Journal of Data Mining and Digital Humanities

EPI sciences overlay journals

An Industrial West? A Mixed-Methods Analysis of Newspapers Discourses about Technology over One Hundred and Ten Years (1830-1940)

Emannuelle Denove[1], Elisa Michelet[1], Germans Savcisens[2], Elena Fernández Fernández[3]

[1]Department of Computer Science, EPFL, Rte Cantonale, 1015 Lausanne, Switzerland
[2]Network Science Institute, Northeastern University, 177 Huntington Ave. Boston, MA 02115, USA
[3]IT Central, University of Zurich, University of Zurich, Schönberggasse 2, 8001 Zürich, Switzerland.

Corresponding author: Elena Fernández Fernández, elena.fernandezfernandez@uzh.ch

Abstract

This article explores information behaviour during one hundred and ten years (1830-1940), using multilingual historic newspapers as a proxy (*Le Figaro, The New York Herald, El Imparcial, Neuer Hamburger Zeitung* and *La Stampa*), seeking to observe to what extent technology has historically acted as a cohesive force across Western societies. Three key technological terms (telephone, gasoline, and iron) are selected as an exploratory research endeavour. Afterwards, a mix-methods approach that combines quantitative and qualitative research methodologies is implemented. In our quantitative analysis, we use a five-step pipeline that includes Topic Modelling (Pachinko Allocation), translation of the topic words into English, Word Embeddings, Ward Hierarchical Clustering, and a directed graph. In our qualitative analysis, we firstly select randomly one newspaper per decade per outlet seeking to observe whether multilingual historical newspapers are comparable objects of analysis (i.e. is their format similar enough to implement meaningful discourse analysis?). Secondly, we also sample randomly a variety of articles containing our selected key terms in order to assess the social impact of technology using a close reading approach. Our quantitative data analysis reveals three main findings: firstly, we detect a trend in information flattening coinciding with the peak of the Second Industrial Revolution (1890 and 1900), as well as a trend of information complexity during the following decades. Secondly, we observe more nuanced patterns of agreement during the Twentieth century, therefore showing how the social and political polarity during that time did not affect technological related discourses. Thirdly, we notice high rates of content similarity across our three selected key terms over our whole observational time, displaying almost identical wording. These findings resonate with our qualitative analysis, where we observe a certain degree of heterogeneity among both newspapers formatting and our selection of articles, yet very subtly. These outcomes make us speculate with the idea that it is possible to trace a shared Western voice in technological related discourses back to two hundred years ago, exposing the agency of technology as a trigger of cultural flattening in terms of information behaviour.

Keywords

Discourse Analysis, Topic Modelling, Topic Evolution, History of Science, Science and Technology in Society

I INTRODUCTION

Contemporary perceptions about technology tend to frame the latest digital revolution as responsible for triggering an unprecedented wave of social changes in Eurocentric, industrialized

societies (Smil Vaclav Smil, *Creating the Twentieth Century: Technical Innovations of 1867-1914 and Their Lasting Impact* (Madrid: Oxford University Press, 2005)). Placed by some authors during the last decades of the twentieth century (Castells Manuel Casatells, *The Rise of the Network Society* (Chichester, West Sussex: Wiley-Blackwell, 2010), Wacjman Judy Wajcman, *Pressed for Time. The Acceleration of Life in Digital Capitalism* (Chicago: University of Chicago Press, 2014)) under the label of post-industrialism, it has been empirically proven that the number of technical innovations arriving to society during the last thirty years is indeed dramatically higher than at any other period of recorded history (Kelly **Kelly**). Yet some authors signal the latest decades of the nineteenth century as a moment when a rapid influx of technological inventions not only profoundly changed social life across the Western world at a coetaneous level but had the capacity of imprinting long-lasting historical and social effects that can be traced up to our present-day times (Smil Smil, *Creating the Twentieth Century: Technical Innovations of 1867-1914 and Their Lasting Impact*):

> That period ranks as history's most remarkable discontinuity not only because of the extensive sweep of its innovations but also because of the rapidity of fundamental advances that were achieved during that time. (6)

The Second Industrial Revolution has been extensively analyzed across countries from historic (Stearns Peter Stearns, *The Industrial Revolution in World History* (New York: Taylor Francis, 2021)), economic (Cipolla Carlo M. Cipolla, *The Fontana Economic History of Europe* (Glasgow: Fontana/Collins, 1973)), and geographic perspectives (King Steven King and Geoffrey Timmis, *Making Sense of the Industrial Revolution* (Manchester: Manchester University Press, 2001)). Coetaneous views expressed by nineteenth century citizens regarding the profound social and historic changes that the Industrial Revolution was imprinting in society have as well received some critical attention, mostly in the field of Social History (King King and Timmis): "We suggest that there was a widely view amongst contemporaries that, irrespective of whether or not they approved, theirs was a society undergoing substantial change. They felt that 'something was happening' much as the Internet gives modern society a feeling of cumulative and fundamental change". (6)

This paper seeks to engage with these academic conversations by contributing with the analysis of a yet unexplored aspect: we aim to observe the social impact of technology by inspecting its agency as a trigger of information homogenization and cultural flattening across time (one hundred and ten years) and space (five different countries: Spain, France, Italy, Germany, and The United States) using multilingual historical newspapers as a proxy. We define information homogenization as a lack of semantic diversity in our data processing, outputted by a five step pipeline that we borrow from Fernández Fernández and Savcisens Elena Fernández Fernández and Germans Savcisens, "A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)," in *Digital Humanities in the Nordic and Baltic Countries 2023. Sustainability, Environment, Community, Data. Online Conference* (CEUR-WS, 2023), https://doi.org/10.5281/zenodo.7742461. It firstly uses topic modelling on a key-term (telephone, gasoline, and iron) filtered sample of our multilingual newspaper corpus, followed by an English translation of the Topic Modelling multilingual words. Afterwards we use word embeddings to detect semantic proximity across those words. Finally, we implement Ward Hierarchical Clustering and a directed graph to accomplish yet another round of semantic affinity filtering and only select topics that are very similar across newspapers. High rates of semantic similarity in our data analysis could be understood as a proxy of technologically related information being reported similarly across our observational

countries. We define cultural flattening as a process of social standardization as a consequence of a similar assimilation of our selected technological-related terms in our different objects of study (multilingual historical newspapers).

Recent work by Fernández Fernández and Savicens Fernández Fernández and Savcisens, "A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)" claims the existence of growing patterns of information agreement in contemporary press discourses about sustainability (1999-2018), as well as the existence of a clearly streamed Western voice that can be traced back up to twenty years ago, yet becoming more pronounced as we approach contemporary times. The authors argue that contemporary newspapers of record across Western countries (*The Times, El País, Le Figaro, The New York Times International, NZZ, La Stampa*) encode very similarly sustainability related discourses, providing highly homogeneous information to multilingual readers of their respective countries. These findings resonate with the ideas presented by other authors such as Hermans and Chomsky Edward S. Herman and Noam Chomsky, *Manufacturing Consent. The Political Economy of the Mass Media* (New York: Pantheon Books, 2002), who also note qualitatively the homogeneity of Western media. However, unlike Fernández Fernández and Savcisens Fernández Fernández and Savcisens, "A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)," Hermans and Chomsky Herman and Chomsky, *Manufacturing Consent. The Political Economy of the Mass Media* argue that Western mass media is controlled by private corporations and Governments seeking to influence public opinion worldwide (the propaganda model, so to speak).

In this paper we seek to engage with these academic conversations, and specifically with the ideas presented by Fernández Fernández and Savcisens Fernández Fernández and Savcisens, "A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)," as we are intrigued to explore a) whether their findings are exclusive to sustainability related discourses or if similar trends in information behaviour could be detected in technologically related ones broadly speaking, and b) if it is possible to note trends of information homogenization as early as the mid-nineteenth century or if it is an exclusive twentieth-first century phenomenon. In order to do so, we quantitatively analyze Western public discourses as recorded by historic multilingual newspapers about technological innovations either arriving to society during the nineteenth century (such as the telephone) or being widely used during this historical time (such as gasoline and iron). We aim to observe to what extend their press coverage was similar or different across time and space during a time-spam of one hundred and ten years (1830-1940). Our observational time coincides with profound social changes that hastened processes of cultural flattening (such as a first peak of globalization during the last decades of the nineteenth century) as well as intense social and political division (the First and Second World Wars, and the interwar period).

We follow the definition of globalization proposed by Stearns Stearns, *The Industrial Revolution in World History*: "Globalization is simply the intensification of contacts among different parts of the world and the creation of networks that, combined with more local factors, increasingly shape human life" (2). We also agree with both Stearns Stearns and Robertson Robbie Robertson, *The Three Waves of Globalization. A History of a Developing Global Consciousness* (London: Zed Books, 2004) proposal that understands globalization as a gradual historic process with roots that can be traced back to 1000 CE, rushed by specific historic moments. And, again in agreement with both of them, we also consider the second half of the nineteenth

century one of the most crucial moments in recent human history, coinciding with the climax of the Second Industrial Revolution (Stearns, *The Industrial Revolution in World History*):

> The industrial revolution effectively caused the modern version of globalization. In turn, globalization, particularly since the mid-nineteenth century, has reshaped the industrial experience in many ways. The connection first became clear in the late nineteenth century, when international trade grew at unprecedented rate. (225)

The nineteenth century witnessed as well the rise of the modern nation around the second half of the nineteenth century following the collective wave of revolutions of 1848 (Hobsbawm Eric Hobsbawm, *The Age of Capital (1848-1875)* (London: Abacus, 1995)). The role of the press as a force of national cohesion and identity building has been discussed by authors like Anderson Benedict Anderson, *Imagined Communities* (London: Verso, 2006), who proposes how during the second half of the nineteenth century, Western citizens would engage in a daily newspaper reading ritual across their respective countries. This collective action of simultaneously reading a same commodified object (newspapers of record) would create a shared feeling of community contributing to foster the creation of a common national identity at a foundational moment of history. In this paper we seek to make a different argument though. While we do not disagree with Anderson, we are mostly interested in observing how discourses about technological innovations were encoded similarly or differently in the press as we consider historic multilingual newspapers an ideal medium where to analyze the social impact of technology as a trigger of cultural flattening.

Thus we use as a proxy a dataset of multilingual historic newspapers (Spanish: *El Imparcial*, French: *Le Figaro*, English: *The New York Herald*, German: *Neuer Hamburger Zeitung*, and Italian: *La Stampa*). We select three technological related terms: gasoline, iron, and telephone, and filter our corpus with them, as a first exploratory approach in the analysis of public discourses about technology. Our selection of these terms is motivated by Smil's ideas Smil, *Creating the Twentieth Century: Technical Innovations of 1867-1914 and Their Lasting Impact*, who signals these technologies as important elements within the categories of combustion engines, new materials and new synthesis, or communication and information, that he uses to segment the Second Industrial Revolution into different thematic sections. In dialogue with Smil *Smil*, we believe that those were mainstream technologies that were widely used across our countries of choice during our observational time, and consequently, should be considered as relevant study cases for a first approximation to the analysis of the agency of technology as a actor of cultural flattening.

Afterwards, and as mentioned, we follow the five-step pipeline proposed by Fernández Fernández and Savcisens Fernández Fernández and Savcisens, "A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)," seeking to track rates of homogeneity versus heterogeneity of information across countries as well as their temporal evolution. This pipeline is written as independently as possible of language and archival type, and can easily be adapted to support more languages and parse other newspapers. We interpret less diversity of topics as a signal of information homogeneity, and speculate with the idea that this uniformity in press discourses could be considered as a shared Western voice that may signal the existence of a common technologically related shared press identity. By extension, we argue that technology could be considered as an element of cultural cohesion across Western societies that can be observed as early as the second half of the nineteenth century and that co-exists and over-performs both global processes of globalization and episodes of social polarization such as the First and Second World Wars.

Additionally, we implement a qualitative research methodology seeking to complement our quantitative data analysis. In order to assess whether multilingual historical newspapers are comparable objects of discourse analysis from a formatting perspective (i.e. newspaper's sections, article length, font size, or advertising), we select randomly one newspaper per decade per publication (a total of forty one newspapers). We also inspect a randomized sample of articles containing our three key terms selected from all our publications. We want to assess qualitatively how discourses reflect the social impact of technology in order to have a better understanding of our dataset, and by extension, implement a better interpretation of our data analysis outcomes.

II STATE OF THE ART

The Industrial Revolution and its impact in society has received a considerable amount of attention across disciplines. Existing state of the art qualitatively has focused its attention on topics such as world history (Stearns Stearns, *The Industrial Revolution in World History*), geography (King King and Timmis, *Making Sense of the Industrial Revolution*), or economic history (Cipolla Cipolla, *The Fontana Economic History of Europe*). Furthermore, the historic specificities of different countries during this period of time have been as well extensively analyzed by scholarly critique (Vilar Juan Bautista Vilar, *La primera revolución industrial española (1827-1869)* (Madrid: Ediciones Itsmo, 1990), Veblen Thorstein Veblen, *Imperial Germany and the Industrial Revolution* (Ann Arbor: The University of Michigan Press, 1966), Gomellini Gianni Toniolo Matteo Gomellini, "The Industrialization of Italy, 1861–1971," ed. Jeffrey Gale Williamson Kevin Hjortshøj O'Rourke, 2017,); just to mention a few).

In recent years, the field of Digital Humanities has contributed with new and highly innovative perspectives to existing qualitative analysis about the Industrial Revolution by using computational methods in the analysis of the past. Yet existing work has been accomplished using predominantly English sources. Under the umbrella of the research project Living with Machines (based at the Alan Turing Institute in London), a variety of new tools, research articles, datasets, and methodologies have been developed to deepen the analysis of the Industrial Revolution in the United Kingdom using Victorian British newspapers as an object of study. Coll Ardanuy et al. **coll** use historic newspapers in English to create a knowledge base in which toponyms appearing in the press are linked to real-life geographic locations. Beleen et al. **2019LivingWM** challenge the utility of digitized historical newspapers as historical objects of analysis of their contemporary time of publication by questioning the fragments of society and reality that they capture and therefore calling into question the readings of the past that they facilitate. Similarly, Beleen et al. **Beleen**, question the bias of the British Newspapers Archive in terms of range (historical coverage), demographic representation, and OCR quality.

In this paper, we seek to dialogue with existing scholarly work, qualitatively and quantitatively, and across fields, in their analysis of the Second Industrial Revolution. Our major contribution lies in the investigation of the agency of technology in shaping information behaviour during the nineteenth and early-twentieth centuries, and specifically, as a trigger of cultural flattening and information homogenization.

Furthermore, and as mentioned, we predominantly dialogue with Fernández Fernández and Savcisens Fernández Fernández and Savcisens, "A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)." The authors use a dataset of multilingual newspapers (English, French, German, Italian, and Spanish), over a time-span of twenty years (1999-2018), using the same pipeline that

we implement in this paper yet targeting sustainability related discourses. Their findings show increasing trends of information homogeneity as we approach contemporary times, relating processes of globalization and information homogenization. Moreover, they also note high rates of information uniformity overall, showcasing a clear Western common voice that can be traced back to the last twenty years. We seek to expand their analysis, and by using one hundred and ten extra years of data (1830-1940 to complement their 1999-2018 observational time) we cover in total almost two hundred years of Western history, therefore contributing to deepening our understanding about the historic role of technology in processes of Western identity formation.

III CORPUS

Our dataset is composed of a variety of historical multilingual newspapers in English (*The New York Herald*), French *Le Figaro*, German (*Neue Hamburger Zeitung*), Spanish (*El Imparcial*), and Italian (*La Stampa*). Our observational time covers one hundred and ten years (1830-1940). During that time some newspapers experienced different ownership phases, and by extension, also changing ideological management.

Le Figaro is known for its historic conservative centered-right political orientation (Kuhn Raymond Kuhn, *The Media in Contemporary France* (Glasgow: Open University Press, 2011)), while *The New York Herald* was a slightly sensationalized outlet, yet the most successful and widely circulated newspaper in the second half of the nineteenth century in the United States (Crouthamel James L. Crouthamel, *Bennett's New York Herald and the Rise of the Popular Press,* (New York: Syracuse University Press, 1989)). *El Imparcial* held progressive liberal views (Magro and Carvajal Angel Bahamonde Magro and Luis Enrique Otero Carvajal, eds., *La sociedad madrileña durante la Restauración (1876-1931). Volumen II. El sistema político de la Restauración. El horizonte cultural. Opinión y medios de información. Conflicto social y clases trabajadoras* (Madrid: Graimo SA, 1989)). *La Stampa* showed a similar political orientation than *Le Figaro*: slightly conservative, leaning towards the social and political views of industrial-capitalist readership. Moreover, during our observational time, it was a somewhat regional newspaper mostly focusing on the Turin area, only becoming one of Italy's newspaper of record during the second half of the twentieth century (Saitta Eugénie Saitta, "The Transformations of Traditional Mass Media Involvement in the Political and Electoral Process. A Case Study on Political Journalism in France and Italy since the 80s.," *ECPR Joint Session. Workshop 9: 'Competitors to Parties in Electoral Politics: The Rise of Non-party Actors'.*, 2006,). *Neue Hamburger Zeitung*, very similarly to *La Stampa*, targeted an educated and intellectually active readership (Tornier Klaus Tornier, *Hamburger Pressegeschichte in Zeitungstiteln vom 17 . bis 20. Jahrhundert* (Norderstedt: Books on Demand, 2021)). So, in spite of being a relatively regional newspaper, it included extensive coverage of international and national affairs. Consequently, although our corpus presents some degree of heterogeneity in terms of scope and readership, we believe that our selected newspapers are similar enough to perform a meaningful cross-cultural comparison to inspect the social impact of technology in Western countries across our observational time.

Our motivation for choosing these newspapers is fundamentally related to availability. The authors of this article are fluent in English, Spanish, French, and German, and their domain knowledge expertise focuses on Western nations where those languages are spoken. Moreover, we are interested in selecting newspapers with a tendency towards politically-centered views across countries with large historical availability covering our observational time (1830-1940) accessible in a data-mining friendly format. We are particularly interested in building a corpus of relatively similar textual objects across countries approximately targeting similar scope and

readership and therefore facilitating a relevant multi-cultural analysis when it comes to identify the agency of technology as a cross-border trigger of information flattening.

While our corpus is not complete (there are some decades missing) this is a common-place scenario in open source historical newspapers datasets, and a challenge that Digital Humanists frequently encounter. Although we are aware of the limitations that our corpus presents when it comes to the articulation of our main thesis (the agency of technology as a trigger of information homogenization and by extension of cultural flattening), we still believe that it provides a relevant medium of analysis where to observe the social impact of technology in Western public discourses.

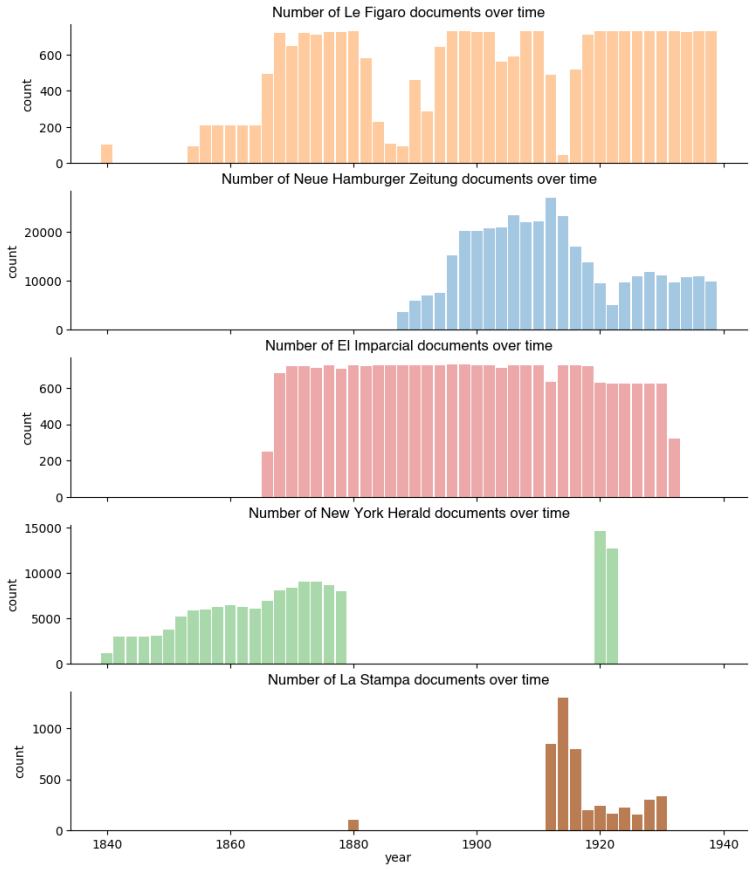

Figure 1: Newspapers Data Availability

Each newspaper has been retrieved from different sources and presents different historical ranges and accessibility. Table 1 describes each newspaper's range of available dates as well as retrieving source, and Figure 1 provides a cross-newspaper comparison of newspaper availability.

IV CORPUS QUALITY

The raw data for these newspapers was gathered by digitising the original issues using Optical Character Recognition (OCR). Since this method is not reliable and regularly results in errors, determining the corpus quality is necessary. To achieve this, Thomas Benchetrit's OCR-qualification pipeline Thomas Bench, *anxietyNews*, https://github.com/ThomasBench/anxietyNews, 2022 was used. It represents text quality as the proportion of words in a document that were correctly digitised, which is determined by checking whether they exist in the

Table 1: Corpus availability and sources

Journal	Time Period	Source
Le Figaro	1860-1920	Gallica
The New York Herald	1840-1888, 1920	Library of Congress
El Imparcial	1860-1930	Biblioteca Nacional de España
La Stampa	1882, 1910-1930	La Stampa Archives (by Bas et al.)
Neuer Hamburger Zeitung	1888-1930	Staats- und Universitaetsbibliothek Hamburg Carl von Ossietzky

respective language's dictionary. Assuming that words are either digitised correctly or not (i.e. no word is digitised as a different valid word), this metric gives us a percentage of overall text correctness. The dictionaries used here are the ones provided by the Enchant spellchecking library. The results are documented in Figure 2.

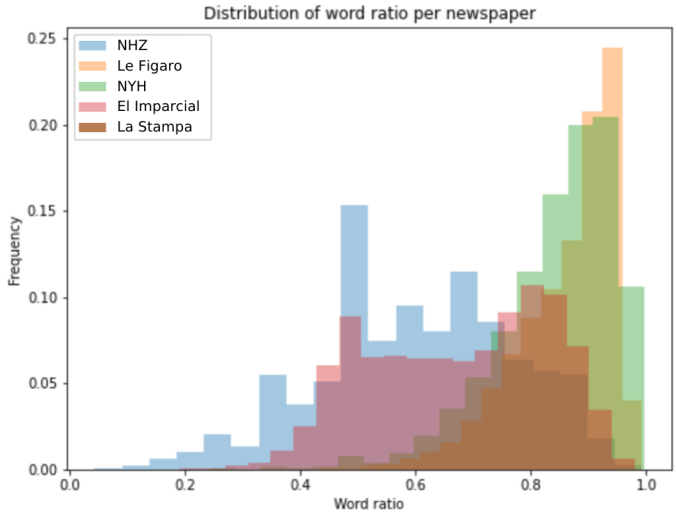

Figure 2: OCR quality of newspapers

V ARTICLE SPLITTING

The data of each newspaper was provided in the form of "documents" which gathered the newspaper text in different ways, namely :

- **Le Figaro** : one "document" is an entire newspaper for a given day, with no clear delimitation between articles.
- **El Imparcial** : like for *Le Figaro*, one "document" is an entire newspaper for a given day. However, very often, different articles are separated by a newline.
- **Neue Hamburger Zeitung** : there is one "document" per article.
- **La Stampa** : there is one "document" per article.
- **New York Herald** : one "document" represents one page of a newspaper issue for a given day. Within these pages, there is no clear delimitation between articles.

For topic modeling, this lack of delimitation between articles is problematic. Since entire newspaper issues, or even a single page of one newspaper issue, often contain many articles that tackle different topics, trying to perform topic modeling on these unaltered documents would lead to very confused and imprecise results. Therefore, a pipeline to split these documents into

individual articles is necessary. These pipelines were determined empirically.

5.1 El Imparcial

Since most articles in the *El Imparcial* dataset are split by a newline, we simply consider every line of each document of length in characters $l > 20$ a distinct article. This method was adopted from Elisa Michelet's master project Elisa Michelet, "An Industrial West? Analyzing Multilingual Newspapers Discourses about Technology during the Second Industrial Revolution (1840-1930)," *EPFL*, 2023,

5.2 Le Figaro

The documents in this dataset contain many newlines that don't necessarily denote new articles. However, new articles are preceded by a newline, and many of their titles are written in uppercase. We therefore consider the beginning of a new article any newline followed by a string of length l whose alphabetic characters are uppercase with a frequency larger than k. Empirically, we determined for best results $l = 30$ and $k = \frac{2}{3}$.

Table 2: Keywords used to filter articles

	Gasoline	Telephone	Iron
French	essence OR petrol	telephone	" fer "
Spanish	gasolina OR petroleo	telefono	" hierro "
German	benzin OR kraftstoff	telefon OR telephon	" eisen "
English	gasoline OR petrol	telephone	" iron "
Italian	benzina	telefono	" fierro "

5.3 The New York Herald

This dataset provided the .xml documents from the OCR as well as their textual interpretation. The XML files contain information about the size in pixels of the detected words. Since most titles of articles are written in a larger font than the rest of the text, this allows us to split articles based on this size information. More specifically, we consider the beginning of a new article any part of the text such that there is a sudden change in font size $diff_{size} < f$. Additionally, we require that the first n words of an article (i.e. the title) have uppercase characters with a frequency larger than k. Empirically, we determined $f = -20$, $n = 2$ and $k = \frac{2}{3}$.

VI KEYWORD DETECTION

To determine the conversations surrounding different technologies in the newspapers of our corpus, the articles were filtered and only those that include the given keyword were kept. The exact words used to filter the newspaper documents are summarized in Table 2.

For simplicity, the words were filtered only in their basic noun form. However, words that include the word root were also included (e.g. "petroleum", "telefonisch"). In the case of gasoline, the synonym "petrol" was also considered, as they are often used interchangeably. Since the keywords were analyzed independently, no special consideration was given to documents that contain multiple keywords. Specifically for the keyword "iron", the filtering was done with a leading and trailing whitespace, as shown in **??**. The word being so short, filtering otherwise led to there being a very large number of unrelated results due to the all words containing "iron" as a subpart.

We believe that our key terms (telephone, gasoline, iron) belong to technologically specialized vocabulary, and therefore, aiming to avoid semantic ambiguity, we have decided to not include any synonyms. We observed, for example, how "essence" (gasoline in French) retrieved a high amount of semantically irrelevant articles due to its polysemic character (for example, it also means perfume). Therefore, we have decided to stick with our selected terms in order to avoid false positive results.

VII METHODS

7.1 Topic Modeling

The goal of this article is to detect the nature and evolution of discussions surrounding technology in different countries over time. In order to observe mainstream discourses in the newspapers in our corpus, we firstly use topic modeling as a proxy to classify documents. Our goal is to a) determine the degree of homogeneity vs heterogeneity in terms of information encoding (i.e. our selected technical terms) across our newspapers of choice, and b) perform a manual reading to the multilingual words (translated into English) outputted by our pipeline. Therefore, Topic Modelling is an adequate method for our research needs.

7.1.1 Pachinko Allocation Model

The Pachinko Allocation Model (PAM) Wei Li and Andrew McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in *Proceedings of the 23rd international conference on Machine learning* (2006), 577–584 was used to detect topics in documents. This model was chosen because it is able to capture correlations between topics. It models the vocabulary as the leaves of a Directed Acyclic Graph (DAG), and the topics as the nodes. Correlations are then represented by the edges between different nodes and leaves of the graph, which can be between words or between topics, the latter resulting in a set of super-topics.

Since PAM performs much better when the number of super-topics and sub-topics is pre-defined Wei Li, David Blei, and Andrew McCallum, "Nonparametric bayes pachinko allocation," *arXiv preprint arXiv:1206.5270*, 2012, a grid search is performed to obtain the optimal parameters. This process is detailed below.

7.1.2 Best parameter calculation

To train a Pachinko model, we require two arguments k_1 and k_2 denoting respectively the number of super-topics and the number of sub-topics in the corpus. Since we do not know these beforehand, we create the model with all possible pairs of k_1 and k_2 with $k_1 \in [1, 2]$ and $k_2 \in [k_1, 14]$. These values were chosen empirically. To determine the best parameters, we compare their coherence value, which is a metric that aims to quantify the coherence and human interpretability of the top words per topic returned by a topic model David Newman et al., "Automatic evaluation of topic coherence," in *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (2010), 100–108. Specifically, the chosen metric is c_V-coherence because its results are the closest to a human interpretation Michelet, "An Industrial West? Analyzing Multilingual Newspapers Discourses about Technology during the Second Industrial Revolution (1840-1930)." By default, the method used to determine the best parameter is the one-standard-error rule, which selects the model whose prediction error is at most one standard error worse than that of the

best model Eva Cantoni et al., "Longitudinal variable selection by cross-validation in the case of many covariates," *Statistics in medicine* 26, no. 4 (2007): 919–930. However, this method is often quite ungenerous and only detects 1 or 2 subtopics. In those cases, we used the parameters of the best model as determined by the grid search. Since the optimal number of topics in a corpus changes based on the documents used to train the model, this calculation was repeated for every combination of keyword, newspaper and time-span.

7.2 Model training

Separate models were trained for every combination of keyword, newspaper and time-span in order to track the evolution of topics across time and newpapers. The pipeline from pre-processed documents to trained model for a given newspaper is:

- splitting the set of articles into time-spans of 10 years
- using Spacy Matthew Honnibal and Ines Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing" (To appear, 2017) to retrieve the lemmas of words in the articles, keeping only lemmas that are not stop words and that are alphanumerical. Lemmas that have less than 4 characters were also discarded, since they often represent noise from the OCR.
- determining the optimal number of super-topics and sub-topics for the given model as described in section 7.1.2
- using Tomotopy's Pachinko Allocation training model bab2min, *Tomotopy*, https://github.com/bab2min/tomotopy, 2019 to train the given model following the pipeline proposed by Germans Savcisens, *Simple Topic Modelling Examples*, https://github.com/carlomarxdk/topic_modelling, 2022 with the previously determined parameters

Finally, each given model was algorithmically described as the set of its $k = 15$ most representative words as determined by the topic modeling. From these keywords, an overarching topic was manually assigned.

Algorithm 1 Global Topic Aggregation

$\mathbf{\Gamma} \in R^{p \times v}$ is a matrix containing all discovered $\boldsymbol{\gamma}_{t,n,d}$ and p is a total number of topics
$\mathcal{A} = \text{distance}(\mathbf{\Gamma}_i, \mathbf{\Gamma}_j)\ \forall\ i,\ j \in \{1, 2, 3, ..., p\}$ ▷ $\mathcal{A} \in R^{p \times p}$ is a distance matrix
for $t = 1 : T$ **do**
  $\mathcal{A}^t = \text{subset}(\mathcal{A}, t)$ ▷ $\boldsymbol{A}^t$ contains only between topics of time spans t
  $thr_t = \text{Silhouette}(\text{Ward}(\mathcal{A}^t))$
  $\mathcal{V}_t = \text{Ward}(\mathcal{A}^t, thr_t)$ ▷ $\mathcal{V}_t$ contains sets of similar inner topics
  $\boldsymbol{\nu}_t \leftarrow \text{AverageSimilarTopics}(\mathcal{V}_t, \mathcal{V})$ ▷ $\boldsymbol{\nu}_t$ contain representations of global topics
end for
$\tilde{\mathcal{V}} = \text{stack}(\boldsymbol{\nu}_t\ \forall\ t \in\ \{1, 2..t\})$

7.3 Topic Similarity

To cluster similar topics originating from different newspapers, we calculate the similarity score between each pair of topics. We do so separately per each time span, dialoguing with Fernández Fernández and Savcisens (Fernández Fernández and Savcisens, "A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)").

First, we start by estimating the similarity between inner topics. To make the multilingual topic vectors $\boldsymbol{\gamma}_{t,n,d}$ comparable, we translate words associated with each topic vector (from

Italian, Spanish, German, and French) into English using the Google Translate API. Since all translations were mapped onto modern English, and the translation model is capable of mapping "old" spellings to their modern counterparts, it is reasonable to assume that the translations allow for linking articles across time regardless of language evolution.

In some cases, a word translates into a phrase. To add these words to a vocabulary, we split the phrase into separate words and equally redistribute the probability associated with the phrase to these separate words.

Second, as the topic similarity measure is based on word embeddings (Nikolaos Aletras and Mark Stevenson, "Measuring the similarity between automatically generated topics," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers* (2014), 22–27), we need to know all the words across every newspaper and time span – we create a global vocabulary set, V. Global vocabulary consists of all the unique English words across every time span and every newspaper. Now we can represent each $\boldsymbol{\gamma}_{t,n,d}$ using the global vocabulary. It might happen that a word that is mentioned in one subset of data (e.g., *cat*) might be missing from the other topic (as it was never mentioned in that particular subset of data). In that case, we update the vocabulary of the $\boldsymbol{\gamma}_{t,n,d}$ and assign a probability of 0 to all newly added words. The aligned vectors can now be used to calculate the similarities between topics.

Further, we stack all $\boldsymbol{\gamma}_{t,n,d}$ into a matrix $\boldsymbol{\Gamma} \in R^{p \times v}$, where p is the total number of topic vectors (across every time span and newspaper), and v is the length of the global vocabulary. $\boldsymbol{\Gamma}$ contains the probabilities of each word in every discovered inner topic.

Using matrix $\boldsymbol{\Gamma}$, we can extract word embeddings, i.e., numerical representations of words (Aletras and Stevenson). Each word, V_i is represented as a vector where dimensions correspond to topics and values correspond to the probability of word V_i in each inner topic (i.e., the i-th column of $\boldsymbol{\Gamma}$).

We calculate the topic similarity based on the *average pairwise cosine similarity* of the N-top[1] words in each topic (Aletras and Stevenson).

7.4 Global Topic Aggregation

To analyze whether newspapers share discussion points over a specific time span, we look at the topic similarity. If multiple inner topics are similar enough, we assume they share similar contexts. The similarity is calculated between each pair of topics (produced by all newspapers over the same time span). We cluster inner topics with the use of the Hierarchical Cluster Analysis (HCA) with Ward's linkage function (Anna Großwendt, Heiko Röglin, and Melanie Schmidt, "Analysis of ward's method," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* (SIAM, 2019), 2939–2957).

If the similarity between topics is above a certain threshold, the HCA model clusters topics together. However, we do not have any prior knowledge of the optimal threshold value. Thus, we vary the threshold and look at the quality of the formed clusters. We use the Silhouette method (Peter J Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics* 20 (1987): 53–65) as a proxy for the clustering quality. This method estimates whether topics are closer to the members

1. $N = 10$ for Gasoline and Telephone; $N = 15$ for Iron

of their own clusters or vice versa. For each time span, we find a separate optimal threshold value.

To make clusters comparable, we create cluster representations by averaging the representations of inner topics that end up in the same cluster (i.e., the average of the corresponding rows in $\mathbf{\Gamma}$). We further refer to those as global topics.

The vectors associated with global topics are stacked into another matrix $\tilde{\mathbf{\Gamma}} \in R^{g \times v}$, where g is the total number of global topics. The representation of a word, V_i, is now based on the matrix of the global topics, $\tilde{\mathbf{\Gamma}}$ (i.e., i-th columns, as in the case with the $\mathbf{\Gamma}$).

7.5 Temporal evolution of topics

We are interested to see how topics change through time – do they disappear, split into multiple discourses, or stay unchanged, etc. To examine the evolution of global topics, we look at the similarities between global topics between the adjacent time spans, t and $t+1$ (i.e., between topics of different time spans). We calculate similarities based on the above-mentioned average pairwise cosine similarity – this time, we use word embeddings from $\tilde{\mathbf{\Gamma}}$. We draw connections between the topics of the adjacent time spans if their similarity is above a certain threshold, t. **?** suggests setting the threshold based on the n-th quantile of the cumulative distribution of similarity scores. When we estimate similarities between topics of every pair of t (current) and $t+1$ (next) time spans. We order the scores, find the 90-th quantile of the cumulative distribution, and draw arrows between the pair of topics only if their similarity is higher than the 90-th quantile [2].

Due to the multi-source nature of the data, our algorithm might capture some noise. To substantiate the results, we manually inspect and correct the noisy output of the algorithm. First, we manually inspect the connection between topics if the similarity falls within the 85-th and 90-th quantile. We draw the arrow if the topics share at least one top word. Second, we remove global topics that consist only of one newspaper. We also remove topics if the values of the top ten words (in $\tilde{\mathbf{\Gamma}}$) are below 0.03[3]. We then manually inspect topics if the values of the top twenty words (in $\tilde{\mathbf{\Gamma}}$) are below 0.05. This procedure helps to remove topics lacking consistency (such as this topic with the following set of the top ten words: *leave, world, think, time, know, look, life, people, man, want*).

Based on the incoming and outgoing arrows, the life of the global topics can progress in several ways: birth, evolution, split, merge, or death. It signifies how public attention and ideas are refined or transformed (Adham Beykikhoshk et al., "Discovering topic structures of a temporally evolving document corpus," *Knowledge and Information Systems* 55 (2018): 599–632). The birth of a topic is characterized by the absence of the incoming arrows - no topic from the previous time span has a similar context. The death of the topic would be the reverse of this case – no topics in the future share a similar context.

If a topic has only one outgoing arrow – it evolves. The topic does not undergo a drastic change. If a topic has multiple out-coming arrows - it splits. The successors reuse similar words and context, but it also involves new words, e.g., the context surrounding these words changes. If a topic has multiple incoming connections – its ancestors merge. The context of the ancestors significantly overlaps. If a topic only has one arrow with a rounded end, the topic dies and does not evolve semantically.

2. we decided on the quantile by manually inspecting graphs
3. we decided on the threshold by manually inspecting the results

VIII RESULTS AND DISCUSSION

After applying our five-step pipeline, and in dialogue with Fernández Fernández and Savcisens Fernández Fernández and Savcisens, "A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)" we use three different metrics in the interpretation of our data analysis: diversification, attention, and geography. We are interested in observing whether topics become more diverse or simplified over time, showing incremental rates of polarity or agreement, and therefore seeking to analyze the plausible origins of a shared Western voice in technological related discourses as early as two hundred years ago (and by extension observing the agency of technology as a trigger of information homogenization and cultural fattening). To measure topic diversification or simplification, we use as a proxy both the labels of the topics (that we manually annotate) as well as broader semantic categories that we implement to create a second classification of topics qualitatively (and we use same colors inside the boxes of the topics to indicate equivalent subject matters). We also quantify which topics receive the highest and lowest rates of attention by counting the number of represented newspapers as we seek to observe conflict-affinity trends over time. Finally, we are interested in measuring the geographic distribution of newspapers present in each topic, as we wonder whether it is possible to detect consistent trends of cultural diversity influencing information behaviour and whether changing political landscapes (i.e. the First and Second World Wars and the interwar period) may have any effect on trends of information affinity.

8.1 Telephone

Figure 3 shows the historical evolution of discourses related to the key term telephone as recorded in our corpus. Our observational time covers 1870-1930, as those are the years where we could find documents containing the word telephone. We have merged 1870 and 1880 into one decade as the number of documents in 1870 was not big enough to implement our pipeline. The telephone was invented in 1876, and that explains the reduced amount of information in the 1870s decade.

Following the same method as Fernández Fernández and Savcisens Fernández Fernández and Savcisens, we use boxes to represent global clusters of topics. Each box includes semantically similar words that we have grouped in the fourth step of our pipeline using word embeddings on the English translation of the multilingual topics that PAM outputted. We provide information of each newspaper represented in each decade and topic using small coloured squares at the bottom of each box (topic) and on top of each date (decade), matching a colour-legend that we include at each figure. We also gather topics qualitatively into semantically similar categories (i.e. real estate, politics), and we indicate this by colouring the background of each box using the same tone. We depict the connections that the directed graph calculates using arrows, and we explain arrow representation (types of connection) in a legend. Additionally, we represent the number of distinct inner topics in each cluster by adding a line of dots above each box. Topic clusters that suffered from under-fitting and therefore included too many semantically different inner topics to assign a coherent label were manually removed. We provide similar figures containing the same information for all our selected three key terms.

Just to provide a reading guide to understand how to interpret our figures using one column as an example (again, in dialogue with Fernández Fernández and Savcisens Fernández Fernández and Savcisens), in the time-spam 1870-1890 (first column in *Fig. 3. Data Analysis of Telephone*), it is possible to observe ten different topics (boxes): crime (two times), military, domestic

service, entertainment, politics (two times), ship industry (two times), and finance. However, we (subjectively) group those ten topics into six different semantic categories that we showcase by colouring the boxes: 1. politics (grey), includes topics military and politics (twice); 2. crime (green), includes the two crime topics; 3. real estate (lilac), includes the domestic service topic; 4. entertainment (lime green), includes the entertainment topic; 5. ship industry (orange), includes the two ship industry topics; and 6. finance (blue), includes the finance topic. We will use these broad semantic categories (and colours) in the rest of the time-spans in all our data visualizations of our three key-terms. In this time-spam (telephone, 1870-1890), the two topics shared by the biggest number of newspapers are finance and domestic service (that include four-out-of-five of the newspapers included in this study: *Neue Hamburger Zeitung, El Imparcial, The New York Herald* and *La Stampa*. The two topics that show the lowest affinity across newspapers are sports (only mentioned by one newspaper) are politics (*Le Figaro*) and ship industry (*Neue Hamburger Zeitung*).

Let's now proceed to the analysis of telephone under the three different criteria of diversification, attention, and geography. In terms of diversification, it is possible to observe a clear trend of information simplification coinciding with the climax of the Second Industrial Revolution around the decades of 1890 and 1900, followed by a progressive fragmentation in the next two decades matching the First World War (1910 decade) and the interwar Period (1920 decade), ending in yet another simplification in the 1930 decade. We observe very little variation of semantic categories over time (seven in total): law and order (grey), domestic service (lilac), entertainment (lime green), ship industry (orange), finance (blue), education (yellow), and First World War (orange). With the exception of education and First World War (both appear only in the 1910 decade), the rest of the topics are repeated across all decades.

We interpret this uniformity as an indication of the similarity of the social impact of this technology (the telephone) in terms of information behaviour across our selection of multilingual newspapers. Our pipeline discriminates semantically different outputs: it only selects the top similar words of the topic models across newspapers. Consequently, these results identify dominant discourses related to our selected key terms (in this case, telephone) during our observational time. Documents where our selected key terms appear in a very different semantic context will produce rare topic models that will not pass the threshold of our word embbedings and ward hierarchical clustering (where we only select words that very similar across different newspapers topics). The fact that our results are quite homogeneous means that all our datasets use those key terms in a very similar way qualitatively. Should there be a lot of semantic variation across newspapers, our pipeline would detect it. The amount of similar words outputted by the topic models would be lower, and consequently, there would be a higher diversity in the topic models. Therefore, semantic uniformity could be correlated with technological social impact uniformity.

While we observe how different topics (i.e. domestic service or real estate) show diverse geographic clustering patterns (sometimes all newspapers appear represented, sometimes only two, and sometimes only one), we do not notice significant semantic differences across topics, indicating the highly homogeneous cultural effect that this technology had across the press of diverse societies in that time-frame. Interestingly, we do not observe any geographic patterns of behaviour (i.e. a Southern European cluster or an English speaking one), and we also do not note any historical variations, even though our selection of countries sided at very different political locations during the First and Second World Wars. The only interesting aspect that we see would be the clustering of *Neuer Hamburger Zeitung* and *The New York Herald* in ship

industry related topics, as those two are the only newspapers headquartered in cities nearby the sea.

In terms of attention, we note five topics that show four-out-of-five newspapers representation: domestic service and finance in 1870, politics both in 1920 and in 1930, and finance in 1870. These results are quite logic, as those are four central semantic categories that are constantly present across our analysis, yet the time distribution patterns contradict the results found by Fernández Fernández and Savcisens Fernández Fernández and Savcisens, "A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)" in contemporary datasets. In there, they noticed a tendency of low-clustering in the first time spans of their study (1999-2003, 2004-2008), followed by an increasing presence of newspapers representation towards the end (2009-2013, 2014-2018).

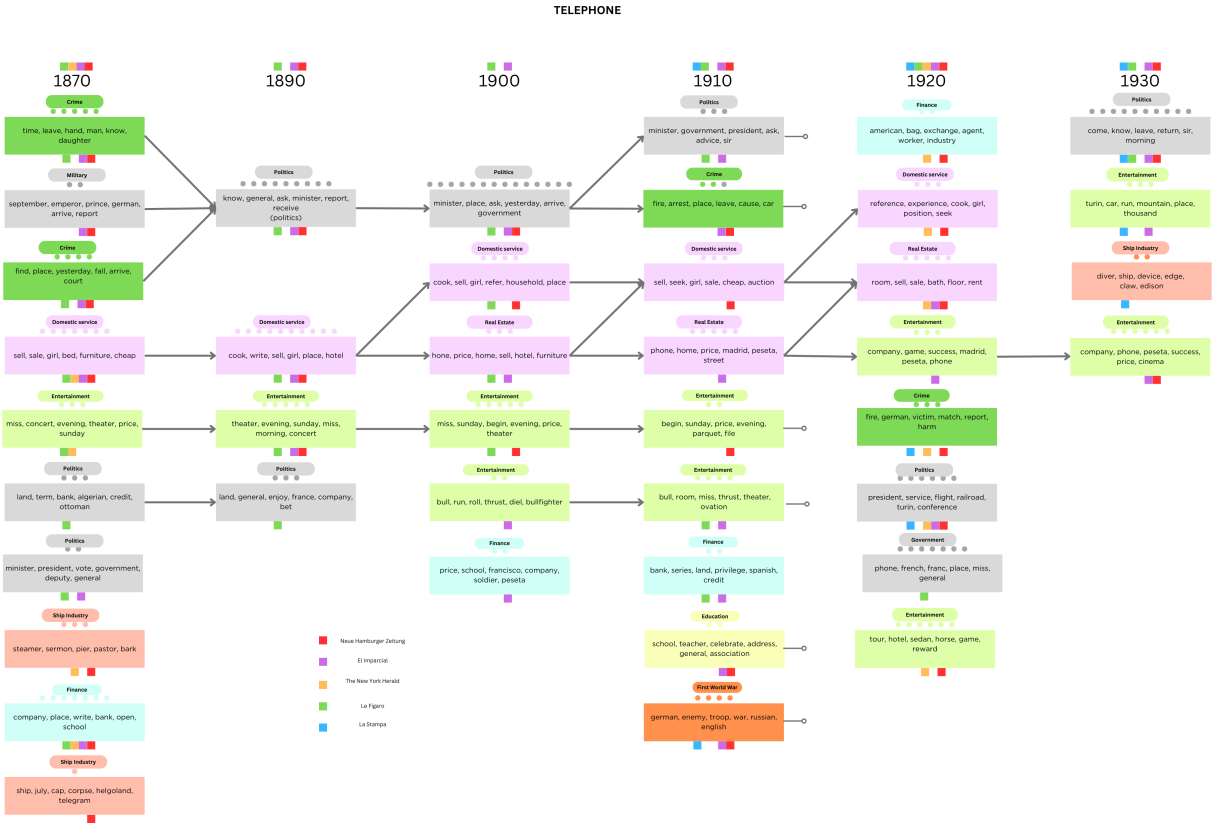

Figure 3: Data Analysis of Telephone

8.2 Gasoline

Gasoline shows very similar results as telephone does across our three different analysis criteria (diversification, geography, and attention). While our observational time is wider (now it covers 1850 to 1930, instead of 1870-1930 as telephone did), we observe a very similar trend of information simplification around 1890 and 1900, a fragmentation around 1910-1920, and another simplification in 1930. The only noticeable difference would be a semantic fragmentation around the decades of 1870 and 1880 (that, as we will explain later on, overlaps with the information behaviour of iron).

Interestingly, topics are very similar to telephone, and we also find the semantic categories

(that we have manually labelled) ship industry (orange), politics (grey), crime (green), entertainment (lime green), finance (blue), and real estate (lilac). On top of those, we detect drug industry (red), transport (fuchsia), and Gold Rush (purple). We additionally note mentions to the First World War in the 1910 decade, but also a report of the Franco-Prussian War in 1870, as well as another mention of war in 1920. We infer that gasoline was used in a military context, and that is why it appears more frequently in conflict-related semantic scenarios than telephone. We also observe a great uniformity in topic distribution: as it happened with telephone, the semantic categories of finance, real estate, politics, and entertainment appear consistently across decades. Interestingly, we note how drug industry only appears during the first half of our observational time (1850-1880), and transport only during the more contemporary decades (1910-1920), therefore showing the real life evolution of the uses of gasoline across countries. Similarly and as we noticed in our analysis of telephone, there is an outlier topic appearing in 1910: Gold Rush (in telephone, it was education).

In terms of geographic diversity, we do not observe any consistent trends (with the exception of ship industry, that still clusters *The New York Herald* and *Neue Hamburger Zeitung*. In our analysis of attention, we do observe remarkably similar trends as in telephone. There are four four-out-of-five topics: Finance (1920 and 1930), Real Estate (1920), and Politics (1930), that are exactly the same ones as in telephone. The only noticeable difference is the historic distribution of these topics: in the case of gasoline they appear towards the end of the observational time, while in the case of telephone, as explained, they appear both at the beginning and at the end. Unlike in telephone, the semantic complexity of the context in which gasoline appears triggers the emergence of many more topics, and that is why we see many more individual topics overall (and this is a trend that we also find in iron). However, the broad semantic categories (that, again, we manually label) remain quite stable over our observational time and across topics, making us speculate with the idea of a shared technological experience that can be detected across different multilingual historic newspapers.

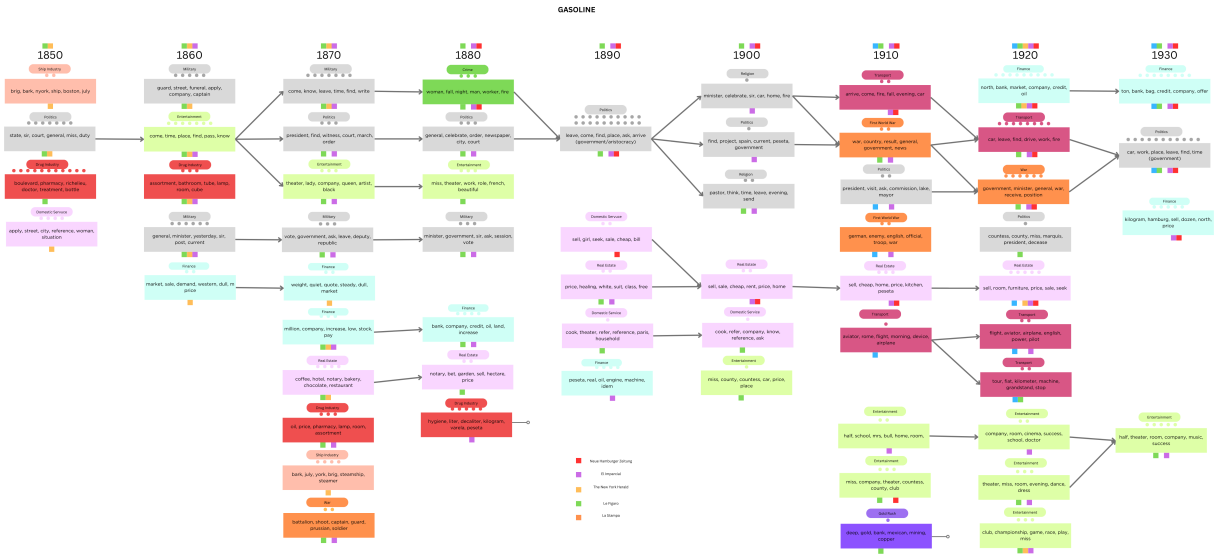

Figure 4: Data Analysis of Gasoline

8.3 Iron

Iron shows a remarkably similar behaviour to gasoline, and also very similar trends to telephone, in both cases across our three different metrics (diversification, geography, and attention).

In terms of diversification, we notice an almost symmetric trend as in gasoline: a fragmentation of information around the decades of 1870-1880 followed by a simplification around the decades of 1890-1900 coinciding with the climax of the Second Industrial Revolution, and another wave of fragmentation coinciding with the First World War and Interwar periods. The most noticeable difference would be a fragmentation of topics in 1930 (against the simplification found in this decade in telephone and gasoline).

Topics are very similar to telephone, and almost identical to gasoline: we also find the general semantic categories ship industry (orange), politics (grey), crime (green), entertainment (lime green), finance (blue), and real estate (lilac) appearing uniformingly across the entire observational time (and, again, those appear both in gasoline and telephone also during the whole observational time). On the top of those, we also find drug industry (red), transport (fuchsia), labour (forest green), and education (yellow). The war topics now proliferate, and we find mentions to the First World War in the 1910 decade, to the Franco-Prussian War in 1870, to the Second World War in 1930, and to some other unknown conflict in 1890. We also deduct, as we did in the case of gasoline, that this term was commonly used in a military context, and that is why it appears more frequently than in the other two terms. Very interestingly, we note how drug industry mirrors the topic behaviour in gasoline and only appears during the first half of our observational time (1830-1880), and transport more frequently during the more contemporary decades (1910-1920), although in this case it also appears during the first decade of our observational time (1830). While telephone and gasoline displayed the outliers gold rush and education in 1910, iron does not, but it shows the topic labour in 1890 and in 1920, as well as the topic of education in 1930, therefore behaving semantically slightly differently (yet still quite similar to the other key terms).

In terms of geography, we do not observe any significant trends other than *The New York Herald* and *Neuer Hamburger Zeitung* clustering in the shipping industry semantic category. However, in terms of attention, we do observe some differences. We see one full representation topic (that contains five-out-of-five newspapers): labour (1920). We also notice two four-out-of-five topics: transport (1910) and war (1910). Therefore, while the temporal distribution is similar to gasoline (more newspaper representation towards the more contemporary times), the topics are different (let's remember than in the case of telephone and gasoline, major representation topics (five-out-of-five or four-out-of-five newspapers represented) were finance, real estate, and politics).

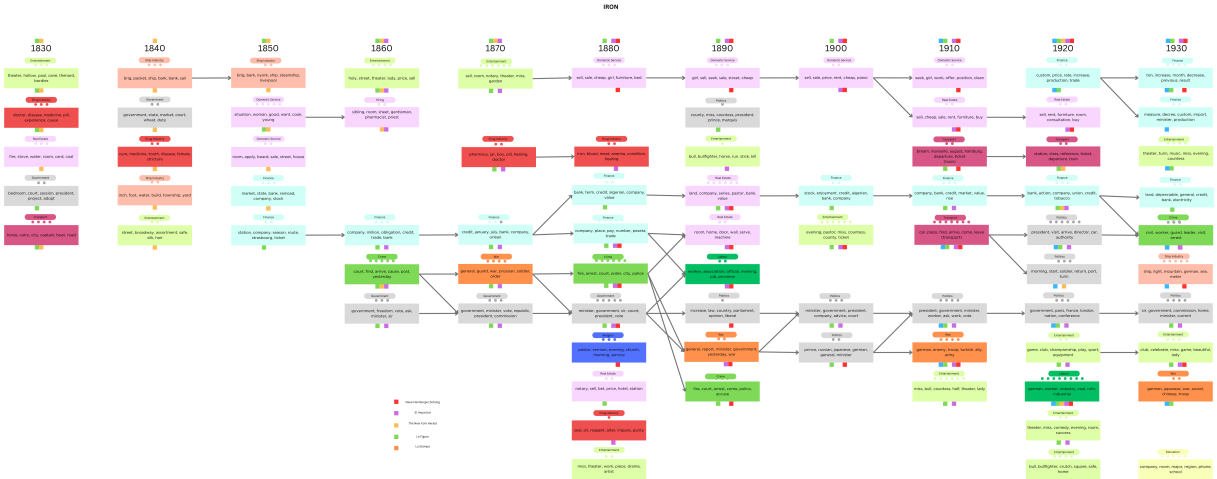

Figure 5: Data Analysis of Iron

IX LIMITATIONS

Since computation is used to analyse large amounts of digitised data, there are some limitations to the obtained results.

Firstly, the data was digitised using OCR, which does not yield a perfect result. Indeed, especially in the case of the *Neue Hamburger Zeitung* and *El Imparcial*, the quality of the OCR strongly limits what can be achieved with the data. The newspaper articles are also incomplete: they do not span the entire time frame evenly. This leads to gaps in our analysis for certain newspapers.

Secondly, most of the newspapers did not digitise their data article-by-article, but rather page-by-page or even issue-by-issue. An experimental pipeline was created to try and split these documents correctly, but it is not perfect. Since different articles touch vastly different topics, not separating them correctly can confuse the topic modelling later-on.

Thirdly, the chosen technologies do not have the same lifetime: the telephone was only invented in the late 1870's, whilst gasoline and iron were being used since the beginning of the corpus (1850 and 1830 respectively). This can make comparison between results more difficult.

Fourthly, our method presents some limitations when it comes to analyzing technologically related rhetorics. While we do believe that our corpus is a good medium of analysis wherein to observe how different countries encode technologically related discourses, our method discriminates a considerable amount of relevant information, as there are several instances in our pipeline (such as step three Ward Hierarchical Clustering, and step four directed graph) that are specifically designed to only select words with the highest rates of semantic similarity across countries (and by words we mean the English translations of the multilingual words belonging to the Pachinko Allocation Topic Modelling Algorithm that we implemented in step one of our pipeline). That means that, while we are able to find a significant amount of shared discourses across newspapers (therefore clearly indicating how these technologies had a similar social impact across countries), our method is not designed to grasp cultured minoritarian discourses. We believe that the fact that the majority of the overlapping topics across papers belong to the semantic fields of politics, finance, or entertainment, is a clear side effect of our method's inner-working (as those tend to be common-place themes frequently reported by newspapers across countries). Yet the high homogeneity of our data analysis in these realms confirms our hypothesis regarding the agency of technology as a trigger of information homogeneity.

Fifthly, it is also relevant to mention that newspapers are commodified objects that are designed to be sold for profit. While newspapers of record tend to report relevant historic events independently of their respective countries of origin, they do also provide some format-specific information about themes such as finance and stock markets, local and international politics, crime, or weather forecast. In summary: they report a very specific segment of reality that shapes the way in which they frame information (and more concretely, technologically related discourses). However, and as our data analysis shows, the amount of overlapping information across all publications is relevantly high, once again showcasing the social impact of technology as a catalyst of cultural flattening.

Finally, and as mentioned, our corpus has some relevant data gaps, inevitably impacting our data analysis and thesis formulation.

X CLOSE READING OF NEWSPAPER SECTIONS AND SELECTED ARTICLES

To have a better understanding of our dataset and to what extent our big data computational analysis is a suitable alternative (or complement) of traditional hermeneutics, we have decided to implement a mix-methods approach by incorporating a close reading analysis of, firstly, a randomized selected number of newspapers where we closely inspect different sections (i.e. finance, politics, entertainment...), and secondly, a random sample of articles containing our target key terms.

10.1 Newspapers sections

We are curious to analyze whether our selection of historic multilingual newspapers are comparable objects of analysis among themselves as we initially suspected that different cultural media traditions were more nuanced during our observational time than nowadays. We are specifically interested to observe and contrast newspapers sections (i.e. finance, politics, sports, science, entertainment...) in our dataset, to ensure that our medium of study (multilingual historical newspapers) is relevant for our research question. Thus we have selected one random newspaper per decade for each newspaper during our observational time, specifically focusing on doing a close reading on each newspaper sections and formatting (i.e. article length or tone of the news). We have used the PDF version of each newspaper available in different online repositories ((The Bibliothèque nationale de France (BNF) for *Le Figaro Bibliothèque Nationale de France*, https://gallica.bnf.fr/ark:/12148/cb34355551z/date, The Europeana Project for *Neue Hamburger Zeitung The Europeana Project*, https://www.europeana.eu/en/collections/organisation/1482250000005967007-state-and-university-library-hamburg-carl-von-ossietzky?page=1&qa=proxy_dc_title%3Aneue%5C%20hamburger%5C%20zeitung, the *Archivio Storico de La Stampa Archivio Storico de La Stampa*, ttp://www.archiviolastampa.it/, the Chronicling America Historical Newspapers Collection hosted by the Library of Congress *Chronicling America Historical Newspapers Collection, hosted by Library of Congress*, https://chroniclingamerica.loc.gov/search/pages/results/?state=&date1=1770&date2=1963&proxtext=new+york+herald&x=0&y=0&dateFilterType=yearRange&rows=20&searchType=basic, and the *Hemeroteca Digital* website for *El Imparcial Hemeroteca Digital*, https://hemerotecadigital.bne.es/hd/es/card?sid=188939 as we seek to inspect newspapers from a material culture perspective, and being able to observe the original formatting of our sources is much more relevant than reading a text data mining version (i.e. txt files) where some visual aspects of the original outlet are harder to analyze. We indicate each newspaper consulted, including its corresponding url, in an appendix.

We observe a certain degree of heterogeneity among newspapers, yet quite contained, as we note how newspapers are quite homogeneous commodified objects overall. *Le Figaro* and *The New York Herald* are the two news outlets that present the biggest differences. *The New York Herald* holds a typical tabloid format: a high degree of condensed information into one page, not clear sections and a significant number of sensationalized news (such as many gruesome details on news about local crime or an emphasis on political corruption when talking about local politicians). We note a progressive streaming of sections as we approach contemporary times (for example, it is possible to notice entire pages dedicated to sports or news after the 1900s), but still quite far away from contemporary notions of newspapers sections. While *The New York Herald* reports international and national news (which is why it is considered as a newspaper of record), it shows a very strong local flavour by containing an uncommonly high number of State of New York contemporary references (advertising, job posts, lost-and-found

notices, real state (seek and wanted) offers, etc). *Le Figaro*, on the other hand, has a very strong literary magazine personality. While there is a very clear evolution towards a tabloid format around the 1890s decade, it still holds certain elements such as article longer length (each article is almost an opinion peace versus the telegraphic tabloid style format of *The New York Herald*), the presence of chapters of novels, a clear interest on reporting the artistic and literary scene, and overall a more intellectual focus. That being said, we observe a very noticeable change around the 1940s (and more visibly in the 1950s) towards a tabloid-style outlet.

The Neue Hamburger Zeitung, *La Stampa*, and *El Imparcial* are very similar and can be found right in the middle between *Le Figaro* and *The New York Herald*. They are slightly more local and less intellectual than *Le Figaro*, yet they consistently include chapters from novels and have longer article length (not as much as *Le Figaro*, but certainly longer than the typical tabloid-style shorter format). At the same time, they report more consistently international news within the realm of the political interests of each respective countries. For example, in the case of *La Stampa*, there is a very big emphasis in reporting news about the Italian colonial empire in Africa or about Italy's foreign policy in countries from the Balkan Region (more evidently so beginning in the 1920s and onwards). The *Neue Hamburger Zeitung*, on the other hand, focuses more on news about the German colonial empire and German foreign policy. Anecdotally the *Neue Hamburger Zeitung* has two editions: a morning (*Morgen Ansgabe*) and an evening one (*Abend Ansgabe*). We notice how the morning edition is a more traditional newspaper (less number of pages, longer articles, more focused on news) and the evening edition is a tabloid-style newspaper (higher number of pages, shorter articles, more advertising). We do observe a trend in these three newspapers towards increasingly becoming a tabloid style newspaper as we approach contemporary times, yet it is a very subtle change (specially in the case of the *Neue Hamburger Zeitung*). *El Imparcial* shows a stronger tendency to report news about Spain and Latin America (and occasionally North Africa) rather than international news, although as the newspaper approaches the twentieth century these become more present.

We also notice how some newspapers have a tendency to report high amounts of local content. For example, *The New York Herald* and *Neue Hamburger Zeitung* (both sea ports) include a high amount of maritime information: ship tours advertising, information about ship lines, names of passengers lists... However, other newspapers such as *Le Figaro* or *El Imparcial* (that, for example, advertise ship lines transporting mail overseas) also include this information in some decades, showing how ships were one of the most important transportation systems of the time.

It is difficult to make strong assertions about newspapers trends in content or sections as we observe a high degree of fluidity over time. For example, we notice seasonal variations in the reported content of the newspapers (i.e. we observe winter-specific news about ski seasons or winter resorts, and the same applies to society pages during the summer). As indicated, historic newspapers during our observational time did not have contemporary media standards regarding sections, and therefore, it is possible to observe random information condensed in every page across all our newspapers (local, national, international, and random advertising).

That being said, we do note an increasing polarization of information coinciding with the first and second World Wars, as well as the interwar period, across our selection of newspapers. While there is a tendency to observe an standardization in terms of form (all newspapers gravitate towards a more nuanced tabloid-format style starting at the turn of the century and more clearly after the 1920s), there is a clear tendency towards reporting news reflecting each country's political interests in a more aggressive way than in previous decades. For example, we

see many news in *Neue Hamburger Zeitung* in the interwar period addressing the topic of the German economic crisis in the 1920s (Hyperinflation in the Weimar Republic). We also notice how, during the First World War, newspapers content varies significantly across publications due to different political interests.

Consequently, while we note some degree of structural heterogeneity across our newspapers, those differences balance each other. As mentioned, some newspapers have a clear tabloid style (*The New York Herald*) others a more literary magazine one (*Le Figaro*), and some others are in between (*Neue Hamburger Zeitung, La Stampa, El Imparcial*). While some newspapers have a very clear local component by putting an emphasis on local advertising, crime, or politics (i.e. *The New York Herald*), almost all our newspapers include at some point this kind of information (local real state, lost-and-found news, job posts...). As mentioned, ships were one of the preferred means of transportation (human and commercial) during our observational time. Consequently, while there is a higher number of ship-related news in newspapers located at sea ports (*Neue Hamburger Zeitung* and *The New York Times*), all newspapers do report some news about ships. Similarly, news about religion (i.e. sermons, or religious related news) also appear more steadily in some publications (*The New York Herald*, *La Stampa*) than in others, yet they appear at some point in the majority of our newspapers.

Therefore, this qualitative approach to the analysis of newspapers formatting emphasizes the relevance of our data analysis: it complements qualitatively our quantitative observations (a relatively stable scenario of information homogeneity across press outlets) that notes how newspapers are not radically different publications across countries over our observational time. Consequently, it is quite logic that media frames about technologically related terms in relatively akin newspapers are similar. However, we do observe a significant polarization of newspapers content during the first four decades of the twentieth century coinciding with the first and second Word Wars as well as the interwar period. Yet our data analysis outputs during this time-frame are quite flat. This confirms our hypothesis about the power of technology as a trigger of cultural flattening across Western countries: technologically related media frames have remained relatively stable during the most turbulent historical periods in recent history.

10.2 Individual Articles

We are also interested in doing a close reading of a randomized selection of articles [4] containing our key terms (telephone, gasoline, iron), to assess qualitatively media frames of technology and the way in which those created an impact in society (we include in an appendix our selection of articles and we also include the original text in an online repository).

Generally speaking, we observe that newspaper articles containing technologically related terms for the most part report the social use of those technologies, and we observe three different instances of this phenomenon. Firstly, we note how the social use of technology is typically embedded in a larger narrative about some other event. For example, in the case of iron, we see how *Le Figaro* reports a trip done by a European aristocrat (Carlos de Borbón y Austria-Este) in

4. Articles used from Le Figaro : figaro1, figaro2, figaro3, figaro4, figaro5, figaro6, figaro7, figaro8, figaro9, figaro10, figaro11
Articles used from the New York Herald : nyh1, nyh2, nyh3, nyh4, nyh5, nyh6, nyh7, nyh8
Articles used from La Stampa : stampa1, stampa2, stampa3, stampa4, stampa5, stampa6, stampa7, stampa8, stampa9, stampa10
Articles used from Neue Hamburger Zeitung : nhz1, nhz2, nhz3, nhz4
Articles used from El Imparcial : ei1, ei2, ei3, ei4, ei5, ei6, ei7

1885 to the Himalayas (that is the main topic of the article). To reach the peak of the mountain, passengers must use an iron-made railway that is considered one of the wonders of the world. Another example would be how *La Stampa* mentions in 1920 how, as a result of a post-office strike (that being the article's main theme), telephone services were suspended creating major social chaos. Similarly, the *Neue Hamburger Zeitung* reports how in 1921 a storage unit got on fire, yet not reaching the gasoline tank.

In a second category we note how these technologies appear in advertising. For example, in the case of telephone, we see how *Le Figaro* includes a telephone number in a real state company advert.

Finally, in a third group, we observe scientific discourses about the technologies in themselves. For example, *La Stampa* provides quite a long article about scientific advances about gasoline in a car industry context in 1928.

We wrap up this mix-methods approach with the idea that, as mentioned in the limitations section, our pipeline is quite efficient in empirically assessing semantic similarity rates in multilingual corpora (the main thesis of this paper), yet inevitably, it leaves out some nuances that a qualitative analysis would reveal. It would be highly interesting to further decompose quantitatively our dataset by using other methodologies such as word embeddings seeking to test how our three key terms (gasoline, telephone, iron) appear represented across different newspapers in order to gain a higher understanding of cultural differences (instead of similarities) of the social use of these technologies, and complement this analysis with a qualitative assessment. Similarly, it would also be relevant to use word embeddings on different thematic areas (i.e. finance, sports, entertainment) to have a deeper sense of how media frames encode our key selected terms similarly or differently.

Yet the fact that, again, our data analysis outcomes displays such a high degree of homogeneity reveals a very similar encoding of technology across Western societies during our observational time necessarily at the article level. This resonates with our qualitative findings about a relatively similar degree of homogeneity in newspapers formatting as well as a high homogeneous content of semantic context (i.e. advertising) where our technological related terms appear encoded in a similar way. Consequently, we believe that our mix-methods approach reinforces our quantitative analysis data outcomes and our observations about the agency of technology as a trigger of cultural flattening in media discourses.

XI CONCLUSION

Our data analysis shows three main findings that appear consistently across our three selected key terms (telephone, gasoline, iron): firstly, we detect a trend of information simplification in the 1890 and 1900 decades, coinciding with the peak of the Second Industrial Revolution in Western Societies. Secondly, we observe a pattern of semantic fragmentation in the first two decades of the twentieth century (1910 and 1920), and in the case of iron, also in 1930, concurring with the First and Second World Wars, as well as the interwar period. That being said, we also observe higher rates of agreement (and by that we mean more topics with either five-out-of-five or four-out-of-five newspaper representation) also during this time frame (1910-1930) in both gasoline and iron, and to a certain extent, also in telephone. Therefore, we notice an increasing amount of information complexity, but not necessarily of geographic polarity, which is quite a significant discovery considering that, during these three decades, our selection of countries aligned with highly different political views during the First and Second World Wars

as well as during the interwar period. Thirdly, we notice a remarkable semantic homogeneity across our three key terms, as there are six semantic categories that consistently appear in all of them (crime (green), politics (grey), real estate (lilac), ship industry (orange), finance (blue), and entertainment (lime green)). Moreover, in the case of iron and gasoline, they also share the majority of the extra semantic categories (drug industry (red), and transport (fuchsia)), as well as their coverage of major military conflicts, which is nearly identical. Indeed, the only independent topics that are found in isolation are gold rush in 1910 in the case of gasoline, and, in the case of iron, religion (1880), and labour (1890 and 1920). The topic of education is found both in telephone (1910) and in iron (1930). As mentioned, our qualitative analysis reinforces this data-driven findings: we do observe a high degree of newspapers formatting similarity over time yet high polarization in content during the first four decades of the twentieth century. Similarly, we note some patterns in how technologies are encoded at the article level upon a close reading of a randomized sample (news reporting the social use of technologies in three different semantic scenarios: as part of a longer narration about some other event, mentions about the use of the technology in itself (i.e. telephone numbers) and as science-related content).

These results make us speculate with the idea that the social impact of technology during our observational time was reflected in press discourses quite homogeneously across different Eurocentric societies permeating in the realms of reality where these technologies imprinted lasting effects (i.e. politics, finance, entertainment, real estate, and ship industry, shared by all key terms; as well as drug industry, transport, and education, shared by two-out-of-three key terms). Moreover, we observe a consistent interference of some elements of the contemporary social fabric at the time (religion and aristocracy) across different topics and semantic categories, quite similarly across our three key terms.

We therefore conclude with the idea that, while multilingual newspapers during the nineteenth and early twentieth century were slightly more heterogeneous than contemporary newspapers of record (therefore reflecting the more pronounced cultural differences across Western societies two hundred years ago), our mix-methods approach shows the transnational social cohesion that the Second Industrial Revolution imprinted in Western societies as recorded by newspapers (that, as mentioned, only report a very specific segment of reality). These findings resonate with existing work on the social impact of technology by Fernández Fernández and Savcisens Fernández Fernández and Savcisens, "A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)" in contemporary newspapers, where the authors show how it is possible to observe a shared Western voice in sustainability related discourses over the last twenty years (1999-2018) that becomes more streamed as we reach contemporary times. This paper complements their findings by illustrating how that voice can be traced back to the early stages of the birth of the modern nation in the mid-nineteenth century, as well as the agency of technology as a force of identity cohesion in press discourses, capable of neutralizing the polarizing effects of two World Wars and an interwar period.

Future lines of research include filtering our dataset with other key technological terms as well as different subjects such as politics, finance, or real estate, seeking to test whether content homogeneity was also present in those semantic realms or if it was an exclusive phenomenon of technological related discourses. We are also interested in adding more Western newspapers from underrepresented regions in our study (i.e. Scandinavian, central, and low-land countries, Australia and Canada) as they all hold open source digitized archives of their major newspapers of record during our observational time. We are also curious in experimenting with other

methodologies (i.e. word embeddings or sentiment analysis) to further decompose tensions between homogeneity or heterogeneity of information across different news outlets and diverse semantic fields.

We have created a repository (https://zenodo.org/records/10657581) where we include scripts, metadata, and the original randomized articles used for qualitative assessment matching this paper. We therefore follow FAIR standards.

XII ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie (MSC) grant agreement No 101024996. The authors would like to thank reviewers for their very useful feedback. The authors would also like to thank Prof. Jerome Baudry and Prof. Julia Flanders for their guidance and useful advice. We would like to thank Thomas Benchetrit for his excellent work on historic newspapers OCR, carried out during his Master Project in the Spring Semester 2022. The authors would like to express as well their gratitude to Clemens Neudecker for his help in the process of locating and acquiring the Multilingual Historic Newspapers datasets, as well as to all the authors of the article "The Corpora They Are a-Changing: a Case Study in Italian Newspapers" (https://aclanthology.org/2021.lchange-1.3/), for their generous gesture of sharing with us a very well curated dataset (and otherwise unavailable) of *La Stampa*. Additionally, the authors would like to thank very much the Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky personnel their help with the *Neue Hamburger Zeitung* related questions, and both the Zurich Zentral Library and the University of Zurich Library personnel for all their help and support.

REFERENCES

Aletras, Nikolaos, and Mark Stevenson. "Measuring the similarity between automatically generated topics." In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, 22–27. 2014.

Anderson, Benedict. *Imagined Communities*. London: Verso, 2006.

Archivio Storico de La Stampa. ttp://www.archiviolastampa.it/.

bab2min. *Tomotopy*. https://github.com/bab2min/tomotopy, 2019.

Bench, Thomas. *anxietyNews*. https://github.com/ThomasBench/anxietyNews, 2022.

Beykikhoshk, Adham, Ognjen Arandjelović, Dinh Phung, and Svetha Venkatesh. "Discovering topic structures of a temporally evolving document corpus." *Knowledge and Information Systems* 55 (2018): 599–632.

Bibliothèque Nationale de France. https://gallica.bnf.fr/ark:/12148/cb34355551z/date.

Cantoni, Eva, Chris Field, J Mills Flemming, and Elvezio Ronchetti. "Longitudinal variable selection by cross-validation in the case of many covariates." *Statistics in medicine* 26, no. 4 (2007): 919–930.

Casatells, Manuel. *The Rise of the Network Society*. Chichester, West Sussex: Wiley-Blackwell, 2010.

Chronicling America Historical Newspapers Collection, hosted by Library of Congress. https://chroniclingamerica.loc.gov/search/pages/results/?state=&date1=1770&date2=1963&proxtext=new+york+herald&x=0&y=0&dateFilterType=yearRange&rows=20&searchType=basic.

Cipolla, Carlo M. *The Fontana Economic History of Europe*. Glasgow: Fontana/Collins, 1973.

Crouthamel, James L. *Bennett's New York Herald and the Rise of the Popular Press,* New York: Syracuse University Press, 1989.

Fernández Fernández, Elena, and Germans Savcisens. "A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)." In *Digital Humanities in the Nordic and Baltic Countries 2023. Sustainability, Environment, Community, Data. Online Conference*. CEUR-WS, 2023. https://doi.org/10.5281/zenodo.7742461.

Großwendt, Anna, Heiko Röglin, and Melanie Schmidt. "Analysis of ward's method." In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2939–2957. SIAM, 2019.

Hemeroteca Digital. https://hemerotecadigital.bne.es/hd/es/card?sid=188939.

Herman, Edward S., and Noam Chomsky. *Manufacturing Consent. The Political Economy of the Mass Media*. New York: Pantheon Books, 2002.

Hobsbawm, Eric. *The Age of Capital (1848-1875)*. London: Abacus, 1995.

Honnibal, Matthew, and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." To appear, 2017.

King, Steven, and Geoffrey Timmis. *Making Sense of the Industrial Revolution*. Manchester: Manchester University Press, 2001.

Kuhn, Raymond. *The Media in Contemporary France*. Glasgow: Open University Press, 2011.

Li, Wei, David Blei, and Andrew McCallum. "Nonparametric bayes pachinko allocation." *arXiv preprint arXiv:1206.5270*, 2012.

Li, Wei, and Andrew McCallum. "Pachinko allocation: DAG-structured mixture models of topic correlations." In *Proceedings of the 23rd international conference on Machine learning*, 577–584. 2006.

Magro, Angel Bahamonde, and Luis Enrique Otero Carvajal, eds. *La sociedad madrileña durante la Restauración (1876-1931). Volumen II. El sistema político de la Restauración. El horizonte cultural. Opinión y medios de información. Conflicto social y clases trabajadoras*. Madrid: Graimo SA, 1989.

Matteo Gomellini, Gianni Toniolo. "The Industrialization of Italy, 1861–1971." Edited by Jeffrey Gale Williamson Kevin Hjortshøj O'Rourke, 2017.

Michelet, Elisa. "An Industrial West? Analyzing Multilingual Newspapers Discourses about Technology during the Second Industrial Revolution (1840-1930)." *EPFL*, 2023.

Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin. "Automatic evaluation of topic coherence." In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 100–108. 2010.

Robertson, Robbie. *The Three Waves of Globalization. A History of a Developing Global Consciousness*. London: Zed Books, 2004.

Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics* 20 (1987): 53–65.

Saitta, Eugénie. "The Transformations of Traditional Mass Media Involvement in the Political and Electoral Process. A Case Study on Political Journalism in France and Italy since the 80s." *ECPR Joint Session. Workshop 9: 'Competitors to Parties in Electoral Politics: The Rise of Non-party Actors'.*, 2006.

Savcisens, Germans. *Simple Topic Modelling Examples*. https://github.com/carlomarxdk/topic_modelling, 2022.

Smil, Vaclav. *Creating the Twentieth Century: Technical Innovations of 1867-1914 and Their Lasting Impact*. Madrid: Oxford University Press, 2005.

Stearns, Peter. *The Industrial Revolution in World History*. New York: Taylor Francis, 2021.

The Europeana Project. https://www.europeana.eu/en/collections/organisation/1482250000005967007-state-and-university-library-hamburg-carl-von-ossietzky?page=1&qa=proxy_dc_title%3Aneue%5C%20hamburger%5C%20zeitung.

Tornier, Klaus. *Hamburger Pressegeschichte in Zeitungstiteln vom 17 . bis 20. Jahrhundert*. Norderstedt: Books on Demand, 2021.

Veblen, Thorstein. *Imperial Germany and the Industrial Revolution*. Ann Arbor: The University of Michigan Press, 1966.

Vilar, Juan Bautista. *La primera revolución industrial española (1827-1869)*. Madrid: Ediciones Itsmo, 1990.

Wajcman, Judy. *Pressed for Time. The Acceleration of Life in Digital Capitalism*. Chicago: University of Chicago Press, 2014.