733507>
2. Frossard, E., Coustaty, M., Doucet, A., Jatowt, A., & Hengchen, S. (2020). *Data for "Dataset for Temporal Analysis of English-French Cognates"* [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.3688086>
3. Hamdi, A., Elvys Linhares Pontes, Boros, E., Nguyen, T. T. H., Hackl, G., Moreno, J. G., & Doucet, A. (2021). *Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers* (V1.0) [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.4573312>
4. Hamdi, Ahmed, Jean-Caurant, Axel, Sidere, Nicolas, Coustaty, Mickaël, & Doucet, Antoine. (2020). *Benchmark for the evaluation of named entity recognition over ancient documents* [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.3877553>
5. Hengchen, Simon, Ros, Ruben, Marjanen, Jani, & Tolonen, Mikko. (2019). *Models for "A data-driven approach to studying changing vocabularies in historical newspaper collections"* (1.0.0) [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.3585027>
6. Linhares Pontes, E., Hamdi, A., Sidere, N., & Doucet, A. (2019). *Benchmark for the evaluation of Named Entity Linking over ancient documents* (0.1) [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.3490332>
7. Michael, Johannes, Weidemann, Max, Laasch, Bastian, & Labahn, Roger. (2020). *Dataset of ICPR 2020 Competition on Text Block Segmentation on a NewsEye Dataset* (1.0) [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.4943581>
8. Muehlberger, G., & Guenter Hackl. (2019). *NewsEye / READ OCR training dataset from Austrian Newspapers (19th C.)* [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.3387368>
9. Muehlberger, Guenter, & Hackl, Guenter. (2020). *NewsEye / READ OCR training dataset from French Newspapers (18th, 19th, early 20th C.)* [Data

- set]. Zenodo. <https://doi.org/10.5281/ZENODO.4293601>
10. Muehlberger, Guenter, & Hackl, Guenter. (2021a). *NewsEye / READ OCR training dataset from Finnish Newspapers (18th, 19th, early 20th C.)* (Version 1) [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.4599471>
  11. Muehlberger, Guenter, & Hackl, Guenter. (2021b). *NewsEye / READ OCR training dataset from Swedish Newspapers (18th, 19th, early 20th C.)* (Version 1) [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.4599623>
  12. Muehlberger, Guenter, & Hackl, Guenter. (2021c). *NewsEye / READ AS training dataset from Austrian Newspapers (19th, early 20th C.)* (Version 2) [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.4693412>
  13. Muehlberger, Guenter, & Hackl, Guenter. (2021d). *NewsEye / READ AS training dataset from French Newspapers (19th, early 20th C.)* (Version 3) [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.4600635>
  14. Muehlberger, Günter, & Hackl, Günter. (2021). *NewsEye / READ AS training dataset from Finnish Newspapers (19th C.)* (Version 2) [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.4600745>
  15. Mutuvi, S., Doucet, A., Lejeune, G., & Odeo, M. (2020). *Data for “A Dataset for Multi-lingual Epidemiological Event Extraction”* [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.3709616>
  16. Rigaud, C., Doucet, A., Coustaty, M., & Moreux, J.-P. (2019). *Dataset of ICDAR 2019 Competition on Post-OCR Text Correction* [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.3515402>
- 

## Theses

1. [Avikainen, J. \(2019\)](#). *A Method for Wavelet-Based Time Series Analysis of Historical Newspapers* [University of Helsinki]. <https://zenodo.org/record/3628263>
  2. Hechl, Stefan Patrick. (2021). *“Wir dürfen wieder Österreicher sein!” Die Rolle der Tagespresse in österreichischen Nation-Building-Prozessen 1945–1948 – eine quantitative Analyse ausgewählter digitaler Zeitungskorpora samt Vorschlägen zur didaktischen Umsetzung* [Universität Innsbruck]. <https://zenodo.org/record/4468295>
  3. Laasch, B. M. (2018). *Wortvektoren* [Universität Rostock]. [http://rosdok.uni-rostock.de/resolve/id/rosdok\\_thesis\\_0000000023](http://rosdok.uni-rostock.de/resolve/id/rosdok_thesis_0000000023)
- 

## Others

1. Hamdi, A., Elvys Linhares Pontes, & Doucet, A. (2021). *Annotation Guidelines for Named Entity Recognition, Entity Linking and Stance Detection* (v3.1). Zenodo. <https://zenodo.org/record/4574198>
2. Kanner, Antti, Mäkelä, Eetu, Marjanen, Jani, Tolonen, Mikko, Oberbichler, Sarah, Duong, Quan, Pivovarova, Lidia, Ali, Dilawar, Verstockt, Steven, Ollion, Étienne, Shen, Rubing, Arnold, Matthias, Brown, David, Adam, Raven, Balasubramanian, Saranya, Charvat, Vera Maria, Füllsack, Manfred, Kleinert, Jörn, Misera, Hanna, ... Lomazow, Steven. (2021). *The Book of*

*Abstracts for What's Past is Prologue: The NewsEye International Conference.* Zenodo. <https://doi.org/10.5281/ZENODO.5167375>

3. *L'intelligence artificielle à la BnF.* (2022). BnF. [http://chroniques.bnf.fr/pdf/Chroniques\\_93.pdf](http://chroniques.bnf.fr/pdf/Chroniques_93.pdf)
4. NewsEye Consortium, Hunter, Tonica, & Chambers, Sally. (2020). *NewsEye Policy Brief.* Zenodo. <https://zenodo.org/record/4291894>
5. Oberbichler, S. (2020). *Using LDA and Jensen-Shannon Distance (JSD) to group similar newspaper articles (v1.0).* Zenodo. <https://zenodo.org/record/3876063>
6. Oberbichler, S., Pfanzelter, E., Hechl, S., & Marjanen, J. (2020, October 16). Doing historical research with digital newspapers – perspectives of DH scholars. *Europeana*, 16. <https://pro.europeana.eu/page/issue-16-newspapers#doing-historical-research-with-digital-newspapers>
7. Omari, N., & Doucet, A. (2020, May 13). Covid-19 et grippe espagnole: Quand la presse du XXe siècle rappelle celle de 2020. *The Conversation*. <https://theconversation.com/covid-19-et-grippe-espagnole-quand-la-presse-du-xx-siecle-rappelle-celle-de-2020-137035>

8. Pivovarova, L., Zosa, E., & Marjanen, J. (2019). *Embeddings built on 19th century newspapers from [Twitter icon] and [Email icon] Zenodo.* <https://zenodo.org/record/3557480>



This project has  
received funding  
from the

European Union's  
Horizon 2020

research and innovation programme  
under grant agreement No 770299.

Copyright © [newseye.eu](https://newseye.eu) 2018  
Theme [Blackrock Digital](#) ([License](#))

[Imprint](#)  
[Cookie Policy](#)