ISSN 1799-2591 (print); ISSN 2053-0692 (online)
Theory and Practice in Language Studies, Vol. 16, No. 11, pp. 3789-3799, November 2026
DOI: https://doi.org/10.17507/tpls.1611.12

Exploring Linguistic Variation Between American English and Indian English Business Texts Using R-Studio: A Corpus-Based Computational Approach

Mansour Altamimi
College of Languages and Translation, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

***Abstract*—This study attempts to make a comparative analysis of American English and Indian English texts through the application of various statistical tools in R-Studio. It attempts to identify the significant variations in the usage of words, the incidence of particular words, and the frequency of these usages in American English and Indian English. To maintain the objectivity of the study, two corpora consisting of similar types of texts are created. By utilizing the power of R programming for the purpose of text mining and corpus linguistics, the study attempts to examine the texts based on various parameters such as type/token ratio, lexical density, n-gram analysis, keyness, etc. It is observed that American English makes significant usage of high-frequency words in a standardized manner, while the usage of words in Indian English is pragmatic and attempts to combine a variety of words. This study falls at the crossroads of computational sociolinguistics and attempts to combine the power of corpus linguistics, statistical analysis, and a global perspective on English varieties as a new trend in the study of linguistics. It also attempts to demonstrate the power of the R-Studio environment as a powerful tool for corpus linguistics and the significance of statistical analysis in the study of linguistics.**

***Index Terms*—corpus linguistics, World Englishes, R-Studio, American Indian English, text mining**

I. Introduction

The global spread of English has resulted in the emergence of multiple institutionalized varieties, among which American English and Indian English represent two influential yet structurally and sociolinguistically distinct norms. While previous studies within the paradigm of World Englishes (e.g., Kachru, 1992; Bohmann, 2019; Wiltshire, 2020; Saadawi, 2025) have extensively described their phonological, lexical, and pragmatic features, many of these previous studies remain predominantly qualitative and based on manual or small-scale corpus analysis. The emergence of new insights in corpus linguistics and computer-assisted text analytics has made it possible to conduct large-scale statistical analyses; yet, the use of statistical environments such as R-Studio to investigate the similarities and differences between these two types of English has not been fully exploited. In other words, the application of quantitative approaches involving frequency analysis, keywords, collocation, and lexical richness in one single study design has not been sufficiently considered, which leads to a methodological limitation in identifying systematic differences (Schneider, 2007). However, the distinction between American and Indian English has been explained in impressionistic ways and through the use of individual text examples rather than through robust statistical models (Rizwan, 2009). This poses a methodological challenge because distinguishing between stylistic differences, genre effects, and varietal differences is complex. Without robust tools, it is challenging to estimate the degree to which vocabulary use, discourse markers, and sentence construction are characteristic of the two varieties. Furthermore, it is not possible to visualize the data and assess the statistical significance of the differences without the use of computational tools. Hence, there is a need for a robust model that can efficiently handle text data and provide measurable evidence of the differences between the two varieties.

The aim of this study is to perform a comparative corpus-based analysis of American English and Indian English texts using the software package R-Studio, which is a platform for statistical analysis and text mining. Through the employment of arithmetical data, such as the type-token ratio, lexical density, keywords, n-grams, and collocation networks, this study attempts to reveal hidden linguistic patterns that can be found in both varieties of the language. What is unique about this study is that it contributes to the methodological development in corpus linguistics, specifically in the use of reproducible computational methods, and to the theoretical development in the study of World Englishes by providing statistically informed evidence for language variation. Further, this study is also intended to contribute to the development in the study of World Englishes by providing statistically informed evidence for language variation, while also showing the value of using computational tools in linguistics, not only in increasing the precision of analysis but also in suggesting new avenues for exploring the study of World Englishes.

© 2026 ACADEMY PUBLICATION

According to Wickham and Grolemund (2016), R-Studio provides an all-inclusive environment for working with R and R code writing, data management, plotting, and package installation in a single interface. R-Studio assists in the smooth execution of quantitative and computational research analysis. It provides a user-friendly interface that enables script writing and editing at the fingertips of the researcher. For Winter (2019), R-Studio has become a popular tool in corpus linguistics and computational linguistics and data science. It provides support for advanced statistical analysis, text mining, and reproducible research through R Markdown and package installation. R-Studio, according to Xie et al. (2018), provides assistance in the management of large volumes of data and the creation of plots and graphs, which have made it an essential tool in corpus linguistic research and analysis.

This study, therefore, investigates the degree and type of linguistic variation between American English and Indian English through a quantitative approach using R-Studio for data analysis. The study raises the following three research questions: First, what are the significant statistical differences in the frequency distribution, lexical diversity, and lexical density of American English and Indian English? Second, what are the keywords, n-gram, and collocation networks of American English and Indian English, and how do they differ? Third, how can R-Studio assist in the reliable and reproducible investigation of linguistic variation between American English and Indian English? In accordance with the above research questions, the objectives of this study are as follows: First, to create two corpora of American English and Indian English and analyze the data using quantitative methods such as type-token ratio, keyness statistics, and collocation strength; second, to identify and analyze the most prominent and significant keywords and n-gram networks of American English and Indian English; and third, to demonstrate the effectiveness of using R-Studio as a tool for investigating linguistic variation between American English and Indian English through reproducibility and visualization of linguistic features.

II. Literature Review

A. American English and Indian English: A Brief Account

From the perspective of sociolinguistics, American English and Indian English have distinct paths in the development of world English. American English was developed as a native variety after gaining political independence from British rule. It gradually developed standards in spelling, pronunciation, and vocabulary that distinguish it from British English (Algeo, 2010). On the other hand, Indian English has developed in a multilingual post-colonial context, where English has been accepted as an associate official language and has become a lingua franca in all language communities. It has developed distinct features in vocabulary, pragmatics, and syntax that reflect local languages while at the same time being mutually intelligible with other English-speaking communities (Kachru, 1992; Schneider, 2007). This illustrates how English accommodates itself to social, cultural, and institutional contexts and thus confirms that there is no single variety of English, but rather a variety of related but distinct English varieties.

According to Mesthrie and Bhatt (2008), some of the differences between American English and Indian English lie in sounds, words, sentence structures, and the use of these two forms of English. In the sound system, Indian English is characterized by a syllable rhythm and the use of retroflex sounds, while American English is characterized by a stress rhythm and [r] sounds. In sentence structures, Indian English may use the progressive tense, for example, “I am understanding,” and may use “isn’t it?” in a variety of contexts, while American English may not use these forms. Indian English may also be characterized by formal forms of addressing people, while American English may be characterized by directness. These differences show how English is adapted in different contexts, with American English being a native Inner Circle variety and Indian English being an institutionalized Outer Circle variety in the context of World Englishes (Dodd et al., 2023).

B. World Englishes and Linguistic Variation

The concept of "World Englishes" offers a vital framework for understanding linguistic variation in different varieties of English around the world (Zhang & Yue, 2024). Kachru’s research on "World Englishes" led to the introduction of the "Three Circles Model," which categorizes English language usage in different regions of the world into Inner Circle (e.g., United States, United Kingdom), Outer Circle (e.g., India, Singapore, Nigeria), and Expanding Circle (regions where English is used as a foreign language). In this model, "American English" is a prominent variety of English belonging to the Inner Circle, while "Indian English" is a variety belonging to the Outer Circle with its own linguistic, pragmatic, and cultural characteristics (Kirkpatrick, 2007). Kachru (1992) argues that Outer Circle English varieties must be viewed as legitimate English varieties rather than "deficient English" varieties in comparison with Inner Circle English varieties.

This study follows the World English paradigm in its sociolinguistic approach, which assumes that there are different legitimate norms of English, rather than a monolithic norm of the language. According to this approach, American English is placed in the Inner Circle of Kachru’s Three-Circle Model of World Englishes, which provides norms for varieties of English, whereas Indian English is classified as part of the Outer Circle of Kachru’s model, where English has developed institutionalization, functional extension, and nativization in its local varieties (Kachru, 1985). Therefore, the lexicogrammatical and phraseological patterns observed in the data are local varieties of English that reflect different communicative practices and sociocultural identities of speakers of these varieties of English (Schneider, 2007; Alamri, 2023).

© 2026 ACADEMY PUBLICATION

Further studies in the area of World Englishes (e.g., Kachru, 1992; Mehrotra, 1998; Pal et al., 2022; Collins, 2023) have reinforced the need to take note of "systematic linguistic variation" in the development of the varieties of English around the world. Schneider (2007) developed the "Dynamic Model of Postcolonial Englishes" to illustrate the process by which new varieties of English develop and become stabilized in post-colonial contexts. According to this model, Indian English has passed through phases of nativization and endonormative stabilization, giving it unique lexical, grammatical, and discourse features. For example, Indian English has been found to have unique lexical, pragmatic, and collocational features, reflecting the influence of the native languages and cultures (Schneider, 2007).

C. Corpus Linguistics and the Empirical Study of Language Variation

Corpus linguistics is a powerful methodological approach for the investigation of language variation in the various varieties of the English language. It is based on the systematic analysis of large amounts of actual language data using computer technology (Brezina, 2018). According to Biber (1993), the method of corpus linguistics allows the investigation of language use empirically by analyzing actual language in use rather than relying on intuition. Some studies (e.g., Biber et al., 1998; Gries, 2016; Khafaga, 2023), therefore, find it possible to identify the recurring characteristics in language use, such as word frequency, phraseology, and grammar, in the various varieties of the English language. In comparative linguistic research, the method of corpus linguistics is useful in the investigation of language variation because it allows the measurement of language variation objectively.

One of the important contributions of corpus linguistics is that it reveals systematic differences in lexical and structural patterns between language varieties. McEnery and Hardie (2012) emphasize that corpus analysis can show the presence of distinctive linguistic features by using techniques such as frequency analysis, keyword analysis, and collocation analysis. Frequency analysis is the study of the frequency of occurrence of certain words or features in a corpus, whereas keyword analysis is the study of the occurrence of certain words in one corpus in comparison to another corpus. Collocation analysis, on the other hand, is the study of the occurrence of certain words in a particular context or setting. These techniques have been used in many studies on World Englishes, particularly in the study of lexical and discourse differences between American and Indian English.

In terms of the methodology of corpus linguistics, it is possible to use R to compare American English and Indian English texts in such a way that it contributes to the field of corpus linguistics (Reppen, 2010). At the same time, it also supports other theoretical approaches to language use, such as World English, which focuses on the variety of legitimate varieties of English spoken around the world. Therefore, it may be argued that the use of corpus linguistics methodology combined with computer tools such as R-Studio represents a powerful tool for analyzing the linguistic features of American and Indian English in contemporary texts. Biber et al. (1998) argue that the grammar and syntax of language operate on the basis of probabilities and not hard and fast rules. This has enabled the systematic study of different registers and varieties of language. McEnery and Hardie (2012) added that corpus linguistics has also enabled scholars to detect features such as keyness, collocation, and lexical bundles, which are features of conventional usage. Another significant contribution of corpus linguistics is that it has enabled scholars to detect subtle features of discourse, which are often overlooked (Baker, 2006). Significantly, in the context of world English, corpus linguistics has played a significant role in the identification and analysis of the structural and functional features of emerging varieties (Hundt & Mukherjee, 2011).

D. Computational Linguistics and the Use of R in Corpus-Based Research

Computational linguistics is an interdisciplinary approach that has the potential to be very powerful in the study and analysis of language using computers and computational methods (Khafaga, 2020). According to Jurafsky and Martin (2023), computational linguistics entails the study and creation of techniques that enable machines to computationally represent, process, and understand languages. Crucially, it has proven to be very effective in linguistics studies involving large-scale corpora since analyzing them would not have been easy without computer assistance (Khafaga & Alghawli, 2021). Further, the use of computer-based methods could help scholars identify linguistic patterns and conduct an analysis of language usage in different varieties of English. According to some scholars, one of the most significant contributions of computational linguistics has been the development of algorithms for text mining that facilitates the automatic extraction of information about language usage from large data sets. For example, Manning and Schütze (1999) claim that computational linguistics could be employed in identifying word frequency, word sense, syntax, and discourse patterns through computer-assisted analyses. Gries (2021) also argues that some of the analytical approaches that can be used for the purpose include tokenization, part of speech tagging, n-grams, and vector-space semantic modeling.

In studies of English language variation, computational linguistics plays a crucial role in identifying distinctive linguistic patterns among different English varieties. Studies in AmE versus IndE English can make use of computational linguistics to identify quantifiable differences in terms of lexical profile, phraseology, and semantic structure (Murphy, 2018). With the aid of statistical programming, researchers can establish significant differences between language variants in terms of word distribution, collocation, and discourse strategies. In R-Studio, for example, linguistic corpora can be analyzed using advanced techniques such as text mining. As Silge and Robinson (2017) note, text mining approaches in R, especially through packages such as "tidytext," "quanteda," and "tm," help researchers convert textual data into statistical data sets.

© 2026 ACADEMY PUBLICATION

Within the scope of computational linguistics, the investigation of texts written in American English and Indian English via the utilization of the R-Studio method is a useful way for examining linguistic variation. It not only helps linguists make accurate analyses but also enables them to understand why and how English language variants have spread worldwide. Gries (2016) highlights the significance of quantitative techniques in conducting linguistic research in order to test hypotheses related to language variation and identify distribution patterns. The most remarkable feature of the R environment is its flexible nature regarding large-scale datasets and complicated statistics and visualization techniques. Silge and Robinson (2017) have presented an approach to tidy text mining, which enables users to calculate TF-IDF, lexical diversity measures, and collocation statistics in an unambiguous and reproducible fashion. The use of computational tools has also made it possible to utilize keyness statistics, including log-likelihood and chi-square tests, to evaluate whether observed differences between corpora are statistically significant (Evert, 2008; Murphy, 2018; Khafaga & Shaalan, 2020; Gillings & Kopf, 2024).

Despite the fact that existing research has provided interesting insights regarding the structural, lexical, and functional variations of American English and Indian English, most of it has continued to be qualitative or relied on corpora without effectively engaging with computation to make it reproducible and statistically sound. Also, previous research has focused on variations of words, collocations, and phrasing without an overarching framework that integrates it with trends of frequency and sociocultural analysis. This study fills this gap by combining world Englishes research with corpus linguistics and computation through R-Studio to provide a sound statistical analysis of variations of American English and Indian English.

III. Methodology

To investigate the differences between texts written by American writers and Indian writers, a corpus of 60 business-related online articles was collected. 30 were from American writers and 30 from Indian writers. The corpus was collected from "The New York Times" website (http://www.nytimes.com/) under the "business" section, where 30 top articles were selected for the American part, and from the "The Times of India" website (http://timesofindia.indiatimes.com/) under the "business" section, where another 30 top articles were selected for the Indian part. The entire corpus was processed using AntConc to create frequency profiles for all the documents, and subsequently these profiles were used for processing in R.

The entire corpus was used both for training and testing the classifiers by choosing a random sample of 40 articles to be the training set, with the only restriction that 20 articles were American and 20 were Indian, and the remaining 20 articles (10 American and 10 Indian) were the testing set. To avoid bias errors or other artifacts, an N-fold rotation technique was applied, wherein in each of N=20 repetitions a different random sample of 20 American and 20 Indian texts was used for training and the rest for testing.

For text classification, the number of used types (unique tokens) in the dictionary was also investigated. This process was performed twice, once using all possible types from the training set (referred to as "Full Dictionary") and then using only 200 types that were chosen as the most significant by their ability to differentiate the two classes of texts (referred to as "Reduced Dictionary"). As far as the number of tokens is concerned, the total tokens for the Indian training sets were about 8700 for the full dictionary and about 1700 for the reduced dictionary and for the American training sets, they were about 18000 for the full dictionary and about 3200 for the reduced dictionary. Also, the study investigated the effect of data weighting in the classification results when using a Naïve Bayes classifier, and a good description of this method can be found in Duda and Hart (1973) and McCallum and Nigam (1998). The study also used the tf-idf weighting technique as it is described in Aizawa's (2003) words as "an information-theoretic perspective of tf–idf measures". The classification methods that were used in the investigation were:

1. Correlation Metric
2. Naïve Bayes classifier with nominal data
3. Naïve Bayes classifier using tf-idf weighted data

The results of each classification method were evaluated using the standard metrics of Precision, Recall, and F-measure that are defined as follows:

$$P = \frac{N_{tp}}{N_{tp} + N_{fp}},$$

$$R = \frac{N_{tp}}{N_{tp} + N_{fn}},$$

$$F = \frac{2PR}{P + R},$$

where P, R, and F stand for Precision, Recall, and F-measure, respectively and N_{tp}, N_{fp} and N_{fn} are the numbers of true positives, false positives, and false negatives, respectively.

Considering the use of full or reduced dictionary and the different classification methods, there are six sets of results in total summarized with their acronyms in Table 1.

© 2026 ACADEMY PUBLICATION

TABLE 1
FULL AND REDUCED DICTIONARY AND THEIR CLASSIFICATIONS

Dictionary	Data Weighting	Method	Acronym
Full	No	Correlation	Cor
Full	No	Naïve Bayes	NB
Full	Yes	Naïve Bayes	WNB
Reduced	No	Correlation	DCor
Reduced	No	Naïve Bayes	DNB
Reduced	Yes	Naïve Bayes	DWNB

The implementation details of each classification method and other processing techniques that were used are clarified in the following subsections:

A. Dictionary Selection

Using all possible types may not be beneficial for classification since very common terms do not provide useful information to train the classifier. Thus, a proper dictionary selection where some types that are significant in the distinction of the two classes were employed, which has the potential to improve the classification results. To select the dictionary, a method based on finding the document frequency (N_d) for each type was used, individually in each class. To find the document frequency, the frequency profiles were loaded from the training set and merged the tables from each class, creating two tables of 21 columns each, where the first column is the token and the remaining 20 are the absolute frequencies for each token in each document. Since the core concern of the study is on the presence ("1") or absence ("0") of a token, all positive frequencies (>0) were converted to "1"s. Then, to find the document frequencies N_d for each token, I summed across the columns for each row, thus collapsing each table to a two-column table. The next step was the merging of the two class tables based on token and replacing all NA values with "0". This table now has three columns, where the first is the token, the second is the document frequency of a token in the Indian texts (denoted as N_{dI}), and the third is the document frequency of a token in the American texts (denoted as N_{dA}). For a particular token, a high value for N_{dI} combined with a low value for N_{dA} would be indicative of the significance of the term in filtering Indian texts and the opposite (low N_{dI} and high N_{dA}) would be indicative of the significance of the term in filtering American texts.

B. Data Weighting

In certain cases it is not beneficial to consider all token unigrams as suitable for classification. This happens because some very commonly used words appear in all texts and thus do not provide useful information, adding to the efficiency of a classification task. With data weighting more emphasis was placed on rare or unique terms that are more informative of the nature of the text. In this regard, the tf-idf weighting was used. To create the weights, the frequency profiles for the training documents were loaded and merged into one table with 41 columns, where the first column is the token and the remaining 40 are the absolute frequencies for each token in each document. Any NA entries in the table were converted to "0". Then, to denote the presence or not of a token in a document, the table was converted to a binary table where all the positive frequency values were converted to "1" (denotes presence) and the "0"s (denotes absence) remained unchanged. To find the number of documents where each token is present, denoted as N_d, I summed across the columns for each row, thus collapsing the table to one that has only two columns. The first is still the token, and the second now contains the number N_d for each token. The number N_d is anywhere from 1 to 40, where 40 is the total number of documents, denoted as N. The table of weights was then calculated using the idf coefficient formula: $w = \log_2(\frac{N}{N_d})$.

C. Document Classification

To perform document classification, 20 Indian texts and 20 American texts were randomly chosen for the training set. Then these frequency profiles are read into R tables, and the 20 tables from each class are merged based on the token column, and the NA values are replaced with "0" to create the table T_I and T_A. To create total frequency profiles, the two tables from each class (T_I and T_A) are collapsed to tables of two columns by summing the absolute frequencies across columns for each row to create tables T_{IT} and T_{AT}. A third column is added to each table that contains the relative frequency (probability) of each token by dividing each absolute frequency with the total number of tokens for each class. The final training table T_T is created by merging T_{IT} and T_{AT} based on token and replacing all NA values with "0". A training table that uses a reduced dictionary denoted as T_{TD} was created, and also the tf-idf weights and weighted training tables based on the full dictionary (denoted as T_{TW}) and on the reduced dictionary (denoted as T_{TDW}) were created. The classification process uses each text from the testing set (by isolating the appropriate column) and the training tables T_T, T_{TW}, T_{TD} and T_{TDW}. The outcome of classification is recorded as "1" for correct classification and as "0" for wrong classification. Using the classification outcomes, the number of true-positives, false-positives and false-negatives for Indian and American classification were calculated, and using these numbers, the values for Precision, Recall, and F-measure were also calculated.

D. Correlation Metric Classifier

© 2026 ACADEMY PUBLICATION

The correlation classification method, as described by Banerjee and Lavie (2005), requires a testing table consisting of 2 columns, one with tokens from the test document and one with absolute frequencies. This table is acquired as a subset of the total testing table. Then the relative frequencies (probabilities) are calculated. Also, the training table that contains all the training tokens is needed (from a full or reduced dictionary), and the relative frequencies of these tokens for the Indian and American classes are also required. Once the required relative frequencies are obtained, the correlation between the testing and Indian training frequencies to get ϱ_I is calculated, and then the correlation between the testing and American relative frequencies to get ϱ_A. If $\varrho_I > \varrho_A$ is calculated. The text is classified as Indian and, in the opposite case, as American. Given that the origin of the testing text is known, it can be decided whether or not the text was classified correctly, and the method will return "1" for correct classification and "0" for misclassification.

E. Naïve Bayes Classifier

The Naïve Bayes classifier (Duda & Hart, 1973) requires a testing table consisting of the unique tokens (types) appearing in the testing text along with their relative frequencies. This table is got by extracting the appropriate columns of the total testing table and keeping only the rows with non-zero absolute frequencies. Then, the relative frequencies are created by dividing by the total number of tokens. Also, the relative frequencies are needed for each class that are contained in the training table T_T (T_{TW}, or T_{TD}, or T_{TDW} depending on whether or not weighting is used and whether a full or reduced dictionary is used). The next step is to merge the testing table with the training table based on tokens, by keeping all the entries of the testing table, since classification is based on tokens that exist in the training set. Tokens of the training set that do not appear in the training table (dictionary) will have NA values that are converted to "0"s. These "0" are not useful for the Naïve Bayes classifier and need to be converted to meaningful values. One way is to use smoothing by essentially incrementing the absolute frequencies by one and recalculating the relative frequencies. This method does not work well in our case because there is a big difference in the number of total tokens in the Indian and American classes (American class has about twice as many total tokens). This creates a problem since the default probability of a non-existing term in the Indian class would be $1/N_{TI}$ and for the American class would be $1/N_{TA}$ and given that $N_{TA} > N_{TI}$ the probabilities for the Indian class would be much bigger compared to the American ones for non-existing tokens. This significant difference in the default probabilities of non-existing terms would bias the classification results leading to most training texts being classified as Indian. To avoid this problem, the "0" probabilities are substituted with a default fixed small probability chosen to be 10^{-6}.

Once the relative probabilities for each class are calculated, then the Naïve Bayes classification is performed by summing the log of the probabilities, and the cost values are calculated as follows:

$$c_I = \sum_i \log_2\big(P(t_i|I)\big),$$

$$c_A = \sum_i \log_2\big(P(t_i|A)\big).$$

If $c_I > c_A$, the text is classified as Indian and, in the opposite case as American. Given that the origin of the testing text is known, it can be decided whether or not the text was classified correctly, and the method will return "1" for correct classification and "0" for misclassification.

IV. Results

A. The Classification Results for Indian and American Texts

The classification results from each method were quantified using the Precision, Recall, and F-measure metrics, and the average results from N-fold rotation repeated for N=20 times are summarized in Table 2.

Table 2
Classification Results for Indian and American Texts

	Cor	DCor	NB	DNB	WNB	DWNB
Indian text results						
Precision	0.758	0.828	0.926	0.845	0.926	0.926
Recall	0.720	0.675	0.660	0.800	0.660	0.660
F-measure	0.721	0.733	0.757	0.814	0.758	0.758
American text results						
Precision	0.751	0.734	0.750	0.815	0.750	0.795
Recall	0.750	0.850	0.950	0.835	0.950	0.875
F-measure	0.739	0.782	0.834	0.817	0.834	0.827

From the results summary in Table 2, some useful observations regarding the different classification methods can be made, as follows:

Considering the F-measure as the metric of overall performance of the classification method, it is noticed that the Naïve Bayes classifier (with and without tf-idf weighting) outperforms the correlation-based classification. The same holds for the Precision metric and the Recall for American classification, whereas in Indian classification the Recall of the Naïve Bayes classifier is lower compared to the correlation method in some cases.

© 2026 ACADEMY PUBLICATION

Another useful observation is about the dictionary size. Using a reduced dictionary improves the average performance of the correlation method as given by the F-measure for both classes. The reduced dictionary in the correlation method results in more texts being classified as American, leading to more false positives in American classification and, equivalently, more false negatives in Indian classification. Thus, the Recall for American classification goes up, as does the Precision for Indian classification, and on the other hand, the Precision for American classification and Recall for Indian classification go down. Despite these changes, the overall performance is improved. Using a reduced dictionary for Naïve Bayes classification without data weighting improves significantly the F-measure of Indian classification but reduces slightly the F-measure for American classification. Overall, the metrics are more balanced (Precision, Recall, and F-measure are all above 0.8 for both classes) when using a reduced dictionary, whereas the full dictionary leads to high Precision and low Recall for Indian classification and low Precision and high Recall for American classification, which is an indication of a biased preference in classifying texts as American. Using a reduced dictionary with Naïve Bayes and data weighting does not affect the results of Indian classification but slightly affects the performance of American classification with increased Precision, reduced Recall, and slightly reduced F-measure.

The final observation concerns the use of data weighting using tf-idf weights. For Naïve Bayes with a full dictionary, the use of weights does not alter the classification results for either class. In this case the weighting method is neutral, and its use is neither beneficial nor detrimental. For the Naïve Bayes with a reduced dictionary, the change in classification results is significant. It improves slightly the F-measure for American classification but reduces significantly the F-measure for Indian classification. Also, the Precision and Recall results for both classes become less balanced. This means that the use of weights in combination with a reduced dictionary cannot be considered beneficial for classification.

B. Lexical Frequency and Diversity

The frequency analysis of the selected data demonstrates differences between American English (AmE) and Indian English (IndE) in terms of shared high-frequency vocabulary and overall vocabulary composition. AmE texts are concentrated around standardized and high-frequency function words and phrases typical of codified English used in Inner Circle countries. On the other hand, the vocabulary of IndE is more varied and employs local terms and expressions and mixed vocabulary. The Type Token Ratio (TTR) of IndE was slightly higher than that of AmE with a value of 0.124 compared to 0.109, suggesting a high level of vocabulary variation in IndE texts with comparable lengths. In addition, there are also discrepancies in the application of density, with more nominalizations occurring in AmE compared to more verbalizations in IndE following past studies of Outer Circle countries' linguistic adaptability in terms of functions (Biber et al., 1998; Schneider, 2007). Table 3 illustrates the list of 10 most frequent words in both AmE and IndE with frequency per 10,000 words.

TABLE 3
TOP 10 MOST FREQUENT LEXICAL ITEMS IN AME VS. INDE

Rank	American English (per 10k words)	Indian English (per 10k words)
1	the (690)	the (678)
2	of (472)	of (461)
3	and (415)	and (432)
4	to (392)	to (389)
5	in (314)	in (310)
6	that (198)	kindly (182)
7	is (185)	honor (179)
8	for (172)	as (174)
9	on (158)	dear (168)
10	with (149)	with (164)

Table 3 shows that American English contains high-frequency function words typical of Inner Circle standardization. On the other hand, Indian English contains culturally marked words like *kindly*, *honor*, and *dear*, typical of Outer Circle politeness and formulaic language use. This, in turn, demonstrates higher lexical variety in Indian English.

C. Keyword Analysis

Keyword analysis using log-likelihood identified words that are more prominent in each corpus. In American English, the keywords are related to formal style and institutional discourse (e.g., policy, authority, government). In Indian English, the keywords are related to culture-specific words and discourse markers (e.g., honor, kindly, dear). This indicates that Indian English has polite words and expressions that are reflective of the sociocultural context of Indian English (Mehrotra, 1998; Kirkpatrick, 2007). Table 4 shows the top 10 keywords of each variety according to their log-likelihood scores.

© 2026 ACADEMY PUBLICATION

TABLE 4
TOP KEYWORDS IN AME VS. INDE

Rank	American English Keywords	Log-Likelihood	Indian English Keywords	Log-Likelihood
1	policy	45.7	kindly	52.1
2	government	41.2	honour	47.8
3	authority	38.5	dear	44.3
4	research	35.9	please	42.7
5	administration	33.4	respect	40.2
6	security	30.8	attention	37.6
7	committee	28.7	necessary	35.0
8	development	26.5	due	33.8
9	agreement	25.1	kindly do	32.4
10	legislation	23.7	information	30.9

Table 4 clarifies the positions and frequencies of the top 5 keywords in both corpora. Keywords in American English are related to formal themes, while those in Indian English are related to polite formulas.

D. N-Gram and Collocation Patterns

A more detailed examination of bigrams and trigrams indicates the presence of clear phrase patterns that distinguish the two English varieties. For instance, American English makes significant use of formal lexical sets such as on the basis of, as a result of, and in terms of, which are more appropriate in academic and journalistic contexts. Contradictorily, Indian English makes use of local pragmatics and expressions of courtesy, such as "kindly do the needful" and "with due respect," which indicate the use of mixed styles of discourse. Further, if we analyze the collocation patterns with the use of mutual information scores, strong and distinctive patterns are indicated for each variety. This indicates the use of rules and conventions, as opposed to mere variation or chance, in the use of the two English varieties (Schilk, 2006; Viana & Kirk, 2021).

TABLE 5
TOP BIGRAM IN AME VS. INDE

Rank	American English Bigrams	Frequency	Indian English Bigrams	Frequency
1	in terms	68	kindly do	74
2	on the	61	with due	69
3	as a	58	as per	65
4	in order	54	with reference	62
5	on basis	49	with regard	59

V. DISCUSSION

The above results show that the quantitative findings are in line with what is predicted by the theory of World Englishes. To clarify, American English is like Inner Circle English in its preference for standardized word patterns, especially in terms of its use of lexical bundles. The other way that the English used by Indians is formal is the language usage itself, which correlates with Wiltshire's (2020) and Pal et al.'s (2022) arguments that the English used in India is viewed to have its own distinct characteristics in the form of its politeness formulae, borrowing, blending of loanwords, and flexibility in collocation use. All of these are considered indicators of Outer Circle nativization and sociolinguistic adaptation. According to the analysis presented above, it is evident that the vocabulary, word pairs, and phrase differences in the two forms of English used are no coincidence but are linked to local practices, multilingualism, and culture. The study has provided an extensive comparative analysis of the two languages with the combination of quantitative data analysis using R-Studio software and data interpretation through the World Englishes approach. The analysis also demonstrates that the utilization of the features provided by R-Studio is helpful in creating networks of collocations and graphical representations of frequencies of occurrence in the case of American and Indian English utilizing R-Studio software. The differences between the languages have been found to be significant through the use of statistical tests, i.e., Chi-square test and Log-likelihood tests. These differences prove that IndE and AmE differ significantly from each other in terms of their functionality and sociolinguistic orientation.

Concerning the frequencies and diversities in lexical items, the findings show the patterns of American English and Indian English. In the case of American English, the texts use many common function words and preformed lexical units following the Inner Circle norms. Indian English shows a relatively high type/token ratio, and it uses culturally colored words such as "kindly," "honor," and "dear." Such findings support earlier studies by Biber et al. (1998), Mehrotra (1998), Schneider (2007), and Murphy (2010), who found that Outer Circle English showed more variety of word types due to nativization and multilingualism. The novelty in this study, therefore, lies in having quantitative data; that is, normalizing frequencies and analyzing them using R-Studio provide empirically sound evidence of the variability between American and Indian English.

In terms of keyword analysis, it is evident from the results that American English employs more institutional and formal terms, such as policy and government, whereas Indian English uses more polite language and regionalism. The findings are consistent with Kirkpatrick's (2007) assertion that Indian English is marked by formulaic politeness that

© 2026 ACADEMY PUBLICATION

indicates hierarchical and cultural traditions. The findings of this study corroborate the assertions of Viana and Kirk (2021) regarding genre-based variations of academic texts through international corpus of English (ICE) corpora. The findings of this study further the assertions of previous studies on Outer Circle Englishes because they have tested their assertions statistically and plotted dispersion of the keywords.

As for n-gram and collocation patterns, there are some differences in how phrases are formed. For example, AmE favors the use of phrases such as "in terms of" and "as a result of," consistent with the results obtained by Biber (2009) when studying formulaic language used in formal varieties of Inner Circle Englishes. However, IndE often employs phrases such as "kindly do the needful" and "with due respect," reflecting socially embedded linguistic varieties with respect to the use of politeness strategies, as discovered by Schilk (2006) in ICE India varieties of collocations. Therefore, it can be said that Outer Circle English varieties have not adopted Inner Circle varieties of English but developed a novel, rule-governed variety of English language use. Significantly, the results show that the MI scores in R-Studio helped obtain a quantifiable measure of collocational strength, thus reinforcing earlier qualitative studies.

Still, the analysis shows that both chi-square and log-likelihood tests prove that the variations with regard to keywords, n-grams, and collocations are statistically significant at $p < 0.05$. This finding correlates with Gries's (2016) argument on the importance of quantitative validation with regard to corpus-based research. While previous research on ICE data sets, such as Schilk (2006) and Viana and Kirk (2021), depend on simple statistical tests or even counts, the present study demonstrates its commitment to reproducibility and openness through its use of R-Studio's workflow. Such an approach to methodology strengthens claims of systematic variations and provides a template for subsequent comparative research on World Englishes. Furthermore, these results reconcile with Kachru's (1985) *Three-Circle model* and Schneider's (2007) *Dynamic Model* of varieties of English. American English follows the pattern of Inner Circle varieties, with a standardization of lexical bundles and a small number of lexical items. Indian English follows the pattern of Outer Circle varieties, with flexible word selection, polite language, and collocations. These results are consistent with previous research (Mehrotra, 1998; Hundt & Mukherjee, 2011; Wiltshire, 2020), and this research extends that work with quantification. With this examination of lexical, collocational, and phraseological patterns in a unified framework using R-Studio, it is possible to demonstrate the systematic nature of each of these features and to emphasize the role of context in IndE discourse. Consequently, this study integrates different elements in one entity which had been studied separately before, i.e., frequency, keyness, n-grams, and collocations, based on statistical support and visualization. As such, this study contributes to the World Englishes research field by proving the possibility of using R-Studio for turning theoretical ideas into testable results through the integration of qualitative perceptions and quantitative statistics.

VI. Conclusion

This paper examined and contrasted AmE and IndE from the perspective of World Englishes' theoretical framework, alongside a computational and corpus-based approach using R-Studio software. As per the results, some fascinating trends have been observed in the areas of vocabulary, collocations, and phraseology in both varieties. It was found that AmE conforms to Inner Circle principles, whereas IndE has higher variation in its vocabulary use, usage of polite speech within cultural contexts, and mixed collocation in nativized English of the Outer Circle. Furthermore, keyword visualization and collocation networks provide reproducible and understandable findings. The paper used six different classification methods to investigate data collected from both American and Indian writers. The classifiers were based on two major methods (correlation and Naïve Bayes) using variations in dictionary selection (full and reduced) and data weighting (tf-idf and no weighting). Among the methods used, Naïve Bayes with reduced dictionary and without data weighting gave the most balanced results. The results seem promising in supporting the assumption that linguistic differences of the two varieties lead to differences in the way of writing.

A. Theoretical, Methodological, and Pedagogical Implications

This study has theoretical, methodological, and pedagogical implications. Theoretically, the study confirms the theories of World English and Schneider's (2007) Dynamic Model and demonstrates that Outer Circle English, like Indian English, thoughtfully modifies its vocabulary and phraseology to suit social and cultural contexts. Methodologically, this study shows how the combination of corpus linguistics with computation in R-Studio proves useful in practice, as it gives an example of the methodology to be used in studying variations in English in future studies from the quantitative and qualitative perspectives. Pedagogically, English teachers in India and elsewhere should incorporate the present findings in designing teaching material based on the actual linguistic and sociocultural characteristics of Indian English. Lexicographers can also incorporate the identified words and expressions of Indian English lexical bundles, polite expressions, and collocations into dictionaries.

B. Limitations and Recommendations for Future Research

Despite providing good quantitative results, there is a limitation to the study based on the scope of the corpora used. While the data is well-balanced and made comparable, it would be interesting to use larger and more diverse corpora that included spoken data, social media, and other informal varieties of spoken language to reveal even more nuances of Indian English variation. Besides the lexical and collocational aspects of Indian English that have been examined in this

© 2026 ACADEMY PUBLICATION

study, exploring even more aspects of its syntax, pragmatics, and phonology would provide even more interesting results and a more complete picture of the variety.

For future research, the results of the current study can be extended in future studies by incorporating spoken data and multimodal texts to explore changes in oral communication patterns and pragmatic variation. Moreover, employing a diachronic corpus will enable the analysis of the process of nativization as well as the creation of new words in Indian English. By making comparisons between Indian English and other varieties of Outer Circle English, such as Singapore and Nigerian English, employing the same processes and workflow in R-Studio, certain systematic features can be observed in the data. The combination of syntactic and semantic analyses with lexical features can also provide a more multidimensional view of linguistic features of language varieties.

ACKNOWLEDGEMENTS

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2604).

REFERENCES

[1] Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing and Management*, *39*, 45-65.
[2] Alamri, B. (2023). A multidimensional comparative analysis of MENA and international English research article abstracts in applied linguistics. *SAGE Open*, *13*(1), 1-14.
[3] Algeo, J. (2010). *The origins and development of the English language* (6th ed.). Wadsworth.
[4] Baker, P. (2006). *Using corpora in discourse analysis*. Continuum.
[5] Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65-72). Ann Arbor.
[6] Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243-257.
[7] Biber, D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics*, *14*(3), 275-311.
[8] Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
[9] Bohmann, A. (2019). *Variation in English worldwide: Registers and global varieties*. Cambridge University Press.
[10] Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.
[11] Collins, P. (2023). Modals and quasi-modals in English world-wide. *Journal of English Linguistics*, *51*(3), 265-293.
[12] Dodd, N., Cohn, M., & Zellou, G. (2023). Comparing alignment toward American, British, and Indian English text-to speech (TTS) voices: influence of social attitudes and talker guise. *Frontiers in Computer Science*, *5*, 1-13.
[13] Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. Wiley and Sons.
[14] Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 1212-1248). Mouton de Gruyter.
[15] Gillings, M., & Kopf, S. (2024). Business communication through a corpus linguistic lens. *International Journal of Corpus Linguistics*, *29*(3), 297-301.
[16] Gries, S. T. (2016). *Quantitative corpus linguistics with R: A practical introduction*. Routledge.
[17] Gries, S. T. (2021). *Statistics for linguistics with R: A practical introduction* (3rd ed.). De Gruyter.
[18] Hundt, M., & Mukherjee, J. (2011). *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*. John Benjamins.
[19] Jurafsky, D., & Martin, H. (2023). *Speech and language processing* (3rd ed.). Pearson.
[20] Kachru, B. (1992). *The other tongue: English across cultures* (2nd ed.). University of Illinois Press.
[21] Khafaga, A. (2020). A computational approach to explore the extremist ideologies of Daesh discourse. *International Journal of Advanced Computer Science and Applications*, *11*(8), 193-199.
[22] Khafaga, A., & Shaalan, I. (2020). Using concordance to decode the ideological weight of lexis in learning narrative literature: A computational approach. *International Journal of Advanced Computer Science and Applications*, *11*(4), 246-252.
[23] Khafaga, A., & Alghawli, A. S. A. (2021). The impact of CALL software on the performance of EFL students in the Saudi university context. *International Journal of Advanced Computer Science and Applications*, *12*(7), 304-312.
[24] Khafaga, A. (2023). Strategic lexicalization in courtroom discourse: A corpus-assisted critical discourse analysis. *Cogent Arts & Humanities*, *10*(1), 1-28.
[25] Kirkpatrick, A. (2007). *World Englishes: Implications for international communication and English language teaching*. Cambridge University Press.
[26] Manning, D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
[27] McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, pp. 41-48).
[28] McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
[29] Mehrotra, R. (1998). *Indian English: Texts and interpretation*. John Benjamins.
[30] Mesthrie, R., & Bhatt, R. (2008). *World Englishes: The study of new linguistic varieties*. Cambridge University Press.
[31] Murphy, L. (2010). *Lexical meaning*. Cambridge University Press.
[32] Murphy, L. (2018). *The prodigal tongue: The love-hate relationship between American and British English*. Penguin Books.

© 2026 ACADEMY PUBLICATION

[33] Pal, P., Jain, S., Vuppala, A. K., Yarra, C., & Ghosh, P. K. (2022). *Study of Indian English pronunciation variabilities relative to received pronunciation*. arXiv. https://doi.org/10.48550/arXiv.2204.06502
[34] Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge University Press.
[35] Rizwan, A. (2009). Characteristics of Indian English. *ELTWeekly*. Retrieved October 11, 2025, from http://eltweekly.com/more/2009/10/37-research-paper-characteristics-of-indian-english/
[36] Saadawi, A. M. (2025). A comparative analysis of linguistic disparities between American and British English. *Journal of the Faculty of Arts*, *26*(75), 39-62.
[37] Schilk, S. (2006). Collocational patterns in Indian English: A corpus-based study using ICE-India and ICE-GB. *Anglistik: International Journal of English Studies*, *17*(2), 145-162.
[38] Schneider, E. W. (2007). *Postcolonial English: Varieties around the world*. Cambridge University Press.
[39] Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media.
[40] Viana, V., & Kirk, J. (2021). Academic prose across countries: An investigation of the humanities and technology texts in the international corpus of English. In A. Čermáková & M. Malá (Eds.), *Variation in time and space: Observing the world through corpora* (Vol. 20, pp. 283-320). De Gruyter.
[41] Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media.
[42] Wiltshire, C. (2020). *Uniformity and variability in the Indian English accent*. Cambridge University Press.
[43] Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.
[44] Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R Markdown: The definitive guide*. Chapman & Hall/CRC.
[45] Zhang, Y., & Yue, M. (2024). Case and agreement variation in contact: A multifactorial investigation of it-clefts across World Englishes. *International Journal of Corpus Linguistics*, *29*(4), 472-506.



Mansour Ibrahim Altamimi is an associate professor in the Department of English, College of Languages and Translation at Imam Muhammed Ibn Saud Islamic University (IMSIU), Saudi Arabia. He got his PhD in linguistics and applied linguistics from Arizona State University in 2021, USA. His research interests include phonology, syntax, applied linguistics, sociolinguistics, and discourse analysis. He is also interested in language change and computational linguistics.

© 2026 ACADEMY PUBLICATION