Using AI to put together the wartime propaganda puzzle

Mari Wigham
Netherlands Institute for Sound and Vision
Hilversum, The Netherlands

Rana Klein
Netherlands Institute for Sound and Vision
Hilversum, The Netherlands

Marjet Brolsma
University of Amsterdam
Amsterdam, The Netherlands

Roeland Ordelman
University of Twente / Netherlands Institute for Sound and Vision
Enschede, The Netherlands

ABSTRACT

The study of wartime propaganda from the past is increasingly relevant in the light of current events [1]. This is underlined by the fact that national leaders embroiled in conflicts refer to incidents from previous wars in their propaganda [1]. Understanding propaganda in the past can therefore contribute to understanding the present day.

A large number of sources of World War II propaganda in the Netherlands has survived. Yet these sources are scattered over isolated silos, residing in fragmentary collections of different media types preserved by different institutions and published on different platforms. Wartime newspapers are found in the Delpher[1] platform from the Dutch Royal Library. The small number of recorded broadcasts from the iconic Dutch radio station Radio Oranje is available as a collection in the CLARIAH Media Suite[2], whereas broadcasts from other wartime broadcasters are included in the Sound & Vision collection in the Media Suite. The written transcripts from Radio Oranje, on the other hand, are stored in the NIOD[3] library [2]. This scattering makes it hard for propaganda researchers to find sources, let alone to see the relationships between them or discover patterns.

There are significant imbalances in the availability of different types of material. In [3] the authors argue that 'these imbalances can be explained by the historical context in which these sources were created as well as by archival policies after 1945'. For example, there are far more Nazified newspaper articles available than anti-Nazi [1], and newspaper sources are available in far larger numbers than radio recordings [3]. These imbalances present researchers with additional challenges, as it is hard to determine if an apparent trend is meaningful, or merely an artefact of the imbalance. The limited availability of radio recordings led the authors of [3] to argue for this imbalance to be partially addressed by the digitisation of archives of related documents such as radio transcripts and monitoring reports, and for their publication in a platform for researchers.

The Media War Matching project took up this call to arms. The NIOD recently digitised two subsets - news bulletins and broadcasts on economics and politics - of the BNO (Berichtendienst Nederlandsche Omroep) [4] collection, which contains transcripts of broadcasts by the BNO radio broadcaster, a radio news service which spread Nazi propaganda. According to [4] the BNO produced about 35% of all transcripts aired by Nazi-controlled Radio Hilversum. These transcripts are particularly precious to researchers as they can be used both as a substitute for the many radio broadcasts that have been lost, and as a proxy to better search and analyse surviving radio broadcasts. The Media War Matching project worked to publish this collection in the CLARIAH Media Suite so that it could be easily searched and browsed. In a broader goal, the project sought to combine the puzzle pieces of radio broadcasts, their transcripts and wartime newspapers from the same date to give researchers a richer and broader picture of wartime propaganda in the Netherlands. Finally, the project looked at how analysis of patterns over these sources could offer added value for propaganda researchers.

In this paper, we first discuss the publication of the BNO collection. We describe the effect of the variable quality of the paper documents on the OCR quality and hence on searchability of the collection. For optimum insight, the pages in the collection should be grouped by the broadcast date, and within that by the individual broadcast titles. However, the dates and titles are not present in the metadata. We discuss our attempts to use NER (named entity recognition) to extract dates and titles, and how the effects of the OCR quality and the variability of document formats made extraction of titles and dates via NER impossible. The similarity of broadcast titles (Radio 1, Radio 2) in combination with the poor OCR quality led us to abandon the attempt to group pages per broadcast within this project. To enable grouping by date despite the OCR quality, we developed an algorithm for date extraction. We present an evaluation of the quality of date extraction, and discuss how we ultimately incorporated both human and automatic date annotations in the published collection. As is discussed in [5], data and tool criticism are essential to good digital humanities scholarship, and so we discuss how we documented the complex provenance of the collection, including date information, so that researchers can grasp its effects on their use of the collection.

Secondly, we explain how we applied speech recognition (ASR) to the wartime radio broadcasts. We discuss the effect of the unique characteristics of the source material on the ASR quality. We explain how the resulting poor quality, in combination with the difficulties in dating the radio transcripts, posed serious challenges for linking radio broadcasts to radio transcripts, which could not be overcome within the scope of this project.

Thirdly, we discuss how the propaganda collections can now be analysed. Despite the lack of links between broadcasts and transcripts, we demonstrate how publishing these and other propaganda collections together in the CLARIAH Media Suite enables the comparison of search results over time across collections, including

[1] https://www.delpher.nl/
[2] https://mediasuite.clariah.nl/
[3] https://www.niod.nl/
[4] Nederlands Instituut voor Oorlogs-, Holocaust- en Genocidestudies, Amsterdam, archief 103 Nederlandse Omroep, inv.no, 302-460

compensation for imbalances in collection size. We also explain how we used the ASR and OCR results to analyse the textual content of the collections. We present the case of analysing words that frequently occur together with the word 'Europa' (Europe) and consider the possibilities and limitations of such analysis given the issues with ASR and OCR quality.

Finally, we conclude by summarising our experiences of the benefits and challenges of using AI to combine propaganda collections. We observe that, while the ideal goal of interlinked collections is yet to be attained, AI has nevertheless enabled researchers to search and analyse propaganda collections in ways that were hitherto impossible. We also identify the key areas for future research to better combine propaganda collections and enable research across them.

Acknowledgement. This work was enabled by the CLARIAH-PLUS project funded by NWO (Grant 184.034.023).

KEYWORDS

speech recognition, OCR, AI, NER, propaganda

REFERENCES

[1] Vincent Kuitenbrouwer and Huub Wijfjes. Media war. *Journal of History*, 135, 2022.
[2] Vincent Kuitenbrouwer. The traces of a media war: Archives of dutch broadcasts from london during the second world war. *Journal of Media History*, 25, 2022.
[3] Vincent Kuitenbrouwer and Marjet Brolsma. Audio on paper: The merits and pitfalls of the dutch digital media archive for studying transnational entanglements during the second world war. *Journal of European Television History Culture*, 12, 2023.
[4] Dick Verkijk. *Radio Hilversum 1940-1945. De omroep in de oorlog.* Uitgeverij de Arbeiderspers, 1974.
[5] Marijn Koolen, Jasmijn van Gorp, and Jaco van Ossenbruggen. Toward a model for digital tool criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities*, 34, 2018.