



---

**An Industrial West? Analyzing  
Multilingual Newspapers Discourses  
about Technology during the Second  
Industrial Revolution (1840-1930)**

---

**Elisa Michelet**

Master Project

at the Laboratory for the History of Science and Technology  
for the completion of the academic degree

Master of Science  
(*M. Sc.*)

in Digital Humanities

submitted on the 20. January 2023 to the  
College of Humanities Faculty from  
Ecole Polytechnique Fédérale de Lausanne

**Professor**

Jérôme Baudry

**Supervisor**

Elena Fernandez Fernandez



# Abstract

---

*While the idea of technology as a driving force of history and society has been extensively studied in the history of the Industrial Revolutions, little attention has been paid to the social perception of those technologies. This thesis focuses on the discourses about technology observed in newspapers contemporary to the Second Industrial Revolution in Western countries. The goal is to highlight tensions between homogeneity and heterogeneity of information around the emerging technologies of that time, using computational methods. Topic Modelling models are applied to bring out the different discourses from the corpus, and the context in which technologies are discussed. To face the challenge of a multi-lingual dataset, Cross-Lingual Word Embeddings Models are used, allowing to compare topics across places and time. Finally, clusters of topics are computed using Network Analysis tools for community detection. General discussions, from the domestic to industrial use of technologies, common to the studied countries emerge from these analysis, as well as others topics specific to their societies.*

*Si l'idée de la technologie comme moteur de l'histoire et de la société a été largement étudiée dans l'histoire des Révolutions Industrielles, peu d'attention a été accordée à la perception sociale de ces-dites technologies. Cette thèse se concentre sur les discours relatifs à la technologie, observés dans les journaux contemporains de la Deuxième Révolution Industrielle dans les pays occidentaux. L'objectif est de mettre en évidence les tensions entre l'homogénéité et l'hétérogénéité des informations autour des technologies émergentes de l'époque, en utilisant des méthodes computationnelles. Des modèles de Topic Modelling sont appliqués pour faire ressortir les différents discours du corpus, et le contexte dans lequel les technologies sont discutées. Pour relever le défi d'un ensemble de données multilingues, des modèles de Cross-Lingual Word Embeddings sont utilisés, permettant de comparer les sujets à travers les lieux et le temps. Enfin, des groupes de discussions sont calculés à l'aide d'outils de Network Analysis pour la détection de communautés. Des discours généraux, de l'utilisation domestique à l'utilisation industrielle des technologies, communes aux pays étudiés émergent de ces analyses, ainsi que d'autres sujets spécifiques à leurs sociétés.*

Keywords : Data Analysis, Natural Language Processing, Computational Linguistics, Second Industrial Revolution, Digital History





# Acknowledgments

---

I am extremely grateful to have shared this journey of a thesis with Elena, who supervised me, always encouraged me, and with whom I had great discussions that I looked forward to every Thursday. Additionally, I would like to extend my sincere thanks to Jérôme, who has been very available and gave precious advice throughout the semester.

Many thanks also go to my friends, the *Abiskopaines*, who never failed to support me, by spending time with me, cooking me food, or just being there in my mind. I am also thankful for my family who shared with me the small break over winter, providing me enough energy to finish this thesis.

Lastly, I would be remiss in not mentioning Taylor Swift, whose albums were the background music of almost all the work in this thesis, and because *I had the time of my life fighting dragons with you*.



# Contents

---

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgments</b>	<b>v</b>
<b>Contents</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 The Second Industrial Revolution . . . . .	1
1.2 Analyzing newspapers discourses about technology . . . . .	3
1.3 Computational Methodologies . . . . .	4
<b>I Corpus</b>	<b>5</b>
<b>2 Corpus Description</b>	<b>7</b>
2.1 The Journals . . . . .	7
2.1.1 Journals Presentations . . . . .	7
2.1.2 Journals data . . . . .	9
2.2 Keywords filtering . . . . .	9
2.2.1 Keywords Choice . . . . .	9
2.2.2 Extracting Articles . . . . .	11
<b>3 Quality Of Corpus</b>	<b>13</b>
3.1 Quality of the Optical Character Recognition . . . . .	13
3.2 Steps for pre-processing . . . . .	14
<b>II Experiments</b>	<b>15</b>
<b>4 Pachinko Allocation Modelling</b>	<b>17</b>
4.1 Motivations . . . . .	17
4.2 Pachinko Allocation Modelling . . . . .	18
4.3 Evaluation method : Coherence of topics . . . . .	19
4.4 Results . . . . .	20

<b>5</b>	<b>Cross-Lingual Word Embeddings</b>	<b>23</b>
5.1	Word Embeddings . . . . .	23
5.1.1	Models Descriptions . . . . .	23
5.1.2	Evaluation Metrics and choice of model . . . . .	25
5.2	Cross-Lingual Word Embeddings : Alignment of models . . . . .	28
5.2.1	Motivation . . . . .	28
5.2.2	Stochastic Gradient Descent Least Squares . . . . .	29
5.2.3	Stochastic Gradient Descent With Orthogonality Constrains . . . . .	29
5.2.4	Wasserstein Procrustes . . . . .	30
5.2.5	Choice of alignment . . . . .	32
5.2.6	Visualisation methods . . . . .	32
<b>6</b>	<b>Network of Cross-Lingual Topics in Time</b>	<b>35</b>
6.1	Creation of Network of Topics . . . . .	35
6.1.1	Word Mover’s Distance . . . . .	35
6.1.2	Network construction . . . . .	36
6.2	Community Detection . . . . .	36
6.2.1	Clauset-Newman-Moore greedy modularity maximization . . . . .	37
6.2.2	Louvain method . . . . .	37
6.3	Results and Evaluation . . . . .	38
6.3.1	Evaluation Metrics . . . . .	38
6.3.2	Results . . . . .	38
	<b>III Discussions</b>	<b>43</b>
<b>7</b>	<b>Coal</b>	<b>45</b>
7.1	Domestic usage . . . . .	45
7.1.1	Daily life . . . . .	45
7.1.2	Medicine . . . . .	45
7.2	Industrial use . . . . .	46
7.2.1	A vital resource . . . . .	46
7.2.2	The coal production by miners . . . . .	47
7.2.3	Coal crisis . . . . .	47
7.3	A source of strategy and tensions . . . . .	48
7.3.1	The control of resources inside countries . . . . .	48
7.3.2	Coal diplomacy and trade between countries . . . . .	49

<b>8</b>	<b>Steel</b>	<b>51</b>
8.1	Steel’s retail applications until the 1880s . . . . .	51
8.1.1	Art . . . . .	51
8.1.2	Fashion . . . . .	51
8.1.3	House items . . . . .	52
8.2	The steel industry . . . . .	52
8.2.1	Steel involved in world-wide trade . . . . .	52
8.2.2	Steel as a war resource . . . . .	53
<b>9</b>	<b>Electricity</b>	<b>55</b>
9.1	Publicity of the scientific discoveries . . . . .	55
9.1.1	Scientific discoveries shared through newspapers . . . . .	55
9.1.2	Public experiments and exhibitions . . . . .	56
9.1.3	Praise of scholars . . . . .	56
9.1.4	Electricity : mystical and religious . . . . .	56
9.1.5	Detractors and critics . . . . .	56
9.2	The revolution of electricity . . . . .	57
9.2.1	Fast human progress . . . . .	57
9.2.2	Implications of electricity in different areas . . . . .	57
9.2.3	Electricity to justify power . . . . .	58
9.3	Industry of electricity . . . . .	59
9.3.1	The essential merchandise . . . . .	59
9.3.2	Private or public governance . . . . .	59
<b>10</b>	<b>Telegraph</b>	<b>61</b>
10.1	The Telecoms Revolution . . . . .	61
10.1.1	Speed, ingenuity, and enthusiasm . . . . .	61
10.1.2	Reduction of distances . . . . .	62
10.1.3	Integration in daily life . . . . .	62
10.2	In wars and politics . . . . .	63
10.2.1	Communications . . . . .	63
10.2.2	Administration and political life . . . . .	64
10.2.3	Strategical in wars and protests . . . . .	64
10.2.4	Part of war organisation . . . . .	65
10.3	The telegraph network development . . . . .	65
10.3.1	Organisation of national and international services . . . . .	66
10.3.2	Regional Disparities . . . . .	67
10.3.3	Cost . . . . .	68
10.3.4	Reliability in the service . . . . .	68

<b>11</b>	<b>Limitations</b>	<b>69</b>
<b>12</b>	<b>Conclusions &amp; Outlook</b>	<b>71</b>
	<b>Bibliography</b>	<b>75</b>
	<b>Appendix</b>	
A	KEYWORD DISTRIBUTION . . . . .	
B	DIRICHLET DISTRIBUTION . . . . .	
C	GIBBS SAMPLING . . . . .	
D	TOPICS EXAMPLE FOR ELECTRICITY ARTICLES IN SPANISH DATASET (1910-1920) . . . . .	
E	BEST PARAMETERS FOR PAM FOR EACH MODEL . . . . .	

*During the Second Industrial Revolution (1870-1914), Western societies witnessed an unprecedented wave of technological innovations arriving to society. As a consequence, profound social changes took place. While this period of time has been extensively analyzed across academic fields such as History, Geography, or Economics, little attention has been given to different cultural perceptions of newspapers discourses about technology. This thesis will fill that research gap using a multilingual dataset of historic newspapers in English, French, German, Italian, and Spanish, and computational methods, specifically, Topic Modelling and Network Analysis tools. The main research goal will consist in the observation of geographic behaviour of data, aiming to detect geographic clusters of discourses across time (1840-1930), and space - five Western countries: Spain, Italy, France, Germany and the United States.*

## 1.1 The Second Industrial Revolution

The Second Industrial Revolution is a term used to describe the period of time in the second half of the XIXth century and the beginning of the XXth century which has seen a number of technological inventions and innovations blossom that lead to wide economic and social changes, in the Western world [MS98]. Exact dates for its beginning and end are consistently in debate, a debate that arguably cannot ever be solved because of the diversity of time and places where the innovations occur. Besides, most of the innovations are building on top of older ones, so there is no direct frontier for a beginning and an end of an innovation. Vaclav Smil [Smi05] argues for an "Age of Synergy" between 1867 and 1914, characterized by an "intensity and competitiveness of the era's quest for innovation", while Robert J. Gordon [Gor00] is more restrictive and states that the 1860-1900 period is when the largest aggregate of innovations was. Some authors like Albert Edward Musson [Mus78] even go as far as saying that the inter-war period is where the actual Second Industrial Revolution stands. The time period selected for this thesis is the most generous one, the study will focus on public opinion between 1840 and 1930.

The particularity of the technological innovations of the period is the impact they have had, either right away or in the next century, onto societies that adopted

them. Vaclav Smil divides them into four fundamental categories [Smi05], further explained below :

- *the formation, diffusion, and standardization of electricity-generating systems and the distribution and uses of this most versatile form of energy*
- *the invention and rapid adoption of internal combustion engines, the dominant prime mover in transportation*
- *the unprecedented pace of the introduction of new materials and industrial chemical syntheses*
- *the birth of a new information age thanks to the new means of communication*

The first significant category is electricity and its broad range of applications. Electricity first started to be theorized in 1826 with André Ampère electro-dynamic theories and in 1827 with George Ohm's complete mathematical theory of electricity [For16] [Amp22] [Ohm27]. About 40 years later, the first reliable dynamo that provides continuous DC power is introduced [Rei17], and is soon subject to incremental improvements. Because of the power it affords, engineers become creative in its applications, from light to transportation to communications.

Then, combustion engines are another great part of the Industrial Revolution. Used primarily in transportation such as cars, boats or railroads, they planted the seed for the modern transportation industries. It first started with relatively inefficient coal-gas-fueled sources of stationary power during the first half of the XIXth century[Smi05]. They then became more and more compatible with industrialisation, with for example Nicolaus Otto's 1876 design of a compression four-stroke engine [Bea14]. Along with the line assembly introduced by Henry Ford [BO91], mass manufacturing became possible in these industry relying on internal combustion engines and saw their production numbers rise.

In parallel, new industrial processes boosted metal and chemicals usage in different industries. As the materials become cheaper to produce, their quality increase. Pretty swiftly, Siemens-Martin furnaces are boosting the production of alloys [Mio97], Hall-Hérault process are allowing to use the lightness of aluminium in transportation [MRB10], and Nobel's dynamite is participating into totally changing the dynamics of wars [Wis08]. All these rapidly introduced new processes are shaping the second Industrial Revolution.



Finally, communications are another important area that profited of the technical advances of that time. Among them are the linotypes in 1885 [Ove02], the chemical processes for paper-making based on wood pulp [Smi05], or convenient and affordable cameras. All these inventions completely revolutionized how the mass media functioned, connecting at an unprecedented speed every part of the world.

## 1.2 Analyzing newspapers discourses about technology

Engaging with Vaclav Smil's periodization of history, this thesis examines discourses about technology as expressed in a dataset of multilingual newspapers in French, Italian, German, Spanish, and English, contemporary to the Second Industrial Revolution (1840-1930). In that context, it is interesting to first do a review of existing notions of public discourses. The concept of publicity in Europe dates back to Ancient Greece, where Agoras were the place of social, political, and mercantile gatherings, and was the place where the different schools of thought were formed [Sen16] [Glo88]. Habermas defines the public sphere as "*a sphere between civil society and the state, in which critical public discussion of matters of general interest was institutionally guaranteed*" [Hab91]. He also argues that a private idea becomes a public opinion when it is put through a rational and critical debate. With early capitalism and the facilitation of long-distance communications, ideas are more easily and rapidly circulating, becoming a part of the word-wide traffic and fed to the public discourse.

While this thesis is informed by Habermas's definition of the public sphere, its main focus will consist in analyzing discourses about technology in newspapers broadly speaking. The research goal that will be pursued in the following pages will be the observation of geographic tensions between homogeneity versus heterogeneity of information as recorded in the selected dataset of multilingual newspapers. The chosen period of analysis, 1840-1920, coincides with the Second Industrial Revolution, labelled by many authors (section 1.1) as one of the most dramatic periods of social change in human history. Triggered by a highly aggressive wave of technological developments arriving to society, it is believed that Western societies were profoundly transformed during this time. Yet little attention has been paid to the differences, or similarities, in which Eurocentric societies experienced that profound social transformation. This thesis will fill that research gap using a collection of multilingual newspapers as an object of analysis and a variety of

computational methodologies.

### **1.3 Computational Methodologies**

The thesis implements a pipeline to study and compare the different discourses in the newspapers, using Natural Language Processing and Network Analysis methodologies. Once the newspapers text data retrieved (Section 2.1) and with isolated relevant articles (Section 2.2), the themes appearing in the newspapers are computed separately for each journal using Pachinko Allocation Modelling (Section 4.4). With these results, the ensuing goal is to create a visualisation that allows to compare across decades and countries the topics, in order to find similarities or differences. To pursue it, different word embeddings models are tested (Section 5.1) and then aligned to create a Cross-Lingual Word Embedding Model (Section 5.2). Finally, with multi-lingual word embeddings and the topics from each decade, the distances between topics are computed and used to create a network of topics, where clusters of similar topics can be detected through community detection algorithms (Section 6).

Part I  
Corpus



# 2

## Corpus Description

---

*Newspapers of record are the primary source of the thesis. This section describes the selected dataset, from the journals and their content, to the specific articles and the methods to extract them.*

### 2.1 The Journals

Because the goal of the thesis is to study the discourses in Eurocentric societies, newspapers from these countries are chosen. The study focuses on five different countries : France, Italy, Germany, Spain, and the United States. Each country has its own *newspapers of records* [MH98]. These newspapers are characterized by having a *certain reputation for consistent attention to accuracy and depth in reporting local as well as international news* [MF80]. The one chosen for the study were also published on a regular basis.

The main goal of this thesis consists in the analysis of newspapers discourses about technology seeking to detect tensions between homogeneity and heterogeneity of information. Therefore newspapers of record, a relatively neutral source of information, provide an ideal medium where to contrast and compare how a same message - public discourse about technology - is decoded. Using a more or less comparatively similar medium of analysis should allow to discern similarities or differences across Western countries. In the following section, a description of each newspaper will be found.

#### 2.1.1 Journals Presentations

##### Le Figaro

*Le Figaro* is a daily french newspaper. Historically satirical, open to discussing politics, and leaning towards socially conservative, the paper describes the Parisian scene while being willing to publish important authors from its time such as Emile Zola in 1880 or Anatole France in 1899, at a time when the journal was among the

most printed ones in Europe. [Ber07].

### **The New York Herald**

The american paper *The New York Herald* at its creation in 1835 was described as an apolitical journal and supporting no party [Cro89]. Before the Civil War, it did endorse the Democratic, Whig, and Republican parties [Bri05]. Its founder, James Gordon Bennett, is known for his sensationalist editorial line, to attract large audiences [CJ73]. The journal covers topics such as finance, sports, foreign events, society affairs, and the theatre [Bri05]. It was the most popular journal in the United States in 1845 [Cro89].

### **El Imparcial**

*El Imparcial* was one of the main published newspaper in Spain at the end of the XIXth century, and politically oriented towards a liberal ideology even though it was not affiliated with any political parties [Mar01]. The paper, that was published in Madrid, circulated in the whole country from March 1867 to May 1933, with around 120.000 prints in the beginning of the XXth century [APZ00].

### **Neue Hamburger Zeitung**

The daily newspaper *Neue Hamburger Zeitung* was in circulation between January 1896 and 1922 in Hamburg, Germany. The journal's content has historically non-partisan positions. In the beginning of the XXth century, around 150.000 copies were circulating in the country [Son06].

### **La Stampa**

*La Stampa* is an italian journal founded in 1867, published in Turin. The paper sometimes discuss politics as it was first an "*important voice in Italy's struggle for liberation and unification*" [Bri17]. It often discusses social issues in Italy, as well as sport events and featured intellectuals such as Luigi Einaudi [De 06]. The newspaper circulated a lot in Italy, with around 100.000 copies per year at the beginning of the XXth century [For12].

### 2.1.2 Journals data

The collection constituted for the thesis varies in time periods, depending on the journal, for reasons of availability. Figure 2.1 summarizes the available time period for each newspaper, along with their source. In addition, not every newspaper of every year has been digitized and then OCR'ed by the above institutions. Figure 2.1 shows the distribution of the gathered newspapers over time.

**Table 2.1:** Summary of the collections' journals, time periods and sources

Journal	Time Period	Source
<i>Le Figaro</i>	1860-1920	Gallica
<i>The New York Herald</i>	1840-1880, 1920	Library of Congress
<i>El Imparcial</i>	1860-1930	Biblioteca Nacional de Espana
<i>La Stampa</i>	1882, 1910-1930	La Stampa Archives, by [Bas+21]
<i>Neue Hamburger Zeitung</i>	1880-1930	Deutsche Digitale Bibliothek

## 2.2 Keywords filtering

### 2.2.1 Keywords Choice

Considering that the intent of the thesis is to study public discourses about the technologies of the Second Industrial Revolution, one now needs to narrow down the dataset to articles that are discussing these technologies. The introduction already gave context for the technology advances during the second half of the XIXth and the beginning of the XXth centuries. For each of the four categories of technologies as detailed in section 1.1, one relevant word is chosen to filter the articles. Indeed, the study focuses on articles that are mentioning at least once one of the four keywords : **coal**, **steel**, **electricity**, or **telegraph**.

#### Coal

The first category of technological change defined in section 1.1 is the "*invention and rapid adoption of internal combustion engines, the dominant prime mover in transportation*". The keyword **coal** will discriminate the articles falling under that scope. Coal is used in the Second Industrial Revolution as a source of power, where the energy produced by heat is transformed into a mechanical one. This technology goes hand-in-hand with steam engines, and allowed many industries to grow faster.



**Figure 2.1:** Pages count distribution across the years for the journals *Le Figaro*, *The New York Herald*, *Neue Hamburger Zeitung*, *La Stampa* and *El Imparcial*

Indeed, coal has been an indicator of industrial growth in Europe since the XVIIIth century [FO20], but during the Second Industrial Revolution many micro-inventions are making coal combustion engines better and better [MS98], so different articles about the usage of coal rather than of the coal innovations are expected.

## Steel

Then, **steel** is chosen as a representative of *the unprecedented pace of the introduction of new materials and industrial chemical syntheses*. With the Bessemer converter, steel takes over from iron by using the impurities in it and its interaction with air. Joel Mokyr and Robert H Strotz [MS98] explain that the process is then gradually improved by Robert Mushet, Percy Gilchrist and Sidney Thomas, and it becomes



cheaper and cheaper to produce steel, of ever better quality. This transforms the modes of production in larger scales, for example in the ships construction.

### Electricity

For the *formation, diffusion, and standardization of electricity-generating systems and the distribution and uses of this most versatile form of energy*, the selected keyword is **electricity**. It started to expand after 1870 with the invention of generators and dynamo to produce current [MS98]. Symbol of a new scientific field and new knowledge, it finds applications in lightning, motors, electric railways, etc. These are the kind of topics expected to be found in articles.

### Telegraph

Finally, the last representative word will be the **telegraph**, illustrating *the birth of a new information age thanks to the new means of communication*. It is an information technology that follows from electricity. The first theories for it are found in the 1830s, but it took several incremental improvements to make it long distance thanks to submarines cables [MS98]. Another noticeable amelioration is wireless telegraphy from the mid 1890s, able to transmit Morse code [MS98].

## 2.2.2 Extracting Articles

With the journals and the words well defined, the articles can be extracted to constitute the dataset. An article of the journal is kept if and only if it contains at least once the studied word, in the language of the journal (see Table 2.2 for the translations). Because the content of each journal is not only made of articles, but also of titles and advertisement, some of them will also end up in the analysis. The steps to extract the texts are explained below, and are different depending on the journal because they do not all come in the same format.

**Table 2.2:** Keywords translation table

United States	France	Spain	Italy	Germany
Coal	Charbon	Carbón	Carbone	Kohle
Steel	Acier	Acero	Acciaio	Stahl
Electricity	Electricité	Electricidad	Elektrizität	Elettricità
Telegraph	Télégraphe	Telégrafo	Telegrafo	Telegraf

### Le Figaro and El Imparcial

The text of the files have been retrieved by Optical Character Recognition (OCR) by their respective sources. OCR'ed files from *Le Figaro* and *El Imparcial* are formatted in a way that between every article in the newspaper page, there is a line break followed by a blank space. It naturally follows that the article's text is retrieved by splitting the text on the line breaks and keeping only the ones that contain the studied word.

### The New York Herald

*The New York Herald* files are not as easily dealt with. There is no blank space in the OCR text files. However, each article starts by its title in capital letters, as illustrated in the example of Figure 2.2. The letters' case is kept through the OCR. With that knowledge, the newspaper's text is split on the capital letters titles, and as before only the articles with the word of interest are kept.

### La Stampa and Neue Hamburger Zeitung

For *La Stampa* and *Neue Hamburger Zeitung*, the OCR'ed newspapers' articles do not have any layout separation. As there is no way to straightforwardly separate them, only the N context-sentences around the keyword are kept, meaning that an "article" will be composed of the N sentences before the word, and N sentences after. N is arbitrarily set to 3, as it has to be small enough to not overstep on another article which could be of a totally different subject and would disturb the smoothness of the analysis.

The figure from Appendix 12.1 shows the distribution of articles extracted with the keywords over time.

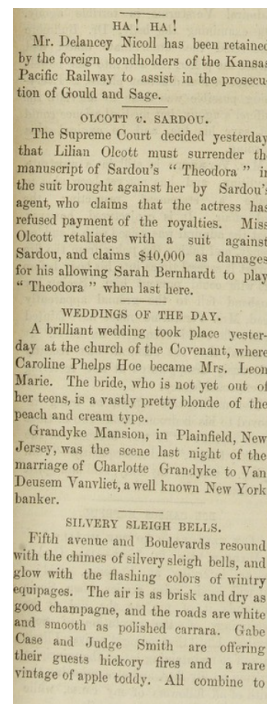


Figure 2.2: *The New York Herald*, excerpt from the issue of the 21-01-1888

# 3

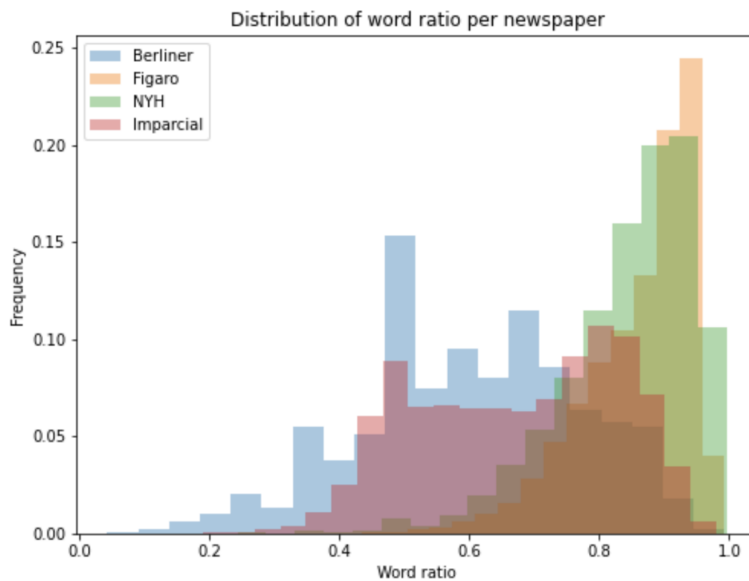
## Quality Of Corpus

---

*With the dataset constituted, this next part discusses the quality of the files constituting the corpus, and describes how they are processed prior to any quantitative analysis.*

### 3.1 Quality of the Optical Character Recognition

Before studying the corpus, it is important to examine the quality of the text files constituting it. The different newspapers of the study are coming from different sources, which are of various quality. The quality of the OCR is assessed by using word ratios, which is a measure defined by the number of tokens belonging to a reference corpora over the number of digitized token recognized by the OCR method used [KC15]. This method has been used by Thomas Benchetrit [BFB22] and Figure 3.1 shows the word ratios for four journals of the dataset.



**Figure 3.1:** Distribution of word ratio per newspaper of the dataset, by Thomas Benchetrit [BFB22]

Figure 3.1 shows that French and English newspapers are of much better quality than the Spanish and German ones. Thomas Benchetrit partly explains it by the difference in model development across languages, and also by their morphological characteristics [BFB22].

## 3.2 Steps for pre-processing

Poor OCR files highlight the importance of pre-processing. With this step the goal is to improve the document's quality by correcting or filtering out words poorly OCR'ed, in order to improve performance [Kan+14]. The difference in OCR quality might however yield to higher variance in the results across languages.

Each text file is processed the following way :

1. Keep the file only if its less than 500000 characters, for efficiency
2. Remove words that are three letters or less, because they often result from OCR mistakes
3. Remove numerical entries
4. Tokenize the text, with the Spacy [HM17] library. We make an "isolating" assumption, that the words do not divide into smaller units [Kan+14].
5. Remove the stopwords, using the stopwords dictionary of the Spacy library [HM17].
6. Lemmatize the words, using again the Spacy library [HM17].

Part II

Experiments



# 4 Pachinko Allocation Modelling

---

*The first goal of the thesis is to detect the different topics discussed around technology in Western newspapers. The chapter below provides explanations and descriptions of the algorithms and models used for this purpose. The topics are detected using Pachinko Allocation Modelling, and successfully capture different discourses across the different regions and period of time.*

## 4.1 Motivations

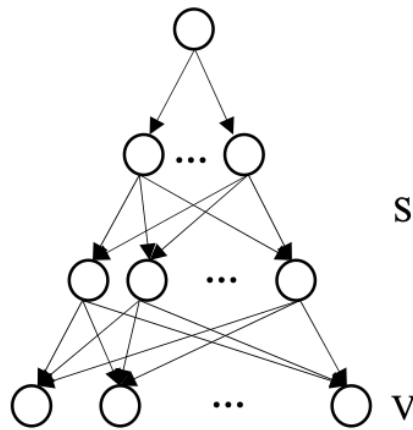
In Natural Language Processing, the field that studies "topics" in corpora is called Topic Modelling. "Topics" are typically clusters of words that semantically make sense together, and that can be used alone or in combination to describe documents. Such statistical methods are useful for revealing the underlying semantic structure in large collection of documents [KB19]. Topic modelling techniques nowadays are using probabilistic approaches, with generative models such as Latent Dirichlet Allocation (LDA) [BNJ03] and Probabilistic Latent Semantic Analysis [Hof13], that describe documents as a mixture of topics. Depending on the type of problem they aim to solve, topic models can be trained in a supervised [MB07] [Res+15] or unsupervised [Guo+16] [KZ21] manner. Some models are also made more complex to incorporate the idea of correlation between the topics [LM06] [BL07], their relationships in time [BL06], or between languages [BB12].

The selected technique for this thesis is Pachinko Allocation Modelling. Like LDA, it assumes a Dirichlet distribution of the words over the topics, but also assumes that there could be different correlations between the topics, which models more accurately real-life data [LM08]. The corpus is challenging as it is both multi-lingual (has five different languages) and temporal (the order of the articles matter). Answer to these problems will be proposed in a following chapter (Chapter 6). First, to prepare the ground of more complex work, topic modelling is done separately on the different journals, by chunks of decades. By applying these methods to the corpora of articles discussing technologies, clusters of discourses should emerge

and with them a beginning of an answer to the thesis' questions. The code for every experiment is available on Github <sup>1</sup>.

## 4.2 Pachinko Allocation Modelling

Pachinko Allocation Modelling (PAM) is a generative topic model that aims to capture correlation between topics. PAM bases its computations on Directed Acyclic Graphs (DAG), where the leaves are words and the interior nodes are topics [LM06]. The topics are drawn over a Dirichlet distribution (see Appendix B for its definition), associated with super-topics, higher in the DAG. Following the steps of Wei Li and Andrew McCallum ([LM06]), the model sticks to a four-level hierarchy graph consisting of a root, a set of super-topics, a set of sub-topics and a word vocabulary (Figure 4.1).



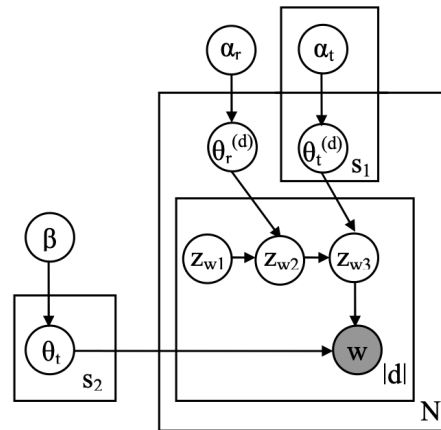
**Figure 4.1:** Four-Level PAM. Represent the DAG structure with a root, super-topics and sub-topics (S) and the vocabulary (V). Figure from Wei Li and Andrew McCallum [LM06]

The algorithm to generate a document is the following : to generate a document  $d$ , the model from Wei Li and Andrew McCallum [LM06] first samples multinomial distributions  $\theta_{t_i}^{(d)}$  of topics  $t_i$ , from  $g_s(\alpha)_s$ , where  $g_s$  is a Dirichlet distribution, and  $\alpha_i$  is a vector with the same dimension as the number of children in  $t_i$ , over its children. Then, it samples a topic path  $\langle root, z_{sp}, z_{sb} \rangle$  of topics nodes, where  $z_{sb}$  is a subtopic, child of  $z_{sp}$ , a supertopic. It is sampled according to the multinomial distribution  $\theta_{z_{sp}}^{(d)}$ . Finally, a word  $w$  is sampled from  $\theta_{z_{sb}}^{(d)}$ . The whole process is

<sup>1</sup> <https://github.com/arobaselisa/industrial-west>



summarized in Figure 4.2. The hidden variables  $\theta$  and  $z$  are estimated using Gibbs sampling [Por+08] (see Appendix C for the details of Gibbs Sampling).



**Figure 4.2:** Graphical model for the Four-Level PAM. Figure from Wei Li and Andrew McCallum [LM06]

PAM is not flexible on the number of topics but works better when the number of relevant themes is known a priori [Zha+15]. Because the number of topics is not known, its optimal value is determined using a grid search. The process is detailed in the following sections. PAM also assumes that the documents are unstructured data, which fails to incorporate changes in time. Moreover, the corpus is made of multi-lingual journals, so a different topic model is used on every decade in every journal, separately. Chapter 6 tackles the problem of unifying everything together.

### 4.3 Evaluation method : Coherence of topics

To find the best hyperparameters for our model (the number of super-topics  $k_1$  and the number of sub-topics  $k_2$ ), the models are evaluated using a **coherence** metric, which is an intrinsic qualitative evaluation of learned topics [New+10]. The goal is to grasp the semantic similarity between the output topics. Many different scores have already been elaborated [RBH15], using pointwise mutual information [AS13] [Bou09] [New+10], Fitelson’s coherence [Fit03], or co-occurrences counts [Mim+11].

The selected metric is the  $c_V$ -coherence, because it was found to correlate the highest with human interpretation [New+10]. The metric compares the words from

each topic to all the other topics, and is based on Normalized Pointwise Mutual Information (NPMI), for which a definition is given below (Definition 4.1).

► **Definition 4.1.** Given  $P(w', w^*)$  the estimated probability of co-occurrence of words  $w'$  and  $w^*$ ,  $P(w)$  the estimated probability of the word  $w$  in the corpus, and a small  $\epsilon > 0$ , the **Normalized Pointwise Mutual Information** is defined as :

$$NPMI(w', w^*) = \frac{\log\left(\frac{P(w', w^*) + \epsilon}{P(w')P(w^*)}\right)}{-\log(P(w', w^*) + \epsilon)}$$

Then, for each topic word  $w_{n,k}$  at index  $n$  in topic  $k$ , a word vector of length  $N$ , the same size as the number of words per topic, is computed :

$$\mathbf{w}_{n,k} = NPMI(w_{n,k}, w_{m,k}) \quad \forall m \in \{1, 2, \dots, N\}$$

And then the topic vector for topic  $k$  is created by adding together all its topic word vectors :

$$\mathbf{t}_k = \sum_{n=1}^N \mathbf{w}_{n,k}$$

Finally, the  $c_V$  coherence score from David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin [New+10] is computed as the average of all cosine similarities between topics (Definition 4.2). The  $c_V$  scores range from 0 to 1, where higher scores mean more coherent topics.

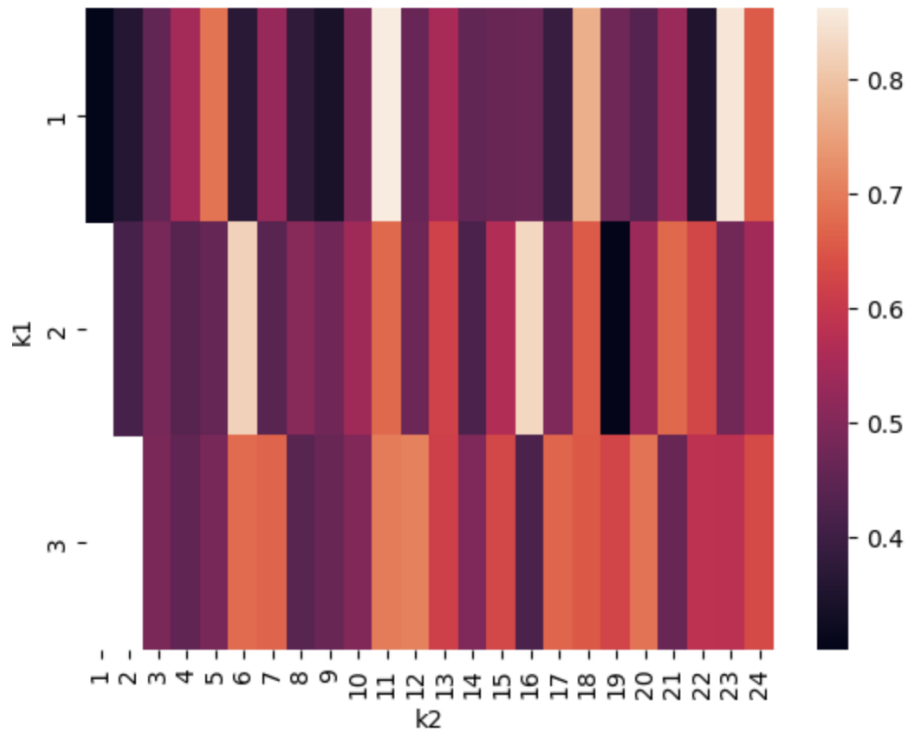
► **Definition 4.2.** Given  $sim(a, b)$  the cosine similarity between vectors  $a$  and  $b$ ,  $N$  the number of words per topic and  $K$  the number of topics, the  **$c_V$  coherence score** is defined as :

$$c_V(w', w^*) = \frac{\sum_{k=1}^K \sum_{n=1}^N sim(\mathbf{w}_{n,k}, \mathbf{t}_k)}{NK}$$

## 4.4 Results

With  $c_V$ -coherence as evaluation score for the topics, a grid search across different values for parameters  $k_1$  and  $k_2$  of the PAM model is done. Using the Spanish

*El Imparcial* newspapers articles containing the word "electricidad" ("electricity") between 1910 and 1920 as an example, Figure 4.3 displays the evaluation results for 69 different models, with  $k_1$  ranging from 1 to 3, and  $k_2$  ranging from 1 to 24.



**Figure 4.3:** Heatmap for the results of the grid-search for optimal parameters  $k_1$  and  $k_2$ , on "electricity" spanish articles, for the decade 1910-1920. The best model is with  $k_1 = 1$  and  $k_2 = 11$ , for a score of 0.86

The operation is repeated on each pair of decade-country. The best parameters for each pair can be found in the Appendix E, and coherence scores associated with them are summarized in Table 4.1. The difference in coherence scores between the topic is relatively the same, whereas the one across languages is quite high, because of the variance in number and quality of documents. In Appendix D can be found the list of topics for the example model of Figure 4.3, whereas the complete list is available on Github<sup>2</sup>.

<sup>2</sup> [https://github.com/arobaselisa/industrial-west/blob/main/data/topics\\_df.csv](https://github.com/arobaselisa/industrial-west/blob/main/data/topics_df.csv)

**Table 4.1:** Mean best coherence scores by country and keywords

<b>Country</b>	<b>Keyword</b>	<b>Mean Coherence</b>
Germany	coal	0.92
	electricity	0.73
	steel	0.93
	telegraph	0.80
Spain	coal	0.78
	electricity	0.74
	steel	0.81
	telegraph	0.79
France	coal	0.75
	electricity	0.69
	steel	0.58
	telegraph	0.77
Italy	coal	0.55
	electricity	0.51
	steel	0.77
	telegraph	0.73
United States	coal	0.83
	electricity	0.91
	steel	0.93
	telegraph	0.92

Pachinko Allocation models that rely on Directed Acyclic Graphs have been successfully used to model topics that are coherent between them. The resulting topics are specific to their decade and journal, therefore the subsequent part of the thesis tackles the problem of finding a way to compare topics across time periods and languages.

# 5 Cross-Lingual Word Embeddings

---

*Working from the list of topics that changes decades to decades, the next part establishes a framework that enables comparisons across decades and languages, using cross-lingual word embeddings. First, different models for word embeddings are tested on the corpus. Then, different methods are compared to perform alignments that move words translations close to one another, creating a common space for each of the languages.*

## 5.1 Word Embeddings

A usual tool to comprehend the semantic context of words in Natural Language Processing are word embeddings models. Word embeddings models allow to represent words as vectors in a multi-dimensional space, where words with similar meanings are close to one another [Tel00]. In the context of the project, word embeddings are meaningful to be able to find similar topics, even when they do not contain exactly the same words, but are close in meaning. The following section develops which model is chosen for the embeddings and how.

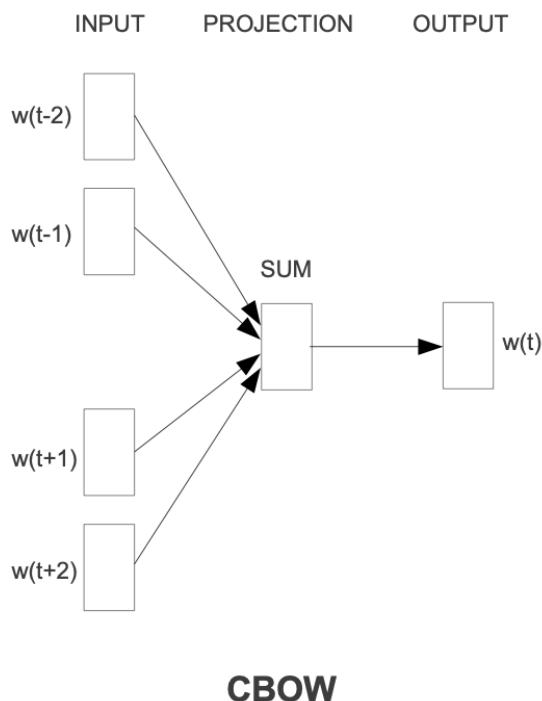
### 5.1.1 Models Descriptions

#### Word2Vec CBOW

The first and most famous word representation is called Word2Vec [Mik+13]. Its creation was motivated by finding a word representation space that preserves the linear regularities among words. For example, the vector *King* minus the vector *Man* plus the vector *Woman* leads to the vector representation of the word *Queen* [Mik+13]. To achieve that, the model takes into account the context of words in the calculation of the projection. It is called a Continuous Bag-of-Words Model (CBOW).

Concretely, the model is made of three layers : the input layer, the hidden projection layer, and the output layer (see Figure 5.1). For a vocabulary of size  $V$  and a word, the input layer takes the  $N$  previous and  $N$  following words to encode them using 1-of- $V$  encoding [BDV00]. Then, the input layer is projected using a shared projection matrix, and the output is the predicted current word. In short,

the distributed representations of context are combined to predict the word in the middle.



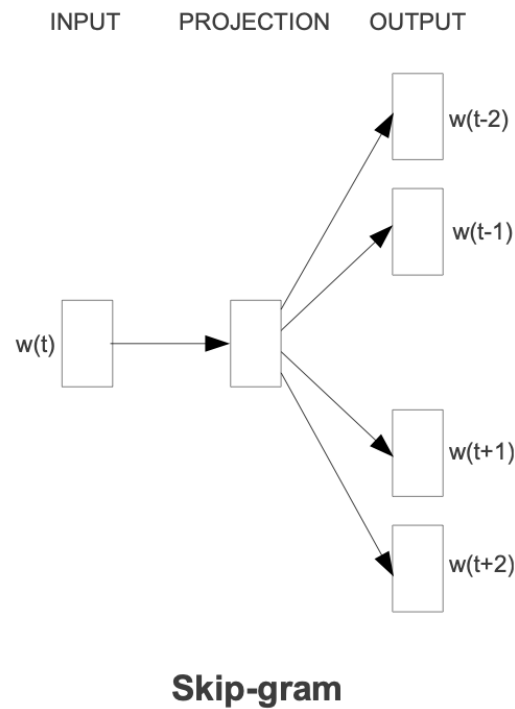
**Figure 5.1:** The CBOW model architecture, predicts the current word based on the context. Figure from [Mik+13]

### Word2Vec Skipgram

The Skipgram Word2Vec model is similar to the CBOW one, but uses the current word as an input to predict the N words before and after (see Figure 5.2). The quality of the prediction increases with range and complexity [Mik+13].

### Fasttext

Fasttext is a different model that is derived from the Skipgram model. Fasttext however takes into account the internal structure of words, representing them by a sum of n-grams. It has the advantage of allowing to share representations



**Figure 5.2:** The Skip-gram model architecture, predicts surrounding words given the current word. Figure from [Mik+13]

between words, and learn reliable representations for rare words, including words that might have been badly OCR'ed. [Boj+17]

### 5.1.2 Evaluation Metrics and choice of model

To choose the model that suits the thesis the most, there needs to be a way to intrinsically evaluate how great the models are for the task [GD16]. Ideally and assuming that words from a same topic are semantically close, great models are representing words from the same topic close to each other. For the later alignment task, it is also crucial that the space has a good geometry, that is has a good spread [Wan+19]. The following section describes three different metrics and evaluates the three different models defined in the previous section.

### Average Compactness of Topics (AvComp)

One metric to assess the geometry of a word space with respect to the topics is the **compactness** [CN16], defined below :

► **Definition 5.1.** Given  $T = \langle w_1, w_2, \dots, w_n \rangle$  a topic, and  $sim(w_i, w_j)$  the cosine similarity between two word vector, the **compactness** of a word  $w$  is defined as :

$$c(w, T) = \frac{1}{n(n-1)} \sum_{w_i \in T \setminus \{w\}} \sum_{\substack{w_j \in T \setminus \{w\} \\ w_j \neq w_i}} sim(w_i, w_j)$$



Then, to get a model average compactness score based on the topics, the following computations are made :

- For each word of each topic, compute its compactness  $c(w, T)$  within its topic  $T$
- Compute the average of the  $c(w)$  over each word from each topic

Because a low cosine similarity means that two word vectors are close, then the lower the score will be, the better an embedding model is with regard to the compactness of the topics.

### mean Outlier Position Percentage (mOPP)

The second method used to measure the quality of the embeddings is the **mean Outlier Position Percentage (mOPP)**. The measure is inspired by the Outlier Position Percentage, which was first introduced by José Camacho-Collados and Roberto Navigli [CN16]. The steps to compute the score are detailed in Algorithm 1. The score computation involves computing the number of times a word new to a topic can be detected as an outlier of the topic. The higher the score is, the better the list of topics is.

### Compactness ratios (CR)

Finally, the last measure that allows to measure how distinct the topics are is the **compactness ratio** (CR, definition 5.2). It simply consists of average of the ratios of the compactness score of words in their topics, and in other topics. The smaller the score is, the better the topics are in that respect.



**Algorithm 1** mean Outlier Position Percentage

---

```

1: Input : Vocabulary  $V$ , List of topics  $\langle T_1, \dots, T_M \rangle$  where  $T_i = \langle w_1, \dots, w_N \rangle$  is
   a list of words
2:  $outliersDetected \leftarrow 0$ 
3: for  $T_i$  in  $\langle T_1, \dots, T_M \rangle$  do
4:   Compute the mean compactness score of the topic  $c_{T_i}$ , as described in 5.1.2
5:   for  $w_i$  in  $V, w_i \notin T_i$  do
6:     Compute the compactness score  $c_{T_i}^*$  of the topic  $T_i^* = T_i + \langle w_i \rangle$ 
7:     if  $c_{T_i}^* > c_{T_i}$  do
8:        $outliersDetected \leftarrow outliersDetected + 1$ 
9:   end for
10: end for
11: Output :  $mOPP \leftarrow \frac{outliersDetected}{(N)*(M-1)}$ 

```

---

► **Definition 5.2.** Given  $T = \langle T_1, \dots, T_M \rangle$  the list of topics  $T_i = \langle w_1, \dots, w_N \rangle$ , the CR score is defined as :

$$CR = \frac{1}{M * N * (M - 1)} \sum_{T_m \in T} \sum_{w \in T_m} \sum_{T_l \neq T_m} \frac{c(w, T_m)}{c(w, T_l)}$$

**Choice of Model**

With the different measures explained, the three models Word2Vec CBOW, Word2Vec Skipgram, and Fasttext described in Section 5.1.1 are trained on data from the "Telegraph" articles, from the journal *Neue Hamburger Zeitung*. The vectors are of size 100, and the models have a window of length 5. The CBOW and Skipgram Word2Vec models are using negative sampling. The results are summarized in table 5.1. The Fasttext performs significantly better according to the AvComp and CR scores, and the Word2Vec CBOW models works slightly better with the mOPP. Because the Fasttext models outperforms the Word2Vec models, they are chosen for the rest of the experiments, and separate models are trained on the French, Italian, Spanish and American datasets.

**Table 5.1:** Evaluation of Word Embedding Models trained on *Telegraph* data using the three different metrics AvComp, mOPP and CR

	AvComp	mOPP	CR
Word2Vec SG	0.72	<b>0.14</b>	0.55
Word2vec CBOW	0.51	0.13	0.35
Fasttext	<b>0.20</b>	0.11	<b>0.07</b>

## 5.2 Cross-Lingual Word Embeddings : Alignment of models

With the idea of comparing topics across languages and decades, the following section aims to create a common semantic space for the five language French, Italian, Spanish, English and German, based on the separate mono-lingual word embedding models.

### 5.2.1 Motivation

The art of finding common semantic spaces for words across different languages is done through Cross-Lingual Word Embeddings Models (CLWEM). A good overview of the field is explained by the survey from Sebastian Ruder, Ivan Vulić, and Anders Søgaard [RVS19]. In an ideal CLWEM, semantic relationships between words are expressed, and translations of words are also close to one another. Different approaches exist to find such embeddings. Monolingual mappings are separately training word embeddings for each language, then find a mapping between them ([MLS13], [Amm+16]). Pseudo-cross-lingual are beforehand creating a cross-lingual corpus and training the cross-lingual model on it ([XG14]). Cross-lingual training uses in parallel the corpus from each language ([HB13]). A model also depends on the training data available, whether the corpus is aligned on a word, sentence, document or lexicon level ([KTB12]).

With the assumption that English, French, Spanish, Italian and German have a similar global semantic structure, meaning that the words for each language usually have an exact translation in the others, the goal is to find a monolingual mapping between them. This claim is supported by [Sen+17]. They computed semantic similarity between languages using word embeddings and representational similarity analysis [Nil+14]. They show, among other things, that the language the most correlated in average to the others is English. The target language is therefore

English for the following experiments. Three different methods for the alignment are tested with the American and French corpora, to be then generalized to the entire corpus.

### 5.2.2 Stochastic Gradient Descent Least Squares

The goal is to find a linear transformation from a source to a target language. The first approach for the alignment will rely on lexicon-level alignment with pairs of words translations. Tomas Mikolov, Quoc V Le, and Ilya Sutskever [MLS13] translate that question to an optimization problem where, with translation pairs  $\{x_i, z_i\}$ , they need to find  $\mathbf{W}$  that minimizes :

$$\sum_{i=1}^N \|\mathbf{W}x_i - z_i\|^2$$

To solve it, Stochastic Gradient Descent (SGD) is used (see Algorithm 2). The gradient is computed using the chain rule.

---

#### Algorithm 2 Stochastic Gradient Descent Least Squares

---

- 1: **Input** : Source embeddings  $\mathbf{X}$ , target embeddings  $\mathbf{Z}$ , learning rate  $\alpha$
- 2: Randomly initialize  $\mathbf{W}$
- 3: **for**  $t = 1$  to  $T$  **do**
- 4:     Compute the gradient :

$$\nabla W = 2 \sum_{i=1}^N \mathbf{w}^T (\mathbf{W}x_i - z_i)$$

- 5:     Update the weight matrix following the rule :

$$W \leftarrow W + \alpha \nabla W$$

- 6: **end for**
- 

### 5.2.3 Stochastic Gradient Descent With Orthogonality Constrains

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin [Xin+15] observed that the distances in word embeddings models are usually computed using cosine similarity,

which differs from the objective function from Section 5.2.2. To tackle this problem, they propose to modify the problem into an optimization problem with a quadratic constraint. All the words have to lay on a sphere of diameter one. This is done by normalizing the vectors when updated, which transforms the inner product into a measurement of the cosine distance, coherent with the distance measurement.

The optimization problem to solve is then :

$$\max_W \sum_i (Wx_i)^T z_i$$

And is solved using gradient descent (Algorithm 3), where the gradient is :

$$\nabla W = \sum_i x_i y_i^T$$

---

**Algorithm 3** Stochastic Gradient Descent With Orthogonality Constrains

---

- 1: **Input** : Source embeddings  $\mathbf{X}$ , target embeddings  $\mathbf{Z}$ , learning rate  $\alpha$
- 2: Randomly initialize  $\mathbf{W}$
- 3: Normalize  $\mathbf{X}$  and  $\mathbf{Z}$
- 4: **for**  $t = 1$  to  $T$  **do**
- 5:     Update the weight matrix following the rule :

$$W \leftarrow W + \alpha \nabla W$$

- 6:     Orthogonalize with constrained quadratic problem, solved by SVD

$$\min_{\bar{W}} ||W - \bar{W}|| \text{ s.t. } \bar{W}^T \bar{W} = I$$

- 7: **end for**
- 

### 5.2.4 Wasserstein Procrustes

Finally, the third method that will be tested is Wasserstein Procrustes, another algorithm that allows to align two embeddings. This algorithm is unsupervised, unlike the two previous ones. Wasserstein Procrustes is based on the joint estimation of an orthogonal matrix and a permutation matrix [GJB19]. It starts with the observation that two problems need to be solved when aligning embeddings : the

original Procrustes problem [Gow75] which is about computing a linear transformation between two sets of points  $\mathbf{X}$  and  $\mathbf{Y}$ , and the problem of finding a one-to-one correspondence between the two sets.

The Procrustes problem can be defined as

$$\min_{\mathbf{Q} \in O_d} \|\mathbf{XQ} - \mathbf{Y}\|_2^2$$

Where  $O$  is the set of orthogonal matrices in order to ensure that the distances are not changed by the transformation. Peter H Schönemann [Sch66] gives the solution to the problem, as :

$$\mathbf{Q}^* = \mathbf{UV}^T$$

Where  $\mathbf{USV}^T$  is the Singular Value Decomposition (SVD) of  $\mathbf{X}^T\mathbf{Y}$

The second problem of finding a one-to-one correspondence between two sets of points can be described as follow :

$$\min_{\mathbf{P} \in P_n} \|\mathbf{X} - \mathbf{PY}\|_2^2$$

Where  $P_n$  is the set of permutation matrices. Then, the authors use the Wasserstein distance to compute distances, defined as :

$$W_2^2(\mathbf{X}, \mathbf{Y}) = \min_{\mathbf{P} \in P_n} \sum_{i,j} P_{ij} \|x_i - y_j\|_2^2$$

Because neither of the linear transformation nor the one-to-one correspondences are known a priori, the global optimization problem can be characterized as a combination of both problems :

$$\min_{\mathbf{Q} \in O_d} W_2^2(\mathbf{XQ}, \mathbf{Y}) = \min_{\mathbf{Q} \in O_d} \min_{\mathbf{P} \in P_n} \|\mathbf{XQ} - \mathbf{PY}\|_2^2$$

Edouard Grave, Armand Joulin, and Quentin Berthet ([GJB19]) propose an algorithm (Algorithm 4) to solve the optimization problem. Note that for a matrix  $\mathbf{M} \in \mathbb{R}^{d \times d}$ , the projection is given by  $\Pi_{O_d}(\mathbf{M}) = \mathbf{UV}^T$ , with  $\mathbf{USV}^T$  the singular value decomposition of  $\mathbf{M}$ .

---

**Algorithm 4** Wasserstein Procrustes Stochastic Optimization

---

- 1: **for**  $t = 1$  to  $T$  **do**
- 2:     Draw  $X_t$  from  $X$  and  $Y_t$  from  $Y$ , of size  $b < n$
- 3:     Compute the optimal matching between  $X_t$  and  $Y_t$  given the current
- 4:     orthogonal matrix  $Q_t$

$$P_t = \operatorname{argmax}_{P \in P_b} \operatorname{tr}(Y_t Q_t^T X_t^T P)$$

- 5:     Compute the gradient  $G_t$  with respect to  $Q$ :

$$G_t = -2X_t^T P_t Y_t$$

- 6:     Perform a gradient step and project on the set of orthogonal matrices :

$$Q_{t+1} = \Pi_{O_d}(Q_t - \alpha G_t)$$

- 7: **end for**
- 

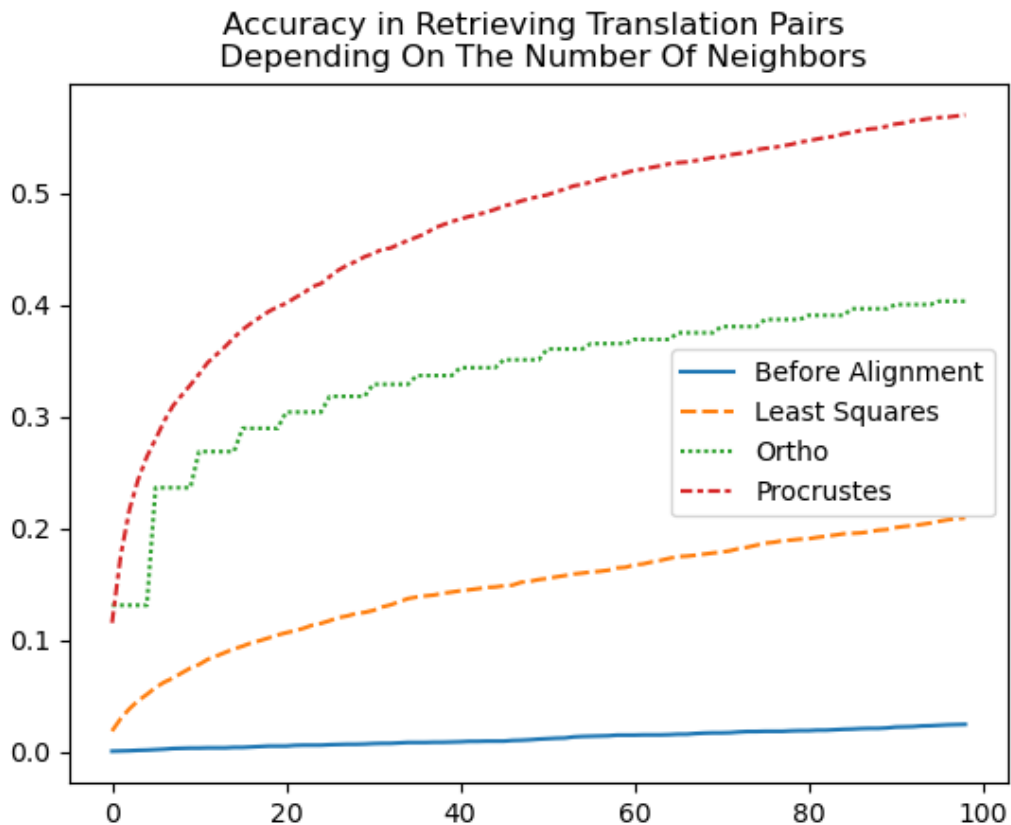
### 5.2.5 Choice of alignment

To discriminate which method of alignment works best for the thesis, algorithms 2, 3, and 4 are tested on the French and American Fasttext embeddings. The French embeddings are projected onto the American ones. For the supervised processes, the training and testing translation data is gathered from the MUSE project [Con+17].

To evaluate the performance of the alignments, the accuracy of an alignment is computed using K-Nearest Neighbors (KNN) for each French word, computing the percentage of times the correct translation falls into the nearest neighbors. Figure 5.3 plots the accuracy of the alignments, depending on the number of neighbors. The three methods are all better than the unaligned embeddings, but the Procrustes method is the one that performs the best. Therefore, all the wordspaces are aligned to the American one, using Algorithm 4.

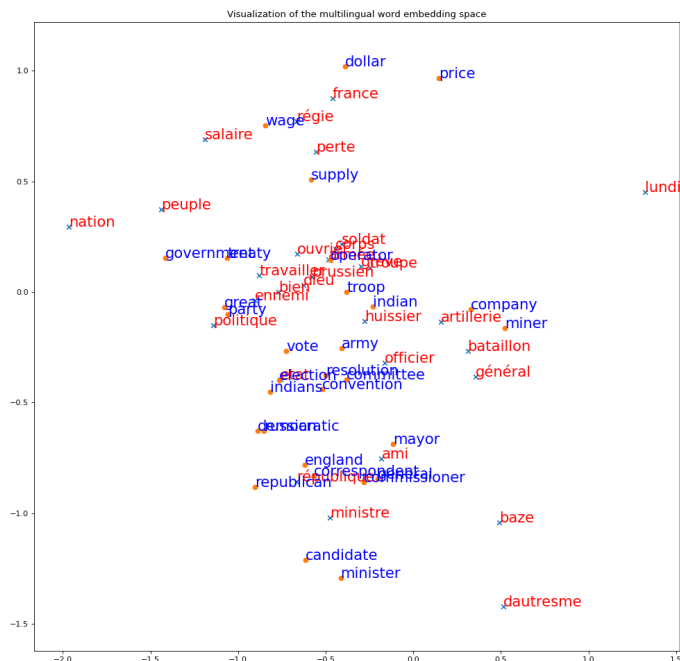
### 5.2.6 Visualisation methods

Finally, to visualise the results of the aligned word embeddings, Principal Component Analysis (PCA) [AW10] is used. It is a multivariate technique that extracts important information from a table to represent it as a set of new orthogonal variables called principal components. In the case of high-dimensional data points, it



**Figure 5.3:** Accuracy of the alignments using KNN to retrieve translation pairs, depending on the number of neighbors

means that PCA is used to find axis that summarize the data the best, which can be used as a projection method in two dimensions. The components are computed such that they maximize the variance of the data. The percentage of variance explained by each of the selected components can assess how well the PCA worked for the data. Figure 5.4 shows a visualisation of the alignment of the Fasttext embeddings using the Wasserstein Procrustes method. Here, the PCA could explain 12% of the variance of the data. It is far from perfect, although it permits to see that clusters of the same notions are indeed close together, for example the words about the government ("minister"/"ministre", "candidate", "republican"/"république") or about money ("salaire"/"wage", "dollar", "perte").



**Figure 5.4:** PCA visualisation of a subset of the aligned word embeddings. The subset consists of words from three french topics (in red) and three american topics (in blue) from the 1870-1880 period

Using Fasttext Word Embeddings for each journal, a cross-lingual space has been created using the unsupervised alignment method Wasserstein Procrustes. In that space, all the words for every articles of the dataset are laying close to words that are semantically close to them, regardless of the language. The comparison of topics across journals becomes now easier, and detailed int the next section.



# 6

## Network of Cross-Lingual Topics in Time

---

*Now that all the multi-lingual words from the corpus lay in the same space, it is possible to compare the topics between the different journals. To proceed, distances are computed using the CLWEM between all the topics. With the distances, a graph is constructed and communities of multi-lingual and multi-temporal topics can be detected, forming clusters of topics that are semantically close and allowing to trace topics recurrence in time and space.*

### 6.1 Creation of Network of Topics

Now that all the words from the vocabulary are represented as vectors in a unified vector space, the goal is to find recurrent topics in time and across countries. In this section, the problem is solved by treating topics as short documents, and finding the closest ones using the previously created word embeddings models.

This procedure is motivated by the observation that two topics might be semantically close, even though they do not have the exact same words. For example, the topics 0 and 1 in the example from Appendix D ("ministro", "gobierno" and "diputado", "presidente") are close to one another, even if they do not have identical words, and both further from topic 6 ("calle", "casa" ["street", "house"]). The distance between topics is computed using the Word Mover's Distance, and used to create networks of topics.

#### 6.1.1 Word Mover's Distance

The Word Mover's Distance (WMD) measures the dissimilarity between two text documents as the minimum amount of distance that the embedded words of one document need to "travel" to reach the embedded words of another document [Kus+15] (Definition 6.1). The topics are compared with each other using this distance.

► **Definition 6.1.** Given  $T = \langle w_1, w_2, \dots, w_n \rangle$  and  $T' = \langle w_{n+1}, w_{n+2}, \dots, w_{2n} \rangle$  two topics, let  $\mathbf{V} \in \mathbf{R}^{n \times n}$  be a (sparse) flow matrix where  $V_{w_i, w_j} \geq 0$  denotes how much of word  $w_i$  in  $T$  travels to word  $w_j$  in  $T'$ . The **WMD** is defined as the distance between the two documents as the minimum (weighted) cumulative cost required to move all words from topic  $T$  to topic  $T'$ .

$$wmd(T, T') = \min_{\mathbf{V} \geq 0} \sum_{i,j=1}^n V_{ij} \|w_i - w_j\|_2$$



### 6.1.2 Network construction

With the WMD metric for the closeness of topics, an undirected multilingual topic-network is constructed. With two arbitrarily set thresholds  $L > 0$  for topics belonging to the same language, and  $L_r > L$  a more relaxed one for topics in different languages, the procedure to create the network is the following way :

1. In the shared embedding space, compute the pairwise distance between every topic
2. Each vertex  $V_{T_i}$  represents a topic  $T_i$
3. Set the threshold  $L_{T_i, T_j}$  to  $L$  if topics  $T_i$  and  $T_j$  are from the same language, set it to  $L_r$  otherwise
4. For every pair of topics  $(T_i, T_j)$ , add an edge between  $V_{T_i}$  and  $V_{T_j}$  if and only if  $wmd(T_i, T_j) \leq L_{T_i, T_j}$

## 6.2 Community Detection

With a network of topics, it is now possible to detect clusters of these topics, to see which countries and period of time are talking about similar subjects. The task is referred to as Community Detection. Communities are clusters of nodes being tightly connected within those groups and weakly connected between them [MMH17]. Different algorithms are taking different approaches to the problems, from optimizing the modularity of the network [CNM04] [NG04] [Blo+08] or using eigenvectors of modularity matrices [New06a], to random walks [PL05], infomaps [RB08] or label propagation [RAK07]. In this thesis, two algorithms optimizing

modularity and well suited for large datasets are tried : the **Clauset-Newman-Moore maximization** and **Louvain method**.

### 6.2.1 Clauset-Newman-Moore greedy modularity maximization

Clauset-Newman-Moore is an algorithm that maximizes in a greedy manner the modularity of a networks communities. The modularity is defined as the deviation of the number of edges within communities from the expected number of edges in a random equivalent network [CNM04]. Mark EJ Newman [New06b] proves that the expected number of edges between two nodes  $v$  and  $w$  with degree  $k_v$  and  $k_w$  in a graph with  $m$  edges is :

$$E[J_{vw}] = \frac{k_v k_w}{2m - 1}$$

From which follows the definition of modularity :

► **Definition 6.2.** For a network of  $m$  edges with adjacency matrix  $A_{vw}$ , with  $k_j$  the degree of node  $j$  and  $c_j$  the community label of node  $j$ , the **modularity** is defined as :

$$Q = \frac{1}{2m} \sum_{vw} [A_{vw} - \frac{k_v k_w}{2m}] \delta(c_v, c_w)$$



The greedy modularity maximization algorithm starts with assigning each node to its own community, and then repeatedly joins pairs of communities that lead to the largest modularity until no further increase in modularity is possible [CNM04].

### 6.2.2 Louvain method

The Louvain method [Blo+08] also aims to optimize the modularity of the communities. It is based on the notion of modularity gain.

► **Definition 6.3.** For a network of  $m$  edges with  $k_j$  the degree of node  $j$ , the modularity gained by isolating node  $i$  is defined as :

$$\nabla Q = \frac{k_{i,in}}{2m} - \gamma \frac{\sum_{tot} k_i}{2m^2}$$



The Louvain method takes two steps. First, like with the Clauset-Newman-Moore optimization, each node is assigned to its own community. Then, the algorithm tries to find the maximum positive modularity gain by moving each node to all of its neighbor communities. If no positive gain is achieved the node remains in its original community [TWV19].

## 6.3 Results and Evaluation

### 6.3.1 Evaluation Metrics

To evaluate the creation of the communities, three different metrics are combined : the modularity, the performance and the coverage of the partitions induced by the communities. The modularity  $Q$  has already been defined in Section 6.2.1, the same definition is used here. It is used as a measure of the strength of the partition. The performance  $P$  of a partition is the ratio of the number of intra-community edges plus inter-community non-edges with the total number of potential edges [For10]. This measures how distinct the communities are. Finally, it is also useful to know the coverage  $C$  of a partition, computed as the ratio of the number of intra-community edges to the total number of edges in the graph [For10]. The suitability of a partition is then computed as a weighted average of these three measures :

$$score = \alpha Q + \beta P + \gamma C$$

where the parameters are set arbitrarily to  $\alpha = 0.70$ ,  $\beta = 0.15$ , and  $\gamma = 0.15$ .

Different thresholds  $L$  and  $L_r$  are tested using a grid search between 0.5 and 2.25 with a step of 0.25, for the two different algorithms (from Sections 6.2.1 and 6.2.2) and for each keyword.

### 6.3.2 Results

#### Grid Search

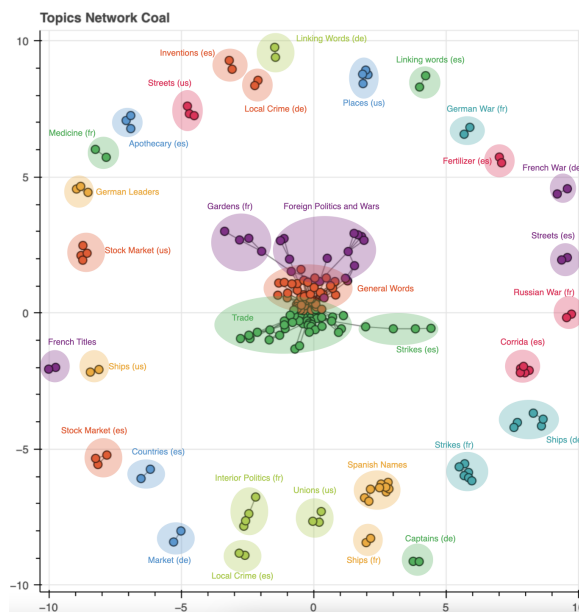
The results of the grid search are summarized in Table 6.1. The Clauset-Newman-Moore algorithm detects communities better for the Telegraph and Coal dataset, whereas the Louvain method is better for the Steel and Electricity corpus. It is also relevant to note that the partitions' qualities differ from keyword to keyword, the communities detected on the Steel data (score=0.492) being significantly better than the partition from the Electricity dataset (score=0.281).

**Table 6.1:** caption

Keyword	Algorithm	Threshold	Threshold relax	Score
Telegraph	Clauset-Newman-Moore	2.0	2.0	0.433
Coal	Clauset-Newman-Moore	2.0	2.0	0.359
Steel	Louvain	1.75	1.75	0.492
Electricity	Louvain	2.25	2.25	0.281

**Networks**

The interactive visualisation of the optimal networks with community labels is available online <sup>3</sup>. Figures 6.1, 6.2, 6.3, and 6.4 also show the networks for each keyword, with names given manually for the communities. A label such as "Corrida (es)" with the country specified means that it is a mono-lingual cluster, otherwise multi-lingual.



**Figure 6.1:** Network of coal discourses during the Second Industrial Revolution, across France, Spain, Italy, Germany and the USA

<sup>3</sup> <https://arobaselisa.github.io/industrial-west/>

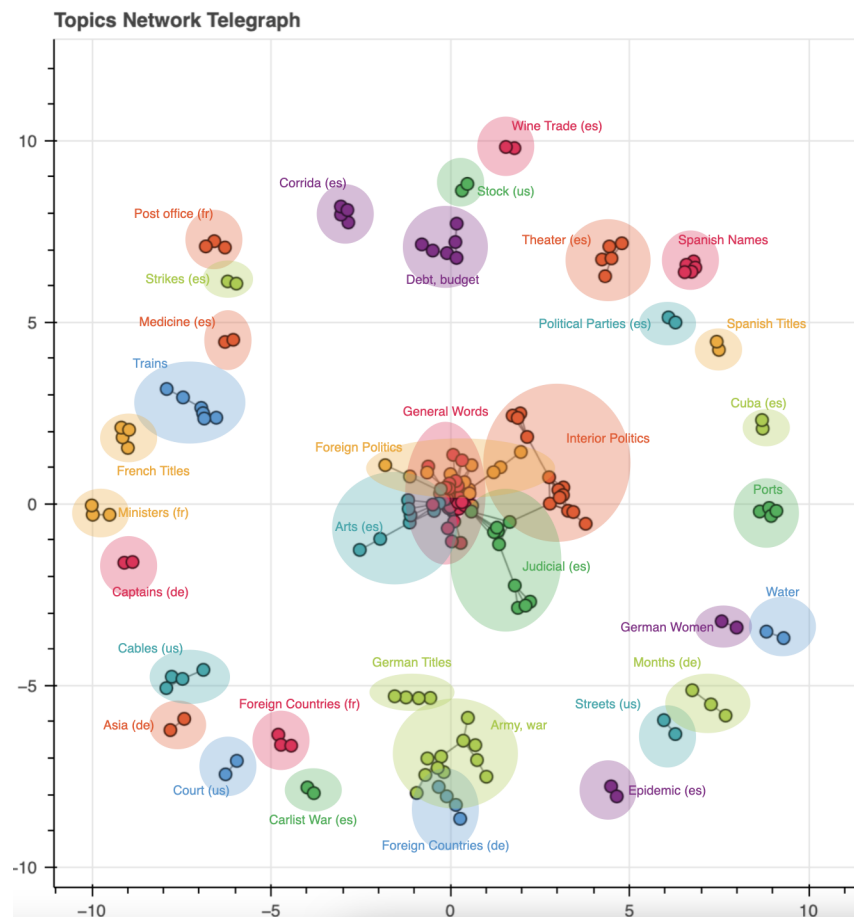
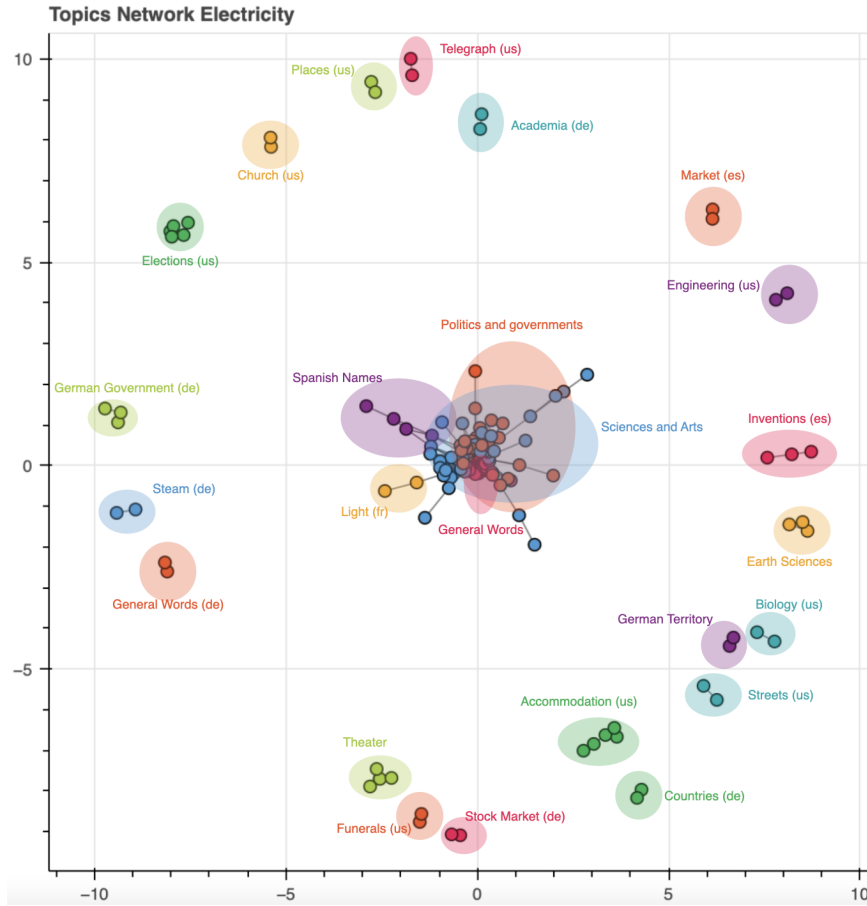


Figure 6.2: Network of telegraph discourses during the Second Industrial Revolution, across France, Spain, Italy, Germany and the USA

To conclude, using Cross-Lingual Word Embedding Models allowed to represent distances between topics, that have been use to create a network. These networks for each keywords have then been used to detect communities of topics, allowing to study topics that are common across countries and time, and also figuring out which topics are unique to specific times and places. The following chapter will take a deep dive into the results from the experiments of this closing chapter.



**Figure 6.3:** Network of *electricity* discourses during the Second Industrial Revolution, across France, Spain, Italy, Germany and the USA

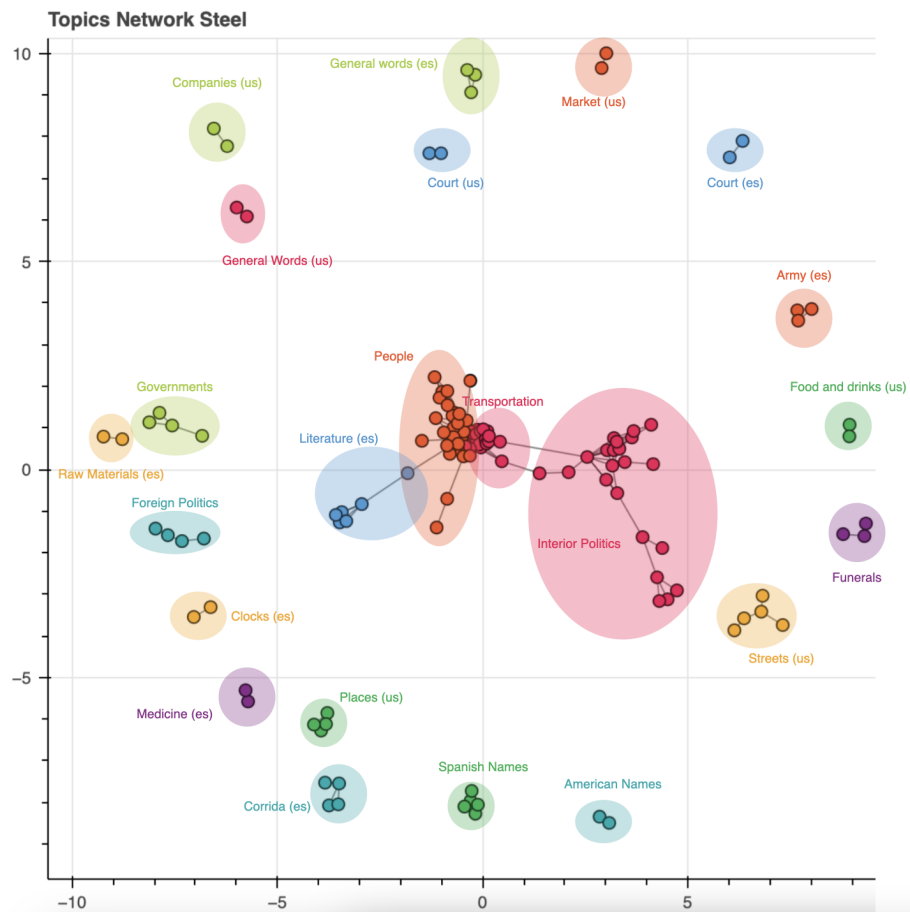


Figure 6.4: Network of *steel* discourses during the Second Industrial Revolution, across France, Spain, Italy, Germany and the USA



Part III

Discussions



*This last part discusses the results for the analysis of the four keywords Coal, Steel, Electricity and Telegraph. The discussions are centered around the results of deep dives into the topic modelling results from Section 4.4, and on the multi-lingual networks of topics from Section 6 (where the clusters results will be referred as "Cluster 'NAME' ").*

## 7.1 Domestic usage

While not novel to the Second Industrial Revolution, coal could be considered broadly speaking a symbol of eighteenth and nineteenth century industrialism across Western societies. It frequently appears in domestic or in medical semantic contexts.

### 7.1.1 Daily life

First and foremost, the analysis depicts coal as it is used by people in their daily life. Coal is mentioned in the articles when used, for example as a pencil for sketches (*"With a charcoal in his hand, he covered the walls of the cabinetmaker's workshop with sketches that were shapeless, but where one could already recognize the organization of a gifted artist."* (*Le Figaro*, 1860-03-01)), or to write on public displays (*"A bad joker had added with coal, at the bottom of the poster: And let us vote for Thiers"* (*Le Figaro*, 1868-09-1)). It is also used as a fertilizer in Spain and France (Cluster "Gardens", "Fertilizer") both at the latest decades of the dataset. All these are examples of the use of coal by private people. But sometimes, coal is also used by Europeans to depict other cultures' daily life as "barbarian", such as for example in French theaters, where an article says : *"Since I attended one of the African performances at the Champs-de-Mars, where they eat frogs and coal, all my illusions about Africa have vanished like magic"* (*Le Figaro*, 1875-10-15).

### 7.1.2 Medicine

Then, coal is also in the conversation of medicine (Clusters "Apothecary", "Medicine"), especially in Spain between 1880 and 1910, and France between 1880 and 1900, with



(a) *Le Figaro*, excerpt from the issue of the 1875-01-10 (b) *The New York Herald*, excerpt from the issue of the 1877-06-16

Figure 7.1:

the topic of Pasteur and its vaccine for the anthrax disease, called "charbon" ("*Mr. Pasteur is going to cure the dog and man of rabies... as he cured the animal and man of anthrax*" (1881-11-23)). The other countries do not touch the subject of medicine with coal.

## 7.2 Industrial use

With the invention of combustion engines, coal becomes a vital resource, produced by miners who are demanding better conditions and go through several crises.

### 7.2.1 A vital resource

Coal becomes a primary resource especially when it comes to heating homes, the article from figure 7.1 (a) talks about a coal merchant that has a lot of orders in winter, which illustrates the demand for coal in 1875 in France. But in many industries, coal is also essential, as explained in a *Le Figaro* article "*First of all, you should know that coal is today as necessary to industry as bread is to man. It has been rightly called "the bread of industry". When coal is lacking, an intense malaise occurs*" (*Le Figaro*, 1875-10-27).

## 7.2.2 The coal production by miners

The people that produce coal are in the very center of the public debate. Occasionally, articles about poor and dangerous working conditions are written ("*Maurice Slmonson, married, badly burned and bruised. His tongue and the root of his mouth are scorched and several fingers and a wrist are broken. He is in a critical condition*" (*The New York Herald*, 1875-12-30)), and more frequently, deaths on the workplace are being reported in the journals ("*Five workers killed and three wounded Parts 18. - A report from Dortmund that during unloading work at the Schelewig Wrake mine there was an explosion of gas from a pile of coal, resulting in the death of five workers and the wounding of three others.*" (*El Imparcial*, 1924-09-19)). Because they are essential to every industry that rely on coal, when striking for higher wages, it paralyzes the whole economy (Cluster "Market") and it is perceived through articles with headlines like figure 7.1 (b), or seen as negative for the economy for example ("*I do not know if the law on strikes will be modified, but it is certain, on the one hand, that it will ruin the French mining industry for the benefit of foreign coals and, on the other hand, that it will not result in any serious improvement in the health of the industry.*" (*Le Figaro* 1893-10-20)). In these articles, coal as a merchandise is seen as good, but the workers are talked about in a negative lens. Geographically, the strikes are talked in the entire Italian and German dataset, since 1860 in the US, 1870 for France, and 1890 for Spain (Clusters "Unions", "Strikes").

## 7.2.3 Coal crisis

The five countries are going through different coal crisis, sometimes due to the strikes in the mines that lead to a shortening in stock ("*The strike continues. Everything is stopped, except in Cransac. Since 1878, we had never seen such a movement. The forge is going to run out of coal incessantly.*" (*Le Figaro*, 1886-02-28)). However, the trade is now global as the following article shows : "*the conclusion of a treaty between the Americans and the Japanese. Some particulars, extracted from the Friend of China, will be found below. The Americans have obtained two ports for trading, and a coal station. Japan yields plenty of coal, and it will be brought from the mines for the use of steamers.*" (*The New York Herald*, 1854-07-02)). That makes the crisis particular because it is all interconnected between the industrialized countries (Clusters "Trade", "Countries", "Market", "Ships").

Le bois et le charbon ne coûtant rien à nos bons fonctionnaires, puisqu'ils leur sont fournis aux frais des contribuables, ces messieurs en profitent pour faire des feux d'enfer dans les cheminées administratives.

Depuis le commencement de l'hiver, les pompiers n'ont pas eu à se déranger moins de *vingt et une* fois pour éteindre les feux de cheminée signalés dans le seul palais du Luxembourg.

C'est là, croyons-nous, un record.

**Le Prix de la Vérité**

Je ne sais si le royaume du ciel est vraiment aux pauvres d'esprit et si la paix sur la terre est aux hommes de bonne volonté, mais je sais bien que le charbon n'est pas pour les gens de bonne foi.

J'ai rencontré une honorable dame, fort exaspérée, et qui m'a dit :

— Je me suis conformée aux règlements, moi ! Bien à contre-cœur, d'ailleurs. Ayant dans ma cave sept cents pauvres kilos de charbon, j'avais dit à mon mari : « Annonce donc n'importe quoi ! » Mais mon mari a tenu à déclarer la vérité. C'est un pauvre homme qui ne veut jamais se mettre dans son tort. Aussi, nous avons tout de suite été récompensés comme il convient. Il nous est venu de la mairie une belle feuille blanche, sur laquelle on lisait :

« Comme suite à votre déclaration, j'ai l'honneur de vous faire connaître qu'il n'y a pas lieu, pour le moment, de vous délivrer une carte de charbon. Le chef de ménage devra présenter une nouvelle déclaration, quand le motif qui s'oppose à la délivrance de la carte aura cessé, c'est-à-dire dans trois mois ! »

Dans trois mois ! vers la mi-décembre ! Nous aurons eu le temps de crever de froid jusque-là, avec nos sept cents kilos de charbon ! Car ce charbon, monsieur, c'est du tout-venant, les trois quarts en poussier, qui ne prend que quand ça veut, à l'aide des quelques morceaux avouables qui y sont mêlés, et qui ne prendra plus du tout, quand ces quelques morceaux seront brûlés.

(a) *Le Figaro*, excerpt from the issue of the 1902-01-22 (b) *Le Figaro*, excerpt from the issue of the 1917-09-24

Figure 7.2:

### 7.3 A source of strategy and tensions

The crisis eventually lead the coal to be a source of tensions and strategies, both inside and in between countries.

#### 7.3.1 The control of resources inside countries

Inside the countries, different approaches are chosen by the different governments and companies (Clusters "Interior Politics"). In the beginning of the XXth century, during the coal crisis, they try to control the amount of resources that any inhabitant can have, as explained in the article in Figure 7.2 (b), or as the article from *The New York Herald* explains "The situation requires centralized responsibility and action" (*The New York Herald* 1922-08-30). It sometimes becomes a source of tension, with for example unfairness of the share felt in the article of Figure 7.2 (a). In Italy, they go as far as portraying people who save coal as heroes ("Here they are gathered around the cannons the true heroes of the war : [...], the modest clerks who did not burn coal this winter because it was too expensive, [...]") (*La Stampa*, 1915)

### 7.3.2 Coal diplomacy and trade between countries

At last but not least, as coal becomes a merchandise in the international market, it occupies an important role in relationships between countries, and the two large clusters "Foreign Politics and Wars" and "Trade" are a demonstration of that. It is a subject that concerns all the five countries, in every decade of the dataset. The coal industry becomes very competitive worldwide (*"all our miners are able to compete indiscriminately, so that the national coals can compete with advantage in quality and price with foreign ones."* (*El Imparcial*, 1867-04-10)), and coal is used as an argument in diplomacy, as indicates the German article *"If, through the fault of the German authorities or industrialists, the production should fall below the present figure, all coal shipments to Germany might be prevented. [...] The German government considers any further discussion of the purpose of the Franco-Belgian invasion to be superfluous. It can only express its astonishment that the French government still believes it can deny the character of its action, which is openly known to the world."* (*Neue Hamburger Zeitung*, 1923-01-20 DE).

To conclude about coal, what stands out from the analysis is its representation in the daily life of every Western society, that also becomes an essential material to many industries, such as the transportation. Some specific themes like the use of the term "coal" in medicine articles are individual to some countries. In parallel, the strikes in mines are a very present subject related to coal, in the five countries, starting a only few decades apart. In general, the analysis allows to see that similar discussions around coal are happening in parallel across countries, or a few decades apart.





---

*The following part discusses the different discourses found during the analysis, how steel is first used as retail, to its industrialization.*

## 8.1 Steel's retail applications until the 1880s

In the articles, steel is often discussed along with tools or objects, not being the main subject. It is particularly seen in the beginning of the dataset, between 1850 and 1880, when talking about engravings, fashion, or objects from the kitchen.

### 8.1.1 Art

Steel is mentioned for the engravings in books (Cluster "Litterature"), such as one that *"is the most complete book of its kind, containing one hundred maps and numerous engravings interspersed in the text, which contains more than 1,200 pages, not counting the chromo types and steel engravings outside the text."* (*Le Figaro*, 1875-12-20) or *"splendid steel plates"* (*The New York Herald*, 1868-11-03) in America. It is more generally a popular medium in art, and articles mention the artworks made with steel for example in *The New York Herald*, *"beautifully illustrated with superb portrait on steel, executed by the holographic process in Paris."* (NY 1855-08-11). The steel is described to the public as a valuable material, for example with the tools that are used to produce art *"Most of these tools, such as gravers and polishing ropes, were of the finest steel and had cost a lot of money"* (*Neue Hamburger Zeitung*, 1889-08-22).

### 8.1.2 Fashion

The other sector that uses steel at the end of the XIXth century is fashion. In the French newspapers, one can read *"Steel satin toilet with steel embroidery."* (*Le Figaro*, 1881-06-13), which is often found in some luxurious clothes. Jewellery also uses steel, as reveal some of the advertisement in the journals, *"Suspender Trimmings. Gold, Silver and Steel Beads, in all Nos. Mother of Pearl, Gold, Silver and Steel Purse Ornaments"* (*The New York Herald*, 1843-08-06). Additionally, *"Blue Steel"* is a color

that was very popular in the 1840s, with 15 different editions of *The New York Herald* mentioning it, out of 22 for the whole dataset.

### 8.1.3 House items

Finally, the last use of steel by individuals is a domestic one. The steel is found in a lot of kitchen objects (Cluster "Food and Drinks"), and stands out in the advertisement section in the United States in newspapers (See Figure 8.1).

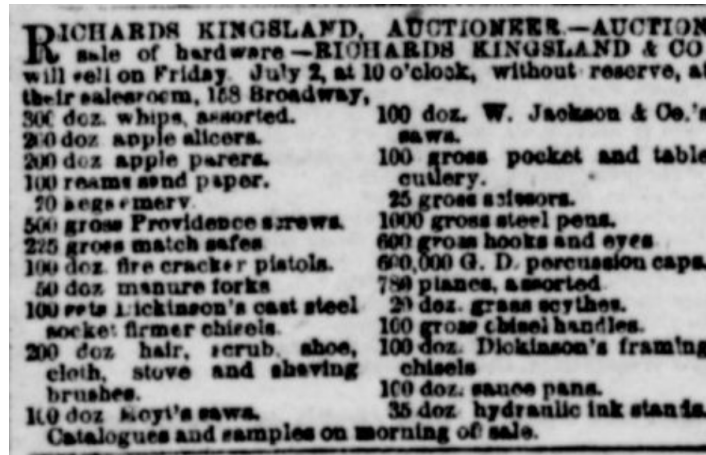


Figure 8.1: *The New York Herald*, excerpt from the issue of the 1858-07-01

## 8.2 The steel industry

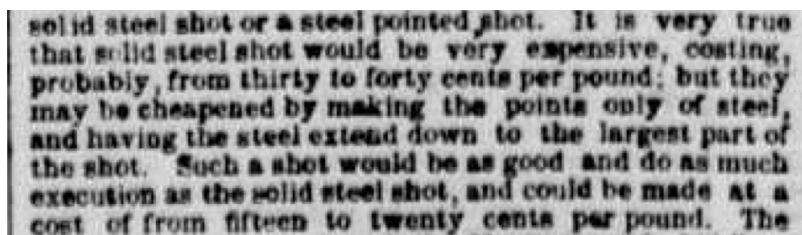
The topics illustrate how, as steel becomes cheaper to produce with the arrivals of new technologies and processes, an enthusiasm is built around it. After the 1880s, the majority of articles about steel becomes related to the industry, in full bloom. It becomes an essential war resource and is involved in world-wide trades, while also becoming a sort of indicator of progress in a society, according to the newspapers studied.

### 8.2.1 Steel involved in world-wide trade

After the different discoveries of steel processes discussed in section 2.2, steel becomes a commodity and the steel industry is expanding (Clusters "Market", "Raw Materials"). In news reports, it goes alongside other types of merchandise, "It is

followed by grains, with 1 million 200 000 tons ; then by oil mineraae, oilseeds, cotton, iron and steel, phosphates, nitrates, timbers, wools, jute, hides, hemp, etc." (*La Stampa*, 1928) or "For Mesnug, Swab, Tin, Nickel, Steel, Knives too., Silver and. Gold. Packs 10 and 20 A" (*Neue Hamburger Zeitung* 1891-11-19). The steel becomes a monitored good "Steel trust stocks were 3 dollars lower, while Betblebems were \$7 higher. At the close of the stock market, the attitude was irregular moderate" (*Neue Hamburger Zeitung*, 1915-11-19).

### 8.2.2 Steel as a war resource



solid steel shot or a steel pointed shot. It is very true that solid steel shot would be very expensive, costing, probably, from thirty to forty cents per pound; but they may be cheapened by making the points only of steel, and having the steel extend down to the largest part of the shot. Such a shot would be as good and do as much execution as the solid steel shot, and could be made at a cost of from fifteen to twenty cents per pound. The

Figure 8.2: *The New York Herald*, excerpt from the issue of the 1862-03-18

Steel is also very present in the news during war periods (Cluster "Army"), especially when governments run out of it and need to import from other countries (Cluster "Transportation", "Companies"). The idea of using steel in war is first evoked in the dataset in American newspapers, but seem very costly at first (See Figure 8.2). Later, with the technical advances, it is massively used in weapons, and the Italian articles are focusing a lot on these, for example talking about the "incendiary howitzers [that] then turn into flashlights, reaching the distance of several kilometers. The steel of the howitzer itself liquefies, causing a brazier, impossible to extinguish" (*La Stampa*, 1916). The other war-related industry that profits a lot from the steel process amelioration is the naval industry, working together hand in hand "strict limitation of naval armament by the chief naval Powers of the world are being: considered thoughtfully here by local representatives of steel and shipbuilding industries." (*The New York Herald* 1921-11-16). Quickly, steel weapons become an indicator of progress for any eurocentric industrialized societies, being sometimes the first to be mentioned by its contemporaries, "it is enough to note the progress of civilization by mentioning four names that sum up contemporary history: Krupp, the inventor of the steel cannon; [...]" (*Le Figaro*, 1875-10-15).

To summarize the discussion around steel, the term appears in the newspapers

discourse in art, fashion and house items before it is becomes a widely industrialized resource, used mainly in war. It stands out from the analysis that, when compared, there is a certain homogeneity of discussions around steel across the Western societies studied.

*The analysis for electricity highlights three main different points : the phenomenon of discoveries around electricity is a public one, the feeling by the general public that they witness a technical revolution, and the industry of electricity.*

## 9.1 Publicity of the scientific discoveries

### 9.1.1 Scientific discoveries shared through newspapers

The analysis shows that the scientific discoveries around electricity were shared through the newspapers, and met with excitement (Clusters "Sciences and Arts", "Innovations", "Engineering"). The article from *El Imparcial* from 1879 in figure 9.1 (a) explains how a thermal regulator for each light-producing wire has been invented, along with a praise of the scientific method. These clusters have topics that belong to the earliest decades in the dataset, before the end of the XIXth century, and in every country studied, showing a relatively homogeneous discourse in these aspects.

Nada, pues, mas barato que la luz de las máquinas magneto-eléctricas: nada menos dependioso que un sol artificial de la intensidad de 4.000 lamparas-carcel, pero la divisibilidad de foco tan de-lumbrador se creía un delirio en 1877, utilizando los medios que la ciencia entonces poseía. Y, sin embargo, ese imposible de ayer es hoy posible; esa pretendida utopía es ya una realidad, y no por la virtud de principios ignorados, sino por la sabia combinación de verdades, acerca de las cuales no cabe en la ciencia lugar a vacilación.

Si una corriente intensa de electricidad, en marcha por un buen conductor de gran grueso, pasa de repente á un hilo muy delgado y de mucha resistencia relativa, como por ejemplo, á un alambre de platino, de iridio ó de otra sustancia difícil de fundir, entonces el alambre delgado se pone incandescente y se hace luminoso. Repítase esto mismo varias veces, y ya está hecha la division de la luz.

Pero ¿quién puede garantir que no crezca de pronto la intensidad de la corriente eléctrica, por aumento rápido de produccion de electricidad en el generador? Y en tal caso, ¿no se fundirán irremisiblemente los hilos?

L'accumulateur électrique, on ne l'ignore pas, n'est autre chose qu'un récipient d'électricité, dans une petite boîte d'un pied de long sur six pouces de large. On emmagasine l'électricité comme sous une cloche, on condense le gaz. C'est M. Planté, qui a inventé l'accumulateur, perfectionné et rendu pratique par M. Faure.

L'électricité emmagasinée, on peut la transporter n'importe où, et la première application pratique de cet admirable agent de force et de lumière a été faite sur la ligne de Londres à Brighton, en présence de toute la presse anglaise, du ministre de la guerre, de sir Arthur Ottway; C. Edwards, membres du Parlement; C. Seymour Grenfell; colonel Moncrieff; monsignor Capel; M. de Calo, ingénieur et de tout ce que Londres renferme en ce moment d'illustrations.

(a) *El Imparcial*, excerpt from the issue of the 1879-02-07 (b) *Le Figaro*, excerpt from the issue of the 1881-12-04

Figure 9.1:

### 9.1.2 Public experiments and exhibitions

Electricity as a scientific discovery and the ones that followed from it were very public. There were public experiments, for capacitors in France for example, where politicians and the press were invited (article in Figure 9.1 (b)). There were also exhibitions to show the latest discoveries, such as the "*Exposition internationale d'électricité*" [International Exhibition of Electricity] (*Le Figaro*, 1881-11-04), that kept the public curious and involved in the innovations surrounding electricity.

### 9.1.3 Praise of scholars

The scholars themselves (Cluster "Academia") are described as heroic figures, praised ("*A politician counts for little in Paris at this time, where, one after the other, the great scientific glories come to visit us. Between Edison, who is leaving us, and Nordenskiöld, who is arriving, a simple minister would have made a rather poor showing; what does an Excellency weigh next to the man who is trying to wrest the last mysteries from electricity*" (*Le Figaro*, 1889-09-14)). These topics are predominant in Germany and in France.

### 9.1.4 Electricity : mystical and religious

For the other side of the ocean, in the United States, electricity is seen through some articles as a religious or mystical force (Cluster "Church"). This theme is only found in the United States, but is consistent through the years. From "*The great agency of the Almighty, in producing the phenomena of the physical world, is electricity*" (*The New York Herald*, 1849-05-30) to "*there is in spiritual matters some such wave as extends, we know not how, in light, in electricity and in heat, so that, for instance, when the faculty of mirth exerts itself in me it sends an electric wave to every man who has a corresponding endowment of mirthfulness, and he feels that wave from my sou vibrate upon his*" (*The New York Herald*, 1878-12-02).

### 9.1.5 Detractors and critics

Finally, the public is not always an avid supporter of electricity. Several articles recount their disgust for electricity and the new technologies in general, "*The reader of these lines must not fail to curse progress and miss the times I reached, and which he undoubtedly does not know from experience; I, too, sometimes wish for a period of half a dozen years like those I still remember, so that the unconscious blandishers of this great century would glorify progress*" (*El Imparcial*, 1874-02-18). Sometimes, articles

are simply stating the dangers of electricity and that the number of deaths by electricity is increasing (Cluster "Funerals") ("*The machine for artificial respiration to those struck by electric current. While the accidents practiced by electricity mark a continuous increase, [...]*") (*La Stampa*, 1928).

## 9.2 The revolution of electricity

The topics given by the analysis highlight a consciousness from the public that electricity is truly a revolution of their time, important by its speed and implications into many areas of their lives.

### 9.2.1 Fast human progress

The speed of progress is also discussed in the articles, for example in Spain where the progress and new policies around electricity are prioritized and everything goes fast ("*we are in the century of electricity and steam, it is not surprising that in only four days a project as important as that of the conversion of the electricity and steam power plant has been discussed and voted in the Congress, it is not surprising that in only four days a project as important as that of the conversion of the electricity and steam power plant has been discussed and voted in the Congress.*") (*El Imparcial*, 1867-07-05)). This rapid technological progress is a vector of hope for the future, to replace old and inconvenient technologies ("*the tramway will circulate illuminated and moved by electricity, provided by the Faure accumulator, and that I hope, in a short time, to have at home my small accumulator, allowing me to work in the evening without having the eyes burned by the light of the gas, or without being obliged to unceasingly go up this deplorable oil lamp which is always extinguished at the precise moment or one needs it.*") (*Le Figaro*, 1881-12-04) (Cluster "Light"). In *Neue Hamburger Zeitung*, they go as far as saying electricity surpassed the Seven Wonders ("*The seven wonders of the ancient world have long since been surpassed by the achievements of our time and can only claim a purely historical interest compared to the modern wonders created by steam, electricity and architecture.*") (*Neue Hamburger Zeitung*, 1889-12-12))

### 9.2.2 Implications of electricity in different areas

The rapid progress brings to light new questions around the applications of electricity. From discourse about its use in death penalty - mostly in the United States (articles in Figure 9.2), to new techniques in science and medicine (Clusters "Earth



**CAPITAL PUNISHMENT.**

**ELECTRICITY IN PLACE OF HEMP—VIEWS OF A MEDICO-ELECTRICIAN ON THE PROPOSED SUBSTITUTE FOR HANGING.**

The subject of carrying out the death penalty through the medium of electricity and of thus substituting a more humane, effective and impressive agency of the law than the present one continues to form a theme of discussion among many thoughtful people. Dr. George M. Hearl, of Twenty-ninth street, well known by his studies in electricity as applied to nervous diseases, was interviewed yesterday by a HERALD reporter.

"What do you think, Doctor," asked the reporter, of the idea of using electricity in place of hemp for carrying out the death penalty?"

"I have thought much of the matter for years and believed that the time would come when such a method of execution would be employed."

"What advantages would you say it possesses?"

"The instantaneousness with which death is accomplished. A man killed by lightning never knows what he dies of, and should we succeed in obtaining an electrical apparatus of sufficient power, and one that could be readily and unerringly controlled, this style of execution would prove the most delicate possible, while inspiring in the multitude a much greater degree of awe and reverence for the law. It is the general belief that drowning is the best and easiest death to die, and that men who so perish die happy, but the first choking sensations in the process of drowning are extremely disagreeable. A man killed by lightning can have no sensation, for in the small fraction of an instant he is as though he had never been."

**LA PEINE DE MORT**  
PAR L'ÉLECTRICITÉ

Les Etats-Unis attendent avec une extrême curiosité le résultat que fournira une récente et capitale invention dont l'essai sera prochainement tenté. Capitale est bien le mot juste, puisqu'il s'agit de l'appareil électrique qui, remplaçant les moyens grossiers usités jusqu'à ces derniers temps, servira à donner la peine de mort. D'un jour à l'autre, l'application de ce châtiment perfectionné sera faite à Auburn, sur un assassin nommé Joseph Kimmler. On ne nous dit pas si ce malheureux se montre flatté de l'honneur grand qui lui est réservé d'inaugurer cette expérience...

A part quelques formalités légales qui restent à expédier, tout est prêt. Les générateurs électriques, dont la puissance mortelle a été éprouvée par divers animaux de grande taille, sont dûment installés. Déjà, comme chez nous, quand une première à sensation se prépare, les journaux américains foisonnent d'indiscretions. C'est à l'un d'eux que nous empruntons les détails suivants :

Voici d'abord comment s'écouleront les dernières heures du condamné. Il sera fixé sur son sort le matin même du jour de l'exécution. S'il en manifeste le désir, des consolations religieuses lui seront accordées. Aussitôt terminée cette suprême formalité les agents de

(a) *The New York Herald*, excerpt from the issue of the 1879-11-26 (b) *Le Figaro*, excerpt from the issue of the 1889-07-16

Figure 9.2:

Science", "Biology"), or for transportation (Clusters "Engineering", "Steam"), electricity is also viewed as an opportunity to redefine broader notions such as nations or journalism (Cluster "Telegraph").

### 9.2.3 Electricity to justify power

Electricity thus becomes a symbol of human progress (*"The great principle which electricity has written on the hitman mind in the word progress, and which are being diffused over the world by steam and the railroad, as the essentials of enlightenment and civilization"* (*The New York Herald*, 1868-07-01), and any culture that do not have it is seen as ancient, and justifiably under the domination of the ones that posses it (*"When a people predominates in such an obvious way in the great manifestations of noble human activity, it ends up creating an intellectual hegemony which, sooner or later, will become political"* (*Le Figaro*, 1889-06-07)).



## 9.3 Industry of electricity

The analysis makes the discourses around the industry of electricity emerge. The articles are seeing the very new industry being built, while electricity becomes an essential resource and its governance sparks debate.

### 9.3.1 The essential merchandise

Electricity is integrated into the world-wide market relatively quickly (Clusters "Market", "Stock Market"). While in the early 1900s France has hopes that this industry will be beneficial for them (*"France will be saved by electricity"* (Le Figaro, 1901-12-14)), it is also the period in Europe where several strikes occur within the industry, by workers demanding greater wages, denying electricity and showcasing how important electricity and their work, is (*"the conduct of the electricians had caused a very great disturbance in the life of Paris"* (Le Figaro, 1907-03-10)).

### 9.3.2 Private or public governance

Electricity discourse finally revolve around who controls its use and trade (Cluster "Politics and Governments"). In the United States for example, in the 1860s the monopoly of electricity by telegraph companies is criticized in public discourses (*"In the first place it should be known that the lines of the Western, Union and American telegraph companies are such lines as are leased and controlled by them, and are the consolidation of the oldest lines of telegraph tn this country. Until a very recent date they have had a complete monopoly of the use of electricity in the conveyance of intelligence, not only for the public and the press, but, since the war began, for the government."* (The New York Herald 1865-06-07)). European governments like in Spain also try to manage the use of electricity (*"The minister seems to have informed them of the complaint formulated by the French ambassador in favor of Mr. Touchet as a result of the measures that the City Council has taken regarding the electricity factory and for turning on the street lights against Touchet's will. "* (El Imparcial 1898-11-08)).

The analysis of electricity discourses shows that, even if the main ideas are common to the five countries France, Spain, Germany, Italy, and the United States, there are some geographical nuances to the debates, where for example spiritual references are more present in the United States than everywhere, or where the praise of scholars is more present in Germany and France.



Finally, this last section discusses the results of the analysis regarding telegraphs, and deep dives into some of the articles representative of certain topics.

## 10.1 The Telecoms Revolution

The technical revolution around telegraph is displayed when looking at the discourses around them, while being consistent throughout the years and across the different regions studied. The discourses also show that the changes coming with the telegraph have clear impacts on the daily life in these societies.

### 10.1.1 Speed, ingenuity, and enthusiasm

The telegraph is usually met with enthusiasm in the newspapers of the five countries. Although the enthusiasm differs from the period of time and the subjects, it is still a steady constant in the articles. In early times, the enthusiasm is carried by the scientific community, in the United States (Figure 10.1 (a)), with inventors claiming patents to what they see will be a big discovery (Cluster "Court"). In parallel in Spain, telegraphs proved themselves to the public by carrying the stock market news with an efficiency that was never seen before (Cluster "Stock"), because *"for the first time since the invention of telegraphy, the closing reports of the London and Liverpool Stock Exchanges were published in New York, not only on the same day, but also a few hours before the event. Thus, a dispatch that could not be dispatched from London before four o'clock in the afternoon, saw the light in the American newspapers at noon. The telegraph had anticipated the sun four hours soon"* (El Imparcial, 1867-03-28). The incremental discoveries are later also a source of enthusiasm. For example in France in 1881, (Figure 10.1 (b)) the discoveries related to the telegraph are shown to the public through exhibitions (Cluster "Foreign Countries", Cluster "Arts"), and seen as great for the cost-reduction it implies. Finally, when the telegraph technology is already well implanted, the speed and ingenuity does not cease to impress thanks to the heir innovation of the telegraph : telephony, talking about *"the miracle of wireless telephony and telegraphy, the word spoken into a telephone in Hamburg or the written word transmitted to the Morse apparatus actually finds the*

**TELEGRAPH PATENTS.**

TO THE EDITOR OF THE HERALD:—  
 No little interest has been excited in telegraph and other scientific circles from the published claim of the late Prof. Page, of Washington, to the merit of the discovery of the powers of the electro-magnetic induction coil, commonly known as the "Ruhmkorf coil," and for which M. Ruhmkorf was awarded a prize of fifty thousand francs by the French government. Prof. Page's claim was pre-

Du côté français, on remarque :  
 L'Exposition du Ministère des postes et télégraphes. Un pavillon beaucoup plus vaste et plus complet que celui exposé par le même ministère au Champ-de-Mars, en 1878. On y admirera les appareils nouveaux qui nous présagent une nouvelle diminution des taxes télégraphiques. Par l'ancien télégraphe à cadran, on ne pouvait expédier que 500 mots à l'heure; par le télégraphe Morse, on peut en expédier 1,000; par les nouveaux systèmes, 4,500. On a trouvé aujourd'hui un système duplex qui permet d'envoyer en même temps deux dépêches par le même fil, soit 9,000 mots à l'heure, et M. Edison exposera pour la première fois un appareil quadruplex qui donnera une moyenne de 18,000 mots.

(a) *The New York Herald*, excerpt from the issue of the 1868-05-22

(b) *Le Figaro*, excerpt from the issue of the 1881-08-09

Figure 10.1:

*same moment everywhere on the surface of the earth, provided that the power of the electrical excitation is strong enough and that no special disturbances occur" (Neue Hamburger Zeitung, 1924-05-02).*

### 10.1.2 Reduction of distances

The enthusiasm is generally accompanied by a feeling of distance reduction, as the little poem from the United States from Figure 10.2 (a) shows (Cluster "Foreign Countries"). The feeling does not disappear with time, in May 1912, an article of *Le Figaro* puts it into words : *"What a mystery, all the same, this wireless telegraphy... \* — Yes, there are no more obstacles, there are no more distances, and their motto is correct: T. S. F. "Tout Se Franchit" ["Everything can be crossed]" (Le Figaro, 1912-05-18).* It is also supported by the fact that throughout the decades and across all countries there are topics falling under the cluster of "Foreign Politics", where the news are coming from all over the world by telegraph.

### 10.1.3 Integration in daily life

Studying the telegraphs technology in the medias in time shows how quickly the telegraph is included into daily life. They are used to talk about any interior or exterior politics all the time (Clusters "Interior Politics", "Foreign Politics"). This was reflected in the dataset by a small sentence stated that the information had been received by telegraph, like *"POR TELEGRAFO"* in the Spanish dataset in which 114 articles are using this sentence, Figure 10.2 (b) is one of them. They become

What a wonderful thing this Telegraph is, indeed! Think of a long speech being delivered in the heart of old Kentucky on one day, and published in the chief city of the Empire State on the next morning—the distance being over eleven hundred miles! Forty years ago a month would have been considered a quick trip between the two points.

Ah! these little “clicks” of the Telegraph—

Though they breathe not a word,  
Their voices are heard  
At a distance no voice could reach;  
And swiftly as thought  
The words are brought,  
And the lightning endowed with speech!

Though seas roll between,  
And lands intervene,  
The absent are close at hand;  
The eye seems to hear,  
And space disappear,  
And Time is compelled to stand.

LAS CONGREGACIONES FRANCESAS  
 POR TELEGRAFO  
 (DE NUESTRO CORRESPONSAL)  
 Paris 2 (10 noche)  
**Protesta de los jesuitas**

Los padres jesuitas provinciales de Tolosa, Paris, Champagne y Lyon han dirigido á los periódicos una larga carta que publican hoy todos los de la tarde.

Constituye esa carta una interesante declaración, y en ella explican los jesuitas los motivos que les impiden solicitar la autorización con arreglo á la nueva ley sobre las congregaciones.

—Esa ley de excepción—dicen—viene á herirnos profundamente en nuestros derechos más esenciales de hombres libres, de ciudadanos, de católicos y de religiosos.

«Nosotros no podemos autorizar con nuestro consentimiento una ley que viola derechos imprescriptibles de la Iglesia.»

Terminan diciendo que, á pesar de la rigurosa resolución que se les obliga á tomar, no conservan ningún rencor contra los que les condenan.—Mar.

- (a) *The New York Herald*, excerpt from the issue of the 1847-11-27
- (b) *El Imparcial*, excerpt from the issue of the 1901-10-03

Figure 10.2:

really quickly integrated into the news infrastructure, so any subject that they would usually talk about is represented, such as the newest theater play, or Corrida (Clusters "Theater", "Corrida"). Another element that support the claim of the telegraph becoming an everyday object is that it becomes being mentioned in advertisement a lot more in Germany around the 1890s, for example with the advertisement for "electric bells, as well as suitable materials for telegraph and telephone systems. Latest price list is given to resellers and installers free of charge." (*Neue Hamburger Zeitung*, 1897-04-15). A last example of the involvement of telegraphs in these societies is its use for weather announcements, that is visible with the Cluster "Water" for Spanish articles.

## 10.2 In wars and politics

The analysis also shows that a wide set of articles are discussing the telegraph and its use as a tool in war and politics.

### 10.2.1 Communications

First, the political life of the country is made public with the use of telegraphs, with articles talking about wars (Clusters "Army, war", "Foreign Politics"). The topic is discussed in every country through every year, as the Cluster "Army, war" illustrate. The model also caught up on more specific war topic, with for example

BOSTON—Arr May 28. P M, ship Kate Howe, Norcross, NOrleans; bark David Nickels, M'Gilvery, Matanzas, 13th inst; brig Caroline, Harding, Savannah; schr Clarinda, Goodwin, York. Old brig Mary Perkins, Nickerson, Philadelphia.  
 Arr 29th, ships Strabo, Cutter, NOrleans; Abalino, Elliott, Mobile; barks J A Hazard, Gardner, Havana, 13th inst; Mary Smith, Smith, NOrleans; Saxony, Howes do; Bay State, Dill, Baltimore; Elm, Taylor, Philadelphia; brigs Sabao, Means, Cardenas, 13th inst via Holmes' Hole; Sarah Williams, Gott, Cienfuegos, 11th inst; Olanda, Dunbar, Darien; Marshall, Ryder, Savannah; Candace, Matthews, Baltimore; Eolus (of NYork), Small, Philadelphia; Erie, Baxter, do; Emma, Baker, do; schrs Vine (Br), Clements, Curacao, 24th ult, via Portland; Lebanon, Drinkwater, Darien; Mary Eliza, Ware, Georgetown, SC; Dirigo, Ober, do; Neptune, Robinson, Philadelphia; White Squall, Aumack, Alexandria; Copia, Sears, Philadelphia; Philadelphia, Perkins, Troy; Perseverance, Terry, Albany; O H Perry, Bullock, NYork; Ocean Star, Thorndike, do; Bay State, Burr, do. [Several of these were incorrectly reported by telegraph.] Sld 28th, steamer City of New York; 29th, schr General Washington.

Agency of Railroads and the Telegraph in War.  
 The present war in Europe shows in a remarkable manner the agency of railroads and the telegraph in the operations and result of such a conflict. This mighty power of our times determines, it may be said, the victory on the side of the belligerent that knows how to use it best. We have heard and read much of the art of war, and, no doubt, it is a great art; but in this age war owes a great deal to scientific discoveries and appliances. Great generals in all ages have understood the importance of celerity of movement and the rapid concentration of a large force at a given point. To strike before the enemy was ready, and with overwhelming numbers, has been the strategy of all celebrated commanders. The first Napoleon was particularly distinguished for this. But it was not till our time—till a few years ago—that armies could be moved by locomotive speed or intelligence communicated with lightning rapidity. Railroads and the telegraph have revolutionized the art of war.

(a) *The New York Herald*, excerpt from the issue of the 1853-05-31

(b) –

Figure 10.3: *The New York Herald*, excerpt from the issue of the 1870-11-21

the cluster "Carlist War", showing how the telegraphs were used for communication in the Spanish civil war of the XIXth century. Additionally, telegraphs also are used throughout the years to keep the public informed about the ships circulating (Cluster "Ports"), for example in the article from Figure 10.3 (a). This theme is recurrent in the American articles, but is not found in any of the European countries.

### 10.2.2 Administration and political life

This power in communications that comes with the telegraph is taken advantage of by the governments and the administration. Diverse ideas are tried out and the different governments are testing processes, especially in France during the 1880-1910 period according to the Cluster "Ministers", but some governments like in Germany are more cautious about this technology in the elections, coming up with a strategy to avoid using it "Since the main purpose of this general organization is to transmit an important message to all league members as quickly as possible, without using the post, the telegraph or any other public office, it is necessary that the selected teams are personally known to the head of the arrondissement." (*Neue Hamburger Zeitung*, 1889-03-15)

### 10.2.3 Strategical in wars and protests

In wars or protests, the physical telegraph structures are also often the target of choice. In European articles, during the whole period some articles mention how



protesters have used the deterioration of the telegraphic installations. The articles from this category prove that communications means are controlled by people who have power, and cutting them from it gives a strategical advantage. Getting cut out of communications was scary, as the article from *El Imparcial* says, *"The news that arrives is alarming, riots have broken out in Rasgiad. In Bulgaria some bands of insurrectionists have appeared. In Rumelia the telegraph has been cut off. Note the importance of these events. It is presumed that its authors are Muslims, irritated by the atrocities committed by the Bulgarians. Demolition of the fortifications in Silistria Square has begun. (El Imparcial, 1879-07-17).* It is also a war strategy, *"The enemy fire broke the wires of the overhead line cable, and we were cut off from both valves, the fault being repaired after two hours by the personnel of Telegrafos Sre.s. López Vicencio and D. Nicolás Romano, helped by their orderly D. Miguel Iglinas."* (*El Imparcial*, 1909-09-11). This techniques were used in many different countries as soon as the communications means became available, for example this piece of article shows how French settlers communications are made more difficult by people cutting the wires : *Bone, April 26. The only telegraph wire linking the Algerian border to Tunis was cut yesterday at Kef. The occupation of this place by the column of General Logerot will soon enable it to be reestablished from Kef to the Algerian border. In the meantime, there will be no further news from Tunis except through a State aviso, which will make a daily service between Tunis and La Calle. (Le Figaro, 1881-04-27).* It is however relevant to note that it was not a public discourse in The New York Herald, unlike the other journals from the dataset.

#### 10.2.4 Part of war organisation

The telegraph slowly becomes an entire part of war and armies between 1890 and 1920 in Spain, France, the United States and Germany (Cluster "Army, war"), especially in the 1910-1920 decade, during which occurred World War I. The telegraphs in war are a part of full autonomous group in the battalions, as narrated in the article of Figure 10.4. They even got some war decorations in Germany *"The Iron Cross The following warriors from Hamburg and the university area were awarded the Iron Cross: [...] Deputy Assistant Telegrapher Pine in the 1st Land- Wahr-Batterie"* (*Neue Hamburger Zeitung*, 1914-11-01).

### 10.3 The telegraph network development

The telegraph in the public discourse does not only take the role of a carrier of information, but the technology itself is also discussed.

**Resumen general**

El resumen general por Armas y Cuerpos es el siguiente:

Infantería.—Sesenta y ocho regimientos de infantería de línea, veintidós batallones de cazadores, brigada disciplinaria.

Caballería.—Veintinueve regimientos, dos escuadrones independientes.

Artillería.—Catorce regimientos montados, un regimiento á caballo, dos regimientos mixtos, cuatro grupos de montaña, dos grupos de obuses de campaña, dos grupos de cañones de 12, dos grupos mixtos de á dos baterías, dos baterías de montaña, un tren de sitio, trece Comandancias de artillería de plaza.

Ingenieros.—Dos regimientos mixtos, ocho batallones de zapadores, un regimiento de Telégrafos, un regimiento de Ferrocarriles, un regimiento de pontoneros, dos compañías de las redes telegráficas de Ceuta y Melilla, tropas de aerostación y alumbrado en campaña, cuatro compañías independientes de zapadores, seis compañías independientes de Telégrafos.

Figure 10.4: *El Imparcial*, excerpt from the issue of the 1911-03-04

### 10.3.1 Organisation of national and international services

The newspapers articles teach another important view of the public : the Second Industrial Revolution is a prosperous period of telegraphic constructions (Clusters "Cables", "Post Office"). It starts in the 1860s in the United State (*"The steamship Great Eastern has commenced to take on board the Franco-American telegraph cable, it is thought sue will start in June next from Brest to lay the cable."* (*The New York Herald*, 1869-01-15)). It is also a discourse seen in the early parts of the German and Spanish datasets, with for example an article that states that the *"Secretary of State von Stevhan remarks that the allied governments have succeeded, through the concessions of the English government, in bringing the Anglo-German cable into state ownership and thus bringing about considerable facilitation of travel Direct telegraphic traffic will be set up between Hamburg and Liverpool in the near future."* (*Neue Hamburger Zeitung*, 1889-02-10) or for Spain (*"The captain general, the telegraph engineers and an inspector in order to choose the point where the cable station of the others must be built, Antilles."* (*El Imparcial*, 1868-05-11)). The cables industries continue to take some place in the news until the end of the XIXth century (see the article in Figure 10.6). Eventually, this topic fades out in the beginning of the XXth century,

**The Government Telegraph.**

A radical organ in this city comes to the aid of the telegraph monopoly and furnishes five reasons why the telegraph business should not be controlled by the government and form a part of the postal system. These reasons, stripped of their verbiage, are in substance as follows:—

I. The public markets in New York are for the most part public auctions, while if left to private enterprise our supply of food would be better and cheaper than it is. Ergo, the telegraph business can be better done by private companies than by the government.

II. The Post Office Department shows a deficit of six million dollars in the business of carrying letters and should not undertake to extend its sphere of operations until it can produce a better balance sheet.

III. That while the government telegraph would cheapen the cost of messages the clamor for lower rates would prevent any profits to the Treasury, and might even necessitate an outlay which might be more profitably spent in cheapening the price of potatoes by establishing government potato-growing farms on a gigantic scale.

IV. That as President Jackson and Postmaster Kendall rifled the mails and burned abolition documents, and as General Jackson and Governor Maroy recommended that the circulation of such documents through the Post Office should be prohibited by law, it would be unsafe to entrust the telegraph to the control of the government, which must always be in the hands of the dominant political party.

V. That as our government is in debt and is hounded by lobbyists for railroad subsidies it had better build railroads than construct telegraphs "where they are not wanted."

Figure 10.5: *The New York Herald*, excerpt from the issue of the 1869-01-15



as it does not appear in the Italian dataset that starts in 1910, nor in any other ones.



**Figure 10.6:** *Le Figaro*, excerpt from the issue of the 1893-10-17

The organisation of the national and international services also come with a creation of new jobs that make the telegraph function, which make other ones obsolete ("The telegraph, the telephone and especially the facility of communications make these civil servants something as old-fashioned as a postmaster or tax collector would be today." (*Le Figaro*, 1907-05-02)). It also questions its governance, with public debates around whether it should be public or private in the United States (Figure 10.5) or with ministries in Europe (Cluster "Ministers"), with more or less success. In Italy, many articles are about complains of the public service, "In short, post, telegraph and telephone are no longer, in Italy, a public service, but public confusion, with enormous damage to the national economic life." (*La Stampa*, 1921)

### 10.3.2 Regional Disparities

On the downsides of the debates, lie its regional disparities in its geographical implementation. An amusing example is how they accommodate : "Since there were neither railways nor telegraphs in the country at that time, the news of the five numbers drawn from one type of drawing after the other could only get through the ordinary mail, which, given the great distance between Copenhagen and Altona, took about two days . For this reason, it was permitted that, if a draw was made in Copenhagen, numbers could still be occupied in Altona on the day after the draw day, since, as was assumed, no news of the numbers drawn the day before in Copenhagen had arrived there at that time could be." (*Neue Hamburger Zeitung*, 1889-08-06)

### 10.3.3 Cost

The cost and budget of the telegraph is also always a public discussion in the newspapers, among every country and every decade (Cluster "Debt, Budget"). For example, in Germany, *"There have also been reductions in postage every year, as well as reductions in telegraph charges, etc. If we were to grant all motions to this budget for postage reductions, salary increases, etc., then 15 million would be necessary, and if we accepted the motions that went further, Singer on salary increases for the messengers and the sub-officials, then a further 9 million would be necessary. That would pretty much use up the postal administration's total surplus of 26 million. Gentlemen, with the best will in the world, not everything that is requested here can be fulfilled. I therefore ask you to reject all of the proposals on the postal budget."* (Neue Hamburger Zeitung, 1889-12-12)

### 10.3.4 Reliability in the service

Finally, some are doubting the reliability of the service, with a fear of falsified news in France for example, with an article saying *"This does not mean that all the dispatches dated from the Far East and given by the agencies are exact. Oh ! No. More than half are false or travesty, and this is understandable when one thinks of the enormous cost of transmitting the word telegraph."* (Le Figaro, 1895-05-03) or in Spain *"the news that have been communicated by telegraph be completely false"* (El Imparcial, 1873-03-15). They might also simply do not trust the organisation itself to give accurate news, because of weather conditions or because of malicious intent, which sometimes cause other material damages (*"The delay in the telegraphic dispatch that we mentioned has caused considerable damage to 4 the person 6 who amounted to 91,000, but of these, 38,106 were moved by animal force"* (El Imparcial, 1869-02-06)).

To conclude on the telegraphs discourses, the technology is met with enthusiasm regarding to its power and applications in daily life, in the five countries studied. It quickly became a tool in war and politics, both as a tool and as strategic attack point. The quality, cost and implementation of the national services are also criticized often by all of the five Western countries. The analysis allowed to highlight the diversity of topics discussed around the telegraphs, illustrating the diversity of contexts and opinions about them, and emphasize the similarities in discussions such as its use in politics, as well as the dissimilarities for example with the singular intensity of the complains about the public service in Italy.

*The main goal of the thesis was to observe trends of data behavior in information about technologies in the dataset, but it still does have limitations in its approach and dataset, which are discussed in this following part.*

### **The Corpus**

To study how society reacted to the new technologies of the Second Industrial Revolution, the thesis used newspapers as its primary source. Newspapers only reflect a fraction of reality, so not all of the discourses about technologies are revealed through the analysis.

### **The Journals' data**

The dataset composed of journals' data is not complete, not every journal spans the entire period of the Second Industrial Revolution, which makes the comparisons harder to achieve. Besides, the thesis quantitative results have a lot of variance between the journals because the OCR quality differs a lot between newspapers.

### **Pachinko Allocation Modelling**

PAM for topic modelling is based on a Bayesian approach that makes a Dirichlet assumption about the distribution of topics in documents. Real-life data has been argued to be more sparse than that [VP15], and non-Bayesian approaches might lead to different topics results. Moreover, the score to evaluate the PAM topics is the coherence score  $c_V$ , which has sometimes been seen to correlate negatively with other coherence measures [Vel17].

### **Cross-Lingual Word Embedding Models**

In the CLWEM experiment, the supervised alignment methods used for ground truth dictionaries from the MUSE project [Con+17], which are dictionaries that potentially differ from the vocabulary of XIXth century's journals. Plus, the alignment processes do not take into account potential OCR errors, yet it has been showed that

there are some. Finally, the choice of target language as English and the evaluation on the French-English alignments might also interfere with the results.

### **Network of Topics**

Finally, the construction and evaluation of the network of topics is very sensible to the hyperparameters and thus influence the cluster's results, so the inferences made from them have to be cautious.

However, while there are some technical limitations in the overall scope of this thesis, its main goal has been achieved: observing trends about social views on technology, and using these results to contribute with knowledge discovery to existing approaches in the field.

The thesis goal was to study cultural differences in information encoding of the same technological concepts across countries, during the Second Industrial Revolution. The analysis has been conducted on newspapers of records and their computational analysis, for the more general purpose of assessing the idea that Industrialization is a defining property of Western societies. While extensive work has already been done on the history of the Industrial Revolution, both generally speaking and in specific countries, little attention was given to the social perception of technology. This thesis has attempted to fill that research gap, questioning what people thought of the technologies themselves and the contexts in which they appear, through newspapers.

For the study, the four words *Coal*, *Steel*, *Electricity* and *Telegraph* were used to represent four categories of massive and fast technological advances that have had lasting impacts on the societies. Their use studied through French, German, Spanish, Italian and American journals are reflecting the public discourses around them. This work contributed to observe how different countries experienced those technologies in a contemporary manner, through clusters of discourses changing over time.

To study the discourses in the historical newspapers at a large scale, Pachinko Allocation Modelling has been used decade by decade and journal by journal. To provide a multi-lingual and cross-temporal analysis of the topics, the journals vocabularies have been transformed into Fasttext Word Embeddings, and all aligned on the American model using the Wasserstein-Procrustes algorithm, in order to have a shared representation for all the words in the entire corpus, regardless of the language. Finally, for the question of similarities and differences of the discourses in time and places, the Word Mover's Distance has been used between topics in the Cross-Lingual space. These distances were used to create a network of topics, detecting communities using the Louvain and Clauset-Newman-Moore algorithms, answering the initial question of finding similarities or differences in topics around technologies in time and across societies.

The results highlighted that the early discourses, before the 1870s, surrounding

coal relate to its domestic usage. It is then globally replaced by its industrial use and quickly the conversation turns to how vital of a resource it is. The miner status and strikes become a central matter, starting from the United States in the 1860s, propagating to France, Germany, Spain and Italy later. Generally, coal is a positive subject in the news when the topic relates to the merchandise, but the discussions are much more negative when talking about the impact on the economy of the strikes, or crisis, and shows that everywhere coal is a source of tension.

The conversation around steel in the dataset is again homogeneous before the 1880s, and is mainly focused on its domestic use. In the United States, the artistic use of steel is emphasised, while in France it is more related to fashion. With the normalisation of industrialisation processes that make steel production cheaper, steel becomes talked about for its world-wide trade and its use in war. That aspect of the discourses is present in the five countries of the corpus.

Regarding electricity, the general sentiment from the public is excitement towards the discoveries that are made very public. In Germany and in France, the scholars are being praised, while in the United States subjects of religion and spirituality are mixed with electricity discourses. Common to every country is the consciousness of it being a revolution, and the witnessing of the construction of a new industry. Topics such as the governance of electricity production and services are also in debates in the United States, between public or private, while in Europe the questions are more about how their government manage the resource.

For the last great technical discovery studied, the telegraph, the discourses are also in general very positive. They are seen as a revolutionary mean of communication, by its speed and power to reduce the distances. The enthusiasm is first seen in the United States from the scientific community, and is spread in Europe quickly after. The discourses about communications in administration and wars with the telegraphs is talked about a lot when there is a war that the country participates in. In parallel, the journals are mentioning the advances in the construction of telegraph infrastructures, talking a lot about the trans-Atlantic cable in the United States. The negative aspects that the newspapers are covering are complains about the reliability, the cost and the quality of the service, the latter especially in Italy.

The thesis opened up the possibility of a comparative framework between newspapers discourses in different societies, and successfully observed trends about social views on technology. To push the project further, an interesting approach to deepen the results would be to utilize Sentiment Analysis methods in order to

get another perspective on the discourses and a quantitative measurement of the sentiment regarding technologies.





# Bibliography

---

- [Amm+16] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. **Massively multilingual word embeddings**. *arXiv preprint arXiv:1602.01925* (2016) (see page 28).
- [Amp22] André-Marie Ampère. **Recueil d'observations electro-dynamiques...** Crochard Paris, 1822 (see page 2).
- [APZ00] Carlos Blanco Aguinaga, Julio Rodríguez Puértolas, and Iris M Zavala. **Historia social de la literatura española (en lengua castellana)**. Vol. 56. Ediciones Akal, 2000 (see page 8).
- [AS13] Nikolaos Aletras and Mark Stevenson. **Evaluating topic coherence using distributional semantics**. In: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*. 2013, 13–22 (see page 19).
- [AW10] Hervé Abdi and Lynne J Williams. **Principal component analysis**. *Wiley interdisciplinary reviews: computational statistics* 2:4 (2010), 433–459 (see page 32).
- [Bas+21] Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. **The corpora they are a-changing: a case study in Italian newspapers**. In: Association for Computational Linguistics (ACL). 2021 (see page 9).
- [BB12] Jordan Boyd-Graber and David Blei. **Multilingual topic models for unaligned text**. *arXiv preprint arXiv:1205.2657* (2012) (see page 17).
- [BDV00] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. **A neural probabilistic language model**. *Advances in neural information processing systems* 13 (2000) (see page 23).
- [Bea14] W Worby Beaumont. **Motor vehicles and motors**. Vol. 1. Cambridge University Press, 2014 (see page 2).
- [Ber07] Alice Bernard. **Le Figaro**. *Vingtième siècle. Revue d'histoire* (2007), 202–204 (see page 8).
- [BFB22] Thomas Benchetrit, Elena Fernandez Fernandez, and Jérôme Baudry. **Quantifying Anxiety During the Second Industrial Revolution (1870-1915)**. In: *Semester Project*. EPFL, Lausanne, Switzerland, 2022 (see pages 13, 14).

- [BL06] David M Blei and John D Lafferty. **Dynamic topic models**. In: *Proceedings of the 23rd international conference on Machine learning*. 2006, 113–120 (see page 17).
- [BL07] David M Blei and John D Lafferty. **A correlated topic model of science**. *The annals of applied statistics* 1:1 (2007), 17–35 (see page 17).
- [Blo+08] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. **Fast unfolding of communities in large networks**. *Journal of statistical mechanics: theory and experiment* 2008:10 (2008), P10008 (see pages 36, 37).
- [BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. **Latent dirichlet allocation**. *Journal of machine Learning research* 3:Jan (2003), 993–1022 (see page 17).
- [BO91] Robert Boyer and André Orléan. **Les transformations des conventions salariales entre théorie et histoire d’Henry Ford au fordisme**. *Revue économique* (1991), 233–272 (see page 2).
- [Boj+17] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. **Enriching word vectors with subword information**. *Transactions of the association for computational linguistics* 5 (2017), 135–146 (see page 25).
- [Bou09] Gerlof Bouma. **Normalized (pointwise) mutual information in collocation extraction**. *Proceedings of GSCL* 30 (2009), 31–40 (see page 19).
- [Bri05] The Editors of Encyclopaedia Britannica. **New York Herald**. *Encyclopedia Britannica* (2005) (see page 8).
- [Bri17] The Editors of Encyclopaedia Britannica. **La Stampa**. *Encyclopedia Britannica* (2017) (see page 8).
- [CJ73] James L Crouthamel and Andrew Jackson. **James Gordon Bennett, the "New York Herald", and the development of newspaper sensationalism**. *New York History* 54:3 (1973), 294–316 (see page 8).
- [CN16] José Camacho-Collados and Roberto Navigli. **Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations**. In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. 2016, 43–50 (see page 26).
- [CNM04] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. **Finding community structure in very large networks**. *Physical review E* 70:6 (2004), 066111 (see pages 36, 37).
- [Con+17] Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. **Word translation without parallel data**. *arXiv preprint arXiv:1710.04087* (2017) (see pages 32, 69).

- [Cro89] James L Crouthamel. **Bennett's New York Herald and the rise of the popular press**. Syracuse University Press, 1989 (see page 8).
- [De 06] Elisabetta De Biasio. **Alfredo Frassati un conservatore illuminato: aspetti biografici editi e inediti**. Vol. 1133. FrancoAngeli, 2006 (see page 8).
- [Fit03] Branden Fitelson. **A probabilistic theory of coherence**. *Analysis* 63:3 (2003), 194–199 (see page 19).
- [FO20] Alan Fernihough and Kevin Hjortshøj O'Rourke. **Coal and the European Industrial Revolution**. *The Economic Journal* 131:635 (Nov. 2020), 1135–1149. ISSN: 0013-0133. DOI: [10.1093/ej/ueaa117](https://doi.org/10.1093/ej/ueaa117). eprint: <https://academic.oup.com/ej/article-pdf/131/635/1135/37008122/ueaa117.pdf>. URL: <https://doi.org/10.1093/ej/ueaa117> (see page 10).
- [For10] Santo Fortunato. **Community detection in graphs**. *Physics reports* 486:3-5 (2010), 75–174 (see page 38).
- [For12] Mauro Forno. **Informazione e potere: storia del giornalismo italiano**. Gius. Laterza & Figli Spa, 2012 (see page 8).
- [For16] Rochelle Forrester. **History of electricity**. Available at SSRN 2876929 (2016) (see page 2).
- [GD16] Anna Gladkova and Aleksandr Drozd. **Intrinsic evaluations of word embeddings: What can we do better?** In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. 2016, 36–42 (see page 25).
- [GJB19] Edouard Grave, Armand Joulin, and Quentin Berthet. **Unsupervised alignment of embeddings with wasserstein procrustes**. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, 1880–1890 (see pages 30, 31).
- [Glo88] Gustave Glotz. **La cité grecque**. Albin Michel, 1988 (see page 3).
- [Gor00] Robert J Gordon. **Does the "new economy" measure up to the great inventions of the past?** *Journal of economic perspectives* 14:4 (2000), 49–74 (see page 1).
- [Gow75] John C Gower. **Generalized procrustes analysis**. *Psychometrika* 40:1 (1975), 33–51 (see page 31).
- [Guo+16] Lei Guo, Chris J Vargo, Zixuan Pan, Weicong Ding, and Prakash Ishwar. **Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling**. *Journalism & Mass Communication Quarterly* 93:2 (2016), 332–359 (see page 17).

- [Hab91] Jurgen Habermas. **The structural transformation of the public sphere: An inquiry into a category of bourgeois society**. MIT press, 1991 (see page 3).
- [HB13] Karl Moritz Hermann and Phil Blunsom. **Multilingual distributed representations without word alignment**. *arXiv preprint arXiv:1312.6173* (2013) (see page 28).
- [HM17] Matthew Honnibal and Ines Montani. **spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing**. To appear. 2017 (see page 14).
- [Hof13] Thomas Hofmann. **Probabilistic latent semantic analysis**. *arXiv preprint arXiv:1301.6705* (2013) (see page 17).
- [Kan+14] Subbu Kannan, Vairaprakash Gurusamy, S Vijayarani, J Ilamathi, Ms Nithya, S Kannan, and V Gurusamy. **Preprocessing techniques for text mining**. *International Journal of Computer Science & Communication Networks* 5:1 (2014), 7–16 (see page 14).
- [KB19] Pooja Kherwa and Poonam Bansal. **Topic modeling: a comprehensive review**. *EAI Endorsed transactions on scalable information systems* 7:24 (2019) (see page 17).
- [KC15] Sunghwan Mac Kim and Steve Cassidy. **Finding Names in Trove: Named Entity Recognition for Australian Historical Newspapers**. In: *Proceedings of the Australasian Language Technology Association Workshop 2015*. Parramatta, Australia, Dec. 2015, 57–65. URL: <https://aclanthology.org/U15-1007> (see page 13).
- [KTB12] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. **Inducing crosslingual distributed representations of words**. In: *Proceedings of COLING 2012*. 2012, 1459–1474 (see page 28).
- [Kus+15] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. **From word embeddings to document distances**. In: *International conference on machine learning*. PMLR. 2015, 957–966 (see page 35).
- [KZ21] Fatemeh Kaveh-Yazdy and Sajjad Zarifzadeh. **Track Iran’s national COVID-19 response committee’s major concerns using two-stage unsupervised topic modeling**. *International journal of medical informatics* 145 (2021), 104309 (see page 17).
- [LM06] 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 (see pages 17–19).

- [LM08] Wei Li and Andrew McCallum. **Pachinko allocation: Scalable mixture models of topic correlations**. *J. of Machine Learning Research*. Submitted (2008) (see page 17).
- [Mar01] Jesús A Martínez. **Historia de la edición en España, 1836-1936**. Marcial Pons Historia, 2001 (see page 8).
- [MB07] Jon Mcauliffe and David Blei. **Supervised topic models**. *Advances in neural information processing systems* 20 (2007) (see page 17).
- [MF80] John C. Merrill and Harold A. Fisher. **The World's Great Dailies, Profiles of Fifty Newspapers**. New York: Hastings House, 1980 (see page 7).
- [MH98] Shannon E Martin and Kathleen A Hansen. **Newspapers of record in a digital age: From hot type to hot link**. Greenwood Publishing Group, 1998 (see page 7).
- [Mik+13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. **Efficient estimation of word representations in vector space**. *arXiv preprint arXiv:1301.3781* (2013) (see pages 23–25).
- [Mim+11] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. **Optimizing semantic coherence in topic models**. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*. 2011, 262–272 (see page 19).
- [Mio97] Philippe Mioche. **Et l'acier créa l'Europe**. *Matériaux pour l'histoire de notre temps* 47:1 (1997), 29–36 (see page 2).
- [MLS13] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. **Exploiting similarities among languages for machine translation**. *arXiv preprint arXiv:1309.4168* (2013) (see pages 28, 29).
- [MMH17] Josiane Mothe, Karen Mkhitarian, and Mariam Haroutunian. **Community detection: Comparison of state of the art algorithms**. In: *2017 Computer Science and Information Technologies (CSIT)*. IEEE. 2017, 125–129 (see page 36).
- [MRB10] Reiza Mukhlis, Muhammad Rhamdhani, and Geoffrey Brooks. **Sidewall materials for the Hall-Héroult process**. *Minerals, Metals and Materials Society/AIME, 420 Commonwealth Dr., P. O. Box 430 Warrendale PA 15086 USA.[npj]. 14-18 Feb* (2010) (see page 2).
- [MS98] Joel Mokyr and Robert H Strotz. **The second industrial revolution, 1870-1914**. *Storia dell'economia Mondiale* 21945:1 (1998) (see pages 1, 10, 11).
- [Mus78] Albert Edward Musson. **The growth of British industry**. London: Batsford, 1978 (see page 1).

- [New+10] David Newman, 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*. 2010, 100–108 (see pages 19, 20).
- [New06a] Mark EJ Newman. **Finding community structure in networks using the eigenvectors of matrices**. *Physical review E* 74:3 (2006), 036104 (see page 36).
- [New06b] Mark EJ Newman. **Modularity and community structure in networks**. *Proceedings of the national academy of sciences* 103:23 (2006), 8577–8582 (see page 37).
- [NG04] Mark EJ Newman and Michelle Girvan. **Finding and evaluating community structure in networks**. *Physical review E* 69:2 (2004), 026113 (see page 36).
- [Nil+14] Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. **A toolbox for representational similarity analysis**. *PLoS computational biology* 10:4 (2014), e1003553 (see page 28).
- [Ohm27] Georg Simon Ohm. **Die galvanische kette: mathematisch**. TH Riemann, 1827 (see page 2).
- [OR64] Ingram Olkin and Herman Rubin. **Multivariate beta distributions and independence properties of the Wishart distribution**. *The Annals of Mathematical Statistics* (1964), 261–269 (see page ).
- [Ove02] John Overholt. **Ottmar Mergenthaler: The Man and His Machine**. *Libraries & Culture* 37:4 (2002), 396–397 (see page 3).
- [PL05] Pascal Pons and Matthieu Latapy. **Computing communities in large networks using random walks**. In: *International symposium on computer and information sciences*. Springer. 2005, 284–293 (see page 36).
- [Por+08] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. **Fast collapsed gibbs sampling for latent dirichlet allocation**. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008, 569–577 (see page 19).
- [RAK07] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. **Near linear time algorithm to detect community structures in large-scale networks**. *Physical review E* 76:3 (2007), 036106 (see page 36).
- [RB08] Martin Rosvall and Carl T Bergstrom. **Maps of random walks on complex networks reveal community structure**. *Proceedings of the national academy of sciences* 105:4 (2008), 1118–1123 (see page 36).



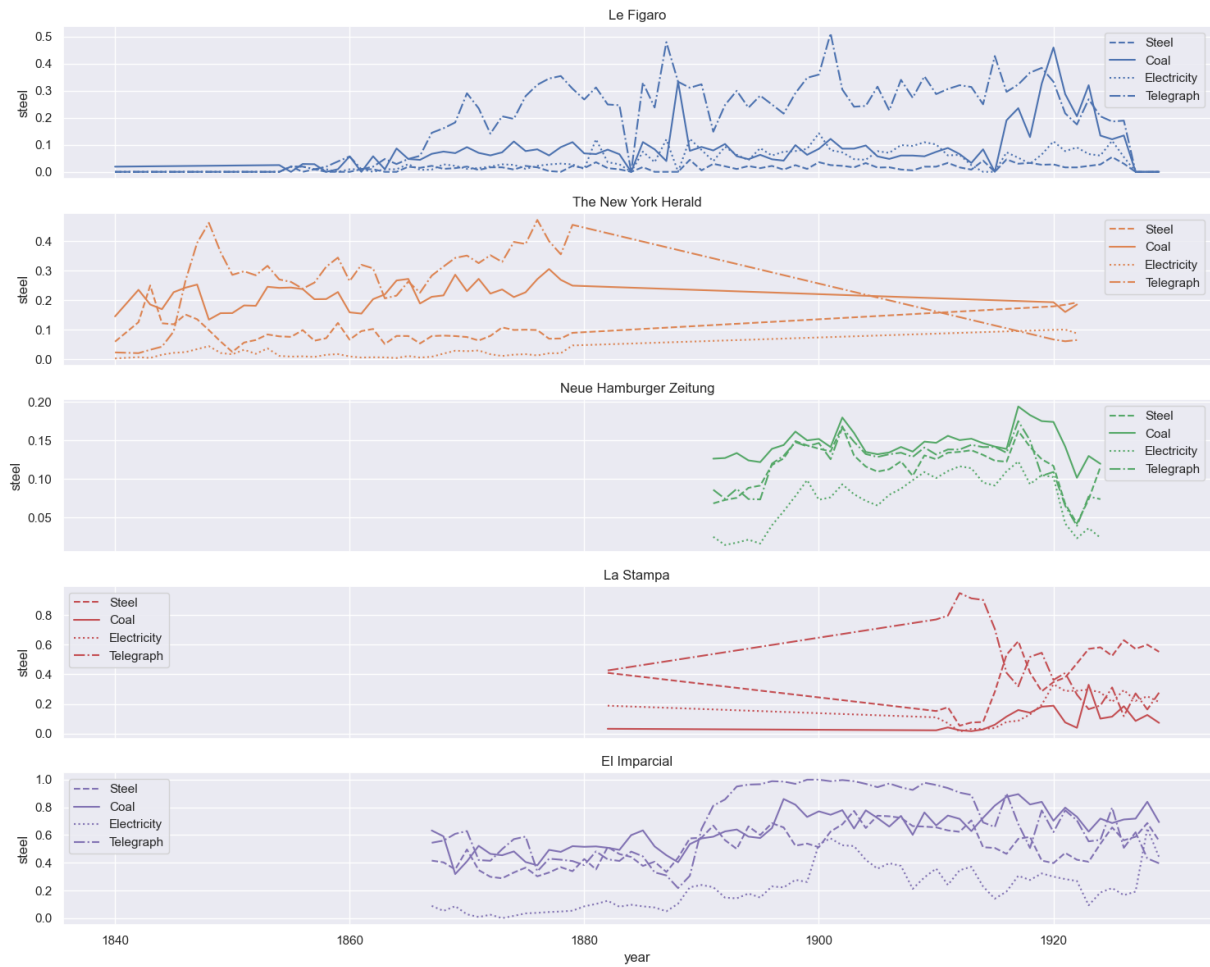
- [RBH15] Michael Röder, Andreas Both, and Alexander Hinneburg. **Exploring the space of topic coherence measures**. In: *Proceedings of the eighth ACM international conference on Web search and data mining*. 2015, 399–408 (see page 19).
- [Rei17] Simón Reif-Acherman. **Ernst Werner Von Siemens and the Early Evolution and Diffusion of Electric Telegraphy [Scanning Our Past]**. *Proceedings of the IEEE* 105:11 (2017), 2274–2284 (see page 2).
- [Res+15] Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. **Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter**. In: *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*. 2015, 99–107 (see page 17).
- [RVS19] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. **A survey of cross-lingual word embedding models**. *Journal of Artificial Intelligence Research* 65 (2019), 569–631 (see page 28).
- [Sch66] Peter H Schönemann. **A generalized solution of the orthogonal procrustes problem**. *Psychometrika* 31:1 (1966), 1–10 (see page 31).
- [Şen+17] Lütfi Kerem Şenel, Veysel Yücesoy, Aykut Koç, and Tolga Çukur. **Measuring cross-lingual semantic similarity across European languages**. In: *2017 40th international conference on telecommunications and signal processing (TSP)*. IEEE. 2017, 359–363 (see page 28).
- [Sen16] Richard Sennett. **Concentrating minds: how the Greeks designed spaces for public debate**. *Democratic Audit UK* (2016) (see page 3).
- [Smi05] Vaclav Smil. **Creating the twentieth century: Technical innovations of 1867-1914 and their lasting impact**. Oxford University Press, 2005 (see pages 1–3).
- [Son06] Christian Sonntag. **Medienkarrieren: biografische Studien über Hamburger Nachkriegsjournalisten 1946-1949**. Vol. 5. Martin Meidenbauer Verlag, 2006 (see page 8).
- [Tel00] Virginia Teller. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. 2000 (see page 23).
- [TWW19] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. **From Louvain to Leiden: guaranteeing well-connected communities**. *Scientific reports* 9:1 (2019), 1–12 (see page 38).
- [Vel17] Velcin. *Dice Group*. <https://github.com/dice-group/Palmetto/issues/12>. 2017 (see page 69).

- [VP15] Konstantin Vorontsov and Anna Potapenko. **Additive regularization of topic models**. *Machine Learning* 101:1 (2015), 303–323 (see page 69).
- [Wan+19] Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo. **Evaluating word embedding models: methods and experimental results**. *APSIPA transactions on signal and information processing* 8 (2019) (see page 25).
- [Wis08] Jaime Wisniak. **The development of Dynamite: From Braconnot to Nobel**. *Educación química* 19:1 (2008), 71–81 (see page 2).
- [XG14] Min Xiao and Yuhong Guo. **Distributed word representation learning for cross-lingual dependency parsing**. In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. 2014, 119–129 (see page 28).
- [Xin+15] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. **Normalized word embedding and orthogonal transform for bilingual word translation**. In: *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2015, 1006–1011 (see page 29).
- [Zha+15] Weizhong Zhao, James J Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding, and Wen Zou. **A heuristic approach to determine an appropriate number of topics in topic modeling**. In: *BMC bioinformatics*. Vol. 16. 13. Springer. 2015, 1–10 (see page 19).



# Appendix

## A KEYWORD DISTRIBUTION



**Figure 12.1:** Distribution of the proportion of articles that contain a specific keyword, across time, for the five different journals

## B DIRICHLET DISTRIBUTION

► **Definition 12.1.** Given the order  $K \geq 2$ , the parameters  $\alpha_1, \dots, \alpha_K > 0$ , and  $B(\alpha)$  the gamma function, the probability density function of the Dirichlet distribution is given by [OR64] :

$$f(x_1, \dots, x_k, \alpha_1, \dots, \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^K x_i^{\alpha_i - 1}$$

Where  $\sum_{i=1}^K x_i = 1$  and  $x_i \in [0, 1]$  and



## C GIBBS SAMPLING

The Gibbs sampling procedure from Wei Li and Andrew McCallum [LM06] is the following :

Gibbs sampling is needed for the estimations of the parameters  $\alpha = \{\alpha_1, \dots, \alpha_k\}$  of the Dirichlet distribution.

For an arbitrary DAG, the goal is to sample a topic path for each word given other variable assignments enumerating all possible paths and calculating their conditional probabilities. The joint probability of a super-topic and a sub-topic is, for a word  $w$ ,  $n_x^d$  the number of occurrences of topics  $t_x$  in document  $d$ ,  $n_{xw}$  the number of occurrences of word  $w$  in sub-topic  $t_x$ , and topic assignments  $\mathbf{z}$ :

$$P(z_{w2} = t_k, w_{w3} = t_p | \mathbf{D}, \mathbf{z}_{-w}, \alpha, \beta) \propto \frac{n_{1k}^{(d)} + \alpha_{1k}}{n_1^d + \sum_{k'} \alpha_{1k'}} \times \frac{n_{kp}^{(d)} + \alpha_{kp}}{n_k^d + \sum_{p'} \alpha_{kp'}} \times \frac{n_{pw} + \beta_w}{n_p + \sum_m \beta_m}$$

The Gibbs sampling iterations are, with  $\alpha_{xy}$  the  $y$ th component of  $\alpha_x$ ,  $N$  the number of document and  $s_2$  the number of subtopics :

$$\begin{aligned} \text{mean}_{xy} &= \frac{1}{N} \times \sum_d \frac{n_{xy}^{(d)}}{n_x^{(d)}} \\ \text{var}_{xy} &= \frac{1}{N} \sum_d \left( \frac{n_{xy}^{(d)}}{n_x^{(d)}} - \text{mean}_{xy} \right)^2 \\ m_{xy} &= \frac{\text{mean}_{xy} \times (1 - \text{mean}_{xy})}{\text{var}_{xy}} - 1 \\ \alpha_{xy} &\propto \text{mean}_{xy} \\ \sum_y \alpha_{xy} &= \frac{1}{5} \exp\left(\frac{\sum_y \log(m_{xy})}{s_2 - 1}\right) \end{aligned}$$

## D TOPICS EXAMPLE FOR ELECTRICITY ARTICLES IN SPANISH DATASET (1910-1920)

1910-1920_elec_es_0	['ministro', 'gobierno', 'señor', 'ción', 'consejo', 'proyecto', 'españa', 'español', 'haber', 'decreto']
1910-1920_elec_es_1	['diputado', 'presidente', 'político', 'declarar', 'comité', 'liberal', 'voto', 'partido', 'hablar', 'jefe']
1910-1920_elec_es_2	['vida', 'ver', 'hombre', 'ferrer', 'ciencia', 'tierra', 'vivir', 'movimiento', 'espíritu', 'materia']
1910-1920_elec_es_3	['salón', 'niño', 'obra', 'mujer', 'teatro', 'amor', 'joven', 'sala', 'padre', 'artístico']
1910-1920_elec_es_4	['pta', 'vendo', 'tiendar', 'pesetas', 'radium', 'precio', 'endo', 'ompro', 'piel', 'curación']
1910-1920_elec_es_5	['alemán', 'enemigo', 'tropa', 'noche', 'inglés', 'oficial', 'buque', 'frente', 'ruso', 'ataque']
1910-1920_elec_es_6	['calle', 'casa', 'santo', 'domicilio', 'socorro', 'plaza', 'juzgado', 'hospital', 'paseo', 'señora']
1910-1920_elec_es_7	['ídem', 'medio', 'doña', 'josé', 'serie', 'cuarto', 'escuela', 'celebrar', 'acto', 'sección']
1910-1920_elec_es_8	['cuyo', 'municipal', 'carbón', 'eléctrico', 'alumbrado', 'alcalde', 'condición', 'ilustre', 'cuerpo', 'director']
1910-1920_elec_es_9	['impuesto', 'millón', 'terreno', 'compañía', 'negocio', 'pesetas', 'municipio', 'valor', 'aumento', 'consumo']
1910-1920_elec_es_10	['huelga', 'obrero', 'civil', 'guardia', 'gobernador', 'conflicto', 'grupo', 'noche', 'fábrica', 'patrono']

**Figure 12.2:** The 11 topics of the Spanish model for the "Electricity" dataset between 1910 and 1920

## E BEST PARAMETERS FOR PAM FOR EACH MODEL

country	keyword	period	k1	k2	corr
fr	telegraph	1860-1869	2	21	0.7885862119058837
fr	coal	1860-1869	3	23	0.7384456445275625
fr	steel	1870-1880	1	17	0.92123691019992828
fr	coal	1870-1880	3	29	0.9126352550043336
fr	telegraph	1870-1880	2	16	0.6198253518750958
fr	elec	1870-1880	2	7	0.4419798248213120
fr	elec	1880-1890	1	13	0.5884356019904117
fr	telegraph	1880-1890	3	22	0.7918821215629578
fr	coal	1880-1890	1	24	0.7489004701375961
fr	steel	1880-1890	1	7	0.649574090540409
fr	steel	1890-1900	1	10	0.5832299202680588
fr	coal	1890-1900	1	24	0.7469907164573669
fr	telegraph	1890-1900	1	16	0.7265876144170761
fr	elec	1890-1900	3	21	0.6976187149683635
fr	elec	1900-1910	1	22	0.7975096762180328
fr	telegraph	1900-1910	1	14	0.8482346475124359
fr	coal	1900-1910	3	18	0.7281996061404546
fr	steel	1900-1910	2	7	0.5189608494285494
fr	steel	1910-1920	2	7	0.5561135046184063
fr	coal	1910-1920	2	24	0.738403594493866
fr	coal	1910-1920	1	22	0.7702499151229858
fr	telegraph	1910-1920	2	19	0.7210405454039575
us	telegraph	1840-1850	1	7	0.9498405575752258
us	telegraph	1850-1860	1	10	0.92980894446373
us	telegraph	1860-1870	1	24	0.8631426095962524
us	telegraph	1870-1880	1	11	0.9564535975456238

country	keyword	period	k1	k2	corr
de	telegraph	1888-1890	1	22	0.7952684044837952
de	telegraph	1890-1900	1	15	0.8587274491786957
de	telegraph	1900-1910	1	14	0.9140062510967256
de	telegraph	1910-1920	3	24	0.7644175797700882
de	telegraph	1920-1930	2	19	0.6897958874702455
es	telegraph	1860-1870	2	24	0.6654401667416096
es	telegraph	1870-1880	2	18	0.7711580902338028
es	telegraph	1880-1890	2	21	0.6520807627588511
es	telegraph	1890-1900	2	15	0.8280601084232331
es	telegraph	1900-1910	1	24	0.9040980756282806
es	telegraph	1910-1920	1	24	0.9392493665218352
it	telegraph	1910-1920	2	7	0.8638278149068356
it	telegraph	1920-1930	2	22	0.5974773705005645
it	coal	1910-1920	1	18	0.4809303909540176
it	coal	1920-1930	1	10	0.6198253588750958
it	steel	1910-1920	3	12	0.7374453445275625
it	steel	1920-1930	2	18	0.7989568322896958
it	elec	1910-1920	1	12	0.3147928400172127
it	elec	1920-1930	2	23	0.7056276317685842
us	coal	1840-1850	1	17	0.8319746077060699
us	coal	1850-1860	1	21	0.8680730342864991
us	coal	1860-1870	2	13	0.8183993980288506
us	coal	1870-1880	2	15	0.838869059085846
us	coal	1920-1930	1	22	0.8161879867315293
us	elec	1840-1850	1	16	0.8883154988288879
us	elec	1850-1860	1	20	0.905642521381378
us	elec	1860-1870	1	24	0.926135754585266
us	elec	1870-1880	1	21	0.9156840324401856
us	elec	1920-1930	2	21	0.916055330634117
us	steel	1840-1850	1	15	0.9806666135787964
us	steel	1850-1860	1	22	0.9125174462795258
us	steel	1860-1870	1	10	0.93187016248703
us	steel	1870-1880	1	17	0.933037942647934
us	steel	1920-1930	1	24	0.9035253405570984

country	keyword	period	k1	k2	corr
de	coal	1880-1890	1	15	0.9235961019992828
de	coal	1890-1900	1	12	0.9623859643936156
de	coal	1900-1910	1	24	0.9467808067798614
de	coal	1910-1920	1	14	0.9252802789211272
de	coal	1920-1930	2	15	0.8648949474096299
de	elec	1880-1890	1	11	0.4419798048213124
de	elec	1890-1900	1	17	0.8483870387077331
de	elec	1900-1910	2	12	0.7485354691743851
de	elec	1910-1920	1	5	0.8375400066375732
de	elec	1920-1930	2	20	0.7489664226770401
de	steel	1880-1890	1	19	0.9870807230472564
de	steel	1890-1900	1	11	0.9817916810512544
de	steel	1900-1910	1	9	0.9746614038944243
de	steel	1910-1920	1	16	0.910264766216278
de	steel	1920-1930	1	16	0.7998746395111084
es	elec	1860-1870	2	21	0.8374889969825745
es	elec	1870-1880	2	19	0.6595143854618073
es	elec	1880-1890	2	23	0.7307646572589874
es	elec	1890-1900	1	18	0.8263696551322937
es	elec	1910-1920	1	11	0.8618036925792694
es	elec	1920-1930	1	23	0.8928063631057739
es	steel	1860-1870	1	10	0.5869528174400329
es	steel	1870-1880	2	16	0.7392612099647522
es	steel	1880-1890	2	17	0.7256784670054912
es	steel	1890-1900	1	19	0.9155355632305144
es	steel	1900-1910	1	21	0.8535748183727264
es	steel	1910-1920	1	24	0.9292645752429962
es	steel	1920-1930	1	20	0.933438104391098
es	coal	1860-1870	1	12	0.7863849699497223
es	coal	1870-1880	2	13	0.8326760441064835
es	coal	1880-1890	1	22	0.7685865879058837
es	coal	1890-1900	1	23	0.8965728938579559
es	coal	1900-1910	2	21	0.8239155493676662
es	coal	1910-1920	1	13	0.9139352560043336
es	coal	1920-1930	1	16	0.8972341179847717

country	keyword	period	k1	k2	corr
it	telegraph	1910-1920	2	7	0.8638278149068356
it	telegraph	1920-1930	2	22	0.5974773705005645
it	coal	1910-1920	1	18	0.4809303909540176
it	coal	1920-1930	1	10	0.6198253588750958
it	steel	1910-1920	3	12	0.7374453445275625
it	steel	1920-1930	2	18	0.7989568322896958
it	elec	1910-1920	1	12	0.3147928400172127
it	elec	1920-1930	2	23	0.7056276317685842
us	coal	1840-1850	1	17	0.8319746077060699
us	coal	1850-1860	1	21	0.8680730342864991
us	coal	1860-1870	2	13	0.8183993980288506
us	coal	1870-1880	2	15	0.838869059085846
us	coal	1920-1930	1	22	0.8161879867315293
us	elec	1840-1850	1	16	0.8883154988288879
us	elec	1850-1860	1	20	0.905642521381378
us	elec	1860-1870	1	24	0.926135754585266
us	elec	1870-1880	1	21	0.9156840324401856
us	elec	1920-1930	2	21	0.916055330634117
us	steel	1840-1850	1	15	0.9806666135787964
us	steel	1850-1860	1	22	0.9125174462795258
us	steel	1860-1870	1	10	0.93187016248703
us	steel	1870-1880	1	17	0.933037942647934
us	steel	1920-1930	1	24	0.9035253405570984
de	coal	1880-1890	1	15	0.9235961019992828
de	coal	1890-1900	1	12	0.9623859643936156
de	coal	1900-1910	1	24	0.9467808067798614
de	coal	1910-1920	1	14	0.9252802789211272
de	coal	1920-1930	2	15	0.8648949474096299



country	keyword	period	k1	k2	corr
de	elec	1880-1890	1	11	0.4419798048213124
de	elec	1890-1900	1	17	0.8483870387077331
de	elec	1900-1910	2	12	0.7485354691743851
de	elec	1910-1920	1	5	0.8375400066375732
de	elec	1920-1930	2	20	0.7489664226770401
de	steel	1880-1890	1	19	0.9870807230472564
de	steel	1890-1900	1	11	0.9817916810512544
de	steel	1900-1910	1	9	0.9746614038944243
de	steel	1910-1920	1	16	0.910264766216278
de	steel	1920-1930	1	16	0.7998746395111084
es	elec	1860-1870	2	21	0.8374889969825745
es	elec	1870-1880	2	19	0.6595143854618073
es	elec	1880-1890	2	23	0.7307646572589874
es	elec	1890-1900	1	18	0.8263696551322937
es	elec	1910-1920	1	11	0.8618036925792694
es	elec	1920-1930	1	23	0.8928063631057739
es	steel	1860-1870	1	10	0.5869528174400329
es	steel	1870-1880	2	16	0.7392612099647522
es	steel	1880-1890	2	17	0.7256784670054912
es	steel	1890-1900	1	19	0.9155355632305144
es	steel	1900-1910	1	21	0.8535748183727264
es	steel	1910-1920	1	24	0.9292645752429962
es	steel	1920-1930	1	20	0.933438104391098
es	coal	1860-1870	1	12	0.7863849699497223
es	coal	1870-1880	2	13	0.8326760441064835
es	coal	1880-1890	1	22	0.7685865879058837
es	coal	1890-1900	1	23	0.8965728938579559
es	coal	1900-1910	2	21	0.8239155493676662
es	coal	1910-1920	1	13	0.9139352560043336
es	coal	1920-1930	1	16	0.8972341179847717