An Industrial West? A Quantitative Analysis of Newspapers Discourses about Technology over Ninety Years (1830-1940)

Emmanuelle Denove[1], Elisa Michelet[2], Germans Savcisens[3], Elena Fernández Fernández[4].
[1 2] Department of Computer Science, EPFL, Rte Cantonale, 1015 Lausanne, Switzerland
[3]Section for Cognitive Systems, Technical University of Denmark, Building 324, Kongens Lyngby 2800, Denmark
[4]Institute of Computational Linguistics, University of Zurich, University of Zurich, Andreas Strasse 15, Zurich, 8050, Switzerland.
{emmanuelle.denove, elisa.michelet}@epfl.ch, gersa@dtu.dk, elena.fernandezfernandez@uzh.ch

***Abstract*—Recent work analyzing the social impact of technology in processes of globalization signals a shared Western voice in technologically related discourses dating back at least twenty years. However, many scholars propose the idea that, as a direct consequence of the Second Industrial Revolution, globalization processes can be traced back at least to the second half of the nineteenth century. Only a few decades later, nevertheless, two of the most divisive historic events ever in human history took place: the First and Second World Wars. In this article we seek to explore information behaviour during one hundred years approximately (1830-1940), using multilingual newspapers of record as a proxy (*Le Figaro, The New York Herald, El Imparcial, Neuer Hamburger Zeitung* and *La Stampa*), to observe to what extent technology acted as a cohesive force across Western societies walking along these different historic happenings. Thus we filter our corpus with three key technological terms (telephone, gasoline, and iron) as an exploratory endeavour. 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. Our data analysis reveals three main findings: firstly, we empirically 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 the whole observational time, displaying almost identical wording. These findings make us speculate with the idea that it is possible to trace back a shared Western voice in technological related discourses back to two hundred years ago.**

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 [1]). Placed by some authors during the last decades of the twentieth century (Castells [2], Wajcman [3]) 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 et al. [4]). 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 [1]):

> 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 [5]), economic (Cipolla [6]), and geographic perspectives (King and Timmis [7]). 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 and Timmis [7]): "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 during ninety years (1830-1940) in terms of information behaviour in order to test its cohesive agency across time and space. 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 [5]: "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 [5] and Robertson [8] 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 [5]):

> 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)

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.

Afterwards we follow the five-step pipeline proposed by Fernández Fernández and Savcisens ([9]), consisting of firstly Topic Modelling (Pachinko Allocation), secondly translation into English of the multilingual words outputted by the topics, thirdly word embeddings, fourthly Ward Hierarchical Clustering, and finally, building a directed graph; that allows us to track rates of homogeneity versus heterogeneity of information across countries as well as their temporal evolution. In dialogue with existing research of the art that claims the existence of growing patterns of information agreement in contemporary press discourses about technology, as well as the existence of a clearly streamed Western voice that can be traced back up to twenty years ago (again, Fernández Fernández and Savcisens ([9])), we seek to analyze shifts of information polarity or agreement in Eurocentric societies over a period of ninety years, aiming to detect whether a common shared Western voice in technological related discourses can be traced back to those foundational times.

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 [5]), geography (King and Timmis [7]), or economic history (Cipolla [6]). Furthermore, the historic specificities of different countries during this period of time have been as well extensively analyzed by scholarly critique (Vilar [10], Veblen [11], Gomellini and Toniolo [12]; 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. [13] 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. [14] 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. [15], 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.

Furthermore, and as mentioned, we dialogue with Fernández Fernández and Savicens [9]. The authors use a dataset of multilingual newspapers (English, French, German, Italian, and Spanish), over a time-spam 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 by using two hundred extra years of data (1830-1940 to complement their 1999-2018 observational time) aiming to deep 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*), predominantly politically neutral, and for the most part newspapers of record during our observational time.

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

TABLE I
CORPUS AVAILABILITY AND SOURCES

Journal	Time Period	Source
Le Figaro	1860-1920	Library of Congress
The New York Herald	1840-1888, 1920	Gallica
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.

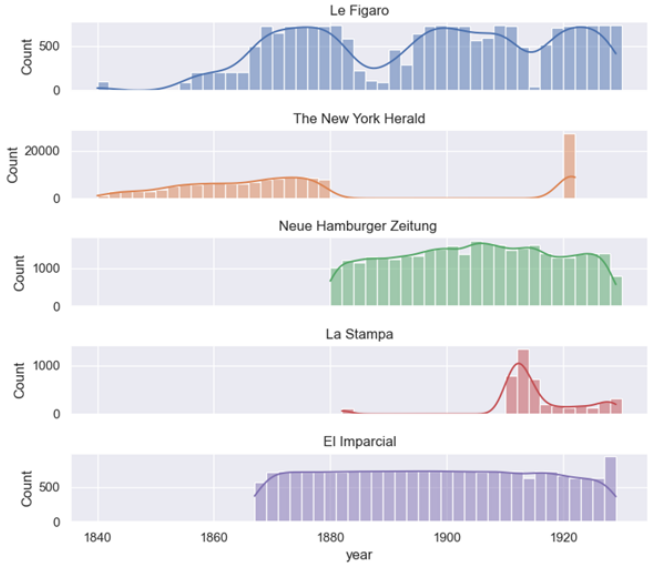

Fig. 1. Newspapers Data Availability

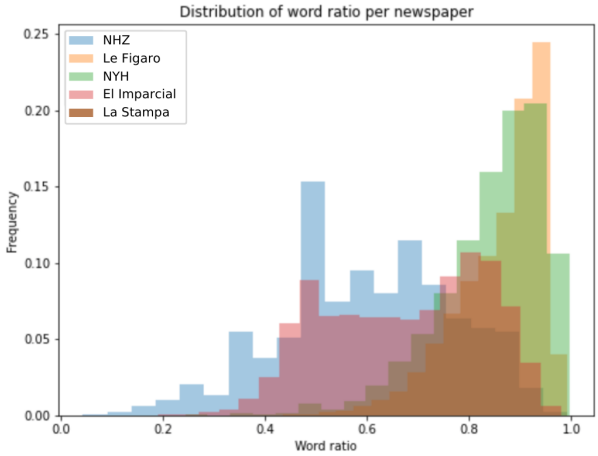

Fig. 2. OCR quality of newspapers

achieve this, Thomas Benchetrit's OCR-qualification pipeline [16] 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 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.

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.

A. 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 [17].

B. 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 $k = \frac{2}{3}$.

C. The New York Herald

This dataset provided the .xml documents from the OCR as well as their textual interpretation. The XML files contain

TABLE II
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 "

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.

VII. METHODS

A. 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 determine how these subjects were discussed in the newspapers in our corpus, we firstly use topic modeling.

1) Pachinko Allocation Model: The Pachinko Allocation Model (PAM) [18] 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 [19], a grid search is performed to obtain the optimal parameters. This process is detailed below.

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 [20]. Specifically, the chosen metric is c_V-coherence because its results are the closest to a human interpretation [17]. 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 [21]. 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.

B. 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 [22] 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 VII-A2
- using Tomotopy's Pachinko Allocation training model [23] to train the given model following the pipeline proposed by [24] with the previously determined parameters

Finally, each given model was represented as the set of its $k = 15$ most representative words as determined by the topic modeling.

C. 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 epoch, dialoguing with Fernández Fernández and Savcisens ([9]).

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.

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.

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 epoch 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
$\tilde{\mathcal{V}} = \text{stack}(\boldsymbol{\nu}_t \; \forall \; t \in \; \{1, 2..t\})$

Second, as the topic similarity measure is based on word embeddings ([25]), we need to know all the words across every newspaper and epoch – we create a global vocabulary set, V. Global vocabulary consists of all the unique English words across every epoch 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 $\mathbf{\Gamma} \in R^{p \times v}$, where p is the total number of topic vectors (across every epoch and newspaper), and v is the length of the global vocabulary. $\mathbf{\Gamma}$ contains the probabilities of each word in every discovered inner topic.

Using matrix $\mathbf{\Gamma}$, we can extract word embeddings, i.e., numerical representations of words ([25]). 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 $\mathbf{\Gamma}$).

We calculate the topic similarity based on the *average pairwise cosine similarity* of the N-top[1] words in each topic ([25]).

D. Global Topic Aggregation

To analyze whether newspapers share discussion points over a specific epoch, 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 epoch). We cluster inner topics with the use of the Hierarchical Cluster Analysis (HCA) with Ward's linkage function ([26]).

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 ([27]) as a proxy for the clustering quality. This method estimates whether topics are closer to the members of their own clusters or vice versa. For each epoch, 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 $\mathbf{\Gamma}$).

E. 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 epochs, t and $t+1$ (i.e., between topics of different epochs). 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 epochs if their similarity is above a certain threshold, t. beykikhoshk2018discovering 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) epochs. 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

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

[2] we decided on the quantile by manually inspecting graphs

[3] we decided on the threshold by manually inspecting the results

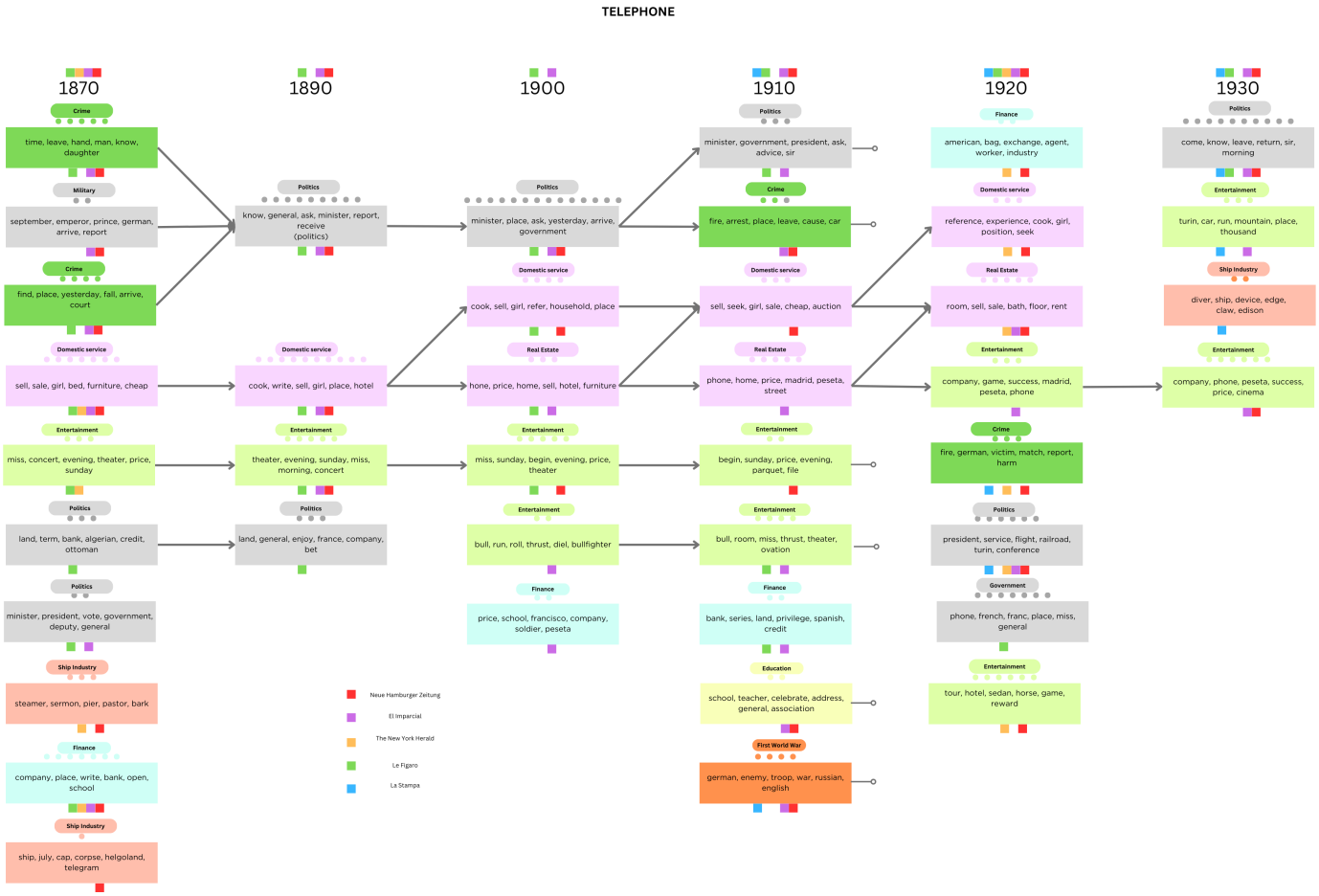

Fig. 3. Data Analysis of Telephone

ideas are refined or transformed ([28]). The birth of a topic is characterized by the absence of the incoming arrows - no topic from the previous epoch 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 outcoming 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.

VIII. Results and Discussion

After applying our five-step pipeline, and in dialogue with Fernández Fernández and Savcisens [9] 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. 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) may have any effect on trends of information affinity. .

A. *Telephone*

Figure 3 shows the historical evolution of discourses about the key term Telephone. Our observational time covers 1870-1930, as those are the years where we could find documents containing the word telephone.

Following the same method as Fernández Fernández and Savcisens [9], 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

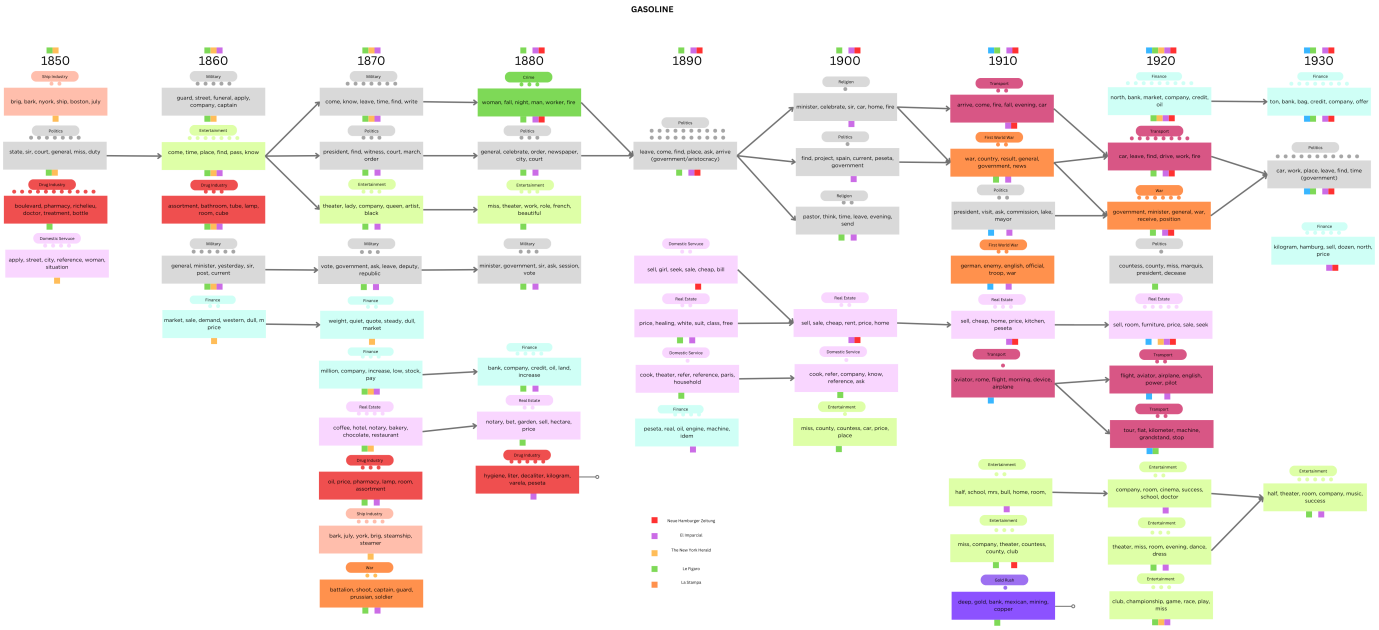

Fig. 4. Data Analysis of Gasoline

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 [9]), in the epoch 1870-1890, 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 seven different semantic categories that we showcase by colouring the boxes: politics (grey), crime (green), real estate (lilac), entertainment (lime green), ship industry (orange), crime (green), and finance (blue). In this epoch, 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) across our selection of countries.

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 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

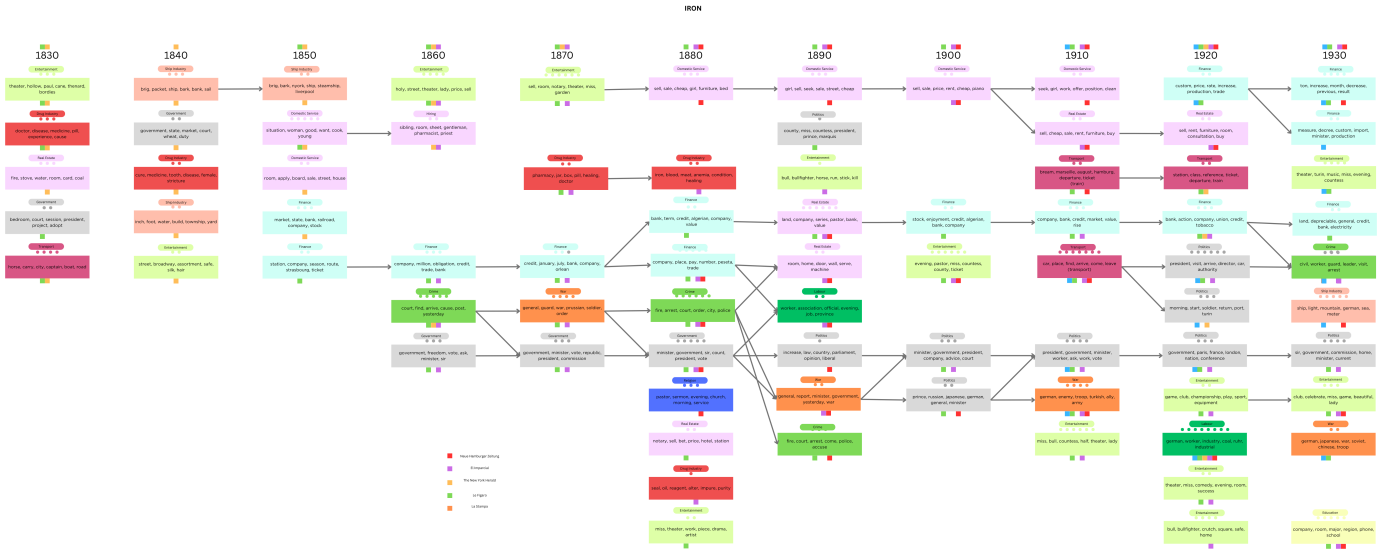

Fig. 5. Data Analysis of Iron

central semantic categories that are constantly present across our analysis, yet the time distribution patterns contradict the results found by Fernández Fernández and Savcisens [9] in contemporary datasets. In there, they noticed a tendency of low-clustering in the first epochs of their study (1999-2003, 2004-2008), followed by an increasing presence of newspapers representation towards the end (2009-2013, 2014-2018).

B. 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.

C. 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 and 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).

D. 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.

Finally, 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.

IX. 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 higher rates 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).

These results make us speculate with the idea that, while historic multilingual newspapers during the nineteenth and early twentieth centuries were quite diverse both in terms of content and form, the social impact of technology during that 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 more heterogeneous than contemporary newspapers of record (therefore reflecting the more pronounced cultural differences across Western societies two hundred years ago), our exploratory data analysis of three technological terms (telephone, gasoline, and iron) shows the transnational social cohesion that the Second Industrial Revolution imprinted in Western societies. These findings resonate with existing work on the social impact of technology by Fernández Fernández and Savcisens ([9]) in contemporary newspapers, where the authors, as already mentioned, 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, 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.

X. 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 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*.

References

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

[2] M. Casatells, *The Rise of the Network Society*. Chichester, West Sussex: Wiley-Blackwell, 2010.

[3] J. Wajcman, *Pressed for Time. The Acceleration of Life in Digital Capitalism*. Chicago: University of Chicago Press, 2014.

[4] A. S. M. T. Bryan Kelly, Dimitris Papanikolaou, "Measuring technological innovation over the long run," *American Economic Review: Insights*, vol. 3, no. 3, pp. 303–320, 2021.

[5] P. Stearns, *The Industrial Revolution in World History*. New York: Taylor Francis, 2021.

[6] C. M. Cipolla, *The Fontana Economic History of Europe*. Glasgow: Fontana/Collins, 1973.

[7] S. King and G. Timmis, *Making Sense of the Industrial Revolution*. Manchester: Manchester University Press, 2001.

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

[9] E. Fernández Fernández and G. 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.

[10] J. B. Vilar, *La primera revolución industrial española (1827-1869)*. Madrid: Ediciones Itsmo, 1990.

[11] T. Veblen, *Imperial Germany and the Industrial Revolution*. Ann Arbor: The University of Michigan Press, 1966.

[12] G. T. Matteo Gomellini, "The industrialization of italy, 1861–1971," 2017.

[13] A. K. D. W. K. H. D. S. Mariona Coll Ardanuy, Katherine McDonough, "Resolving places, past and present: toponym resolution in historical british newspapers using multiple resources," *GIR '19: Proceedings of the 13th Workshop on Geographic Information Retrieval.*, pp. 1–6, 2019.

[14] D. B. M. C. A. E. G. J. H. J. L. K. M. M. R. G. T. D. V. S. D. C. W. Kaspar Beelen, Ruth Ahnert, "Living with machines: Exploring bias in the british newspaper archive," 2019.

[15] K. H. B. M. G. C. Kaspar Beleen, Mariona Coll Ardanuy, "Assessing the impact of ocr quality on downstream nlp tasks," *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pp. 484–496, 2020.

[16] T. Bench, "anxietynews," https://github.com/ThomasBench/anxietyNews, 2022.

[17] E. Michelet, "An industrial west? analyzing multilingual newspapers discourses about technology during the second industrial revolution (1840-1930)," *EPFL*, 2023.

[18] W. Li and A. McCallum, "Pachinko allocation: Dag-structured mixture models of topic correlations," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 577–584.

[19] W. Li, D. Blei, and A. McCallum, "Nonparametric bayes pachinko allocation," *arXiv preprint arXiv:1206.5270*, 2012.

[20] D. Newman, J. H. Lau, K. Grieser, and T. 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*, 2010, pp. 100–108.

[21] E. Cantoni, C. Field, J. Mills Flemming, and E. Ronchetti, "Longitudinal variable selection by cross-validation in the case of many covariates," *Statistics in medicine*, vol. 26, no. 4, pp. 919–930, 2007.

[22] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

[23] bab2min, "Tomotopy," https://github.com/bab2min/tomotopy, 2019.

[24] G. Savcisens, "Simple topic modelling examples," https://github.com/carlomarxdk/topic_modelling, 2022.

[25] N. Aletras and M. 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, pp. 22–27.

[26] A. Großwendt, H. Röglin, and M. Schmidt, "Analysis of ward's method," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2019, pp. 2939–2957.

[27] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[28] A. Beykikhoshk, O. Arandjelović, D. Phung, and S. Venkatesh, "Discovering topic structures of a temporally evolving document corpus," *Knowledge and Information Systems*, vol. 55, pp. 599–632, 2018.