Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2, 1245-1259
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i2.4746

Automatic author attribution of different languages: A review

Smriti Priya Medhi[1,2*], Shikhar Kumar Sarma[1]
[1]Department of Information Technology, Gauhati University, Guwahati, Assam, India; sp.medhi26@gmail.com (S.P.M.).
[2]Department of Computer Science and Engineering, Assam Don Bosco University, Guwahati, Assam, India.

Abstract: This paper presents a comprehensive analysis of the evolution of writing styles over a century and its implications for author identification. We analyze a diverse corpus of texts spanning a hundred years, focusing on linguistic features such as vocabulary, syntax, and discourse patterns. Our findings reveal significant shifts in writing styles over time, influenced by cultural, technological, and social factors. Furthermore, we investigate the impact of these changes on the task of author identification, a crucial area in forensic linguistics and stylometry. We demonstrate how traditional methods of authorship attribution may be challenged by the evolving nature of language use. However, we also highlight the potential for machine learning techniques, such as deep learning models, to adapt to these changes and improve author identification accuracy. Overall, this study sheds light on the dynamic nature of writing styles and the challenges and opportunities they present for author identification in the digital age, with special importance to low-resource languages.

Keywords: *Author identification, Linguistics, Writing styles.*

1. Introduction

The written word shows how smart and creative humans are. It keeps our history safe and changes as our society changes. In the last hundred years, writing has changed a lot, showing how our culture, technology, and education have changed too. When we look at how writing styles have changed over time, we see a beautiful mix of new ideas and ways of writing. Studying writing styles isn't just about understanding how languages are used. It also shows us what's important to a society. It tells us how language has changed to fit new ideas, technologies, and ways of thinking. From the fancy writing of the early 1900s to the quick, to-the-point writing we see today, writing styles show us what each time period was like.

These changes are important, especially when trying to figure out who wrote something based on how it's written. Writing styles can be very different in different languages, shaped by how the language is structured, the culture it comes from, and the time in history when it was used. Each language has its own special way of writing that affects how ideas are put into words. Things like grammar, vocabulary, how formal or polite the writing is, and the culture and history behind the language all play a part in shaping its writing style.

Old ways of figuring out who wrote something, using fixed language features, might struggle to keep up with how writing styles change. But with machine learning, especially deep learning models, we might be able to do a better job of identifying authors, even as writing styles change.

Through this study, we want to uncover the complex relationship between writing styles and figuring out who wrote something over the last hundred years. By looking closely at how language has changed in each time period, we hope to understand the patterns of change and similarities that have shaped our writing. Through this exploration, we hope to not only learn more about writing styles but

History: Received: 4 December 2024; Revised: 17 January 2025; Accepted: 20 January 2025; Published: 12 February 2025
* Correspondence: sp.medhi26@gmail.com

also find and understand the ways how AI models learn to figure out who wrote something in a world where language is always changing.

2. Study

2.1. Graphical Abstract

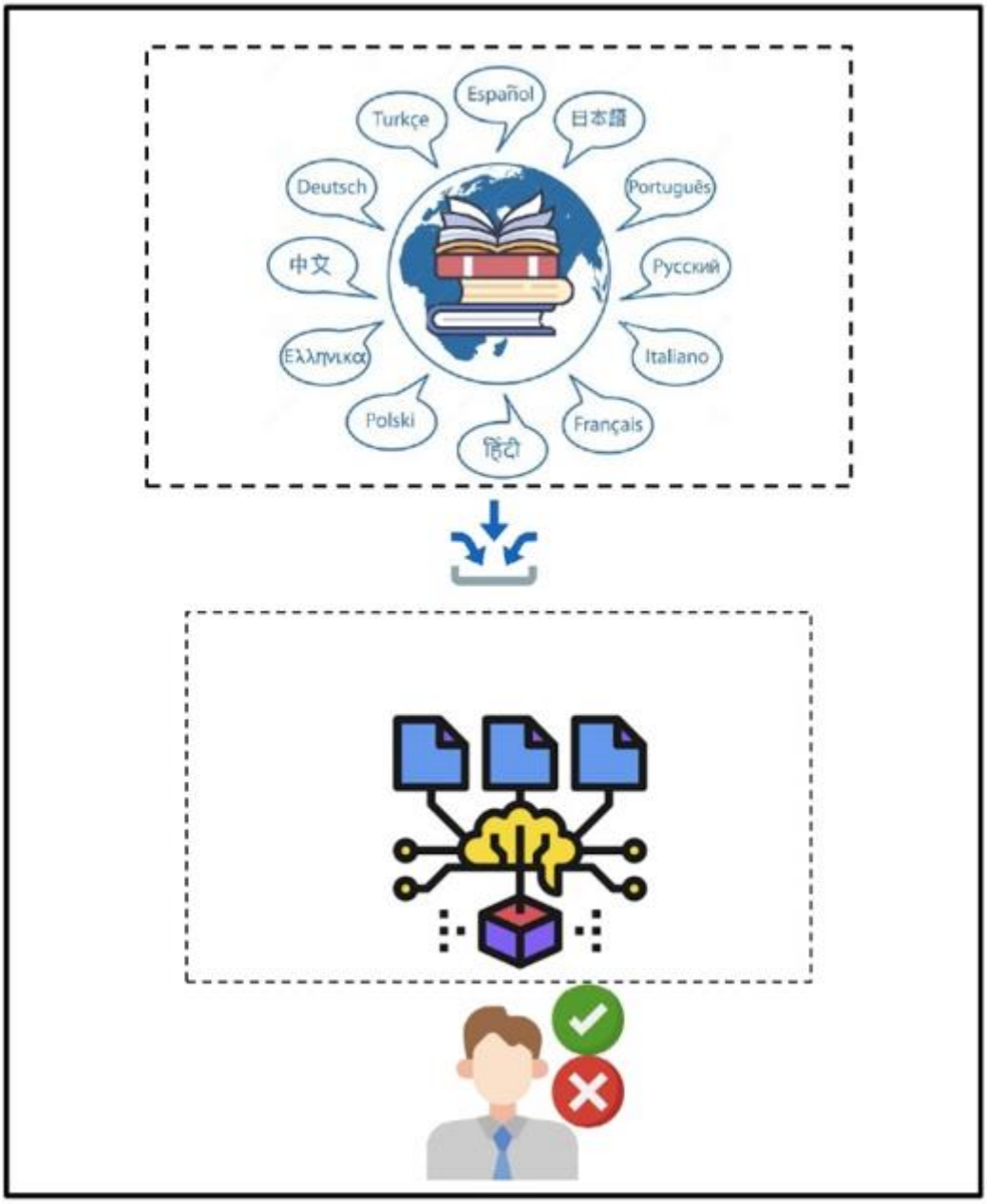

Figure 1.
A graphical illustration of authorship attribution methodology using machine intelligence.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 1245-1259, 2025
DOI: 10.55214/25768484.v9i2.4746

2.2. Origin of Language

The origin of language is shrouded in mystery, but several hypotheses like continuity, discontinuity and theories like tool-making, social interaction, gesture, mirror-neuron etc. have emerged to explain how human language might have originated. These theories are speculative, as they are based on indirect evidence from various fields like anthropology, evolutionary biology, cognitive psychology, and linguistics. The exact timing of the origin of language is still a matter of debate and speculation, as there are no direct records from the period when language first began. However, most estimates place the emergence of language somewhere between 50,000 to 100,000 years ago, with some theories suggesting that it could be even older, up to 200,000 years ago or more.

2.3. Evolution of Language

Language, which is believed to be a subset of cognition saw its evolution from speech to writing through a long and complex process that began with simple symbolic representations and gradually developed into the sophisticated writing systems we use today.

2.3.1. Early Writing Forms and Scripts

Starting with the token system, which represented goods were typically used for accounting purposes dates back to 8000-3000 BCE. They were purely symbolic and had no direct relation to spoken language. With further developments, symbols began to represent not just tangible objects but also concepts and ideas. Pictograms (pictures representing objects) and ideograms (symbols representing ideas or actions) are considered some of the earliest forms of writing, with some evolving from the token system.

Over time, these pictorial representations became more stylized and abstract, allowing for a more versatile and efficient way to convey information. This is evident in the cuneiform script of Mesopotamia and Egyptian hieroglyphs. Further abstraction led to the creation of syllabaries, where symbols represented syllables rather than entire words or concepts. This allowed for the representation of a vast range of words with a limited set of symbols. The next major step was the development of phonetic writing systems that could represent the sounds of spoken language, including vowels and consonants. This allowed for the precise transcription of speech.

The alphabet represents a significant evolution in writing, with each letter corresponding to a single phoneme, or sound, of the spoken language. The first true alphabet, containing both consonants and vowels, is believed to be the Greek alphabet, which evolved from the Phoenician script. As writing became more widespread, it was standardized within cultures and began to be used for a broader range of purposes beyond accounting, such as for literature, law, and personal correspondence.

Finally, technological advancements such as the invention of the printing press greatly increased the accessibility and distribution of written materials, solidifying the importance of writing in human societies. All this information is nicely curated in [1].

Thus, looking back in time, we can say that majorly three distinct scripts were developed. The Cuneiform, script, which was invented in Sumer, Mesopotamia c. 3500 BCE, hieroglyphics sometime prior to the Early Dynastic Period in Egypt (c. 3150-2613 BCE), and Sanskrit in India during the Vedic Period (c. 1500 to c. 500 BCE). Writing was later adopted by other cultures enabling the development of civilization [2].

The evolution of written scripts has been marked by innovations that have increased the efficiency and ease of writing. For example, the Proto-Sinaitic script is considered an important historical development because it led to the creation of the first alphabets, which greatly simplified writing compared to the complex cuneiform and hieroglyphic scripts that preceded it [3].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 1245-1259, 2025
DOI: 10.55214/25768484.v9i2.4746

Figure 2.
From left to right, cuneiform script, hieroglyphics script, Sanskrit script.

2.4. Evolution of Writing Styles

Having the understanding about the history and origins of language and different writing forms and scripts, this section now dwells on the transitional phase when people realized that the style of writing can be studied to understand the usage of different vocabulary by different writers and also how each different style can be used as a signature to further analyse their writing style of different authors and writers.

Taking into account the influence of the Renaissance, an era of European cultural renewal marked by a return to interest in classical education and morals. Humanism, which emphasized the value of individual expression and the study of classical texts, emerged during this time. As a result, literature in common tongues flourished, diverging from the Latin that had dominated academic publications. The Industrial Revolution brought another shift. The demands of a fast-paced, industrialized society were met by a more direct and succinct writing style. The two World Wars in the 20th century had a significant impact on writing. Literature began to reflect a sense of loss and disillusionment, which gave rise to movements like postmodernism and modernism. Possibly the most revolutionary has been the digital revolution. Writing has been made more accessible by the internet, which now lets anyone with internet access share their ideas with a worldwide audience encouraged a more direct and informal writing style [4].

2.5. Study of Stylometry

The ground-breaking stylometric research [5] was predicated on word-length distribution histograms from different writers. The histograms for various languages and authors (Dickens vs. Thackeray) using the same language differed significantly between these papers. "Word spectrums," or "characteristic curves," are a visual depiction of a word arrangement based on length and relative frequency of occurrence, as suggested by Mendenhall, who also suggested using this method to analyze a composition. These manually calculated curves might then be used to compare author writing style models and possibly even to distinguish between different authors' writings. The study in Holmes [6] suggests another reliable stylistic marker that utilizes the advantage of function words and syntactic feature distributions as they are not conscious creations of the author and hence tends to preserve the uniqueness of an author. The same goes for equivalent style markers: function words [7] n-grams of syntactic labels from partial parsing [8] frequencies of rewrite rules [9] n-grams of parts-of-speech [10] and functional lexical features [11]. The table below discusses some of the pioneering work that lay the foundations of research in the field of stylometry and authorship attribution.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 1245-1259, 2025
DOI: 10.55214/25768484.v9i2.4746

Table 1.
Contributions of key works in stylometry and authorship attribution.

Year	Title	Summary	Findings	Future Work	Reference
1984	Applied Bayesian and Classical Inference: The Case of the Federalist Papers	Explores the use of Bayesian and classical inference for authorship disputes of the Federalist Papers, focusing on word frequency analysis.	Bayesian methods remain a reference point for authorship attribution.	Extending Bayesian inference to more complex cases and historical documents, improving computational techniques, exploring alternative statistical models.	Mosteller and Wallace [12].
1991	*Re-counting Plato: A Computer Analysis of Plato's Style*	Uses computational stylometry to analyze the chronology of Plato's works by counting letter frequencies.	Methodological flaws noted in letter-counting and issues in genre distinction.	Refining stylometric methods, addressing limitations in genre differentiation, cross-author comparisons, and using a wider range of linguistic features.	Corlett [13]
1992	*Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information*	Discusses the challenges of using statistical methods for authorship attribution and the need for interpretative frameworks.	Introduced Delta, a key statistical method for stylometric analysis.	Refining statistical models, expanding datasets, integrating statistical analysis with literary scholarship, and encouraging interdisciplinary collaboration between literary critics and statisticians.	Burrows [14]
1994	*Authorship Attribution*	Outlines various methods used to quantify literary style, focusing on lexical units like word length and function words for authorship attribution.	Multivariate statistical methods offer a comprehensive understanding of stylistic differences.	Future work involves refining traditional models, improving multivariate methods, and incorporating more advanced probabilistic models for better authorship attribution accuracy.	Holmes [6]
1998	*Measuring Linguistic Complexity: The Morphological Tier*	Proposes an information-theoretic method to measure morphological complexity by altering morphological information and testing across language corpora.	Languages like Finnish and Russian exhibit higher morphological complexity compared to English and Maori.	Refining the method for syntactic complexity and exploring its application to sub-languages. Further testing across diverse languages is also suggested to confirm findings.	Juola [15].
2002	*Automatically Categorizing Written Texts by Author Gender*	Uses machine learning to determine the gender of an author with 80% accuracy and differentiate fiction from non-fiction with 98%	Identified significant differences in male and female writing styles, particularly in function word use.	Expanding the model to include other demographic features, improving algorithms, and applying the technique to other languages.	Koppel, et al. [16]

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 1245-1259, 2025
DOI: 10.55214/25768484.v9i2.4746

		accuracy, based on lexical and syntactic features.			
2007	*Authorship Attribution*	Surveys traditional and modern methods of authorship attribution, discussing the application of feature sets like lexical and syntactic features to infer authorship.	Identifies challenges posed by small datasets, multilingual texts, and different writing styles.	Refining algorithms to handle short/noisy texts, exploring deeper syntactic/semantic features, and expanding authorship attribution to domains like music, art, and programming.	Juola [17]
2009	*A Survey of Modern Authorship Attribution Methods*	Reviews computational methods of authorship attribution, from traditional stylometry to modern machine learning approaches, emphasizing the evolution of text representation features.	Outlines the benefits of objective evaluation using benchmark datasets and the growing application in real-world domains.	Exploring more advanced stylometric features, handling short/noisy texts, improving models for imbalanced corpora, and integrating machine learning algorithms like deep learning.	Stamatatos [18]

Table 1. Contributions of Key works in Stylometry and Authorship Attribution.

2.5.1. Study of Stylometry with respect to Author Identification in World Languages

Typically, these experiments were done in English language considering different test beds like emails [19, 20] newspaper articles, forum messages from newsgroup discussion threads [21-23] poems [24-26] books [27-31] scientific journals [32, 33] chats [34-37] etc. However, gradually, this study was noted to be done across other world languages also like Arabic [21] Nigerian [38] Brazilian [39] Hebrew [40] Russian [41] Romanian [42] Chinese [43] Persian [44, 45].

Stylometric analysis was used in Abbasi and Chen [19] and Burrows [20] to analyze English-language emails that were extracted from publicly accessible corpuses and discussion boards. Ultimately, one hundred authors were selected from each of the aforementioned datasets. The idea was to find people based on their style of writing. The enlarged feature set's accuracy was 94%. The writers were identified and categorized using SVM, and the similarity detection process was further enhanced by PCA and KL Transformers. The information was collected from various sources, including newsgroup discussions, emails, and online forums [21-23]. The feature set considered in this study included lexical, syntactic, structural, and content-specific features. The writers primarily focused on Arabic and English languages, selecting C4.5, SVM, and decision-tree based methods. The accuracy scores differed between the datasets, with the C4.5 approach scoring an average of 87.8% and SVM 93% for the English dataset. In contrast, the scores were 67.5% for C4.5 and 91.9% for SVM for the Arabic dataset.

Another study Gamon [46] involved extracting sentences no longer than 50 words from books. The authors created a feature set consisting of seven different types and employed linear SVM for classification. They divided their experiments into two types: one considering standalone features and

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 1245-1259, 2025
DOI: 10.55214/25768484.v9i2.4746

the other considering an ensemble. The former achieved an accuracy of 54%, while the latter technique yielded an accuracy of 97.57%.

For author identification in literary short texts, as detailed in [8] an SVM classifier was employed, focusing on works by authors Anne and Charlotte Brontë. The authors introduced bigrams of syntactic labels as a novel classification feature, which outperformed prior approaches. They noted that accuracy in these tasks can be improved by using brief texts, no longer than 200 words, and considering a variety of features.

In a study using newspaper articles as a dataset [47] a set of 22 style markers was created to train a classification model, achieving an accuracy of 81%. However, increasing the number of style markers to 72 did not result in a proportional increase in accuracy. The authors concluded that extracting deep-level style markers requires texts of no more than 1000 words, and their technique is best suited for understanding the linguistic style of a single author rather than multiple authors.

Similar attempts were made for Brazilian [39] and Nigerian languages [38] focusing on authorship attribution in different contexts.

Avram and Oltean [42] focused on authorship attribution in the Romanian language, using lexical, syntactic, and semantic features, as well as machine learning algorithms like Naive Bayes and Decision Trees, with high accuracy.

A study Automatic Authorship Attribution [40] identified authors through modern Hebrew passages, using SVM and Markov chains. While the blog corpus achieved 75% accuracy, the literary corpus using SVM displayed a higher accuracy of 98%.

Other works in Persian [44, 45] Russian [41] and Chinese [43] languages also attempted authorship attribution using various computational methods and achieved varying levels of accuracy.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 1245-1259, 2025
DOI: 10.55214/25768484.v9i2.4746

Table 2.
Author attribution in some world languages.

Test Beds	Language (s)	Study/Reference	Dataset	Techniques	Feature Set	Accuracy/Results
Emails	English	Mosteller and Wallace [12] and Corlett [13]	Publicly accessible corpuses, discussion boards	SVM, PCA, KL Transformers	Lexical, syntactic, structural, content-specific	94%
Newspaper Articles	English	Automatic Authorship Attribution [40]	Newspaper articles	SVM	22 style markers	81%
Forum Messages	English	Burrows [14]; Juola [15] And Koppel, et al. [16]	Newsgroup discussion threads	SVM, Decision Trees, C4.5	Lexical, syntactic, structural, content-specific	C4.5: 87.8% (English), 67.5% (Arabic); SVM: 93% (English), 91.9% (Arabic)
Poems	English	Juola [17]; Schmandt-Besserat and Erard [1] and [19]				
Books	English	[20]; Abbasi and Chen [21]; Argamon, et al. [22]; Litvinova, et al. [23] and Ahmed, et al. [24]	Extracted sentences (max 50 words)	Linear SVM	Seven feature types	Standalone: 54%, Ensemble: 97.57%
Scientific Journals	English	Guzmán-Cabrera [25] and Gallagher and Li [26]	-	-	-	-
Chats	English	Zhao and Zobel [27]; Mehri, et al. [28]; Somers and Tweedie [29] and Koppel, et al. [30]	-	-	-	-
Literary Short Texts	English	Hirst and Feiguina [8]	Works by Anne & Charlotte Brontë	SVM	Bigrams of syntactic labels	97%
Newspaper Articles	English	Automatic Authorship Attribution [40]	Texts (no more than 1000 words)	Classification model	72 style markers	81%
Emails, Forum Messages	Arabic	Burrows [14]; Juola [15] And Koppel, et al. [16]	Newsgroup discussions, online forums	SVM, C4.5	Lexical, syntactic, structural, content-specific	SVM: 91.9%, C4.5: 67.5%
Newspaper Articles	Romanian	Cristani, et al. [35]	Lexical, syntactic, semantic features	Naive Bayes, Decision Trees	-	High accuracy
Literary Passages, Blogs	Hebrew	Lissoni and Montobbio [33]	Modern Hebrew	SVM, Markov Chains	-	75% (blogs), 98% (literary corpus)
Literary Passages	Brazilian, Nigerian	Binongo [31] and Brand, et al. [32]	-	-	-	-
Literary Passages	Russian, Persian, Chinese	Inches, et al. [34]; Layton, et al. [36]; Misra, et al. [37]	-	Various computational methods	-	Varying accuracy

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 1245-1259, 2025
DOI: 10.55214/25768484.v9i2.4746

		and Ayogu and Olutayo [38]				

2.5.2. Study of Stylometry in Indian Languages

India is a land of diverse linguistic groups that play a crucial role in preserving the country's culture and traditions. With reference to the Language Atlas of India, a total of 121 languages have been identified out of which 22 were recognized as scheduled and 99 were categorized under non-scheduled languages [48]. Linguistically, India can be categorized into six major language regions: Central, East, North East, South, Central Himalayan, and Peninsular, each characterized by a rich variety of languages. The languages spoken in India are generally classified into several families, including Indo-European (with a notable branch being Indo-Iranian), Dravidian, Austroasiatic (specifically Munda), and Sino-Tibetan (particularly Tibeto-Burman). The Indo-Aryan language group is part of the larger Indo-European family. In the Indian subcontinent, each of the four major language families—Indo-Aryan, Dravidian, Austro-Asiatic, and Tibeto-Burman—is associated with one of the four main ethnic groups: Negrito, Australoid, Mongoloid, and Caucasoid. These language families are seen as the surviving representatives of a historical tradition of language families that originated in South Asia [49].

One of the very initial attempts to study the statistical components behind the morphology of one of the root Indian Language, Sanskrit was done by Gussner [50] in the year 1976. During that time, the primary languages that saw its dawn were Sanskrit, Persian, Hindi, Urdu and English. The instructional material in educational institutions were mainly in Sanskrit [51]. However post-independence there was seen a minor language shift of the Sanskrit language [52].

2.5.3. Study of Stylometry of Indian Languages Post Independence and its Evolution

In the years following independence, Indian linguistic studies focused more on descriptive and structural linguistics, with relatively little attention given to stylometry. The lack of computational resources and digitized corpora for Indian languages contributed to this. However, foundational work was laid during this period by scholars interested in the mathematical aspects of language, particularly in metrics like frequency analysis of phonemes and morphemes in Sanskrit and other classical languages. The advent of personal computers and increased accessibility to digital resources in the 1990s catalysed the evolution of stylometric studies in Indian languages. Key developments in terms of corpus development, algorithm innovations, and stylometric research of Indian languages were also seen.

2.6. Advent of NLP And Its Influence on the Study of Stylometry

The earlier works of NLP were basically clattered and although the contributions like Chomsky's grammar [53] which was later modified by Hockett and Schooneveld [54] were noteworthy Johri, et al. [55]. Also after this, in the year 1969, tokens were discovered in a legendary work by Schank and Tesler [56]. However, due to lack of computational resources and required awareness and understanding in this field, there was no progressive advancements in this field. But with the advent of Natural Language Processing (NLP) the field of stylometry, was revolutionized with more sophisticated and efficient tools to analyze text and study its properties. Stylometry has traditionally been concerned with the statistical analysis of literary style, focusing on elements such as word usage, sentence structure, and rhythm.

NLP research prior to 1970 laid the foundation for the field but was relatively limited in scope due to the computational constraints and the emerging understanding of language processing by machines. Early work in this period primarily focused on syntactic analysis and simple rule-based systems rather than the more sophisticated, statistical, and machine learning-based approaches that came later. Some of them included Turing's Work on Machine Intelligence [57] Machine Translation [58] ELIZA [59] Augmented Transition Networks [60] Latent Semantic Analysis [61] Context-Free Grammars and Parsing Algorithms [62] SHRDLU [63] Definitional Dictionaries and Lexical Resources [64] and

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 1245-1259, 2025
DOI: 10.55214/25768484.v9i2.4746

Early Semantic Networks [65]. Research during this period set the stage for the more sophisticated statistical, machine learning, and neural network approaches that emerged in the 1970s and beyond.

Post 1970, the society saw some major evolution took place from traditional rule-based study to complex statistical and machine learning models which led to more and more sophisticated research to be carried out. At this time Hidden Markov Models (HMMs) [66] became popular for tasks such as part-of-speech tagging and speech recognition, representing language as a sequence of states with probabilistic transitions. Also, companies like IBM and Bell Labs began working on large-scale speech recognition systems using HMMs and statistical methods. The availability of large corpora, such as the Brown Corpus [67] and later Penn Treebank [68] acted as catalysts by enabling researchers to develop and evaluate statistical models on real-world data. This in turn provided the foundation for training machine learning models. Later the years, in the early 1990s, IBM's researchers, including Peter Brown, introduced the first practical statistical machine translation systems [69]. They treated translation as a probabilistic process, focusing on sentence alignment and word-based translation models. These models replaced earlier rule-based approaches, and SMT became the dominant paradigm for machine translation for over two decades. In the same year itself, supervised machine learning began to dominate NLP tasks. Researchers used algorithms such as decision trees, maximum entropy models, and support vector machines (SVMs) for tasks like named entity recognition (NER), text classification, and parsing [70]. Another paradigm shift came with the advent of deep learning architectures. Deep learning transformed NLP in the early 2010s. Traditional machine learning approaches were based on manually engineered features, but neural networks learned these features automatically from data. The development of Word2Vec by Google in 2013 [71] introduced word embeddings, where words are represented as dense vectors in a continuous vector space. This revolutionized how NLP models handled semantic information by capturing word similarity based on context. Although the concept of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were stated back in the year 1990s by Elman [72] and but they became popular during this phase for processing sequential data like text and were used for tasks like machine translation, text generation, and sentiment analysis.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 1245-1259, 2025
DOI: 10.55214/25768484.v9i2.4746

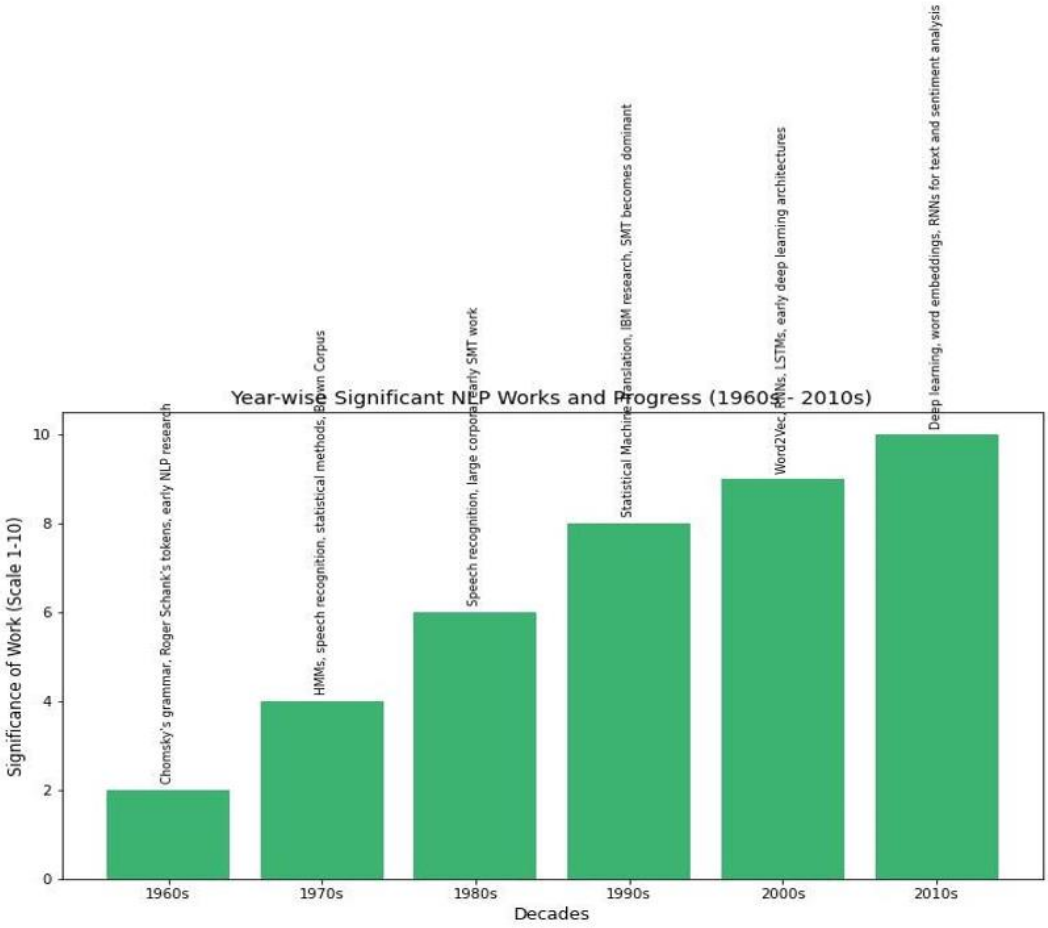

Figure 3.
Year-wise illustration of significant works in NLP.

2.6.1. Study on Authorship Attribution wrt Indian languages (Low Resource)

A model in Lissoni and Montobbio [33] used a few lexical n-grams as the differentiating element, along with 3,000 passages from three Bengali authors, to create an authorship classification system. The study explored feature selection, ranking, and analysis, as well as the relationship between test accuracy and training data volume. Supervised and unsupervised author classification approaches in Hindi literature were explored using lexical unigrams, bigrams, trigrams, and the concept of MDA to construct a feature vector. The study emphasized that the granularity of a dataset significantly impacts classification scores, with a greater number of diverse feature vectors leading to higher accuracy. To categorize authors of Kannada texts, Cristani, et al. [35] employed a machine learning algorithm based on the author's profile, achieving an accuracy score of 88%. Similarly, in Oliveira Jr, et al. [39] authorship attribution in the Telugu language was studied using word n-grams, character n-grams, syntactic features, and machine learning algorithms, accurately identifying authors of Telugu texts.

3. Analysis and Discussion

The paper covers a wide spectrum of research conducted in the field of author attribution across multiple languages, utilizing various techniques like SVM, decision trees, and deep learning models. The results showcase different accuracies across datasets, with some approaches achieving over 90% accuracy. The study highlights how traditional stylometric methods (based on word frequency, function

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 1245-1259, 2025
DOI: 10.55214/25768484.v9i2.4746

words, etc.) are still relevant but are often outperformed by machine learning techniques. Feature sets like lexical, syntactic, structural, and content-specific markers proved useful in different contexts and languages, though performance varied significantly based on language complexity, corpus size, and the writing styles involved. The transition from traditional rule-based approaches to statistical machine learning and deep learning methods has enabled more robust and scalable author attribution techniques. The advent of deep learning, in particular, has facilitated higher accuracy in low-resource languages by automating the feature extraction process. However, the analysis reveals challenges in maintaining high accuracy across multilingual datasets, especially for short or noisy texts, where classification accuracy tends to drop.

4. Conclusion and Future Work

Authorship attribution has advanced considerably with the integration of machine learning and computational models. Despite challenges posed by multilingual and low-resource languages, the results indicate that deep learning models, in particular, hold the potential for improving accuracy across diverse corpora. The paper emphasizes the dynamic nature of language and how the evolving nature of writing styles influences author attribution. It concludes that, while traditional methods face difficulties, modern machine learning and deep learning techniques can effectively adapt and improve author identification across different languages and genres.

5. Limitations

This paper highlights significant advancements in authorship attribution using machine learning, but it also presents several limitations. First, there is a limited focus on multilingual and cross-lingual authorship attribution, despite the growing importance of analyzing mixed-language texts. Second, while low-resource languages like Kannada and Telugu are explored, the study relies on relatively small datasets, limiting the generalizability of its findings. Larger datasets could provide better insights into these languages' unique features. Finally, although the paper provides in-depth experimental analysis, there is a lack of discussion on real-world applications, such as in forensic linguistics or legal contexts, which would demonstrate the practical utility of the authorship attribution methods. Addressing these limitations could strengthen future research in the field.

6. Ethics Statement

In conducting this review on authorship attribution across various languages, the research adhered to ethical guidelines, ensuring that all data and methodologies used respected the privacy and intellectual property of the original authors. Publicly available datasets were utilized to avoid any violation of personal or proprietary information. The study focused solely on text analysis, and no human participants were involved, mitigating concerns around consent or privacy violations. The research also acknowledges the limitations of machine learning models, particularly in their biases toward languages with more digital resources. To address this, future work aims to develop more inclusive methods that consider the linguistic diversity and complexity of underrepresented languages, thereby reducing inequity in computational research. The authors are committed to advancing ethical, fair, and transparent methods in the field of authorship attribution.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 1245-1259, 2025
DOI: 10.55214/25768484.v9i2.4746

Acknowledgements:
The authors would like to extend their heartfelt gratitude to the NLP and Language Technology Lab, Department of IT, Gauhati University, for their invaluable support and guidance throughout the process of writing this article by putting forward their valuable insights and data for this review. We are also grateful to our colleagues for their critical feedback and support throughout the development of this paper. Their expertise and suggestions helped shape the analysis and direction of the study.

References
[1] D. Schmandt-Besserat and M. Erard, "Origins and forms of writing," *Handb. Reserch Write Hist. Social Scholar Individ Text*, vol. 30, pp. 7–26, 2009. https://doi.org/10.4324/9781410616470-10
[2] J. J. Mark and J. Van Der Crabbe, "Script," Retrieved: https://www.worldhistory.org/script. [Accessed 2023.
[3] M. Gross, "The evolution of writing," *Current Biology*, vol. 22, no. 23, pp. R981-R984, 2012. https://doi.org/10.1016/j.cub.2012.11.032
[4] Writing Styles, "Writing styles: A historical evolution," Retrieved: https://essaysleader.com/the-evolution-of-writing-styles-through-history/. [Accessed April 30], 2024.
[5] C. M. Thomas, "The characteristic curves of composition. History of Information," Retrieved: https://www.historyofinformation.com/detail.php?id=4120. [Accessed 1887.
[6] D. I. Holmes, "Authorship attribution," *Computers and the Humanities*, vol. 28, pp. 87-106, 1994. https://doi.org/10.1007/BF01830689
[7] A. M. Garcia and J. C. Martin, "Function words in authorship attribution studies," *Literary and Linguistic Computing*, vol. 22, no. 1, pp. 49-66, 2006. https://doi.org/10.1093/llc/fql048
[8] G. Hirst and O. g. Feiguina, "Bigrams of syntactic labels for authorship discrimination of short texts," *Literary and Linguistic Computing*, vol. 22, no. 4, pp. 405-417, 2007. https://doi.org/10.1093/llc/fqm023
[9] H. Baayen, H. Van Halteren, and F. Tweedie, "Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution," *Literary and Linguistic Computing*, vol. 11, no. 3, pp. 121-132, 1996. https://doi.org/10.1093/llc/11.3.121
[10] J. Diederich, J. Kindermann, E. Leopold, and G. Paass, "Authorship attribution with support vector machines," *Applied Intelligence*, vol. 19, pp. 109-123, 2003. https://doi.org/10.1023/A:1023824908771
[11] S. Argamon, C. Whitelaw, P. Chase, S. R. Hota, N. Garg, and S. Levitan, "Stylistic text classification using functional lexical features," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 6, pp. 802-822, 2007.
[12] F. Mosteller and D. L. Wallace, "Applied Bayesian and classical inference: The case of the Federalist papers. Springer Science & Business Media," Retrieved: https://books.google.com/books?hl=en&lr=&id=LJXaBwAAQBAJ&oi=fnd&pg=PA6&dq=Mosteller,+F.,+%26+Wallace,+D.+L.+(2000).+Applied+Bayesian+and+Classical+Inference:+The+Case+of+the+Federalist+Papers.+Springer-erlag.&ots=1LzcMyXdYY&sig=9lO0XKcc7NBK07Qt7yxkGx8hrNU. [Accessed Oct. 22, 2024], 2012.
[13] T. Corlett, "Re-counting Plato: A computer analysis of Plato's style." JSTOR," Retrieved: https://www.jstor.org/stable/2348241. [Accessed Oct. 22, 2024], 1991.
[14] J. F. Burrows, "Not unles you ask nicely: The interpretative nexus between analysis and information," *Literary and Linguistic Computing*, vol. 7, no. 2, pp. 91-109, 1992.
[15] P. Juola, "Measuring linguistic complexity: The morphological tier," *Journal of Quantitative Linguistics*, vol. 5, no. 3, pp. 206-213, 1998.
[16] M. Koppel, S. Argamon, and A. R. Shimoni, "Automatically categorizing written texts by author gender," *Literary and linguistic computing*, vol. 17, no. 4, pp. 401-412, 2002.
[17] P. Juola, "Authorship attribution," *Foundations and Trends® in Information Retrieval*, vol. 1, no. 3, pp. 233-334, 2008. https://doi.org/10.1561/1500000005
[18] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for information Science and Technology*, vol. 60, no. 3, pp. 538-556, 2009. https://doi.org/10.1002/asi.21001
[19] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 2, pp. 1-29, 2008. https://doi.org/10.1145/1344411.1344413
[20] J. Burrows, "'Delta': A measure of stylistic difference and a guide to likely authorship," *Literary and Linguistic Computing*, vol. 17, no. 3, pp. 267-287, 2002. https://doi.org/10.1093/llc/17.3.267

[21] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group web forum messages," *IEEE Intelligent Systems*, vol. 20, no. 5, pp. 67-75, 2005.

[22] S. Argamon, M. Šarić, and S. S. Stein, "Style mining of electronic messages for multiple authorship discrimination: First results," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003, pp. 475-480.

[23] T. Litvinova, O. Litvinova, and P. Panicheva, "Authorship attribution of Russian forum posts with different types of n-gram features," in *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, 2019, pp. 9-14.

[24] A.-F. Ahmed, R. Mohamed, and B. Mostafa, "Machine learning for authorship attribution in Arabic poetry," *International Journal Future Comput. Commun*, vol. 6, no. 2, pp. 42-46, 2017. https://doi.org/10.18178/ijfcc.2017.6.2.486

[25] R. Guzmán-Cabrera, "Authorship attribution of Spanish poems using n-grams and the web as corpus," *Journal of Intelligent & Fuzzy Systems*, vol. 39, no. 2, pp. 2391-2396, 2020. https://doi.org/10.3233/JIFS-179899

[26] C. Gallagher and Y. Li, "Text categorization for authorship attribution in english poetry," in *Intelligent Computing: Proceedings of the 2018 Computing Conference, Springer International Publishing*, 2019, vol. 1, pp. 249-261.

[27] Y. Zhao and J. Zobel, "Searching with style: Authorship attribution in classic literature," in *ACM International Conference Proceeding Series*, 2007, vol. 244, pp. 59-68.

[28] A. Mehri, A. H. Darooneh, and A. Shariati, "The complex networks approach for authorship attribution of books," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 7, pp. 2429-2437, 2012. https://doi.org/10.1016/j.physa.2011.12.011

[29] H. Somers and F. Tweedie, "Authorship attribution and pastiche," *Computers and the Humanities*, vol. 37, pp. 407-429, 2003.

[30] M. Koppel, J. Schler, and S. Argamon, "Computational methods in authorship attribution," *Journal of the American Society for information Science and Technology*, vol. 60, no. 1, pp. 9-26, 2009. https://doi.org/10.1002/asi.20961

[31] J. N. G. Binongo, "Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution," *Chance*, vol. 16, no. 2, pp. 9-17, 2003. https://doi.org/10.1080/09332480.2003.10554843

[32] A. Brand, L. Allen, M. Altman, M. Hlava, and J. Scott, "Beyond authorship: Attribution, contribution, collaboration, and credit," *Learned Publishing*, vol. 28, no. 2, 2015. https://doi.org/10.1087/20150211

[33] F. Lissoni and F. Montobbio, "Guest authors or ghost inventors? Inventorship and authorship attribution in academic science," *Evaluation Review*, vol. 39, no. 1, pp. 19-45, 2015. https://doi.org/10.1177/0193841X13517234

[34] G. Inches, M. Harvey, and F. Crestani, "Finding participants in a chat: Authorship attribution for conversational documents," presented at the 2013 International Conference on Social Computing. IEEE, 2013.

[35] M. Cristani, G. Roffo, C. Segalin, L. Bazzani, A. Vinciarelli, and V. Murino, "Conversationally-inspired stylometric features for authorship attribution in instant messaging," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 1121-1124.

[36] R. Layton, P. Watters, and R. Dazeley, "Authorship attribution for twitter in 140 characters or less," presented at the 2010 Second Cybercrime and Trustworthy Computing Workshop, Ballarat, Australia: IEEE. https://doi.org/10.1109/CTC.2010.17, 2010.

[37] K. Misra, H. Devarapalli, T. R. Ringenberg, and J. T. Rayz, "Authorship analysis of online predatory conversations using character level convolution neural networks," presented at the 2019 IEEE International Conference on Systems, Man and Cybernetics, IEEE, 2019.

[38] I. I. Ayogu and V. A. Olutayo, "Authorship attribution using rough sets based feature selection techniques," *International Journal of Computer Applications*, vol. 152, no. 6, pp. 38-46, 2016. https://doi.org/10.5120/ijca2016911889

[39] W. Oliveira Jr, E. Justino, and L. S. Oliveira, "Comparing compression models for authorship attribution," *Forensic Science International*, vol. 228, no. 1-3, pp. 100-104, 2013. https://doi.org/10.1016/j.forsciint.2013.02.025

[40] Automatic Authorship Attribution, "Automatic authorship attribution in the hebrew Bible and other literary texts | Digital Humanities," Retrieved: https://live-digital-humanities-berkeley.pantheon.berkeley.edu/projects/automatic-authorship-attribution-hebrew-bible-and-other-literary-texts. [Accessed May 02, 2024], 2024.

[41] E. Reisi and H. Mahboob Farimani, "Authorship attribution in historical and literary texts by a deep learning classifier," *Journal of Applied Intelligent Systems and Information Sciences*, vol. 1, no. 2, pp. 118-127, 2020. https://doi.org/10.22034/jaisis.2021.269735.1018

[42] S. M. Avram and M. Oltean, "A comparison of several AI techniques for authorship attribution on Romanian texts," *Mathematics*, vol. 10, no. 23, p. 4589, 2020. https://doi.org/10.3390/math10234589

[43] H. Wang and A. Riddell, "CCTAA: A reproducible corpus for Chinese authorship attribution research," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds., Marseille, France: European Language Resources Association*, 2022, pp. 5889–5893.

[44] R. Ramezani, "A language-independent authorship attribution approach for author identification of text documents," *Expert Systems with Applications*, vol. 180, p. 115139, 2021. https://doi.org/10.1016/j.eswa.2021.115139

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 1245-1259, 2025
DOI: 10.55214/25768484.v9i2.4746

[45] R. M. Dabagh, "Authorship attribution and statistical text analysis," *Metodoloski Zvezki*, vol. 4, no. 2, p. 149, 2007. https://doi.org/10.51936/uvjx7198

[46] M. Gamon, "Linguistic correlates of style: authorship classification with deep linguistic analysis features," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland: COLING*, 2004, pp. 611–617.

[47] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Computer-based authorship attribution without lexical measures," *Computers and the Humanities*, vol. 35, no. 2, pp. 193-214, 2001. https://doi.org/10.1023/A:1002681919510

[48] C. o. I.-L. A.-. India, "Census of India 2011," Retrieved: https://censusindia.gov.in/nada/index.php/catalog/42561. [Accessed 2011.

[49] S. Kulkarni-Joshi, "Linguistic history and language diversity in India: Views and counterviews," *Journal of Biosciences*, vol. 44, no. 3, p. 62, 2019. https://doi.org/10.1007/s12038-019-9879-1

[50] R. E. Gussner, "A stylometric study of the authorship of seventeen sanskrit hymns attributed to Śaṅkara," *Journal of the American Oriental Society*, vol. 96, no. 2, pp. 259-267, 1976. https://doi.org/10.2307/599828

[51] B. B. Kachru, "Language policy in South Asia," *Annual Review of Applied Linguistics*, vol. 2, pp. 60-85, 1981. https://doi.org/10.1017/S0267190500000258

[52] W. Fase, K. Jaspaert, and S. Kroon, *Maintenance and loss of minority languages*. Amsterdam: John Benjamins Publishing, 1992.

[53] R. B. Lees, "Review of syntactic structures," *Language*, vol. 33, no. 3, pp. 375–408, 1957. https://doi.org/10.2307/411160

[54] C. F. Hockett and C. H. Schooneveld, *Language, mathematics, and linguistics*. The Hague: Mouton, 1967.

[55] P. Johri, S. K. Khatri, A. T. Al-Taani, M. Sabharwal, S. Suvanov, and A. Kumar, "Natural language processing: History, evolution, application, and future work," in *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020*, 2021: Springer, pp. 365-375.

[56] R. C. Schank and L. Tesler, "A conceptual dependency parser for natural language," in *International Conference on Computational Linguistics COLING 1969: Preprint No. 2*, 1969.

[57] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950. https://doi.org/10.1093/mind/LIX.236.433

[58] W. J. Hutchins, *Machine translation: Past, present and future. in Ellis Horwood series in computers and their applications*. Chichester: Ellis Horwood, 1986.

[59] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36-45, 1966. https://doi.org/10.1145/365153.365168

[60] W. A. Woods, "Transition network grammars for natural language analysis," *Communications of the ACM*, vol. 13, no. 10, pp. 591-606, 1970. https://doi.org/10.1145/355598.362773

[61] D. K. Harman, *The second text retrieval conference (TREC-2)*. Gaithersburg, MD: US Department of Commerce, National Institute of Standards and Technology, 1994.

[62] T. Kasami, "An efficient recognition and syntax-analysis algorithm for context-free languages," Coordinated Science Laboratory Report No. R-257, 1966.

[63] J. S. Brown, *Terry Winograd. Understanding natural language*. New York: Academic Press, 1973.

[64] S. Dierk, "The SMART retrieval system: Experiments in automatic document processing—Gerard Salton, Ed.(Englewood Cliffs, NJ: Prentice-Hall, 1971, 556 pp., $15.00)," *IEEE Transactions on Professional Communication*, vol. PC-15, no. 1, pp. 17-17, 1972. https://doi.org/10.1109/TPC.1972.6591971

[65] M. R. Quillian, *Semantic memory (AFCRL-66-189)*. Bedford, MA: Air Force Cambridge Research Laboratories, Office of Aerospace Research, United States Air Force, 1966.

[66] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989. https://doi.org/10.1109/5.18626

[67] Michigan Corpus of Academic Spoken English (MICASE), "CoRD," Retrieved: https://varieng.helsinki.fi/CoRD/corpora/MICASE/. [Accessed 2001.

[68] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of English: The Penn Treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313-330, 1993. https://doi.org/10.21236/ada273556

[69] P. F. Brown *et al.*, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no. 2, pp. 79-85, 1990.

[70] C. Manning and H. Schutze, *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press, 1999.

[71] T. Mikolov, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, vol. 3781, 2013. https://doi.org/10.48550/arXiv.1301.3781

[72] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179-211, 1990. https://doi.org/10.1207/s15516709cog1402_1

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 1245-1259, 2025
DOI: 10.55214/25768484.v9i2.4746