Digital Scholarship in the Humanities, 2023, **00**, 1–19
https://doi.org/10.1093/llc/fqad049
Full Paper

In search of founding era registers: automatic modeling of registers from the corpus of Founding Era American English

Liina Repo[1,*], Brett Hashimoto [2], Veronika Laippala[1]

[1]School of Languages and Translation Studies, University of Turku, FI-20014 University of Turku, Finland
[2]Department of Linguistics, Brigham Young University, Provo, UT 84602, USA

*Corresponding author. School of Languages and Translation Studies, University of Turku, Arcanuminkuja 1, 20500, Turku, Finland. E-mail: tlkrep@utu.fi

Abstract

Registers are situationally defined text varieties, such as letters, essays, or news articles, that are considered to be one of the most important predictors of linguistic variation. Often historical databases of language lack register information, which could greatly enhance their usability (e.g. Early English Books Online). This article examines register variation in Late Modern English and automatic register identification in historical corpora. We model register variation in the corpus of Founding Era American English (COFEA) and develop machine-learning methods for automatic register identification in COFEA. We also extract and analyze the most significant grammatical characteristics estimated by the classifier for the best-predicted registers and found that letters and journals in the 1700s were characterized by informational density. The chosen method enables us to learn more about registers in the Founding Era. We show that some registers can be reliably identified from COFEA, the best overall performance achieved by the deep learning model Bidirectional Encoder Representations from Transformers with an F1-score of 97 per cent. This suggests that deep learning models could be utilized in other studies concerned with historical language and its automatic classification.

Keywords: Late Modern English; register; text classification; historical natural language processing; BERT.

1. Introduction

Historical language is often studied with the aid of large corpora. For instance, the Early English Books Online (EEBO, https://www.proquest.com/eebo) database of early printed works from 1473 to 1700 claims to include 'almost every work printed—[f]rom the first book printed in English through to the ages of Spencer and Shakespeare and of the English Civil War' (ProQuest 2021). Large corpora offer an important source for studying language use in a historical context but, as Mouritsen (2010) notes, they can be equally useful for legal scholars interpreting legal texts. Large historical corpora provide useful sources, for instance, for judges and legal scholars who want to reveal how a word has been used in a legal text at the time it was written, that is, the 'original meaning' of a word (see Slocum 2015). Corpus linguistics can aid the identification of the most frequent senses of words which can give insight into the typical use of a word.

Although corpus linguistics has been used in legal interpretation, the lack of corpora suited for this has hindered the process of adopting corpus linguistics as a commonly used tool in legal studies, as Hashimoto (2023) notes. One such corpus is the corpus of Founding Era American English (COFEA). COFEA is useful for legal scholars as it represents language from the late 1700s and much of foundational American legislation was written during that time. COFEA has already been used in the legal community to study the original meaning of the Constitution and other Founding Era statutory laws. Various terms and parts of the Constitution have been investigated, for example, in court cases (e.g. Carpenter v. United States 2018) and law review articles (Barclay, Earley, and Boone 2019; Baron 2019; Cunningham and Egbert 2020). For a list of cases and articles that have used COFEA thus far, see Hashimoto (2023).

Databases of historical documents often lack detailed metadata, such as register. Registers are text varieties that are associated with particular situations of use and communicative purposes (Biber and Conrad 2019). A register, such as a news article or a recipe, is thus described by its situational context and characterized by the linguistic features typical of the register. Situational context and variation refer to the ways in which linguistic characteristics of language are shaped by the communicative situation in which they are used. These include parameters such as relations among participants (e.g. addressor, addressee), mode (e.g. speech, writing), medium (e.g. printed, handwritten, radio), and communicative purpose (e.g. entertaining, persuading, informing) of the text. The linguistic features are functionally motivated, pervasive lexical and grammatical characteristics that appear throughout the text and serve important communicative functions in the register (Biber and Conrad 2019). Register features are also distributional. The features are not unique but can appear in any register, although they are substantially more common within the target register (Biber and Conrad 2019). For registers in academic settings, see, for example, Conrad (1996); in historical settings, see, for example, Taavitsainen and Pahta (2004), and on the web, see, for example, Biber and Egbert (2016).

The relevance of registers has been acknowledged in historical linguistics and in relation to historical corpora (Wright 1994). Certainly, even minor register differences can greatly affect language use (including word meaning) (see, e.g. Biber and Gray 2013). Register information would considerably enhance the usability of many databases and corpora, especially in corpus linguistics, as registers are considered as one of the most important predictors of linguistic variation and affect the way texts are interpreted (see, e.g. Biber 2012). The register distribution of a historical corpus can be used in analyzing frequencies of linguistic variant forms across registers and time, showing possible change. Many corpora that are currently available have only limited information on the extralinguistic features of texts or registers and are in need of grammatical or other annotation (Kytö 2019).

This study aims to examine register variation in Late Modern English and develop machine-learning methods for identifying registers in historical corpora, specifically in COFEA that presently lacks register information. In this study, we (1) explore the possibility of automatically identifying registers from the whole corpus with a sub-corpus of COFEA and (2) examine the registers in COFEA by analyzing the most important discriminative features estimated by a classifier for the registers. This allows us to gain insights into the linguistic characteristics of the registers and the functioning of the classifier. Adding register information to COFEA will increase its applicability and validity, providing particular support to the generalizability of the studies based on COFEA.

Automatic register identification is a type of text classification task where a classifier builds a classification model based on manually annotated examples, which can then be used to identify the register of new texts (Sebastiani 2002; Argamon 2019). This kind of automatic register identification is an example of supervised machine-learning, where algorithms learn to associate example data and its corresponding target responses, such as classes or tags, and can then predict the correct response with new and unseen data. Text classification is one of the most prominent tasks in natural language processing (NLP) and a notable problem in information, computer, and library science. Previous studies on automatic register identification have applied various machine-learning methods, such as support vector machines (SVMs; see, e.g. Sharoff, Wu, and Markert 2010; Pritsos and Stamatatos 2018) and different deep neural networks (see, e.g. Worsham and Kalita 2018; Sharoff 2021).

Deep learning methods utilize deep neural network architectures to build language models that have been pre-trained on large amounts of unlabeled text. The field of NLP has been greatly advanced by the use of pre-trained deep learning language models in tasks like recognition of hate speech (Mishra and Mishra 2019) and machine translation (Bahdanau, Cho, and Bengio 2015). Deep learning classifiers generally offer high performance, whereas linear classifiers are generally more transparent and their decisions are easier to understand (Linardatos, Papastefanopoulos, and Kotsiantis 2021). Furthermore, it is not certain how deep learning classifiers perform with historical language, as they have usually been trained with modern language. We train deep learning models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019) and XLM-RoBERTa (XLM-R) (Conneau et al. 2019), as well as linear classifiers, such as SVMs (Vapnik 1998) in order to predict registers in COFEA and to examine their linguistic characteristics. By understanding the linguistic basis of the classifier's decisions, we can gain insights into the features that distinguish each register, which can aid in the development of more efficient classifiers.

This article is divided in the following way: Section 2 provides an overview of previous work in register classification, focusing first on early and late modern registers, and then on automatic identification of registers with machine-learning methods. Section 3 is concerned with the material and methods used for this study. Section 4 introduces the findings of the study by presenting classification results and linguistic analysis of

the best-predicted registers. Section 5 concludes this article.

2. Previous studies on register modeling

Registers can be analyzed by their linguistic characteristics and these characteristics can be used to model and automatically identify registers. This section begins by introducing previous studies on early and late modern registers, and then moves on to discuss automatic register identification with traditional machine-learning and deep learning models, tying these to the present study in Section 2.3.

2.1 Studies on early and late modern registers

The concept of register is relevant in historical corpus linguistics, as register is a powerful factor in shaping language use and linguistic and stylistic change (Kytö 2019). The increasing interest in language variation and advances in corpus linguistics after the 1980s have provided a great deal of empirical research on patterns of linguistic variation with the aid of large electronic corpora. In historical linguistics, registers are often considered as possible agents of change, but research has also focused on registers themselves as the object of study. Focusing only on specific registers has been proposed to give better insights into historical language use than trying to represent an entire historical language variety at a given time (Kytö 2019). Although registers and genres are considered to be important across history, Early and Late Modern English registers are especially rewarding objects of study as registers begin to diversify and become better represented from the 16th century onward (Kytö 2019).

The introduction of the multidimensional analysis (MDA) approach (Biber 1988) marked an important advance in historical register studies, paving the way for a novel way of studying historical register variation. Biber (1988) studied the cooccurrence patterns of several linguistic features serving as the foundation of dimensions of variation that were interpreted to reflect different functions of communication in speech and writing. Biber and Finegan (1989) empirically identified three dimensions of linguistic variation that highlight the differences of literate and oral registers. MDA (and other similar statistical methods) have then been applied to study register characteristics at a specific time in different languages and discourse domains. MDA studies have been applied to inspect the trends of development, especially in English and varieties of English, but also the historical development of other languages. For instance, González-Álvarez and Pérez-Guerra (1998) studied the evolution of five English genres from the 15th to the 17th centuries with MDA, finding that science and drama evolve more oral characteristics, whereas education and fiction shift toward literate characterization, while letters do not show any significant change. Geisler (2002) inspected the development of English registers throughout the 19th century, the results showing that generally there is a dichotomy between non-expository and expository registers: the non-expository registers, Fiction, Letters, and Drama, become less abstract, while the expository register, History, becomes more narrative and elaborated. Both studies by González-Álvarez and Pérez-Guerra (1998) and Geisler (2002) then compared their results with those obtained by Biber and Finegan (1989, 1992, 1997) for the same genres from later centuries. Taavitsainen (1997) studied Early Modern English genres and text types and inspected personal affect, the expression of attitudes, feelings, and emotions, in fiction and non-fiction texts with corpus-based statistical analyses, and found four different qualities of personal affect that were all characteristic of early fiction. For a comprehensive list of MDA studies, historical and other, see, for example, Biber and Conrad (2019).

Although studies focusing on NLP and machine-learning methods with historical data are sparse, the 'historicization' of NLP (Hosseini et al. 2021) has been of interest in recent research, and machine-learning methods have also been used in modeling and predicting registers from historical texts. For example, Taavitsainen and Schneider (2019) traced scholastic thought style in Middle English as well as Early and Late Modern English medical texts by using logistic regression as a text classification method. They concluded that their chosen method proved fruitful in detecting scholastic features and that scholasticism lasted for centuries, although the development of scholastic tradition underwent some changes. Machine-learning methods can provide valuable tools for studying historical texts, and we aim to follow Taavitsainen and Schneider (2019) in promoting the use of machine-learning methods in historical register classification. Automatic identification of registers from historical data is still a rather understudied area and in need of suited historical corpora and information on how machine-learning models deal with historical language. The next section aims to give an overview on how machine learning has been used in register identification tasks with modern language data.

2.2 Automatic register identification with machine learning

This section discusses previous studies on automatic register identification with machine-learning models.

In many areas, such as the legal industry, text categorization is still often done manually by human experts (Wan et al. 2019). As text categorization is

very time and resource consuming, automatic text classification methods would reduce human effort considerably. In recent years, machine-learning methods, especially deep neural networks, have achieved remarkable results on text classification tasks. Researchers are continuously improving existing classification methods and developing new ones to achieve higher computational efficiency.

2.2.1 Traditional supervised machine learning

Machine learning is an application of artificial intelligence where algorithms have the ability to automatically learn by experience. This article is concerned with supervised machine learning which refers to machine-learning models trained with labeled data, which enables the model to have an idea of the relationship between input and output and extend its knowledge to new and unseen data. The term traditional machine learning is used to distinguish these models from deep learning models that are introduced in the next section.

Traditional supervised learning algorithms include, for instance, SVMs (Vapnik 1998) that can be used for classification and regression analysis. In a classification situation, an SVM is used to predict whether something belongs to a certain class. The objective of the classification algorithm is to find an optimal separating hyperplane that distinctly divides the data set into classes. The algorithm uses support vectors, which are datapoints nearest to the separating hyperplane from each class, to determine the optimal hyperplane. The optimal hyperplane represents largest separation between classes and maximizes the distance between the hyperplane and the support vectors of each class.

SVMs have been used in various register identification tasks, such as web genre identification (see, e.g. Pritsos and Stamatatos 2018), and in text classification in general (Joachims 1998). Linear SVMs can be used to estimate and rank the importance of a feature to the model performance (Guyon and Elisseeff 2003) which enables the evaluation of discriminative significance of different linguistic features for different registers (Kyröläinen and Laippala 2023). Each feature is assigned a weight, which is used in the model's decision function to predict the class label. Being able to identify the most important features and their contributions to the model performance allows us to understand the linguistic characteristics of the data and to assess the discriminative significance of different linguistic features for different registers, as demonstrated by Laippala et al. (2021) in their study of automatic register identification on the unrestricted web.

2.2.2 Deep learning models

Deep learning is a specialized type of machine learning where algorithms analyze data with artificial neural networks (ANNs) that are layered structures of algorithms inspired by the biological neural network of human brains. This leads to a more capable learning process than that of traditional machine-learning processes introduced in the previous section.

Deep learning language models are based on ANN architectures and are trained on a large corpus of unannotated text. These models can then be fine-tuned to suit specific data and tasks, such as text classification or register identification. The use of pre-trained deep learning models has advanced many NLP tasks relating to text classification in recent years, such as sentiment analysis (Li et al. 2019) and various natural language understanding tasks (Devlin et al. 2019). Below we present two deep learning language models used in this study, namely BERT and XLM-R, and discuss previous studies on automatic register identification with deep learning.

BERT, (Devlin et al. 2019) is a state-of-the-art deep-learning language representation model for English. BERT's architecture is a multi-layer bidirectional Transformer encoder that is based on the original Transformer architecture (Vaswani *et al.* 2017), which is an attention mechanism that learns contextual relations between words in a text. BERT has several Transformers stacked into layers: twelve layers in the base model and twenty-four layers in the large model. BERT is pre-trained on a large corpus consisting of English Wikipedia texts (2,500 M words) and the Books Corpus (800 M words) (Zhu et al. 2015). BERT has 30,000 token vocabularies that consist of word pieces that represent words or other common character sequences, such as prefixes or affixes. BERT is capable of working with a maximum of 512 tokens at a time, which is related to memory limitations of the Transformer architecture, which may affect register identification of longer texts as the model is not able to access the entire text at once. BERT is designed to pre-train bidirectional representations from unlabeled text. Instead of simply analyzing from left-to-right or right-to-left like most text analyzing models, the Transformer encoder is used to analyze text from both directions by randomly masking 15 per cent of the word piece tokens in a sequence and predicting the masked tokens by analyzing the surrounding text with the help of Transformers. When published, BERT outperformed the state-of-the-art systems of that time on many benchmark datasets (Devlin et al. 2019) and has received much attention since; there is now even a field of research relating to BERT (BERTology; Rogers, Kovaleva, and Rumshisky 2020).

In addition to BERT, we use XLM-R (Conneau et al. 2019), which is a multilingual language model based on the Cross-lingual Language Model (Lample and Conneau 2018) and the RoBERTa model (Liu et al.

2019), which, in turn, shares the same architecture as BERT. XLM-R is pre-trained on a massive multilingual corpus comprising 2.5 TB of filtered Common Crawl data (Wenzek et al. 2020) on monolingual texts in 100 languages. Because XLM-R has been shown to outperform monolingual baselines that rely on pre-trained models, as well as multilingual BERT (Conneau et al. 2019; Libovický, Rosa, and Fraser 2019; Tanase, Cercel, and Chiru 2020), we chose the model as a point of reference. Additionally, our data include small amounts of text in other languages, which facilitate the choice of XLM-R.

Deep learning models have brought great advances on register identification from large corpora. For instance, Laippala et al. (2019) presented an online register corpus on Finnish and demonstrated that web registers can be identified in a cross-lingual setting. Repo et al. (2021) extended the possibilities of web register identification and inspected cross-lingual transfer on French, Finnish, Swedish, and English web registers with BERT and XLM-R, among others, and showed that registers can be identified reliably even when the training and testing languages are different.

2.3 The present study

While there have been many MDA studies on Early and Late Modern English registers, as outlined in Section 2.1, these studies have been inspecting manually annotated registers and register dimensions. The purpose of the present study is to examine register variation in Late Modern English by exploring automatic register identification in historical corpora, specifically in COFEA, and by examining the most important linguistic characteristics of the registers as identified by the classifier. COFEA is a register diversified corpus that also exhibits situational variation. Differences in situational context lead to differences in the linguistic features of the registers, such as the use of specialized vocabulary, different types of sentence structures, and specific discourse organization. We expect the situational variation to correspond to registerial variation, which provides us with a resource for exploring to what extent the register classes can be automatically identified and what kinds of linguistic characteristics are typical of them.

An SVM allows us to assess the discriminative significance of different linguistic features, as mentioned in Section 2.2.1. This is important in our study, as our aim is also to examine the registers in the corpus and the role of different linguistic features. We train a linear classifier and extract the grammatical characteristics estimated by the classifier. We then analyze the most important grammatical characteristics estimated for each register. We hope to validate our methods by looking at evaluation measures and classification scores, together with the inspection of the classifier prediction mistakes, as well as the linguistic characteristics of the registers estimated by the classifier for the registers.

3. Material and methods

This section discusses the COFEA. We present a subcorpus of COFEA used as material in the present study and describe the machine-learning methods used for register classification. Additionally, we outline the methods of analysis used to evaluate the classification methods and detail the linguistic analysis of the classified data.

3.1 Corpus of Founding Era American English

Currently, the COFEA consists of 127,840 texts and over 140 million words (https://lawcorpus.byu.edu/cofea). COFEA is a historical corpus intended for scholars as a source for studying language use in the Late Modern English period leading up to the ratification of the Constitution. COFEA was created primarily by originalist and textualist legal and linguistic researchers for the purpose of investigating constitutional interpretation issues. Data in the corpus were selected to be representative of registers that were relevant for this purpose. Although COFEA aims to represent written American English from the Founding Era 1760–1799, it should be noted that the corpus is not fully representative of all aspects of language use due to structural biases in the society and the preservation of historical texts. For more details, see Hashimoto (2023).

COFEA contains texts from America's Founding Fathers, legal sources as well as from ordinary people of the day, and it includes texts from various registers, such as letters, newspapers, personal records, fiction, non-fiction books, sermons, debates, legal cases, and other legal material. The texts in COFEA come from seven different sources that are presented in Fig. 1 together with the number of words in these source databases.

3.2 Data in the present study

The data used in the present study are sub-corpus of COFEA, comprising 112,107 texts from the Founders Online dataset. The dataset has been classified with hand-coded annotations on seventeen register categories listed in Table 1. In addition to these register categories, there are also two technical categories that are *Not English* and *No Text*, that include texts written in other languages than English (mainly French) and empty texts or texts that do not include actual content but rather metatext from the original source database. *Undetermined* represents texts that did not fit into the

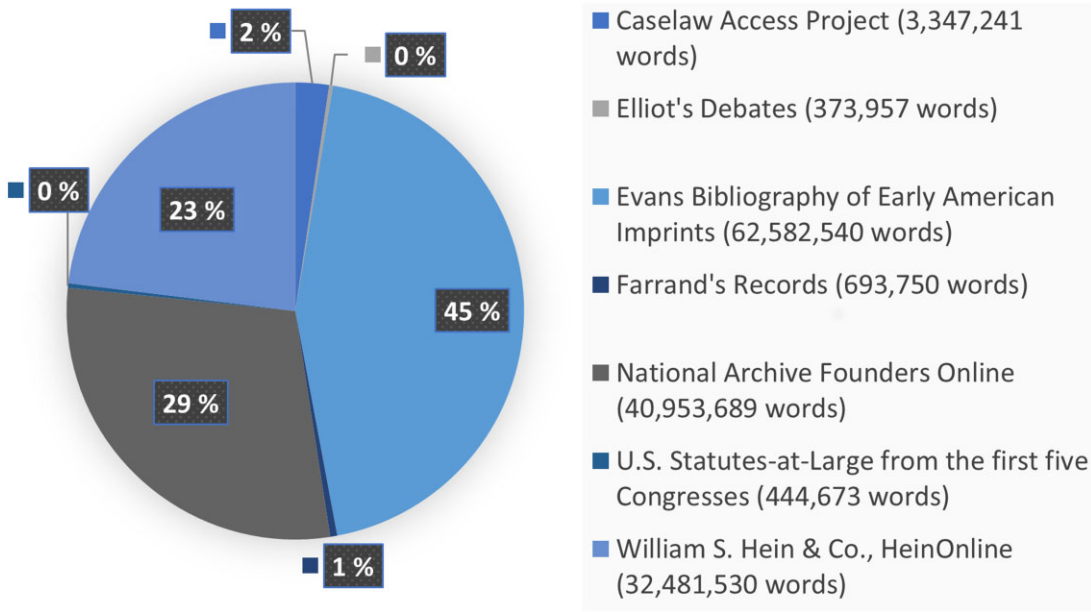

Figure 1. The distribution of words in the source databases forming COFEA.

Table 1. Distribution of register classes, number of words in each class, and mean text length in words.

Register	Number of texts	Number of words	Mean text length	Register	Number of texts	Number of words	Mean text length
Letter	92,671	33,167,014	366	Contract	297	127,567	442
Not English	5,308	1,630,626	314	Cases	233	218,318	956
Personal Record	5,147	1,441,178	303	Speech	149	163,668	1,130
No text	3,716	703,016	202	Laws and statutes	111	111,828	1,046
Proceedings	1,423	902,379	656	Pamphlet	90	240,669	2,827
Undetermined	1,357	814,319	607	Debate	55	40,310	748
Essay	430	939,840	2,194	Treaty	44	99,546	2,163
Proposal	400	327,525	843	Non-narrative non-fiction book	12	15,302	1,278
Book	338	178,783	538	Narrative non-fiction book	10	5,597	563
Newspaper	316	377,040	1,214				
				All	112,107	41,504,525	968

other register categories, such as poems, epitaphs, and songs. The data include both personal texts, such as letters and diary entries, and less personal texts, such as books and pamphlets, as well as more formal legal texts, such as treaties and legal cases. The register classes follow established conventions and are typical for a corpus of historical language [see, for instance, the register classes in ARCHER—A Representative Corpus of Historical English Registers; Biber, Finegan, and Atkinson (1994); https://www.projects.alc.manchester.ac.uk/archer/].

Register categories for COFEA were created by undergoing three main stages: (1) category generation and refinement, (2) hand annotation, and (3) accuracy checking. The register categories were generated by iteratively subsampling from the corpus and defining labels based on the observed texts. A complete list of the register categories and the codebook for training the human coders in identifying these categories can be found at Hashimoto (n.d.). The Founders Online texts were then sent to two trained, independent coders who annotated the text for its register category. If there was disagreement from the two raters, a third rater was brought in to break the tie. Register categorization accuracy was also manually checked by one of the research team for texts where the coders' annotations did not agree, or there were other uncertainties. The categories arose out of iterative discussions by a group of

legal and linguistic scholars for the long-term purpose of legal interpretation.

Table 1 presents the distribution of the register classes in the dataset.

As can be seen from Table 1, by far the most frequent register is *Letter*, which covers 84 per cent of the texts with over 92,000 texts, while the least frequent registers are *Debate*, *Treaty*, *Non-Narrative Non-Fiction Book*, and *Narrative Non-Fiction Book* with only fifty-five, forty-four, twelve, and ten texts, respectively. Table 1 also shows the mean length of texts in each register class. On average, the number of words in a single text is 968, while some registers, such as *No Text* (202 words) or *Personal Record* (303 words), have much shorter texts and some registers such as *Pamphlet* (2,827 words) and *Essay* (2,194 words) have much longer texts. The corpus features texts of varying length as the shortest texts in the data were under ten words, whereas the longest texts had over 34,000 words. The uneven number of texts in each register class is related to their distribution in the original datasets and the annotation process of the corpus which did not entail a sampling procedure aimed at achieving a balanced dataset.

3.3 Experimental setup

This section presents the practical implementation of the study and the experimental setup of the models introduced in Section 2.2. We use two different types of classifiers: two deep learning models, BERT and XLM-R, and a traditional machine-learning model SVM.

For the traditional machine-learning method SVM, we used three different feature sets representing lexical and grammatical information in the experiments. First, we used lexical features that consist of words as they appear in the texts. Second, we used grammatical tags associated with each word, and third, we used a combination of the first two, treating the words and the associated grammatical tags as separate features. We study how well registers can be predicted from the corpus with these different feature sets. Best results on register identification have been reached with word-based features, such as bag-of-words and word trigrams (Pritsos and Stamatatos 2018) or character four-grams as binary features (Sharoff, Wu, and Markert 2010). The predictive power of other than textual elements, such as html markup, image count, or links count, has also been explored in, for example, Lim, Lee, and Kim (2005) and Levering, Cutler, and Yu (2008). Grammatical features have also been used in automatic register identification, though the performance of models trained only on grammatical information tends to be poorer (see Petrenz and Webber 2011). Grammatical tags, however, do not reflect the topic of the text but rather the genre or register the text represents (Finn and Kushmerick 2006), and are considered to have register specific functional associations (Biber 1988; Biber and Conrad 2019). Grammatical features are widely recognized as essential components of registers and form the basis of register studies. Laippala et al. (2021) show that adding grammatical information increases the stability of the classifier and that the importance of grammar varies across registers.

For the analysis of grammatical characteristics of the registers, the Northern Arizona University (NAU) Tagger was used to generate a grammatical tag for each word in the data. The grammatical tags created by the tagger follow Biber's (1988) work, where these specific tags have been deemed important for registerial variation and are often used in register studies. The grammatical information provided by the tagger includes part-of-speech labels as well as other information, such as person, number, and semantic domain. Table 2 provides an example of the grammatical tags created by the tagger. For the classification with grammatical features, the tags were represented as n-grams, with all possible n-grams generated from a sequence of tags. For instance, the word 'be' in Table 2 would be represented with six separate n-gram features: monograms VL, INF, and COP, bigrams VL-INF and INF-COP, and a trigram VL-INF-COP. The grammatical tags were used as classifier features to gain insights into the linguistic characteristics of the registers in COFEA.

Before the classification tasks, the data were preprocessed by removing punctuation marks, duplicate texts, and empty texts. Additionally, some registers were excluded from the analysis: *Non-Narrative Non-Fiction Books* and *Narrative Non-Fiction Books* due to the small number of texts, and *Undetermined* due to unclear register labels.

For the deep learning models, the data were divided into training (70 per cent), development (10 per cent), and test (20 per cent) data with stratified random sampling in order to ensure equal class distribution between the datasets. We used large, cased models (where the letter case of the word is preserved) of BERT and XLM-R (TensorFlow versions) through the Huggingface Transformers library (Wolf et al. 2020). For BERT and XLM-R, the maximum sequence length of 512 tokens was used, meaning that texts longer than the maximum sequence length were truncated at the end. To find the best hyperparameters (i.e. parameters, whose value is manually set prior to the learning process), we tested different hyperparameter combinations that follow the recommendations by Devlin et al. (2019) and our own pilot experiments. We used a grid search, which iterates through every combination of the specified parameters, on learning rate (8e−6–6e−5), which controls the model weights, and the number of training epochs (three to seven), which

Table 2. An example of grammatical tags by the NAU Tagger.

Word	Part of speech	Pronoun type/ verb category	Number/ syntactic role	Person/ valency	Type of pronoun	Semantic domain
I	P (pronoun)	PER (personal)	SING (singular)	1 (first)	S (subject)	
shall	VM (modal verb)					PRDN (prediction)
be	VL (lexical verb)	INF (infinitive)		COP (copular)		
glad	J (adjective)		PRDV (predicative)			

determines the number of times the entire input data are worked through. Due to available GPU memory, we used a fixed batch size of seven (the number of training examples to work through before updating the model internal parameters). The development data were used for model development and hyperparameter optimization, whereas the test data were only used for the final experiments.

The SVM in the present study was implemented in Python's Scikit-Learn library (http://scikit-learn.org/; Pedregosa et al. 2011). We trained a linear SVM to predict the register classes based on the lexical and grammatical features, and a combination of both, as discussed above. For each feature set, the classifier was run fifty times with the best regularization parameter *C* value determined with grid search. This parameter controls the balance between the model's accuracy and generalizability. To avoid overfitting, we used the default L2 penalty parameter that controls the model complexity. Following Kyröläinen and Laippala (2023), we used a minimum document frequency of 0.05, that is, we ignore terms that appear in less than 5 per cent of the documents, to minimize data sparsity. Between every run the data were randomly divided into training (80 per cent) and testing (20 per cent) set. During each of the runs, the top 200 positive features for each register were recorded. The aim of this set-up is to evaluate the stability of the features estimated by the model. The fifty sampling rounds enabled the estimation of selection frequency of the features, which refers to how often a feature was ranked among the top features throughout the rounds. This offers a way to inspect how well a feature represents a stable property of a certain register, as Laippala et al. (2021) note. A secondary aim is to measure the variation in the model's performance, which is often overlooked in classification studies and more difficult to calculate with more traditional techniques, such as cross-validation.

All classification tasks mentioned above were performed as multi-class text classification, where each text can have only one register label. Three classification metrics were applied to quantify the performance of the models: precision, recall, and F1-score. Precision indicates what proportion of positive identifications was actually correct, whereas recall tells what proportion of correct actual positives was correctly identified. F1-score is the balanced and harmonic mean of the first two measures.

4. Classification results and analysis

In this section, we begin by presenting the classification results for the different models and register specific classification scores in Section 4.1. We then inspect the classification mistakes in order to gain insight into the performance of the models. Section 4.2 inspects the SVM model qualitatively through inspecting the discriminative features that the model estimates for the registers.

4.1 Classification results

This section presents the classification results of the five different models used in the study: SVMs with three different feature sets (lexical, syntactic, and a combination of both) as well as BERT and XLM-R (lexical features). The deep learning models are based on a predefined vocabulary, as discussed in Section 2.2.2, and are not trained with explicit syntactic information.

Table 3 presents the classification results with the models discussed in the previous sections, namely SVM, BERT, and XLM-R, as well as the classification results with syntactic features using SVM. For computational reasons, the measure of variation in model performance was calculated based on a different number of repetitions. Specifically, Table 3 reports the mean and standard deviation of micro-averaged F1-score (proportion of correctly classified observations) for fifty repetitions for the SVMs and three repetitions for the deep learning models.

As Table 3 shows, the deep learning models BERT (F1 = 97.0 per cent) and XLM-R (F1 = 96.8 per cent) perform over two percentage points better than the

Table 3. Classification results with SVM ($n = 50$), BERT ($n = 3$), and XLM-R ($n = 3$).

	Lexical features		Syntactic features		Lexical + syntactic features	
	F1 (per cent)	Std.	F1 (per cent)	Std.	F1 (per cent)	Std.
SVM	94.59	0.0006	87.43	0.0007	94.62	0.0005
BERT	97.01	0.0008	–	–	–	–
XLM-R	96.83	0.0010	–	–	–	–

traditional supervised machine-learning method SVM (F1 = 94 per cent) when classifying with lexical features. The SVM classifier did not perform as well with only the syntactic information (F1 = 87 per cent), which could be expected as previous studies have showed that grammatical information by itself may not be sufficient for identifying all registers equally well (Laippala et al. 2021). Since the data are unbalanced, the models may be biased toward correctly classifying the majority class *Letter*. Therefore, it is also important to get a more comprehensive view of the models' performance on different register classes. All results, however, outperform the Zero Rule baseline accuracy of 84 per cent. The Zero Rule algorithm always predicts the class that has the greatest number of observations in the training data. Even though the pre-trained deep learning language models have not been trained with Late Modern English language, the data are still sufficiently close to contemporary English to improve the results from the simple feature based linear SVM classifier.

Registers differ in terms of how well they are linguistically defined, which has a direct effect on how well they can be identified automatically, as shown in previous research (Biber and Egbert 2016, 2018; Laippala et al. 2021). Thus, it is important to inspect the individual registers and how well each of them can be identified. Table 4 presents the register specific classification results with the best performing model BERT as an average of three runs.

As Table 4 shows, the best identified register is *Letter*, with the F1-score of 99 per cent. Letters form the majority of the data, and there are enough texts of this register for the model to learn to identify them correctly, but the high F1-score indicates that the register is also linguistically well-defined and distinct from the other registers. Although much of the data are letters, Table 4 shows that the model is able to identify other register classes with reasonable performance. The technical class *Not English* (F1 = 97 per cent) is also relatively easy to identify. It is the second largest class in the data and expectedly easy to identify as all the texts are similar in that they are written in another language than English. For instance, *Personal Record* is identified with 91 per cent F1-score which also indicates that the texts share identifiable similarities across the register and the texts are relatively easy to separate from other registers.

Laws and Statutes has the lowest F1-score of 39 per cent, which is not surprising as this class has very few texts in the data (see Table 1 and Fig. 2), and thus the deep learning model may not have enough examples in the training data to learn the distinct features of the register. However, the lower classification scores could also indicate something about the uniformity of the texts in different registers. Some registers have more similar texts within them and are thus easier to identify, whereas other registers with inner variation may not be so well defined, as mentioned above.

The register specific differences are further illustrated in Fig. 2, which provides a scatter plot of the F1-score in relation to the size of the training data. Figure 2 displays the tendency of registers with fewer texts to have a lower F1-score, as with the classes in the lower left-hand corner, such as *Laws and Statutes*, *Debate*, and *Speech* that have only few instances in the training data. However, toward the upper left-hand corner of the figure, there are register classes such as *Book*, *Cases*, and *Contract* that have moderate to high F1-scores despite having only few texts in the training data. This would indicate that these registers would have some distinct linguistic characteristics for them to be identified with a higher classification rate. In the upper right corner of Fig. 2 is the register *Letter* with the highest number of texts as well as the highest classification score, followed by *Not English*, *Personal Record*, and *No Text* with high classification results but much less texts. These registers are also likely to have distinct and well-defined characteristics.

Figure 3 further illustrates the classification results of individual registers by presenting a heatmap of a confusion matrix on the predictions of the best performing model BERT. The actual classes are shown in the rows and predicted classes are shown in the columns. The cells in the diagonal show the number of correctly classified texts. Table 4 and Fig. 2 give information on how well a register class is predicted and Fig. 3 complements this information by revealing the class predictions. Studying the classifier's decisions and finding possible explanations for the mismatches between predicted labels and the correct register labels can help to ensure the validity of the classifier.

Table 4. Register specific classification scores with BERT ($n = 3$).

	Precision (per cent)	Recall (per cent)	F1-score (per cent)	Number of texts
Book	88.95	92.04	90.47	67
Cases	81.48	66.67	73.28	46
Contract	70.49	75.71	72.79	59
Debate	79.05	39.39	52.31	11
Essay	66.33	59.61	62.68	85
Letter	98.69	98.79	98.74	18,527
Laws and Statutes	41.79	36.36	38.77	22
Newspaper	62.78	50.79	56.12	63
Not English	96.02	97.55	96.78	1,062
No Text	87.68	93.28	90.39	724
Pamphlet	73.08	50.00	59.36	18
Personal Record	92.51	89.71	91.09	972
Proposal	68.43	61.67	64.78	80
Proceedings	72.27	72.77	72.49	284
Speech	64.72	46.67	54.15	30
Treaty	69.72	48.15	56.43	9

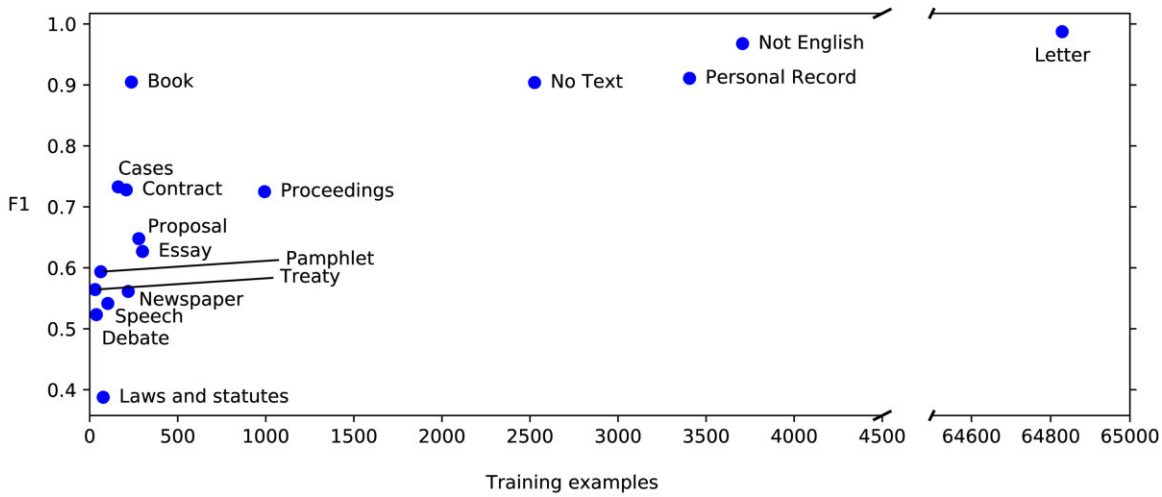

Figure 2. Class-wise F1-scores in contrast to the number of training examples.

As can be seen in Fig. 3, when a text is misclassified, it is most often predicted to be a *Letter*, as in the case of *Proceedings*, 16 per cent of which are misclassified as *Letter*. Additionally, *Personal Record* and *No Text* are also fairly commonly misclassified as *Letter*. Texts in *Letter* are quite consistently predicted correctly, but if they are misclassified, it is to *No Text* or *Proceedings*. This would suggest that these registers share some characteristics that confuse the classifier. On the other hand, Fig. 3 shows that texts in the register classes *Book*, *Cases*, and *Contract* are quite consistently predicted, further supporting the idea of their well-defined and characteristic features.

Examples (1) and (2) provide text examples of *Proceeding* and *No Text*, respectively, that are predicted incorrectly by the classifier as belonging to the register *Letter*. Example (1) is a hospital report describing the admittance and discharge of patients. Example (2) is a summary of a letter from Willink Van Staphorst to Thomas Jefferson.

(1) 4 June 1752
During our months visitation John Poor being cured, was discharged; and Katherine Shannon's child being also cured, she was discharged with the Child.

It was agreed to admit Angus McDonnel a dropsical person, if his friends will engage to indemnify the hospital and City from all charges that may accrue on his death or removal to the place of his residence.

It was also agreed to admit a Lunatick from New York at the rate of 10s. per week. We received three pounds from Dr. Thomas Bond on account of

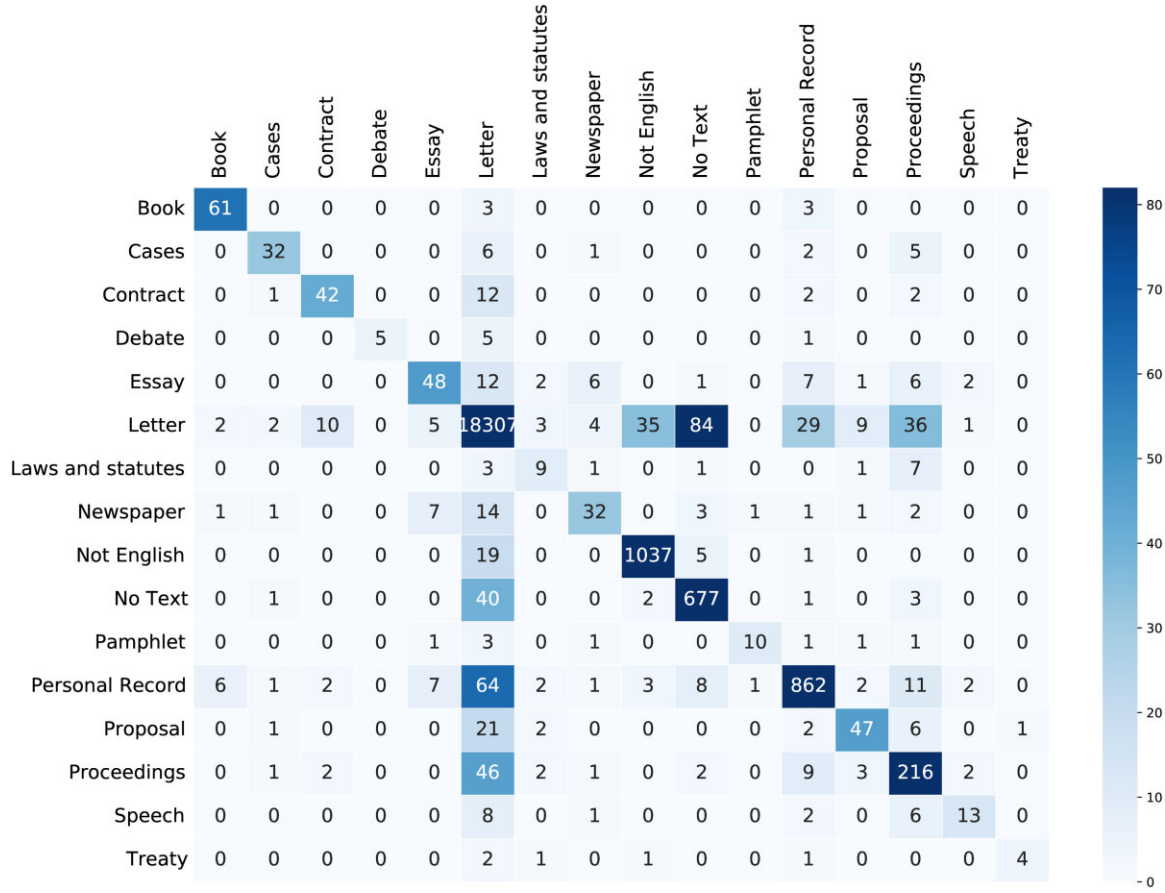

Figure 3. Heatmap confusion matrix presenting multi-class classification results with BERT. Rows represent the actual class label instances and columns represent the classifier's class label predictions. Numbers represent the actual instances in the test data.

the board of Thomas Stow, which we delivered to the Matron.

Henry Fetter being cured, was discharged.

The Physicians have attended constantly on the days of our visitation.

B. Franklin
Is: Jones
(Franklin/01-04-02-0112)

(2) Amsterdam, 30 December 1791. Having received TJ's 26 July letter about his draft of 1,000 dollars in favor of Gouverneur Morris, they paid it at the rate of 50 stivers per dollar and, as TJ directed, charged it at 2,500 florins to the Secretary of State's public account.

(Jefferson/01-27-02-0768).

Misclassifications between the register labels and the classifier's predictions not only provide information about the model's performance but also about the coherence of the corpora, as a misclassification may be due to the classifier making a mistake, or because a text represents some other register more than the one it has been assigned to. For instance, as with Example (2), many of the texts in the *No Text* register misclassified as *Letter* are in fact summaries or modern editions of Founding-Era letters. This information may not be immediately clear from the text itself, but rather comes from the metadata surrounding the text, such as the titles, which are not available to the classifier. Overall, while the BERT model shows promising results in identifying registers, the issues of generalizability should be considered due to the varying degrees of accuracy in identifying different registers and the potential for misclassifications.

4.2 Analysis of *Letter* and *Personal Record*

This section inspects the discriminative features estimated by the SVM to gain qualitative insights into the register classes. We focus on two of the largest and best-predicted (see Section 4) registers, *Letter* and *Personal Record*. Although *Not English* has the second-best classification score, it is not considered

here to be a meaningful register, but instead a technical category. We analyze the grammatical characteristics of *Letter* and *Personal Record* based on the most important grammatical features estimated by the SVM. Although the SVM is not the best performing model, it enables us to inspect the grammatical features of the texts, following the traditional line of register studies, as mentioned in Section 3.3. These features can further clarify the classifier's decisions and the quality of the classification, allowing us to determine whether the classifier's learning is effective. To obtain a better understanding of the grammatical features, we recorded the most significant features for each register, as described in Section 3.3.

Table 5 presents the thirty most discriminative features estimated by the SVM classifier for the register *Letter*. All of the discriminative features of *Letter* can be seen in Appendix 1. As can be seen in Table 5, many of the discriminative features in *Letter* are verbal, and include, for instance, features related to past tense, possibility verbs, auxiliary verbs, passive verbs, and verb complement clauses. It is also noteworthy that many features are related to copular verbs. These include, for instance, third person contracted copular verbs and present participle copular verb complement clauses. In addition to verbal features, the register *Letter* is characterized by past participle attributive and predicative adjectives, plural common nouns, proper nouns, indefinite pronouns, and third person singular pronouns. Other discriminative features found for *Letter* are stranded prepositions, subordinators of concession, letters of the alphabet, and extraposing.

The register *Letter* includes both private and professional letters. Late Modern English letters display interpersonal communication that is characterized by formulaic greetings and closing salutations, much like modern letters. These characteristics are also portrayed by the estimated features, as Example (3), which contain closing salutations with copular verbs and letters of the alphabet in bold, illustrates.

(3) (a) I **am** Dear Sir with great Respect your most Obedient and Very Humble Servt. **W** Heath
(b) I **am** indeed affectionately your Friend **J. A.**

Features that have been linked to interpersonal and interactive communication of letters (Biber and Finegan 1989) can also be seen in Table 5. These include features expressing personal involvement, such as copular verbs, and features expressing stance, such as possibility modals. However, though some features related to personal interaction can be seen in Table 5, there are also features representing more informational characteristics, such as passives and different types of verb complement clauses. Some informational characteristics are illustrated in Example (4) that is a text sample of a correctly classified *Letter*.

Excerpt from a letter (passives underlined; past participle attributive adjectives in italics; proper nouns in bold)
(4) As the **Article** of **Powder** is much wanted to carry on the operations vs the ministerial **Army**, and as the british **Ministry**, have taken every **Step** that human **Nature** could divise to prevent the **Americans** obtaining So essential an **Article**; it is humbly Submitted to the **Wisdom**, of the cont. **Congress**, whether it will not be prudent to Supply yourselves with that **Article** at the **Expence** of the *said* **Ministry**, by taking it whenever you can get it.

Table 5. Thirty most frequent discriminative features of *Letter*.

Tag	#	Tag	#
Past participle clause	50	Plural common noun	39
Passive clause	50	Base form of the verb	38
Past participle attributive adjective	49	Third person singular pronoun	36
Indefinite pronoun	49	*be* as a lexical verb	36
Singular letter of the alphabet	48	*be* as a copular verb	36
Third person contracted copular verb	46	Present participle verb complement clause	36
Past perfect auxiliary verb	45	Present participle verb complement clause with a copular verb	36
Number as a determiner	45	*to* complement clause	35
Third person contracted lexical verb	45	Premodifying singular common noun	34
Subordinating conjunction of concession	43	Semi-modal of possibility	34
Phrasal verb	43	Copular verb	33
be as an auxiliary verb	43	Proper noun	33
Extraposed verb complement clause	42	Extraposed clause	33
Past participle predicative adjective	42	Past perfect lexical verb	32
Stranded preposition	40	Definite article	31

Note: # represents the number of times the feature was used to discriminate *Letter* from other registers in the model.

It is thus further recommended that 2 **Vessells** properly mann'd be sent to the **Island** of **Antigua**, one of which may anchor at **Old Road** on the South **Side** of the *Said* **Island** (where there are only a few **Houses**) in the **Evening** under Dutch **Colours**; passing for a **Vessell** bound on a *forced* **Trade**, to the **French Islands**; in the night you may land, and take away all the powder; there being not above one or two **Persons**, in the fort to prevent it. As Soon as the **Powder** is obtained the **Vessell** may proceed down to **Johnsons Point Fort**, at the **S. W.** point of the **Island**; and take what is there; there being only a **Single Matross** in the *Said* **Fort;** the other **Vessell** must be commanded by a prudent **Man**; well acquainted with the **Bar** and **Harbour** at **St. Johns.**
(Adams/06-03-02-0100-0002)

Passives are very frequent in this short text excerpt which exhibits six occurrences. The example also showcases the frequent use of proper nouns that are related to specific people and places in this professional letter. These proper nouns are very information focused and related to specific arrangements and instructions. Capitalization of nouns was at its height at the end of the 17th century and early 18th century, and some authors used it in emphasizing any noun they considered important (Crystal 2003). The text sample also features past participle attributive adjectives, such as *said* in *the said Ministry*. Interestingly, the participial adjective construction 'the said X' with anaphoric reference appears three times in this short text excerpt. This could indicate a desire for precision typical of informative language use.

In earlier multidimensional studies, Late Modern English letters have been found to have expository, descriptive, and argumentative features (Biber 2001; Monaco 2017). The lack of features concerning personal involvement and highly informative characteristics have been explained by letters being used in the transmission of scientific knowledge and their didactic character in the late modern period (Monaco 2017). The informativeness and descriptiveness of letters can also be seen in the present study, although there are also more personal and less professional letters in the data, as suggested by some of the features in Table 5.

For comparison, Table 6 presents the thirty most discriminative features estimated by the classifier for the register *Personal Record*.

Many of the discriminative features in the *Personal Record* register are nominal, whereas *Letter* has more verbal features. The classifier has also estimated for *Personal Record* similar verbal features related to past tense, copular, and auxiliary verbs, but also agentless passives and necessity modals. The nominal discriminative features include many features related to plural and singular common nouns, predicative and attributive adjectives, temporal nouns, and determiners. Plural personal pronouns are also estimated as significant features, as well as conjuncts, such as subordinators of concession, coordinating conjunctions, WH-subordinators, and phrasal coordination.

The register *Personal Record* includes journal entries and other personal documentation on events and personal ideas worth recording for oneself, and thus it seems reasonable to witness so many nominal features and past tense use as these describe what, when, and where something occurred. These features are typical in narrative texts describing or commenting on events and actions. In addition to narration, the discriminative features in *Personal Record*, such as the presence of complex noun phrase structures (e.g. common nouns, attributive adjectives, post-determiners, and phrasal coordination), many types of conjuncts, and agentless passives, would indicate that journal entries were characterized by informational density. This observation is consistent with earlier research on language use in personal journal entries, which has found similar features (Biber 2001). The agreement between previous research and the results generated by the SVM provide evidence for the validity of the model in identifying the linguistic patterns in this register. As most of the *Personal Record* entries have been written by men in our sub-corpus of COFEA, the informational density in these journals may be related to the gender of the writer. Earlier research has shown that men's writing has a more informational focus and a more nominal style (Biber and Burges 2000; Degaetano-Ortlieb, Säily, and Bizzoni 2021). Example (5) is an excerpt of a correctly classified *Personal Record* that illustrates some of the features characterizing informational density.

Excerpt from a personal record (passives underlined; temporal nouns in italics; past tense verbs in bold; adjectives in wavy underline)
(5) **Dined** at Judge Sargeant's, with Mr. and Mrs. Shaw. Mr. Porter and his lady are there upon a visit from Rye: with a child about six *weeks* old, which forsooth immediately after dinner must be produced, and **was** handed about from one to another; and very shrewd discoveries **were** made of its resemblance to all the family by turns, whereas in fact it **did** resemble nothing but chaos. How much is the merciful author of nature to be adored for implanting in the heart of man a passion stronger than the power of reason, which affords delight to the parent at the sight of his offspring even at a *Time*, when to every other person it must be disgusting. Yet it appears to me, that parents would do wisely in keeping their children out of sight at least untill they are a *year* old, for I cannot see what

Table 6. Thirty most frequent discriminative features of *Personal Record*.

Tag	#	Tag	#
Past participle attributive adjectives	50	Modal verb of necessity	40
Cardinal number	49	Attributive adjectives	39
Past participle predicative adjectives	48	Quantifying post-determiner	39
Predicative adjectives	48	Past participle auxiliary verb	39
Agentless passive multiword verb	48	Plural noun	37
Singular temporal noun	47	Definite article	37
Singular common noun	46	Past participle lexical verb	34
Possessive determiner	46	Common noun (neutral for number)	33
Plural personal pronouns	45	Pronoun	33
Post-determiner	42	Copular verb	31
Perfect multiword verbs	41	Phrasal coordination	30
Multiword conjunctions	41	Multiword preposition	29
WH-subordinating conjunctions	41	Extraposed clause	26
Subordinating conjunctions of concession	41	Adverbial past participle clause	26
Contracted copular verb	41	Prepositional verb	25

Note: # represents the number of times the feature was used to discriminate *Personal Record* from other registers in the model.

> satisfaction, either sensual or intellectual can be derived from seeing a misshapen, bawling, slobbering infant, unless to persons particularly interested. We **drank** tea likewise at the judge's, and **return'd** home between 7 and 8 in the *evening*. Leonard White **came** up to give me a letter for his chum.
>
> (Adams/03-02-02-0002-0008-0023)

Though the text sample is from a personal diary, it exemplifies the frequent use of passives. The text excerpt also features many nouns, such as temporal nouns, as well as singular and plural common nouns, adjectives as well as past tense verbs. Interestingly, the journal entry is not characterized by the frequent use of first-person subject pronouns, but rather their omission (*Dined at Judge Sargeant's*) as well as plural personal pronouns when describing past events (*We drank tea*).

In a MDA of 18th century registers, personal journals have been found to have highly literate and narrative characterizations (Biber 2001). The narration of journals is rather impersonal, characterized by passives, omissions of subject pronouns, and the use of full nouns rather than third person pronouns when referring to someone else than the writer or the person addressed, which is different from the dynamic narration of, for example, fiction (Biber 2001). These same characterizations can also be seen in the present analysis and in Example (5).

5. Discussion and conclusions

The aim of this article was to examine the registers in COFEA to gain insights into the linguistic characteristics of the registers as well as to explore the automatic identification of registers in COFEA. For the inspection of linguistic features, we trained a linear classifier and analyzed the most important grammatical features estimated by the classifier to examine the registers and their characteristics in COFEA. For the register identification, we applied three machine-learning methods that were BERT, XLM-R, and SVM, and trained the classifiers with a sub-corpus of COFEA to see how well the registers are predicted. We experimented with three different feature sets in predicting the registers and evaluated the predictions by analyzing the register specific classification results, the classifier prediction mistakes, as well as the linguistic characteristics of the two best predicted registers.

The potential of register studies is greatly enhanced through new corpora with register information and linguistic annotation, especially when considering the increased interest in large corpora and big data among different research areas, such as, linguistics, history, and gender studies, as Kytö (2019) notes. We studied the automatic register identification in COFEA to make the corpus and its register information more applicable and versatile for many kinds of research.

The best performance was achieved by the BERT Large model with an F1-score of 97 per cent. Classification with grammatical information was more difficult, as also shown in previous studies. Despite being pre-trained on contemporary language that differs from the late modern variety used in testing, the deep learning methods outperformed the traditional machine-learning method SVM. This suggests that these deep learning methods could also be utilized in other studies concerning historical language and its automatic classification.

The analysis of register specific classifications showed that *Letter* was the best identified register class.

A high classification score for the individual registers with many instances in the training data, such as *Letter*, *Personal Record*, *No Text*, and *Not English* could indicate that these registers are linguistically well-defined and are relatively easy to distinguish from the other registers. The number of examples on one register class in the training data were not, however, the most meaningful aspect in the classification as some registers, such as *Book*, do not have very many examples in the training data but are quite correctly classified. This would also indicate that these registers have distinct linguistic characteristics.

The linguistic inspection of the two best predicted register classes revealed that letters in the late 1700s, like modern letters, were addressed to a specific individual. However, instead of the modern interaction and involvement in letters, letters in COFEA exhibited more descriptive and argumentative style, which is likely attributed to their professional tone. Additionally, journal entries in the late 1700s seem to be characterized by informational density, which is different from modern diaries and journals that one would expect to be characterized by personal involvement and stance expressions. A modern journal is more introspective in nature and has developed many features related to spoken communication. Unlike modern registers, both *Letter* and *Personal Record* were found to be characterized by the use of passive constructions. This supports earlier findings of most 18th century registers commonly using the passive voice (Biber 2001). Letters and journals of late 1700s have evolved into much more spoken-like and interactive language use in contemporary registers.

This initial sub-corpus classification naturally leaves many perspectives for future work. The next step would be extending the register classification to the whole corpus to obtain more practical utility from COFEA. However, it should be noted that due to the nature of the data sampling and the unbalanced register distribution, there could be some issues with the generalizability of the results to the whole corpus. The classification could also be used as a pre-processing step, for example, in identifying the technical classes *No Text* and *Not English*.

BERT is limited by its ability to model only 512 word tokens at a time, although it has been shown that the beginnings of texts exhibit the most predictive power (Laippala et al. 2023). This naturally means that all the information provided by longer texts cannot be used by BERT. Future work could explore the predictive power of different parts of longer texts and models adapted to long documents to see whether these would affect register specific classification performance.

Acknowledgements

We thank CSC—IT Centre for Science for computational recourses.

Conflict of interest statement. None declared.

Funding

None declared.

References

Primary Sources

"22d.," Founders Online, National Archives, https://founders.archives.gov/documents/Adams/03-02-02-0002-0008-0023. [Original source: Robert Taylor, J., and Friedlaender, M., eds (1981) *The Adams Papers, Diary of John Quincy Adams*, Vol. **2**, March 1786 – December 1788, pp. 278–9. Cambridge, MA: Harvard University Press.]

"Enclosure: A Proposal Regarding the Procurement of Powder, 12 October 1775," Founders Online, National Archives, https://founders.archives.gov/documents/Adams/06-03-02-0100-0002. [Original source: Taylor, R. J., ed. (1979) *The Adams Papers, Papers of John Adams*, Vol. **3**, May 1775 – January 1776, pp. 197–8. Cambridge, MA: Harvard University Press.]

"Pennsylvania Hospital: Report of the Weekly Committee, 4 June 1752," Founders Online, National Archives, https://founders.archives.gov/documents/Franklin/01-04-02-0112. [Original source: Labaree, L. W., ed. (1961) *The Papers of Benjamin Franklin*, Vol. **4**, July 1, 1750, through June 30, 1753, pp. 320–1. New Haven: Yale University Press.]

"To Thomas Jefferson from Willink, Van Staphorst & Hubbard, 30 December 1791," Founders Online, National Archives, https://founders.archives.gov/documents/Jefferson/01-27-02-0768. [Original source: Catanzariti, J., ed. (1997) *The Papers of Thomas Jefferson*, Vol. **27**, 1 September–31 December 1793, p. 812. Princeton: Princeton University Press]

Secondary Sources

Argamon, S. E. (2019) 'Register in Computational Language Research', *Register Studies*, **1**: 100–35.

Bahdanau, D., Cho, K. H., and Bengio, Y. (2015) 'Neural Machine Translation by Jointly Learning to Align and Translate', *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, pp. 1–15, San Diego, CA, USA.

Barclay, S., Earley, B., and Boone, A. (2019) 'Original Meaning and the Establishment Clause: A Corpus Linguistics Analysis', *Arizona Law Review*, **61**: 505–60.

Baron, D. (2019) 'Corpus Evidence Illuminates the Meaning of Bear Arms', *Hastings Constitutional Law Quarterly*, **46**: 509–22.

Biber, D. (1988) *Variation across Speech and Writing*. Cambridge: CUP.

Biber, D. (2001) 'Dimensions of Variation among 18th Century Registers'. In: Diller, H.-J. and Görlach, M. (eds) *Towards a History of English as a History of Genres*, pp. 89–110. Heidelberg: C. Winter.

Biber, D. (2012) 'Register as a Predictor of Linguistic Variation', *Corpus Linguistics and Linguistic Theory*, **8**: 9–37.

Biber, D., and Burges, J. (2000) 'Historical Change in the Language Use of Women and Men', *Journal of English Linguistics*, **28**: 21–37.

Biber, D., and Conrad, S. (2019). *Register, Genre, and Style*, 2nd edn. Cambridge: CUP.

Biber, D., and Egbert, J. (2016) 'Register Variation on the Searchable Web: A Multi-Dimensional Analysis', *Journal of English Linguistics*, **44**: 95–137. https://doi.org/10.1177/0075424216628955.

Biber, D., and Egbert, J. (2018) *Register Variation Online*. Cambridge: CUP. https://doi.org/10.1017/9781316388228.

Biber, D., and Finegan, E. (1989). Drift and the Evolution of English Style: A History of Three Genres, *Language*, **65**: 487–517.

Biber, D., and Finegan, E. (1992). 'The Linguistic Evolution of Five Written and Speech-Based English Genres from the Seventeenth to the Twentieth Centuries'. In: Rissanen, M., Ihalainen, O., Nevalainen, T. and Taavitsainen, I. (eds), *History of Englishes: New Methods and Interpretations in Historical Linguistics*, pp. 688–704. Berlin: Mouton de Gruyter.

Biber, D., and Finegan, E. (1997) 'Diachronic Relations among Speech-Based and Written Registers in English', in: Nevalainen, T. and Kahlas-Tarkka, L. (eds) *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*, pp. 253–75. Helsinki: Société Neophilologique.

Biber, D., Finegan, F., and Atkinson, D. (1994) 'ARCHER and Its Challenges: Compiling and Exploring A Representative Corpus of Historical English Registers', *Creating and Using English Language Corpora, Papers from the 14th International Conference on English Language Research on Computerized Corpora*, pp. 1–13. Zurich, Switzerland.

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

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

Conneau, A. et al. (2019) 'Unsupervised Cross-Lingual Representation Learning', *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 31–8. Florence, Italy. https://doi.org/10.18653/v1/p19-4007

Conrad, S. (1996) 'Investigating Academic Texts with Corpus-Based Techniques: An Example from Biology', *Linguistics and Education*, **8**: 299–326.

Crystal, D. (2003) *The Cambridge Encyclopedia of the English Language*, 2nd edn. Cambridge: CUP.

Cunningham, C., and Egbert, J. (2020) 'Using Empirical Data to Investigate the Original Meaning of "Emolument" in the Constitution', *Georgia State University Law Review*, **36**: 465–89.

Degaetano-Ortlieb, S., Säily, T., and Bizzoni, Y. (2021). 'Register Adaptation vs. Innovation Across Situational Contexts: 18th Century Women in Transition', *Frontiers of Artificial Intelligence*, **4**: 1–22.

Devlin, J. *et al.* (2019) 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding', *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (AACL HLT 2019)*, pp. 4171–86, Minneapolis, Minnesota, USA.

Finn, A., and Kushmerick, N. (2006) 'Learning to Classify Documents According to Genre', *Journal of the American Society for Information Science and Technology*, **51**: 1506–18. https://doi.org/10.1002/asi.20427

Geisler, C. (2002) 'Investigating Register Variation in Nineteenth-Century English: A Multi-Dimensional Comparison'. In: Reppen, R., Fitzmaurice, S. M., and Biber, D. (eds) *Using Corpora to Explore Linguistic Variation*, pp. 249–71. Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/scl.9.17gei

González-Álvarez, D., and Pérez-Guerra, J. (1998) 'Texting the Written Evidence: On Register Analysis in Late Middle English and Early Modern English', *Text & Talk*, **18**: 321–48. https://doi.org/10.1515/text.1.1998.18.3.321

Guyon, I., and Elisseeff, A. (2003) 'An Introduction to Variable and Feature Selection', *Journal of Machine Learning Research*, **3**: 1157–82. https://doi.org/10.1016/j.aca.2011.07.027

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

Hashimoto, B. (n.d.). *Data: COFEA*. https://sites.google.com/site/brettjameshashimoto/data, accessed 19 Jan. 2023.

Hosseini, K. *et al.* (2021) 'Neural Language Models for Nineteenth-Century English', *Journal of Open Humanities Data*, **7**: 1–6. http://doi.org/10.5334/johd.48

Joachims, T. (1998) 'Text Categorization with Support Vector Machines: Learning with Many Relevant Features'. In: Nédellec, C. and Rouveirol, C. (eds) *Machine Learning: ECML-98. Lecture Notes in Computer Science*, pp. 137–42. Berlin: Springer.

Kyröläinen, A.-J., and Laippala, V. (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, M. (2019) 'Register in Historical Linguistics', *Register Studies*, **1**: 136–67. https://doi.org/10.1075/rs.18011.kyt

Laippala, V. *et al.* (2019) 'Toward Multilingual Identification of Online Registers', *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pp. 292–7. Turku, Finland.

Laippala, V. *et al.* (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. https://doi.org/10.1007/s10579-020-09519-z

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

Lample, G., and Conneau, A. (2018) 'Cross-lingual Language Model Pretraining', https://arxiv.org/abs/1901.07291. Accessed 3 July 2021.

Levering, R., Cutler, M., and Yu, L. (2008) 'Using Visual Features for Fine-Grained Genre Classification of Web Pages', *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, pp. 1–10. Waikoloa, Hawaii. https://doi.org/10.1109/HICSS.2008.488

Li, X. *et al.* (2019) 'Exploiting BERT for End-to-End Aspect-Based Sentiment Analysis', *Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text*, pp. 34–41. Hong Kong, China. https://doi.org/10.18653/v1/d19-5505

Libovický, J., Rosa, R., and Fraser, A. (2019) 'How Language-Neutral Is Multilingual BERT?', *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001. Florence, Italy.

Lim, C. S., Lee, K. J., and Kim, G. C. (2005) 'Multiple Sets of Features for Automatic Genre Classification of Web Documents', *Information Processing and Management*, **41**: 1263–76. https://doi.org/10.1016/j.ipm.2004.06.004

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). 'Explainable AI: A Review of Machine Learning Interpretability Methods', *Entropy*, **23**, 18.

Liu, Y. *et al.* (2019) 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', http://arxiv.org/abs/1907.11692 (accessed 3 July 2021).

Mishra, S., and Mishra, S. (2019) '3Idiots at HASOC 2019: Fine-Tuning Transformer Neural Networks for Hate Speech Identification in Indo-European Languages', *CEUR Workshop Proceedings*, Vol. 2517, pp. 208–13. Kolkata, India.

Monaco, L. M. (2017) 'A Multidimensional Analysis of Late Modern English Scientific Texts from the Coruña Corpus', PhD thesis, The University of A Coruña.

Mouritsen, S. C. (2010) 'The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning', *Brigham Young University Law Review*, **87**: 1915–78. http://ssrn.com/abstract=1753333 (accessed 11 May 2021).

Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, **12**: 2825–30.

Petrenz, P., and Webber, B. (2011) 'Stable Classification of Text Genres', *Computational Linguistics*, **37**: 385–93. https://doi.org/10.1162/COLI_a_00052

Pritsos, D., and Stamatatos, E. (2018) 'Open Set Evaluation of Web Genre Identification', *Language Resources and Evaluation*, **52**: 949–68. https://doi.org/10.1007/s10579-018-9418-y

ProQuest (2021) 'Early English Books Online', https://www.proquest.com/eebo, accessed 14 September 2021.

Repo, L. *et al.* (2021) 'Beyond the English Web: Zero-Shot Cross-Lingual and Lightweight Monolingual Classification of Registers', *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, Kyiv, Ukraine (Online)*, pp. 183–91.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020) 'A Primer in BERTology: What We Know About How BERT Works', *Transactions of the Association for Computational Linguistics*, **8**: 842–66. https://doi.org/10.1162/tacl_a_00349

Sebastiani, F. (2002) 'Machine Learning in Automated Text Categorization', *ACM Computing Surveys*, **34**: 1–47.

Sharoff, S. (2021) 'Genre Annotation for the Web', *Register Studies*, **3**: 1–32. https://doi.org/10.1075/rs.19015.sha

Sharoff, S., Wu, Z., and Markert, K. (2010) 'The Web Library of Babel: Evaluating Genre Collections', *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 3063–70. Valletta, Malta.

Slocum, B. G. (2015) *Ordinary Meaning: A Theory of the Most Fundamental Principle of Legal Interpretation*. Chicago: University of Chicago Press.

Taavitsainen, I. (1997) 'Genre Conventions: Personal Affect in Fiction and Non-Fiction in Early Modern English', in: Rissanen, M., Kytö, M., and Heikkonen K. (eds) *English in Transition: Corpus-Based Studies in Linguistic Variation and Genre Styles*, pp. 185–266. Berlin: Mouton de Gruyter.

Taavitsainen, I., and Pahta, P. (eds) (2004) *Medical and Scientific Writing in Late Medieval English*. Cambridge: CUP.

Taavitsainen, I., and Schneider, G. (2019) 'Scholastic Argumentation in Early English Medical Writing and Its Afterlife: New Corpus Evidence', in: Suhr, C., Nevalainen, T., and Taavitsainen, I. (eds) *From Data to Evidence in English Language Research*, pp. 191–221. Boston: BRILL.

Tanase, M.-A., Cercel, D.-C., and Chiru, C. (2020) 'UPB at SemEval-2020 Task 12: Multilingual Offensive Language Detection on Social Media by Fine-Tuning a Variety of BERT-Based Models', *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 2222–31. Barcelona (online). https://doi.org/10.18653/v1/2020.semeval-1.296

Vapnik, V. (1998) *Statistical Learning Theory*. New York: Wiley.

Vaswani, A. *et al.* (2017) 'Attention is All You Need', *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pp. 6000–10. Red Hook, NY, USA.

Wan, L. *et al.* (2019) 'Long-Length Legal Document Classification', http://arxiv.org/abs/1912.06905 (accessed 3 July 2021).

Wenzek, G. *et al.* (2020) 'CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data', *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4003–12. Marseille, France.

Wolf, T. *et al.* (2020) 'Transformers: State-of-the-Art Natural Language Processing', *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Punta Cana, Dominican Republic (online).

Worsham, J., and Kalita, J. (2018) 'Genre Identification and the Compositional Effect of Genre in Literature', *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1963–73. Santa Fe, New Mexico, USA.

Wright, S. (1994) 'The Place of Genre in the Corpus', in: Rissanen, M., Kytö, M., and Wright, S. (eds) *Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, St Catharine's College Cambridge, 25–27 March 1993*, pp. 101–7. Amsterdam: Rodopi.

Zhu, Y. *et al.* (2015) 'Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books', *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pp. 19–27. Santiago, Chile. https://doi.org/10.1109/ICCV.2015.11

Appendix 1

All discriminative features for *Letter*.

Letter				
en_cls 50	nec_multi1 28	vl_ing_comp_prep_multi1 11	im 4	per_v 1
psv 50	multi2 27	cls_advs 11	p_im 4	plur_1 1
en_atrb 49	vm__multi3 27	cls 10	va_ed 4	plur_pmod 1
p_indef 49	wh_comp_verb 27	obj 10	to_adj 3	prf_multi1 1
zz_sing 48	en_prf 26	part 10	c__srd_tht 3	ing_rel 1
z_cop_cont 46	en_psv 25	part_phrv 10	c_crd_phrs 3	ing_rel_subj 1
va_en_prf 45	z_multi1 25	ppvb 10	i_ppvb 3	p_wh_comp 1
more_pmod 45	pr_u 24	pmod_multi2 9	vm__pos 3	ing_cop 1
vl_z_cop_cont 45	wh 24	pmod 9	more 3	vl_ing_cop 1
c_srd_conc 43	comp_verb 24	vm 9	p_per_sing_2 3	prep_multi1 1
r_part_phrv_multi2 43	deg_multi1 23	c__srd_multi2 9	p_per_sing_2_x 3	splt_multi2 1
va_am 43	comp_verb_binf 23	cop 8	per_sing_2 3	c__srd_tht_rel 1
verb_ext 42	r_splt_deg 22	en_sing 8	per_sing_2_x 3	ing_comp_verb 1
en_prdv 42	vl_en_cls_advl 22	n_en_sing 8	vl_ing_comp_prep 3	n_u 1
i_strn 40	vm__multi2 21	1_x 8	d_impr 3	impr_3_v 1
cm_plur 39	c_srd_wh_cls 21	va_z 8	d_impr_3 3	p_wh_comp_verb 1
bf 38	i__multi3 21	plur 7	d_per_2 3	sing 1
sing_3 36	r__splt_multi2 20	atrb 7	d_per_2_v 3	va_ing_comp 1
vl_am 36	j_en 19	est 6	vl_ing 3	verb 1
vl_am_cop 36	ing_prg 19	inf_multi1 6	binf 3	c__srd_multi3 1
ing_cop_verb 36	prg 19	p_per_plur_1 6	verb_binf 3	er 1
vl_ing_cop_verb 36	vl_en_psv_agls_multi1 18	splt_multi1 6	n_cm_sing 3	va_en_psv_agls 1
vl_inf_comp_verb 35	inf_comp_verb_binf 17	va_en_psv 6	i__multi2 3	vl_z 1
cm_sing_pmod 34	ed_cont 17	sing_2 5	i_multi1 3	
pos_multi2 34	p_per_2 16	vl_ed_multi1 5	srd_multi3 2	
vl_inf_comp 34	phrs 16	advs 4	3_ext 2	
am_cop 33	en_prf_multi1 15	to_adj_ext 4	c_crd_multi1 2	
n__pr 33	cm_sing 14	neut 4	crd_multi1 2	
ext 33	r_splt_multi1 14	pr_plur 4	n_ing 2	
vl_en_prf 32	d_sing_pro 14	vm__pos_multi2 4	srd_multi2 2	
n_pr 32	to_comp_adj 13	prf 4	vl_ing_prg 2	
d_at_def 31	va_ing_comp_prep 13	sing_2_x 4	wh_ever 2	
ing_comp 31	c__crd_multi3 12	d_impr_3_v 4	advl 2	
at_def 30	crd_multi3 12	cm_neut 4	r_est 2	
v 30	prdv 12	n_en 4	d_plur 1	
splt_advs 29	ing_comp_prep 12	srd_wh 4	prdn 1	
r_multi1 28	vl_en_prf_multi1 11	en_cls_advl 4	p_per_v 1	

All discriminative features for *Personal Record*.

Personal Record				
j_en_atrb 50	2_v 21	plur_nom 9	ever 5	srd_wh_rel_obj 2
one 49	comp_verb 21	vl_inf_comp_verb_binf 8	impr_3 4	p_wh_rel_obj_pied 2
j_en_prdv 48	per_2_v 21	c__srd_sing 8	bf_multi1 4	rel_obj_pied 2
vl_en_psv_agls_multi1 48	agls 20	n_cm_plur_nom_pmod 8	pos_multi2 4	wh_rel_obj_pied 2
j_prdv 48	srd_wh_cls 19	rel_obj 8	srd_cnd 4	r_part 2
sing_time 47	vm_pos 19	srd_sing 8	va_bf 4	r_part_phrv 2
cm_sing 46	vl_ing_rel 19	wh_rel_obj 8	2_x 4	n_pr_plur 1
d_per_sing 46	vl_ing_rel_subj 19	ing_sing_pmod 8	n_pr_sing_time 4	to_comp_adj 1
per_plur 45	va_en_psv_agls 19	n_ing_sing_pmod 8	n_pr_tit_pmod 4	splt_deg 1
d_prdv 42	vl_z_multi1 18	vm_inf 8	pr_sing_time 4	cm_sing_time 1
c__crd_multi2 41	n_en 18	noun 8	pr_tit_pmod 4	cop 1
crd_multi2 41	r_splt_time 18	r__splt 8	psv 4	multi1 1
c_srd_wh 41	vm__multi2 17	srd_tht 8	vl_ing 4	n_cm_sing_time 1
cop_cont 41	j_prdv_er 16	tht 8	c_srd_sing 3	n_pr_sing_nom 1
prf_multi1 41	pr_plur_nom_pmod 16	i_strn 7	day 3	n_pr_tit 1
vm_nec 40	vl_ed 16	cm_plur_nom_pmod 7	en_psv_agls 3	n_u 1
atrb 39	n_pr_plur_nom_pmod 15	ing_sing 7	indef 3	p_wh_comp_verb 1
va_en 39	wh_cls 15	n_ing_sing 7	n_pr_plur_pmod 3	pr_sing_nom 1
n_cm_plur 39	n_ing_plur 14	wh_rel_subj 7	d_a 3	pr_tit 1
p_per_plur 39	r__multi2 14	r_a 7	ed_multi1 3	va 1
c_srd_conc 38	en_prf 13	c_srd_cnd 6	3_ext 3	vl_en_psv 1
d_plur_a 37	ing_plur 13	ing_prdv 6	pr 3	2 1
plur_a 37	zz 13	j_ing_prdv 6	to_comp_noun 3	cls_cnd 1
vl_en 34	va_ed_cont 12	pr_plur 6	d_sing_pro 3	d_plur_pro_a 1
n_plur 33	wh_rel 12	vm_prdn 6	d_plur_pro 3	multi3 1
p 33	z 12	binf 6	plur_pro 3	n_pr_sing_pmod 1
cop_verb 31	c_srd_tht 12	r_splt_multi1 6	c__srd 3	plur_pro_a 1
phrs 30	cm_sing_pmod 12	srd_cls 6	sing_more_pmod 3	pr_sing_pmod 1
at_def 29	wh_comp_verb 12	verb_binf 6	en_psv_by 2	p_per_sing_3 1
d_at_def 29	c__srd_multi2 11	advl 6	va_ing 2	r_g 1
i_multi1 29	va_ing_rel 11	plur_time 6	d_est 2	r_g_wh 1
n_cm_neut 27	va_ing_rel_subj 11	thdel 6	part_phrv_multi2 2	vl_ing_comp_prep_multi1 1
cm_neut 26	va_ed 11	en_prf_multi1 5	n_ 2	c_srd_multi1 1
en_cls_advl 26	vl_ing_prg 10	ing_comp_prep 5	n_cm 2	pmod_multi2 1
verb_ext 26	to_inf 9	va_am 5	pr_neut 2	srd_multi1 1
i_ppvb 25	deg 9	cls_advl 5	vl_en_cls_advl 2	vl_ing_comp_verb 1
p_wh_comp 25	vl_en_cls 9	vl_ed_multi1 5	en 2	comp_prep_multi1 1
comp_prep 24	r_part_phrv_multi2 9	c_srd_cls 5	plur_pmod 2	ing_comp_prep_multi1 1
vm_ 24	vl_en_psv_agls 9	n_pr_neut 5	to_adj 2	n_pr 1
r_time 22	adj_ext 9	splt_multi1 5	to_adj_ext 2	
psv_agls 21	splt_time 9	n__pr 5	c_srd_wh_rel_obj 2	