ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

Teaching the Principles of Text Compression and Exploring Gender Differences in Eradicating Information Redundancy

[1] Revaz Tabatadze, [2] Maia Chkheidze

[1] [2] The University of Georgia, Tbilisi, Georgia

Corresponding Author Email: [1] revaz.tabatadze@ug.edu.ge, [2] m.chkheidze@ug.edu.ge

***Abstract*— The research on the phenomenon of text compression lies in response to the ever-increasing demands of the modern information society. These demands are intricately tied to the efficient utilization of knowledge and the continuous pursuit of enhancing the methods for processing information.**

Hence, the relevance of this research is derived from the active manifestation of the need to provide students with the skills required for compressing information.

The significance of this research stems from the tangible demonstration of the imperative to equip students with the essential aptitude for information compression. The primary aim of this study is to discern the distinct features pertaining to various principles employed in the compression of information.

The primary analytical approach employed in this research is comparative and content analyses. The research predominantly relies on the inductive method of analysis, wherein the investigation progresses from specific linguistic observations towards discerning systematic relationships among them.

Within the frame of the research, the qualitative survey conducted in 2022-23 at the university of Georgia involved a sample of 50 text constructors (undergraduate and graduate students studying English Philology). The survey encompassed one embedded unit, ensuring gender balance, and the respondents were selected using a randomizing principle. The research material, comprising scientific articles, was provided in English.

Upon analyzing the collected data, certain regularities emerged:

1.Abstracts generated by different individuals exhibit subjectivity.

2.The subjectivity of the abstracts generated by individuals is influenced by the gender of the text constructors.

The research findings have substantiated the significance of redundancy when examining the phenomenon of text compression.

***Index Terms*— original text, compressed version, text compression, informational redundancy, gender differences, principle of least effort, principle of economy.**

I. INTRODUCTION

In contemporary scientific communication, there has been a substantial surge in the volume of information and knowledge. This trend has been further exacerbated by the advent of the internet network, posing a pressing challenge of coping with the rapid growth of information. Keeping pace with new information has become increasingly difficult for specialists.

In modern linguistics, one crucial objective is to uncover the patterns underlying systemic description and generation of texts. It is worth noting that discerning these patterns plays a pivotal role in understanding the mechanism of text compression.

By delving into the specifics of text compression, a deeper comprehension of the text can be attained. An abstract can be deemed well-constructed if the text constructor can effectively condense the primary (original) text while preserving its core content.

Defining the category of understanding (comprehension) proves to be quite challenging. This difficulty stems from the fact that this phenomenon pertains more to unconscious or partially conscious processes rather than formal-logical processes of human activity.

As a result of understanding (comprehension), the text constructor undergoes a sequential transformation in the mental structure of the text. In other words, there is a cognitive process of moving from one textual element to another, leading to the logical reconstruction of the text structure in the mind of the text constructor.

The existence of a primary (original) text is a prerequisite for the existence of a secondary document (abstract). The secondary document, or abstract, is intrinsically linked to the primary text, as it is constructed based on it. The secondary text must possess all the markers of textuality and reflect the principal stages of text generation.

Information compression imposes specific requirements on the text constructor. These requirements encompass semantic alignment of the abstract with the primary text's main content and compactness in relation to the compressed text's volume.

Information compression entails the condensation of signifiers while maintaining the integrity of the signified. The

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

notion of text norm is employed to establish the boundary for compression. It is important to acknowledge that the text norm varies across different texts. Nevertheless, there exists a shared semantic indicator associated with the text norm. Essentially, the communicative function of a linguistic unit must not be compromised. The norm signifies a harmonious alignment of textual elements, encompassing both the form of the text and the intentions of the author or text constructor, in accordance with the expectations of the recipient. This represents the most effective mode of interaction among components to achieve the intended outcome.

II. METHODOLOGY

The reduction of text volume is achieved through language compression methods [1]. Traditionally, text compression is regarded as a simplification of the surface structure of a text. This simplification arises from the principles of speech economy, genre requirements, and characteristics of the medium of information transmission. The surface structure of the text is streamlined by increasing the informativeness of language units and eliminating elements that can be inferred from the text without compromising its informational content.

According to the information theory, both the original text and the compressed version contain the same information about the described object. However, it is worth noting that compression, by reducing the number of linguistic units, significantly impacts the meaningful aspect of the text, rearranging its grammatical and semantic structure. During the compression process, compressed components delegate their functions to the uncompressed counterparts. The functional load of these components differs from their role in the uncompressed, original text [3].

The amount of information to be eliminated depends on the text's norms and style. It is important to emphasize that the mentioned type of information varies across texts. Nevertheless, there is a universal criterion for all texts of this nature: speech units should not lose their communicative significance or function. Otherwise, text compression can lead to communication breakdown, hindering the intended purpose of the text [2].

To compress the original text, the following actions were taken:

1. Separation of primary and secondary information from each other.
2. Elimination of redundant and secondary information, such as lists of objects, phenomena, and facts. Redundant and secondary information can be removed by omitting words, phrases, sentence fragments, or entire sentences.
3. Compression of the original information through generalization, replacing specific details with more general terms.

The following methods are employed to compress the primary text:

1. Exclusion, which involves eliminating repetitions, explanatory constructions, homogeneous sentence parts, examples, quotations, and details. These elements do not significantly impact the specificity of the author's thinking, reasoning, or description. Consequently, removing such units does not impede the adequate comprehension of the main idea of the text. Synonyms fall into this category.
2. Generalization, which entails summarizing sentences, sentence fragments, facts, and related events. For instance, homogeneous entities can be replaced with a generic term, and hyponyms can be substituted with hypernyms. The generalization technique involves lexical transformations in the text.
3. Simplification, which encompasses combining multiple sentences into one, replacing a sentence or its part with a pronoun, converting a complex sentence into a simple one, breaking down a complex sentence into simpler sentences, replacing a sentence fragment with a synonym, and replacing direct statements with indirect ones. This technique involves grammatical transformations in the text.

Every method of information compression employed in the research is founded upon the information source model, particularly the model of information redundancy.

III. BODY

To compress information, it is necessary to possess knowledge about the type of information that can undergo compression. Without such information, it is impossible to make informed decisions regarding the transformations that can reduce the text size. This type of information is utilized during the compression process as well as in the subsequent expansion or decompression of the information. Building or parameterizing the information redundancy model can also occur during the compression step. Methods that allow for the transformation of the information redundancy model based on data are referred to as adaptive methods.

The compression mechanism serves a crucial role not only in saving text space but also in encoding information. Through compression, information in the text can be encoded in condensed and compact language structures. The outcome of compression is a product with a minimalist form and maximum content volume.

Compression fulfills several important functions. Firstly, it enables the economical allocation of text space. It should be noted that the compact nature of the text necessitates the frequent use of elliptical constructions. With elliptical constructions, maximal informativeness can be achieved in a minimalist form. This outcome is attained by omitting units that can be easily inferred from the context without significant loss.

Compression involves not only condensing structures but also omitting elements that can be inferred from the context

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

without significant loss. The compression mechanism initiates the inference process because the more condensed a structure is, the more cognitive effort is required to decode it. Thus, the process of inference becomes vital in decoding the encoded information.

As explicit verbal information becomes more condensed, the importance of implicit information integrated into the text increases. Implications acquire greater semantic weight and serve as guides in the process of text perception. It's important to note that the compression mechanism necessitates the reliance on background knowledge. Background knowledge aids the recipient in understanding the implications conveyed by the author. The gap between "what is said" and "what is understood" is bridged through implicit inferences, utilizing the knowledge base stored in the listener's or reader's memory. During compression, the meaning of the text is not readily apparent on the surface but requires abstraction and extraction. Background knowledge, which represents a collection of shared meanings within a linguistic and cultural community, functions as a "guide" to the meaning of the text.

The compression mechanism also influences the number of possible interpretations of the text. Compression creates the effect of "incomplete speech," requiring active participation from the recipient in deciphering the intended meaning. The recipient doesn't receive a pre-defined meaning but abstracts the meaning from the information based on their own mental inventory. Each recipient imparts their individual meaning or interpretation onto the text, enriching its semantic space from their own perspective. Different recipients may have different interpretations, expanding the understanding of the text beyond the author's original intention or the intended audience's perspective.

In summary, the compression mechanism is primarily employed as a tool for saving the text space. It also facilitates the encoding of information into compact structures, heightening the significance of processes such as implication and inference, and giving rise to diverse interpretations of the text.

IV. RESULTS AND DISCUSSION

When discussing the fundamental principles of text compression, the principle of minimum effort emerges as a prominent factor. However, a comprehensive analysis of the compression issue in terms of meaning expression necessitates the quality control of message compression to avert potential communication conflicts.

In essence, compression primarily reflects the regulation of language behavior through principles such as the principle of least effort and the principle of economy. Notably, the principle of minimum effort in linguistics was initially distinguished by Courtenay (Some General Remarks on Linguistics and Language. Inaugural lecture given at St. Petersburg in Dec. 1870 (1972, p. 49-80), while French linguist Martine further developed the concept based on extensive linguistic material. Martine posited the idea of human maturity in minimizing mental and physical exertion, stating, "The term “economy” encompasses everything: the elimination of unnecessary distinctions, the emergence of new distinctions, and the preservation of existing situations. Linguistic economy is a synthesis of active forces" [4], which aims to eliminate extraneous information and prevent information overload. Naturally, the rules established by Grice incorporate the inclusion of essential information for effective dialogue.

Determining criteria of significance specific to individuals or communication contexts is an essential factor when considering the balance between informativeness and expressiveness in a compressed message.

Undoubtedly, linguistic compression should be approached in a comprehensive manner due to its multidimensional nature. Confirming its multidimensionality, linguistic compression as a phenomenon unveils the essence and specificity of extralinguistic events through its distinct techniques. Human thinking and language exhibit a tendency to conserve effort and resources, summarizing and generalizing depicted events. When describing various life situations, elementary thoughts and the sentences expressing them can be condensed in various ways. Linguistic compression is a product of human cognition and the constraints of time, as it is impossible to convey every detail of actual events through language.

Compression is not solely a linguistic technique or a means of organizing thought in terms of content and expression but primarily represents a cognitive mechanism for organizing and representing reality, thereby confirming the inseparable connection between language, and thought.

We conducted a qualitative survey involving a sample of 50 text constructors, consisting of undergraduate and graduate students studying English Philology at the University of Georgia during the fall academic year of 2022-23. The survey encompassed one embedded unit, ensuring gender balance, and the respondents were selected using a randomizing principle. The research material, comprising scientific articles, was provided in English.

The same text to be compressed was given to all text constructors. Within the frames of the article, we present only two passages from the original text and some illustrative models of the compressed versions.

Original Passage # 1

“Globalization” - certainly no word in recent memory has meant so many different things to different people and has evoked as much emotion. Some see it as nirvana - a blessed state of universal peace and prosperity - while others condemn it as a new kind of chaos.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

Original Passage # 2

The few intrepid adventurers and travelers of past centuries who brought distant societies together have given way to the waves and tides of migrants, as well as hundreds of millions of tourists jetting around the world. All these comings and goings deepen and broaden the connections among far parts of the world and facilitate the transmission of goods, ideas and cultures.

Illustrative models of the abstracts generated by female text constructors:

1. In the abstracts generated by 14 female text constructors, the text space was not reduced due to extensive use of paraphrasis: e.g., *"the term "Globalization" **elicits diverse interpretations and evokes strong** emotions. It is viewed by some as a utopian vision of worldwide harmony and affluence, while others perceive it as a disruptive force leading to chaos".*
2. In the abstracts generated by 7 female text constructors, some explanatory comments were preserved: e.g., *"Historically, the term "Globalization" **meant many different things to** different people. Some see it as nirvana - **a blessed state of universal peace and prosperity**, while others consider it to be a new kind of chaos".*
3. In the abstracts generated by 4 female text constructors, stylistic devices were applied: e.g., *"Globalization" means so many different things to different people. For example, some see it as nirvana while others believe that it leads us to **a tangled mess**.*

Table 1 – Percentage of the use of paraphrasis, explanatory comments and stylistic devices in the abstracts produced by female text constructors.

	paraphrasis	explanatory comments	stylistic devices
Female	56 %	28 %	16 %

Illustrative models of the abstracts generated by male text constructors:

1. In the abstracts generated by 19 male text constructors, the text space was reduced due to minimal use of paraphrasis: e.g., *"**Globalization means many different things to different people**".*
2. In the abstracts generated by 8 male text constructors, some explanatory comments were preserved: e.g., *"Globalization" means so many different things to different people. Some see it as nirvana - **the promised land**, while others condemn it as a chaos".*
3. In the abstracts generated by male text constructors, no stylistic devices were evidenced.

Table 2 – Percentage of the use of paraphrasis, explanatory comments and stylistic devices in the abstracts produced by male text constructors.

	paraphrasis	explanatory comments	stylistic devices
Male	76 %	24 %	0 %

Illustrative models of the abstracts generated by female text constructors:

1. In the abstracts generated by 15 female text constructors, the text space was not reduced due to extensive use of substitutes: e.g., *"The few **fearless** adventurers of past centuries who brought distant societies together have given way to **a great number of** refugees and immigrants fleeing across borders, as well as **jetting around the world**. All **this mobility expands** the connections among far parts of the world and eases the transmission of goods, ideas and cultures".*
2. In the abstracts generated by 7 female text constructors, some figurative elements were preserved: e.g., *"The few intrepid adventurers and travelers of past centuries who brought distant societies together have given way to **thousands and even millions of refugees and immigrants fleeing across borders**, as well as hundreds of millions of tourists jetting around the world. All these comings and goings deepen and broaden the connections among far parts of the world and facilitate the transmission of goods, ideas and cultures".*
3. In the abstracts generated by 3 female text constructors, some explanatory comments were preserved: e.g., *"The few intrepid adventurers and travelers of past centuries who brought distant societies together have given way to the waves and tides of migrants, as well as hundreds of millions of tourists jetting around the world. All these comings and goings deepen and broaden the connections among far parts of the world and **pave the way for** the transmission of goods, ideas and cultures".*

Table 3 – Percentage of the extensive use of substitutes, figurative elements and explanatory comments in the abstracts produced by female text constructors.

	paraphrasis	explanatory comments	stylistic devices
Female	60 %	28 %	12 %

Illustrative models of the abstracts generated by male text constructors:

1. In the abstracts generated by 23 male text constructors, the text space was reduced due to the minimal use of paraphrasis: e.g., *"**Many travelers brought communities together resulting in broadening** the connections of the world among far parts as well as **facilitating the transmission** of goods, ideas and cultures".*

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

2. In the abstracts generated by 2 male text constructors, some explanatory comments were preserved: e.g., "*All this constant movement broadens the connections among far parts of the world and facilitates the transmission of goods, ideas and cultures*".
3. In the abstracts generated by male text constructors, no stylistic devices were evidenced.

Table 4 – Percentage of the use of paraphrasis, explanatory comments and stylistic devices in the abstracts produced by male text constructors.

	Paraphrasis	Explanatory Comments	Stylistic Devices
Male	92 %	8 %	0 %

As a result of the analysis of the abstracts generated by female text constructors, some regularities emerged:

1. The text space was not reduced due to the extensive use of paraphrasis.
2. Some explanatory comments were preserved.
3. Stylistic devices were applied.
4. The text space was not reduced due to extensive use of substitutes.
5. Some figurative elements were preserved.

While the analysis of the abstracts generated by male text constructors evidenced the following regularities:

1. The text space was reduced due to minimal use of paraphrasis.
2. Some explanatory comments were preserved.
3. No stylistic devices were evidenced.

V. CONCLUSION

The comparative and content analyses of the abstracts generated by female and male text constructors exhibit different degree of subjectivity and individuality. The subjectivity and individuality of the abstracts generated by text constructors is marked by gender differences.

The research findings have substantiated the significance of redundancy when examining the phenomenon of linguistic compression.

These principles form the basis for the development of a compression model. The development of a model for scientific text compression involves the modeling of the compression process.

Hence, compression can be viewed as a fundamental cognitive operation that initiates the process of information condensation and influences the selection of concise language structures. The cognitive foundation of the compression procedure is the universal principle of economy.

REFERENCES

[1] Borko, H. Abstracting Concepts and Methods / H. Borko, C. Bernier. New York: Academic Press. 1975.

[2] Hutchins, J. Summarization: Some problems and methods / J. Hutchins // Proceedings Informatics 9: Meaning the frontier of Informatics. — London: Aslib. 1987.

[3] Mann W., Matthiessen Ch., & Thompson S. Rhetorical structure theory and text analysis. // Discourse Description. Amsterdam: Benjamins. 1992.

[4] Martinet, A. Éléments De Linguistique Générale. volume 349 de la Collection Armand Colin. 1960.