TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Machine Learning in Modeling Historical Registers – A New Perspective to Text Linguistics

Liina Repo

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. B OSA – TOM. 759 | HUMANIORA | TURKU 2026

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

MACHINE LEARNING IN MODELING HISTORICAL REGISTERS – A NEW PERSPECTIVE TO TEXT LINGUISTICS

Liina Repo

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. B OSA – TOM. 759 | HUMANIORA | TURKU 2026

University of Turku

Faculty of Humanities
School of Languages and Translation Studies
Digital Language Studies
Doctoral Program in Languages and Translation Studies, Utuling

Supervised by

Professor Veronika Laippala
University of Turku
Turku, Finland

Assistant Professor Brett Hashimoto
Brigham Young University
Provo, Utah, USA

Reviewed by

Assistant Professor Daniel Keller
Western Kentucky University
Bowling Green, Kentucky, USA

Docent Turo Hiltunen
University of Helsinki
Helsinki, Finland

Opponent

Assistant Professor Daniel Keller
Western Kentucky University
Bowling Green, Kentucky, USA

ISBN 978-952-02-0514-0 (PRINT)
ISBN 978-952-02-0515-7 (PDF)
ISSN 0082-6987 (Print)
ISSN 2343-3191 (Online)
Painosalama, Turku, Finland 2026

UNIVERSITY OF TURKU
Faculty of Humanities
School of Languages and Translation Studies
Digital Language Studies
LIINA REPO: Machine Learning in Modeling Historical Registers – A New Perspective to Text Linguistics
Doctoral Dissertation, 173 pp.
Doctoral Program in Languages and Translation Studies, Utuling
January 2026

ABSTRACT

This dissertation explores the insights into historical linguistic variation that can be gained through automatically identifying registers in large historical corpora, as well as the role of register variation in shaping these insights. Registers, i.e., situationally defined text varieties, are central to interpreting linguistic variation. This thesis investigates how existing annotated resources can be leveraged to enrich unannotated datasets, how variation between and within registers affects prediction reliability, and how linguistically interpretable features can deepen our understanding of register-specific language use.

Across three studies, this thesis integrates supervised machine learning with qualitative feature analysis, training models on the manually annotated Corpus of Founding Era American English (COFEA) and applying them to the large, heterogeneous Eighteenth Century Collections Online (ECCO). Study I models register variation within COFEA and demonstrates the feasibility of automatic register classification for historical texts, with feature analyses confirming that the model acquires meaningful, register-specific patterns (e.g., verbal and interpersonal features in letters). Study II extends the classification to ECCO, showing that models trained on COFEA generalize to ECCO for well-defined registers (e.g., letters, cases) but face challenges with hybrid categories, corpus-specific differences, and OCR-induced noise. Model explainability method Integrated Gradients highlights shared situational and linguistic cues behind both correct predictions and systematic misclassifications. Study III shifts focus to intra-document variation, demonstrating that text beginnings are most reliable for register prediction and that models capture stable, meaningful linguistic patterns across text segments. Keyword analyses confirm stable, linguistically motivated cues (e.g., interpersonal and informational features in letters) that persist across text parts.

Together, the studies offer new methods for enriching historical corpora with register information. Moreover, the results clarify how register variation shapes model behavior and deliver interpretable linguistic insights that strengthen corpus usability for research in historical linguistics, legal history, and digital humanities.

KEYWORDS: eighteenth-century English, historical computational linguistics, register, metadata annotation, digital humanities

TURUN YLIOPISTO
Humanistinen tiedekunta
Kieli- ja käännöstieteiden laitos
Digitaalinen kielentutkimus
LIINA REPO: Machine Learning in Modeling Historical Registers – A New Perspective to Text Linguistics
Väitöskirja, 173 s.
Kieli- ja käännöstieteiden tohtoriohjelma, Utuling
Tammikuu 2026

TIIVISTELMÄ

Väitöskirja tarkastelee, millaisia näkökulmia historiallisen kielen variaatioon saadaan, kun suurten historiallisten tekstikorpusten rekisterit tunnistetaan automaattisesti, sekä miten rekisterivaihtelu vaikuttaa näihin havaintoihin. Rekisterit, eli tilanteisesti määrittyvät tekstilajit, ovat keskeisiä kielen vaihtelun tulkinnassa. Tutkimus selvittää, miten luokiteltuja aineistoja voidaan hyödyntää luokittelemattomien aineistojen täydentämisessä, miten rekisterien välinen ja sisäinen vaihtelu vaikuttaa ennusteiden luotettavuuteen, sekä miten kielelliset piirteet voivat syventää ymmärrystä rekisterikohtaisesta kielenkäytöstä.

Väitöskirjan kolme osatutkimusta yhdistävät koneoppimista ja kielenpiirteiden laadullista tarkastelua. Koneoppimismallit koulutetaan käsin luokitellulla Corpus of Founding Era American English (COFEA) -korpuksella ja niitä sovelletaan laajaan, luokittelemattomaan Eighteenth Century Collections Online (ECCO) -aineistoon. Osatutkimus I mallintaa rekisterivariaatiota COFEA:ssa ja osoittaa automaattisen luokittelun mahdollisuuden historiallisilla teksteillä. Kielenpiirteiden analyysi puolestaan paljastaa rekisterikohtaisia säännönmukaisuuksia, kuten verbaalisia ja interpersoonallisia piirteitä kirjeissä. Osatutkimus II laajentaa luokittelun ECCO:on ja osoittaa, että COFEA:lla koulutetut mallit toimivat hyvin joissakin rekistereissä (esim. kirjeet, oikeustapaukset), mutta haasteita tuovat hybridiluokat, korpuskohtaiset erot ja OCR-virheet. Mallien selitettävyysmenetelmä Integrated Gradients paljastaa, mitkä kielelliset piirteet ovat yhteisiä sekä onnistuneille ennusteille että toistuville virheille. Osatutkimus III tarkastelee dokumenttien sisäistä vaihtelua ja osoittaa, että tekstien alut ennustuvat luotettavimmin ja että mallit tunnistavat johdonmukaisia ja merkityksellisiä kielellisiä piirteitä eri tekstiosissa. Avainsana-analyysi vahvistaa, että tietyt kielelliset piirteet (kuten interpersoonalliset ja informatiiviset piirteet kirjeissä) säilyvät läpi tekstin.

Tutkimus tarjoaa menetelmiä historiallisten korpusten rikastamiseen rekisteri-tiedolla ja osoittaa rekisterivariaation vaikutuksen koneoppimismalleihin. Samalla se tarkastelee, miten korpusten käytettävyyttä voidaan parantaa historiallisessa kielentutkimuksessa, oikeushistoriassa ja digitaalisissa ihmistieteissä.

ASIASANAT: 1700-luvun englanti, historiallinen laskennallinen kielitiede, rekisteri, metadatan annotointi, digitaaliset ihmistieteet

Acknowledgements

This PhD journey has been much more than a writing process; it has been a continuous endeavor of acquiring knowledge and developing expertise. When I began, my understanding of the topic was limited, but through dedicated study and the guidance of highly knowledgeable people around me, I was able to grasp concepts that were entirely new to me. This dissertation is the result of not only my efforts but also the support, encouragement, and generosity of many people and institutions.

First and foremost, I extend my sincerest gratitude to my thesis supervisors, Professor Veronika Laippala and Assistant Professor Brett Hashimoto, for their unwavering guidance, support, and encouragement throughout this journey. Their constructive feedback, insightful observations, and expert advice have been invaluable in shaping my research. I am particularly grateful to Professor Laippala for her constant encouragement over the years and for being the first person to suggest that I pursue PhD studies. At the time, I was working on her project and hadn't even considered doing a doctorate. Thank you, Veronika, for believing in me and for opening a door I didn't even know existed.

I am also deeply thankful to my pre-examiners, Assistant Professor Daniel Keller (Western Kentucky University) and Docent Turo Hiltunen (University of Helsinki), for their thoughtful and constructive feedback. I especially wish to thank Assistant Professor Keller for kindly agreeing to serve as the opponent in the public examination of this thesis.

My appreciation extends to all members of the TurkuNLP research group, who kept me informed and inspired on topics related to machine learning, and to the HELDIG community at the University of Helsinki, whose assistance was crucial for my second paper.

I am profoundly grateful for the financial support that made this work possible. My thanks go to the Finnish Cultural Foundation, the Otto A. Malm Foundation, the Emil Aaltonen Foundation, Savonlinna Rotary Club, the TOP Foundation, and especially the University of Turku School of Languages and Translation Studies doctoral program Utuling. Their trust in my research enabled me to complete this thesis. I also acknowledge the grants provided by Utuling, the Langnet network, the

EC2U Alliance, and the Turku University Foundation, which allowed me to attend conferences, summer schools, and research visits that enriched my academic experience.

My daily life in the academic circle would not have been the same without my fellow doctoral candidates, colleagues, and friends who helped me to carry on. Thank you for the countless discussions, lunches, coffees, and travels – not to mention the shared laughs and sorrows – over the past years. My gratitude for these moments goes to Anna Ristilä, Frederike Schram, Hanna-Mari Kupari, Jenna Saarni, Saara Hellström, Selcen Erten Johansson, Otto Tarkka and Valtteri Skantsi. Hanna-Mari, Anna, and Saara, thank you for the numerous coffee sessions and continuous peer-support you have given me. Saara, thank you for being my second pair of eyes and for reading all my drafts and applications with such care. Jenna and Frederike, I wish to thank you for becoming such good friends during this journey. The countless conversations, moments of joy, and little escapes from work made these years so much more enjoyable.

My family members and friends have eagerly followed my writing process and supported me to their best ability. Thank you, Kati, Samppa, and Ida, for your support throughout these years and for making my life so much more enjoyable.

Finally, with all my heart, I thank you, Lassi, for always being there for me.

December 2025

Liina Repo

Table of Contents

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text as follows:

Study I Repo, Liina, Brett Hashimoto, and Veronika Laippala. 2023. In Search of Founding Era Registers: Automatic Modeling of Registers from the Corpus of Founding Era American English. *Digital Scholarship in the Humanities*, Volume 38, Issue 4: 1659–1677. https://doi.org/10.1093/llc/fqad049

Study II Repo, Liina, Brett Hashimoto, Aatu Liimatta, Lassi Saario, Tanja Säily, Iiro Tiihonen, Mikko Tolonen, and Veronika Laippala. 2024. Towards Automatic Register Classification in Unrestricted Databases of Historical English. In *Linguistics across Disciplinary Borders: The March of Data*, eds. S. Coats and V. Laippala: 97–126. https://doi.org/10.5040/9781350362291.0011

Study III Repo, Liina, Brett Hashimoto, and Veronika Laippala. Accepted for publication. From Unannotated to Annotated: The Impact of Text Internal Variation on Register Predictions of Long Historical Documents. *International Journal of Corpus Linguistics*.

List of Co-Authored Publications Not Included in This Thesis

In addition to the publications included in this thesis, the following co-authored peer-reviewed publications, in reverse chronological order, were produced during the doctoral studies but are not included in this doctoral dissertation:

Henriksson, Erik, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson, Anni Eskelinen, Liina Repo, and Veronika Laippala. 2024. From Discrete to Continuous Classes: A Situational Analysis of Multilingual Web Registers with LLM Annotations. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, 308–318, Miami, USA. Association for Computational Linguistics.

Myntti, Amanda, Liina Repo, Elian Freyermuth, Antti Kanner, Veronika Laippala, and Erik Henriksson. 2024. Intersecting Register and Genre: Understanding the Contents of Web-Crawled Corpora. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, 386–397, Miami, USA. Association for Computational Linguistics.

Rastas, Iiro, Yann Ciarán Ryan, Iiro Tiihonen, Mohammadreza Qaraei, Liina Repo, Rohit Babbar, Eetu Mäkelä, Mikko Tolonen, and Filip Ginter. 2022. Explainable Publication Year Prediction of Eighteenth Century Texts with the BERT Model. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, 68–77, Dublin, Ireland. Association for Computational Linguistics.

Repo, Liina, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. Beyond the English Web: Zero-Shot Cross-Lingual and Lightweight Monolingual Classification of

Registers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 183–191, Online. Association for Computational Linguistics.

Laippala, Veronika, Samuel Rönnqvist, Saara Hellström, Juhani Luotolahti, Liina Repo, Anna Salmela, Valtteri Skantsi, and Sampo Pyysalo. 2020. From Web Crawl to Clean Register-Annotated Corpora. In *Proceedings of the 12th Web as Corpus Workshop*, 14–22, Marseille, France. European Language Resources Association.

1 Introduction

In recent years, the humanities have seen a shift toward embracing big data (Ekbia et al. 2015; Cha 2023). Over the years, the field has significantly advanced through the digitization of vast quantities of historical documents, such as the digitized historical newspaper collection of the National Library of Finland. These digitization campaigns have revolutionized access to historical texts, providing researchers with various resources for studying historical language and texts. For example, resources like Eighteenth Century Collections Online (ECCO), which includes over 200,000 texts from the 18th century, offer unprecedented opportunities for studying historical language use, cultural trends, and textual practices across disciplines, including linguistics, history, literature, and law.

This influx of data presents both challenges and opportunities for researchers, as the sheer volume of available information often lacks established methods for effective analysis. The growing abundance of digital resources, combined with advancements in computing power, has made the use of quantitative methods in historical and linguistic disciplines more feasible, though still complex to implement due to issues like incomplete digital collections and the need for specialized analytical skills (Milligan 2022). However, despite the potential of these databases and collections, many of them lack detailed metadata, particularly regarding registers.

Registers, defined as situationally specific text varieties, such as letters, essays, or news articles, are key predictors of linguistic variation and are essential for contextualizing and interpreting modern and historical texts (Biber 2012; Wright 1994). Registers provide insights into the situational contexts and communicative purposes of language use, thus shaping our interpretation of texts (Biber & Gray 2013). The analysis of registers is crucial for understanding how language functions across different domains (Biber 1988; Biber & Conrad 2019). The absence of register information limits the usability of these databases for comprehensive linguistic analysis, highlighting the need for methods to automatically identify registers in historical corpora and to reveal the linguistic patterns that define and differentiate registers in historical English.

Although register variation has been extensively studied in modern corpora, research on early and late modern English registers remains relatively scarce. This gap is particularly evident in large historical databases like ECCO, which lack register annotations. ECCO and other large historical databases and corpora offer opportunities to explore linguistic trends and patterns from the past. For instance, they can be used to investigate how lexical associations in historical print corpora reveal the social and semantic construction of concepts like "Protestant" in eighteenth-century Britain (Regan 2022), or how reused text passages can be systematically identified to uncover intertextual relationships and the circulation of ideas across genres (Mahadevan et al. 2025). These corpora contain vast arrays of texts that reflect the linguistic, social, political, and cultural contexts of their time.

However, the manual annotation of registers in such extensive datasets is impractical due to the sheer volume of data. In particular, register identification requires expert knowledge and careful analysis of situational and linguistic features, making it a time-consuming and labor-intensive task. This process involves reading and categorizing each text, which is not feasible for extensive datasets like ECCO. Therefore, there is a need for automatic methods to identify and classify registers in historical texts, thereby enhancing the usability of large historical corpora. At the same time, automatic annotation methods face a bottleneck in that they require high-quality, manually annotated training data, which are costly and time-consuming to produce (Jenset & McGillivray 2017). This requirement creates a vicious cycle in which the lack of annotated data hampers the development of automatic methods, and the impracticality of manual annotation prevents the creation of the necessary training data. Additionally, the challenge is compounded by the fact that we often do not know what registers are present in these corpora or what their linguistic characteristics look like. This lack of knowledge makes it difficult to define annotation targets and train models effectively. As a result, there is a pressing need for automatic approaches that can reliably identify and classify registers in historical corpora, while overcoming the challenges posed by data heterogeneity and linguistic variation. Addressing this issue is essential for enhancing the usability of large-scale historical datasets and for advancing research in historical corpus linguistics.

This doctoral dissertation addresses the challenge of register annotation in historical corpora, with a special focus on leveraging existing annotated data while also considering the challenges posed by linguistic variation within late modern English texts. The study focuses on two corpora: the Corpus of Founding Era American English (COFEA), which contains manually annotated texts spanning a wide range of registers, including newspapers, personal journals, and court cases, and ECCO, a large but unannotated corpus. These corpora differ in structure, content, and metadata availability, thus offering valuable testbeds for exploring register classification across heterogeneous datasets.

Methodologically, this research introduces a new approach that combines qualitative linguistic analysis with quantitative machine learning (ML) techniques. This work employs both pre-neural models and deep learning language models to study register variation in ECCO and COFEA. These models are trained on COFEA and applied to ECCO to evaluate their ability to generalize across corpora. This study emphasizes the importance of interpreting predictions through linguistic features to ensure that the models not only perform well but also provide meaningful insights into register-specific language use. This approach has benefits for the linguistic analysis of register, enabling a deeper understanding of historical language variation and usage.

The focus of the research is two-fold. First, it explores ways by which we can use existing manually annotated data to train ML models for automatic register identification. Second, it investigates the impact of linguistic variation within and between registers on the accuracy of these models. By integrating qualitative and quantitative approaches, this research not only aims to enhance the accuracy of register classification but also to uncover the underlying linguistic features characterizing different registers in the late modern era. Overall, this study contributes to the field of historical linguistics by generating insights that can be applied when processing or analyzing other historical corpora and studying unannotated corpora.

1.1 Research Questions

This thesis and the included publications investigate the automatic identification and analysis of registers in historical corpora and how linguistic variation affects these registers and their predictions. Specifically, the following research questions (RQs) guide this study:

- RQ1: How can existing manually annotated register data from a curated corpus like COFEA be leveraged to automatically annotate registers in a heterogeneous corpus like ECCO?
- RQ2: How does variation between and within registers affect the register predictions?
- RQ3: How can the examination of linguistic features enhance the understanding of register-specific language use and contribute to metadata enrichment in large historical corpora?

RQ1 explores the transferability and applicability of manually annotated register data from one corpus to another from the same era and the process by which curated data can be used to enrich unannotated datasets from the same historical period. RQ1 also reflects a practical concern: how to make use of limited annotated resources to

improve the usability of much larger corpora. Following that, RQ2 attempts to understand the impacts of register variation, both across different registers and within individual register categories, on the reliability of register predictions. Finally, RQ3 investigates how the detailed examination of linguistic features can enhance our comprehension of register-specific language use, thereby providing a more nuanced understanding of historical language variation.

By identifying consistent linguistic patterns associated with specific registers, this study supports the broader goal of making sense of large, messy, and heterogeneous datasets, such as ECCO. This has direct implications for bringing more materials within reach for researchers by improving the searchability, interpretability, and usability of existing historical text collections. Doing so enables more targeted exploration and facilitates better interdisciplinary research across linguistics, history, and digital humanities. More broadly, the RQs can be summed up as exploring what kinds of insights into historical linguistic variation can be gained through the automatic identification of registers and how register variation influences these insights.

This dissertation consists of three studies, including two journal articles and one book chapter in an edited volume, along with this thesis summary. While the individual studies inevitably overlap in their coverage of the RQs, Table 1 highlights which questions are most relevant to each study. The individual studies are presented in more detail in Section 4.3, and the research questions are revisited in Section 5.1 in light of the aforementioned studies.

Table 1. Overview of the research questions in relation to the included studies.

RESEARCH QUESTIONS	RQ1	RQ2	RQ3
Study I		X	X
Study II	X	X	X
Study III	X	X	X

1.2 Structure of the Thesis

As mentioned previously, this dissertation consists of three previously published studies and a summary. The main research is presented in the studies. The structure of the thesis is as follows. Chapter 2 provides a theoretical foundation by exploring the concept of *register*, distinguishing it from *genre*, and reviewing previous research in the field, including corpus-based approaches and historical perspectives. Chapter 3 shifts focus to computational methods, offering an overview of ML in language studies from traditional techniques to recent advances in deep learning and large

language models (LLMs). It also discusses model explainability and the role of these models in analyzing historical texts and register variation. Chapter 4 presents the material and methodology used in this thesis. It describes the corpora, addresses challenges related to historical data, and outlines the data preparation and model training processes. This chapter also includes an overview of the three peer-reviewed papers comprising the core of the thesis. Chapter 5 synthesizes the findings, revisits the RQs, discusses limitations, and proposes directions for future research. Finally, Chapter 6 concludes the thesis.

2 Register

Understanding how language varies across different contexts is a fundamental line of research in corpus linguistics. One of the most influential frameworks for capturing this variation is the concept of *register*, which refers to situationally and culturally defined varieties of language that are shaped by the purpose, audience, and communicative setting of a speech or a piece of text. Registers help explain why the same language can produce vastly different linguistic patterns depending on the context. Register is a key factor in predicting linguistic and functional variation in general contexts (e.g., Biber 1988) and within specific subdomains, such as blogs (Grieve et al. 2011), university student writings (Nesi & Gardner 2012), or historical news (Biber & Gray 2013). Register analysis bridges the gap between linguistic theory and empirical data by providing a framework for interpreting the functional dimensions of language variation that can be observed in corpora, thereby allowing researchers to explore how linguistic choices are systematically influenced by contextual factors.

This chapter outlines the concept of *register* as a foundational element in corpus-based studies of linguistic variation. It begins by defining what a register is, then distinguishes it from related concepts such as *genre*, and finally reviews key studies that have shaped the current understanding of register variation across different domains.

2.1 What is a Register?

The term "register" began to be used by early twentieth-century anthropological linguists, who applied the term in their studies of situational, social, and descriptive analyses of non-Western languages and social contexts (Biber & Finegan 1994: 4–5). By the late 1950s and 1960s, scholars were interested in how linguistic forms in various languages were influenced by communicative purposes and situational contexts, such as the study of American English address terms (Brown & Ford 1961) and conversational language (Crystal & Davy 1969). To date, the term "register" has become widely used and is often understood to mean language varieties related to specific situational contexts and functional language use (Biber & Conrad 2019; Crystal 1991; Ferguson 1994). The precise definition of register is not universally

agreed upon, and various scholars and theoretical frameworks have offered differing interpretations of the term.

Among the theoretical frameworks, Biber's (2012) text-linguistic framework, which refers to a theoretical and methodological approach for studying register variation, offers a well-established approach that is particularly relevant for the current thesis. This framework uses quantitative methods to study registers from a comparative perspective and is based on two fundamental claims: 1) linguistic features are functionally motivated and 2) linguistic features arise from the communicative needs of different situations (Biber 2019). Registers within the text-linguistic framework encompass functionally motivated variations in vocabulary, syntax, and style. These features emerge more frequently in certain situational contexts than others, as they are shaped by the communicative situation and are better suited to that specific context (Biber & Conrad 2019). These contextual factors include the relationships among participants (e.g., speaker and listener), the mode (e.g., spoken or written), the medium (e.g., print, handwriting, or radio), and the communicative purpose (e.g., entertaining, persuading, or informing) of the text.

Recently, a growing body of work in the text-linguistic framework has conceptualized registers as culturally constructed categories of text rather than scientifically defined ones, placing emphasis on the fact that their recognition is rooted in shared linguistic and social conventions rather than fixed linguistic criteria (Biber et al. 2020; Biber & Egbert 2023). Registers are often identified and labeled based on shared social understandings of communicative situations, such as academic writing, news reporting, or recipes. Although these register categories are typically characterized by typical situational and linguistic features, the categories are not defined by these characteristics (Biber & Egbert 2023). Recently, research on registers has also acknowledged that register boundaries are not always clear-cut, and there is a possibility that some texts do not represent any culturally recognized register category, while others may have features of multiple registers (Biber & Egbert 2018, 2023). These kinds of hybrid texts emerge particularly in contexts wherein communicative practices are fluid or evolving, such as in digital discourse on the web (Santini 2007). For instance, a blog post may combine elements of narrative and informational writing and employ persuasive language.

In the text-linguistic framework, register features are inherently distributional due to their functional nature, meaning that the features can appear in any register, but are more common in specific registers (Biber & Conrad 2019). Registers are characterized by linguistic elements that are distributed throughout the text, frequently occurring in particular registers because they match the situational context and communicative goals. However, studies have also begun to analyze registers in situationally and linguistically continuous terms, leading to the formulation of a more bottom-up approach. Registers can be thought of existing on a continuum and

texts are coded for the extent to which they exhibit features associated with various registers (Biber et al. 2020; Henriksson et al. 2024a).

Aside from the text-linguistic framework, another complementary view is systemic functional linguistics (SFL), which assumes that registers are dynamic and evolving configurations of language shaped by the social and situational contexts in which communication takes place (Halliday & Hasan 1985; Matthiessen 2019). In SFL, registers are viewed as configurations of contextually motivated language choices that can be divided into three key variables: field (types of social activities), tenor (relationships between participants), and mode (the channel of communication; Halliday & Hasan 1985). These variables not only shape the lexical and grammatical choices of the writer or speaker but also guide how meaning is constructed and conveyed for different communicative purposes. For example, fairy tales typically follow recognizable story patterns that are associated with specific contextual variables, such as the field involves imaginary characters, the tenor is engaging or cautionary, and the mode reflects language that is accessible to children (Rose 2023).

Despite the differences between the text-linguistic framework and SFL's view on registers, as well as other scholars' conceptualization of the term (see, e.g., Reid 1956; Ferguson 1968), the overarching characteristics are shared. In general, these definitions agree that registers can be defined by linguistic or situationally contextual characteristics. In this thesis, I will be using Biber's text-linguistic framework from a corpus linguistics perspective.

2.2 Register versus Genre

The distinction between genre and register is often unclear, and these are often treated as synonymous in studies that model different text varieties due to their overlapping definitions (for an overview, see, e.g., Lee 2001). Registers are commonly used in corpus and text-linguistic studies (e.g., Biber & Conrad 2019), whereas genres are typically used in literary and discourse analysis (e.g., Swales 1990), as well as text classification research (e.g., Sharoff 2018).

From a discourse analytical perspective, specifically within the systemic functional linguistics framework, genre is a higher-level construct that can be understood as culturally shared patterns of meaning that guide how texts (and other semiotic modalities, such as gestures and images) are produced and understood within a community (Martin & Rose 2008). Genre is realized through different contextual and functional variables, field, tenor, and mode, that are called registers. Therefore, within this framework, registers serve as the linguistic means through which genres are enacted. While genre provides the overarching social purpose and structural organization of a text, register determines the specific linguistic choices that align with the situational context.

In corpus linguistics, specifically in text linguistics studies, the approach tends to be more data driven. Unlike in the hierarchical framework in SFL, corpus-based studies treat registers and genres as separate entities. Biber and Conrad (2019), for example, differentiate genre from register by highlighting the former's focus on specialized expressions and the rhetorical organization of texts. Unlike register-based features, genre-specific linguistic elements are not pervasive throughout a text; instead, they appear as specialized expressions in specific parts of the text, such as the conventional ways to begin or end a letter (Biber & Conrad 2019). Therefore, genre features are conventional, whereas register features are functional.

In this dissertation, we follow the long tradition of corpus linguistics and examine text varieties through their linguistic and situational features from a register perspective. By focusing on register rather than genre, the thesis emphasizes functional variation over structural conventions. This approach allows for a more nuanced analysis of historical texts, especially those that do not conform neatly to established genre categories but still exhibit distinct communicative functions. However, when discussing previous works, we include studies that refer to either *genre* or *register*, recognizing that many past works do not clearly differentiate between the two terms.

2.3 Previous Research on Registers

This section provides an overview of the foundational and evolving research on registers in corpus linguistic studies. It outlines the theoretical underpinnings of register studies, explores patterns of variation across and within registers, presents key methodological frameworks for studying registers, and highlights important contributions to the study of historical registers.

2.3.1 Foundations of Register Studies in Corpus Linguistics

Rather than relying on introspective or anecdotal evidence, register studies are grounded in empirical data drawn from carefully constructed corpora. In particular, corpus linguistic register studies rely on large corpora that are systematically compiled and manually annotated with the aim of representing a wide range of communicative situations. These corpora are designed to capture a broad spectrum of language use and typically include texts from multiple registers, such as academic prose, news articles, fiction, letters, drama, speeches, and sermons.

One of the foundational contributions to this field was the development of multi-register corpora, such as the Brown Corpus (see Francis & Kucera 1967) and ARCHER: A Representative Corpus of Historical English Registers. These corpora were compiled with the explicit goal of representing a wide range of text types, each

selected to reflect distinct communicative purposes. The Brown Corpus and ARCHER each include one to three million words across twelve to fifteen different registers. The inclusion of diverse registers enables researchers to systematically compare linguistic patterns and to identify how language use varies in accordance with situational factors. Notably, representativeness and balance have long been fundamental principles in the construction of corpora for register studies.

Each register is characterized by a unique distribution of lexico-grammatical features that reflect the communicative function of the register. These features include, for instance, different grammatical structures (e.g., past tense or passives), lexical choices (e.g., stance markers), and discourse patterns (e.g., referential expressions; Biber 2012). The distribution of these features across registers is not random but reflects the communicative goals and constraints of each context. For example, academic writing tends to favor nominalizations and passive constructions to convey objectivity and abstraction, while conversation often features contractions, first-person pronouns, and interactive markers (Biber 2006; Biber et al. 2002).

In addition to the availability of register-specific corpora, Biber's (1988) introduction of the multidimensional analysis (MDA) approach was an important advancement in the study of registers, as it provided a novel method for exploring contemporary and historical register variation. Biber (1988) examined the co-occurrence patterns of various linguistic features, which formed the basis for dimensions of variation. In turn, these dimensions were interpreted to reflect different communicative functions in speech and writing. The MDA approach is structured into three steps: 1) describing the situational characteristics of the register, 2) detailing the linguistic features of the register, and 3) interpreting the functional relationship between these characteristics and features (Biber & Conrad 2019). The MDA is inherently based on the text-linguistic register framework, as it systematically analyzes the situational and linguistic characteristics of registers to uncover the functional motivations behind language use in varying contexts.

In addition to MDA, Geometric Multivariate Analysis (GMA; Diwersy et al. 2014), which is based on MDA, has emerged as a method for interpreting and visualizing register variation across corpora. As such, GMA enables researchers to map texts into a multidimensional space based on linguistic features, revealing patterns of variation that are often obscured in purely statistical models. This technique has been applied to study register variation across varieties of English. For example, Neumann and Evert (2021) used GMA to distinguish registers in ICE corpora, while Frenken et al. (2025) confirmed the method's stability and replicability across nine English varieties.

Recent developments have also emphasized the use of keyness analysis and keyword extraction as alternative or supplementary methods employed in register analysis. These techniques are particularly useful in identifying salient linguistic

markers that distinguish registers, especially in large and diverse corpora. Keyness analysis, traditionally based on frequency comparisons between target and reference corpora, has evolved to incorporate dispersion-based metrics, thus offering a more nuanced view of lexical prominence. For example, Gries (2021) proposed a two-dimensional approach combining frequency and dispersion to improve the identification of register-specific keywords. Similarly, Egbert and Biber (2019) advocated for dispersion-sensitive keyword analysis to better capture the distributional characteristics of lexical items across texts. Keyword-based methods have also been applied to specialized domains. For instance, Larsen et al. (2025) used text-dispersion keyness analysis to develop subregister-specific word lists for legal contracts, thereby demonstrating how register-sensitive keyword extraction can enhance domain-specific vocabulary resources. Another approach is the sibling-texts keyword analysis (Cvrček & Berrocal 2025), which distinguishes between topic-related and register-related keywords by comparing parallel keyword analyses across texts with similar register profiles.

While MDA provides a holistic view of variation, it requires extensive manual coding and statistical expertise. Keyword-based methods can offer more targeted insights, especially when combined with statistical modeling, but may overlook deeper functional relationships. A growing body of research has applied ML techniques to register analysis, enabling automatic classification and large-scale annotation. This line of research underscores the ongoing evolution of register studies in the digital age — a development that will be discussed in more detail in Chapter 3.5. By combining traditional corpus methods with computational approaches, this study aims to uncover patterns of variation that are not easily captured through manual analysis alone. The integration of keyword-based and computational methods enables this thesis to address complex questions about register diversity and internal variation, thus contributing to more inclusive and scalable models of register analysis.

2.3.2 Register Variation

The study of registers has evolved over time, with various approaches contributing to our understanding of language variation across different contexts. Corpus linguistics has played a central role in this development, enabling studies concentrating on various aspects of variation. For example, some studies have inspected diachronic change in registers throughout a given period, whereas others have studied synchronically specific registers in more detail.

The specificity of the definition of a register can also vary, as studies have examined broader registers, such as newspaper texts, or more specific ones, such as editorial columns. Defining registers is possible at multiple levels; for instance, a

register within academic prose can be narrowly defined as textbooks, or even more specifically as the methods sections of research papers (see, e.g., Egbert 2015; Gries 2006).

The granularity or specificity of the definition of a register depends on the extent to which a study examines the situational characteristics of individual texts. At its widest scope, register analysis has explored differences between spoken and written language, uncovering fundamental parameters of variation, such as interactivity, planning, and contextual dependency (Biber 2012). These macro-level comparisons have been instrumental in establishing foundational dimensions of linguistic variation, such as informational density, narrative structure, and degree of involvement. Additionally, research has frequently addressed more specific text varieties, such as lectures, office hours, textbooks, or course management texts (Biber 2006). Some approaches have taken this even further, conceptualizing register at highly detailed levels. For example, studies have explored disciplinary variation in journal articles, revealing how fields such as biology, history, and linguistics differ in their use of passives, stance markers, and nominalizations (Gray 2015). Similarly, student writing has been examined across registers, such as argumentative essays, lab reports, and reflective writing, thereby highlighting how register and educational level influence linguistic choices (Staples et al. 2016).

In the literature, registers have often been studied as separate entities, with linguistic features analyzed within isolated categories, such as academic writing or conversation. While this approach has generated valuable insights, it has also led to a somewhat fragmented understanding of register variation. Recent work has emphasized the need to consider continuous situational spaces and overlapping features across registers, challenging rigid boundaries and encouraging more nuanced models of register variation (Biber et al. 2020)

While individual registers and the variation between registers have been extensively studied, recent research has revealed that significant intra-linguistic variation also exists within register categories typically viewed as well-defined (e.g., Egbert & Gracheva 2023; Goulart et al. 2022). Certain registers, such as academic prose, conversation, news reporting, and fiction, are typically considered stable due to their clearly defined communicative purposes and consistent linguistic patterns. This dual perspective of variation, both between registers and within individual registers, offers complementary insights in which each register possesses unique linguistic features, yet individual texts within the same register can still exhibit considerable variation (Biber & Egbert 2023).

Although variation within individual registers is acknowledged, document-internal register variation remains largely unexplored. This is especially relevant for long-form texts, such as historical literary works, novels, or full-length academic monographs, that often encompass multiple communicative purposes and stylistic

shifts across different sections. For example, a novel may alternate between narrative passages, character dialogue, and embedded documents (e.g., letters or newspaper clippings), each exhibiting distinct linguistic features and potentially aligning with different registers. Similarly, a historical treatise might blend expository, argumentative, and anecdotal segments, reflecting shifts in stance, audience engagement, and information density. These internal shifts challenge the assumption of register homogeneity within a single document and call for more fine-grained, segment-level analyses (e.g., Egbert & Mahlberg 2020). Study III in the current thesis, which focuses on this very issue, examines register variation within long historical texts and explores how internal shifts in linguistic features reflect changes in register.

2.3.3 Historical Perspectives on Register Studies

Historical register studies have played a crucial role in understanding how language use evolves across time and communicative contexts. Several studies have especially used MDA to analyze the evolution of English registers over different periods. One of the foundational studies in this area is that of Biber and Finegan (1989), who examined the diachronic development of three registers (letters, essays, and fiction) from 1650 to 1950. Their analysis revealed distinct patterns of linguistic change, demonstrating how registers diverge in their use of features, such as elaboration, involvement, and narrative style. Later, Biber and Finegan (1997) expanded their earlier diachronic studies (Biber & Finegan 1989, 1992) to include additional registers, such as drama and medical writing, thus illustrating how different communicative domains evolve in different directions.

González-Álvarez and Pérez-Guerra (1998) extended the historical scope further back, analyzing five English registers from the fifteenth to the seventeenth centuries. Their findings highlighted a shift in communicative orientation in which science and drama developed more oral-like features, while education and fiction became increasingly literate. Interestingly, the register of letters remained relatively stable, suggesting that some communicative functions maintain continuity despite the occurrence of broader linguistic shifts. Geisler (2002) focused on the nineteenth century, identifying a divide between non-expository and expository registers. They found that non-expository registers (e.g., fiction, letters, and drama) tended to become more concrete and interactional, whereas the expository register history evolved to be elaborated and more narrative. This contrast underscores how register development is shaped not only by linguistic trends but also by changing social and epistemological functions of the text.

More recent research has emphasized the value of register-specific analysis in historical linguistics. For instance, Kytö (2019) argued that focusing on individual

registers provides a more accurate picture of historical language use than attempting to generalize across entire periods. This approach has been adopted in studies of specialized registers, such as English medical writing (Taavitsainen et al. 2022), letters (Bergs 2004), and law reports (Fanego et al. 2017). Additionally, Culpeper and Kytö (2010) have contributed to the study of historical spoken interaction by emphasizing the importance of speech-related genres, such as dialogues and trial transcripts. Their work complements the traditionally text-focused orientation of historical corpus linguistics and highlights the need to consider orality in diachronic studies.

Efforts have also been made to compile multi-genre corpora that support the study of linguistic variation across registers and time periods. For instance, the Historical Corpus of Dutch (Van De Voorde et al. 2023) offers a richly layered dataset that encompasses a wide range of genres from Early and Late Modern Dutch. Similarly, Murphy (2019) developed a genre classification scheme for Early English Books Online, focusing on Shakespearean texts and their contemporaries. This effort highlights the need for more granular classification within large historical text collections. However, a discrepancy remains between the registers that have been extensively researched and those that are potentially available in historical archives and corpora. Most corpus-based studies have focused on a relatively narrow set of registers, such as fiction, letters, drama, and scientific writing, largely because these registers are well-preserved, accessible, and frequently included in curated corpora, including ARCHER, the Helsinki Corpus, and the Corpus of English Dialogues.

However, historical archives contain a much broader spectrum of textual materials that remain underexplored even though they could be considered registers. These include administrative records, commercial documents, technical manuals, ephemeral texts (e.g., pamphlets and posters), household guides, travel writing, and marginalia. Such materials often reflect everyday language use, specialized discourse communities, and non-elite communicative practices, offering valuable insights into linguistic variation that is not otherwise captured by literary or scholarly genres. Recent research has begun to acknowledge this gap. For example, Pescuma et al. (2023) emphasizes the need to investigate register phenomena across diverse modalities, time periods, and cultural contexts, calling for broader empirical coverage to include registers that have traditionally been excluded from corpus design due to classification challenges or lack of standardization.

Building on these insights, the present study contributes to the expanding field of historical register analysis by exploring how computational approaches, particularly ML, can be used to study and identify registers in large-scale historical corpora. This work argues that while traditional corpus-based methods have laid the groundwork for understanding register variation, ML offers new possibilities for uncovering patterns in underexplored registers. This approach not only aligns with

recent calls for more inclusive and empirically grounded studies of register diversity but also provides a methodological bridge between historical linguistics and contemporary computational techniques.

3 Machine Learning in Language Studies

ML algorithms usually predict patterns or make predictions from data. These algorithms are generally divided into two categories: supervised and unsupervised learning. Supervised learning involves training a model on a labeled dataset in which each instance has input features paired with an output label. The objective is to learn a mapping function from inputs to outputs that can be generalized to predict the labels of unseen instances (Bishop 2006). The process of supervised learning typically involves three steps: 1) selecting relevant features from data, 2) training the machine learning model on labelled examples, and 3) using the trained model to predict new and unlabeled data. Supervised learning can be used for tasks, such as part-of-speech tagging, where the model learns from a corpus of labeled sentences to predict the part of speech for each word in new sentences (see, e.g., Màrquez et al. 2000; Plank et al. 2016). Other examples of supervised learning include, for instance, error detection in which models are trained to classify and detect errors in learner language (see, e.g., Yannakoudakis et al. 2011; Farag et al. 2018), sentiment analysis in which models are trained to classify text as positive or negative (see, e.g., Pang et al. 2002; Xu et al. 2019), and grammatical relation identification in which models are trained to identify functional relations between constituents in a clause (see, e.g., Chen & Manning 2014; Tiedemann & Agić 2016).

Unsupervised learning, in contrast, involves training an algorithm on a dataset without labels. The aim is to find structures or groupings within a set of data points based on their similarities or co-occurrences (Bishop 2006). The main applications of unsupervised learning include clustering, association rule learning, and dimensionality reduction. Clustering is the process in which unlabeled data, such as texts or sentences, are grouped into clusters based on the similarities or differences the model finds in the data (see, e.g., Zhang et al. 2010). Association rule learning is a technique that finds associations between parameters of a dataset, such as in market basket analysis or identification of word co-occurrences (see, e.g., Ghous et al. 2023). Dimensionality reduction algorithms decrease the number of features while retaining the essential structure and information within the data (see, e.g., van der Maaten et al. 2009).

The key differences between supervised and unsupervised learning algorithms lie in the data requirements and objectives: Supervised learning focuses on predictive accuracy with labeled data, while unsupervised learning focuses on discovering patterns or structures in unlabeled data. On the one hand, supervised learning methods generally provide high accuracy but the cost- and labor-intensive creation of manually labelled data poses challenges when large amounts of data are required to train the models (Huang & Zhao 2024). On the other hand, unsupervised learning algorithms may be more cost effective and useful in exploratory analysis, but algorithms rely on assumptions about the data, and the usefulness and meaningfulness of the results depend largely on the interpretation of the researcher (Kolge & Jadhav 2018).

As a middle ground between supervised and unsupervised learning, semi-supervised learning leverages a small set of labeled data while also utilizing a large amount of unlabeled data to improve accuracy to reduce reliance on labeled data. Another approach to ML is reinforcement learning in which the algorithm learns from interacting with its environment. Rather than learning from labeled examples, the algorithm is rewarded or penalized based on its actions (Bishop 2006). Through trial and error and by exploring different actions, the algorithm learns which actions produce the best long-term outcomes.

Having introduced different types of ML techniques, the remainder of this chapter delves into the development, advancements, and implications of ML, especially in relation to historical linguistics, beginning with the introduction of pre-neural approaches of ML in Section 3.1. This is followed by an exploration of LLMs and deep learning architectures in Section 3.2. Section 3.3 discusses the nature and importance of model explainability, while Section 3.4. presents the applications of deep learning language models to historical data. Finally, Section 3.5 concludes by discussing register studies using deep learning language models.

3.1 Pre-Neural Machine Learning

Before deep neural networks, statistical approaches to corpus linguistics enabled significant advances in computational corpus research. Pre-neural ML approaches typically rely on manually engineered features, such as part-of-speech tags or syntactic dependencies, that have been manually encoded to the data. These feature-driven approaches are often limited by the quality of the data and their labels. Nevertheless, algorithms such as decision trees, naïve Bayes classifiers and Support Vector Machines (SVMs) offer powerful tools for analyzing language data on a larger scale. These tools facilitate the automatic classification of linguistic features, such as part-of-speech tagging (Màrquez & Rodríguez 1998); the identification of patterns in syntactic and semantic structures, such as word sense disambiguation

(Dixit et al. 2015); and the development of predictive models for text classification (Hsu 2020).

Among pre-neural models, SVMs (Vapnik 1998) were particularly important in text classification and regression tasks. In a classification situation, SVMs help predict the category to which a datapoint belongs. The objective is to find the optimal line (or hyperplane) that separates different categories in the data. This line is chosen based on the closest points from each category, called support vectors. The optimal hyperplane maximizes the distance between these points and the line, thus ensuring a clear separation between categories (Vapnik 1998).

In corpus linguistics, linear SVMs were often preferred for their interpretability and efficiency. Linear SVMs assume a linear relationship between the input feature and the output variable, which aligns well with many text categorization tasks that tend to be linearly separable. Because SVMs learn linear threshold functions, they are particularly well-suited for identifying clear boundaries between textual categories (Joachims 1998).

The robustness of SVMs to irrelevant features and their ability to generalize well from limited annotated data made them a popular choice prior to the rise of neural models. While the present thesis primarily explores more advanced models for register classification, understanding the principles and applications of SVMs, as detailed in Study I, offers valuable perspective on the evolution and comparative effectiveness of different classification techniques. In particular, linear SVMs can be used to rank how important each feature is for the model's performance (Guyon & Elisseeff 2003) and evaluate the importance of different linguistic features in distinguishing between classes (Kyröläinen & Laippala 2023). Each feature is assigned a weight, which the model utilizes to determine the class label. By identifying the most important features and their impacts on model performance, we can understand the data's linguistic characteristics and evaluate the importance of different linguistic features.

SVMs have been successfully applied in a large body of corpus linguistic research. For example, SVMs have been used in identifying different types of text, such as web genres (e.g. Pritsos & Stamatatos 2018) and in performing general text classification (Joachims 1998). Laippala et al. (2021) also used SVMs to automatically distinguish between various registers across the unrestricted web. Beyond genre and register classification, SVMs have been employed in information extraction tasks by identifying entities and relationships within unstructured text (Isozaki & Kazawa 2002; Hong 2005). SVMs have also been used in authorship attribution tasks to distinguish authors with varying linguistic features (full word forms and grammatical tags) in German newspapers (Diederich et al. 2003). SVMs are particularly well-suited for linguistic research because they are known for their strong generalization capabilities (Sun et al. 2009), even when dealing with high-

dimensional textual data. Hence, SVMs are deemed effective for analyzing different linguistic features, such as part-of-speech distributions, lexical features, and dependency relations.

While pre-neural ML methods relied on manually engineered features, the introduction of the word2Vec (Mikolov et al. 2013) technique marked an important shift towards neural ML. Word2Vec and similar methods, such as GloVe (Pennington et al. 2014), represented a new paradigm in which the model automatically learned vector representations of words from raw text and captured semantic and syntactic relations without the need for manual feature engineering. Following the distributional hypothesis (Harris 1954; Miller & Charles 1991), these vector representations of words, or embeddings, positioned a word in a continuous vector space in which words used in similar context gained similar representations. In corpus linguistic research, embedding models have been used to study the semantic similarity of words (Schnabel et al. 2015), diachronic shift in word meanings (Hamilton et al. 2016), and collocational patterns by identifying statistically significant word pairings and their semantic associations (Weeds et al. 2004). Overall, these new models foreshadowed the deep learning methods that would redefine computational approaches to language and text research.

3.2 Deep Learning and Large Language Models

The rise of deep learning in natural language processing (NLP) represented a significant change in ML approaches. Where pre-neural models relied on manually engineered features, deep learning introduced architectures capable of learning hierarchical representations of linguistic data directly from raw text. These advances in ML have made deep learning language models indispensable in NLP and digital humanities. Deep learning involves the use of artificial neural networks composed of multiple layers. Each layer transforms the input data it receives into increasingly abstract representations, allowing the network to gradually capture more complex features and patterns and eventually learn and make decisions (Madhiarasan & Louzazni 2022). Early deep learning architectures, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), offered improvements in modeling sequential or structured linguistic data, but their impact was limited by various training difficulties (Mikhaeil et al. 2022; Li et al. 2022). These limitations hindered the performance of language models on key NLP tasks such as machine translation, summarization, and question answering (Sutskever et al. 2014).

A breakthrough came with the introduction of the Transformer architecture (Vaswani et al. 2017). Unlike earlier models that processed sequences step-by-step, Transformers brought into practice a mechanism known as “self-attention”, which allowed a model to look at all words in the sentence at once and weigh the relevance

of each word regardless of their position in the sentence (Vaswani et al. 2017). As this approach moved beyond assigning single, fixed embedding to each word, this enabled the generation of context-sensitive representations. This was particularly impactful in terms of handling linguistic phenomena (e.g., polysemy) and affirming the notion that the meaning of a word could now be inferred from its surrounding context (Garí Soler & Apidianaki 2021).

The emergence of the Transformer architecture has led to the development of numerous pre-trained language models that are first trained on extensive raw text data coming from the internet. After the initial pre-training, the models are usually fine-tuned with smaller datasets for specific tasks. This *transfer learning* method enables the model to use existing knowledge and adapt to a wide range of different tasks, such as part-of-speech tagging (Chiche & Yitagesu 2022), sentiment analysis (Chan et al. 2023; Lyu et al. 2024), named entity recognition (Amin & Neumann 2021), or genre and register classification (Kuzman et al. 2023; Chen & Wang 2023).

The massive datasets the models are trained on are sourced from platforms, such as Wikipedia, Common Crawl, social media, and news websites, although the exact composition of these datasets is often opaque, particularly for proprietary models. Furthermore, most of these data are heavily dominated by the English language, which reflects its availability and accessibility. The predominance of English and Western-centric content in training corpora has led to models that perform disproportionately well on high-resource languages and cultural contexts. In comparison, some models tend to underperform on low-resource or marginalized languages that often lack the large, digitized corpora necessary for effective model training. This issue has prompted growing interest in developing language models for low-resource languages and in creating more balanced datasets that reflect global linguistic diversity (Joshi et al. 2020).

Nevertheless, pre-trained, contextualized language models based on the Transformer architecture have become essential tools for NLP research. One of the first models to apply the Transformer to bidirectional (looking at both the left and right context of a word simultaneously) language understanding was the Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. 2019). BERT's success has not only led to its widespread adoption but also paved the way for the development of numerous other pre-trained language models, such as RoBERTa (Liu et al. 2019), ELECTRA (Clark et al. 2020), and DistilBERT (Sanh et al. 2020), which build upon BERT's architecture while introducing various improvements.

After BERT's introduction, a growing body of research known as BERTology has emerged, mainly analyzing and understanding the internal workings of BERT and its variants, resulting in over 150 studies in less than a year (Rogers et al. 2020). The adaptability of Transformer-based models has facilitated their integration into

multilingual and cross-lingual NLP tasks. For instance, models such as mBERT (Devlin et al. 2019) and XLM-R (Conneau et al., 2020) have extended the capabilities of their monolingual counterparts by being trained on diverse linguistic corpora, allowing them to perform well across multiple languages.

Building on these foundations, more recent developments have shifted toward generative LLMs, such as GPT-4 (OpenAI et al. 2023) and LLaMA (Touvron et al. 2023). These models can produce coherent and contextually appropriate text from natural language input, making them highly versatile across various tasks, including summarization, question answering, and creative writing. An important development in working with these models is prompting, which involves crafting input instructions to modify the model's output. Prompting has become central to interacting with generative models, enabling users to perform complex tasks without the need for additional training or fine-tuning. This thesis, however, does not make use of generative models. This decision and its implications are further discussed in Section 5.3.

Meanwhile, corpus linguistics research has also benefited from LLMs. For instance, in machine translation, deep learning models offer context-sensitive outputs that align more closely with natural language use than earlier rule-based or statistical approaches (Lai 2024; Zhang et al. 2024). These models also facilitate automatic corpus annotation, such as part-of-speech tagging (Bobojonova et al. 2025) and semantic role labeling (Li et al. 2025). In discourse analysis, these models assist in detecting discourse markers (Mishev et al. 2023) and rhetorical structures (Kim et al. 2025) across large datasets. Furthermore, because of their capacity to model language variation over time, they are considered valuable tools for diachronic studies, including the analysis of lexical semantic shifts (Periti & Montanelli 2024).

Deep learning and LLMs have dramatically expanded methodological possibilities for corpus linguistics, thus offering powerful tools for analyzing, annotating, and interpreting linguistic data at scale. However, these complex models make it difficult to interpret which linguistic information is being encoded or why certain predictions are made (Pan et al. 2023). This limitation raises questions about transparency, validity, and reproducibility, which are central concerns in linguistics. Although several techniques have been explored to address this issue, no universally accepted solution has emerged. The following section explores in more detail the various ways to address these concerns by focusing on model explainability.

3.3 Model Explainability

Contemporary language models are often referred to as "black boxes" due to their complex and opaque nature. These models' internal workings are not easily

interpretable, thus highlighting the need for model explainability[1] methods. *Model explainability* refers to the ability to understand and interpret ML model decisions. These explainability methods help researchers uncover the rationale behind a model's decisions ensuring greater transparency and trust. In text classification tasks like register classification model explainability is important for understanding why and how the model makes its predictions. Explainability techniques help to see which words or phrases in the text are important and influence the model's decisions.

Simple models, such as SVM or linear regression, were some of the first interpretable models used in ML. These models are generally easier to interpret and feature easy-to-follow decision-making processes. The models are often considered inherently interpretable, as their structure allows for a global understanding of how input features influence predictions. In the case of linear SVMs, the weight assigned to each feature directly reflects its contribution to the decision boundary, making it possible to trace and explain the model's reasoning, as explained in Section 3.1. For instance, linear SVMs have been used to identify which semantic features contribute to the perceived complexity of online health materials, with lexical difficulty and syntactic complexity linked to more difficult texts, as well as features like personal pronouns or common verbs that are associated with easier ones (Xie et al. 2021).

At the same time, techniques such as Abstract Feature Importance (AFI) offer fast, dataset-independent measures of feature relevance (Pal et al. 2022), while visualization tools like color-based nomograms make model decisions more accessible to non-experts (Van Belle et al. 2016). Because of their transparency, SVMs are considered a valuable tool not only for classification tasks but also for linguistic analysis, wherein understanding which features drive distinctions between registers or genres is important.

However, with the rise of more complex neural network models, several methods for explaining the models' decisions have been developed to shed light on their inner workings. In particular, these models are explained locally with post-hoc explanation methods that inform us how each feature in the data individually affects the model's decisions and the results, rather than understand the whole behavior of the model. To this end, several model-agnostic explanation methods have been developed to explain any ML models, including Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). LIME (Ribeiro et al. 2016) works by creating a simple and interpretable surrogate model that approximates how the complex black box model behaves around a specific data point. This helps in

[1] Although model *interpretability* and *explainability* are sometimes distinguished in the literature, this work uses explainability to emphasize methods that help clarify the behavior of complex language models, without delving into interpretability as a separate concept.

understanding the most important features for a specific prediction and the model's local decision-making process. LIME has been widely used in tasks like medical text classification, sentiment analysis, and fake news detection (Lebeña et al. 2024; Kiefer 2022; Ahmed et al. 2022). SHAP (Lundberg & Lee 2017) uses Shapley values from game theory to optimally assign importance values to each input feature and determines how much each feature affects the model's predictions. SHAP has been applied to visualize how individual words or tokens influence outcomes in various tasks, such as sentiment analysis, spam detection, and document categorization (Zhang et al. 2023; Occhipinti et al. 2022; Zhao et al. 2020). While SHAP provides more robust and consistent results, it can be computationally intensive, whereas LIME is faster but may suffer from instability and sensitivity to sampling choices (Dey & Chowdhury 2024).

Another widely used explainability method, Integrated Gradients (IG; Sundararajan et al. 2017), integrates the model's output gradients with respect to its inputs and assigns contributions to each feature involved in the prediction. Unlike SHAP and LIME, IG is a model-specific technique that mainly works for neural network models by measuring how changes in input features affect the model's prediction, starting from a baseline input and gradually moving to the actual input. Hence, IG is especially useful for deep learning models by helping highlight which words or tokens have the most influence on the final decision (Sundararajan et al. 2017). IG has been used in tasks like sentiment analysis, text classification, and detection of sociopsychological semantic markers (Sanyal & Ren 2021; Ribeiro et al. 2024; Aghababaei et al. 2025).

Building on these explainability techniques, recent approaches like the Stable Attribution Class Explanation method (SACX; Rönnqvist et al. 2022) aim to address the limitations of deep learning models in generalizing across text classification tasks (Laippala et al. 2021; Petrenz & Webber 2011). In particular, SACX enhances interpretability with stable explanations by aggregating IG attributions across documents and models, thereby identifying consistently predictive features. By aggregating attributions across documents, the method identifies the most influential keywords associated with each class. In turn, the process of aggregating across multiple models highlights consistently assigned keywords that demonstrate stability and reliability in classification. The resulting extracted keywords offer valuable insights into the defining linguistic characteristics of each text class (i.e., registers in the context of this thesis). Incorporating methods like SACX into historical register research highlights the critical role of explainability in uncovering model behavior and underlying language patterns. Such integration not only enhances the interpretability of classification outcomes but also supports more transparent analyses of historical language data. Section 4.2.2 presents the practical details of the model explainability methods (IG and SACX) used in the present thesis.

3.4 Deep Learning Language Models in the Study of Historical Texts

While most deep learning language models are pre-trained on modern English web data, the growing availability of digitized historical corpora has opened new possibilities for training models that are more sensitive to historical language use. This shift is facilitated by the creation of several large-scale historical datasets, such as a corpus of historical Swedish newspapers from 1645 to 1926 (Adesam et al. 2019), an English multi-genre corpus covering the period from 1810 to 2009 (Alatrash et al. 2020), and a collection of Korean travel diaries between the sixteenth and the nineteenth centuries (Yang et al. 2023).

The study of historical texts is affected by linguistic challenges (e.g., spelling variation, vocabulary change, and differing orthography) that traditional NLP tools may struggle to handle effectively (Piotrowski 2012). In response, and with the recent availability of large digitized historical corpora in various languages, researchers have begun developing deep learning language models that are specifically tailored to specific languages and time periods. Many of these historically specific LLMs were first trained for English varieties. For instance, Manjavacas and Fonteyn (2021) presented a Transformer-based model that was pre-trained on texts from 1450 to 1950 obtained from numerous historical text collections. Likewise, Rastas et al. (2022) developed a Transformer-based model that was pre-trained on eighteenth-century texts from the ECCO dataset. Other languages have also been of recent interest in the development of Transformer-based language models, including early modern French (Gabay et al. 2022), Dutch from 1500 to 1950 (Manjavacas & Fonteyn 2022b), and German from Old High German (c. 750) to New High German (c. 1999; Beck & Köllner 2023). Additionally, multilingual historical models have been introduced to support cross-linguistic research. For example, Schweter et al. (2022) developed a Transformer-based multilingual language model that was pre-trained on German, English, French, Finnish, and Swedish texts from the seventeenth to the twentieth centuries.

With the rise of generative LLMs, historical language processing has shown both promises and limitations. While these models can handle tasks like part-of-speech tagging in Old Occitan, their performance has been found to be hindered by non-standard orthography and weak cross-lingual transfer learning (Schöffel et al. 2025). While LLMs have outperformed traditional handwriting recognition systems and offer near-human accuracy on eighteenth- and nineteenth-century English texts (Humphries et al. 2024), they struggle with the post-correction of "noisy" historical transcripts of 10 European languages ranging from tenth century Greek to twentieth-century French (Boros et al. 2024). In translation and summarization tasks, LLMs like GPT-4 have produced strong results for Latin and Early New High German (Volk et al. 2024), outperforming earlier approaches. However, safety filters in these

models can unintentionally block sensitive content, thus preventing full translations for languages like Ottoman Turkish (Tekgurler 2025).

Research has shown that models specifically pre-trained on historical data from scratch outperform general-purpose models for applications involving historical texts (Beck & Köllner 2023; Manjavacas & Fonteyn 2022a, 2022b). This is likely due to their ability to internalize the unique linguistic patterns and domain-specific features of historical language use, which are often poorly represented in modern corpora. While modern LLMs such as GPT-4 are powerful and versatile, domain-specific models still seem to have an advantage in specialized historical applications (Chen et al. 2024). However, pre-training models from scratch is computationally expensive and resource-intensive, making it inaccessible for many researchers and institutions. As a result, fine-tuning existing models or using smaller, targeted pre-training strategies remains a practical alternative.

3.5 Studying Registers with Deep Learning Language Models

In quantitative corpus linguistics, Biber's text-linguistic register framework is a widely used methodology for analyzing text varieties. While it is primarily known for its application in the MDA of registers, this framework has also been used for automatic text classification where genre perspective is often used.

Several studies have also used deep learning to identify web texts in different languages. For example, Laippala et al. (2023) explored automatic register classification in English web texts, Skantsi and Laippala (2023) inspected the web registers in Finnish, whereas Repo et al. (2021) studied Swedish and French web registers. Rönnqvist et al. (2021) explored multilingual fine-tuning by combining data from several languages, thereby yielding modest improvements over monolingual training. Their approach also enhanced zero-shot performance when using multilingual models instead of English-only ones. Laippala et al. (2022) extended this work by applying the best-performing multilingual model to eight additional languages, thus confirming its generalizability across diverse linguistic contexts. Furthermore, Henriksson et al. (2024) introduced the multilingual CORE corpora, a dataset comprising over 72,000 documents annotated with a 25-class register taxonomy across 16 languages. By fine-tuning Transformer models in monolingual, multilingual, and zero-shot settings, they demonstrate that multilingual models outperform monolingual ones, especially for languages with limited training data.

Although web text studies are plentiful, deep learning models have been utilized with other kinds of data. For instance, literary genre classification using large collections of full-length texts from Project Gutenberg have been explored with

different deep learning approaches, demonstrating that these models can distinguish between genres such as fiction, poetry, and drama by modeling the structural and sequential characteristics of entire texts (Goyal & Prem Prakash 2022). Worsham and Kalita (2018) introduced the Gutenberg Dataset for genre identification and emphasized the importance of modeling the hierarchical and compositional nature of long literary documents.

The abovementioned studies exemplify the strengths of deep learning approaches in register classification, particularly in multilingual and low-resource settings. They also demonstrate how models can generalize across languages, enabling zero-shot and multilingual classification without requiring extensive manual annotation in each target language. This scalability and adaptability mark an advantage over traditional linguistic methods, which typically require language-specific feature engineering and expert-driven annotation. However, linguistic approaches offer certain degree of interpretability and theoretical grounding that ML learning models often lack. While deep learning models can classify registers with high accuracy, they do so without explicitly modeling communicative purpose or linguistic function. This limits their usefulness for explanatory or theory-driven research, in which understanding why a text belongs to a particular register is as important as performing the classification itself. Additionally, deep learning models often struggle with extremely long texts, given that maintaining coherence and capturing long-range dependencies across hundreds of thousands of words remains a significant challenge for many of the model architectures. Moreover, predictive modeling offers benefits like allowing for the comparison of model accuracies and performance across corpora (Shmueli 2010). The current thesis reflects such distinctions and considers how predictive and explanatory approaches can complement each other in register analysis.

4 Materials and Methodology

This chapter outlines the empirical foundation and analytical procedures of this study. It begins by introducing the two primary corpora used, highlighting their composition, historical scope, and relevance to the study's RQs. Particular attention is given to the challenges posed by historical data, including issues of digitization, orthographic variation, and metadata inconsistencies. A brief overview of previous linguistic research conducted on these corpora situates the present study within the broader scholarly context. The second part of the chapter details the methodological framework, including data preparation steps, feature extraction, and the application of ML models for classification and interpretation. Finally, the chapter provides an overview of the three peer-reviewed studies comprising the core of the dissertation.

4.1 Material

This section provides a detailed overview of the materials used in this thesis. The data came from two different late modern English corpora – COFEA and ECCO. Chapters 4.1.1 and 4.1.2 introduce the two corpora in more detail, respectively, whereas Chapter 4.1.3 discusses some challenges related to the data. Chapter 4.1.4 introduces previous studies that have utilized COFEA and ECCO.

4.1.1 COFEA

COFEA is a relatively recently created corpus that encompasses written works from 1760 to 1799, that come from a variety of sources and registers. The corpus, which aims to represent American English from the Founding Era, has been constructed while considering corpus-based legal interpretation of documents from the Founding Era (Hashimoto 2023). In particular, the original meaning of the US Constitution is of interest for the legal community, and COFEA has been used in these studies for quantitative and generalizable evidence for word meaning (Hashimoto 2023). These and other studies using COFEA are discussed in more detail in Chapter 4.1.4.

COFEA includes over 127,000 texts, containing a total of over 138 million words. To accurately reflect Founding Era American English for legal interpretation and to ensure comprehensive coverage, researchers recognized the need for a large

and register-diverse corpus. The collection process of the texts was assisted by subject area experts who aided in finding databases with Founding Era texts of various registers (Hashimoto 2023). Currently COFEA includes texts from Evans Early Imprint Series (Evans), the National Archives Founders Papers Online project (Founders), Hein Online's U.S. Treaties and Agreements Library, U.S. Congressional Documents, American Indian Law Collection, and Session Laws Library (Hein), Farrand's Records of the Federal Convention of 1787 (Farrand's), US Statutes at Large, Elliot's Debates (Elliot's), and Harvard Law School's Caselaw Access Project (CAP). Evans includes a variety of documents from both ordinary Americans and the elites, whereas the Founders database houses the correspondence, papers, and personal writings of the Founding Fathers. Hein, US Statutes at Large, and CAP all feature various legislative documents. Farrand's and Elliot's collections include texts documenting the creation of the US Constitution. These databases were selected for their diverse language use, including legal and nonlegal materials, and texts ranging from highly literate to highly oral, thus ensuring a representative and generalizable sample for studying language from the Founding era (Hashimoto 2023).

Although the texts in COFEA were collected from databases with a diverse range of texts and registers, the texts within these databases have not been specifically annotated to indicate their respective registers. Thus, the texts were manually annotated by trained coders. As described in Hashimoto (2023), the register annotation process consisted of three comprehensive steps: 1) generating and refining register categories, 2) conducting manual annotation, and 3) performing accuracy verification. The annotations were double coded, and a third coder was involved to resolve any disagreements between the coders. The full list of register categories can be found in Table 2, and the codebook for training human coders on these categories is available at Hashimoto (n.d.). The data includes personal texts (e.g., letters and diary entries), less personal texts (e.g., books and pamphlets), and more formal legal texts (e.g., treaties and legal cases). There are also two technical categories, namely, Not English, which includes texts written in languages other than English (primarily French), and No Text, which includes empty entries or those containing only metadata from the original source database. Additionally, a separate category called Undetermined covers texts that do not align with the other registers, such as poems, epitaphs, and songs. At the time of the publication of the studies in this thesis, the manual annotation of registers in COFEA was an ongoing process. The data used in Studies I, II, and III include a subset of COFEA texts. A detailed description of this subset is provided in Chapter 4.2.1.

Table 2. Register categories in COFEA.

Book	Law and Statute	Proceedings	Narrative Fiction Book
Cases	Letter	Proposal	Narrative Non-fiction Book
Contract	Newspaper	Speech	Non-narrative Non-fiction Book
Debate	Pamphlet	Treaty	Not English
Essay	Personal Record	Undetermined	No Text

The registers in COFEA are broadly similar to those found in other large historical corpora from the Founding Era, such as ARCHER, the Corpus of Early American Literature (CEAL; Höglund & Syrjänen 2016), and the Corpus of Late Modern British and American English Prose (COLMOBAENG; Fanego 2012). For example, ARCHER features journals/diaries, legal opinions, scientific prose, personal letters, and news reportage, while CEAL includes political writings, religious texts, fiction, and biographies. COLMOBAENG, though more limited in scope, focuses on prose like philosophical and academic writing, and on fictional and historical prose.

Most of the texts in COFEA were already in machine-readable format when collected and were digitized using various optical character recognition (OCR) technologies, but texts from Founders and Evans were manually transcribed (Hashimoto 2023). The digitization processes introduced various OCR errors to the texts due to the age and diverse format and quality of the documents, leading the corpus creators to use custom Python scripts to automatically correct some recurring errors (Hashimoto 2023). Despite the correction, various OCR errors remain in the texts. The potential issues of OCR errors are discussed further in Chapter 4.1.4.

4.1.2 ECCO

ECCO houses over 200,000 texts printed between 1701 and 1800, making it the largest online repository of works from this period (Gregg 2020). ECCO is a database with a rich and varied history that affects its composition. ECCO was first published in 2003 as a digitized collection of facsimiles, but it originated from a microfilm collection produced between 1982 and early 2000s, whose contents, in turn, were selected based on a catalogue of books called the Eighteenth-Century Short Title Catalogue from the 1970s (eventually renamed the English Short Title Catalogue or ESTC; Gregg 2020). ECCO was made by scanning these microfilm reels to black and white digital images of the book pages and then using OCR engines to automatically produce text transcripts of each page. The quality of the transcripts varies greatly due to different scanning methods and programs used at different times. The problems caused by this process will be further discussed in Chapter 4.1.3.

While the ESTC has catalogued works from two thousand public and private libraries from all over the world, the microfilming series is based on only around twenty libraries, the most notable of which is the British Library, with approximately 67% of all books in ECCO and 83% of Part I being reproductions sourced from the British Library's holdings (Holahan 2021; Spedding 2011). Other notable source libraries include the Bodleian, Harvard, Cambridge, Huntington, and the National Library of Scotland.

ECCO is claimed to include "every significant English and foreign language title printed in the United Kingdom during the eighteenth century, along with thousands of important works from the Americas" (Gale 2016). However, ECCO is a collection that continues to grow. Following the initial release of ECCO with over 135,000 texts in 2003, Part II of ECCO was released in 2009, which added some 50,000 titles to the collection. At present, Part III, with an additional 90,000 titles, is also planned to be launched in the future. As a whole, the collection has been evaluated to contain over 50% of the titles included in the ESTC, which is the most comprehensible metadata source for British publications from the early modern era (Tolonen et al. 2021).

ECCO includes a wide range of printed materials, from brief pamphlets to complete books, covering diverse fields, such as religion, philosophy, law, fine arts, literature, and language. The database, however, lacks internal order and metadata on the texts, making it rather difficult to query. The texts in ECCO are not annotated with register information but are organized into eight subject categories derived from the original categorization of the microfilm reels (Gregg 2020). These eight categories are 1) Religion and Philosophy; 2) Literature and Language; 3) History and Geography; 4) Fine Arts and Antiquities; 5) Social Science; 6) Science, Technology, and Medicine; 7) Law (Criminal, National, and International); and 8) General Reference and Miscellaneous. According to Holahan (2021) the top 3 categories in ECCO are Literature and Language with 28% of the collection, Religion and Philosophy (25%), and Social Sciences (18%).

For more information on the development and composition of ECCO, see especially Gregg (2020), Holahan (2021), and Tolonen et al. (2021; 2022). Studies II and III use texts from the ECCO dataset. Chapter 4.2.1 provides a detailed description of this subset used in the two studies.

4.1.3 Challenges in Data

COFEA and ECCO are not without their limitations. For instance, both are heavily biased toward the language of white, educated, and literate men. Many historical corpora and databases are skewed towards including texts from dominant sociocultural groups (Gregg 2020; Hauswedell et al. 2020). This skew often results

in the underrepresentation of marginalized communities, including women, people of color, and those from lower socioeconomic backgrounds (Earhart 2012). COFEA aims to represent American English from the Founding Era, but it is especially skewed towards texts from individuals involved in the drafting of the US Constitution due to their language being well preserved and often deemed more crucial for Constitutional interpretation (Hashimoto 2023).

ECCO, with its layered history, was not designed to be a balanced and carefully selected collection of texts, let alone a text corpus. Consequently, there are many skews in the data (e.g., Tolonen et al. 2021, 2022). For instance, ECCO contains mainly texts printed in England, and, to a lesser extent in Ireland and Scotland. Additionally, ECCO shows a preference for the late eighteenth-century texts, with a noticeable increase in publications during the final two decades. Furthermore, exact dates for some documents are not certain, resulting in estimates peaking at even five and ten years.

With regards to the present thesis, OCR errors are an important factor to consider. Both COFEA and ECCO contain, to varying degrees, OCR errors resulting from different causes. For instance, the quality of British printing in the eighteenth century was often quite poor, leading to transcription errors from broken, poorly inked, or warped lines of type (Wendorf 2022). Additionally, eighteenth-century texts include typographical features, such as the long "s" (⟨ſ⟩), ligatures, and ornate capital letters, that have caused challenges for the OCR software of the time. In long-term scanning processes, such as for ECCO, the tools and OCR software used have changed, leading to varying levels of noise in different parts of the database. Tolonen et al. (2021) note that OCR errors are especially notable in Part I of ECCO, sometimes to the degree that the digitized text is even incomprehensible to humans. Various projects have made an effort to increase the accuracy of the texts in ECCO that have undergone OCR. For instance, notable efforts include the ongoing Eighteenth Century Collections Online Text Creation Partnership (ECCO-TCP), which has manually produced transcriptions for some 3,000 texts in ECCO. Additionally, the Early Modern OCR Project (eMOP) has developed new algorithms and software to improve OCR for early modern texts, and a crowd-sourced manual OCR correction tool TypeWright by 18thConnect has been created to mitigate OCR problems. COFEA is also manually being rated for OCR accuracy so that low accuracy texts can be targeted for manual correction (Hashimoto 2023). The automatic post-correction of OCR errors in historical texts with LLMs has been an interest in recent research, but the results vary (Kanerva et al. 2025). This topic will be revisited in Chapter 5.3 when discussing the limitations of this thesis.

Despite the availability of ECCO-TCP, which offers manually corrected transcriptions, its coverage is limited in terms of the number of texts, and they are all from ECCO Part I. To ensure a sufficiently large and diverse dataset for robust

analysis, the present study opted to use the un-corrected texts from the whole ECCO dataset.

4.1.4 Previous Research on ECCO and COFEA

This section provides a brief overview of previous research on ECCO and COFEA. ECCO has been the focus of various studies that have investigated the composition of the database itself or used it as data for various linguistic research questions. Meanwhile, COFEA is a relatively new corpus that has been used in studies conducting legal research rather than linguistic research.

ECCO is valued for its massive scale and representation of eighteenth-century anglophone print culture and has attracted significant research attention. Several studies have focused on the composition and nature of ECCO. For instance, Tolonen et al. (2021; 2022) provide a foundational overview of the corpus, examining the diversity and imbalance of genres, publication metadata, and reprint practices. Their analyses reveal that a substantial portion of ECCO consists of reprints, thus leading to duplication that skews corpus composition. The overrepresentation of certain texts and genres, such as petitions and bills, and the exclusion of genres like almanacs and proclamations also contribute to an uneven genre balance. Their works emphasize the need for corpus criticism and awareness, especially given ECCO's reliance on OCR-processed texts and limited metadata (see Holahan 2021; Spedding 2011; Gregg 2020).

At the same time, more linguistically targeted studies have explored lexical and stylistic variation. Tiihonen et al. (2023), for instance, applied regression modeling to trace Scotticisms across Scottish and English publications in ECCO, revealing insights into regional linguistic identity during the Enlightenment period. Textual reuse has been another major focus, with Rosson et al. (2023) and Ryan et al. (2023) developing sequence-alignment and intertextuality tools to map shared language across texts. Their works enhance our understanding of discourse circulation and literary influence in early print culture. Additionally, Regan (2022) studied semantic change and knowledge transmission by analyzing diachronic shifts in the distribution of some words in ECCO, highlighting how evolving lexical associations reflect broader intellectual trends in eighteenth-century Britain.

ECCO's uneven OCR quality and inconsistent metadata remain critical concerns. For instance, Hill and Hengchen (2019) compared linguistic metrics in raw OCR data and it's cleaned counterpart in ECCO, demonstrating a significant impact on the quantitative analysis of historical documents. Nevertheless, despite problems with OCR, recent studies have demonstrated that many quantitative methods that are commonly used in corpus linguistics and digital humanities remain reasonably accurate when applied to the database (Hill & Hengchen 2019; Liimatta et al. 2023).

In contrast to ECCO's broader scope, COFEA provides a more controlled and balanced view of American English in the late eighteenth century. The corpus is designed to support inquiries into legal language and civic discourse, and COFEA has been used especially in the legal community to study the original meaning of the Constitution and other statutory laws in the Founding Era. For instance, Cunningham and Egbert (2020) studied the original meaning of the term "emolument" by examining over 2,800 instances of the word in COFEA using multiple computational methods to determine its typical usage during the Founding Era. Miller and Obelgoner (2020) tracked the original meaning of the phrase "executive power" as used in the Constitution. Using COFEA, they analyzed linguistic patterns and historical discourse surrounding the term. Lee and Phillips (2019) applied corpus linguistic methods to investigate the phrase "natural born citizen" using COFEA to analyze collocations, frequency data, and contextual usage, thereby determining whether the phrase was treated as a legal term of art or an ordinary expression.

The influence of this corpus extends beyond academia. COFEA has been referenced in two cases before the United States Supreme Court (Carpenter v. United States 2018; Lucia v. SEC 2018) and two circuit court cases (District of Columbia v. Trump 2019; Wright v. Spaulding 2019). These references underscore COFEA's growing role in judicial reasoning and its credibility as a source for originalist interpretation. For a more thorough list of publications using COFEA, see Hashimoto (2023).

4.2 Methodology

This study employs a multi-faceted methodological framework that integrates corpus linguistics with NLP techniques to investigate register variation in late modern English texts. The approach is grounded in the use of domain-specific language models and text classification, complemented by model interpretability techniques and keyword extraction procedures. The methodology is designed to ensure linguistic interpretability, aligning with best practices in both corpus-based and computational linguistic research.

Historical corpora such ECCO are characterized by their heterogeneity, lack of consistent metadata, and linguistic variation across time, registers, and communicative purposes, as mentioned earlier. Thus, traditional rule-based approaches, and those requiring extensive manual annotation are insufficient for capturing the nuanced, often implicit register distinctions in such data. In this regard, ML offers a scalable, inductive framework capable of learning patterns directly from the data, without requiring exhaustive manual annotation or rigid a priori categorization.

SVMs were employed in the early stages of this research to establish interpretable and robust baselines for register identification. As mentioned in Chapter 3.1, SVMs are well-suited for high-dimensional linguistic data and have demonstrated strong performance in text classification tasks. In addition, their ability to provide clear decision boundaries and feature weights made them particularly valuable for initial experiments during which interpretability and diagnostic insights were prioritized.

Register distinctions often hinge on contextual cues, discourse structure, and subtle shifts in communicative function that are not easily captured by pre-neural ML models and rule-based corpus linguistic approaches. To address this, I used Transformer-based models, especially fine-tuned variants of BERT, which are designed to capture the linguistic contexts of words. Register is inherently a multi-dimensional construct, influenced by lexico-grammatical choices, discourse structure, and communicative intent. Transformer-based models that are capable of learning from raw text are uniquely positioned to uncover linguistic patterns that correspond to register distinctions.

4.2.1 Data Preparation

For the experiments in Studies I–III, I prepared a sub-corpus of COFEA that includes over 110,000 texts that were hand-coded for various registers introduced in Chapter 4.1.1. Before the classification, the dataset was pre-processed by removing punctuation, duplicate entries, and empty texts. Some registers with a very small number of texts, that is non-narrative non-fiction books and narrative non-fiction books, were excluded from the analysis, along with those in the Undetermined category due to ambiguous labeling.

For the SVM classification in Study I, the COFEA sub-corpus was grammatically tagged with the Northern Arizona University (NAU) Tagger. The grammatical tags generated by the NAU Tagger are based on Biber's (1988) work, which identified specific grammatical features deemed important for registerial variation. The tagger provides detailed grammatical information, including part-of-speech labels, person, number, and semantic domain.

For the register classification of ECCO, two coders manually annotated 300 texts across 15 different register categories. This annotation process was iterative and developed from the bottom up. However, the register categories and guidelines from COFEA were referenced when applicable. The coders independently annotated the texts, and their work was cross verified to ensure consistency. By comparing these annotations, we can measure how much the annotators agree with each other, which is known as inter-annotator agreement (IAA). Disagreements between annotators can be solved by establishing a consolidated consensus set of annotations. In this case,

measuring alignment between each annotator and the consensus set can be done using metrics, such as accuracy, which represents the proportion of similarly annotated examples out of all examples. For measuring agreement between two annotators without a resolved consensus, a Cohen's Kappa (Cohen 1960) measure is often used, as it also considers agreement by chance. Before unifying the annotations, we measured the IAA using Cohen's Kappa coefficient, resulting in a score of 0.67, which indicates a fairly good level of agreement. As the ECCO texts were drawn from published volumes, they often contain paratextual elements, such as title pages and forewords, that can interfere with register classification. To ensure that the model focused on the main body of the text, all paratextual materials preceding the core content were removed from the 300-text subsample during preprocessing.

Figure 1 presents the proportion of hand-coded registers in the ECCO and COFEA subsamples. As shown in Figure 1, the register category letter is the most common register in COFEA, comprising over 80% of the dataset, whereas treatise is the most common register category in the ECCO subsample. While ECCO and COFEA largely share the same register categories, the texts within these categories can differ significantly across the datasets. These differences may stem not only from dataset composition, but also from the situational characteristics of the registers and their production circumstances, as well as the inherent complexity of manual annotation, including inconsistencies among annotators and annotation procedures.

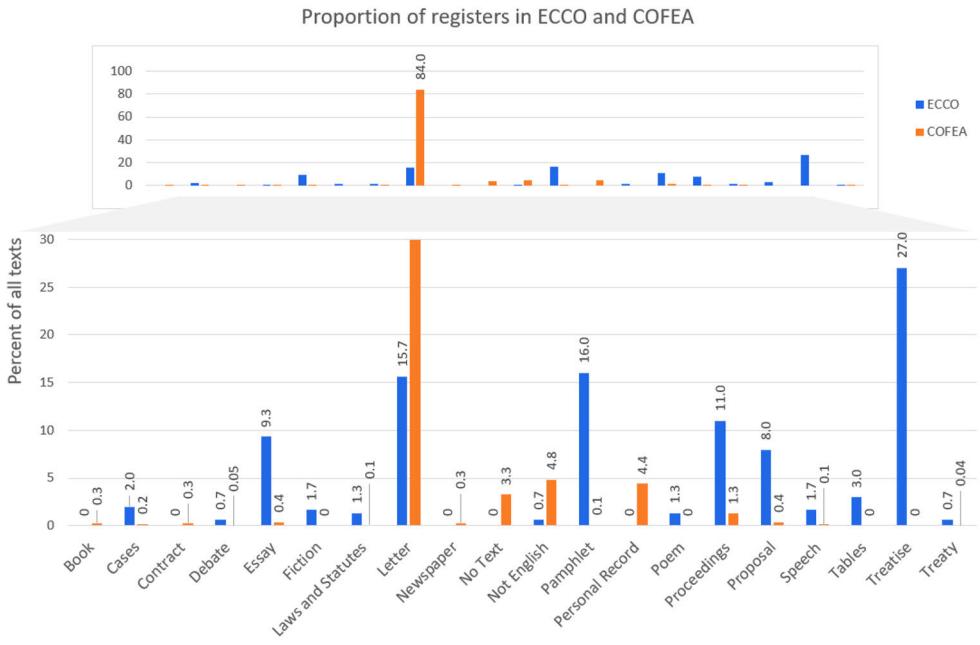

Figure 1. Proportion of manually annotated registers in the ECCO and COFEA subsamples. Adapted from Repo et al. (2024).

For the keyword analysis in Study III, the ECCO subsample was further expanded by extracting unannotated texts and automatically assigning them a register with a model that was fine-tuned with COFEA texts. To support register modeling and analysis of different text sections, each document in ECCO and COFEA datasets was divided into five equal-length segments: beginning, beginning-middle, middle, middle-end, and end. This procedure is further detailed in Study III.

4.2.2 Model Training and Interpretation of Model Decisions

Given the complexity of linguistic variation across registers and the historical nature of the data, Transformer-based language models were selected for their good performance in capturing contextual and syntactic nuances. In particular, we used the Transformer-based language model BERT (Devlin et al. 2019) due to its well-documented effectiveness in recognizing contemporary language registers (Laippala et al. 2022). In addition to BERT, we used another deep learning language model, XLM-RoBERTa (XLM-R; Conneau et al. 2020), which is a multilingual language model based on the Cross-lingual Language Model (Conneau & Lample 2018) and the RoBERTa model (Liu et al. 2019). XLM-R has been shown to outperform monolingual baselines that rely on pre-trained models and other multilingual models (Conneau et al. 2020; Libovický et al. 2019; Tanase et al. 2020). However, the core classification model used in the present study is ECCO-BERT (Rastas et al. 2022), a Transformer-based language model pre-trained on eighteenth century English texts from the ECCO database. ECCO-BERT shares the architecture of the BERT base (cased) model, with its main distinction being an adjusted vocabulary size specifically designed to align with the historical language used in the ECCO dataset. ECCO-BERT's domain-specific training makes it especially effective for our analysis, which also relies on the ECCO database.

All models were trained using PyTorch (Paszke et al. 2017) and the Hugging Face Transformers library (Wolf et al. 2020). Training was conducted on high-performance computing clusters equipped with NVIDIA V100 GPUs, thus allowing for efficient fine-tuning of large models on substantial historical datasets. This process was carried out using the supercomputing resources provided by CSC – IT Center for Science. Hyperparameters such as learning rate, batch size, and number of epochs were optimized using grid search and early stopping based on validation loss[2]. To enhance

[2] *Learning rate* controls the model's weights during training; *batch size* is the number of samples processed before the model's parameters are updated; number of *epochs* refers to how many times the entire dataset is used to train the model; *early stopping* halts training when performance on a validation set stops improving; and *validation loss* measures model error on unseen data during training.

model performance and generalizability, Optuna (Akiba et al. 2019), a Bayesian hyperparameter optimization framework, was also employed. The specific hyperparameter values used in the papers are described in more detail in Studies I–III. This automated tuning ensured that each model was trained under optimal conditions, thus reducing the risk of overfitting and improving classification accuracy.

Apart from the deep learning models, we also utilized a linear SVM to classify lexical and grammatical features in the COFEA data. The SVM was implemented in Python's Scikit-Learn library[3] (Pedregosa et al. 2011). We used an SVM model as it provides a way to analyze the importance of linguistic features in distinguishing between different classes. We used lexical features consisting of words as they appeared in the texts and grammatical tags associated with each word in our classification. The most effective results in register identification have been achieved using word-based features, such as bag-of-words and word trigrams (Pritsos & Stamatatos 2018), as well as binary character four-grams (Sharoff et al. 2010). In contrast, grammatical tags do not convey topical content but instead reflect the genre or register of a text (Finn & Kushmerick 2006) and are believed to have functional associations specific to particular registers (Biber 1988; Biber & Conrad 2019). Grammatical features are widely acknowledged as fundamental to register analysis and form the foundation of register studies. Laippala et al. (2021) demonstrate that incorporating grammatical information enhances classifier stability and that the influence of grammar varies across different registers. The grammatical tagging process and the technical implementation of the SCV classifier are introduced in more detail in Study I.

To further gain insights into the linguistic features considered important in distinguishing the registers in ECCO and COFEA, we employed two model explanation methods, IG and SACX. We applied the IG interpretability method (Sundararajan et al. 2017) to compute feature attributions, thus revealing the contribution of each feature to the model's prediction. This approach helped uncover the linguistic features that differentiate the registers from one another. Additionally, we used SACX for keyword analysis that also allowed us to identify the most important words. The keyword analysis procedure is detailed in Study III.

4.3 Overview of the Publications

The next three sections present the studies included in this thesis in more detail. In Study I, we laid the groundwork for automatic register classification in COFEA,

[3] http://scikit-learn.org/

demonstrating that registers could be effectively modeled using linguistically interpretable features. Building upon the ideas of Study I, Study II extends the approach to unannotated corpora by applying COFEA-trained models to ECCO texts, thereby introducing model interpretability through IG to assess the linguistic plausibility of predictions and exploring the feasibility of cross-corpus register classification. Study III further advances this line of inquiry by addressing the challenge of document-internal variation in long texts, evaluating how different text segments affect model performance and using the SACX method to identify linguistically meaningful features across document sections. Collectively, these studies contribute to the creation of a scalable, linguistically grounded framework for enriching historical text databases with register metadata, supporting both computational and humanities-based research.

I served as the first author and lead contributor in all the papers. In the following subsections, I briefly describe the contents of the papers and the division of labor in the collaborative work conducted for this dissertation project.

4.3.1 Study I

In Search of Founding Era Registers: Automatic Modeling of Registers from the Corpus of Founding Era American English (Repo et al. 2023)

This article models register variation in COFEA and explores the possibility of automatically identifying registers from a corpus of late modern English texts. The study is situated within the broader context of historical linguistics and corpus-based research, in which one of the persistent limitations is the absence of register annotations in historical corpora. The central RQ of this study is whether modern ML techniques can be effectively applied to model registers in historical English texts. Additionally, we aimed to inspect which linguistic features are most indicative of different registers in eighteenth-century American English. These questions are motivated by the need to make historical corpora more accessible, especially for researchers in the fields of linguistics, legal history, and digital humanities.

To address these questions, we used a computational approach that combines supervised ML with linguistic feature analysis. For our training and evaluation dataset, we selected a subset of COFEA that includes texts manually labeled with register information. Several classification models were tested, including traditional ML algorithms, such as SVMs, as well as deep-learning models, namely, BERT and XLM-R. For the SVM, we experimented with three different feature sets – lexical features, grammatical features, and a combination of the two. Additionally, we extracted and analyzed the most important discriminative features estimated by the SVM classifier for the registers. This approach allowed us to learn more about the

linguistic characteristics of the registers in COFEA and the functioning of the classifier.

The results of the study are very promising. The best performing model, BERT, achieved a micro F1-score of 97%, representing a level of performance that is noteworthy given the challenges posed by historical language data, which often include archaic vocabulary, non-standard spelling, and syntactic variation. However, some of the register classes that appeared less frequently in the data were not as well identified by the model. This difficulty in identifying some of the registers could indicate a potential issue about the uniformity of the texts in registers.

In the detailed analysis of the linguistic features contributing to the classification, we found that letters in COFEA are characterized by many verbal features, such as possibility, auxiliary, and copular verbs as well as features related to past tense. Additionally, past participle attributive and predicative adjectives, plural common nouns, proper nouns, indefinite pronouns, and third-person singular pronouns also characterized the letters. Furthermore, letters in COFEA seem to be more descriptive and argumentative in style than other letters in the seventeenth century, likely due to their professional nature. In earlier studies, letters have been linked to interpersonal and interactive communication featuring, for example, personal involvement and stance taking (Biber & Finegan 1989). Interpersonal communication is well characterized by formulaic opening and leavetaking expressions deemed typical of the early and late modern letters, much like modern letters (Elsweiler & Ronan 2023). In earlier multidimensional studies, letters have also been found to have expository, descriptive, and argumentative features (Biber 2001; Monaco 2017).

We also found that many of the discriminative features in the journals were nominal, such as plural and singular common nouns, temporal nouns, predicative and attributive adjectives, and determiners. These features are commonly found in narrative texts that recount or reflect on events and actions. The presence of complex noun phrase structures, many types of conjuncts, and agentless passives indicate that journals are characterized by high informational density, which is in line with the findings of previous studies (Biber 2001).

Methodologically, Study I demonstrates that automatic register classification is not only feasible but also effective, even for historical texts. The ability to identify registers based on linguistically motivated features brings new linguistic information about registers and enhances the value of corpora like COFEA for a wide range of research applications. This makes it possible for legal historians, for instance, to more easily isolate legal texts from personal correspondence, while linguists can explore variation in language use across different social and communicative contexts.

The design and the article's main idea was a joint effort between all the authors. I was responsible for data preprocessing, model implementation, and evaluation. Furthermore, I ran all the experiments and conducted the linguistic feature analysis.

All authors contributed to the interpretation of the feature analysis. Finally, I wrote the first draft, and all the authors revised the subsequent versions.

4.3.2 Study II

Toward Automatic Register Classification in Unrestricted Databases of Historical English (Repo et al. 2024)

Study II extends the register annotations to the ECCO corpus. The study addresses the complex challenge of identifying registers in large, unannotated collections of late modern English texts. Unlike Study I, which focused on curated corpora with clearly labeled registers, Study II explores how register classification can be achieved in a more heterogeneous and unrestricted dataset. Because ECCO does not have existing register annotations, we explored the possibility of predicting registers for unannotated, lengthy historical texts using COFEA registers for the automatic annotation of ECCO registers. The central RQs revolve around whether automatic methods can reliably identify registers in historical English texts without prior annotation, which linguistic features are most effective for distinguishing these registers, and how such methods can be scaled across different corpora. This is particularly relevant when considering how similar the registers are across different corpora. The paper argues that variation in registers may arise not only from differences in register definitions or corpus composition, but also from historical shifts in language use, annotation practices, and the communicative functions of texts within each dataset.

We hand-coded a set of 300 texts in ECCO for their register and tested how well a model trained with COFEA registers can predict registers for ECCO texts. We used the model that was found to be best in predicting COFEA registers in Study I. Additionally, we used ECCO-BERT, which has been specifically pre-trained on ECCO texts. We also used the IG model interpretability method to analyze which features the model relies on when determining a text's register, and whether it focuses on similar features across different datasets. Furthermore, we inspected the shared situational features between misclassified and correct registers to better understand the sources of confusion in the model's predictions.

The results reveal that models trained on COFEA can generalize to ECCO to a meaningful extent. Some of the register classes in ECCO, such as letters and cases, were well identified. This result indicates that texts within these registers are well-defined and share notable similarities with their corresponding categories in COFEA. However, other registers were not identified at all due to various reasons, such as categories not shared by both corpora, differences in the registerial features compared to the corresponding register in COFEA, or a prior register distribution the model had learned from COFEA. The results highlighted the challenges posed by

hybrid texts, shared situational characteristics. Historical texts also present additional challenges due to non-standardized spelling and a wide range of OCR-related errors originating from the diversity of the source materials (Baron & Rayson 2008; Hill & Hengchen 2019). The OCR errors and the challenges they pose are further discussed in Chapter 5.3.

Many of the misclassifications in the data appeared to be linguistically and functionally motivated, reflecting the overlapping characteristics commonly associated with different registers. For instance, we found that the register classes essay and proceeding have frequent misclassifications to personal records due to the shared typical features among the classes. At the same time, some registers, such as proposals, were frequently misclassified as essays. From a linguistic standpoint, many of the misclassified proposals contain evaluative language and stance-taking features that are typical of essay texts in COFEA. Additionally, some texts shared clear characteristics of many registers. This issue is particularly evident in the case of proposal texts, which were frequently misclassified as letters. The confusion stems from the fact that many proposals were written in letter format, causing them to exhibit linguistic features typical of both registers. The presence of shared features across registers contributed to the classifier's confusion, thus highlighting the inherent complexity of register identification.

Automatic register annotation requires a corpus that fully represents ECCO's register diversity. However, building such a resource is hindered by limited knowledge of ECCO's registers and the intensive manual effort required by annotation. Thus, this study explored whether existing register annotations can serve as a starting point to overcome this challenge and opened the door to automated annotation of historical corpora.

The idea of the paper was a joint effort between V. Laippala and myself, while further idea development was done jointly by all authors. Data annotation was jointly completed by L. Saario and myself. I conducted all the experiments and wrote the first draft. All authors contributed to the finalization of the paper.

4.3.3 Study III

From Unannotated to Annotated: The Impact of Text Internal Variation on Register Predictions of Long Historical Documents (Repo et al., forthcoming)

In this paper, we further investigated the reliability and linguistic motivation of automatic register classification in long, unannotated historical texts in ECCO. We studied how well a model trained on annotated data from COFEA can generalize to ECCO texts, particularly in capturing consistent register features across different sections of a document. While register prediction has advanced to the point of being

used for metadata generation, its effectiveness in handling internal variation within lengthy documents remains uncertain, particularly given the linguistic complexity and contextual shifts that such variation entails. Thus, we shifted the focus from inter-document classification to intra-document variation. The central RQs focused on how internal textual variation affected register predictions and whether the predictions reflected pervasive, register-specific linguistic characteristics.

We analyzed model performance across different text segments, such as beginnings and middle sections, by fine-tuning BERT-based models with COFEA data, after which we predicted registers for the ECCO documents annotated in Study II. Dividing each document into five equal-length segments allowed us to assess which parts of a document are most effectively predicted and how internal variation influences classification accuracy. To expand the dataset for more robust analysis, we applied the best-performing model to unannotated ECCO texts, selecting those with consistent predictions across segments. We then employed the SACX method to extract and rank keywords that contributed to the model's decisions. Finally, we fine-tuned and tested multiple models across all text segments, resulting in 36 model-data combinations.

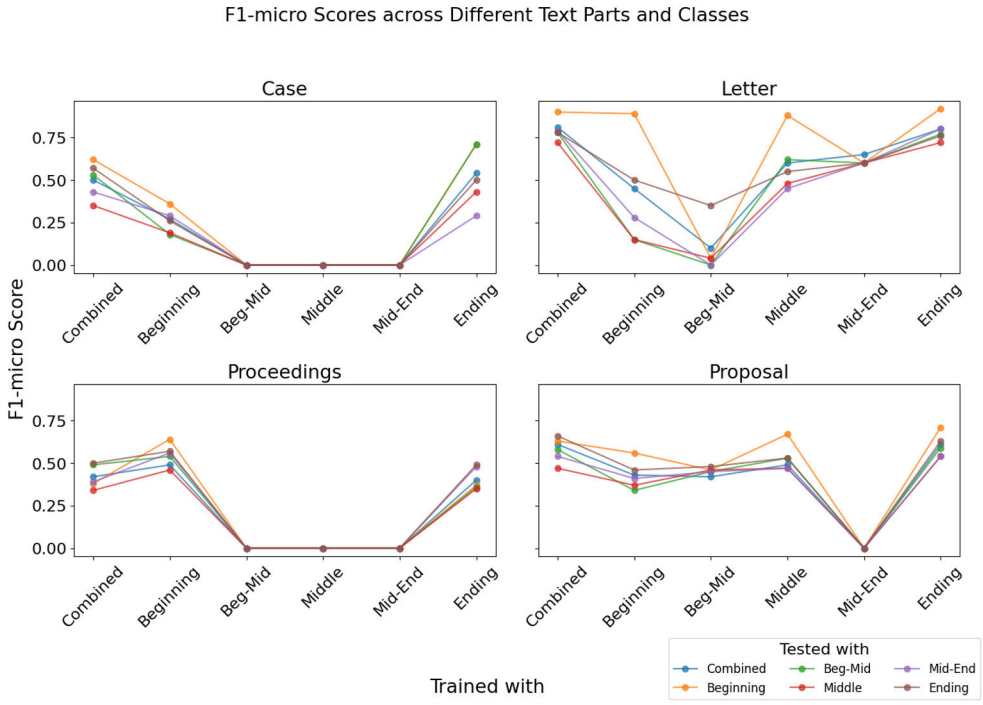

Figure 2. Register-specific micro-averaged F1 scores across various training and testing configurations. Each data point corresponds to a combination of training and testing datasets. (Repo et al., forthcoming).

Our results showed that text beginnings consistently yielded the most reliable register predictions. Notably, a model trained on the ending segments and tested on beginnings achieved the best results. In comparison, models tested on the middle sections of texts performed significantly worse, suggesting that these segments contained more linguistic variability or less distinctive register features, making classification more difficult. Additionally, models trained on all five text segments or on endings demonstrated the most consistent and balanced performance across all registers and text parts. Figure 2 shows the register-specific results of various training and testing configurations for four register classes. The results align with earlier research on web data, which concluded that the beginnings of texts tend to exhibit the greatest predictive power (Laippala et al. 2023).

The keyword analysis further confirmed that the models captured pervasive, linguistically meaningful features consistent across text parts, supporting the notion regarding the reliability of register predictions in complex historical documents. Furthermore, the features present across the entire register class of letters illustrate a balance between formality, politeness, humility, and emotional expression. These keywords frequently include personal pronouns, attributive adjectives, and common nouns. The prominence of attributive adjectives, public verbs, and nouns also reflects characteristics of informational language, whereas the use of personal pronouns, the verb "be" as a main verb, and mental verbs underscores the involved and interpersonal nature of letters. These patterns align with Biber's (1988) framework on involved and informational language production. Previous research has also identified involved features in Early Modern letters (Biber & Finegan 1989), while letters from the eighteenth and nineteenth centuries have been found to exhibit less involvement (Biber & Finegan 1989; Biber 2001). Informational traits, in turn, have been linked to the role of letters in spreading scientific knowledge and their didactic function during the late modern period (Monaco 2017).

The stable keywords in the case register class are dominated by legal terminologies, thereby functioning as markers of specialized discourse within legal contexts. Historical legislative language is characterized by precision and the use of formulaic expressions to facilitate comprehension and interpretation (Lehto 2010). Most of the keywords occurring across all sections are common nouns referring to the various roles of participants in legal cases. A typical feature of Modern English legal documents is the practice of beginning each item or sentence by identifying the legal actor and their action, followed by more detailed qualifications (Lehto 2010). Likewise, the stable keywords in the proceeding register are dominated by legal terminologies, thus reflecting the formal and procedural character of these texts. Many of the nouns refer to various bodies, positions, and the outcomes of meetings. These findings highlight the comprehensive training data that capture the full range of internal variation in long historical documents.

The design and the main idea were conceived together by V. Laippala and myself. Additionally, I was responsible for running all experiments, including data preprocessing, model training, and evaluation. I also conducted the linguistic feature analyses and interpretability assessments. I wrote the first draft of the article. Subsequent versions were revised by all the authors.

5 Synthesis and Discussion

The three studies presented in this thesis collectively not only contribute to the advancement of automatic register analysis in historical English corpora, but also enrich linguistic research by providing corpus-based analyses of register variation and facilitating interdisciplinary studies of language use in historical contexts. In the following sections, the overarching research aims are revisited through a synthesis of the empirical findings and their relation to existing literature. First, the RQs are addressed in light of the results from the three studies. This is followed by a reflection on the limitations encountered during the research. Finally, the chapter concludes with an outline of directions for future work, particularly in light of recent developments in large-scale language modeling.

5.1 Revisiting the Research Questions

This section discusses the RQs presented in Section 1.1 in light of the three papers' results. To summarize, I discovered that register annotations from a curated corpus like COFEA can be repurposed to automatically annotate registers in a large and heterogeneous corpus such as ECCO. Furthermore, I found that register variation between different registers and within individual documents significantly affects classification performance. In particular, registers with clear communicative functions and distinctive linguistic features were easier to identify, while hybrid and ambiguous texts posed challenges. Next, I will go through the RQs and summarize the relevant results from the papers included in this thesis. I will also discuss how the results align with previous research in the field.

RQ1: How can existing manually annotated register data from a curated corpus like COFEA be leveraged to automatically annotate registers in a heterogeneous corpus like ECCO?

RQ1 addresses the foundational challenge of register annotation in historical databases – the absence of consistent metadata and the difficulty of applying linguistic categorization across diverse textual collections. In this thesis, I demonstrate that manually annotated register data from the curated COFEA corpus

can be repurposed to support automatic register annotation in ECCO, a corpus that is both significantly larger and more heterogeneous in its composition.

Study I established a foundation by modeling register variation in COFEA, a corpus with clearly defined and manually annotated registers. Study I demonstrated that modern ML techniques can effectively capture register distinctions in eighteenth-century American English, identifying linguistic features that are strongly indicative of specific registers. These features form the basis for the development of predictive models capable of generalizing beyond the original corpus.

Building on this foundation, Study II explored the transferability of register annotations from COFEA to ECCO. Unlike COFEA, ECCO comprises a vast and diverse collection of texts without existing register labels. Despite this heterogeneity, the models trained on COFEA achieved promising results in predicting registers across ECCO documents, although the registers with identical labels may reflect different communicative settings in these corpora. This finding demonstrates that curated register annotations can serve not only as a descriptive resource but also as a functional tool for annotation transfer. Thus, the findings support the idea that register annotations sufficiently share similar features to be scaled across corpora, provided that the source data are sufficiently representative and the models are carefully tuned. This approach aligns with broader efforts in historical sociolinguistics and digital humanities to make large-scale textual data more accessible and analyzable (Kytö 2019; McGillivray 2021).

At the same time, the use of Transformer-based models, such as BERT and its historical adaptations, reflects a broader methodological shift in corpus linguistics toward deep learning and contextualized language representations. These models offer improved performance in capturing subtle linguistic patterns and have shown particular promise in domain-specific and diachronic corpora (Hou & Smith 2023; Zeng 2024). Their application in the present thesis underscores the potential of integrating linguistic theory with computational tools to address longstanding challenges in corpus annotation.

Study III further validated the proposed approach by examining the consistency of register predictions across different sections of ECCO texts. Notably, the beginning sections of documents yielded stable and reliable register classifications, as has been noted in earlier research (Laippala et al. 2023). This finding suggests that register-specific features are not only transferable but also contextually persistent. This insight has practical implications for corpus expansion by demonstrating that identifying documents with consistent register predictions facilitates the semi-automatic annotation of new texts, thereby enriching the dataset and improving model performance iteratively. Furthermore, this finding reinforces the idea that register-specific features learned from COFEA are not only transferable but also

stable across textual contexts. By applying the best-performing models to unannotated ECCO texts, we identified documents with consistent register predictions, which were then used to expand the dataset and refine the models further.

RQ2: How does variation between and within registers affect the register predictions?

RQ2 examines how variation between and within registers influences the performance and reliability of automatic register classification in historical corpora, addressing the assumption that register is a stable property of a text. Drawing on the results of the three studies, the discussion is organized around variation between registers and within registers.

Previous studies on register theory have emphasized that registers are culturally recognized categories of texts defined by shared communicative purposes and situational contexts (Biber & Conrad 2019; Biber et al. 2020). In corpus-based studies, registers have typically been treated as discrete and stable categories, enabling comparative linguistic analysis across domains (e.g., Biber 1988; Nesi & Gardner 2012). However, recent research has begun to challenge this view, noting that register boundaries are often fluid and that hybrid texts may exhibit features of multiple registers (Biber & Egbert 2023).

Study I demonstrated that letters and journals exhibited strong linguistic signals, resulting in their easy identification. These registers were consistently identified with high accuracy in the COFEA corpus, with the BERT model achieving a micro F1-score of 97%. The success of these predictions suggests that registers with well-defined communicative functions and distinctive linguistic characteristics facilitate more efficient automatic classification. This finding supports earlier claims that registers are associated with recurring linguistic characteristics reflecting their functional roles in the text (Biber 1995).

However, Study II revealed that this predictability did not generalize systematically across corpora. When models trained on COFEA were applied to ECCO, some registers were well identified (e.g., letters), while others were misclassified or not detected at all. In part, these discrepancies may be explained by OCR noise and limitations in the training data's ability to represent the register variation present in ECCO. However, they also point to inter-register ambiguity, particularly in corpora like ECCO in which register boundaries are less clearly defined and texts often exhibit hybrid characteristics due to the nature of the texts and books often including collections of texts or paratextual elements.

In Study II, misclassifications were frequently linguistically and functionally motivated, reflecting shared situational features across registers and functional

elements more typical of the misclassified register. Additionally, texts combining, for example, argumentative and descriptive elements challenged the model's ability to assign a single register label, thus underscoring the fluidity of register categories in historical texts. These results echo findings from previous studies (e.g., Grieve et al. 2011), which indicate that hybrid texts, especially in evolving communicative environments, challenge traditional classification schemes. The presence of shared linguistic features across registers led to systematic confusion in model predictions, thus highlighting the limitations of treating registers as rigid categories.

Study III shifted our focus to internal document variation, mainly investigating how register predictions vary across different sections of a single text. By segmenting ECCO documents into five parts, the study found that text beginnings yielded the most reliable register predictions, while middle sections were significantly less predictable. This finding aligns with earlier studies on web texts, which found that openings tend to be the most predictive for register classification (Laippala et al. 2023). Study III also showed that models trained on ending segments or full-text combinations resulted in the most balanced performance across all document parts. This positional effect suggests that register signals are not evenly distributed throughout a document and that certain sections carry stronger linguistic and functional cues. In all, these findings challenge the assumption that register is a stable, document-level property and instead support a view of registers as locally realized and context-sensitive, especially in long historical texts. This finding aligns with the recent calls to move beyond document-level assumptions and consider register as a locally realized phenomenon (Egbert & Gracheva 2023; Goulart et al. 2022; Biber & Egbert 2023).

RQ3: How can the examination of linguistic features enhance the understanding of register-specific language use and contribute to metadata enrichment in large historical corpora?

RQ3 addresses a central concern in historical corpus linguistics: the potential of linguistic feature analysis to deepen our understanding of register-specific language use and to support metadata enrichment in large, unannotated corpora, thus allowing us to make sense of large, messy, and heterogenous databases. Across the three studies, this question has been explored through increasingly complex scenarios from curated corpora with manual annotations to heterogeneous, unannotated corpora, and finally to intra-document variation within long texts. While RQ1 and RQ2 addressed the feasibility and variability of automatic register classification, RQ3 explored how linguistic features themselves can deepen our understanding of register-specific language use and serve as a foundation for enriching metadata in historical corpora.

The findings in this thesis challenge the notion that metadata must be externally imposed or manually curated. Instead, they demonstrate that metadata can be inferred from the linguistic structures of the texts themselves, provided that the features are systematically analyzed and interpreted. This notion aligns with the broader movement in digital humanities toward data-driven enrichment, in which computational methods are used to uncover latent structures in historical datasets (Piotrowski 2012; Hill & Hengchen 2019).

The interpretability of linguistic features also opens up new avenues for functional genre analysis. Rather than treating registers as fixed categories, the thesis shows that they can be understood as emergent patterns of language use, mainly shaped by communicative purpose and textual context. For example, the consistent presence of legal terminologies across segments of case and proceeding texts suggests that these registers are defined not just by topic but by institutional discourse practices. Similarly, the interpersonal features found in letters, such as personal pronouns and stance-taking verbs, reflect the functional adaptations of personal reference and opinions in interpersonal interactions that persist across time. These insights contribute to a more nuanced understanding of historical genres, thus addressing calls in genre theory to view genres as social actions rather than static forms (Miller 1984; Devitt 2004).

The thesis also highlights the temporal and positional dynamics of register signals. The finding that text beginnings are more predictive than middle sections suggests that authors may foreground register-relevant cues early in the text, or that annotators of the text possibly focus more on the beginnings of the texts when deciding the register of the text. This finding has implications for corpus design and annotation strategies: Rather than requiring full-text analysis, metadata enrichment could be efficiently achieved by focusing on strategic textual zones. This resonates with earlier work on genre detection in web texts (Laippala et al. 2023) and also extends it to historical corpora, where textual conventions may differ significantly.

In terms of metadata enrichment, the implications are twofold. First, linguistic features provide a scalable method for annotating large corpora, thus reducing reliance on manual labor and enabling broader access to historical data. Second, they offer interpretive depth, allowing researchers to explore not just what a text is, but how it functions within its historical and social context. The integration of keyword extraction methods such as SACX further enhances the interpretability of the models, allowing for a linguistically motivated understanding of classification decisions. This finding aligns with recent calls in explainable artificial intelligence (AI) and computational humanities to ensure that model outputs are not only accurate but also transparent and grounded in linguistic theory (Belinkov & Glass 2019).

ML methods excel at capturing subtle and context-dependent features that reflect the functional roles of registers. Beyond its role in classification, ML offers powerful

tools for linguistic research by facilitating the discovery of latent structures in textual data, which may not be apparent through manual analysis. The ability to extract and analyze discriminative features through interpretability techniques bridges computational predictions with linguistic theory and reinforces the idea that registers are associated with consistent linguistic characteristics, thus reflecting their communicative purposes. This notion not only enhances our understanding of historical discourse but also contributes to the development of more nuanced and data-driven approaches to genre and register theory. Furthermore, the scalability and efficiency brought by ML make it possible to annotate and analyze millions of words or documents, which is an essential advantage for handling historical corpora where manual annotation is often unfeasible.

5.2 Limitations

While the results of this thesis demonstrate the potential of register analysis to enrich our understanding of historical corpora, several aspects invite critical reflection. These limitations arise from the technical challenges of working with historical data and the methodological choices made during the study. In particular, working with historical texts digitized from printed sources requires careful attention to issues of textual integrity, representativeness, and annotation consistency. Moreover, the analytical choices made in modeling register categories and interpreting linguistic features inevitably shape the scope and reliability of the findings.

One of the most persistent challenges in working with historical corpora like ECCO is the quality of the source texts, which are often digitized from printed materials using different OCR software. OCR errors ranging from misrecognized characters to missing or fragmented words can significantly distort the linguistic features extracted from the texts. These distortions affect the surface-level features, such as word frequency, spelling variants, and deeper syntactic or grammatical patterns deemed crucial for register classification.

The variability and unpredictability of errors across ECCO documents also remain a source of uncertainty. This issue is particularly acute in texts with complex layouts, marginalia, or degraded print quality. As previous studies have shown (Baron & Rayson 2008; Hill & Hengchen 2019), OCR noise can introduce bias into linguistic analyses and reduce model performance. Thus, future studies should explore more robust OCR correction techniques or integrate confidence scores from OCR engines to filter unreliable text segments.

Another limitation stems from the decision to treat register classification as a single-label task, in which every document is assigned just one register label. This choice was made to maintain interpretability and consistency across data and analyses, as the texts in COFEA were originally annotated with just a single register

label per text. However, this approach does not fully reflect the hybrid nature of many historical texts, which often combine multiple communicative functions and could plausibly belong to more than one register. Although we considered multilabel classification wherein each text can have several register labels, we ultimately rejected this due to several concerns. First, the lack of multilabel annotations in COFEA would have required speculative labeling, thus undermining the reliability of the training data. Second, multilabel models tend to be more complex and less interpretable, which would have conflicted with this study'd emphasis on linguistic transparency. Nonetheless, the results suggest that a multilabel approach may be more appropriate. Future research could explore semi-supervised or probabilistic models that allow for register overlap without sacrificing interpretability.

Additionally, the manual annotation of ECCO texts was necessary for evaluation, but it was limited in scope due to time and resource constraints. A larger annotated sample could have provided a more robust basis for assessing model generalizability and for understanding the full range of register variation in ECCO. In hindsight, a larger collaborative annotation effort might have helped scale this process and diversify the perspectives involved in defining registers.

Furthermore, the curated nature of COFEA, while beneficial for model training, also introduced register imbalance. In particular, the register class letter was overrepresented in the COFEA dataset due to original sampling strategy for the annotation. Other registers, such as contracts, had relatively few examples. Given that underrepresented registers were more prone to misclassification and less likely to yield stable linguistic patterns, such an imbalance affected both model performance and feature analysis. Several strategies could have been employed to address this issue. For instance, data augmentation techniques, such as paraphrasing or synthetic text generation, might have helped balance the dataset. Alternatively, sampling methods or weighted loss functions could have been used to reduce bias during model training. However, these approaches carry their own risks, such as the introduction of artificial patterns or the distortion of historical authenticity. In this thesis, the decision was made to prioritize data integrity over balance. However, for Study III, the number of the largest class was capped to the second largest class to ensure a more balanced dataset. Hence, future studies should consider more nuanced strategies for handling register imbalance, especially when scaling to larger or more diverse corpora.

Beyond these core limitations, several broader issues merit attention. In particular, the reliance on supervised learning assumes that register categories are stable and well-defined, which may not hold true across different historical periods or cultural contexts. Moreover, the linguistic features used in the models are shaped by contemporary linguistic theory and may not fully capture the communicative norms of the eighteenth century. This, therefore, raises questions about diachronic

validity and the need for historically sensitive feature engineering. Finally, while the current study focused on English-language corpora, the methods and findings may not be generalized to other languages or multilingual corpora in which register conventions and linguistic structures differ significantly. Thus, cross-linguistic validation would be a valuable direction for future research.

5.3 Future Works

The findings of this thesis open several avenues for future research in register analysis and corpus enrichment, particularly in the context of historical corpora. While the current work has demonstrated the viability of supervised classification using curated linguistic features, there are other methodological and technological opportunities to expand and refine this approach.

For example, future research could benefit from the use of generative LLMs, such as GPT-style models, which have become increasingly prominent in NLP and digital humanities. At the time this research began, such models were either unavailable or not yet mature enough to be reliably applied. However, their use raises several concerns from a corpus linguistics perspective. The most widely used generative models are closed commercial systems, offering no public access to their training data or model architecture. This lack of transparency poses a challenge for scientific inquiry, as it is difficult to assess what linguistic knowledge these models encode, how they process historical variation, or what biases they may introduce.

From a corpus linguistics standpoint, the abovementioned issue also raises the question of what new knowledge can be gained from studying systems whose internal workings are inaccessible. Although open-source models with comparable performance are now emerging, the research conducted in this thesis predates their availability. Thus, future studies could explore the controlled use of generative models for tasks, such as register annotation, document summarization, or register hypothesis generation, provided that these outputs are critically evaluated and supplemented with feature-based analysis. Hybrid approaches that combine the interpretability of traditional models with the flexibility of generative systems may also offer a promising middle ground.

Furthermore, the results of this thesis point to the need for more flexible register models, particularly in light of the hybrid and multifunctional nature of many historical texts. The decision to use single-label classification was motivated by the lack of multilabel annotations and the desire for interpretability. However, future studies should consider multilabel or probabilistic classification frameworks that allow texts to be associated with multiple registers or degrees of register features, thereby capturing the complexity of texts that serve multiple communicative purposes. For instance, categories known to be hybrid (e.g., proposals in letter

format) reflect the cultural reality of overlapping practices. This approach, however, would require rethinking annotation practices and developing new guidelines for identifying overlapping communicative functions. It may also involve the use of unsupervised or semi-supervised methods to detect register mixtures in unannotated corpora. Despite the difficulties involved, such approaches would better reflect the fluidity of genre boundaries in historical writing and support more nuanced metadata enrichment.

Finally, the findings from the intra-document analysis suggest that textual structure and position play a significant role in register signaling. Future research could thus incorporate structural metadata, such as headings, paratextual elements, or document layout, into classification models. Doing so would allow for a more precise identification of register cues and support section-level annotation, which may be especially useful for processing or analyzing long or composite texts.

6 Conclusions

This thesis set out to address three interrelated RQs concerning register identification, variation, and linguistic feature analysis in historical English corpora. Through the case studies, this work demonstrated that computational methods can be applied to model and predict registers in curated and heterogeneous corpora, and that linguistic features offer a robust foundation for understanding register-specific language use and enriching metadata in large historical datasets.

Over the course of this research, my understanding of register has evolved significantly. Initially, I approached register as a relatively stable and discrete property of texts, suitable for classification and comparative analysis. Throughout the dissertation, I have adopted the text linguistic perspective in which register refers to functionally motivated, situationally grounded variation manifested in distributions of lexico-grammatical features, whereas genre concerns conventionalized rhetorical organizations and specialized expressions that tend to be localized to certain text parts (e.g., openings/closings of letters, formulaic headings). This distinction, following a long tradition in corpus linguistics, underpins the analytical design across Studies I–III. However, working with ECCO's heterogeneous and often ambiguous texts revealed the fluidity and complexity of register categories, especially in historical contexts. I came to see register not as a fixed label but as a dynamic interplay of linguistic features, communicative intent, and textual context.

In light of previous research, this work contributes to a more nuanced understanding of register as a cultural and functional construct, often shaped by social conventions and textual contexts. This finding supports the shift toward viewing registers not as fixed scientific categories, but as flexible, culturally mediated frameworks for organizing language use, which is an approach that has been increasingly adopted in contemporary register theory. At the same time, the tradition retains the claim that registers display functional correspondences between situations and linguistic features.

The empirical pipeline in the present study implicitly treats registers as cultural categories insofar as the labels stem from historically situated annotation schemes. At the same time, our explanations seek functional correspondence by showing how

specific lexico-grammatical distributions plausibly map to communicative purposes (e.g., interpersonal involvement in letters; institutional precision in legal cases). Importantly, Study II's cross corpus transfer and Study III's intra document diagnostics provide evidence about the stability (or lack thereof) of these cultural categories and correspondences. For example, letters and cases transfer well from COFEA to ECCO, suggesting that these categories maintain broadly similar functional correspondences across late modern domains and repositories. However, proposals and proceedings show frequent confusion with letters or personal records, indicating category overlap or genre staging that may shift within an evolving print culture. Such overlaps caution against assuming uniform, time invariant correspondences.

Crucially, in classification settings the genre–register distinction is not straightforward. Models tend to attend to highly localized conventions, such as opening formulae or salutations, that behave like genre markers rather than register pervasive features. These surface cues can temporarily dominate predictions, yielding confident classifications that reflect genre staging rather than the underlying functional distribution of lexico-grammatical traits. Practically, this means that apparent model accuracy can be inflated by positional or formulaic signals unless we diagnose how and where attributions arise.

Study III provides the clearest evidence that model interpretability and keyword analysis can function as practical diagnostics for separating pervasive register features from localized, convention-based genre features. The misclassifications observed in Study II's proposal letters reinforce this point, showing how surface conventions can mimic functional signals, which in turn highlights the need for IG/SACX audits to ensure that genre staging is not mistaken for register.

Segmenting long ECCO documents into five equal parts revealed that register cues are strongest at beginnings but, when a text consistently instantiates a single register, those cues also persist across segments. In letters, for instance, the recurrent presence of personal pronouns, attributive adjectives, public and mental verbs, and common nouns aligns with interpersonal and informational functions identified in multidimensional and corpus-based studies (Biber 1988; Biber & Finegan 1989; Biber 2001). IG attributions highlight the tokens that systematically drive predictions; SACX aggregates those attributions to identify stable keywords that recur across a register and across segments. In our data, such stability is evident for letters and legal texts, where legal terminology functions as a marker of specialized institutional discourse. By contrast, highly localized features are tied to rhetorical moves and are therefore more sensibly treated as genre markers.

Practically, the methods offer an operational test: register features appear with persistent attribution across segments and across many texts of a class; genre features produce attribution spikes confined to particular structural zones (openings, closings,

front matter). This distinction is already visible in our misclassification patterns in Study II, where proposals written in letter format were often predicted as letters: the genre staging temporarily dominated the surface signal, but the document internal distribution of features (interpersonal vs. informational balance; stance taking) remained the decisive register evidence once examined with IG/SACX.

Adopting segment-based evaluation reframes register as a locally realized property rather than a uniformly document-level constant. Beginnings are consistently more predictive than middles, both in our historical data and in contemporary web studies (Laippala et al. 2023), which is likely due to conventionalized front loading of purpose, audience orientation, and footing in eighteenth century print. The middles of long documents often house mixed functions, such as exposition, narration, digression, embedded documents (e.g., letters in novels), which produces lower predictability. This pattern dovetails with calls to treat register boundaries as fuzzy, acknowledges hybridization within documents, and invites segment level annotation and analysis.

The insights from the results of this study have important implications for the design of register annotation systems in historical corpora. In particular, these insights suggest that static, document-level labels may be insufficient for capturing the complexity of register variation and that dynamic, context-sensitive approaches are needed. The findings from intra-document analysis specifically highlight how register signals vary across text parts, with beginnings often carrying the strongest cues. This finding therefore calls for annotation strategies that are sensitive to positional and contextual variation, rather than assuming uniformity across entire documents.

Furthermore, the interpretability analyses (e.g., IG and SACX) confirm that models rely on linguistically meaningful features, reinforcing the value of automatic classification not only for NLP applications but also for linguistic research into historical discourse structure. Specifically, these methods bridged the gap between computational modeling and theoretical insight, allowing for a more transparent understanding of how registers are identified and what linguistic traits define them. Conceptually, we can distinguish register pervasive signals from genre localized moves with empirical precision, and we can measure where cultural categories and functional mappings appear stable, where they overlap, and where they drift. Practically, these advances pave the way toward transparent, explainable metadata enrichment at scale.

Additionally, my view of corpus linguistics has expanded from a primarily descriptive and statistical discipline to one that can be considered innovative from interdisciplinary and methodological perspectives. The integration of deep learning models, interpretability techniques, and linguistic theory has demonstrated how computational tools can not only scale analysis but also deepen our understanding of

language use. This thesis has therefore reinforced the importance of linguistic interpretability, historical sensitivity, and contextual nuance in corpus-based research.

From an interdisciplinary standpoint, this thesis offers tools and conceptual insights for fields beyond linguistics. In digital humanities, for example, the ability to automatically annotate registers in large corpora supports metadata enrichment and facilitates new modes of textual analysis. In legal history, the identification of legal registers facilitates the targeted exploration of institutional discourse. In sociolinguistics, the interpersonal and informational features found in letters and journals contribute to understanding how language encodes social relationships and epistemic stance. Overall, these applications underscore the value of integrating computational methods with linguistic theory to make historical data more accessible and analyzable for researchers in various fields.

Abbreviations

AFI	Abstract Feature Importance
AI	Artificial Intelligence
ARCHER	A Representative Corpus of Historical English Registers
BERT	Bidirectional Encoder Representations from Transformers
CEAL	Corpus of Early American Literature
CNN	Convolutional Neural Network
COFEA	Corpus of Founding Era American English
ECCO	Eighteenth Century Collections Online
ECCO-TCP	Eighteenth Century Collections Online Text Creation Partnership
eMOP	Early Modern OCR Project
ESTC	English Short Title Catalogue
GMA	Geometric Multivariate Analysis
IAA	Inter-Annotator Agreement
IG	Integrated Gradients
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
ML	Machine Learning
MDA	Multidimensional Analysis
NAU Tagger	Northern Arizona University Tagger
NLP	Natural Language Processing
OCR	Optical Character Recognition
RQ	Research Question
RNN	Recurrent Neural Network
SACX	Stable Attribution Class Explanation method
SFL	Systemic Functional Linguistics
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
XLM-R	XLM-RoBERTa

List of References

Adesam, Yvonne, Dana Dannélls, and Nina Tahmasebi. 2019. “Exploring the Quality of the Digital Historical Newspaper Archive KubHist.” D*igital Humanities in the Nordic and Baltic Countries Publications* 2 (1): 9–17.

Aghababaei, Ali, Jan Nikadon, Magdalena Formanowicz, et al. 2025. “Application of Integrated Gradients Explainability to Sociopsychological Semantic Markers.” *arXiv* preprint, March 6. https://doi.org/10.48550/arXiv.2503.04989.

Ahmed, Mumtahina, Mohammad Shahadat Hossain, Raihan Ul Islam, and Karl Andersson. 2022. “Explainable Text Classification Model for COVID-19 Fake News Detection.” *Journal of Internet Services and Information Security* 12 (2): 51–69. https://doi.org/10.22667/JISIS.2022.05.31.051.

Akiba, Takuya, Shinya Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. “Optuna: A Next-Generation Hyperparameter Optimization Framework.” In P*roceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. New York: Association for Computing Machinery.

Alatrash, Reem, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. “CCOHA: Clean Corpus of Historical American English.” In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6958–6966, Marseille, France. European Language Resources Association.

Amin, Saadullah, and Guenter Neumann. 2021. “T2NER: Transformers Based Transfer Learning Framework for Named Entity Recognition.” In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 212–20. https://doi.org/10.18653/v1/2021.eacl-demos.25.

Baron, Alistair, and Paul Rayson. 2008. “Automatic Standardisation of Texts Containing Spelling Variation.” In *Proceedings of the Corpus Linguistics Conference CL2009*, University of Liverpool, UK, July 20–23, 2009. Edited by Michaela Mahlberg, Victorina González-Díaz, and Catherine Smith. Article #314.

Beck, Christin, and Marisa Köllner. 2023. “GHisBERT – Training BERT from Scratch for Lexical Semantic Investigations across Historical German Language Stages.” In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, 33–45. https://doi.org/10.18653/v1/2023.lchange-1.4.

Belinkov, Yonatan and James Glass. 2019. “Analysis Methods in Neural Language Processing: A Survey.” *Transactions of the Association for Computational Linguistics* 7: 49–72.

Van Belle, Vanya, Ben Van Calster, Sabine Van Huffel, Johan A. K. Suykens, and Paulo Lisboa. 2016. “Explaining Support Vector Machines: A Color Based Nomogram.” *PLOS ONE* 11 (10): e0164568. https://doi.org/10.1371/journal.pone.0164568.

Bergs, Alexander T. 2004. “Letters: A new approach to text typology.” *Journal of Historical Pragmatics* 5 (2): 207–227.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, Douglas. 2001. “Dimensions of Variation among Eighteenth-Century Speech-Based and Written Registers.” In *Variation in English: Multi-Dimensional Studies*, edited by Susan Conrad and Douglas Biber, 200–14. London: Longman.

Biber, Douglas. 2006. *University Language: A Corpus-Based Study of Spoken and Written Registers*. Studies in Corpus Linguistics, vol. 23. Amsterdam: John Benjamins.

Biber, Douglas. 2012. “Register as a Predictor of Linguistic Variation.” *Corpus Linguistics and Linguistic Theory* 8 (1): 9–37.

Biber, Douglas. 2019. “Text-Linguistic Approaches to Register Variation.” *Register Studies* 1 (1): 42–75. https://doi.org/10.1075/rs.18007.bib.

Biber, Douglas, and Susan Conrad. 2019. *Register, Genre, and Style*. 2nd ed. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd, and Marie Helt. 2002. “Speaking and Writing in the University: A Multidimensional Comparison.” *TESOL Quarterly* 36 (1): 9. https://doi.org/10.2307/3588359.

Biber, Douglas, and Edward Finegan. 1989. “Drift and the Evolution of English Style: A History of Three Genres.” *Language* 65 (3): 487–517.

Biber, Douglas, and Edward Finegan. 1992. “The Linguistic Evolution of Five Written and Speech-Based English Genres from the 17th to the 20th Centuries.” In *History of Englishes: New Methods and Interpretations in Historical Linguistics*, edited by Matti Rissanen, Ossi Ihalainen, Terttu Nevalainen, and Irma Taavitsainen, 688–704. Berlin: Mouton de Gruyter.

Biber, Douglas, and Edward Finegan, eds. 1994. *Sociolinguistic Perspectives on Register*. Oxford: Oxford University Press.

Biber, Douglas, and Edward Finegan. 1997. “Diachronic Relations among Speech-Based and Written Registers in English.” In *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*, edited by Terttu Nevalainen and Leena Kahlas-Tarkka, 253–75. Helsinki: Société Néophilologique.

Biber, Douglas, and Jesse Egbert. 2018. *Register Variation Online*. Cambridge: Cambridge University Press.

Biber, Douglas, and Jesse Egbert. 2023. “What Is a Register?: Accounting for Linguistic and Situational Variation within – and Outside of – Textual Varieties.” *Register Studies* 5 (1): 1–22. https://doi.org/10.1075/rs.00004.bib.

Biber, Douglas, Jesse Egbert, and Daniel Keller. 2020. “Reconceptualizing Register in a Continuous Situational Space.” *Corpus Linguistics and Linguistic Theory* 16 (3): 581–616. https://doi.org/10.1515/cllt-2018-0086.

Biber, Douglas, and Bethany Gray. 2013. “Being Specific about Historical Change: The Influence of Sub-Register.” *Journal of English Linguistics* 41 (2): 104–34. https://doi.org/10.1177/0075424212472509.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer.

Bobojonova, Latofat, Arofat Akhundjanova, Phil Ostheimer, and Sophie Fellenz. 2025. “BBPOS: BERT-Based Part-of-Speech Tagging for Uzbek.” *arXiv* preprint, January 17. https://doi.org/10.48550/arXiv.2501.10107.

Boros, Emanuela, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. 2024. “Post-Correction of Historical Text Transcripts with Large Language Models: An Exploratory Study.” In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (LaTeCH-CLfL 2024), 133–59. St. Julians, Malta: Association for Computational Linguistics.

Brown, Roger, and Marguerite Ford. 1961. “Address in American English.” *The Journal of Abnormal and Social Psychology* 62 (2): 375–85. https://doi.org/10.1037/h0042862.

Carpenter v. United States, No. 16–402, 585 U.S. ____ (2018).

Cha, Javier. 2023. “Big Data Studies: The Humanities in Uncharted Waters.” *Korean Studies* 47 (1): 274–99. https://doi.org/10.1353/ks.2023.a908625.

Chan, Jireh Yi-Le, Khean Thye Bea, Steven Mun Hong Leow, Seuk Wai Phoong, and Wai Khuen Cheng. 2023. “State of the Art: A Review of Sentiment Analysis Based on Sequential Transfer Learning.” *Artificial Intelligence Review* 56 (1): 749–80. https://doi.org/10.1007/s10462-022-10183-8.

Chen, Danqi, and Christopher Manning. 2014. “A Fast and Accurate Dependency Parser Using Neural Networks.” In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 740–50. https://doi.org/10.3115/v1/D14-1082.

Chen, Jessica, and Karen Wang. 2023. *Bert-Powered Book Genre Classification*. Stanford University CS224N project report. Accessed January 13, 2025. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/final-reports/final-report-169280003.pdf

Chen, Yuqi, Sixuan Li, Ying Li, and Mohammad Atari. 2024. “Surveying the Dead Minds: Historical-Psychological Text Analysis with Contextualized Construct Representation (CCR) for Classical Chinese.” In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2597–615. https://doi.org/10.18653/v1/2024.emnlp-main.151.

Chiche, Alebachew, and Betselot Yitagesu. 2022. “Part of Speech Tagging: A Systematic Review of Deep Learning and Machine Learning Approaches.” *Journal of Big Data* 9 (1). https://doi.org/10.1186/s40537-022-00561-y.

Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. “ELECTRA: Pre-Training Text Encoders as Discriminators Rather Than Generators.” *arXiv* preprint, March 23. https://doi.org/10.48550/arXiv.2003.10555.

Cohen, Jacob. 1960. “A Coefficient of Agreement for Nominal Scales.” *Educational and Psychological Measurement* 20 (1): 37–46.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. “Unsupervised Cross-lingual Representation Learning at Scale.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://aclanthology.org/2020.acl-main.747/.

Conneau, Alexis, and Guillaume Lample. 2019. “Cross-Lingual Language Model Pretraining.” In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Article 634, 7059–69. Red Hook, NY: Curran Associates Inc.

Crystal, David. 1991. *A Dictionary of Linguistics and Phonetics*. Oxford, UK: Basil Blackwell.

Crystal, David, and Derek Davy. 1969. *Investigating English Style*. London and New York: Routledge.

Culpeper, Jonathan, and Merja Kytö. 2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: Cambridge University Press.

Cunningham, Clark D., and Jesse Egbert. 2020. “Using Empirical Data to Investigate the Original Meaning of ‘Emolument’ in the Constitution.” *Georgia State University Law Review* 36 (5).

Cvrček, Václav, and Martina Berrocal. 2025. “Sibling-Texts Keyword Analysis: Exploring Topic and Register Keywords.” *Digital Scholarship in the Humanities*, May 16, fqaf037. https://doi.org/10.1093/llc/fqaf037.

Devitt, Amy J. 2004. *Writing Genres*. Carbondale: Southern Illinois University Press.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” In *Proceedings of NAACL HLT 2019: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers), 4171–86.

Dey, Sudipta, and Tathagata Roy Chowdhury. 2024. “A Comparative Survey of SHAP and LIME: Explaining Machine Learning Models for Transparent AI.” *International Journal of Innovative Technology* 11 (6): 827–35. https://ijirt.org/article?manuscript=169215.

Diederich, Joachim, Jörg Kindermann, Edda Leopold, And Gerhard Paass. 2003. "Authorship Attribution with Support Vector Machines." *Applied Intelligence* 19 (1/2): 109–23. https://doi.org/10.1023/A:1023824908771.

District of Columbia v. Trump, No. 18–2488. (4th Cir. 2019).

Diwersy, Sascha, Stefan Evert, and Stella Neumann. 2014. "A Weakly Supervised Multivariate Approach to the Study of Language Variation." In *Aggregating Dialectology, Typology, and Register Analysis*, edited by Benedikt Szmrecsanyi and Bernhard Wälchli, 174–204. Berlin: De Gruyter Mouton.

Dixit, Vimal, Kamlesh Dutta, and Pardeep Singh. 2015. "Word Sense Disambiguation and Its Approaches." *CPUH-Research Journal* 1 (2): 54–58. ISSN (Online): 2455-6076.

Earhart, Amy E. 2012. "Can Information Be Unfettered? Race and the New Digital Humanities Canon." In *Debates in the Digital Humanities*, edited by Matthew K. Gold. Minneapolis: University of Minnesota Press. https://doi.org/10.5749/minnesota/9780816677948.003.0030.

Egbert, Jesse. 2015. "Publication Type and Discipline Variation in Published Academic Writing: Investigating Statistical Interaction in Corpus Data." *International Journal of Corpus Linguistics* 20 (1): 1–29. https://doi.org/10.1075/ijcl.20.1.01egb.

Egbert, Jesse, and Doug Biber. 2019. "Incorporating Text Dispersion into Keyword Analyses." *Corpora* 14 (1): 77–104. https://doi.org/10.3366/cor.2019.0162.

Egbert, Jesse, and Maria Gracheva. 2023. "Linguistic Variation within Registers: Granularity in Textual Units and Situational Parameters." *Corpus Linguistics and Linguistic Theory* 19 (1): 115–43.

Egbert, Jesse, and Michaela Mahlberg. 2020. "Fiction – One Register or Two?: Speech and Narration in Novels." *Register Studies* 2 (1): 72–101.

Ekbia, Hamid, Michael Mattioli, Inna Kouper, et al. 2015. "Big Data, Bigger Dilemmas: A Critical Review." *Journal of the Association for Information Science and Technology* 66 (8): 1523–45. https://doi.org/10.1002/asi.23294.

Elsweiler, Christine, and Patricia Ronan. 2023. "From I Am, with Sincere Regard, Your Most Obedient Servant to Yours Sincerely: The Simplification of Leavetaking Formulae in 18th-Century Scottish and Irish English Letters." *ICAME Journal* 47 (1): 1–17. https://doi.org/10.2478/icame-2023-0001.

Fanego, Teresa. 2012. "COLMOBAENG: 'A Corpus of Late Modern British and American English Prose.'" In *Creation and Use of Historical English Corpora in Spain*, edited by N. Vázquez, 101–17. Newcastle upon Tyne: Cambridge Scholars Publishing.

Fanego, Teresa, Paula Rodríguez-Puente, María José López-Couso, Belén Méndez-Naya, Pilar Núñez-Pertejo, Cristina Blanco-García, and Isabel Tamaredo. 2017. "The Corpus of Historical English Law Reports 1535–1999 (CHELAR): A Resource for Analysing the Development of English Legal Discourse." *ICAME Journal* 41 (1): 53–82. https://doi.org/10.1515/icame-2017-0003.

Farag, Youmna, Helen Yannakoudakis, and Ted Briscoe. 2018. "Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long Papers), 263–71. https://doi.org/10.18653/v1/N18-1024.

Ferguson, Charles A. 1968. "Language Development." In *Language Problems of Developing Nations*, edited by Joshua A. Fishman, Charles A. Ferguson, and Jyotirindra Das Gupta, 27–36. New York: John Wiley & Sons.

Ferguson, Charles A. 1994. "Dialect, Register and Genre: Working Assumptions about Conventionalization." In *Sociolinguistic Perspectives on Register*, edited by Douglas Biber and Edward Finegan, 15–30. New York: Oxford University Press.

Finn, Aidan, and Nicholas Kushmerick. 2006. "Learning to Classify Documents According to Genre." *Journal of the American Society for Information Science and Technology* 57 (11): 1506–18. https://doi.org/10.1002/asi.20427.

Francis, W. Nelson, and Henry Kucera. 1967. *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Frenken, Florian, Stephanie Evert, Gerold Schneider, and Stella Neumann. 2025. "How Stable Are Multivariate Findings about Register Variation across Varieties of English? On the Replicability of Geometric Multivariate Analysis." *ICAME Journal* 49 (1): 23–45. https://doi.org/10.2478/icame-2025-0003.

Gabay, Simon, Pedro Ortiz Suarez, Alexandre Bartz, et al. 2022. "From FreEM to D'AlemBERT: A Large Corpus and a Language Model for Early Modern French." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* pages 3367–3374, Marseille, France. European Language Resources Association.

Gale. 2016. *Eighteenth Century Collections Online: The Most Comprehensive Online Library of English and Foreign-Language Titles Printed in the United Kingdom during the Eighteenth Century, plus Thousands of Important Works Printed in English Elsewhere*. Accessed February 15, 2023. https://www.gale.com/binaries/content/assets/gale-usen/primary-sources/eighteenth-century-collections-online/ecco-roll-fold-2016web.pdf

Garí Soler, Aina, and Marianna Apidianaki. 2021. "Let's Play Mono - Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses." *Transactions of the Association for Computational Linguistics* 9 (August): 825–44. https://doi.org/10.1162/tacl_a_00400.

Geisler, Cheryl. 2002. "Investigating Register Variation in Nineteenth-Century English." In *Using Corpora to Explore Linguistic Variation*, edited by Randi Reppen, Susan M. Fitzmaurice, and Douglas Biber, 249–71. Amsterdam: John Benjamins Publishing Company.

Ghous, Hamid, Mubasher Malik, and Iqra Rehman. 2023. "Deep Learning Based Market Basket Analysis Using Association Rules." *KIET Journal of Computing and Information Sciences* 6 (2): 14–34. https://doi.org/10.51153/kjcis.v6i2.166.

González-Álvarez, Dolores, and Javier Pérez-Guerra. 1998. "Texting the Written Evidence: On Register Analysis in Late Middle English and Early Modern English." *Text* 18 (3): 321–48.

Goulart, Larissa, Douglas Biber, and Randi Reppen. 2022. "In This Essay, I Will …: Examining Variation of Communicative Purpose in Student Written Genres." *Journal of English for Academic Purposes* 59: 101159. https://doi.org/10.1016/j.jeap.2022.101159.

Goyal, Anshaj, and V. Prem Prakash. 2022. "Statistical and Deep Learning Approaches for Literary Genre Classification." In *Advances in Data and Information Sciences* vol. 318, edited by Shailesh Tiwari, Munesh C. Trivedi, Mohan Lal Kolhe, K.K. Mishra, and Brajesh Kumar Singh. Lecture Notes in Networks and Systems. Singapore: Springer. https://doi.org/10.1007/978-981-16-5689-7_26.

Gray, Bethany. 2015. *Linguistic Variation in Research Articles: When Discipline Tells Only Part of the Story*. 1st ed. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/scl.71.

Gregg, Stephen H. 2020. *Old Books and Digital Publishing: Eighteenth-Century Collections Online.* https://doi.org/10.1017/9781108767415.

Gries Stefan Th. 2006. "Exploring variability within and between corpora: some methodological considerations." *Corpora* 1 (2): 109–51.

Gries, Stefan Th. 2021. "A New Approach to (Key) Keywords Analysis: Using Frequency, and Now Also Dispersion." *Research in Corpus Linguistics* 9 (2): 1–33. https://doi.org/10.32714/ricl.09.02.02.

Grieve, Jack, Douglas Biber, Eric Friginal, and Tanya Nekrasova. 2011. "Variation among Blogs: A Multi-Dimensional Analysis." In *Genres on the Web: Computational Models and Empirical Studies*, edited by Alexander Mehler, Serge Sharoff, and Marina Santini, 303–22. Berlin: Springer.

Guyon, Isabelle, and André Elisseeff. 2003. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research* 3: 1157–82.

Halliday, M. A. K., and Ruqaiya Hasan. 1985. *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*. Oxford: Oxford University Press.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." In *Proceedings of the 54th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), 1489–1501. https://doi.org/10.18653/v1/P16-1141.

Harris, Zellig. 1954. "Distributional Structure." *Word* 10 (23): 146–62.

Hashimoto, Brett. 2023. "Corpus of Founding Era American English: Designing a Corpus for Interpreting the United States Constitution." *Corpora* 18 (1): 1–14. https://doi.org/10.3366/cor.2023.0270.

Hashimoto, Brett. n.d. "Data: COFEA." https://sites.google.com/site/brettjameshashimoto/data. Accessed January 19, 2023.

Hauswedell, Tessa, Julianne Nyhan, M. H. Beals, Melissa Terras, and Emily Bell. 2020. "Of Global Reach yet of Situated Contexts: An Examination of the Implicit and Explicit Selection Criteria That Shape Digital Archives of Historical Newspapers." *Archival Science* 20 (2): 139–65. https://doi.org/10.1007/s10502-020-09332-1.

Henriksson, Erik, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson, et al. 2024a. "From Discrete to Continuous Classes: A Situational Analysis of Multilingual Web Registers with LLM Annotations." In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, 308–18. https://doi.org/10.18653/v1/2024.nlp4dh-1.30.

Henriksson, Erik, Amanda Myntti, Saara Hellström, Anni Eskelinen, Selcen Erten-Johansson, and Veronika Laippala. 2024b. "Automatic Register Identification for the Open Web Using Multilingual Deep Learning." *arXiv* preprint, December 10. https://doi.org/10.48550/arXiv.2406.19892.

Hill, Mark J., and Simon Hengchen. 2019. "Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study." *Digital Scholarship in the Humanities* 34 (4): 825–43. https://doi.org/10.1093/llc/fqz024.

Höglund, Mikko, and Kaj Syrjänen. 2016. "Corpus of Early American Literature." *ICAME Journal* 40 (1): 17–38. https://doi.org/10.1515/icame-2016-0003.

Holahan, Cassidy. 2021. "Rummaging in the Dark: ECCO as Opaque Digital Archive." *Eighteenth-Century Studies* 54 (4): 803–26. https://doi.org/10.1353/ecs.2021.0093.

Hong, Gumwon. 2005. "Relation Extraction Using Support Vector Machine." In *Natural Language Processing – IJCNLP 2005*, edited by Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong, Vol. 3651. Lecture Notes in Computer Science. Berlin: Springer. https://doi.org/10.1007/11562214_33.

Hou, Liwen, and David Smith. 2023. "Detecting Syntactic Change with Pre-Trained Transformer Models." In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3564–74. https://doi.org/10.18653/v1/2023.findings-emnlp.230.

Hsu, Bi-Min. 2020. "Comparison of Supervised Classification Models on Textual Data." *Mathematics* 8 (5): 851. https://doi.org/10.3390/math8050851.

Huang, Qianyu, and Tongfang Zhao. 2024. "Data Collection and Labeling Techniques for Machine Learning." *arXiv* preprint, June 19. https://doi.org/10.48550/arXiv.2407.12793.

Humphries, Mark, Lianne C. Leddy, Quinn Downton, et al. 2024. "Unlocking the Archives: Using Large Language Models to Transcribe Handwritten Historical Documents." *arXiv* preprint, November 2. https://doi.org/10.48550/arXiv.2411.03340.

Isozaki, Hideki, and Hideto Kazawa. 2002. "Efficient Support Vector Classifiers for Named Entity Recognition." In *Proceedings of the 19th International Conference on Computational Linguistics* 1: 1–7. https://doi.org/10.3115/1072228.1072282.

Jenset, Gard B., and Barbara McGillivray. 2017. *Quantitative Historical Linguistics: A Corpus Framework*. Oxford: Oxford University Press.

Joachims, Thorsten. 1998. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. https://doi.org/10.17877/DE290R-5097.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, 6282–93. https://doi.org/10.18653/v1/2020.acl-main.560.

Kanerva, Jenna, Cassandra Ledins, Siiri Käpyaho, and Filip Ginter. 2025. "OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches." *arXiv* preprint, February 3. https://doi.org/10.48550/arXiv.2502.01205.

Kiefer, Sebastian. 2022. "CaSE: Explaining Text Classifications by Fusion of Local Surrogate Explanation Models with Contextual and Semantic Knowledge." *Information Fusion* 77 (January): 184–95. https://doi.org/10.1016/j.inffus.2021.07.014.

Kim, Zae Myung, Anand Ramachandran, Farideh Tavazoee, Joo-Kyung Kim, Oleg Rokhlenko, and Dongyeop Kang. 2025. "Align to Structure: Aligning Large Language Models with Structural Information." *arXiv* preprint, April 4. https://doi.org/10.48550/arXiv.2504.03622.

Kolge, Bhagyashree, and Lavina Jadhav. 2018. "Comparison between Supervised Learning and Unsupervised Learning." *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 5 (6): 2157–9.

Kuzman, Taja, Igor Mozetič, and Nikola Ljubešić. 2023. "Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models." *Machine Learning and Knowledge Extraction* 5 (3): 1149–75. https://doi.org/10.3390/make5030059.

Kyröläinen, Aki-Juhani, and Veronika Laippala. 2023. "Predictive Keywords: Using Machine Learning to Explain Document Characteristics." *Frontiers in Artificial Intelligence* 5: 975729. https://doi.org/10.3389/frai.2022.975729.

Kytö, Merja. 2019. "Register in Historical Linguistics." *Register Studies* 1 (1): 136–67.

Lai, Kaijie. 2024. "Research on Key Technologies of Neural Machine Translation Based on Deep Learning." In *Proceedings of the 1st International Conference on Modern Logistics and Supply Chain Management*, 60–65. https://doi.org/10.5220/0013231400004558.

Laippala, Veronika, Jesse Egbert, Douglas Biber, and Aki-Juhani Kyröläinen. 2021. "Exploring the Role of Lexis and Grammar for the Stable Identification of Register in an Unrestricted Corpus of Web Documents." *Language Resources and Evaluation* 55: 757–88.

Laippala, Veronika, Samuel Rönnqvist, Miika Oinonen, et al. 2023. "Register Identification from the Unrestricted Open Web Using the Corpus of Online Registers of English." *Language Resources and Evaluation* 57 (3): 1045–79. https://doi.org/10.1007/s10579-022-09624-1.

Laippala, Veronika, Anna Salmela, Samuel Rönnqvist, Alham F. Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, Valtteri Skantsi, Lintang Sutawika, and Sampo Pyysalo. 2022. "Towards Better Structured and Less Noisy Web Data: OSCAR with Register Annotations." In P*roceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, 215–21. Gyeongju, Republic of Korea: Association for Computational Linguistics.

Larsen, Kyra, McKayla Lindman, Brett Hashimoto, Elizabeth Hanks, and Jesse Egbert. 2025. "A Register Approach to Specialized Word List Creation: Using Keywords to Supplement the Contracts Word List." *Register Studies*, ahead of print, May 2. https://doi.org/10.1075/rs.25004.lar.

Lebeña, Nuria, Alicia Pérez, and Arantza Casillas. 2024. "Quantifying Decision Support Level of Explainable Automatic Classification of Diagnoses in Spanish Medical Records." *Computers in Biology and Medicine* 182 (November): 109127. https://doi.org/10.1016/j.compbiomed.2024.109127.

Lee, David. 2002. "Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle." *Language Learning and Technology* 5. https://doi.org/10.64152/10125/44565.

Lee, Thomas R., and Jud Campbell Phillips. 2019. "Data-Driven Originalism." *University of Pennsylvania Law Review* 167 (2): 261–335.

Lehto, Anu. 2010. “Complexity in National Legislation of the Early Modern English Period.” *Journal of Historical Pragmatics* 11 (2): 277–300.

Li, Xinxin, Huiyao Chen, Chengjun Liu, et al. 2025. “LLMs Can Also Do Well! Breaking Barriers in Semantic Role Labeling via Large Language Models.” In *Findings of the Association for Computational Linguistics: ACL 2025*, 23162–23180, Vienna, Austria. Association for Computational Linguistics.

Li, Zewen, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. 2022. “A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects.” *IEEE Transactions on Neural Networks and Learning Systems* 33 (12): 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827.

Libovický, Jindřich, Rudolf Rosa, and Alexander Fraser. 2019. “How Language-Neutral Is Multilingual BERT?” *arXiv* preprint. http://arxiv.org/abs/1911.03310.

Liimatta, Aatu, Yann Ryan, Tanja Säily, and Mikko Tolonen. 2023. “Results from Rough Data? The Large-Scale Study of Early Modern Historiography with Multi-Dimensional Register Analysis.” *Digital Humanities in the Nordic and Baltic Countries Publications* 5 (1): 297–312. https://doi.org/10.5617/dhnbpub.10668.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” *arXiv* preprint. https://doi.org/10.48550/arXiv.1907.11692.

Lucia v. SEC, 585 U.S. ____ (2018).

Lundberg, Scott M., and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” *arXiv* preprint. November 25. https://doi.org/10.48550/arXiv.1705.07874.

Lyu, Zhiheng, Zhijing Jin, Fernando Gonzalez, Rada Mihalcea, Bernhard Schölkopf, and Mrinmaya Sachan. 2024. “On the Causal Nature of Sentiment Analysis.” *arXiv* preprint, April 17. http://arxiv.org/abs/2404.11055.

van der Maaten, Laurens, Eric Postma, and Jaap van den Herik. 2009. “Dimensionality Reduction: A Comparative Review.” Tilburg University Technical Report, TiCC-TR 2009-005. https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf.

Madhiarasan, Manogaran, and Mohamed Louzazni. 2022. “Analysis of Artificial Neural Network: Architecture, Types, and Forecasting Applications.” *Journal of Electrical and Computer Engineering* 2022 (April): 1–23. https://doi.org/10.1155/2022/5416722.

Mahadevan, Ananth, Michael Mathioudakis, Eetu Mäkelä, and Mikko Tolonen. 2025. “Text Reuse in Large Historical Corpora: Insights from the Optimization of a Data Science System.” *International Journal of Data Science and Analytics*, ahead of print, April 7. https://doi.org/10.1007/s41060-025-00742-x.

Manjavacas, Enrique, and Lauren Fonteyn. 2021. “LinguisticsMacBERTh: Development and Evaluation of a Historically Pre-Trained Language Model for English (1450–1950).” In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, 23–36.

Manjavacas, Enrique, and Lauren Fonteyn. 2022a. “Adapting vs. Pre-Training Language Models for Historical Languages.” *Journal of Data Mining and Digital Humanities*, jdmdh:9152. https://doi.org/10.46298/jdmdh.9152.

Manjavacas, Enrique, and Lauren Fonteyn. 2022b. “Non-Parametric Word Sense Disambiguation for Historical Languages.” In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, 123–34. Association for Computational Linguistics.

Màrquez, Lluís, Lluís Padró, and Horacio Rodríguez. 2000. “A Machine Learning Approach to POS Tagging.” *Machine Learning* 39: 59–91. https://doi.org/10.1023/A:1007673816718

Màrquez, Lluís, and Horacio Rodríguez. 1998. “Part-of-Speech Tagging Using Decision Trees.” In *Machine Learning: ECML-98*, edited by Claire Nédellec and Céline Rouveirol, Vol. 1398. Lecture Notes in Computer Science. Berlin: Springer. https://doi.org/10.1007/BFb0026668.

Martin, J. R., and David Rose. 2008. *Genre Relations: Mapping Culture*. London: Equinox.

Matthiessen, Christian M. I. M. 2019. “Register in Systemic Functional Linguistics.” *Register Studies* 1 (1): 10–41. https://doi.org/10.1075/rs.18010.mat.

McGillivray, Barbara. 2021. “Computational Methods for Semantic Analysis of Historical Texts.” In *Routledge International Handbook of Research Methods in Digital Humanities*, edited by Kristen Schuster and Stuart Dunn, 391–414. London & New York: Routledge.

Mikhaeil, Jonas M., Zahra Monfared, and Daniel Durstewitz. 2022. “On the Difficulty of Learning Chaotic Dynamics with RNNs.” *arXiv* preprint, October 6. https://doi.org/10.48550/arXiv.2110.07238.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1–12.

Milligan, Ian. 2022. *The Transformation of Historical Research in the Digital Age*. 1st ed. Cambridge: Cambridge University Press.

Miller, Carolyn R. 1984. “Genre as Social Action.” *Quarterly Journal of Speech*, 70 (2): 151–167. https://doi.org/10.1080/00335638409383686.

Miller, E., and H. Obelgoner. 2020. “Effective but Limited: A Corpus Linguistic Analysis of the Original Public Meaning of Executive Power.” *Georgia State Law Review* 36 (5): 607–64.

Miller, G. A., and W. G. Charles. 1991. “Contextual Correlates of Semantic Similarity.” *Language and Cognitive Processes* 6 (1): 1–28.

Mishev, Kostadin, Ss Cyril, Chaya Liebeskind, and Elena-Simona Apostol. 2023. “Validation of Language Agnostic Models for Discourse Marker Detection.” In *Proceedings of the 4th Conference on Language, Data and Knowledge*, 434–439, Vienna, Austria. NOVA CLUNL, Portugal.

Monaco, Leida Maria. 2017. *A Multidimensional Analysis of Late Modern English Scientific Texts from the Coruña Corpus*. Doctoral thesis, Universidade da Coruña. RUC – Repositorio Institucional da Universidade da Coruña. https://ruc.udc.es/entities/publication/b0fc199f-e82a-4850-8872-f3bcbfd6d9fc.

Murphy, Sean. 2019. “Shakespeare and His Contemporaries: Designing a Genre Classification Scheme for Early English Books Online 1560–1640.” *ICAME Journal* 43 (1): 59–82.

Nesi, Hilary, and Sheena Gardner. 2012. *Genres across the Disciplines: Student Writing in Higher Education*. Cambridge: Cambridge University Press.

Neumann, Stella, and Stefan Evert. 2021. “A Register Variation Perspective on Varieties of English.” In *Corpus-based Approaches to Register Variation*, edited by Elena Seoane and Douglas Biber, 143–78. Amsterdam: Benjamins.

Occhipinti, Annalisa, Louis Rogers, and Claudio Angione. 2022. “A Pipeline and Comparative Study of 12 Machine Learning Models for Text Classification.” *Expert Systems with Applications* 201 (September): 117193.

OpenAI, Josh Achiam, Steven Adler, et al. 2023. “GPT-4 Technical Report.” *arXiv* preprint, March 4. http://arxiv.org/abs/2303.08774.

Pal, Abhinandan, Francesco Ranzato, Caterina Urban, and Marco Zanella. 2022. “Abstract Interpretation-Based Feature Importance for SVMs.” *arXiv* preprint, October 22. https://doi.org/10.48550/arXiv.2210.12456.

Pan, Shirui, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. “Unifying Large Language Models and Knowledge Graphs: A Roadmap.” *IEEE Transactions on Knowledge and Data Engineering* 36 (7): 3580–99.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. “Thumbs Up?: Sentiment Classification Using Machine Learning Techniques.” In P*roceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP ’02)* 10: 79–86.

Paszke, Adam, Sam Gross, Soumith Chintala, et al. 2017. “Automatic Differentiation in PyTorch.” *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "Glove: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–43. https://doi.org/10.3115/v1/D14-1162.

Periti, Francesco, and Stefano Montanelli. 2024. "Lexical Semantic Change through Large Language Models: A Survey." *ACM Computing Surveys* 56 (11): 1–38.

Pescuma, Valentina N., Dina Serova, Julia Lukassek, et al. 2023. "Situating Language Register across the Ages, Languages, Modalities, and Cultural Aspects: Evidence from Complementary Methods." *Frontiers in Psychology* 13 (January): 964658. https://doi.org/10.3389/fpsyg.2022.964658.

Petrenz, Philipp, and Bonnie Webber. 2011. "Stable Classification of Text Genres." *Computational Linguistics* 37 (2): 385–93.

Piotrowski, Michael. 2012. *Natural Language Processing for Historical Texts*. San Rafael, CA: Morgan & Claypool.

Plank, Barbara, Anders Søgaard, and Yoav Goldberg. 2016. "Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 412–18.

Pritsos, D., and E. Stamatatos. 2018. "Open Set Evaluation of Web Genre Identification." *Language Resources and Evaluation* 52 (4): 949–68.

Rastas, Iiro, Yann Ciarán Ryan, Iiro Tiihonen, et al. 2022. "Explainable Publication Year Prediction of Eighteenth Century Texts with the BERT Model." In P*roceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, 68–77.

Regan, John. 2022. "Semantic Change and Knowledge Transmission: Some Case Studies in the Distribution of Words in ECCO." *Digital Scholarship in the Humanities* 37 (4): 1141–56.

Reid, T. B. W. 1956. "Linguistics, Structuralism, Philology." *Archivum Linguisticum* 8: 28–37.

Repo, Liina, Valtteri Skantsi, Samuel Rönnqvist, et al. 2021. "Beyond the English Web: Zero-Shot Cross-Lingual and Lightweight Monolingual Classification of Registers." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 183–191, Online. Association for Computational Linguistics.

Repo, Liina, Brett Hashimoto, and Veronika Laippala. 2023. "In Search of Founding Era Registers: Automatic Modelling of Registers from the Corpus of Founding Era American English." Digital Scholarship in the Humanities 38 (4): 1659–77.

Repo, Liina, Brett Hashimoto, and Veronika Laippala. Forthcoming. "From Unannotated to Annotated: The Impact of Text Internal Variation on Register Predictions of Long Historical Documents." International Journal of Corpus Linguistics.

Repo, Liina, Brett Hashimoto, Aatu Liimatta, Lassi Saario, Tanja Säily, Iiro Tiihonen, Mikko Tolonen, and Veronika Laippala. 2024. "Towards Automatic Register Classification in Unrestricted Databases of Historical English." In *Linguistics across Disciplinary Borders: The March of Data*, edited by Steven Coats and Veronika Laippala, 97–126. London: Bloomsbury Academic.

Ribeiro, Marco, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 97–101, San Diego, California. Association for Computational Linguistics.

Ribeiro, Marina, Bárbara Malcorra, Natália B. Mota, Rodrigo Wilkens, Lilian C. Hubner, and César Rennó-Costa. 2024. "A Methodology for Explainable Large Language Models with Integrated Gradients and Linguistic Analysis in Text Classification." *arXiv* preprint, September 30. https://doi.org/10.48550/arXiv.2410.00250.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. “A Primer in BERTology: What We Know about How BERT Works.” *Transactions of the Association for Computational Linguistics*, 8: 842–866.

Rose, David. 2023. “Genre, Register and Discourse in Systemic Functional Linguistics.” In *The Routledge Handbook of Discourse Analysis*, 2nd ed., edited by Michael Handford and James Paul Gee, 328–45. London: Routledge.

Rönnqvist, Samuel, Valtteri Skantsi, Miika Oinonen, and Veronika Laippala. 2021. “Multilingual and Zero-Shot Is Closing in on Monolingual Web Register Classification.” In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 157–65. Reykjavik, Iceland (Online): Linköping University Electronic Press.

Rönnqvist, Samuel, Aki-Juhani Kyröläinen, Amanda Myntti, Filip Ginter, and Veronika Laippala. 2022. “Explaining Classes through Stable Word Attributions.” *Findings of the Association for Computational Linguistics: ACL 2022*, 1063–74.

Rosson, David, Eetu Mäkelä, Ville Vaara, Ananth Mahadevan, Yann Ryan, and Mikko Tolonen. 2023. “Reception Reader: Exploring Text Reuse in Early Modern British Publications.” *Journal of Open Humanities Data* 9 (5): 1–11.

Ryan, Yann, Ananth Mahadevan, and Mikko Tolonen. 2023. “A Comparative Text Similarity Analysis of the Works of Bernard Mandeville.” *Digital Enlightenment Studies* 1 (1). https://doi.org/10.61147/des.6.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. “DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter.” *arXiv* preprint, March 1. https://doi.org/10.48550/arXiv.1910.01108.

Santini, Marina. 2007. “Characterizing Genres of Web Pages: Genre Hybridism and Individualization.” In *Proceedings of the Annual Hawaii International Conference on System Sciences*, January. https://doi.org/10.1109/HICSS.2007.124.

Sanyal, Soumya, and Xiang Ren. 2021. “Discretized Integrated Gradients for Explaining Language Models.” In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10285–99.

Schnabel, Tobias, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. “Evaluation Methods for Unsupervised Word Embeddings.” In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 298–307.

Schöffel, Matthias, Marinus Wiedner, Esteban Garces Arias, Paula Ruppert, Christian Heumann, and Matthias Aßenmacher. 2025. “Modern Models, Medieval Texts: A POS Tagging Study of Old Occitan.” *arXiv* preprint, March 10. https://doi.org/10.48550/arXiv.2503.07827.

Schweter, Stefan, Luisa März, Katharina Schmid, and Erion Çano. 2022. “hmBERT: Historical Multilingual Language Models for Named Entity Recognition.” *arXiv* preprint, July 1. https://doi.org/10.48550/arXiv.2205.15575.

Sharoff, Serge, Zhili Wu, and Katja Markert. 2010. “The Web Library of Babel: Evaluating Genre Collections.” In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta. ELRA.

Sharoff, Serge 2018. “Functional Text Dimensions for the Annotation of Web Corpora.” *Corpora* 13 (1): 65–95.

Shmueli, Galit. 2010. “To Explain or to Predict?” *Statistical Science* 25 (3): 289–310.

Skantsi, Valtteri, and Veronika Laippala. 2023. “Analyzing the Unrestricted Web: The Finnish Corpus of Online Registers.” *Nordic Journal of Linguistics* 48 (1): 1–31.

Spedding, Patrick. 2011. “‘The New Machine’: Discovering the Limits of ECCO.” *Eighteenth-Century Studies* 44 (4): 437–53.

Staples, Shelley, Jesse Egbert, Douglas Biber, and Bethany Gray. 2016. “Academic Writing Development at the University Level: Phrasal and Clausal Complexity across Level of Study, Discipline, and Genre.” *Written Communication* 33 (2): 149–83.

Sun, Aixin, Ee-Peng Lim, and Ying Liu. 2009. “On Strategies for Imbalanced Text Classification Using SVM: A Comparative Study.” *Decision Support Systems* 48 (1): 191–201.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. “Axiomatic Attribution for Deep Networks.” In *Proceedings of the 34th International Conference on Machine Learning (ICML)* 7: 5109–18.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. “Sequence to Sequence Learning with Neural Networks.” In P*roceedings of the 28th International Conference on Neural Information Processing Systems* - Volume 2 (NIPS’14), Vol. 2. MIT Press, Cambridge, MA, USA, 3104–3112.

Swales, John. 1990. *Genre Analysis: English in Academic and Research Setting*. Cambridge: Cambridge University Press.

Taavitsainen, Irma, Turo Hiltunen, Jeremy J. Smith, and Carla Suhr. 2022. *Genre in English Medical Writing, 1500–1820: Sociocultural Contexts of Production and Use*. Edited by Irma Taavitsainen. 1st ed. Cambridge: Cambridge University Press.

Tanase, Mircea-Adrian, Dumitru-Clementin Cercel, and Costin Chiru. 2020. “UPB at SemEval-2020 Task 12: Multilingual Offensive Language Detection on Social Media by Fine-Tuning a Variety of BERT-Based Models.” In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2222–31.

Tekgurler, Merve. 2025. “LLMs for Translation: Historical, Low-Resourced Languages and Contemporary AI Models.” *arXiv* preprint, March 14. https://doi.org/10.48550/arXiv.2503.11898.

Tiedemann, Jörg, and Zeljko Agić. 2016. “Synthetic Treebanking for Cross-Lingual Dependency Parsing.” *Journal of Artificial Intelligence Research* 55 (January): 209–48.

Tiihonen, Iiro, Aatu Liimatta, Lidia Pivovarova, Tanja Säily, and Mikko Tolonen. 2023. “Measuring the Distribution of Hume’s Scotticisms in the ECCO Collection.” In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, 36–44, Tokyo, Japan. Association for Computational Linguistics.

Tolonen, Mikko, Eetu Mäkelä, Ali Ijaz, and Leo Lahti. 2021. “Corpus Linguistics and Eighteenth Century Collections Online (ECCO).” *Research in Corpus Linguistics* 9 (1): 19–34.

Tolonen, Mikko, Eetu Mäkelä, and Leo Lahti. 2022. “The Anatomy of Eighteenth Century Collections Online (ECCO).” *Eighteenth-Century Studies* 56 (1): 95–123.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, et al. 2023. “LLaMA: Open and Efficient Foundation Language Models.” *arXiv* preprint, February 27. https://doi.org/10.48550/arXiv.2302.13971.

Vapnik, Vladimir. 1998. *Statistical Learning Theory*. New York: Wiley.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2017. “Attention Is All You Need.” *arXiv* preprint, June 12. https://doi.org/10.48550/arXiv.1706.03762.

Volk, Martin, Dominic Philipp Fischer, Lukas Fischer, Patricia Scheurer, and Phillip Benjamin Ströbel. 2024. “LLM-Based Machine Translation and Summarization for Latin.” In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, 122–128, Torino, Italia. ELRA and ICCL.

Van De Voorde, Iris, Gijsbert Rutten, Rik Vosters, Marijke van der Wal, and Wim Vandenbussche. 2023. “Historical Corpus of Dutch: A New Multi-Genre Corpus of Early and Late Modern Dutch.” *Taal en Tongval* 75 (1): 114–32.

Weeds, Julie, David Weir, and Diana McCarthy. 2004. “Characterising Measures of Lexical Distributional Similarity.” In *Proceedings of the 20th International Conference on Computational Linguistics (COLING ’04)*, 1015–22.

Wendorf, Richard. 2022. *Printing History and Cultural Change: Fashioning the Modern English Text in Eighteenth-Century Britain*. Oxford: Oxford University Press.

Wolf, Thomas, Lysandre Debut, Victor Sanh, et al. 2020. “Transformers: State-of-the-Art Natural Language Processing.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45, Online. Association for Computational Linguistics.

Worsham, Joseph, and Jugal Kalita. 2018. “Genre Identification and the Compositional Effect of Genre in Literature.” In *Proceedings of the 27th International Conference on Computational Linguistics*, 1963–1973, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Wright, Susan. 1994. “The Place of Genre in the Corpus.” In *Corpora Across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora*, edited by M. Kytö, M. Rissanen, and S. Wright, 101–7. Amsterdam: Rodopi.

Wright v. Spaulding. No. 17–4257. (6th Cir. 2019).

Xie, Wenxiu, Christine Ji, Tianyong Hao, and Chi-Yin Chow. 2021. “Predicting the Easiness and Complexity of English Health Materials for International Tertiary Students with Linguistically Enhanced Machine Learning Algorithms: Development and Validation Study.” *JMIR Medical Informatics* 9 (10): e25110.

Xu, Hu, Bing Liu, Lei Shu, and Philip Yu. 2019. “BERT Post-Training for Review Reading Comprehension and Aspect-Based Sentiment Analysis.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Yang, Soyoung, Minseok Choi, Youngwoo Cho, and Jaegul Choo. 2023. “HistRED: A Historical Document-Level Relation Extraction Dataset.” In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3207–24. Toronto: ACL.

Yannakoudakis, Helen, Ted Briscoe, and Ben Medlock. 2011. “A New Dataset and Method for Automatically Grading ESOL Texts.” In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Zeng, Yifan. 2024. “HistoLens: An LLM-Powered Framework for Multi-Layered Analysis of Historical Texts – A Case Application of Yantie Lun.” *arXiv* preprint, November 15. https://doi.org/10.48550/arXiv.2411.09978.

Zhang, Yin, Rong Jin, and Zhi-Hua Zhou. 2010. “Understanding Bag-of-Words Model: A Statistical Framework.” *International Journal of Machine Learning and Cybernetics* 1 (1–4): 43–52.

Zhang, K., L. Feng, and X. Yu. 2023. “Shap-PreBiNT: A Sentiment Analysis Model Based on Optimized Transformer.” In *Web and Big Data. APWeb-WAIM 2022. Lecture Notes in Computer Science*, vol. 13422. Cham: Springer.

Zhang, Yuanchi, Yile Wang, Zijun Liu, et al. 2024. “Enhancing Multilingual Capabilities of Large Language Models through Self-Distillation from Resource-Rich Languages.” In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11189–204.

Zhao, Wei, Tarun Joshi, Vijayan N. Nair, and Agus Sudjianto. 2020. “SHAP Values for Explaining CNN-Based Text Classification Models.” *arXiv* preprint, August 26. https://doi.org/10.48550/arXiv.2008.11825.

Declaration of AI Use

In preparing this dissertation, generative artificial intelligence (AI) features in Grammarly and ChatGPT (3.5 and 4.0) were used for improving spelling, grammar, and readability. Concensus: AI Search Engine for Research and arXiv Xplorer were used to locate relevant research articles, which were then verified through manual review of the original source materials. The author has personally reviewed all materials and assumes full responsibility for the accuracy and integrity of the content.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2026

ISBN 978-952-02-0514-0 (PRINT)
ISBN 978-952-02-0515-7 (PDF)
ISSN 0082-6987 (Print)
ISSN 2343-3191 (Online)