

Is There a Patent Genre? A Computational Text Analysis of French Patents, 1903–1940

Matti La Mela^{1,2} , and Yunting Xie³ 

¹ Paris Institute for Advanced Study, France

² Department of ALM, Uppsala University, Sweden

³ Department of Business Studies, Uppsala University, Sweden

Abstract

Patents are legal documents that grant temporary exclusive rights over inventions and describe their technical features. They have become an important source for studying innovation and technological change, including through computational analyses of patent texts. Yet patent applications are not only repositories of technical knowledge, but also shaped by institutional and linguistic conventions that structure their writing. This paper approaches patents as a rhetorical genre by analysing a corpus of French patent descriptions from 1903–1940. Using computational text analysis, it examines both the stability and variation of patent language. First, we analyse the persistence of frequent bigrams over time to identify repeating lexical patterns. Second, we measure lexical differences between patents from different countries filed in France using Jensen–Shannon distance. The results show a stable core of word constructions in patent descriptions. We also observe differences between countries that show little systematic convergence over time. We find persistent closeness within linguistic regions, for instance between US and UK patents and among German-language or Romance-language countries. This points to the persistence of shared linguistic conventions and patent practices that survive translation into French. The findings highlight the importance of approaching patents as historically situated textual artifacts.

Keywords: innovation, genre, patent history, computational text analysis, France

1 Introduction

Patents are temporary exclusive rights granted to inventors over the commercial use of their inventions. In the patent application, which ultimately becomes the published patent, the applicant describes in text and drawings the invention and the novel elements that the patent protects. Patents can be seen as embodying state-of-the-art technical knowledge, which has made them one of the key sources for studying innovation and technological change ([14], for limitations of patents, see [22]). With the digitization of patent sources internationally (e.g. [5; 28]), computational text analysis approaches have been used to discern novelty from patent texts and to identify breakthrough patents among the thousands granted over the years (e.g. [18; 20]). In these computational approaches, patents are typically treated as containers of technical information expressed in the text, setting aside the textual conventions that shape how this information is written.

At the same time, patent applications must conform to formal expectations from patent offices, and since inventions relate to previous work and existing problems, they refer internally to the existing body of patent texts. In this sense, patents form a specific genre; Burk and Reyman [7, p. 181] elaborate how “patents constitute a distinct genre, not only in the more traditional rhetorical sense of displaying uniquely characteristic documentary features, but in the more modern sense of association with a discrete and identifiable community that has shaped and has been shaped by the

Matti La Mela, and Yunting Xie. “Is There a Patent Genre? A Computational Text Analysis of French Patents, 1903–1940.” In: *Computational Cultural Science Workshop 2026*, ed. by Suzanne Mpouli, Jean-Baptiste Camps, Florian Cafiero, and Chahan Vidal-Gorène. Vol. 6. Anthology of Computers and the Humanities. 2026, 57–72. <https://doi.org/10.63744/I4y3L7pBTOyo>.

© 2026 by the authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

document.” How does this affect computational analyses that treat patents simply as a corpus of technical text, given that the language is structured and contextually bound within the community of patent practitioners?

This paper studies the patent genre computationally through French patent texts from the early twentieth century, treating patents as shaped by both repetition of textual elements shared within an expert community and variation arising from drafting choices. We examine the patent genre by analysing its stability through bigrams in the full corpus, assuming that recurring multi-word patterns reflect shared drafting conventions rather than technical content only. Variation is examined through lexical distributions measured by the Jensen-Shannon distance. As one possible source of variation, or the existence of separate national patent genres, we investigate patents filed in France by patentees from different countries.

If patents form a genre shaped by a national patent community, international patenting offers a useful test case. Patenting expanded and became increasingly international from the late nineteenth century (e.g. [19]). Inventors needed to file patent applications in every country where they sought to protect their interests [2]; they relied on local patent agents familiar with the local jurisdiction and language [1; 24; 27]. Applicants were aware of institutional differences, and could draft strategically across jurisdictions, for instance by disclosing less information in some [17; 26]. However, at times the same text could be reused directly between countries with similar patent cultures and a shared language, such as Sweden and Finland [2]. International harmonization accelerated from the 1880s onward with the first international agreements.¹

The paper asks whether patenting in France required adhering to certain linguistic conventions shared within the French patent system. It also examines whether patents originating abroad carried traces of drafting conventions from their home jurisdiction, or broader practices of using technical language, producing measurable differences between countries of origin even after translation into French. Whether such differences diminish over time is an open question: as Gooday and Wilf [13] emphasize, the harmonization of patent systems has remained partial, and national patent systems have persisted alongside it well into the twentieth century.

The paper makes three contributions. First, it revisits computational approaches that take patents as containers of technical knowledge and instead approaches them as historical artifacts that constitute a specific textual style. This contextualization is relevant for understanding applicant choices, differences between patent systems, and what counts as technical novelty presented in patent applications. Second, the paper provides computational evidence for the persistence of national patent cultures. Third, it discusses the quality of the current French patent data, which is adequate for large-scale textual analysis, but would benefit from re-OCR and metadata reconstruction. Together, these contributions highlight the importance of contextual awareness and source quality in computational analysis of historical (textual) materials (for related critical approaches to historical text as data, see e.g. [16]).

2 Theory and Previous Work

Genre has been used across linguistics and media studies to visual arts and literary studies. In general, genre has been used to classify literature and works of art into broader types [9]. In other words, genre refers to shared conventions of content and form within a group of works. At the same time, genres have a dynamic element, as each new work both follows and contributes to the genre. Neale [23, p. 56] emphasizes that genres can be understood as processes: “dominated by repetition, but [...] also marked fundamentally by difference, variation, and change.” As Neale continues, this processual interaction takes place at three levels: expectations, the generic corpus,

¹ For example, the International Paris Convention was established in 1883, but countries such as Germany, Austria–Hungary, and Canada, only joined the Convention in the early twentieth century.

and the norms that govern both.

Besides the processual nature, genre theorists emphasize the communicative situations that genres make possible. Miller [21] conceptualizes genre as social action: rhetorical action that acquires meaning from the situations where it arises and connects private intentions with socially set rules. In other words, genre gives order to language-use: communication occurring in recurring situations, participants familiar with these contexts can make sense of others' communication and structure their communication through these situations [3]. Burk and Reyman [7, pp. 178–179] theorize the patent as a rhetorical genre, typified through dimensions such as “purpose”, “content”, or “participants”.

This paper approaches genre through its features of continuity (repetition and expectations) and variation (difference and authorial choice). Texts within a genre follow conventions recognized by the community, but also depart from them. If patents constitute a genre, we should expect to find recurring linguistic patterns across documents that reflect shared rhetorical conventions, such as introducing and describing the invention in standardized ways. Such patterns of continuity and variation can be detected computationally, by examining both what persists across the corpus and what differs between the groups of patents.

The historical patent genre has been studied by Groom and Grieve [15], who investigate rhetorical moves in 130 British patent specifications (1711–1860) to understand why genres change. They show that certain coded elements are very stable in the patent genre, particularly structural features such as the act of granting it, invention description, and witnesses statement or signature. They also show that legal reforms influence how patent genre evolves. In the current paper, we identify such elements computationally by looking at recurring textual patterns. OCR limitations prevent us from discerning the different sections of the patent documents, but we can still observe structural stability. Figure 1 presents an example that shows structural elements of a patent document, which remain relatively constant across the period.

Related to patents, a strand of research has examined genre in financial statements [25], which, like patents address multiple audiences including government authorities, business competitors, or potential partners. Rutherford studies financial reporting using corpus linguistic methods, and identifies differences in financial reporting across company types. Similarly, we investigate recurring patterns and textual similarity, but focus on patents filed in France by applicants from different countries. Another genre type that can be contrasted with patents is the scientific article, which similarly situates new knowledge in relation to existing research.

Finally, similar to this paper, there is a growing body of literature that approaches historical patent texts computationally. These studies apply natural language processing and machine learning methods to examine aspects such as novelty and technological change [18; 20] and patent classification [6].

3 Data and Method

For the analysis, we use a corpus of patents from France between 1903 and 1940. The data is available through the National Institute of Intellectual Property (INPI) and the data services of the European Patent Office (EPO). For this paper, we downloaded patent data—bibliographic information and patent descriptions—through the EPO API.² The bibliographic information includes names of actors, their country of residence, grant and application date, and technical classification. Our analysis begins with the first complete year in the EPO dataset, 1903, and ends with the beginning of the Second World War in 1940. Appendix A presents yearly French patent statistics and discusses the coverage, quality, and textual scope of the EPO data, also in comparison with other

² See the EPO Open Patent Services documentation: <https://www.epo.org/en/searching-for-patents/data/web-services/ops>

RÉPUBLIQUE FRANÇAISE.
OFFICE NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE.

BREVET D'INVENTION.

V. — Machines.

3. — ORGANES, ACCESSOIRES ET ENTRETIEN DES MACHINES.

N° 420.004

Mécanisme de commande applicable aux mouvements d'ailes des aéroplanes.

M. René TAMPIER résidant en France (Seine).

Demandé le 10 novembre 1909.

Délivré le 11 novembre 1910. — Publié le 20 janvier 1911.

[Brevet d'invention dont la délivrance a été ajournée en exécution de l'art. 11 § 7 de la loi du 5 juillet 1844 modifiée par la loi du 7 avril 1902.]

L'invention a pour objet une disposition mécanique permettant de donner à un arbre des orientations quelconques dans l'espace par l'intermédiaire d'un système de trois axes rectangulaires reliés entre eux par des transmissions convenablement disposées et commandées, telles que des roues dentées par exemple. Les arbres à orienter pourront être, par exemple, ceux qui servent de supports aux surfaces portantes et gouvernantes des aéroplanes, auxquels on pourra donner ainsi tous les mouvements nécessaires pour la direction et les changements d'allure.

Aux dessins annexés :

La fig. 1 est une vue en plan du mécanisme faisant l'objet de l'invention;

La fig. 2 est une coupe prise suivant la ligne A-A de la fig. 1;

La fig. 3 est une vue à plus petite échelle du dispositif de commande;

La fig. 4 représente l'application du mécanisme précédent à l'obtention d'un mouvement similaire.

L'appareil comprend essentiellement trois arbres 1, 2, 3, disposés perpendiculairement les uns aux autres et reliés entre eux de la manière suivante : l'arbre 1, qui sert de support à tout l'appareil, est fixé au bâti de toute

façon convenable; l'appareil est symétrique par rapport à l'arbre 1. L'arbre 2 est solidaire, à l'une de ses extrémités, d'une douille 4 pouvant tourner sur l'arbre 1 et portant à son autre extrémité une douille 5 dans laquelle tourne l'arbre 3. La douille 5 est supportée par un berceau 5' prenant appui par des manchons sur l'arbre 1. Un arbre 6, solidaire de l'arbre 3, supporte la carcasse 7 de l'aile. Une roue 8 à denture conique est fixée sur l'arbre 1 et engrène avec une roue à double denture conique 9 montée folle sur l'arbre 2; cette roue 9 engrène d'autre part avec un segment de roue 10 à denture conique clavetée sur l'arbre 3. Une roue à denture conique 11, montée folle sur l'arbre 6, engrène avec une roue 12 à denture conique montée folle sur l'arbre 3, qui engrène elle-même avec la roue 9. Les deux berceaux 5' s'appuient sur l'arbre 1 par l'intermédiaire de manchons 13 et 14 montés concentriquement.

Des roues dentées 15 et 16 sont fixées respectivement sur un des systèmes de manchons 13 et 14.

Un cadre triangulaire gauche, formé par exemple de tubes 17, est supporté à l'un de ses sommets par l'arbre 1 et porte, à ses autres sommets, deux axes 18 sur lesquels

Prix du fascicule : 1 franc

Figure 1: An example patent: The first page of the French patent no 420.004 granted in November 1910

aggregated datasets such as PatentCity [4].

The corpus was preprocessed in several steps. Patents of addition (E-series) were excluded to focus on main patent applications. First, we addressed gaps in the bibliographic metadata. Missing publication years in the EPO metadata were inferred from patent numbers, which follow a chronological sequence. The EPO data lacks country information after 1921, which was supplemented using PatentCity metadata. Second, the patent texts were processed (see Appendix A for XML tags and document lengths). Patent texts were stripped of XML tags (present in 46% of documents), related to formatting, unrecognized characters or graphical elements. Patent descriptions containing 50 tokens or fewer were excluded to remove incomplete OCR output. The patent texts were tokenized, converted to lowercase, accents were removed (“é” → “e”), and reduced to alphanumeric tokens. Finally, very rare tokens were filtered out to reduce OCR noise, removing tokens that appeared in less than four patents. After preprocessing, the corpus contains 458,816 patents. Because OCR coverage varies, with substantial gaps in 1920, 1921, 1929, and 1930 (see Appendix A), the analysis examines language change by period rather than by year.

To analyse the linguistic characteristics of the patent corpus, we apply two approaches that capture aspects of textual stability and variation. First, we analyse the most common bigrams in the patent descriptions and examine their persistence over time. Bigrams capture two-word combinations, that can correspond to formulaic expressions used in patent writing. Bigrams, and ngrams more broadly, have been used to study similarities in genre and register between texts (e.g. [10]). Several filtering steps were applied to focus on the language of patent descriptions (see Appendix

B); trigrams were also investigated, but bigram results are presented given the paper’s scope. The most frequent bigrams were examined iteratively to remove administrative vocabulary originating from patent headers (present only in part of the material due to the OCR process), such as “published”, “patent”, or “intellectual property”. In addition, bigrams with little semantic relevance consisting of function words only, such as “qu ils” and “et dont”, were removed. Bigrams containing articles (e.g. “la”, “un”, “du”) were also removed, as these reduce to unigrams in practice. The filtering vocabularies were drawn from standard French function-word lists. After filtering, bigrams were extracted and counted by time periods. To measure the stability of the lexical patterns, we examine the overlap and persistence of bigrams over time.

Second, we examine variation in patent texts filed in France by patentees from different countries (the texts are thus written in French and follow the French application process). We use patents from residents of countries that patented the most in France, and which also belong to different linguistic areas: France (domestic residents), Germany, United States, United Kingdom, Switzerland, Belgium, Italy, Austria, and Sweden. Figure 2 shows how the shares of the patents in our corpus from these countries evolved over time, and Table 1 reports the patent counts per country in the two periods used for the Jensen-Shannon analysis.

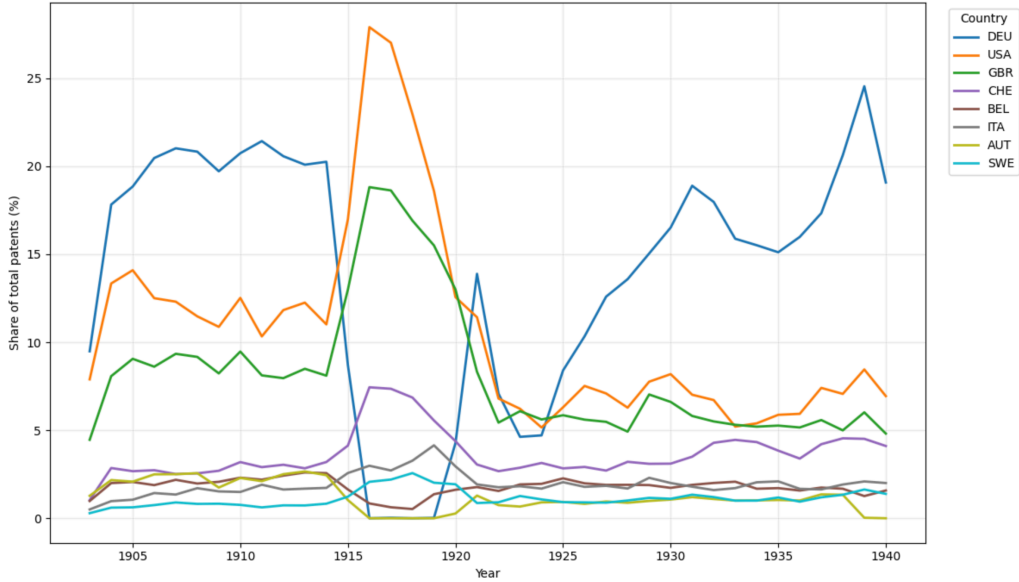


Figure 2: Share of foreign patents in France, 1903–1940. Shares are calculated relative to all patents in the study corpus, for the countries with most patents.

We compute the Jensen-Shannon distance to measure lexical similarity between patent texts from different countries. The Jensen-Shannon distance is calculated from the probability distributions of tokens in two corpora. The two distributions (P and Q) represent the frequencies of tokens in the patent corpora that are compared. For each token (i), the probability in each corpus (p_i and q_i) is calculated from its relative frequency. The distributions are then compared, and values range between 0 and 1, where 0 indicates complete similarity (identical token distributions) and 1 indicates complete divergence.

$$d_{JS}(P, Q) = \sqrt{\frac{1}{2} \sum_i p_i \log_2 \left(\frac{p_i}{m_i} \right) + \frac{1}{2} \sum_i q_i \log_2 \left(\frac{q_i}{m_i} \right)}$$

$$\text{where } m_i = \frac{1}{2}(p_i + q_i)$$

Table 1: Number of patents per country in the early (1903–1912) and late (1931–1940) periods used in Jensen-Shannon distance analysis.

Country	1903–1912	1931–1940
France	56,785	62,971
Germany	22,919	23,970
United States	13,995	8,782
United Kingdom	9,860	7,326
Switzerland	3,142	5,569
Belgium	2,414	2,418
Italy	1,642	2,490
Austria	2,609	1,364
Sweden	825	1,618

4 Results

4.1 Stability in Patent Documents through Bigrams

We observe stability of the lexical combinations used in patent descriptions through the overlap of the most frequent bigrams between the periods analysed. Table 2 presents two measures of bigram stability across the periods. The first column describes the rolling overlap of the one hundred most common bigrams: the first period is 1903–1909, and the overlap is calculated between consecutive periods. A value of 0.800 for 1910–1915 means that 80% of the bigrams from the first period (1903–1909) are also present in the following period (1910–1915). The values remain high across the entire study period, ranging between 0.75 and 0.94, which suggests that patent texts rely on a stable set of recurring lexical patterns. In addition, we find that the stability of the bigrams increases over time: in the 1930s, the rolling overlap is around 90%. A temporary decrease appears around the war years (visible in the period 1921–1925), which probably reflects a shift in the types of patent being filed, as also suggested by changes in patent length shown in Figure 4 (Appendix A). The change in OCR processing of documents around 1920 may also contribute to the differences (see Appendix A).

Table 2: Stability of frequent bigrams across periods. “Rolling” is the share of top-100 bigrams from the previous period that also appear in the current period’s top-100. The “vs. 1903–1909” columns present the share of Top-k bigrams from 1903–1909 (the baseline period) that appear in the current period’s top-100.

Period	Rolling	vs. 1903–1909		
	Top-100	Top-20	Top-50	Top-100
1903–1909	–	1.000	1.000	1.000
1910–1915	0.800	0.850	0.880	0.800
1916–1920	0.840	0.850	0.880	0.710
1921–1925	0.750	0.800	0.840	0.710
1926–1930	0.940	0.800	0.820	0.680
1931–1935	0.890	0.700	0.740	0.640
1936–1940	0.890	0.700	0.740	0.610

The remaining columns of Table 2 present the share of the 20, 50, and 100 most frequent bigrams from 1903–1909 that still appear among the top-100 in later periods; For example, the value

0.700 for Top-20 in 1936–1940 indicates that 70% of the earliest period’s (1903–1909) Top-20 bigrams remained among the Top-100 in 1936–1940. The results show that even the full Top-100 bigrams from 1903–1909 persists strongly: 61% are still present in 1936–1940. The vocabulary gradually changes, creating increasing “distance” from the original distribution. As with the rolling overlap in the first column, patterns appear more stable after the First World War.

Table 3: Most persistent bigrams in French patents. The full list of the 100 most frequent bigrams is in Appendix B.

Rank	Bigram	Patents (n)	Share
1	presente invention	243641	64.94
2	invention a	190389	50.75
3	par exemple	187956	50.10
4	a titre	181334	48.33
5	au moyen	175432	46.76
6	pour objet	161689	43.10
7	fig est	142206	37.90
8	dessin annexe	131859	35.15
9	se trouve	129936	34.63
10	invention est	125935	33.57
11	facon a	112151	29.89
12	fig et	111436	29.70
13	par suite	111272	29.66
14	vue en	105645	28.16
15	maniere a	99189	26.44

Table 3 further illustrates the persistence of bigrams across the corpus, showing the 15 most frequent bigrams (extended list in Appendix B). Several expressions clearly belong to the language of patents: “présente invention” and “pour objet” introduce or frame the invention, whereas “fig est” refers to describing the invention and linking the text to the accompanying figures illustrating the invention.

4.2 Change in Patent Documents through Jensen-Shannon Distance

Tables 4 and 5 report the Jensen-Shannon distances between France and the main foreign patentees for the early (1903–1912) and late (1931–1940) periods. Distances were calculated using random samples of 800 patents per country, with the sample size set by the smallest available corpus (Swedish patents, 1903–1912; see Table 1). To assess robustness, the calculation was repeated with ten different random samples; the values reported are means of the samples, with standard deviations below 0.005 in all cells. Although the Jensen-Shannon distance is calculated on probability distributions, these distributions are estimated from finite text samples. For example, larger corpora may include more rare words, and can in this way increase the measured distance relative to smaller corpora. We also calculated distances for shorter consecutive periods (3-years, 5-years), to test converge over time, but found little systematic evidence; we therefore present two benchmark periods.

The Tables 4 and 5 suggest little or no convergence over time between French patent descriptions and those originating from abroad. For example, the Jensen-Shannon distance between France and the United Kingdom remains almost unchanged across the two periods, at 0.225 in 1903–1912 and 0.223 in 1931–1940. The distance between France and the United States decreases only slightly, from 0.239 to 0.226.

Table 4: Jensen-Shannon distances between national patent corpora (1903–1912). Note. Values are means of 10 random samples of 800 patents per country. Standard deviations for the samples were below 0.005 in all cells.

	FRA	DEU	USA	GBR	CHE	BEL	ITA	AUT	SWE
FRA		0.231	0.239	0.225	0.226	0.205	0.218	0.221	0.247
DEU			0.249	0.234	0.205	0.223	0.229	0.196	0.219
USA				0.181	0.242	0.244	0.242	0.230	0.234
GBR					0.231	0.227	0.228	0.219	0.223
CHE						0.225	0.229	0.205	0.228
BEL							0.217	0.217	0.234
ITA								0.219	0.231
AUT									0.212
SWE									

Table 5: Jensen-Shannon distances between national patent corpora (1931–1940). Note. Values are means of 10 random samples of 800 patents per country. Standard deviations for the samples were below 0.005 in all cells.

	FRA	DEU	USA	GBR	CHE	BEL	ITA	AUT	SWE
FRA		0.248	0.226	0.223	0.233	0.199	0.210	0.242	0.231
DEU			0.230	0.233	0.199	0.241	0.240	0.201	0.229
USA				0.175	0.229	0.225	0.221	0.229	0.205
GBR					0.224	0.222	0.218	0.226	0.207
CHE						0.230	0.226	0.210	0.218
BEL							0.208	0.230	0.226
ITA								0.229	0.219
AUT									0.228
SWE									

Instead, the distances reveal stable relationships between certain national patents (also visible in shorter periodizations we tested). For example, the UK and US patents are consistently among the most similar pair. In the early period, the distance between the United States and the United Kingdom is 0.181, and in the later period it remains similar at 0.175. By contrast, the distance between France and Germany is much larger, and even increases over time, from 0.231 in 1903–1912 to 0.248 in 1931–1940.

These patterns of national closeness likely reflect similarities in languages (and language use) that persist even through translation. A similar pattern can be observed for Germany, Austria, and also Switzerland, which show smaller distances from one another. Proximity can also be observed among Romance-language countries, such as Italy and Belgium, whose patent texts appear comparatively close to the French patents. Sweden appears as a relative outlier. Notably, Swedish patent texts align over time more closely with those of the United States and the United Kingdom rather than with German-language countries, despite linguistic proximity. This pattern suggests that factors other than language may play a role, including differences in technological specialization (and thus terminology used), the involvement of patent agents or translators, or similarities in patent application practices. Previous studies also show that international patenting patterns across countries remained relatively stable during this period [11].

5 Concluding Discussion

This paper has examined whether patents can be understood as forming a rhetorical genre. From this perspective, patent texts are not only technical descriptions but forms of communication shaped by institutional expectations and shared writing conventions in the patent community. We investigated both the stability and variation of patent language by analysing French patent descriptions from 1903–1940 with computational methods. If patents constitute a genre, we expected to observe recurring linguistic patterns related to patenting shared across documents, but also variation reflecting the choices of the actors drafting the texts and differences between national contexts.

The bigram analysis shows that patent descriptions have a stable core of lexical constructions that persist throughout the years analysed. These expressions reflect rhetorical functions of patents such as framing the invention or referring to accompanying figures. At the same time, the Jensen-Shannon distances between patents filed in France by applicants from different countries show differences that, while relatively small, show little or no systematic convergence over time. This may imply that the patent genre was already relatively stable at the beginning of the twentieth century, or rather, that national patent genres existed. This finding resonates with Gooday and Wilf [13], who emphasize how national patent cultures persisted alongside international harmonization efforts. Importantly, we find that certain patent cultures remain consistently closer to one another: patents from the United Kingdom and the United States, for example, remain among the most similar over time, even when written in French. Whether the same patterns can be observed in other jurisdictions, such as US and UK patents filed in Italy or Sweden, is a question for further comparative work.

Overall, French patents appear to have a stable core of patent rhetoric, around which variation occurs both in the surrounding bigrams and in the broader vocabulary. The national differences likely reflect linguistic proximity and similarities in patent culture, but also other factors such as technological specialization or the role of intermediating actors like patent agents or translators. Patent descriptions are thus not neutral carriers of technical information, but texts representing the inventions written according to conventions recognized within the patenting community. Such factors should be considered in computational analyses, which often treat patent documents primarily as data rather than as historically situated texts.

We suggest three directions for future work on patents as a rhetorical genre. First, long-term patent data from several countries should be used to examine the dynamics of national patent languages beyond French: given the partial nature of patent system harmonization, do we find convergence in a broader selection of cases? Second, further studies should investigate the roots of convergence between patent texts within the same linguistic region. Do we see multiple “patent cultures” carried by language communities, or differences resulting from distinct technology paths and long-term technological lock-in [8]? Accounting for technological specialization between countries can help explain the differences in Jensen-Shannon distances found here, and comparison with other contemporary technical texts (engineering and industry journals) could situate changes specific to the patent genre in relation to shifts in technical French at the time. Third, future research should examine the role of translation in international patenting. As Thomas Edison noted at the turn of the twentieth century, one key challenge for inventors was to translate their idea not only into other languages but also into their corresponding legal norms [17]. What role did translations produced by technology intermediaries play in the making of patent genres? Since the relevance of patents ultimately lies in communicating disclosed technologies to the public, including potential buyers and possible competitors, the clarity and rhetorical framing of the text reflect the intentions of the patentee and contribute to variation within the genre.

6 Acknowledgements

This paper benefited from a fellowship at the Paris Institute for Advanced Study as part of the Riksbankens Jubileumsfond Chair.

References

- [1] Andersson, David E. and La Mela, Matti. “Nordic Networks: Patent Agents and the Business of Technology Intermediation in Sweden and Finland, 1860–1910”. In: *Scandinavian Economic History Review* 68, no. 1 (2020), pp. 45–65.
- [2] Andersson, David E., La Mela, Matti, and Tell, Fredrik. “Family First: Defining, Constructing, and Applying Historical Patent Families”. In: *Explorations in Economic History* 94 (2024), p. 101627.
- [3] Bazerman, Charles. “Genre as Social Action”. In: *The Routledge handbook of discourse analysis*. Routledge, 2013, pp. 226–238.
- [4] Bergeaud, Antonin and Verluise, Cyril. “A New Dataset to Study a Century of Innovation in Europe and in the US”. In: *Research Policy* 53, no. 1 (2024), p. 104903.
- [5] Berkes, Enrico, Chen, Matthew Lee, and Tranchero, Matteo. “300 Years of British Patents”. In: *Research Policy* 55, no. 1 (2026), p. 105347.
- [6] Billington, Stephen D and Hanna, Alan J. “That’s Classified! Inventing a New Patent Taxonomy”. In: *Industrial and Corporate Change* 30, no. 3 (2021), pp. 678–705.
- [7] Burk, Dan L. and Reyman, Jessica. “Patents as Genre: A Prospectus”. In: *Law & Literature* 26, no. 2 (2014), pp. 163–190.
- [8] Cantwell, John. “Technological Lock-in of Large Firms since the Interwar Period”. In: *European Review of Economic History* 4, no. 2 (2000), pp. 147–174.
- [9] Chandler, Daniel. “An Introduction to Genre Theory”. [WWW document], accessed 6 March 2026. 1997.
- [10] Crossley, Scott A. and Louwerse, Max M. “Multi-dimensional Register Classification Using Bigrams”. In: *International Journal of Corpus Linguistics* 12, no. 4 (2007), pp. 453–478.
- [11] Degner, Harald and Streb, Jochen. “Foreign Patenting in Germany, 1877–1932”. In: *Organizing Global Technology Flows*. Routledge, 2013, pp. 17–38.
- [12] Federico, Pasquale J. “Historical Patent Statistics of 44 Nations, 1791 to 1961.” (ZA8437; Version 1.0.0) [Data set]. GESIS, Cologne. 2011. DOI: 10.4232/1.10297.
- [13] Gooday, Graeme and Wilf, Steven. *Patent Cultures: Diversity and Harmonization in Historical Perspective*. Vol. 52. Cambridge University Press, 2020.
- [14] Griliches, Zvi. “Patent Statistics as Economic Indicators: A Survey”. In: *Journal of Economic Literature* 28, no. 4 (1990), pp. 1661–1707.
- [15] Groom, Nicholas and Grieve, Jack. “The Evolution of a Legal Genre: Rhetorical Moves in British Patent Specifications, 1711 to 1860”. In: *Studies in Corpus Linguistics*, ed. by Teresa Fanego and Paul Rodriguez-Puente. Amsterdam: John Benjamins Publishing Company, 2019, pp. 201–234.
- [16] Guldi, Jo. *The Dangerous Art of Text Mining: A Methodology for Digital History*. Cambridge University Press, 2023.
- [17] Israel, Paul. “‘Claim the Earth’: Protecting Edison’s Inventions at Home and Abroad”. In: *Knowledge Management and Intellectual Property*. Edward Elgar Publishing, 2013, pp. 19–43.

- [18] Kelly, Bryan, Papanikolaou, Dimitris, Seru, Amit, and Taddy, Matt. “Measuring Technological Innovation over the Long Run”. In: *American Economic Review: Insights* 3, no. 3 (2021), pp. 303–320.
- [19] Khan, B Zorina. “Selling Ideas: An International Perspective on Patenting and Markets for Technological Innovations, 1790–1930”. In: *Business History Review* 87, no. 1 (2013), pp. 39–68.
- [20] La Mela, Matti, Frankemölle, Jonas, and Tell, Fredrik. “Novelty and Impact: Using Document Similarity to Study Important Inventions in Historical Swedish Patents, 1890–1929”. In: *Digital Humanities in the Baltic and Nordic Countries 8th Conference, May 27-31, 2024, Reykjavík*. Vol. 7. University of Oslo Library. 2025, pp. 1–15.
- [21] Miller, Carolyn R. “Genre as Social Action”. In: *Quarterly Journal of Speech* 70, no. 2 (1984), pp. 151–167.
- [22] Moser, Petra. “Patents and Innovation: Evidence from Economic History”. In: *Journal of Economic Perspectives* 27, no. 1 (2013), pp. 23–44.
- [23] Neale, Steve. “Questions of Genre”. In: *Screen* 31, no. 1 (1990), pp. 45–66.
- [24] Pretel, David and Sáiz, Patricio. “Patent Agents in the European Periphery: Spain (1826–1902)”. In: *History of Technology* 31 (2012), pp. 97–114.
- [25] Rutherford, Brian A. “A Genre-theoretic Approach to Financial Reporting Research”. In: *The British Accounting Review* 45, no. 4 (2013), pp. 297–310.
- [26] Sáiz, Patricio and Amengual, Ruben. “Do Patents Enable Disclosure? Strategic Innovation Management of the Four-stroke Engine”. In: *Industrial and Corporate Change* 27, no. 6 (2018), pp. 975–997.
- [27] Xie, Yunting. “Handle All Matters with Speed and Courtesy” : *Japanese Patent Agents in International Patenting Activities (1922-1940)*. MA thesis. Uppsala University, Department of ALM, 2024.
- [28] Xie, Yunting, La Mela, Matti, and Tell, Fredrik. “Multimodal LLM-assisted Information Extraction from Historical Documents: The Case of Swedish Patent Cards (1945-1975) and ChatGPT”. In: *The 9th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2025), March 5–7, 2025, Tartu, Estonia*. University of Oslo Library. 2025, pp. 1–15.

A Appendix: French patent data from 1902 onwards and its quality

France had established its modern patent institution during the French Revolution, but a new centralized patent office, the Office National de la Propriété Industrielle (ONPI), was created in 1901 in response to the increasing number of patent applications. Following this change, patents formed a new administrative collection starting in 1902, which has since been digitized and published separately.³

The data collection starts with the French patent FR317502A, granted on 8 September 1902; the patent numbering follows the order of grant. In addition to the main patent documents (A-series), we also collected the E-series of patent documents, which are additions made to previously granted patents. The European Patent Office (EPO) dataset forms the basis for aggregated patent resources such as PatentCity patent metadata. However, these resources do not contain the patent texts, and it appears that PatentCity does not include the E-series⁴. The E-series also begins in

³ Patents preceding 1902 are available in the historical French patents database of the INPI, but have not been OCR or HTR processed. <https://archives.inpi.fr/>

⁴ In its documentation, PatentCity mentions only the A-series, see: https://www.longtermproductivity.com/perso/Online_Appendix_Patentcity.pdf

1902. As discussed below, the EPO data is not complete for all years, and a comprehensive research database of French patents would require further reconstruction. For the purposes of this paper—a linguistic analysis of patent texts that focuses on longer-term patterns rather than individual years—the coverage of the INPI/EPO dataset is sufficient.

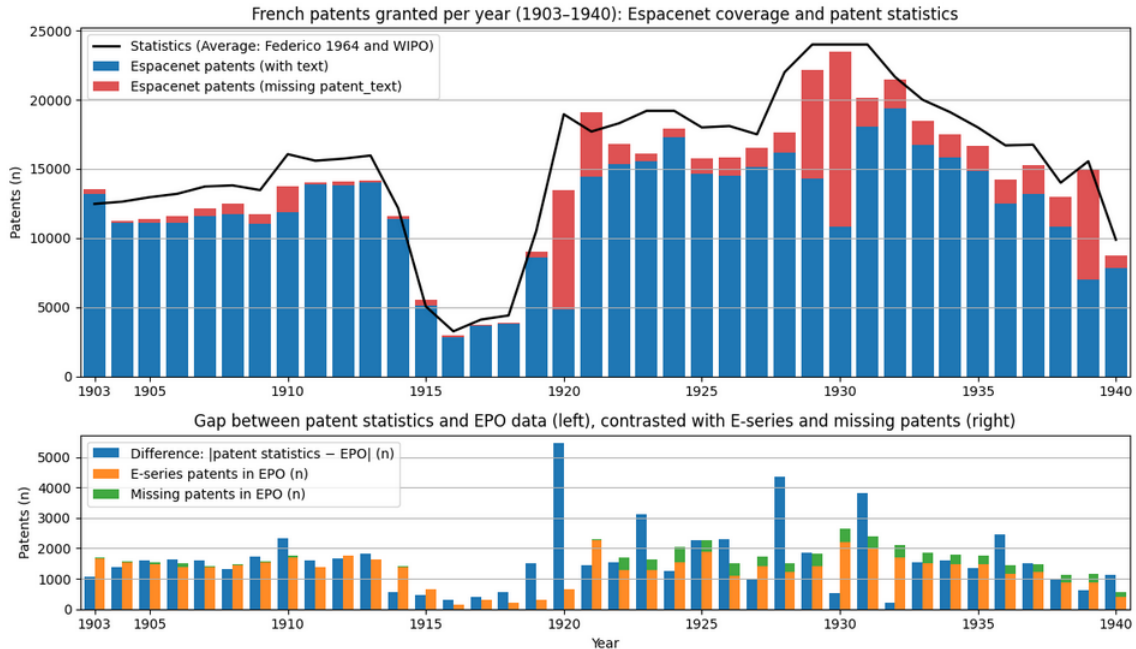


Figure 3: Description of French patent corpus (EPO) in relation to established patent statistics, 1903–1940.

Figure 3 displays the coverage of the EPO dataset from 1903 until 1940. The figure shows the EPO data as bars and the yearly patent statistics from two established sources as a line (their average, even though in most cases they match): the World Intellectual Property Office historical statistics⁵ and International Patent Statistics by Federico [12]. We observe that the coverage follows the general trend of yearly granted patents, although there are gaps for some years such as 1920, 1928, and 1931.

In the lower panel of Figure 3, we visualize the gap between the EPO data and the patent statistics, and next to it what could explain the gap. As mentioned above, granted patents include patents of addition (the E-series), which are plotted with orange bars. Following the numbering of the patents, we also identify missing patents in the EPO dataset, marked in green. Although further investigation is required, it appears that the international statistics also include the patents of addition, which helps explain the gap between the granted new patents and the yearly statistics. When the granted patents (A-series), patents of addition (E-series), and missing patents are summed for 1903–1940, the total reaches 587,598 patents, compared with 586,814 and 588,488 patents reported by WIPO and [12] respectively. This corresponds to coverage rates of 100.13% and 99.85%. These differences suggest that, in addition to missing patents, some publication-date metadata in the EPO bibliographic data may be inaccurate.

For our analysis, we exclude patents of addition, as we focus on the textual style of the main patent applications. A more pressing question concerns the quality of the patent texts produced through OCR from the original patent documents. As shown in Figure 3, in some years, such as

⁵ The dataset is available at https://www.wipo.int/documents/2948119/3215563/wipo_ip_historical_data.zip

1920, 1921, 1929, and 1930, a large share of the patents do not include the OCR text.⁶

To address this issue, the analysis examines language change in periods rather than by year. We also observe that the patent descriptions contain annotations created during OCR processing: 46% of the documents contain XML tags. These XML tags were removed from the texts for the analysis. Interestingly, most of the tags relate to formatting, although the tag `<img>` also appears. We find These tags can result from OCR attempts to capture formulas or graphical elements that could not be recognized as text, though most relate to OCR errors in reading individual characters or lines. notably, patents before 1920 rarely contain `<img>` tags (less than 0.01%), whereas after 1920 approximately 30% of patent documents include this tag. This pattern may indicate changes in OCR quality or scanning procedures for the later period.

Finally, figure 4 shows document lengths over time for patents that contain a patent text in the EPO dataset, after XML tags have been removed. The figure reports average and median lengths and the quartile range. Overall, document length increases over time. We also detect an upward peak in document length during the war years in 1915–1918, visible in both the average and the median. This coincides with a drop in the number of granted patents (see Figure 3), suggesting a change in the types of patents filed in these years. The issue requires further investigation, but could signal that “popular” patenting decreased during the war. Some patents are surprisingly short, with only around 150 tokens. To reduce OCR errors and remove incomplete documents, we exclude patent descriptions containing 50 tokens or less from the corpus.

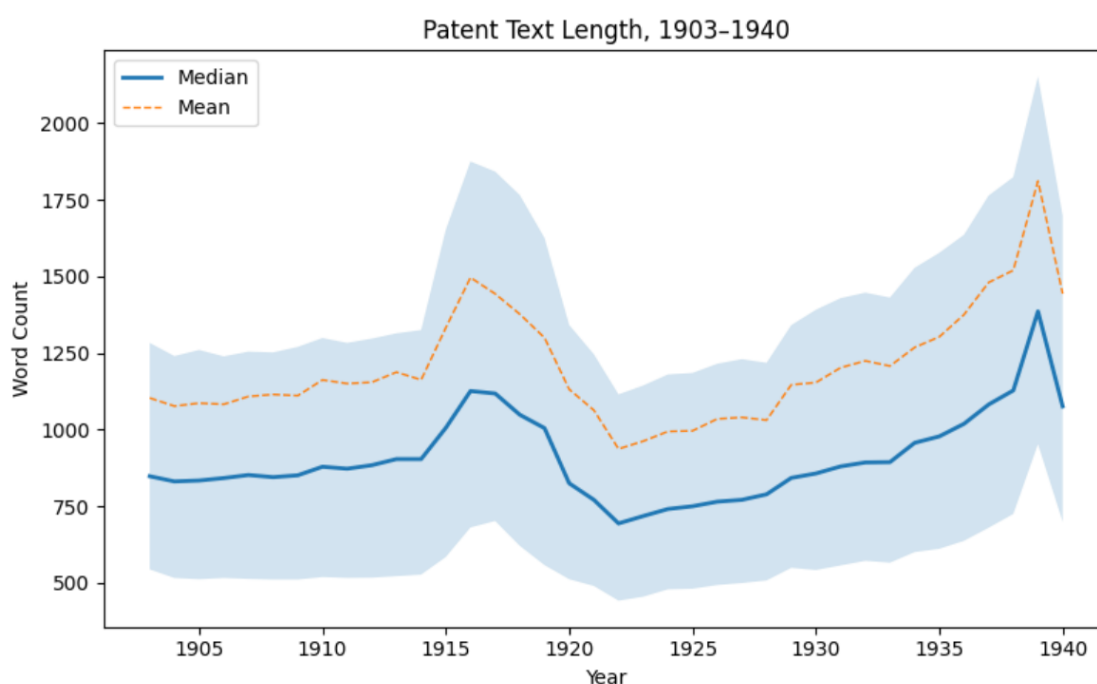


Figure 4: Patent document length in 1903–1940. Note. The shaded area covers the 25% and 75% percentiles.

⁶ We can only speculate about the reason, but it may be related to the quality of the original documents for these years, which could also explain the issues in publication-year metadata visible in the gaps.

B Appendix: Top 100 most frequent bigrams in French patents (1903–1940) and filtering vocabularies

The bigram analysis applied three filtering vocabularies, drawn from standard French function-word lists and refined through inspection of the most frequent bigrams in the corpus. All tokens are lowercased and accent-stripped to match the corpus preprocessing. Metadata tokens (administrative vocabulary from patent headers), filtered if one term of the bigram matched: *publie, delivre, prix, fascicule, office, national, franc, brevet, propriete, industrielle, residant, france, republique, francaise, vente, imprimerie, adresser*. Function words, drawn from standard French function-word lists, and filtered if both terms of the bigram matched: *a, ainsi, au, aux, avec, c, ce, celle, celui, ces, cet, cette, ci, d, dans, de, des, dessous, dessus, dont, elle, elles, en, entre, est, et, ete, etre, il, ils, jusqu, l, la, laquelle, le, lequel, les, lesquelles, lesquels, leur, leurs, lorsqu, lorsque, n, ne, on, ou, outre, par, pas, peut, peuvent, pour, qu, que, qui, rapport, s, sa, se, ses, son, sont, sorte, sur, tandis, tel, telle, telles, tels, travers, un, une, y*. Articles, filtered if one term of the bigram matched: *de, des, du, d, l, la, le, les, un, une*.

Table 6: Top 100 most frequent bigrams in French patents (1903–1940). “Periods” describes the number of seven study periods in which the bigram appears among the period’s 100 most frequent.

Rank	Bigram	Patents (n)	Share (%)	Periods
1	presente invention	243641	64.94	7
2	invention a	190389	50.75	7
3	par exemple	187956	50.10	7
4	a titre	181334	48.33	7
5	au moyen	175432	46.76	7
6	pour objet	161689	43.10	7
7	fig est	142206	37.90	7
8	dessin annexe	131859	35.15	7
9	se trouve	129936	34.63	7
10	invention est	125935	33.57	7
11	facon a	112151	29.89	7
12	fig et	111436	29.70	7
13	par suite	111272	29.66	7
14	vue en	105645	28.16	7
15	maniere a	99189	26.44	7
16	en coupe	94726	25.25	7
17	a dire	92937	24.77	7
18	ce dernier	92849	24.75	7
19	fig i	90639	24.16	5
20	autre part	89833	23.94	7
21	par procuration	87460	23.31	3
22	memes temps	85664	22.83	7
23	en meme	85504	22.79	7
24	plus grande	83008	22.12	7
25	en plan	80413	21.43	7
26	i et	79957	21.31	6
27	i o	74699	19.91	5
28	en forme	73466	19.58	7

Continued on next page

Table 6: Top 100 most frequent bigrams in French patents (1903–1940). “Periods” describes the number of seven study periods in which the bigram appears among the period’s 100 most frequent.

Rank	Bigram	Patents (n)	Share (%)	Periods
29	cette invention	73443	19.58	5
30	au lieu	72864	19.42	7
31	partie superieure	72496	19.32	7
32	cet effet	70719	18.85	7
33	i est	69670	18.57	5
34	en vue	69412	18.50	7
35	constitue par	69154	18.43	6
36	en elevation	68876	18.36	7
37	ou autre	68862	18.35	7
38	destine a	68724	18.32	7
39	au dessin	66351	17.69	5
40	facon que	66315	17.68	7
41	mise en	66021	17.60	7
42	fig a	64649	17.23	7
43	en contact	64114	17.09	7
44	ce cas	63772	17.00	6
45	grace a	63219	16.85	5
46	bien entendu	62384	16.63	4
47	fait que	62131	16.56	5
48	a i	62053	16.54	3
49	par consequent	61576	16.41	7
50	comme on	61499	16.39	7
51	partie inferieure	60299	16.07	7
52	represente a	59879	15.96	4
53	ce dispositif	59638	15.90	6
54	pour but	59632	15.89	3
55	en place	59089	15.75	7
56	qui permet	57964	15.45	4
57	on obtient	57943	15.44	4
58	ou plusieurs	57924	15.44	7
59	est fixe	57663	15.37	7
60	contact avec	57498	15.33	7
61	cette derniere	56706	15.11	6
62	maniere que	56644	15.10	7
63	doit etre	56592	15.08	5
64	fig montre	56501	15.06	7
65	pour permettre	56063	14.94	6
66	son extremite	55738	14.86	7
67	tout en	55482	14.79	4
68	ces deux	54471	14.52	4
69	est represente	54395	14.50	6
70	se trouvent	54334	14.48	7
71	comme il	54132	14.43	7

Continued on next page

Table 6: Top 100 most frequent bigrams in French patents (1903–1940). “Periods” describes the number of seven study periods in which the bigram appears among the period’s 100 most frequent.

Rank	Bigram	Patents (n)	Share (%)	Periods
72	est monte	53796	14.34	6
73	servant a	53447	14.25	7
74	on fait	52936	14.11	7
75	sa position	52862	14.09	7
76	dessin ci	52569	14.01	0
77	coupe verticale	51035	13.60	4
78	consiste en	51028	13.60	2
79	dessins annexes	50930	13.57	0
80	invention consiste	50847	13.55	1
81	pouvant etre	50700	13.51	2
82	ou moins	50598	13.49	4
83	plus ou	50565	13.48	5
84	b et	50323	13.41	4
85	lui meme	50029	13.33	1
86	ce genre	49913	13.30	4
87	au point	49727	13.25	5
88	est fixee	49591	13.22	4
89	au moment	49510	13.20	6
90	ou bien	48800	13.01	5
91	se produit	48674	12.97	4
92	sa partie	48666	12.97	4
93	monte sur	48613	12.96	5
94	fig represente	48265	12.86	6
95	cette disposition	48149	12.83	1
96	est relative	48016	12.80	0
97	destinee a	47699	12.71	0
98	se compose	47339	12.62	1
99	qui fait	47330	12.62	0
100	si on	47174	12.57	5