Metalinguistic Indexing: A Quantitative Framework for Analyzing Oppositional Semantics in Computer-Mediated Communication

Alexander Barkovich
Department of Print Technology and Media Communications
Belarusian State Technological University
Minsk, Belarus
albark@tut.by, 0000-0001-8469-8431

***Abstract*—The paper examines the shallowness of semantic data processing from a metalinguistic perspective as a fundamental problem. The generalization of empirical material into metadata is considered as an important stage in knowledge generation. It is a highly relevant vector for improving Artificial Intelligence. The oppositional dimension of data significance was characterized and the *Oppositionality Index* was presented. Essentially, it provides an instrument for algorithmically compatible and discrete parameterization of communication data in large volumes of computer-mediated content.**

***Keywords—Computational Linguistics, computer-mediated communication, Artificial Intelligence, semantic analysis, metalinguistic indexing, Oppositionality Index, modeling*.**

I. INTRODUCTION

Quantitative approaches to communication analysis have enabled the processing and exploration of data on a scale previously unimaginable [1, 2, 3]. At the same time, the gap between rapidly developing practice, armed with statistical tools, and traditionally inert theory has long been evident. This led A. Einstein, back in 1942, to call his time an era of "…perfection of means and confusion of goals." [4: 113]. The trend was noted quite astutely: despite significant quantitative advances, computer-aided science has always retained an inherent need for equally reliable qualitative, or meaningful, support. The meaningful aspect of communication is classically embodied in *semantics* as the problem domain of meaning, or "value". In the field of computer science, the slang understanding of semantics is significantly narrower, limited to formal requirements for the conformity of program "text" to predefined templates. However, this does not preclude its broad interpretation, which is important for our study.

In reality, there is a certain tension between the possibilities of computer-aided quantification and the subtle demands of a qualitative understanding of communication. This allowed N. Carr to note the meaningful "shallows" of modern communication scientific support [5]. The most important principle of scientific activity—*systematicity*—presupposes a harmonious relationship between the formal (quantitative) and meaningful (qualitative) aspects of *knowledge*. It is no coincidence that this very state of affairs preoccupied the creator of the World Wide Web, T. Berners-Lee. A short time, in historical terms, after the establishment of an effective communication network, in 2006, a highly indicative meaningful ("value") vector of its development was proclaimed—the *Semantic Web*:

> "However, the Semantic Web, a vision of extending and adding value to the Web, is intended to exploit the possibilities of logical assertion over linked relational data to allow the automation of much information processing" [6: 5].

Taking into account the semantic specifics of computer-mediated communication contributes to the formation of a holistic understanding of the content of communication based on the *metalinguistic paradigm*. And one of the most important tools of the metalinguistic paradigm is the *indexing* technique.

An equally important cornerstone in this regard is the *oppositional nature* of communication structuring, fundamental to computer technologies, which are entirely based on this principle through binarity. In this regard, the method of *metalinguistic indexing* through the *Oppositionality Index* has significant potential. This paper proposes an operationalization of qualitative semantic concepts for quantitative analysis.

II. DISCUSSION

The Semantic Challenge of Computer-Mediated Communication: beyond Lexical Surface to Artificially Structured Meaning

The transformative impact of computing fundamentally reshaped communication [7: 154]. Indeed, the revolutionary character of the computer in improving human relations is comparable to major inventions such as the "wheel" or the mastery of "fire". The contemporary communicational landscape is characterized by significant *computerized data* expansion.

As human interaction becomes increasingly computerized, its base, *language*, undergoes profound changes. No living language has remained untouched by these developments in the social and cultural spheres, as evidenced:

> "Language is changing due to the computerization of human interaction. No living language has remained uninvolved in the developments in the social and cultural spheres:

millions of texts are stored on servers operating on the Internet" [8: 3].

Moreover, many *new codes* are actively used here, especially artificial ones, and their range is constantly expanding. This diversity has led to the emergence of computer-mediated communication as a distinct language continuum [9, 10, 11]. The communication that occurs when people interact with each other through networked computers is a broad *multi-modal domain* with a specific metalinguistic superstructure:

> "Computer networks are often considered a medium of communication distinct from writing and speaking" [12: 614].

In natural language, the shallowness of semantic interpretation is compensated by the breadth of context and paralinguistic means. There is substantial evidence that users cannot compensate textually for the absence of auditory and gestural cues in computer-mediated communication, particularly. The causal relationship here is that the inherent limitations of the medium (being primarily text-only) and the compensatory strategies employed by users inspire the application of some new tools for data processing. Participants in these digital interactions even perceive computer-mediated *subjectivity* as different from the traditional one, sometimes as a hybrid. And a new personalization, *Artificial Intelligence*, is formed with its own unique limitations and potential:

> "But AI also has the potential to improve human communication by augmenting our natural ability to communicate with one another and improving the affordances of such interactions in computer-mediated communication channels" [13: 99].

However, the new cognitive situation in computer-mediated circumstances still presupposes a meaningful interpretation of communication, but the problem lies in the *depth* of such analysis [14, 15, 16]. Particularly, despite the vast quantities of textual data readily available within computer-mediated communication environments, a significant challenge persists: the semantic representation of language in these contexts often remains predominantly *lexical*. But traditional lexical analysis, situationally designed for the computerized form of communication, struggles to capture some implicit nuances of textual expressions, thereby limiting their richness of meaning [17, 18]. And this complicates the way to achieve rich quality of *metadata* generalization. The investigation of text without considering the context is suitable only for the analysis of dictionary definitions, since it does not have regard to the communicational meaning of words—the information [19].

This situation has generated a high demand for the creation and development of formal-language compatible analytical *frameworks*, a need that extends beyond the requirements for various quantitative applications. This underscores the imperative for adopted approaches that can provide the qualitative richness of metadata and knowledge. And the sheer volume of online interaction necessitates *modeling* data in its processing. The kind of such modeling is *metalinguistic indexing*.

III. METHODOLOGY

Metalinguistic Indexing: A Framework for Enhanced Semantic Analysis

Identifying and modeling the content of communication is the real path to the ergonomic and effective development of scientific support of computer-mediated communication. In this relation, the *semantic enforcement* of the metalinguistic coverage of communication content is substantiated as ensuring a high-quality interpretation of data, which allows for significant optimization of Artificial Intelligence programs [20, 21].

To overcome the limitations of current approaches to Natural Language Processing and to bridge the gap between quantitative linguistic analysis and the consideration of essential qualitative parameters of language, metalinguistic indexing is a promising technique. This approach involves the parameterization of annotated texts, particularly those found in computer-mediated communication environments, through the application of meta-language indicators, or indexes. The fundamental objective in this context is the strategic creation of a robust *instrumental base*—founded on the frequency and distribution of linguistic items [22]. This entails adapting and improving quantitatively compatible qualitative instrumentarium.

The common framework of Computational Linguistics and Information Theory provides such scientifically approved set of tools. The practice of computer-mediated communication, being determined by its formalized content, necessitates a *syncretic study* of data [23, 24, 25]. This emphasis on the specific nature of computer-mediated communication ensures that metalinguistic indexing is not a limited methodological innovation but is deeply embedded within an objective frame that specifically addresses the unique characteristics of communication. It is referring to the interplay between a linguistic form and its empirical distribution, ensuring the practical feasibility of metalinguistic indexing, which increases the value of the corresponding processing of data.

While the semantics of computer-mediated data often present challenges in terms of transparency for computer systems, the digital environment itself offers significant advantages for communication analysis. This environment comprises an immense and continuously expanding *volume* of texts and media artifacts, a quantity already considerable when compared to traditional communication, and one that is constantly growing due to the separate storage of each speech fragment. The scale of language data available on the Internet is staggering, numbering in millions.

Programs built upon statistical models of communication are uniquely capable of successfully processing huge arrays of adopted texts that come with *specific markup*. This capability is already evident in the functioning of the Internet and various text corpora. Such a foundation allows for the modeling of representative meta-descriptions, particularly those derived from primarily annotated data material. Particularly, the unprecedented scale of computer-mediated communication, a defining characteristic of the computerized data continuum, is precisely what enables the effective application of related methodology for "…very extensive collections of transcribed utterances or written texts." [26: 1].

This approach provides the scientific foundation necessary for metalinguistic indexing. The formalized nature of computer-mediated communication makes metalinguistic indexing a useful instrument, thereby creating the *optimal conditions* for developing and validating an admissible level of communication analysis and synthesis.

IV. RESULTS

Quantifying Oppositional Quality: The Oppositionality Index

Metalinguistic indexing serves as a powerful tool for representing and verifying the *oppositional quality* inherent in communication, which ultimately contributes to the high-quality processing of information-compatible semantics. At a fundamental level, oppositional (or binary) relations are recognized as key organizers of semantic structure in any form of communication. For the purposes of this framework, an ***Oppositionality Index*** (***I_o***) is operationally defined as a relationship between a pair of linguistic items (e.g., words, phrases) that are semantically related and functionally contrastive within a given context, such as synonyms, antonyms, paronyms, or similar words. Their significance becomes even more pronounced when considering the scale of a polycode or multi-modal communicational environment.

This focus on opposition is deeply rooted in foundational linguistic theory. As F. de Saussure posited:

> "Language makes sense only through the differences and contrasts (*binary oppositions*) that it sets up. These differences and contrasts are the structure out of which meanings are made" [27: 120].

This framework reveals that the *Oppositionality Index* (I_o) is not merely a statistical tool for quantifying frequency. Instead, it represents an attempt to operationalize a *fundamental principle* of the computer mediation of communication. If meanings are substantially different, then quantifying these differences—these oppositions—becomes a direct pathway to understanding the underlying semantic logic. This elevates I_o from a simple metric to a reasonable tool for investigating the very fabric of communicational meaning.

The *Oppositionality Index*, is formally defined as an indicator that quantifies the ratio of the frequency of occurrence (token count) of one oppositional element to the frequency of occurrence of another element, both of which are used within the same text or text corpus. Formally, for two oppositional elements, *a* and *b*, within a given corpus, the *Oppositionality Index* of *a* with relation to *b* is defined as:

$$I_o(a,b) = frequency(a) / frequency(b),$$

where frequencies are counts of tokens in some representative arrays of data.

This index provides a means for the discrete substantiation of primary linguistic generalizations, thereby enabling synthetic modeling and the derivation of *quantitatively compatible* semantics. Such a tool is particularly valuable for addressing issues that remain unsolvable for current computer programs, such as programmatically determining choices between different categories of tokens.

The application of the *Oppositionality Index* (I_o) is particularly well-suited to computer-mediated text wholes, hyper-texts, and Internet discourse. These digital linguistic environments need for *demanding modeling* due to their inherent characteristics, especially their immense size. This is best realized not through uncontrolled internet searches, but through the analysis of large-scale, structured text corpora, which provide stable and replicable data. It addresses a significant challenge encountered in traditional linguistic analysis: the problem of the representativeness threshold. Unlike smaller, manually curated texts, the vastness of digitalized data ensures a high degree of communicational representativeness.

The concept of the *Oppositionality Index* (I_o) draws a parallel with established metrics in Computational Linguistics, such as the *Type-Token Ratio* (*TTR*):

> "We determine the *type-token ratio* by dividing the number of types in a corpus by the number of tokens. The result is sometimes multiplied by 100 to express the type-token ratio as a percentage. This allows us to measure vocabulary variation between corpora—the closer the result is to 1 (or 100 if it's a percentage), the greater the vocabulary variation; the further the result is from 100, the less the vocabulary variation" [26: 50].

TTR is a widely recognized indicator used to quantify vocabulary diversity within a text or corpus. Just as *TTR* provides a quantitative measure of lexical richness, I_o offers a quantitative measure of oppositional quality. This parallel suggests that computerized text arrays are amenable to similar processing for various other data parameters, further solidifying the methodological foundation for metalinguistic indexing.

In addition to the parallel with *TTR*, the *Oppositionality Index* can be positioned within the broader family of computer-aided metrics. Traditional well-known measures such as *Mutual Information* (*MI*), *Pointwise Mutual Information* (*PMI*), or *Log-Likelihood Scores* (*LLS*) are primarily oriented towards collocational strength and the associative proximity of communication units [28, 29, 30]. By contrast, I_o explicitly targets differential semantic value within established oppositional sets, rather than general co-occurrence probabilities. This makes the index particularly suitable for cases where the analyst is interested not in "how strongly items co-occur," but in "how strongly they compete" as alternative data options in similar contexts. As a result, the *Oppositionality Index* complements, rather than replaces, existing distributional metrics and provides an additional dimension for semantic modeling.

The computer-aided communicational environment, therefore, is not just a source of statistical data but is the optimal subject for studying oppositionality. The inherent characteristics of computer-mediated communication—its vast volume, ready accessibility, and machine-readability—directly facilitate the application of quantitative measures like I_o. Such studies would be prohibitively difficult or yield less representative results with traditional, smaller-scale communication material. This implies that the computer data is not only transforming communication itself but also providing the means for a more scientifically grounded understanding of its fundamental structures.

V. EXPERIMENTAL EVALUATION

Empirical Analysis of Oppositional Semantics: Communicational—Communicative—Communicatory

The English language presents a complex oppositional landscape with the triad *communicational—communicative—communicatory.* An analysis using token frequencies from the *Corpus of Global Web-Based English* (*GloWbE*), a publicly available *1.9*-billion-word corpus

representing English as used in 20 countries, reveals the following verified token counts:

- *communicative*: *34,120*;
- *communicatory*: *1,212*;
- *communicational*: *498*.

These values were obtained directly from the official *GloWbE* interface (*https://www.english-corpora.org/glowbe/*) on 8 November 2025 and reflect actual word usage in real-world web texts, not uncontrolled web search "hit counts."

To analyze this multi-component opposition, a ***Compound Oppositionality Index*** (***I_oc***) is introduced, as an option for considering non-binary oppositions, too. This index is designed for situations where oppositionality involves more than two main components, providing a more sophisticated processing of the relationships. The individual dyad I_o values within this triad are as follows:

- *communicational—communicative*: $I_o \approx 0.015$ (*498 / 34,120*);
- *communicational—communicatory*: $I_o \approx 0.411$ (*498 / 1,212*);
- *communicative—communicational*: $I_o \approx 68.51$ (*34,120 / 498*);
- *communicative—communicatory*: $I_o \approx 28.15$ (*34,120 / 1,212*);
- *communicatory—communicational*: $I_o \approx 2.43$ (*1,212 / 498*);
- *communicatory—communicative*: $I_o \approx 0.035$ (*1,212 / 34,120*).

Compound Oppositionality Index (I_o*c*) for *communicational—communicative—communicatory* is processed by summing the I_o values for dyads where each component appears first:

- for *communicational*: *0.015 + 0.411 = 0.426*;
- for *communicative*: *68.51 + 28.15 = 96.66*;
- for *communicatory*: *2.43 + 0.035 = 2.465*.

The final I_o*c* representation is *0.426—96.66—2.465*, clearly confirming the leading position of *communicative*, with an I_o*c* of approximately *96.66*.

English speech, in this context, demonstrates a variable presentation of "communicational" semantics. While *communicational* might seem grammatically correct, *communicatory* and especially *communicative* hold significant oppositional priority, relegating *communicational* to the third place. This priority is particularly noteworthy in computer-mediated communication, where the efficiency of binary oppositionality holds considerable weight. The analysis of the English triad, particularly the strong dominance of *communicative* and the low frequency of *communicational*, is consistent with actual corpus data. The lexeme *communicational* is phonetically longer and graphically longer by three characters than *communicative*, which may partly explain the preference for the more concise and euphonious form. This finding suggests the robustness of an ergonomic principle in its selection within speech. Additionally, regardless of the specific language, the inherent characteristics of computer-mediated communication—such as typing speed, visual presentation, and the dynamics of digital interaction—create a selective pressure favoring shorter, more euphonious, and graphically concise items. This suggests a universal underlying trend in language adaptation to digital environments, where efficiency and ease of use influence lexical choices more profoundly than traditional semantic or derivational logic alone.

VI. CONCLUSION

The empirical analyses presented herein suggest that speech practice does not always align perfectly with the quantitative meta-descriptions of communication. Instead, it appears to adhere more steadily to the general underlying qualitative logic of semantics as it is manifested in actual usage.

The computer-mediated communication environment is characterized by the flow of enormous amounts of data. However, the vast majority of this data is only nominally or superficially semantically interpreted. The linguistic dimension of communication demonstrates that, to date, the semantic interpretation of data has been limited by the resources of dictionaries. The verifiability of semantic processing can be significantly improved by providing it with a level of oppositional interpretation of metadata, as confirmed by the functional relevance of the *Oppositionality Index* (I_o). The index, presented in this paper, is an example of a relevant metalinguistic tool for verifying the substantive dimension of communication and enabling an empirically grounded, knowledge-based interpretation of data.

The rapid and dynamic evolution of data in computer-mediated environments often contradicts traditional, often slowly updated, research canons. This necessitates new, data-driven approaches like metalinguistic indexing to accurately reflect contemporary language reality. This suggests a future where science, particularly in digital contexts, will increasingly rely on real-time data analysis over static resources.

The instrumentarium of metalinguistic indexing and its specialized tools, particularly the *Oppositionality Index*, demonstrate advanced linguistic persuasiveness in probing peculiar semantic issues within computer-mediated communication. These tools are not merely descriptive; they offer a well-developed analytical framework for understanding the complex collisions of unit choice and meaningful functionality of the communication environment.

REFERENCES

[1] A. Belletti and C. Chesi, "A syntactic approach toward the interpretation of some distributional frequencies: comparing relative clauses in Italian corpora and in elicited production," in Rivista di Grammatica Generativa, No. 36, 2014, pp. 1–28.

[2] J. Craenenbroeck, M. van Koppen, and A. van den Bosch, "A quantitative-theoretical analysis of syntactic microvariation: Word order in Dutch verb clusters," in Language, No. 95(2), 2019, pp. 333–370.

[3] A. Massaro and G. Samo, "Prompting Metalinguistic Awareness in Large Language Models: ChatGPT and Bias Effects on the Grammar of Italian and Italian Varieties," in Verbum, No. 14, 2023, pp. 1–11.

[4] A. Einstein, "Out of My Later Years," New York: Philosophical Library, 1950.

[5] N. Carr, "The Shallows: What the Internet Is Doing to Our Brains," New York: Norton, W.,W. & Company, Inc., 2010.

[6] T. Berners-Lee, W. Hall, J. A. Hendler, K. O'Hara, N. Shadbolt, and D. J. Weitzner, "A Framework for Web Science," in Foundations and Trends in Web Science, Vol. 1, No. 1, 2006, pp 1–130.

[7] A. Barkovich, "Computer Discourse? On the Specifics of the Language Meta Descriptions" [Компьютерный дискурс? О специфике метаописаний языка], in Visnyk of Zaporizhzhya

National University. Philological Sciences, No. 2, 2015, pp. 153–161.

[8] A. Barkovich, “Informational Linguistics: The New Communicational Reality,” Cambridge: Cambridge Scholars Publishing, 2023.

[9] C. Eliasmith, “How to build a brain: from function to implementation,” in Synthese, No. 159(3), 2007, pp. 373–388.

[10] A. L. Gonzales and J. T. Hancock, “Identity shift in computer-mediated environments,” in Media Psychology, No. 11, 2008, pp. 167–185.

[11] R. Katzir, “Why large language models are poor theories of human linguistic cognition. A reply to Piantadosi,” in Biolinguistics, No. 17, 2023, Article e13153.

[12] S. C. Herring, “Computer-mediated discourse,” in Handbook of Discourse Analysis, editors D. Tannen, D. Schiffrin and H. Hamilton, Oxford : Blackwell, 2001, pp. 612–634.

[13] J. T. Hancock, M. Naaman, and K. Levy, “AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations,” in Journal of Computer-Mediated Communication, No. 25(1), 2020, pp. 89–100.

[14] J. Donath, “Signals in Social Supernets,” in Journal of Computer-Mediated Communication, No. 13(1), 2007, pp. 231–251.

[15] J. Camacho-Collados and M. T. Pilehvar, “From word to sense embeddings: A survey on vector representations of meaning,” in Journal of Artificial Intelligence Research, No. 63, 2018, pp. 743–788.

[16] T. Linzen and M. Baroni, “Syntactic structure from deep learning,” in Annual Review of Linguistics, No. 7, 2021, pp. 195–212.

[17] M. Khalifa and V. Liu, “Semantic network representation of computer-mediated discussions: Conceptual facilitation form and knowledge acquisition,” in Omega, No. 36(2), 2008, pp. 252–266.

[18] N. Kobyshev, H. Riemenschneider, and L. van Gool, “Matching features correctly through semantic understanding,” in Proceedings of 3DV, 2014, pp. 472–479.

[19] A. Barkovich, “Informational Linguistics: Computer, Internet, Artificial Intelligence and Language,” in IEEE 1st International Conference on Artificial Intelligence in Information and Communication (ICAIIC 2019), pp. 8–13.

[20] P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun, “Predicting Deeper into the Future of Semantic Segmentation,” in Proceedings of ICCV 2017, pp. 648–657.

[21] A. Tao, K. Sapra, and B. Catanzaro, “Hierarchical multi-scale attention for semantic segmentation,” 2020, arXiv preprint arXiv:2005.10821.

[22] Z. Jiang, J. Araki, H. Ding, and G. Neubig, “How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering,” in Transactions of the Association for Computational Linguistics, No. 9, 2021, pp. 962–977.

[23] M. L. Anderson, B. Lee, J. Go, S. Li, B. Sutandio, and L. Y. Zhou, in “On the Types, Frequency, Uses and Characteristics of Meta-language in Conversation,” in Proceedings of the 28th Annual Conference of the Cognitive Science Society, Vancouver, 2006, pp. 973–978.

[24] M. A. Riordan and R. J. Kreuz, “Cues in computer-mediated communication: A corpus analysis,” in Computers in Human Behavior, No. 26(6), 2010, pp. 1806–1817.

[25] A. Roberts, C. Raffel, and N. Shazeer, “How much knowledge can you pack into the parameters of a language model?” in Proceedings of EMNLP 2020, pp. 5418–5426.

[26] T. McEnery and A. Hardie, “Corpus Linguistics: Method, Theory and Practice,” Cambridge: Cambridge University Press, 2012.

[27] F. de Saussure, “Course in General Linguistics,” New York: Philosophical Library, 1959.

[28] K. W. Church and P. Hanks, “Word association norms, mutual information, and lexicography,” in Computational Linguistics, No. 16(1), 1990, pp. 22–29.

[29] P. D. Turney, “Mining the Web for synonyms: PMI-IR versus LSA on TOEFL,” in Proceedings of ECML 2001, pp. 491–502.

[30] T. Dunning, “Accurate methods for the statistics of surprise and coincidence,” in Computational Linguistics, No. 19(1), 1993, pp. 61–74.