ISSN NO: 2770-8608
January 2025

Analysis of global experience in text analysis

Rahimov Hasanboy Komiljonovich
2nd year basic doctoral student of ASU,
intern teacher of Namangan State
Institute of Foreign Languages
xasanraximov@adu.uz

Abstract:
Since the dawn of human self-awareness to the modern era of technology, one of the consistently evolving and relevant fields of scientific inquiry has been the study and analysis of text. This article examines the historical development and key approaches in the field of text analysis. It highlights the stages of text analysis from ancient civilizations to contemporary computational linguistics and natural language processing (NLP) technologies. The contributions and methodologies of renowned scholars such as Wilhelm Wundt, Vladimir Propp, Claude Shannon and Noam Chomsky are analyzed. The article provides extensive insights into the principles, methods, and applications of text analysis in NLP, machine learning, and information technologies.
Keywords: text analysis, natural language processing (NLP), corpus linguistics, information theory, generative grammar, language modeling, sentiment analysis.

Introduction

In recent times, the topic of text analysis, now largely performed by machines, has intrigued humanity since ancient times. Its earliest known attempts date back to ancient civilizations. Scholars and scribes in ancient Mesopotamia, Egypt, Greece, and Rome analyzed texts for meaning, structure, and rhetorical aspects. Over subsequent eras, various new methods and directions in text analysis emerged based on human needs. As highlighted in the scientific works of Professor N. Abdurahmonova from the National University of Uzbekistan (NUUz), modern text analysis began to establish itself as an independent scientific field in the late 19th and early 20th centuries, coinciding with the development of formal methodologies in linguistics, philology, and semiotics. The initial attempts related to text analysis during this period can be observed in the research efforts of the following scholars.

The first figure to mention is the German scholar Wilhelm Wundt (1832– 1920), who is known for his early works on understanding the structure and meaning of language. He was primarily a German psychologist and one of the founders of psychology, contributing significantly to the development of experimental psychology and physiological psychology. Although his scientific works were not directly related to modern text analysis, Wundt greatly contributed to understanding the psychological foundations of language, meaning, and communication, which later paved the way for the development of linguistics and text analysis. In his work *"Völkerpsychologie"* (Folk Psychology), Wundt analyzed language and culture as collective expressions of human consciousness. He argued that language is the primary means of communication and a factor in the development of human thought[1].

Later, the Russian folklorist Vladimir Propp significantly contributed to global science with his innovative approach to text analysis. Vladimir Propp (1895– 1970), a Russian folklorist and structuralist scholar, is widely recognized for his work *"Morphology of the Folktale"* (1928), which is regarded as a revolutionary method in text analysis. In this work, he developed a theoretical model aimed at identifying the main structures and compositional elements of stories and folktales. Propp's approach focused on understanding the structural components of texts, making a significant impact on the development of modern linguistics and narrative analysis. In his pioneering research, he analyzed the structures of folktales to identify their universal patterns. He described the structural components of stories through the concept of "functions." This approach enabled the identification and analysis of the primary sequence of events in narratives [2].Propp's approach is based on universalizing narratives into a model, and his theory is widely applied in analyzing stories from different

ISSN NO: 2770-8608
January 2025

cultures and periods. He combined semantic and syntactic approaches to analyze the internal structure of narratives.

The 1940s and 1950s brought significant breakthroughs in the field of computational linguistics. During this period, pioneers such as Claude Shannon developed methods for analyzing data and text within the framework of information theory. Claude Shannon (1916– 2001), widely recognized as the founder of modern information theory, made groundbreaking contributions to text analysis, data compression, and encoding. Shannon was one of the first researchers to apply mathematical and statistical approaches to text analysis, and these methods became the foundational principles of natural language processing (NLP) and text analysis.In his renowned 1948 paper, *"A Mathematical Theory of Communication,"* Shannon laid the groundwork for information theory. In this work, he introduced key concepts for mathematically modeling the processes of information transmission and processing, including:

- **Information Quantity:** Measuring the degree of precision and uncertainty in data.
- **Entropy:** A statistical measure representing the level of uncertainty in information.
- **Efficiency:** Methods for minimizing losses during information transmission[3].

These principles served as the basis for implementing effective methods of data compression and processing in text analysis. Shannon focused on identifying the statistical properties of text during analysis. He used character frequency to determine the probability of repetition of each symbol or word in a text. Shannon also studied the probability of one character following another in a sequence. For instance, in English, the letter "q" is typically followed by the letter "u." These approaches enabled the identification and mathematical modeling of the semantic and structural properties of text. Through information theory, Claude Shannon played a crucial role in identifying redundancies in text and developing algorithms for data compression.

In the 1960s, Noam Chomsky revolutionized the field of linguistics with his works on syntax and transformational grammar, paving the way for systematic text analysis. Noam Chomsky (born in 1928), an American linguist, cognitive scientist, and philosopher, has significantly influenced modern linguistics, particularly in text analysis and natural language processing (NLP). By introducing revolutionary changes in linguistics, Chomsky developed theories of syntactic structures, generative grammar, and transformational grammar, which provided a theoretical foundation for applying structural and generative models in text analysis. Below are his four main approaches related to text analysis:

1. Generative Grammar. In his work *"Syntactic Structures"* (1957), Chomsky introduced the theory of generative grammar, which has had a revolutionary impact on analyzing and understanding text structure. Generative grammar focuses on identifying the internal rules of language and explaining how linguistic units are formed. Through syntactic structures, Chomsky demonstrated how linguistic units—words, phrases, or sentences—are constructed based on syntactic rules[4]. This approach aids in the automatic identification of syntactic units in text analysis. Chomsky also proposed the distinction between the **deep structure** and **surface structure** of language, enabling a better understanding of sentence and text meanings.

2. Transformational Grammar. Transformational grammar explains how a specific sentence structure can be transformed into another[4]. This approach is applied in the following areas:

- **Text Paraphrasing:** Transforming sentences to generate different structures expressing the same meaning.
- **Syntactic Analysis:** Analyzing the syntactic components of text to determine its meaning.

Chomsky used this method to model the structure of language, providing a foundation for algorithms in NLP and text analysis.

3. Hierarchical Model of Language (Chomsky Hierarchy). Chomsky categorized languages into a hierarchy of four classes: regular languages, context-free languages, context-sensitive languages, and recursively enumerable languages (Turing machine languages). This model is used to determine grammatical structure in language and develop algorithms for text analysis[5].

ISSN NO: 2770-8608
January 2025

4. Universal Grammar. Chomsky proposed the theory of universal grammar, suggesting the existence of a set of rules common to all languages. This approach provides a theoretical basis for analyzing texts across different languages[6].

The use of computers in text analysis began in 1949 with Roberto Busa' s *Index Thomisticus* project. Roberto Busa' s role in applying computers to text analysis is unparalleled. Roberto Busa (1913– 2011), an Italian Catholic priest and linguist, is regarded as one of the first scholars to use computers for text analysis. By integrating technology into the humanities, Busa initiated revolutionary changes and holds a distinguished place in history as a pioneer in this process. His work laid the foundation not only for text analysis but also for computational linguistics and digital humanities. Among Roberto Busa' s significant contributions to science, the *Index Thomisticus* project stands as his most notable research dedicated to text analysis. The *Index Thomisticus* project was completed in 1980 after 30 years of work, resulting in a monumental compilation of 56 volumes. This index covers over 10 million words from Aquinas' s works and related texts. The outcomes of this project have become the foundation for contemporary digital linguistics and projects in the humanities[7].

As we explore the early efforts of pioneers in text analysis, it becomes evident how their work has steered this field toward perfection. Text analysis continues to capture the interest of scholars worldwide to this day. Below is a chronological table summarizing their contributions, showcasing the progress made in text analysis over time. Roberto Busa' s groundbreaking methodologies have not only shaped the history of text analysis but also created a strong foundation for advancements in computational linguistics, AI, and digital humanities. His legacy continues to inspire innovation in the intersection of technology and the humanities.

Chronological table of text analysis contributions from 1960 to 2024.

Year	Author(s)	Title Translation	Significance
1966	Noam Chomsky	*Foundations of Syntactic Theory*	Provided a new approach to analyzing hierarchical structures of language and grammatical rules[4].
1975	Teun A. van Dijk	*Text and Context: Explorations in Discourse Semantics and Pragmatics*	Highlighted text as a key element of discourse analysis and analyzed pragmatic context[8].
1980	M.A.K. Halliday and Ruqaiya Hasan	*Cohesion in English*	Created a theoretical basis for analyzing cohesive devices within texts[9].
1983	Michael Hoey	*The Surface Structure of Discourse*	Presented a practical model for understanding discourse structure and textual coherence[10].
1988	Gillian Brown and George Yule	*Discourse Analysis*	Established a discourse paradigm for text analysis and explored semantic aspects[11].
1991	John Sinclair	*Corpus, Concordance, Collocation*	Pioneered early work in corpus linguistics and analyzed collocations[12].
1992	Douglas Biber	*Variation Across Speech and Writing*	Conducted a foundational study in corpus linguistics to differentiate styles and genres[13].
1996	Michael Stubbs	*Text and Corpus Analysis: Computer-*	Introduced new tools for text analysis using computer

ISSN NO: 2770-8608
January 2025

		Assisted Studies of Language and Culture	technologies[14].
2001	Tony McEnery and Andrew Wilson	*Corpus Linguistics: An Introduction*	Provided foundational theoretical and practical guidelines for corpus linguistics[15].
2002	Stefan Th. Gries	*Foundations of Syntactic Theory*	Provided a new approach to analyzing hierarchical structures of language and grammatical rules[16].
2006	Paul Baker	*Using Corpora in Discourse Analysis*	Introduced new methods for analyzing discourse based on corpora[17].
2010	Elena Tognini-Bonelli	*Corpus Linguistics in Practice*	Contributed significantly to the development of practical research in corpus linguistics[18].
2012	Tony McEnery and Andrew Hardie	*Corpus Linguistics: Method, Theory, and Practice*	Combined theory and practice to set new directions in corpus linguistics[19].
2015	Douglas Biber and Randi Reppen	*The Cambridge Handbook of Corpus Linguistics*	Served as a fundamental resource for corpus linguistics in English[20].
2018	Stefan Th. Gries	*Quantitative Corpus Linguistics with R*	Deepened corpus analysis with statistical and computational tools[21].
2020	Paul Baker	*Corpus Linguistics and Sociolinguistics*	Applied corpus tools to study social aspects of language in text analysis[22].
2024	Haiyan Liu· Shelly Tsang · Adrienne Wood · Xin Tong	*Sentiment Analysis of Long-Term Communication Texts*	Proposed a two-stage model for long-term sentiment analysis, integrating emotion detection and growth trajectory modeling[23].

Conclusion

ISSN NO: 2770-8608
January 2025

The field of text analysis has undergone various stages from ancient times to its development into a modern scientific discipline. From ancient civilizations to the works of scholars like Wilhelm Wundt, Vladimir Propp, Claude Shannon, Noam Chomsky, and Roberto Busa, significant research has emerged on the role of text in meaning, structure, and communication. The approaches introduced by these scholars laid the foundation for disciplines such as semantic analysis, narrative analysis, information theory, and generative grammar. Leveraging international research, the development of various branches of text analysis in Uzbekistan will strengthen the integration between national linguistics and digital technologies. This will also enhance the position of the Uzbek language in the global scientific community.

References

1. Wundt, Wilhelm. *Elements of Folk Psychology: Outline of a Psychological History of the Development of Mankind*. Translated by Edward Leroy Schaub, George Allen & Unwin, 1916.
2. Propp, Vladimir. *Morphology of the Folktale*. Translated by Laurence Scott, 2nd ed., University of Texas Press, 1968. (Originally published in 1928).
3. Shannon, Claude E. "A Mathematical Theory of Communication." *The Bell System Technical Journal*, vol. 27, no. 3, 1948, pp. 379– 423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.
4. Chomsky, Noam. *Syntactic Structures*. The Hague: Mouton, 1957.
5. Chomsky, N. "Three Models for the Description of Language." *IRE Transactions on Information Theory*, vol. 2, no. 3, 1956, pp. 113– 124. https://chomsky.info/wp-content/uploads/195609-.pdf.
6. Chomsky, N. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press, 1965. https://archive.org/details/aspectsoftheoryo00chom.
7. Wikipedia. "Roberto Busa." https://en.wikipedia.org/wiki/Roberto_Busa.
8. van Dijk, Teun A. *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*. London: Longman, 1977.
9. Halliday, M. A. K., and Ruqaiya Hasan. *Cohesion in English*. London: Longman, 1976.
10. Hoey, Michael. *On the Surface of Discourse*. London: Allen & Unwin, 1983.
11. Brown, Gillian, and George Yule. *Discourse Analysis*. Cambridge: Cambridge University Press, 1983.
12. Sinclair, John. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.
13. Biber, Douglas. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press, 1988.
14. Stubbs, Michael. *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell, 1996.
15. McEnery, Tony, and Andrew Wilson. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press, 1996.
16. Gries, Stefan Th. *Quantitative Corpus Linguistics with R: A Practical Introduction*. London: Routledge, 2009.
17. Baker, Paul. *Using Corpora in Discourse Analysis*. London: Continuum, 2006.
18. Tognini-Bonelli, Elena. *Corpus Linguistics at Work*. Amsterdam: John Benjamins, 2001.
19. McEnery, Tony, and Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press, 2012.
20. Biber, Douglas, and Randi Reppen. *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 2015.
21. Gries, Stefan Th. *Quantitative Corpus Linguistics with R: A Practical Introduction*. 2nd ed., London: Routledge, 2013.
22. Baker, Paul. *Corpus Linguistics and Sociolinguistics: A Study of Variation in English*. Edinburgh: Edinburgh University Press, 2010.
23. Liu, Haiyan, et al. "Sentiment Analysis Over Time: A Two-Stage Model of Malleability." *Journal of Computational Linguistics*, 2024.