



Language, Power and Imperial Administration: Tracing Linguistic Change through Colonial Records and Public Institutions

Antonella Guida

Department for Humanistic, Scientific and Social Innovation (DIUSS), University of Basilicata, 75100 Matera, Italy

Received: 09/06/2026 Revised: 10/07/2026 Acceptance: 17/07/2026 Published: 25/07/2026

Abstract

Language played a fundamental role in the operation of colonial governance by shaping administrative communication, institutional authority, and documentary practices. This study examined linguistic patterns in colonial-era texts to investigate the relationship between language, imperial administration, and public institutions. A historical corpus-based research design was employed using metadata from the Project Gutenberg collection. Following systematic screening, metadata cleaning, publication-year verification, and full-text retrieval, a corpus of 25 colonial texts comprising approximately 2.17 million words was constructed. The texts were analysed using computational linguistic techniques, including lexical preprocessing, dictionary-based vocabulary classification, TF-IDF analysis, descriptive statistics, and non-parametric tests. The findings revealed consistent patterns of administrative, legal, educational, and institutional vocabulary across the analysed corpus. Institutional categories exhibited distinct linguistic profiles, reflecting their respective administrative functions, while statistical analysis indicated no significant temporal variation in the Administrative Language Index across historical periods. The results suggest that colonial administrations relied on standardized linguistic conventions that supported bureaucratic continuity and institutional governance. By integrating archival screening with corpus linguistics, the study demonstrates the value of computational approaches for examining historical documentary collections. The proposed methodological framework contributes to historical linguistics and colonial studies by providing a transparent approach for analysing administrative language and offers a foundation for future investigations using larger and more diverse archival corpora.

Keywords: colonial administration, corpus linguistics, historical archives, language policy, linguistic change

Introduction

Language has always and everywhere been one of the most potent means of establishing, exercising and consolidating political power. In addition to the communication role, language has been used by governments to create laws, standardisation of administrative practices, control of institutions and development of systems of governance that impact the social, cultural and political relationships. Linguistic practices are not only used in day-to-day communication but are also the tools used to organise imperial government, to record the presence of the empire on its territory, and to validate state power in colonial situations. Language imperialism is an example of the use of language by the dominant political power to In studies related to English as an imperial language, language was used as a symbol of political power, administrative legitimacy as well as communication media (Namiq & Zafarghandi, 2025).

Written records were a key part of the administration of colonial governments over widely divergent areas and much was recorded, including laws, regulations, census, correspondence, education policy, and administration reports. These records evolved into invaluable tools for documentation, legal enforcement, and governance on their sprawling colonies, spread across a wide geography (Siddique, 2024). Colonial archives were not neutral documents which stored historical information, but were rather a reflection of politics, administration and power structures within imperial governance (Burak et al., 2022). Language used in these documents is therefore telling in terms of the imperial institutions' use of power, and the development of imperial bureaucratic practices and public administration.

There is an increasing interest in the relationship between language and governance in the study of language policy and multilingual administration. The research shows that, often, governmental institutions set up formal language policies to make administration easier, communication smoother and political integration with linguistically diverse groups more effective (Gao & Moratto, 2024). History indicates that many of today's language policies are derived from institutional practices from past periods of imperial expansion and colonial rule (Laitin & Ramachandran, 2025). Such developments suggest that the language decisions are not just matters of culture and education, but are related to political goals and administrative structure.

Administrative language also was a key part of the decision-making process as it related to legal authority and institutional legitimacy. Standardised forms of communication through official correspondence, legal documents and codified rules bolstered bureaucratic control and judicial governance. The formalisation of administrative discourse allowed the colonial governments to control the public institutions and also built political accountability and legal administrative systems (Jayaraj & Navas, 2024). Bureaucratic institutions were some of the key agents that negotiated citizenship, governance and relations between colonial powers and local communities, hence, (Berda, 2022). Institutional processes similar to these have been noticed in other imperial borderlands, which created long-term political and social legacies (Popescu, 2023).

While a considerable amount of research has been done into linguistic imperialism, language policy and colonial administration, much less has been done so far to analyze historical colonial documents systematically with corpus-based methods, with the aim of observing how linguistic features changed over the history of colonial documents. These studies have tended to focus on theoretical analysis, historical interpretation or policy analysis, with fewer quantitative analyses of administrative vocabulary within a wide range of colonial documents (Makoni et al., 2023; Pratama, 2025). Meanwhile, studies on language variation have shown that historical texts can capture 'measurable' processes of lexical change which can be used to show over time institutional and sociopolitical changes in general (Gonzales, 2026). Bilingualism and postcolonial language development studies also indicate that past use of administrative language also shapes current linguistic structures and identities (Warditz, 2026). These observations emphasize the need to draw in the historical analysis of the corpus together with the institutional and archival approaches, so as to be able to better comprehend the interplay of language, governance, and imperial administration.

In this context, the current study aims at (re)discovering language patterns in colonial era texts through a methodical and comparative analysis based on a historical screening of archives and a computational analysis of a text corpus. The study aims to investigate the use of administrative language in the past

and in various institutional types to gain a sense of the role language played within the administration of colonial governance and public institutions, and to offer empirical evidence of the linguistic features that are maintained in documentary historical sources.

Research Objectives

- To construct a historical corpus of colonial-era texts representing public institutions and administrative governance.
- To examine variations in administrative and institutional language across different colonial historical periods.
- To evaluate linguistic patterns associated with colonial governance using corpus-based computational analysis.

2. Methodology

2.1 Research Design

To examine linguistic patterns relating to colonial administration and public institutions, a historical corpus-based research design was used. The mixed analytical approach involved archival text selection, computational linguistic analysis, descriptive statistics and non-parametric inferential testing. The aim was to explore differences in administrative and institutional terms both within and between colonial texts and periods. The source dataset comprised bibliographic data without a specified dataset of colonial records; the analytical corpus was formed as systematic, metadata-based, publication year verified, full-text retrieved, preprocessed, and linguistically analysed. The data source and the development of the corpus. Data source and Corpus development.

2.2 Data Source and Corpus Development

The analysis was based on the Gutenberg metadata dataset, and initially 15,331 entries were available that included metadata of the book titles, authors, links, and bookshelf categories (Bejan, 2021). Duplicate and incomplete records were removed leaving 9,073 unique records. Records were filtered with the aid of these thematic keywords: imperial governance, public administration, law and judiciary, education and language, revenue and economy, census and ethnography, as well as the colonial geographical identifiers. This process identified 45 potentially relevant texts. The publication years were found using the Open Library web service and 34 of the records were successfully dated. Documents were included only if they had been published between 1700 and 1947, leaving a final usable collection of 25 documents. These texts were divided into three groups of early, middle and late colonial period, and downloaded from Project Gutenberg for analysis. The final corpus consisted of some 2.17 million words.

2.3 Text Preparation and Preprocessing

Downloaded texts were cleaned by removing Project Gutenberg headers, footers, and licensing information. Inconsistent spacing was standardised and line breaks and punctuation were added before tokenization. The texts retained were of at least 5,000 words. The corpus was then lowercased, stop words and non-alphabetic tokens were filtered out and the remaining words were lemmatized using the WordNet lemmatize. Processed tokens were then used to create analytical strings of text for vocabulary analysis, and raw and processed word counts were kept for descriptive analysis.

2.4 Linguistic Indicators

Five linguistic categories were created based on information from the dictionary for assessing imperial authority, administrative control, law and regulation, education and language, and population classification. The number of occurrences of a term was normalized by the number of words processed and multiplied by 1000 to make the set of numbers uniform irrespective of the document lengths. Lexical diversity was determined by the type-token ratio and an Administrative Language Index (ALI) was derived as the average of the five standardised frequencies of the categories. Institutional categories were determined based on the overall topic identified through the metadata screening process.

2.5 Data Analysis

The analyses are done in python using Google colab. Descriptive statistics analysed the composition of the corpus, the distribution of institutions, historical periods, lexical diversity and administrative language indicators. Annual averages, rolling means and heatmaps were used to analyse temporal trends and institutional linguistic profiles, respectively. Distinctive terms were found for the different historical periods by TF-IDF analysis. Kruskal–Wallis test was used for inferential comparisons between the Administrative Language Index for each colonial period and Spearman's rank correlation test was used to determine temporal trends. A 5% level of significance was used to test for statistical significance. No statistically significant differences were found between historical periods and/or temporal trends in the Administrative Language Index.

3. Results

3.1 Corpus Identification and Historical Coverage

The corpus construction process was a systematic reduction of the initial Project Gutenberg metadata collection, to retrieve texts on colonial administration, language, and public institutions. The initial data set comprised of 15,331 bibliographic records (Table 1). 9,073 unique records were kept for thematic screening, after discarding duplicates and incomplete records. A filter based on keywords (institutional, linguistic and colonial) retrieved 45 potentially relevant records. Through using Open Library metadata, publication years could be retrieved for 34 texts, and in order to keep the historical authenticity of the colonial period, only documents published between 1700 and 1947 have been selected. The 25 eligible texts were all successfully downloaded from Project Gutenberg; no texts failed the minimum text-length requirement of 5000 words. This multi-stage selection process helped to guarantee that the corpus was of historically relevant content, and that the selection was transparent in the sense of inclusion/exclusion of documents. The progression from the original metadata collection to the final analytical corpus is summarised in Table 1. Figure 1 presents the distribution of the final analytical corpus across historical periods and institutional categories, highlighting the predominance of late colonial texts and the institutional composition of the analysed documents.

Table 1. Corpus Screening and Selection Process

Screening Stage	Number of Records	Retention (%)
Original metadata records	15,331	100.00
Records retained after metadata cleaning	9,073	59.18
Colonial and institutional candidates	45	0.29
Records with identified publication year	34	0.22
Records dated between 1700 and 1947	25	0.16
Texts successfully downloaded	25	0.16
Texts retained after length screening	25	0.16

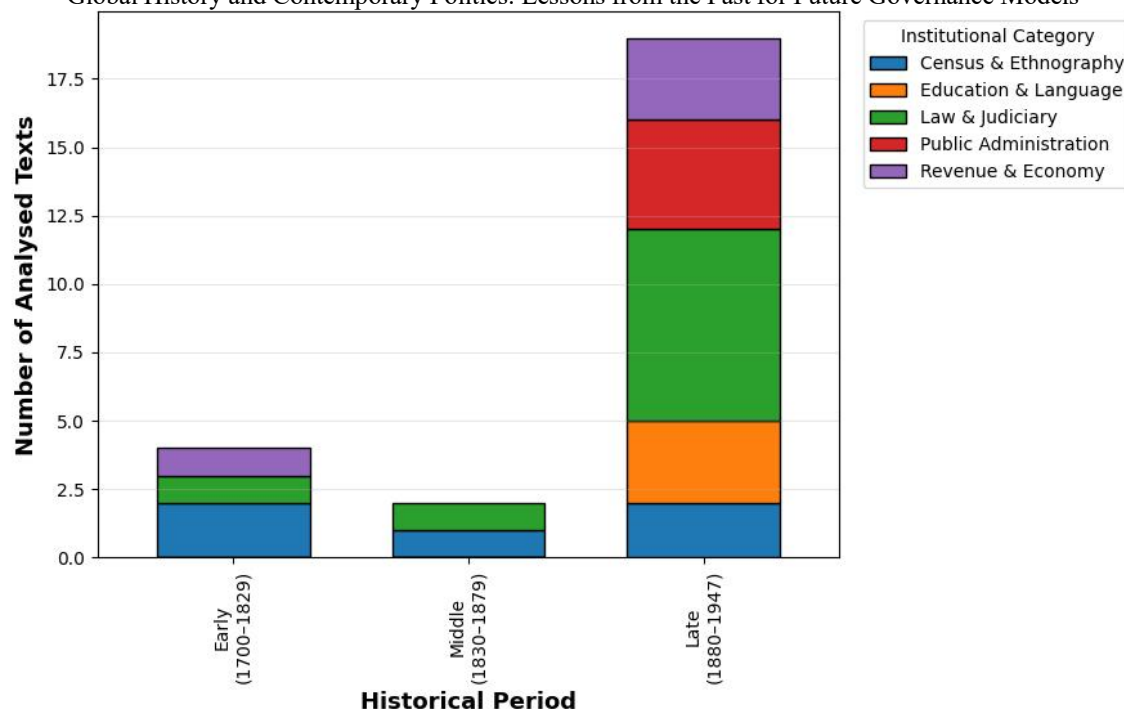


Figure 1. Distribution of the Final Analytical Corpus across Historical Periods and Institutional Categories

3.2 Distribution of Colonial Records across Historical Periods and Institutions

The final analytical corpus consisted of 25 colonial texts spread over three historical periods and five institutional types. Table 2 shows that most of the documents ($n = 19$) were from the late colonial period (1880–1947), followed by the early colonial period (1700–1829; $n = 4$) and the middle colonial period (1830–1879; $n = 2$). In terms of institutional classification, the law and judiciary ($n = 9$), census and ethnography ($n = 5$), public administration ($n = 4$), revenue and economy ($n = 4$), and education and language ($n = 3$) are the largest institutional categories. The period of the late colonial era exhibited the highest number of institutional categories, with the presence of all five categories, while the other periods had a smaller number of institutional categories. This distribution served as a foundation for further comparisons of linguistic features between the historical periods and between the institutions.

Table 2. Distribution of Analysed Texts by Historical Period and Institutional Category

Historical Period	Census & Ethnography	Education & Language	Law & Judiciary	Public Administration	Revenue & Economy	Total
Early Colonial Period	2	0	1	0	1	4
Middle Colonial Period	1	0	1	0	0	2
Late Colonial Period	2	3	7	4	3	19
Total	5	3	9	4	4	25

3.3 Temporal Patterns of Administrative Language

Document length, lexical diversity and Administrative Language Index were used to compare the linguistic features of the corpus in the three historical periods. Table 3 reveals that the early colonial period had the highest median Administrative Language Index, followed by the middle and late colonial periods, suggesting that there was not a significant increase or decrease in the use of administrative vocabulary across the three periods. There were relatively small differences in the mean values for all periods, there was no significant difference in lexical diversity. Similarly, the median lengths of documents also varied but showed no obvious temporal trend, indicating that the variation in administrative language was not driven by document size. Figure 2 shows the annual distribution of the Administrative Language Index, as well as a centred rolling mean. There were moderate changes in the number of individual publications, but a sustained increase or decline was not observed throughout the study period. These are descriptive findings, showing that the density of administrative/institutional language was generally similar across the analysed corpus throughout the colonial period.

Table 3. Linguistic Characteristics of the Corpus across Historical Periods

Historical Period	Documents	Mean Word Count	Mean Lexical Diversity	Mean Administrative Index	Median Administrative Index
Early Colonial Period (1700–1829)	4	121,169.750	0.159	5.573	4.361
Middle Colonial Period (1830–1879)	2	85,106.500	0.214	3.814	3.814
Late Colonial Period (1880–1947)	19	79,511.684	0.214	5.626	3.798

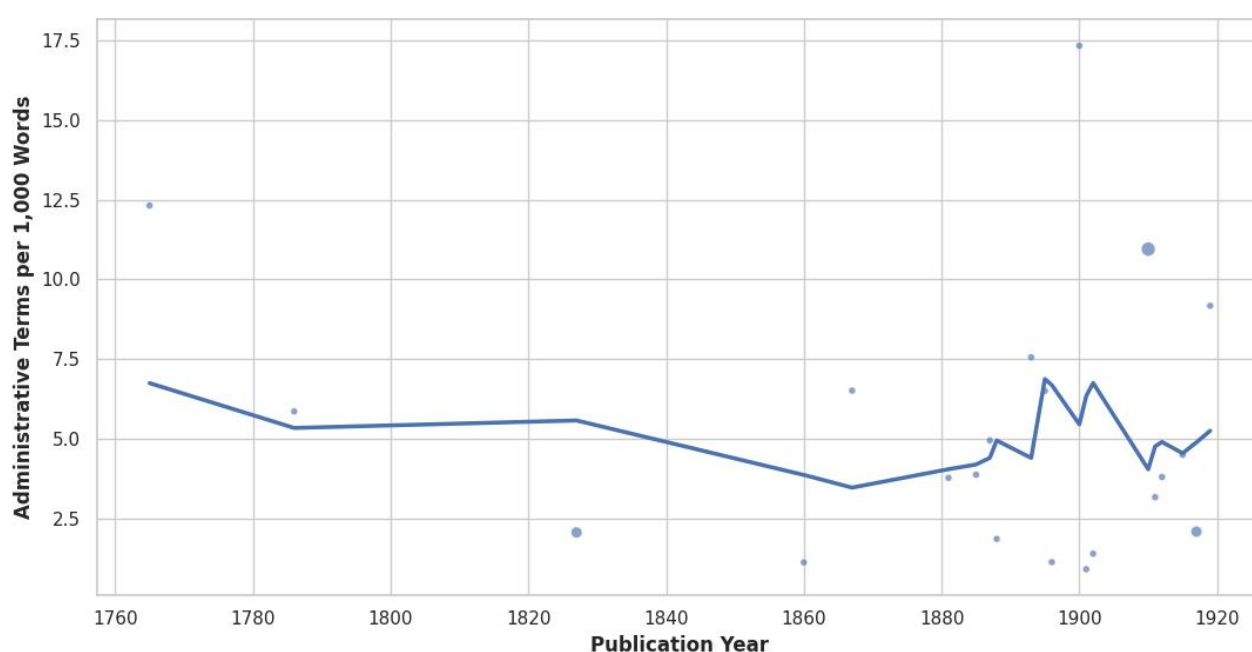


Figure 2. Temporal Change in the Administrative Language Index across Colonial Publications

3.4 Institutional Linguistic Profiles

The five predetermined lexical categories: imperial authority, administrative control, law and regulation, education and language, and population classification were used to analyze institutional differences in linguistic patterns. The mean frequency of each category was determined for each group of institutions and is displayed in Figure 3. There were clear linguistic differences between institutions in terms of the focus on administrative and governance-related terminology. The law and judiciary category contained the most legal and regulatory terms and public administration had relatively more terms related to administrative control. Census and ethnography had a higher frequency for the population classification terminology, while education and language records had a higher frequency for the educational and language expressions. The category of revenue and economy revealed a good balance between administrative and economic terminology. In general, the institutional heatmap shows that there are significant differences in lexical emphasis between institutional categories, meaning that the texts analysed contained specific patterns of administrative language in relation to the thematic areas.

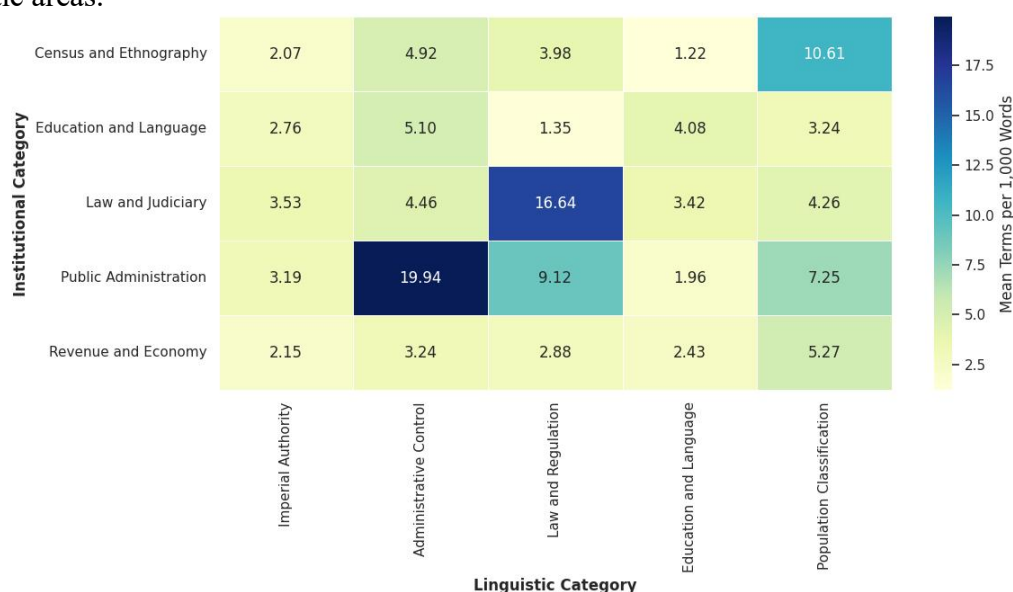


Figure 3. Institutional Distribution of Administrative and Institutional Vocabulary across the Analysed Corpus

3.5 Statistical and Lexical Evidence of Linguistic Change

Table 4 gives the statistical analysis of the linguistic changes from one time-period to another. Results of Kruskal–Wallis test indicated that the difference in the Administrative Language Index was not statistically significant between the early colonial period, the middle colonial period, and the late colonial period ($H = 0.2509$, $p = 0.8821$). Likewise, Spearman's correlation showed no correlation between publication year and Administrative Language Index ($\rho = 0.0293$, $p = 0.8896$). These findings indicate that the density of administrative vocabulary in the analysed corpus did not change significantly nor across the different publications' year nor across the different historical periods. In addition to the statistical analysis, the top TF-IDF terms of each colonial period are shown in Figure 4. After text preprocessing and term weighting, the visualization brings attention to the most unique weighted vocabulary in the early, middle, and late colonial corpora. The extracted terms offer a comparative view of the lexical characteristics of the historical periods, as well as showing differences in the salience of words of the period being represented in the chosen texts.

Table 4. Statistical Tests of Temporal and Period-Based Linguistic Change

Statistical Analysis	Test Statistic	p-value	Interpretation
Kruskal–Wallis Test (Administrative Language Index across historical periods)	0.2509	0.8821	No significant difference
Spearman's Rank Correlation (Publication Year vs. Administrative Language Index)	0.0293	0.8896	No significant association

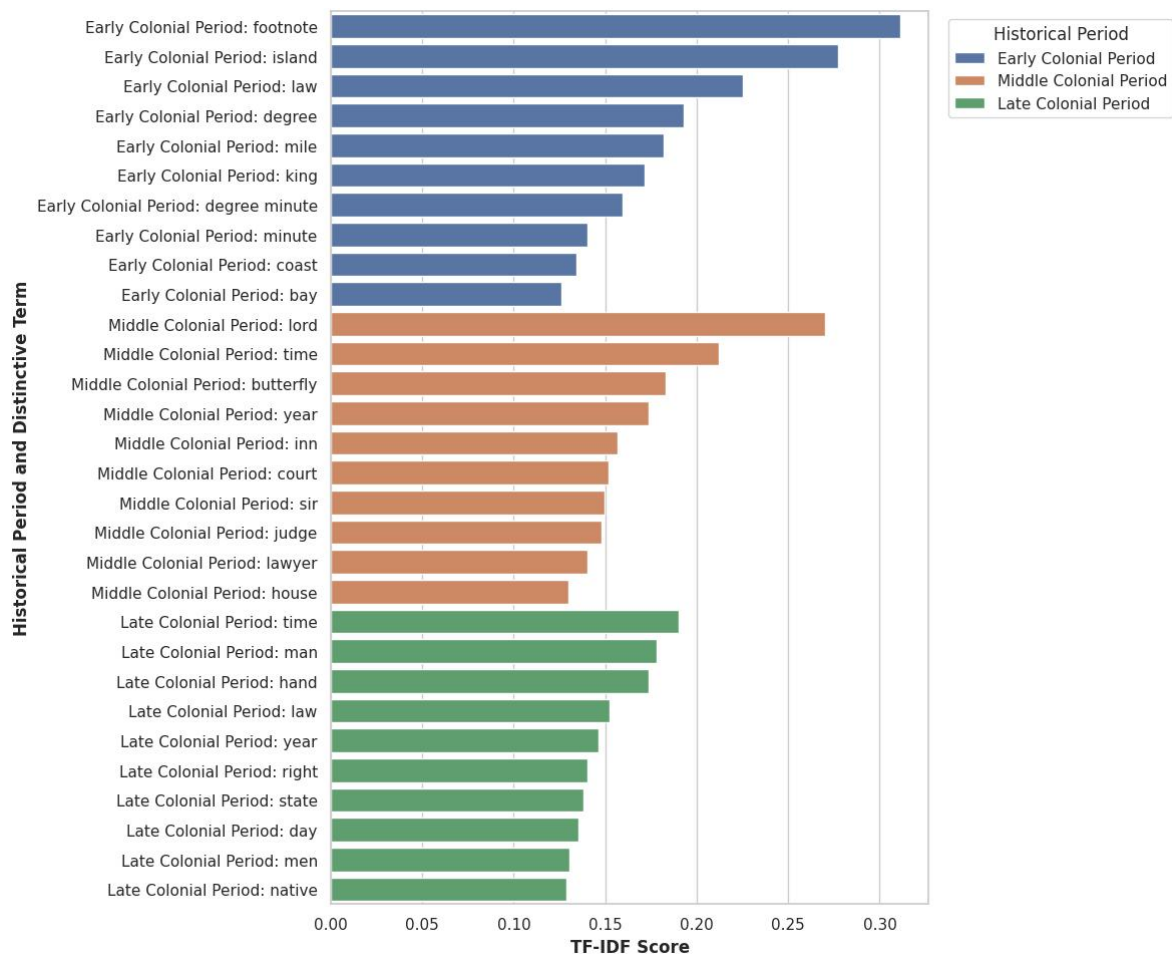


Figure 4. Distinctive Administrative and Institutional Terms across Historical Periods Based on TF-IDF Analysis

4. Discussion

The present study looked at language in colonial era texts to address the question of language, imperial rule, and public institutions. The corpus construction process showed that a minority of the historical texts available actually focused on the governance and institutional administration of the colonies, thus emphasizing the selective nature of archival evidence left from colonial time. This illustrates the idea that colonial archives were not necessarily "historical" archives; they are rather a filter of documentation, shaped by administrative forms and structures of power (Hilden, 2022). The use of institutional records also reinforces the idea that imperial governments created documentation mainly for the purposes of good governance, control of territories and the continuity of the bureaucracy, and not necessarily to offer a balanced historical representation (Jordan, 2024).

Based on descriptive analysis, the findings indicated that there were not much different vocabulary in administration from the early colonial period to middle colonial period and late colonial period. While some changes were noted, statistical analysis revealed no significant differences in Administrative Language Index over time. This discovery indicates that language usage during colonial rule was relatively stable during the time period examined. This continuity also demonstrates the need for institutional uniformity of administrative communication, allowing governments to assert and institutionalize legal control and organizational continuity across varying political and geographical environments. Comparable dynamics can be observed across various forms of language in colonial and post-colonial governance and administrative practices, which served to assert institutional dominance over heterogeneous populations (Young 2023). Additionally, the continuity of formal linguistic structures is linked to the fact that the language policy sometimes changes slowly, as administrative structure needs continuity and standardized communication instead of quick language change (Zeng & Li, 2023).

Another significant finding of the study was the institutional variation. The heatmap analysis revealed

that law and judiciary, public administration, education and ethnographic records had different linguistic profiles, meaning that administrative words suggested a varying colonial discourse and instead, they suggested that the use of the words in administrative practices related to the function of the institutions. In legal documents, legal jargon was naturally used, and in educational documents, language related to teaching and administration was used. The differences are examples of the way in which colonial institutions have developed a specialized use of language to serve their various administrative needs. Analogous institutional differentiation can be seen in the colonial educational systems in which schools and professional institutions emerge as key channels to transfer official language and administrative knowledge (Cho & Robert, 2024). Similarly, the educational institutions of the past were also responsible for forming Intellectual and Administrative tradition, and they helped to standardize the language in the governance of the colonies (Suman, 2023).

The statistical tests failed to find any temporal changes but the lexical analyses revealed identifiable patterns of administrative and institutional vocabulary in historical documents. The result suggests that computational corpus methods could still be applied to study collections of historical documents, as they allow for systematic evidence of language features that are not necessarily apparent from a historical interpretation alone. The non-significant statistical results do not seem to signify that there was no language change but rather the small and uneven historical corpus that was analysed in the present study. As a result, quantitative analysis of corpora should be used in conjunction with qualitative historical research, especially in the case of collections which are highly unevenly available and institutionally represented, such as archives.

The results also have implications for colonial language policies as they affect the continuing influence. Vocabulary of administration developed under colonial rule often continued to be used in public institutions after political independence, for example in the law, the administration of education and government communication. Present-day research also shows that the educational practice, official communication, and linguistic inclusion in many post-colonial societies are still impacted by colonial language policies (Mehboob & Zafar, 2025). In some cases, these historical legacies have led to the marginalisation of indigenous language, and colonial administrative language(s) continue to dominate in some settings (Naveed et al., 2025). Recent bibliometric evidence also shows that there is a negotiation of the balance between the inherited colonial language structure and more inclusive language policies in higher education systems throughout the Global South (Banerjee & Pan, 2024).

Overall, the study shows that corpus analysis of historical texts is a viable method for the analysis of the relationship between language, institutional governance and imperial administration. The research combines archival screening and computational linguistic analysis to provide empirical evidence of the administrative language that is maintained in the colonial texts and to provide a picture of the impact of institutional functions on language use. These results add to the scholarship on colonial documentary traditions and serve as building blocks for further research on colonial documentary traditions with larger archival corpora and more varied historical collections.

5. Conclusion

This study investigated the correlation between language, imperial administration and public institution with a historic corpus-based analysis of colonial era texts. The study showed how administrative language was embedded in the colonial documentary practices and institutional communication through the use of archival screening and computational linguistic techniques. While no temporal differences were found in the statistical analysis of the Administrative Language Index between the historical periods analysed, there were nevertheless interesting patterns in the use of administrative, legal, educational and institutional language within the selected corpus. These patterns suggest that the colonial administration of government was based on a standardized use of language to facilitate continuity in government, legal systems, and institutional authority. The study is a contribution to historical linguistics and to colonial studies in general in that it demonstrates the value of combining a corpus analysis with archival research for the study of historical documentation evidence. The systematic approach to the construction of the corpus and the use of a dictionary as the basis for the linguistic framework offer a clear methodology that can be transferred to other historical collections and institutional records. Moreover, the results of the institutional comparisons reveal that

each governmental unit of the colonies had its own linguistic profile which depended on its administrative role and responsibilities. These contributions notwithstanding, the study was limited by the relatively small number of texts of historical relevance found after screening the corpus, as well as the uneven distribution of texts over the colonial periods. Further studies may be conducted to enlarge the database by incorporating official government documents from archives, legal records, educational documents, and colonial documents in various languages and from diverse imperial situations. Adding more sophisticated uses of advanced NLP tools to larger collections from archives would further enhance the identification of linguistic change and a more complete picture of the function of language as a persistent means of imperial governance and institutional power.

References

1. Awal, A. (2023). The evolution of linguistic rights throughout history and the major milestones. *Indones. J. Adv. Res*, 2(9), 1317-1334.
2. Banerjee, S., & Pan, A. (2024). From colonial legacies to linguistic inclusion: A BERTopic enhanced bibliometric insight into global south higher education. *IEEE Access*, 12, 117418-117435.
3. Bejan, M. (2021). 15000 Gutenberg Books [Data set]. Kaggle. <https://www.kaggle.com/datasets/mateibejan/15000-gutenberg-books>
4. Berda, Y. (2022). *Colonial bureaucracy and contemporary citizenship*. Cambridge University Press.
5. Bettney Heidt, E. (2023). A critical examination of language ideologies and policies in an international school in Colombia. *Teachers College Record*, 125(7-8), 276-299.
6. Burak, G., Rothman, E. N., & Ferguson, H. (2022). Toward early modern archivality: The perils of history in the age of neo-Eurocentrism. *Comparative Studies in Society and History*, 64(3), 541-575.
7. Cho, H., & Robert, M. (2024). Medical schools in empires: connecting the dots. *Medical history*, 68(2), 110-127.
8. Gao, Y., & Moratto, R. (2024). Language and translation policies in China's multilingual governance: A study of the early and mid-Qing dynasty. *Language Policy*, 23(2), 169-192.
9. Gonzales, W. D. W. (2026). Variation and change in Philippine languages: Trends, challenges, and pathways forward. *Asia-Pacific Language Variation*, 12(1), 72-111.
10. Hilden, I. (2022). *Absent Presences in the Colonial Archive: Dealing with the Berlin Sound Archive's Acoustic Legacies*. Leuven University Press.
11. Jayaraj, T., & Navas, K. C. (2024). Semantics of power: Written communication, formal Documentation and codified law in British Malabar. *International Journal for the Semiotics of Law-*Revue internationale de Sémiotique juridique**, 37(7), 2151-2174.
12. Jordan, B. (2024). *Imperial power, provincial government, and the emergence of Roman Asia, 133 BCE-14 CE*. Oxford University Press.
13. Laitin, D. D., & Ramachandran, R. (2025). The historical sources of language policy. *The Journal of Politics*, 87(3), 967-981.
14. Lukianenko, N. (2024). Language and power: linguistic imperialism. *International Science Journal of Education & Linguistics*, 3(5), 41-49.
15. Makoni, S., Severo, C., & Abdelhay, A. (2023). Postcolonial language policy and planning and the limits of the notion of the modern state. *Annual Review of Linguistics*, 9(1), 483-496.
16. Mehboob, T., & Zafar, R. (2025). Language policy and medium of instruction in pakistan's education system. *ASSAJ*, 3(01), 1495-1509.
17. Namiq, F., & Zafarghandi, A. M. (2025). English as an imperial language: a qualitative study of Iraqi postgraduate EFL students. *Discover Education*, 4(1), 406.
18. Naveed, H., Mehdi, M., & Bhar, S. K. (2025). Linguistic Genocide of Indigenous Languages in Pakistan and India: A Comparative Analysis of Language Policies and Practices. In *Forum for Linguistic Studies* (Vol. 7, No. 9, pp. 698-720).
19. Popescu, B. G. (2023). *Imperial borderlands: institutions and legacies of the Habsburg military frontier*. Cambridge University Press.

20. Pratama, R. K. (2025). Power, Language, and Identity: A Foucauldian Analysis of Post-Colonial Language Education Policy in Indonesia. *Acta Paedagogica Vilnensia*, 55, 23-39.
21. Siddique, A. K. (2024). *The archive of empire: knowledge, conquest, and the making of the early modern British world*. Yale University Press.
22. Suman, A. K. (2023). Trailblazers Of Knowledge: Old Delhi College's Enduring Influence On India's Intellectual Landscape. *Journal of Language and Linguistic Studies*, 16(1).
23. Warditz, V. (2026). Modelling postcolonial transferred bilingualism: Russian in multilingual post-Soviet diasporas. *International Journal of Bilingualism*, 13670069261419590.
24. Young, C. (2023). Ethnicity and the colonial and post-colonial state in Africa. In *Ethnic groups and the state* (pp. 57-93). Routledge.
25. Zeng, J., & Li, X. (2023). Ideologies underlying language policy and planning in the Philippines. *Humanities and Social Sciences Communications*, 10(1), 405.