Methods and Tools of Quantification in Historical Research

Napoleonic Employment Applications as a Case Study

Valentina Dal Cin
Università Ca' Foscari Venezia, Italia

Abstract With the advent of Digital Humanities offering innovative tools for historical research, this chapter evaluates their benefits and drawbacks using the Napoleonic Employment Applications project (NapApps) as a case study. After examining data structuring and categorization, it analyses 800 application letters from candidates for the Napoleonic administration to investigate aspects of professionalization. Focusing in particular on willingness to relocate and associated rhetorical strategies, the study identifies specific trends and distinctive vocabulary patterns. Following an evaluation of various text mining techniques along with their strengths and limitations, it explores methods that integrate text analysis with metadata about the authors. It concludes by arguing that this integrated approach provides a robust means for investigating historical phenomena.

Keywords Digital humanities. Quantitative methods. Text mining. Employment applications. Napoleonic Europe.

1 Introduction

The recent rising number of projects and publications within the field of Digital Humanities has significantly impacted historical research. Contemporary historical publications have extensively addressed

e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2024-07-08 | Accepted 2024-08-28 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1//007

both theoretical and methodological dimensions, offering researchers detailed guidance on utilizing new digital tools.[1] This contribution presents methods suitable for individual historical research endeavours, utilizing the Napoleonic Employment Applications project (NapApps) as a case study to analyse the advantages and limitations of different methodologies.

Following an introduction to the research questions, archival sources, and criteria for sample selection, the focus transitions to the considerations involved in categorization choices and the methods for identifying correlations between variables. Subsequently, the discussion explores text-mining techniques, emphasizing those that integrate text analysis with metadata essential for prosopographic research.

2 Research Questions, Sources, and Sample Selection

Before discussing any methods and tools, it is essential to introduce the historiographical questions that this research aims to address. The NapApps project focuses on applications for employment submitted by individuals seeking to join the Napoleonic administration at the beginning of the nineteenth century.[2] While the *Declaration of the Rights of Man and of the Citizen* of 1789 opened public employment to all, regardless of social status, it is assumed that the Napoleonic era not only maintained this principle but also consolidated its meritocratic essence.[3]

Although individuals did not enter the State administration through public competition but rather through sovereign appointments, recruitment criteria considered certain professional attributes. For example, the rule that a prefect (i.e., the head of administration in each department) should not be appointed in his place of residence, to avoid conflicts between public and private interests, implied that candidates had to be prepared to relocate. Additionally,

[1] On recent methodological reflections and discussion, see Salmi 2021; Crymble 2021; Lässig 2021, 5-34; Guildi 2020, 327-46; Story et al. 2020, 1337-46; Robertson 2016, 289-307. Examples of recent guides are Lemercier, Zalc 2019; Blaney et al. 2021; Corfield, Hitchcock 2022; Graham et al. 2022.

[2] The project *Napoleonic Job Applications: from Personal Pleas to Modern Curriculum Vitae in Early 19th-Century Europe* (NapApps) has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 101018470.

[3] The close connection between meritocracy and the Napoleonic regime has been extensively debated. For further insights cf. Ellis 1997, 50-1; Grab 2003, 21-3, 43, 59; Forrest 2011, 140-1, 202-3. For an in-depth examination of meritocracy spanning the revolutionary period to the Napoleonic era, particularly within the military context, consult Blaufarb 2002.

since the French Empire included non-French speaking departments, knowledge of the local language (i.e., Italian or German) was a significant factor for the government in selecting suitable candidates. However, as the highest local administrative officials were also expected to represent the government properly and lend prestige to their roles, a comfortable economic situation was an additional implicit criterion. This often resulted in the integration of local elite members into the administration to secure their support.[4]

These various elements can be traced within employment applications to determine whether and to what extent candidates understood and internalized these mechanisms and underlying values.[5] The content of the applications and the lexical choices made by the candidates can be related to their characteristics (i.e., age, place of origin, professional background, the employment requested and its geographical location, and the addressee of the application) to understand which features were influential and to what extent. This aspect of the research lends itself to quantitative analysis. Nevertheless, outlining a background of general trends does not exclude – and indeed fosters – the need to focus on specific cases qualitatively to understand in detail the reasons behind the behaviour of the candidates and the choices of the government.

Since the administrative positions under examination are of high and medium-high profiles, the candidates largely belonged to the upper or upper-middle classes of society, situating this research within the broader field of studies on the Napoleonic notables and their interaction with the government, especially in the departments annexed to the French Empire.[6] For this analysis, the focus has been restricted to the departments in the Italian peninsula and those within the *ancienne France*, to allow a comparative study of the application writing styles of Italian and French candidates.

The sources for the quantitative part of the research project are preserved at the Archives nationales de France in a collection covering the period 1789-1878.[7] The collection consists of 178 boxes ar-

[4] For prefects within the French Empire, Karila-Cohen 2021 and Whitcomb 1974. Regarding the practice of appointing prefects outside their native departments, Woloch 2001b, 53 and Broers 2016, 350. For studies examining language skills, cf. Woolf 1991, 73; McCain 2018, 206; Lignereux 2019, 61-3, 184-8. Broers (1996, 138-41) explores the question of professionalism, while Lignereux (2019) provides insights into the imperial dimension of careers during the Napoleonic age.

[5] Being oriented toward a goal, which is getting hired, the texts of employment applications reveal the criteria their authors considered relevant to reach it (Cohen 2017, 102; Lignereux 2012, 109).

[6] On Napoleonic notables, Bergeron, Chaussinand-Nogaret 1979; Woloch 2001a, 67-74; Dunne 2007, 61-78; Levati 2009, 215-28.

[7] Archives Nationales de France (AN), F/1dII, *Demandes diverses*, A1; A3; A4; B1; B2; B5; B6; B9; B11; B13; B15; B19; B26; B28; B29; C1; C2; C3; C4; C6; C8; C10; C11; C12;

ranged alphabetically. To target employment applications from the Napoleonic era, the most straightforward approach is to collect them from a random sample of boxes. However, to account for the candidates' place of origin, instead of choosing the boxes randomly, those containing at least one folder with a surname likely of Italian origin dated between 1799 and 1815 were selected. This resulted in a sample of 85 boxes (48% of the total). This choice was made to account for the smaller sample size of candidates from the Italian peninsula compared to those from France and furthermore to ensure the existence of a sufficient number of candidates from the Italian peninsula for meaningful comparisons.[8] Then, within each box, a choice was made to select only the folders of those who applied for the positions of prefect, subprefect, secretary-general of the prefecture, and prefectural counselor, as these are the main positions within the Napoleonic departmental administration.[9] The outcome was a sample of 330 candidates, authors of 800 job applications.[10] Even if the selection process was not random, the guiding principle does not bias the analysis, since – place of origin excluded – all other features of the candidates found in the 85 boxes are random.

C14; D1; D2; D4; D8; D21; F1; F3; F5; F6; G1; G2; G5; G6; G10; H2; I1; L1; L6; L15; L18; M1; M3; M4; M6; M8; M9; M10; M11; M12; M13; N1; N2; O1; P1; P4; P7; P8; P11; R1; R2; R3; R4; R5; R10; S1; S2; S3; S5; S6; S7; S8; T4; T5; T6; V1; V5; V7; Z1.

[8] Oversampling of minorities is addressed by Lemercier and Zalc, who recommend selecting samples of similar sizes (2019, 43). In the case presented here, French candidates constitute 59% of the sample (196 individuals), having submitted 456 employment applications (57% of the total), while Italian candidates make up 41% of the sample (134 individuals), with 344 employment applications (43% of the total). Although the sample sizes are not identical, they are sufficiently comparable. These comparisons can be performed using the chi-square statistical test, which will be discussed later. A preliminary version of the sample, consisting of 300 applicants and 690 applications (analysed in Dal Cin 2023, 53-68), has been supplemented with additional transcriptions. The full text of the transcriptions and the spreadsheet underlying the analyses is accessible in the Zenodo open access archive: `https://doi.org/10.5281/zenodo.13686359`.

[9] For a general overview of these positions, and their different duties see Godechot 1985, 586-99 and Lentz et al. 2008, 158-9, 521-2, 569, 593-4.

[10] As a rule, a sample size of at least 300 items is considered sufficient to obtain reliable results in historical research. Lemercier, Zalc 2019, 39.

3 Categorization, Quantification, and Correlation Detection

Once the sample for the quantitative analysis had been defined, the next step in the research was to extract information about the candidates and their job applications. In order to duly analyse the applications, a sort of a relational database was created with the help of two spreadsheets: the first encompassing the candidates who were assigned a unique ID and the second the job applications marked by the respective candidate ID, and associated with a progressive number for each application starting from 1 (e.g., a candidate with ID 10, would have a job application numbered 10.1). Each job application was then associated with a series of features, including date and place of writing, the requested job, its location (if specified), and the recipient. Other characteristics relate to the author of the application, including age, place of origin, and professional background. Each of these characteristics, called variables, was allocated its own column. They were associated with a column containing the full text of each job application letter.

In the first phase, the information was ingested into the candidates' spreadsheet in a discursive manner, following what was reported in the source. In the second phase, the information was categorized and ingested into the applications' spreadsheet. Categorization is necessary because, as Claire Lemercier and Claire Zalc explain, "quantification is only possible on standardized, as opposed to 'raw' data". They also add that prosopography – the approach adopted in the project – "always begins with categorization" (Lemercier, Zalc 2019, 34). Rather than categorizing the information in the candidates' spreadsheet, the focus was on the applications' spreadsheet as a primary dataset. This choice was justified by the intention to avoid anachronisms, since many features of the candidates could change over time. For example, a candidate's professional background at the time of the first application could be different from that of two or five years later. The same applies to the desired position, which could change over time. Therefore, using employment applications as the focal point of analysis meant contextualizing each action of a historical actor in a specific timeframe.

Though time-consuming, the process of ingesting and categorizing data is an opportunity to rethink how to interrogate the sources and to reexamine research hypotheses (Lemercier, Zalc 2019, 55-6). Consequently, categorization revision occurred several times while attention was paid to avoid losing information. Refining the granularity depended on the nature of analysis to be carried out. For example, a column was added specifying the department in which candidates requested to be employed (if mentioned) and another indicating its geographic area. Regrettably, these columns, like other columns within the spreadsheet, contain a number of cases of missing data due to the unavailability of certain information in both primary sources and secondary literature (Lemercier, Zalc 2019, 73). Common to historical research, the issue of missing data does not significantly alter the outcomes of quantitative analyses when its occurrence is restricted. This also applies to errors; while a few errors might have a notable impact on a small subset of cases, they typically do not disrupt broader trends calculated from a larger pool of cases.

Hence, for scholars inclined towards a quantitative approach or those seeking to criticize it, a grasp of statistics is indispensable. Lemercier and Zalc argue that "contingency tables and chi-square tests are the most important tools for historians to learn" (2019, 73). Indeed, contingency tables serve as the primary instrument for quantification, as they concurrently present the values of multiple variables (e.g., age class and desired position), thereby revealing any potential correlations. Illustrating this point with an example from the project is elucidating. As previously noted, the inquiry revolves around discerning the presence or absence of professional inclination among candidates who authored a sample of 800 employment applications. One indicative parameter of this inclination is the candidate's readiness to relocate. The archetype of the Weberian-style professional civil servant is presumed to be prepared to discharge duties wherever the State mandates.[11] Conversely, a preference to remain in one's home department suggests a profile akin to that of a local notable, willing to collaborate with the government only under the condition of retaining ties to the milieu from which his social prominence emanates.

The question is particularly intriguing within the Napoleonic era, as Aurélien Lignereux aptly articulates through the concept of the "marché impérial de l'emploi public", offering a lens to examine the impact of border expansion on social behaviour. The imperial expanse brought forth augmented employment prospects for those amenable to relocation, albeit amidst heightened competition (Lignereux 2019, 22-3). In his prosopographical analysis of the *Impériaux* – French nationals employed beyond their national boundaries – Lignereux delineates four applicant categories: those seeking employment in proximity of their home department, those explicitly seeking roles in distant locales, those vying for employment in newly annexed territories following the Empire's expansion, and those exhibiting no preference in a bid to enhance their employment prospects (105-7).

[11] The main features of the professional civil servant are described in Weber 1981, 58-66. The link between mobility and professionalism in the Napoleonic administration has been underlined by Michael Broers 1996, 141; 2005, 199-201; 2016, 356.

In contrast, the present analysis adopts a two-tiered approach: the first tier categorizes applications based on whether the candidate specifies a position within his home department, while the second tier differentiates applications for positions within the candidate's local area (including his home department and neighbouring departments) from others. Applications for positions within the candidate's home department account for 19.5% of the total. This percentage gains greater significance when considered alongside other variables. Consequently, the ensuing contingency tables relate the requested department to the place of origin variable, examining disparities between candidates from the *ancienne France* and those from the annexed departments of the Italian peninsula.

Table 1 Requested location in relation to candidates' place of origin (%)

	French departments	**Italian departments**	**Total**
Candidate's own department	10.1	9.4	19.5
Elsewhere or unspecified	46.5	34	80.5
Total	56.6	43.4	100 (N = 777)

Table 2 Requested location in relation to candidates' place of origin (%)*

	French departments	**Italian peninsula**	**Total**
Candidate's own area (home department and neighbouring departments)	18.5	19.7	38.2
Elsewhere or unspecified	38	23.8	61.8
Total	56.5	43.5	100 (N = 782)

* The variables are significantly correlated at the 5% level, as determined by the p-value from a chi-square test.

Only the disparities outlined in Table 2 are statistically significant according to the chi-square test [tab. 2].[12] This test compares observed frequencies with those expected under the null hypothesis,

[12] In Table 1, the total is 777 because precise information about the candidate's department was unavailable for 23 out of 800 applications. In Table 2, the total is 782 because, although the department was not explicitly mentioned in some cases, it was evident that five candidates were from Piedmont. Therefore, it was possible to determine their requests for employment within the same area.

assuming independence of the variables. If the difference exceeds a conventional threshold, indicated by a p-value below 0.05, implying a probability of less than 5%, the variables are deemed correlated. This indicates an imbalance in the distribution of values beyond what chance alone would produce, given the sample size. However, correlation does not imply causation, as two correlated variables could both be influenced by a third variable. Therefore, cautious interpretation is imperative. While the test is necessary, it is insufficient on its own. Since no correlation has been detected in Table 1, this indicates that there was no significant disparity between candidates from *ancienne France* and those from the Italian peninsula in terms of requesting a position within their own department **[tab. 1]**. Table 2 refines this observation, demonstrating that a correlation exists when the local area is considered **[tab. 2]**. Italian candidates exhibited a greater propensity to request positions within their own area, with Piedmontese seeking employment in Piedmont, Ligurians in Liguria, and so forth, reflecting the enduring influence of Ancien Régime Italian statal boundaries on societal mindset.

A similar "psychological barrier" existed for the French (Lignereux 2019, 107). Despite French candidates' greater willingness to relocate beyond their neighbouring departments compared to Italian applicants, a significant portion still specified positions within the historical borders of France.[13] Thus, further investigation into the specific geographic preferences of those explicitly citing distant locations is warranted. Among the French applications, 13.4% mentioned positions outside France, with 6.1% specifying departments in Italy, 4.8% indicating departments in Belgium, Germany, and the Netherlands collectively, and 2.2% referencing departments in Spain, Switzerland or the Illyrian Provinces. Conversely, 11.3% of Italian applications mentioned positions outside the Italian peninsula, with 9% specifying French departments, 1.2% mentioning Belgian or German departments, and 1.7% referencing departments in Spain, Switzerland, or in the Illyrian Provinces. The most notable disparity between the two groups is the French candidates' greater propensity to seek employment in Belgian, German, and Dutch departments compared to their Italian counterparts. However, since this difference arises from small percentages, each below 5%, a closer examination of individual applications is necessary. A detailed analysis reveals that French applications specifying positions in such areas were submitted by fifteen candidates. Of these, only eight could be

[13] Among the applications submitted by French candidates, 56% mentioning a department outside their area cited one that remained within the traditional borders of France. Conversely, for Italians, the percentage referring to Italian departments outside their area is 53%. This calculation excludes applications where the desired location was unspecified.

considered genuine relocators, while the remainder sought employment in departments where they could leverage their cultural background – particularly their knowledge of German if they came from the eastern regions of France – and family connections. Among these candidates is Felix Barthélemy, an auditor to the Council of State in 1811, who sought employment in the recently annexed Bouches de l'Elbe department, with Hamburg as its chief town, hailing from the Meurthe department in the East of France, drawing on his prior experience in the Left Bank of the Rhine and his knowledge of German.[14] This familiarity with the requested location is mirrored by Etienne de Falaiseau, who sought to become prefect in the Netherlands.[15] A branch of his family had embraced Protestantism centuries prior and settled in the United Provinces after the revocation of the Edict of Nantes. This old connection was revitalized in recent years when Falaiseau and his wife abandoned France to escape the excesses of the revolution and spent some time in the Netherlands as *émigrés* (De Zuylen de Nyevelt 1893, 350-7). The outcomes of this analysis align with observations made by Lignereux concerning the *Impériaux*, albeit his study encompasses not only officials from the prefectural administration. He noted that the second category among the four he identified – applicants explicitly seeking distant locations – comprised a limited number of individuals. Some candidates within this category justified their requests based on prior familiarity with the mentioned location, with many citing family reasons such as having a relative serving or exerting influence there (Lignereux 2019, 105-6). Therefore, the inclination of French candidates to apply for distant positions more frequently than Italian candidates is nuanced by these considerations.

This example highlights the importance of integrating quantitative and qualitative analyses to discern significant disparities through statistical tests. Such tests are vital for identifying correlations warranting attention and guiding qualitative analysis. Hence, deliberating on the results of these tests, alongside highlighting the absolute numbers underlying the percentages, stands as essential research validation practices akin to citing archival sources. Prior to delving further into the relationship between professionalism and relocation by scrutinizing candidates' vocabulary, it is imperative to acquire a comprehensive understanding of text mining techniques best suited to the corpus and research inquiries.

[14] Paris, 27 August 1811. To the minister of the Interior (AN, F/1dII/B5, folder Barthélemy). On his life and career refer to Barthélemy 1885, Lignereux 2019, 34-5, 213, and Van der Burg 2021, 126, 136.

[15] Paris, 24 July 1810. To the minister of the Interior (AN, F/1dII/D4, folder De Falaiseau).

4 Text Mining Methods: An Overview

In a recent article, Claire Lemercier (2019) underscored the persistent deferral of the intersection between history and quantitative text analysis, noting that "those who count words continue to do so more and more in the absence of historians". Indeed, historians encounter challenges in familiarizing themselves with a variety of methods characterized by non-uniform terminology (e.g., quantitative text analysis, text mining, natural language processing, distant reading, lexicography, etc.). Furthermore, mastering available tools often proves arduous without specialized backgrounds, as many are tailored for computational linguistics or computer science (Sinclair, Rockwell 2016, 288). However, both collective and individual projects have shown interest in these methods, assessing their relevance to historical research.

4.1 Counting Words

These range from basic word clouds and frequency graphs generated using Google Books Ngram to advanced machine learning algorithms. Each method has its own advantages and limitations, making careful selection crucial. As a result, there are often suggestions to explore multiple tools to ensure comprehensive analysis (Sinclair, Rockwell 2016, 288). While Google Books Ngrams and word clouds offer ease of creation and utility for visualizing broad features in presentations, they lack the robustness required to support rigorous scholarly arguments due to inherent biases.[16] More robust are text mining and machine learning methods. The former falls under unsupervised learning, focusing on exploration and discovery, particularly adept at clustering elements based on shared features. The latter, a supervised learning method, facilitates prediction by classifying elements according to researcher-defined properties (Jockers, Underwood 2016, 293; Nanni, Kuemper, Ponzetto 2016, 66-7).

In text mining document similarity can be computed based on textual attributes, such as the relative frequency of most common words. Word count serves as a fundamental statistic in text analysis. However, beyond merely tallying the most frequent words, identifying the most distinctive words – those significantly overrepresented in specific texts or groups defined by metadata – often yields more insightful results. Words that are infrequently used within the corpus may emerge as salient in texts by a particular author or written

[16] Biases of Google Books NGrams are discussed in Romein et al. 2020, 304-6 and in Lemercier, Zalc 2019, 146.

in a specific year. Analysing co-occurrences, or the frequency of two terms appearing together, represents a valuable method for further exploration. Voyant, a web-based reading and analysis environment, provides two tools for this purpose. The Correlations tool enables the exploration of term frequency synchronicity, while the Corpus Collocates tool showcases terms frequently appearing alongside a researcher-defined keyword, elucidating contextual associations.[17] The output generated by this tool resembles the display of keyword in context (KWIC), which is a prominent feature of software such as AntConc (Anthony 2005, 729). Although Voyant and AntConc possess notable advantages – freeware, rapidity, and ease of manipulation – they also entail significant drawbacks (Andersen 2022, 138-9). One such limitation, particularly relevant to historical research, is the inability to incorporate metadata. When it is essential to link texts with temporal, authorial, or geographical metadata, manual corpus segmentation becomes necessary. This process, though crucial, is exceedingly time-consuming, especially when dealing with complex metadata structures that encompass numerous variables for analysis.

4.2 Topic Modelling

Another prevalent unsupervised technique is topic modelling. This method, based on an algorithm that lacks semantic understanding of words but calculates the probability of their co-occurrence, identifies various topics within a corpus and presents them through lists of their most distinctive words. While topic modelling can be executed using various tools, it is often associated with MALLET (Machine Learning for Language Toolkit), a Java-based package based on Latent Dirichlet Allocation (LDA).[18] Despite predominantly being employed by literary scholars, historians have also begun to explore topic modelling.[19] However, prior to its application to a corpus, thorough consideration of its numerous pitfalls is imperative. The number of topics must be predetermined, a parameter devoid of definitive rules, necessitating researchers to iterate the process using different numbers until obtaining satisfactory results. Additionally, researchers must compile a list of stopwords – words that frequently occur

[17] https://voyant-tools.org/docs/#!/guide/correlations and https://voyant-tools.org/docs/#!/guide/corpuscollocates.

[18] Latent Dirichlet allocation (LDA) was developed in the early 2000s by a group of researchers led by David Blei.

[19] A survey on the application of topic modelling in historical research is provided by Brauer, Fridlund 2013, 152-63.

but lack substantive meaning (e.g., articles and conjunctions) – to enhance result significance. This occasionally leads researchers to add words with meaning to this list if their excessive frequency compromises output quality, a decision subject to debate. Moreover, topics generated by the algorithm in the form of word lists lack labels, mandating researchers to assign them.[20] This task is challenging as words often are inconsistent, with only a fraction sharing common features. These considerations underscore the significant role of human decisions and interpretation in a seemingly objective technique.

However, this does not discount the utility of Voyant, AntConc, and MALLET for historians. In her study on the transnational history of psychiatry, Eva Andersen integrated AntConc and MALLET with Histogram to devise a potent search tool aiding navigation through her extensive corpus of over 300,000 pages of psychiatric journals, uncovering unnoticed trends and pertinent segments for close reading (Andersen 2022, 131-57). Heidi Hakkarainen and Zuhair Iftikhar utilized topic modelling on a corpus comprising nearly 100 texts from the years 1829 to 1850 to probe the discourse surrounding the concept of humanism (*Humanismus*) in the German-speaking press. Recognizing the pivotal role of historical context in conceptual understanding, they employed topic modelling dynamically, segmenting their corpus into distinct time frames and organizing keywords chronologically. This approach corroborated Reinhart Koselleck's assertion regarding the *Sattelzeit* – a period wherein previously static phenomena became viewed as dynamic processes due to the heightened significance of temporal terms like *Zeit* (time) or *Zukunft* (future) by mid-century. However, Hakkarainen and Iftikhar cautioned that "the output of a topic modelling process is not a result in itself and needs to be studied further for reliable conclusions" (Hakkarainen, Iftikhar 2020, 269, 272). To furnish a bespoke solution for historical research, particularly for studies transitioning from exploratory analysis to hypothesis-testing endeavours, Federico Nanni, Hiram Kumper, and Simone Paolo Ponzetto advocated for a suite of semi-supervised computational methods. These techniques, both knowledge- and data-driven, facilitate fruitful synergy between algorithmic computational power and researcher domain expertise. Specifically, they advocate applying semi-supervised topic modelling algorithms utilized in natural language processing to historical research. These approaches permit the incorporation of metadata associated with texts into topic detection, as demonstrated by Labeled LDA, or allow researchers to manually define a list of relevant words to "guide the topic model in a specific direction", as exemplified by Seeded LDA

[20] For a detailed explanation of the mechanics of topic modelling, Graham et al. 2022, 115-54.

(Nanni, Kuemper, Ponzetto 2016, 69-71). Despite showcasing the efficacy of this approach through its application to a corpus of approximately 1,000 legal books from the seventeenth and eighteenth centuries, the adoption of these more intricate variants of topic modelling appears still limited (Nanni, Kuemper, Ponzetto 2016, 73-4).

In summary, the most widely used unsupervised techniques harbour significant heuristic potential when historians aim to explore large corpora that are impractical to read entirely. However, their utility diminishes when applied to relatively small corpora or to address precise research questions. Even Voyant's most useful feature – statistics on word frequency and co-occurrence – may require to be supplemented with other tools if the integration of metadata is indispensable for addressing research inquiries.

4.3 Machine Learning Methods

Before delving into this aspect further in the subsequent section, a brief overview of the possibilities afforded by machine learning methods, particularly supervised learning techniques, is warranted. These algorithms are trained on annotated samples to make predictions. However, their primary limitation lies in their demand for substantial amounts of training data to construct reliable models, alongside significant computational resources. An example is Word2Vec, an algorithm utilizing a neural network model to generate word embeddings. Since its introduction in 2013, it has gained traction among humanists interested in natural language processing. As elucidated by Melvin Wevers and Marijn Koolen, "a Word Embedding Model (WEM) contains semantic and syntactic information" derived from word distribution based on their co-occurrence frequency (Wevers, Koolen 2020, 226). By discerning relationships between words, the model facilitates contextual analysis (embeddings), synonym identification, and semantic change tracking. Consequently, a word embedding model empowers researchers exploring concepts like democracy to search not only for texts explicitly mentioning the keyword but also for texts contextualizing its usage, significantly enhancing information retrieval capabilities. Moreover, word embedding models prove beneficial for conceptual history and historical semantics due to their capacity to trace shifts in meaning, echoing the ideas of Reinhart Koselleck (Wevers, Koolen 2020, 227, 232, 238).

For instance, diachronic word embeddings were employed by Rocco Tripodi, Massimo Warglien, Simon Levis Sullam, and Deborah Paci to scrutinize the chronological evolution of antisemitic language. Through training the algorithm on a corpus comprising resources containing keywords related to Jews published between 1789 and 1914, digitized in the online library of the Bibliothèque Nationale de

France (Gallica), encompassing 54,403 books and 245,188 periodical issues, they traced the "evolution of antisemitic bias in the religious, economic, socio-political, racial, ethic and conspirational domains". This study not only qualitatively confirmed the "chronological development of antisemitic moments identified by historians" but also unveiled "an unexpected peak in adverse bias between 1855 and 1866, in connection with the French Second Empire" (Tripodi et al. 2019, 115-25). Conducted within the framework of the European project ODYCCEUS (Opinion Dynamics and Cultural Conflict in European Spaces), this study underscores a significant aspect of employing machine learning methods: the formidable challenges historians face in independently harnessing these tools, necessitating substantial support from research centres or large-scale collaborative projects.[21]

Nevertheless, despite their potency, word embedding models demand specific conditions seldom met in historical research. As previously noted, one pivotal condition is access to extensive corpora. It has been posited that, for robust representation, Word2Vec necessitates a corpus comprising at least 100 million words per time slice, a vocabulary of approximately one to two million distinct words, and texts featuring frequently co-occurring words. Furthermore, the corpus should incorporate a substantial number of words of interest; otherwise, the analysis output regarding semantic change may prove unreliable (Wevers, Koolen 2020, 233).

Given that the necessity for a large corpus can be mitigated if the corpus is relatively homogeneous, containing texts within a narrow domain, scholars have attempted to apply the algorithm to limited datasets. This was exemplified by Ekaterina Kamlovskaya, who investigated a collection of Indigenous Australian autobiographical narratives. Her initial findings corroborated qualitative research observations on the prevalence of introduced sports due to early Western cultural imposition and underscored the association between sports and words linked to self-esteem, suggesting that word embedding models represent "a promising method", albeit one that requires cautious utilization due to the absence of an optimal combination of parameters, necessitating researcher-driven choices (Kamlovskaya 2022, 93, 102-3, 105).

Unfortunately, in most cases, historians lack digitized text corpora meeting the necessary criteria for processing with this algorithm. Furthermore, the largest available corpora exhibit biases toward certain source types, predominantly comprising parliamentary debates, newspapers, books, and journals, particularly from the nineteenth

[21] https://www.odycceus.eu/.

century onward.[22] For preceding centuries and varied collections, digitized primary sources remain scarce, potentially biasing future investigations. Should an increasing number of history students develop an interest in machine learning methods, their focus would naturally gravitate toward the study of accessible digitized sources, potentially exacerbating the divergence between 'traditional historians' and 'digital historians', whose topics of interest – not just methods – could significantly differ. Consequently, the selection of materials for digitization by public and private institutions will wield a profound influence on the trajectory of historical research.

While their adoption in Digital Humanities remains limited, there is burgeoning interest in applying neural network models to predictive tasks such as text reconstruction, authorship attribution, and sentiment analysis. For instance, sentiment analysis was conducted by Tobias Blanke, Michael Bryant, and Mark Hedges employing recurrent neural networks to analyse 1,882 textual transcripts of interviews with Holocaust survivors from the United States Holocaust Memorial Museum. Following the creation of a training corpus using a blend of supervised and unsupervised techniques – including dictionary-based sentiment analysis and recurrent neural networks – they qualitatively and quantitatively assessed the results, determining that the latter technique outperformed significantly. However, given that "one of the major criticisms of neural networks is that they make it difficult to understand why they arrive at conclusions", the authors combined word statistics and algorithms to elucidate the neural network's decision-making process in tagging positive and negative testimonies. This multifaceted approach enabled them to comprehend certain errors, with the neural network being misled by family relations, erroneously associating them with positive memories even when occurring within a negative context (Blanke, Bryant, Hedges 2020, 26, 29-31). Employing predictive techniques on historical data, an approach termed 'Predicting the Past', remains contingent on the availability of extensive datasets. Moreover, while sentiment analysis is increasingly refined, it continues to attract substantial critique (Samuel, Rozzi, Palle 2022).

[22] Recent historical research employing text mining techniques has relied on such corpora. See, for example, Buongiorno et al. 2022 and Bunout, Ehrmann, Clavert 2023. While this study does not address these challenges, it is important to note that even when digitized corpora are available, significant issues must be considered, including OCR quality, alternative word spellings, abbreviations, polysemy, changing word usage, idioms, misspellings, and omissions (Oberbichler, Pfanzelter 2023, 136).

5 Combining Text Mining With Metadata Analysis

Combining text mining with metadata analysis presents a middle ground between simplistic approaches like Google Ngram and advanced techniques based on neural networks. Instead of relying on complex algorithms that might seem like a ‘black box’, historians could benefit from the statistics available in most text mining software. One particularly useful statistical measure is Term Frequency-Inverse Document Frequency (TF-IDF), first introduced by Karen Spärck Jones in 1972 for information retrieval. TF-IDF assesses the importance of a term within a corpus of documents by assigning higher weight to terms that are rare in the corpus yet highly discriminative for specific texts, and lower weight to terms that are common across the corpus and less useful for distinguishing between texts. In essence, a higher TF-IDF score indicates that a term is more distinctive for a certain text compared to all other texts in the corpus. This measure can be computed for individual texts as well as for texts aggregated based on metadata such as author, year, place, subject, etc. (Guldi 2022, 908). In a recent research project, Jo Guldi utilized a statistical measure called Term Frequency-Inverse Period Frequency (TF-IPF) on British parliamentary debates. This measure helped detect the most distinctive words in temporal divisions spanning from twenty years down to a single day. By altering the scale of analysis, Guldi revealed both long-term trends and previously overlooked short-term concerns, suggesting that TF-IPF can mitigate implicit biases and stimulate new research questions about lesser-known events (Guldi 2022, 895-911).[23]

Both TF-IDF and TF-IPF revolve around the concept of distinctiveness. However, distinctiveness can be approached in various ways, including through the comparison of word frequencies. Relative frequency, which measures the proportion of a word’s occurrences relative to the total number of words in a text, allows for comparisons across texts of different sizes. Although the idea of counting words and identifying co-occurrences is not new, it remains valuable in historical research. For instance, Jacques Guilhaumou’s studies on revolutionary discourse in the 1980s already employed counts and statistics (Guilhaumou 1986, 27-46). More recently, Luca Scholz analysed over 20,000 legal dissertations from seventeenth-century German universities, identifying trends in topics such as marriage, debt, and property law (Scholz 2022, 297-327). Cesare Vetter, Marco Marin, and Elisabetta Gon studied Maximilien Robespierre’s vocabulary and rhetoric, measuring the distinctiveness of the most

[23] New insights on this topic and general guidance on text mining techniques suitable to historical research are provided in Guldi 2023.

frequent socio-political words across different time periods (Vetter, Marin, Gon 2015, 96-110).

Drawing on the concept of distinctiveness, the Napoleonic Employment Applications project investigated the correlation between candidates' willingness to relocate and their professionalism. Employing an inductive approach grounded in relative frequencies, the study compared the vocabulary of candidates according to their preferred employment locations.[24] The corpus of 800 applications was divided into two distinct sub-corpora based on metadata provided by the applications' spreadsheet, which associated each application's text with specific details such as the desired place of employment. One sub-corpus comprised applications from candidates explicitly seeking positions within their home department, while the other consisted of applications from candidates who did not specify their home department as their desired place of employment. To facilitate the application of text mining methods, all words underwent standardization to contemporary spellings and lemmatization to reduce morphological variations to a singular lemma (Cortelazzo 2013, 302). To identify distinctive words based on relative frequencies, two key operations were conducted. Initially, only words exhibiting significant overuse in each sub-corpus, as determined through a chi-square test performed on relative frequencies, were retained. Consequently, the resulting list comprised words distinctive to each sub-corpus. Subsequently, this list was divided based on whether the highest relative frequency of each word occurred in the first or second sub-corpus. The following bubble chart illustrates the results of this analysis, with the size of the bubbles representing the higher frequency of words and their distinctiveness indicated by colour. The darker the colour, the more distinctive the word is of that sub-corpus **[fig. 1]**.[25]

[24] For a deductive approach centred on word lists that define thematic vocabularies, cf. Dal Cin 2023, 62-8.

[25] The relative frequencies of words are calculated per 10,000 words, excluding both names of people and places.

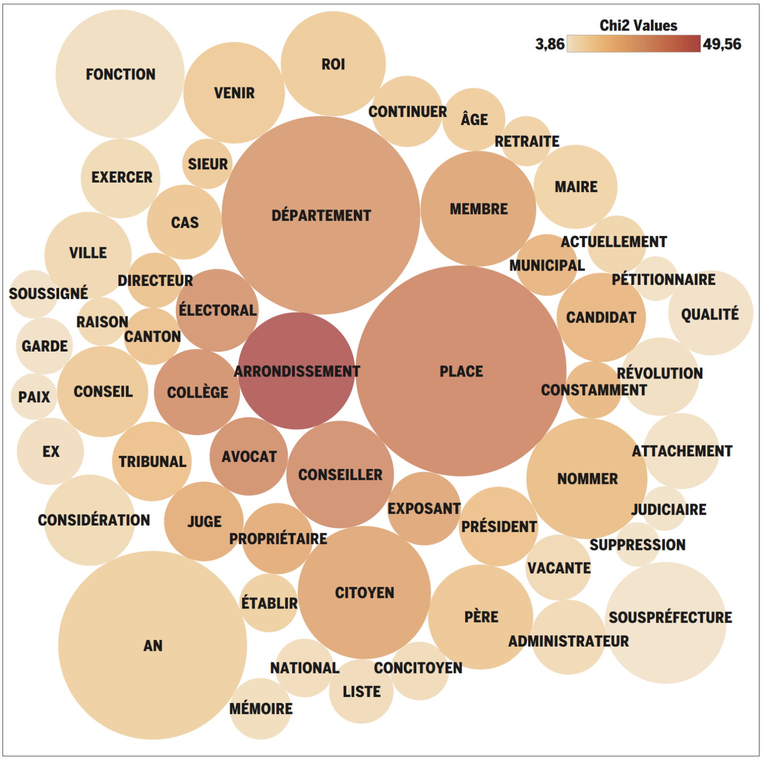

Figure 1 Most distinctive words used by candidates applying for employment within their own department

The larger bubbles indicate that words such as *place*, *département*, and *an* have the highest relative frequency. However, their distinctiveness varies, as indicated by their different colours. To further analyse this set of words, it is beneficial to categorize them according to their respective domains. Many words describe administrative boundaries, such as *arrondissement*, *département*, *canton*, *ville*. Other refer to institutions and positions embedded in specific territorial contexts, such as *collège*, *electoral*, *président* (often referring to the president of electoral colleges), *conseil*, *conseiller*, *municipal*, *maire*, *garde* (referring to the national guard).[26] Some words relate to the candidate's professional situation or background, such as *avocat*, *membre*, *directeur*, *juge*, *judiciaire*, *tribunal*, *candidat* (often referring to the *Corps législatif*), *paix* (referring to the position of *juge de paix*), *sous-préfecture*, *administrateur*, *suppression*, *fonction*, *exercer*.

[26] For a description of the role of electoral colleges and other Napoleonic departmental institutions, Lentz et al. 2008, 121-3, 154, 167.

Other words express a temporal dimension, distinguishing between the candidate's past experiences and their present situation. For instance, *an*, *continuer*, *constamment*, *actuellement*, *retraite*, and *ex*. *Propriétaire* and *liste* clearly reference the candidate's wealth and social status, indicating their inclusion on the list of notables or the six hundred largest taxpayers of their department. *Considération* and *concitoyen* suggest the candidate's social standing and the respect they enjoyed as evidence of their notability.

In essence, the words in the list delineate the characteristics of a local notable, highlighting their role as landowners and substantial taxpayers, along with their participation in departmental or municipal institutions. These individuals expressed readiness to engage with the government on the condition that they could uphold their social, familial, and economic connections within their local environment. Further elucidation can be attained by juxtaposing this list of distinctive words with that derived from the other sub-corpus, consisting of applications for positions beyond the candidate's department [fig. 2].

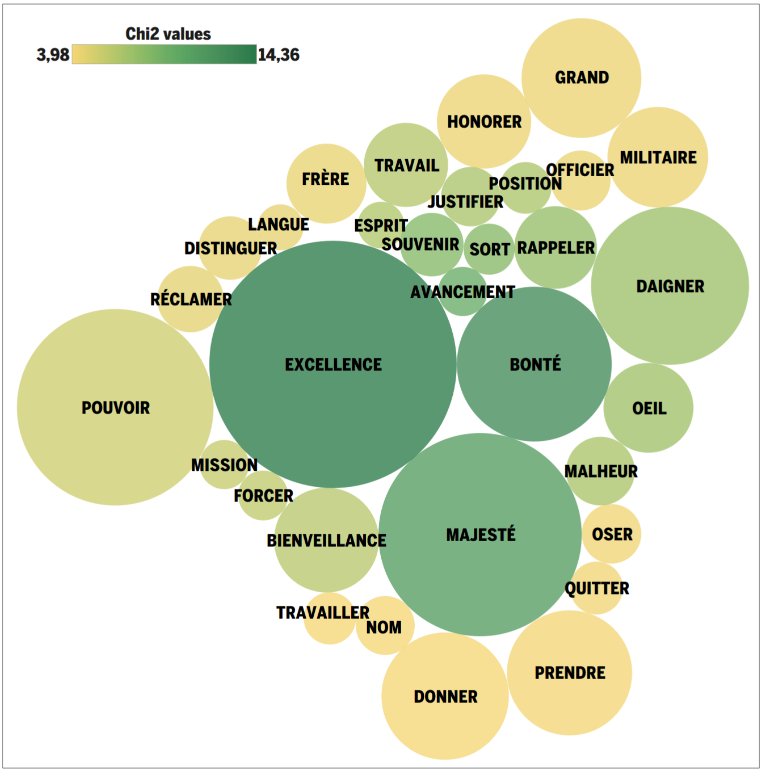

Figure 2 Most distinctive words used by candidates applying for employment outside their own department

In this instance, the word list offers a less distinct portrayal, given the greater heterogeneity of this sub-corpus compared to the previous one. The most frequent words, denoted by larger bubble sizes (*excellence*, *majesté, pouvoir, daigner, bonté, donner, prendre*) offer limited insight into the specific content of these applications. However, one word stands out as particularly distinctive, despite not being among the most frequent: *avancement*. Its absence in the other sub-corpus suggests that no candidate believed they could advance their career within their own department. Other words in the list pertain to duties performed and professional backgrounds, such as *mission*, *travail*, *travailler*, *langue*, *officier*, *militaire*. Yet, there are also numerous words conveying a paternal and benevolent expectation directed towards Napoleon or the Interior Minister, considering the applicant's misfortunes (*bonté*, *bienveillance*, *œil*, *malheur*, *souvenir*, *quitter*, *sort*, *forcer*). Regarding the term *frère*, it refers to either the candidate's brother's role or a recommendation provided by him. Since this term is more frequently used by French applicants, its inclusion in the distinctive words of this sub-corpus is likely due to their slightly higher presence within it (Dal Cin 2023, 63-4).

Within these sub-corpora, two key terms stand out as particularly distinctive: *propriétaire* in the first sub-corpus, and *avancement* in the second. The latter is notable for its reference to career prospects, a significant aspect given the absence of standardized rules for advancement in the French prefectural administration during the nineteenth century (Karila-Cohen 2021, 152-7). The former is significant because it indicates the candidates' intention to emphasize their ties to a specific area, along with their socio-economic status. For example, André-Paul Sain Rousset, the mayor of the Midi district of Lyon, expressed his desire for a position in proximity to the Rhône department, facilitating oversight of his vineyards in Vaux sur Villefranche, which formed a large portion of the estate he was to pass on to his heirs.[27]

The most frequent co-occurrences of *avancement* within a five-word distance include *excellence*, *monseigneur*, *vouloir*, *pouvoir*, *supplier*, *plaire*, *majesté*, *daigner*, *obtenir*, *service*, *administratif*, *place*, *bonté*, *prendre*, *administration*, *carrière*, *solliciter, demander*. The most associated co-occurrences, considering both how often two words co-occur and how often they do not, are *ci-joint*, *administratif*, *carrière*, *oser*, *supplier*, *obtenir*, *solliciter*, *accorder*, *fournir*, *aspirer*, *demander*. Six words appear in both lists: *carrière*, *administratif, supplier, obtenir, solliciter, demander*, indicating the applicant's reference to their career. The most frequent co-occurrences of *propriétaire* within a five-word distance include *département*, *monseigneur*, *an*, *agé*,

[27] Lyon, 27 September 1807. To the ministry of the Interior (AN, F/1dII/S2, folder Sain Rousset).

ancien, *général*, *place*, *natif*, *père*, *membre*, *famille*, *excellence*, *plaire*, *fil*, *arrondissement*, *solliciter*, *service*. The most associated co-occurrences are *natif*, *agé*, *département*, *feu*, *père*, *fil*, *commune*, *canton*, *ancien*, *avocat*, *marier*, *arrondissement*. Both lists include terms related to administrative geography (*département*, *arrondissement*, *canton*, *commune*) and family (*père*, *famille*, *fil*, *marier*).[28]

To achieve a comprehensive understanding of the context surrounding both *avancement* and *propriétaire*, it is imperative to analyse the positions sought by applications containing these terms,[29] along with correlations with requested locations and applicants' age classes.[30]

[28] Even if they frequently co-occur, names of people and places were disregarded. Hence, TF-IDF was not employed to identify the most distinctive words in the two sub-corpora, as it primarily identified names.

[29] Each occurrence was examined to determine its context. Instances where *avancement* referred to Napoleon's rise to power or individuals other than the applicant were excluded from the count. Similarly, instances of *propriétaire* highlighting the opposite of wealthiness were disregarded. Applications containing terms such as *aisance*, *fortune*, *ressource*, *subsistance*, *revenue* (wealth, fortune, resource, subsistence, income), *imposé* (referring to the list of the six hundred largest taxpayers), and *contribuable* (referring to the same context) were added to the count when emphasizing the candidate's wealth.

[30] In Tables 3-5, the total is never 800 because a single application could request multiple positions, corresponding to the total number of positions applied for. The number of applications considered is 777, consistent with Table 7. This excludes 23 applications for which the candidate's department is unknown. The total in Table 6 is 539 due to some unknown birthdates. All these contingency tables present values in percentages to facilitate comparisons; however, chi-square tests were performed on row counts. Instead of analysing correlations between pairs of variables, multiple correspondence analysis (MCA), a type of factor analysis, could be employed. MCA is often used in prosopographical research as it provides a general overview of the relationships between all variables, allowing for the detection of correlations without creating numerous contingency tables (Lemercier, Zalc 2019, 88-9, 97). However, MCA was not used in this analysis because the presence of missing data influences the results, as they correlate with each other. Various solutions to this problem exist, but they result in a loss of information.

Table 3 Requested position in relation to mention of the word *avancement* (%)*

	Prefect	Subprefect	Secretary General	Prefectural counselor	Other	Total
Avancement mentioned	2.4 (+)	0.9 (–)	0.7 (=)	0.1 (–)	0.3 (–)	4.5
Not mentioned	19.9 (–)	42.6 (+)	13.9 (=)	5.3 (+)	13.8 (+)	95.5
Total	22.3	43.5	14.6	5.4	14.2	100 (N = 873)

* In this table and the subsequent ones, an asterisk denotes a significant correlation between the variables at the 5% level, as determined by the p-value from a chi-square test. The symbols plus (+) and minus (–) indicate that the observed frequencies are higher or lower than the expected frequencies, respectively. An equal sign (=) signifies no significant difference between observed and expected frequencies. Expected frequencies are calculated under the assumption that the variables are independent. This means that the expected number for each cell is derived from the overall proportions in the table, considering both the row and column totals. If the variables were independent, the observed distribution would match these expected values.

Table 4 Requested position in relation to the mention of the candidate's wealth (%)*

	Prefect	Subprefect	Secretary General	Prefectural counselor	Other	Total
Wealth mentioned	1.8 (–)	7 (+)	1.9 (=)	1 (+)	0.5 (–)	12.3
Not mentioned	20.5 (+)	36.7 (–)	12.7 (=)	4.4 (–)	13.5 (+)	87.7
Total	22.3	43.6	14.7	5.4	14	100 (N = 873)

Table 5 Requested position in relation to requested location (%)*

	Prefect	Subprefect	Secretary General	Prefectural counselor	Other	Total
Candidate's own department	0.8 (–)	9.4 (+)	3.1 (+)	3.2 (+)	1.7 (–)	18.2
Elsewhere or unspecified	21.5 (+)	34.2 (–)	11.6 (–)	2.2 (–)	12.3 (+)	81.8
Total	22.3	43.6	14.7	5.4	14	100 (N = 873)

Table 6 Departments requested by age group (%)*

	20s	30s	40s	50s	60s	Total
Candidate's own department	0.4 (–)	4.8 (–)	3.3 (=)	5.2 (+)	0.6 (=)	14.3
Elsewhere or unspecified	11.1 (+)	37.3 (+)	23.9 (=)	10.9 (–)	2.4 (=)	85.7
Total	11.5	42.1	27.3	16.1	3	100 (N = 539)

Table 7 Department requested in relation to the mention of the candidate's wealth (%)*

	Wealth mentioned	Not mentioned	Total
Candidate's own department	4.4 (+)	15.1 (–)	19.4
Elsewhere or unspecified	7.9 (–)	72.7 (+)	80.6

All contingency tables show a statistically highly significant correlation, with a probability of less than 0.1 percent that the observed distribution of values is due to chance. Table 3 indicates a higher likelihood for applications containing the word *avancement* to be aimed at obtaining the role of prefect. This is attributed not only to its higher position in the hierarchy, prompting subprefects and secretaries general to aspire for promotion, but also to the establishment, beginning in 1810, of various classes of prefectures with differing salary levels (Karila-Cohen 2021, 152-4). Table 4 displays a different trend for applications in which candidates mentioned their wealth, showing a greater inclination towards subprefect and prefectural counselor roles, and less towards the position of prefect. Table 5 complements these findings by illustrating that applications requesting a position within the candidate's department were more likely to target subprefect and prefectural counselor roles, with very few applications for prefect positions specifying the candidate's own department **[tabs 3-5]**. These results align with the differing nature of these positions and governmental policies. As previously mentioned, the government typically refrained from appointing a prefect in their own department to maintain independence from local influences. In contrast, prefectural counselors were often chosen from local notables, offering a position with a low salary and limited career prospects but the advantages of avoiding relocation and gaining recognition under the new regime (Tulard, Tulard 2014, 97-126). Subprefects could be selected from either the local elite or career civil servants (Thoral 2010, 65-6), with the latter group becoming more prominent after 1809 when a decree reserved a quarter of vacant subprefectures for auditors at the Council of State (Durand 1958, 24-5). Table 6 shows that candidates in their twenties and thirties were more inclined to relocate, while those in their fifties were more likely to seek positions within their own department to enhance their social prestige. Table 7 corroborates previous findings, indicating a greater tendency for applications mentioning the applicant's wealth to specify their own department, whereas those willing to relocate exhibited a limited inclination to do so **[tabs 6-7]**.[31]

A couple of examples illustrate the duality underscored by this quantitative analysis. Charles-Claude Rambaud Brosse, a 53-year-old vice-mayor of the Midi district of Lyon, applied for the position of prefectural counsellor in his own department, Rhône. His application was grounded in his administrative involvement in Lyon, membership in the department's electoral college, and status as a landowner in both the city and the countryside.[32] Conversely, Etienne Charles Garnier, a 32-year-old who had previously served as a bureau chief at the Seine prefecture before becoming a secretary general in the Swiss department of Leman from 1801 to 1809, sought advancement by applying for vacant prefectural or similar administrative positions in Italy (Laharie, Lamoussière 1998, 336). He cited family hardships due to living expenses and climate, but also highlighted his past military service and administrative experience.[33] Garnier explained that he applied for a different role because his current tasks offered no possibility for advancement.[34] His case, coupled with the presence of the word *malheur* in the list of distinctive terms within applications for positions outside the candidate's department, illustrates the potential coexistence of career considerations and personal hardships within the same texts. Contrary to intuition, the inclusion of vocabulary related to 'sufferings' does not necessarily imply the absence of a professional mindset.

Since the subsets of candidates utilizing the vocabulary of wealth and the term *avancement* were selected to illustrate those seeking positions within their own department and those who were not, the correlations shown were expected **[tabs 3-5, 7]**. However, this quantitative analysis reveals additional trends. The position of subprefect emerges as the most sought-after, attracting 43.6% of all applications.

[31] The tendency is underlined by the result of the chi-square test, showing which percentages are higher than expected.

[32] Lyon, 17 March 1804. To the minister of the Interior (AN, F/1dII/R1, folder Rambaud Brosse).

[33] Genève, 24 March 1806. To the minister of the Interior (AN, F/1dII/G2, folder Garnier).

[34] "Ses fonctions ne lui fournissent pas dans les attributions qui y sont attachées les moyens de mériter par son travail de l'avancement dans la carrière". Cf. 15 October 1805 (AN, F/1dII/G2, folder Garnier). This letter addressed to the minister of the Interior is written by the candidate's wife. Therefore, it is not included in the sample analysed quantitatively.

Given that candidates rarely used the term *avancement* in this context, it suggests that this role was perceived as an entry-level, not requiring prior administrative experience. For example, Armand-Joseph-Louis Randon Saint-Marcel, a 25-year-old from Isère, applied in 1805 to become subprefect in Vouziers, Ardennes department, citing only his family connection with Napoleon's *capitaine des chasses* and his fervent desire to serve the emperor, following the example of his relatives.[35] Similarly, Claude-François Groshenry d'Emagny, a 26-year-old rentier from Besançon, Doubs department, applied in 1810 for the position of subprefect of Bergerac, Dordogne department, mentioning only his and his brother's military service.[36] Despite the increased prominence given to auditors in subprefect appointments after 1809, aimed at enhancing professionalization, the position continued to attract a variety of applicants from outside the administration.[37] Therefore, it would be beneficial to further investigate the dual nature of this role, sometimes entrusted to local notables and sometimes to career officials, but this analysis is beyond the scope of this study.

In addition to illuminating the heterogeneous profiles of candidates vying for the role of subprefect, the amalgamation of quantitative analysis of prosopographic data and vocabulary usage effectively delineated the profile of local notables inclined to serve the government exclusively within their own department. The distinctive terms they employed primarily related to local administrative positions, representative roles, and their status as landowners. The findings of the contingency tables resonate with analyses conducted on all French notables, indicating that a majority of them fell between the ages of 40 and 60, with 24.55% classified as property owners, 18.12% as local administrators, and 15.76% as civil servants by 1810.[38] Notably, this last percentage reveals an overlap between the profiles of notables and civil servants. However, while a prefect was likely to be a member of his department's electoral college and affluent enough to possess landed properties, a notable did not necessarily

[35] 3 January 1805. To Napoleon (AN, F/1dII/R1, folder Randon Saint Marcel).

[36] Paris, 8 and 11 October 1810. To the minister of the Interior (AN, F/1dII/G10, folder Groshenry d'Emagny).

[37] After 1809, applications for the role of subprefect dropped from 35% to 15%. However, this decline is part of a broader trend observed across the entire sample, indicating no correlation with the 1809 decree.

[38] These percentages refer to the approximately seventy thousand members of electoral colleges analysed in Bergeron, Chaussinand-Nogaret 1979, 14, 43. Piedmontese and Ligurian members of departmental electoral colleges mostly belonged to the same age class. Cf. Violardo 1995, 85; Beaurepaire-Hernandez 2014, 350-1.

aspire to a career in officialdom.[39] This elucidates why no application for a position within the candidate's own department includes the word *avancement*. This finding is consistent with Lignereux's analysis of the careers of the *Impériaux*, who considered appointments outside France as promotions in 40% of cases. It suggests that candidates rightly perceived career advancement as contingent upon seeking positions beyond their department.[40] Even if this underscores the internalization of certain professional behaviours, it is worth noting that, when mentioned, even departments outside national borders were typically not entirely unfamiliar to applicants requesting them. Furthermore, a rhetorical appeal grounded in hardships was utilized to pursue career progression, alongside references to prior services rendered. The convergence of these factors emphasizes that Napoleonic applications served as a nexus between Ancien Régime pleas and contemporary employment submissions, yet the transition to the latter was far from complete.

6 Conclusion

The various text analysis methods delineated in this study each harbour distinct strengths and weaknesses, rendering them apt for addressing particular inquiries while less suitable for others. Topic modelling and word embedding algorithms, for instance, excel in navigating vast corpora typically encountered in large-scale collaborative ventures. Conversely, statistical analyses anchored in the notion of distinctiveness can yield robust findings even when applied to smaller corpora, a common scenario in individual projects. Moreover, statistical analysis readily accommodates metadata.

Regardless of their level of complexity, the choice between these methods ultimately hinges on a decision between an inductive and a deductive approach. Rather than relying on exploratory techniques, the NapApps project adopted an analytical stance to scrutinize employment applications from the Napoleonic era, specifically probing candidates' inclination to relocate as an indicator of professional behaviour. This inquiry entailed examining correlations among several variables pertinent to both candidates and applications, such as age,

[39] "An elite, a nobility cannot be conceived at the beginning of this century without landed property" (Tulard 1975, 222). Members of the departmental electoral colleges were chosen among the six hundred largest taxpayers. However, those to be appointed were not necessary the wealthiest, since their position, social background, and reputation were also considered (Tulard 1975, 224-5).

[40] Although this percentage was lower for personnel of the prefectures (1/3 of first appointments), it would increase if identical positions in better-paid, higher-class prefectures were also considered promotions (Lignereux 2019, 199-202).

department of origin, position sought, and desired department. Further analysis of the latter variable involved comparing the most distinctive words used by candidates applying for positions within their own department versus those who did not. Rather than prioritizing visualization, which, despite its strengths in Digital Humanities, can sometimes be misleading,[41] the focus was on statistical analysis, complemented by an examination of individual cases.

Subjected to the critique that targeted quantitative history from the 1960s to the 1980s and its subsequent decline in later decades, statistics should nonetheless be regarded as an asset by historians, particularly within the realm of digital projects. This should be embraced without apprehension of diluting the essence of the historian's craft.[42] As François Furet noted in a 1982 article reflecting on the distinction between narrative history and problem-oriented history, statistical analysis facilitates description rather than interpretation and explanation. Even when history is approached through a problem-oriented lens, historians are tasked with interpreting the mechanisms "through which a probable pattern of collective behaviour – the very one revealed by data analysis – is manifested in individual behaviour during a given period" (Furet 1984, 93-4, 99). This necessitates a dialogue between quantitative and qualitative methods, macro-analysis, and micro-analysis. Numerous studies have showcased the fruitful interaction between these two methods of analysis and scales of observation (Karila-Cohen et al. 2018, 782-3).[43] It is precisely through the integration of general phenomena and individual trajectories that historians can underscore the relevance of their work in the era of 'big data'.

[41] On visualization as a narrative technique, Hullman, Diakopoulos 2011, 2231-40.

[42] Handbooks tailored for historians seeking to acquaint themselves with statistics encompass works such as those by Haskins, Jeffrey 2011, as well as Hudson, Ishizu 2016. For insights into the trajectory of quantitative history in the United States, Ruggles 2021, 1-25. To explore historical perspectives on criticism and evolving methodologies in quantification, particularly in Europe and France, consult Lemercier, Zalc 2013, 135-64.

[43] Examples are present in the same issue of the *Annales*.

Bibliography

Andersen, E. (2022). "From Search to Digital Search. An Exploration Through the Transnational History of Psychiatry". Fickers, Tatarinov 2022, 131-57. https://doi.org/10.1515/9783110723991-007

Anthony, L. (2005). "AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom". *Proceedings. IEEE International Professional Communication Conference* (Limerick, 2005), 729-37. https://doi.org/10.1109/ipcc.2005.1494244

Barthélemy, H.-C.-F. (1885). *Souvenirs d'un ancien préfet (1787-1848)*. Paris: E. Dentu.

Beaurepaire-Hernandez, A. (2014). "Un modèle de notable européen? Les 'masses de Granit' des départements liguriens et leur intégration au système impérial". Antoine, F. et al. (éds), *L'Empire napoléonien. Une expérience européenne?*. Paris: Armand Colin, 347-58.

Bergeron, L.; Chaussinand-Nogaret, G. (1979). *Les 'masses de granit'. Cent mille notables du Premier Empire*. Paris: École des hautes études en sciences sociales.

Blaney, J. et al. (2021). *Doing Digital History: A Beginner's Guide to Working with Text as Data*. Manchester: Manchester University Press. https://doi.org/10.7765/9781526157713

Blanke, T.; Bryant, M.; Hedges, M. (2020). "Understanding Memories of the Holocaust. A New Approach to Neural Networks in the Digital Humanities". *Digital Scholarship in the Humanities*, 35(1), 17-33. https://doi.org/10.1093/llc/fqy082

Blaufarb, R. (2002). *The French Army (1750-1820). Careers, Talent, Merit*. Manchester; New York: Manchester University Press.

Brauer, R.; Fridlund, M. (2013). "Historizing Topic Models: A Distant Reading of Topic Modeling Texts Within Historical Studies". Nikiforova, L.V.; Nikiforova, N.V. (eds), *Cultural Research in the Context of "Digital Humanities" = Proceedings of International Conference* (St. Petersburg, 3-5 October 2013). St. Petersburg: Herzen State Pedagogical University & Publishing House Asterion, 152-63.

Broers, M. (1996). *Europe Under Napoleon 1799-1815*. London; New York: Arnold.

Broers, M. (2005). *The Napoleonic Empire in Italy, 1796-1814. Cultural Imperialism in a European Context?*. New York: Palgrave Macmillan.

Broers, M. (2016). "'Les Enfants du Siècle': An Empire of Young Professionals and the Creation of a Bureaucratic, Imperial Ethos in Napoleonic Europe". Crooks, P.; Parsons, T.H. (eds), *Empires and Bureaucracy in World History: From Late Antiquity to the Twentieth Century*. Cambridge: Cambridge University Press, 344-63. http://dx.doi.org/10.1017/cbo9781316694312.015

Bunout, E.; Ehrmann, M.; Clavert, F. (eds) (2023). *Digitised Newspapers. A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology*. Berlin; Boston: De Gruyter. https://doi.org/10.1515/9783110729214

Buongiorno, S. et al. (2022). *The Hansard 19th-Century British Parliamentary Debates with Improved Speaker Names: Parsed Debates, N-Gram Counts, Special Vocabulary, Collocates, and Topics*. Harvard Dataverse, V2. http://doi.org/10.7910/DVN/ZCYJH8

Cohen, D. (2017). "Commis et fonctionnaires, entre service du public et droits de l'individu, de 1792 à l'an IV". *Annales historiques de la Révolution française*, 3, 101-17.

Corfield, P.J.; Hitchcock, T. (2022). "Using Technology Creatively: Digital history". Corfield, P.J.; Hitchcock, T. (eds), *Becoming a Historian. An Informal Guide*. London: University of London Press; Institute of Historical Research, 93-102.

Cortelazzo, M.A. (2013). "Metodi qualitativi e quantitativi di analisi dei testi". *Contemporanea*, 16(2), 299-310.

Crymble, A. (2021). *Technology and the Historian: Transformations in the Digital Age*. Urbana: University of Illinois Press.
https://doi.org/10.5406/j.ctv1k03s73

Dal Cin, V. (2023). "Candidarsi a un impiego in età napoleonica. Riflessioni a partire da una ricerca in corso". *Passato e presente*, 119, 53-68.

De Zuylen de Nyevelt, S.I. (1893). "The Exile of the Marquise the Falaiseau". *The Living Age*, 198(3), 350-7.

Dunne, J. (2007). "Power on the Periphery: Elite-State Relations in the Napoleonic Empire". Dwyer, P.G.; Forrest, A. (eds), *Napoleon and His Empire. Europe, 1804-1814*. Houndmills; Basingstoke; New York: Palgrave Macmillan, 61-78.

Durand, C. (1958). *Les auditeurs au Conseil d'état de 1803 à 1814*. Aix-en-Provence: La Pensée Universitaire.

Ellis, G. (1997). *Napoleon*. London; New York: Longman.

Fickers, A.; Tatarinov, J. (eds) (2022). *Digital History and Hermeneutics. Between Theory and Practice*. Berlin; Boston: De Gruyter.

Forrest, A. (2011). *Napoleon*. London: Quercus.

Furet, F. (1984). "From Narrative History to Problem-oriented History". Furet, F. (ed.), *In the Workshop of History*. Chicago; London: The University of Chicago Press, 54-67. Transl. of: *L'atelier de l'histoire*. Paris: Flammarion.

Godechot, J. [1951] (1985). *Les institutions de la France sous la Révolution et l'Empire*. 3rd ed. Paris: Presses Universitaires de France.

Grab, A. (2003). *Napoleon and the Transformation of Europe*. New York: Palgrave Macmillan.

Graham, S. et al. (2022). "Topic Modeling: A Hands-On Adventure in Big Data". Graham, S. et al. (eds), *Exploring Big Historical Data: the Historian's Macroscope*. 2nd ed. Hackensack (NJ): World Scientific, 115-54.
https://doi.org/10.1142/9789811243042_0004

Graham, S. et al. (2022). *Exploring Big Historical Data: the Historian's Macroscope*. 2nd ed. Hackensack (NJ): World Scientific.

Guilhaumou, J. (1986). "L'historien du discours et la lexicométrie. Étude d'une série chronologique: le 'Père Duchesne' d'Hébert (Juillet 1793-Mars 1794)". *Histoire & Mesure*, 1(3-4), 27-46.
https://doi.org/10.3406/hism.1986.1529

Guildi, J. (2020). "The Common Landscape of Digital History: Universal Methods, Global Borderlands, Longue-Durée History, and Critical Thinking about Approaches and Institutions". Fridlund, M.; Oiva, M.; Paju, P. (eds), *Digital Histories. Emergent Approaches Within the New Digital History*. Helsinki: Helsinki University Press, 327-46.
https://doi.org/10.33134/HUP-5-18

Guldi, J. (2022). "The Algorithm. Mapping Long-Term Trends and Short-Term Change at Multiple Scales of Time". *The American Historical Review*, 127(2), 895-911.
https://doi.org/10.1093/ahr/rhac160

Guldi, J. (2023). *The Dangerous Art of Text Mining*. Cambridge: Cambridge University Press.

Hakkarainen, H.; Iftikhar, Z. (2020). "The Many Themes of Humanism: Topic Modelling Humanism Discourse in Early 19th-Century German-Language Press". Fridlund, M.; Oiva, M.; Paju, P. (eds), *Digital Histories. Emergent Approaches within the New Digital History*. Helsinki: Helsinki University Press, 259-77.
https://doi.org/10.2307/j.ctv1c9hpt8.20

Haskins, L.; Jeffrey, K. [1990] (2011). *Understanding Quantitative History*. Eugene: Resource Publications.

Hudson, P.; Ishizu, M. (2016). *History by Numbers: An Introduction to Quantitative Approaches*. London: Bloomsbury Publishing.

Hullman, J.; Diakopoulos, N. (2011). “Visualization Rhetoric: Framing Effects in Narrative Visualization”. *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2231-40.
https://doi.org/10.1109/TVCG.2011.255

Jockers, M.L.; Underwood, T. (2016). “Text-Mining the Humanities”. Schreibman, S.; Siemens, R.; Unsworth, J. (eds), *A New Companion to Digital Humanities*. Malden: Wiley Blackwell, 291-306.
https://doi.org/10.1002/9781118680605.ch20

Kamlovskaya, E. (2022). “Exploring a Corpus of Indigenous Australian Autobiographical Works With Word Embedding Modeling”. Fickers, Tatarinov 2022, 87-108.

Karila-Cohen, K. et al. (2018). “Nouvelles cuisines de l’histoire quantitative”. *Annales. Histoire, Sciences sociales*, 73(4), 773-83.

Karila-Cohen, P. (2021). *Monsieur le Préfet. Incarner l’État dans la France du XIXe siècle*. Ceyzérieu: Champ Vallon.

Laharie, P.; Lamoussière, C. (1998). *Le Personnel de l’administration préfectorale, 1800-1880. Répertoires nominatif et territorial*. Paris: Centre historique des Archives Nationales.

Lässig, S. (2021). “Digital History. Challenges and Opportunities for the Profession”. *Geschichte und Gesellschaft*, 47(1), 5-34.
https://doi.org/10.13109/gege.2021.47.1.5

Lemercier, C. (2019). “L’analisi testuale”. Paci, D. (a cura di), *La storia in digitale. Teorie e metodologie*. Milano: Unicopli, 293-4.

Lemercier, C.; Zalc, C. (2013). “Le sens de la mesure: nouveaux usages de la quantification”. Granger, C. (éd.), *À quoi pensent les historiens? Faire de l’histoire au XXIe siècle*. Paris: Autrement, 135-64.

Lemercier, C.; Zalc, C. (2019). *Quantitative Methods in the Humanities: an Introduction*. Charlottesville: University of Virginia Press.
https://doi.org/10.2307/j.ctvbqs963.6

Lentz, T. et al. (2008). *Quand Napoléon inventait la France: dictionnaire des institutions politiques, administratives et de cour du Consulat et de l’Empire*. Paris: Tallandier.

Levati, S. (2009). “Les notables napoléoniens: du cas français à celui italien”. *Rives méditerranéennes*, 32-3, 215-28.
https://doi.org/10.4000/rives.2969

Lignereux, A. (2012). *Servir Napoléon. Policiers et gendarmes dans les départements annexés (1796-1814)*. Seyssel: Champ Vallon.

Lignereux, A. (2019). *Les Impériaux. Administrer et habiter l’Europe de Napoléon*. Paris: Fayard.

McCain, S. (2018). *The Language Question under Napoleon, 1799-1814*. Cham: Palgrave Macmillan.

Moretti, F. (2013). *Distant Reading*. London; New York: Verso.

Moretti, F. (2022). *Falso movimento: la svolta quantitativa nello studio della letteratura*. Milano: Nottetempo.

Nanni, F.; Kuemper, H.; Ponzetto, S.P. (2016). “Semi-Supervised Textual Analysis and Historical Research Helping Each Other: Some Thoughts and Observations”. *International Journal of Humanities and Arts Computing*, 10(1), 63-77.
https://doi.org/10.3366/ijhac.2016.0160

Oberbichler, S.; Pfanzelter, E. (2023). “Tracing Discourses in Digital Newspaper Collections. A Contribution to Digital Hermeneutics while Investigating ‘Return Migration’ in Historical Press Coverage”. Bunout, Ehrmann, Clavert 2023, 126-52.
https://doi.org/10.1515/9783110729214-007

Robertson, S. (2016). “The Differences Between Digital Humanities and Digital History”. Gold, M.K.; Klein, L.F. (eds), *Debates in the Digital Humanities 2016*. Minneapolis: University of Minnesota Press, 289-307.

Romein, C.A. et al. (2020). “State of the Field: Digital History”. *History*, 105(365), 291-312. https://doi.org/10.1111/1468-229X.12969

Ruggles, S. (2021). “The Revival of Quantification: Reflections on Old New Histories”. *Social Science History*, 45, 1-25. https://doi.org/10.1017/ssh.2020.44

Salmi, H. (2021). *What is Digital History?* Cambridge: Polity Press.

Samuel, J.; Rozzi, G.; Palle, R. (2022). “The Dark Side of Sentiment Analysis: An Exploratory Review Using Lexicons, Dictionaries, and a Statistical Monkey and Chimp”. http://dx.doi.org/10.2139/ssrn.4000087

Scholz, L. (2022). “A Distant Reading of Legal Dissertations from German Universities in the Seventeenth Century”. *The Historical Journal*, 65, 297-327. https://doi.org/10.1017/S0018246X2100011X

Sinclair, S.; Rockwell, G. (2016). “Text Analysis and Visualization: Making Meaning Count”. Schreibman, S.; Siemens, R.; Unsworth, J. (eds), *A New Companion to Digital Humanities*. Malden: Wiley Blackwell, 274-90. https://doi.org/10.1002/9781118680605.ch19

Story, D.J. et al. (2020). “History’s Future in the Age of the Internet”. *The American Historical Review*, 125(4), 1337-46. https://doi.org/10.1093/ahr/rhaa477

Thoral, M.C. (2010). *L’émergence du pouvoir local: le département de l’Isère face à la centralisation napoléonienne (1800-1837)*. Rennes: Presses Universitaires de Rennes.

Tripodi, R. et al. (2019).“Tracing Antisemitic Language Through Diachronic Embedding Projections: France 1789-1914”. *Proceedings of the 1st International Workshop on Computational Approaches to Historical* Language *Change*. Firenze: Association for Computational Linguistics, 115-25. https://doi.org/10.18653/v1/W19-4715

Tulard, J. (1975). “Les notables impériaux”. Chaussinand-Nogaret, G. (ed.), *Une histoire des élites, 1700-1848*. Paris; La Haye: Mouton éditeur, 218-34.

Tulard, J.; Tulard, M.-J. (2014). *Napoléon et 40 millions de sujets. La centralisation et le Premier Empire*. Paris: Editions Tallandier.

Van der Burg, M. (2021). *Napoleonic Governance in the Netherlands and NorthWest Germany: Conquest, Incorporation, and Integration*. Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-030-66658-3

Vetter, C.; Marin, M.; Gon, E. (2015). *Dictionnaire Robespierre. Lexicométrie et usages langagiers. Outils pour une histoire du lexique de l’Incorruptible*, t. 1. Trieste: Edizioni Università di Trieste.

Violardo, M. (1995). *Il notabilato piemontese da Napoleone a Carlo Alberto*. Torino: Comitato di Torino dell’Istituto per la storia del Risorgimento italiano.

Weber, M. (1981). *Economia e società*. Vol. 4, *Sociologia politica*. Milano: Edizioni di Comunità. Transl. of: *Wirtschaft und Gesellschaft*. Tübingen: Mohr.

Wevers, M.; Koolen, M. (2020). “Digital Begriffsgeschichte: Tracing Semantic Change Using Word Embeddings”. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(4), 226-43. https://doi.org/10.1080/01615440.2020.1760157

Whitcomb, E.A. (1974). “Napoleon’s Prefects”. *The American Historical Review*, 79, 1089-118. https://doi.org/10.2307/1869564

Woloch, I. (2001a). “The Napoleonic Regime and French Society”. Dwyer, P.G. (ed.), *Napoleon and Europe*. New York: Longman, 60-78.

Woloch, I. (2001b). *Napoleon and His Collaborators. The Making of a Dictatorship*. New York; London: Norton & Company.

Woolf, S. (1991). *Napoleon's Integration of Europe*. London; New York: Routledge.