Journal of Systemics, Cybernetics and Informatics (2025) 23(6), 13-19
https://doi.org/10.54808/JSCI.23.06.13

AI-Enhanced Transdisciplinary Data Encoding for LLMs Training

Rusudan MAKHACHASHVILI
Borys Grinchenko Kyiv Metropolitan University,
Bulvarno-Kudryavska-st., 18/2,
Kyiv, Ukraine

Natalia BOBER
Borys Grinchenko Kyiv Metropolitan University,
Bulvarno-Kudryavska-st., 18/2,
Kyiv, Ukraine

ABSTRACT

The rapid advancement of artificial intelligence (AI) has reshaped linguistic data encoding, particularly for Large Language Models (LLMs). AI-driven annotation techniques enable efficient lexical processing, semantic disambiguation, and automated neology tagging, refining computational language modeling across transdisciplinary domains.

This study explores AI-enhanced methodologies for encoding linguistic data for LLM training. AI-assisted lexicographic workflows enable LLMs to dynamically adjust to linguistic evolution while ensuring scalable annotation across diverse transdisciplinary corpora. LLMs trained on transdisciplinary lexicons can generate cross-modal language interpretations, refining machine-generated discourse across domains. The inquiry objective is the investigation of the innovative philosophic aspects cyberspace through the lenses of the language development processes as it informs AI models elaboration, LLMs training, and digital communication. The study design is the disclosure of cyberspace as an ontology model and as a logosphere model. Two data encoding projects, developed by the authors, serve as foundational elements for this investigation.

A methodology and AI-augmented, AI-performed protocols of computer vocabulary innovative elements phenomenological features identification is introduced supplying the template for a new study field – phenomenological, AI-enhanced digital neology, neography and neosemiotics. Transdisciplinary educational applications of these approaches to data encoding, include: training AI-enhanced NLP models for transdisciplinary communication; developing standardized linguistic annotation protocols**,** ensuring interoperability across AI-driven lexicographic systems; integrating transdisciplinary discourse structures into machine-learning lexicons**,** refining AI adaptive language comprehension.

Keywords: Large Language Models, Generative AI, AI-driven Linguistics, Linguistic Philosophy, Innovative Logosphere of Cyberspace, Data Encoding; Digital Communication

1. INTRODUCTION

At the turn of the XX-XXI centuries, as an essential product of civilization, computer reality has been gradually separated into an independent existential whole, within which digital media serve not only as a means of transmitting information or interaction, but fulfill their own world-building, sense-building and, consequently, logo-generative potential [12; 31]. Cyberspace, henceforth, is an object of study of a wide range of academic branches – philosophy of modern humanities, psychology, sociology, cultural studies, etc.

By virtue of objective historical and geopolitical context (the cybernetization, globalization, informatization of society, the Globalization of culture – [17; 16; 41] in the new millenium modern English, is a priority communicative medium of primary speech coding, speech and meta-language (terminology) representation of cyberspace elements and structures mapping [7; 32; 37]. Methodological perspectives of the modern view of natural language activity in the area of advanced technologies, include a macro-factor of the vocabulary functional updates [20; 23; 24], a cognitive structure, a segment of supranational worldview, a discursive digital communicative medium [8; 40], that gets extrapolated onto computer mediated discourse and terminology of European and Oriental languages alike**.**

Conditions for the development of modern globalized civilization determine the expansion and refinement of the paradigm of views on the theoretical principles of determining the groundwork and characteristics of the consolidation of the world order, its perception in culture, collective social consciousness and natural language.

Cyberspace stands as an integral environment, demanding new cognition and perception ways via complex philosophic, cultural, social, linguistic approaches, providing unlimited opportunities for human intellect, language development and research.

Given the conceptual system of identification of onto-mental and linguo-mental complex formations to identify constructs of reality, cyberspace and its innovative linguistic shell can be located in the coordinates of such paradigms: 1) philosophy - as *a particular type of substance* – material and ideal reality in the multitude of its forms; a meta-negentropy (the term after Nagib Callaos [4]); 2) anthropology – as an environment for actualization of post-humanistic forms of anthropogenesis; 3) psychology – as psychosomatic and emotional plane of a personality functioning; 4) sociology – as a system of multi-tiered and multi-directional social and communicative relations.

The rapid advancement of artificial intelligence (AI) has reshaped linguistic data encoding, particularly for Large Language Models (LLMs). AI-driven annotation techniques enable efficient lexical processing, semantic disambiguation, and

[1] Peer-editor: Nataliia Lazebna, Hab. Doc., Würzburg University, Germany

automated neology tagging, refining computational language modeling across transdisciplinary domains.

This study explores AI-enhanced methodologies for encoding linguistic data for LLM training, focusing on: neological vocabulary processing through structured multimodal annotation; emotive profiling to optimize sentiment classification; automated and semi-automated tagging for transdisciplinary contextual integration.

By integrating multi-layered corpus tagging and semantic matrices, this approach to data encoding provides high-precision linguistic data for training adaptive and multilingual LLMs.

AI-assisted lexicographic workflows enable LLMs to dynamically adjust to linguistic evolution while ensuring scalable annotation across diverse transdisciplinary corpora. LLMs trained on transdisciplinary lexicons can generate cross-modal language interpretations, refining machine-generated discourse across domains.

Linguophilosophic approach to the study of language encoding in digital communication allows to efficiently investigate lingual manifestation of cyberspace ontology (namely space and time dimensions), to closely study the generic categories and dimensions of cyberanthroposphere, to denote its existential anthropocentric character. Linguistic philosophy of cyberspace is construed in the study through deterministic phenomenological correlation of innovative language of digital technologies with: 1) constructive elements of the substance - space, time, phenomenon, essence; 2) constructive elements of knowledge / cognition - information, episteme, concept, concept; 3) constructive elements of human consciousness - identification, identity, individuality.

The *methodology* is based upon the hypothesis of the cyberspace-related language terminological nature. The dual systematization character of terminology determined the analysis of both linguistic and external (ontological, anthropological, cognitive, social) paradigmatic parameters of cybervocabulary. Due to its polydimensional nature the term acquires the unique, philosophic status (the entity of Being and Language respectively).

The cyber-term as a specific intralingual and extralingual phenomenon turns out to be the means of perception and comprehension to a degree as well as the encoding and categorization source of the modern cyberspace and technosphere. The introduced approach to defining the cyberterm presents a key to comprehending the underlying mechanisms of linguistic actualization, perception and processing of cyberspace.

The study of groundwork principles of universality and philosophic interdisciplinary of natural language development in cyberspace is a parcel of the framework project [25] *TRANSITION: Transformation, Network, Society and Education*.

Two data encoding projects, developed by the authors, serve as foundational elements for this investigation: 1) The Cyber-Speak Project – an AI-enhanced digital neography initiative focusing on computational terminology across European and Asian languages relevant to cyberspace and emerging technologies; 2) The British National Corpus (BNC) Emotive Profiling – a matrix-tagging approach to structuring emotional-state semantics, revealing the interplay between computational markup, corpus annotation, and lexical categorization for sentiment analysis of machine-readable data.

2. FINDINGS

Digital Realm as a Transdisciplinary Ontology

Theoretical issues of holistic, multidimensional language modeling and the reality of its individual areas (one of which within the framework of the linguistic culture at the beginning of the XXI century is the area of advanced computer technology) determined the interaction of a number of concepts that consistently connect ontological system features of sophisticated reality objects, attribute their reception and interpretation (in the field of individual and collective mind), implementation, consolidation and relay the results of the interaction of these features. In this regard, a fundamental dimension of being [1] is defined as a heterogeneous hyperonymic concept that can summarize the polydimensional signs of world order: A world that exists; world that is not subject to direct reception [11; 13], but is in reality; The imaginary, unrealistic world (for example, the idea of perfect, mythological images [15]); A reality that exists objectively, independently of human consciousness (nature, objective laws of world order [6]); A common mode of existence of human society, culture and civilization.

Thus, we can see that within the paradigm of Western traditions, the foundation of theoretical and conceptual perspective parameterization holistic reality modeling is the synthesis and anthropometric ontological principles. Plane integration of aforementioned principles can be considered the system of psycho-mental epistemic concepts that are part of the semantic field of the term "world view" and its multi-substrat and hierarchically heterogeneous derivatives.

That way, world-view is identified as a holistic set of philosophical knowledge of the world [27; 36], which formed during the evaluation results of reality by the knowledge subject. The subjects or bearers of the world picture are individuals and social or professional groups, and ethnic or religious communities. The subject forms a picture of the world based on their own feelings, perceptions, ideas, forms of thought and consciousness. Accordingly, the picture of the world as a result of accumulation of subjective experience, knowledge of world order, based on feelings, beliefs, perceptions and thinking of the individual and the human community, dictates rules of behavior, meaning system [19; 39], which affect the formation and generalization of concepts. Thus, it is determined that mapping framework in the system of modern humanities determines the definition of such leading features of this concept: Systemic images (and links between them); Visual representations of the world and human place in it, information about the relationship between a human and reality (human and nature, human and society, human and human, human and technology); Attitudes of people, and their beliefs, ideals, principles, and knowledge of, meanings and spiritual guidance. Any significant changes in worldview entail changes in a complex system of these elements. The outlook as a consecutive and causative result of the interaction and interpenetration signs of previous constructs and a collective system of ideas about general categories of space, time and movement. According to researchers, the basic elements - the so-called "frame picture of the world" [41] - is a set of first principles or considerations of fundamental assumptions about reality substant parameters and its parts. They cannot be realized by human mind, but are embedded in the picture of the world, because it is necessary to interpret any situation in life, to determine the meaning and to assess what is happening. Some of them, such as motion, causality, identity, time and space may be understood as a priori within the realm of human experience. According to the correspondence principle there are distinguished the following types of reality mapping: The real world - this is an objective physical reality [34]; Illusory picture of the world (the term by Erich Fromm [10]) – the accumulation of distorted, unstructured information in the individual and collective consciousness.

The world view, as a consolidated, multi-dimensional, quasi-Gnostic model of world-built features the following

characteristics: It defines the specific mode of perception and interpretation of events and phenomena; It is the foundation of worldview, based on which people act in the world; It has historically conditioned properties, implying constant dynamic changes of all its subjects' world view.

Language as a particular way of understanding and mapping of reality is partly universal, partly nationally specific [29]. Hence, linguistic picture of the world is as a result of a certain way of reflection of reality in the mind through the lens of language and national, differential historical and cultural features of its speakers.

Each language reflects a natural way of holistic perception [30; 35] and the organization (conceptualization) of the world. The views expressed therein mentioned consist of a single frame of varying degrees of abstraction, which is extrapolated, as mandatory, in individual and collective consciousness of native speakers. Thus, linguistic picture of the world as a set of ideas about the world is a retrospectively and prospectively (based on principles of thinking historicism) arranged integral image of the world, shaped by all parties involved in human mental activity. Language world is a historically constituted communal knowledge, displayed in the language set of ideas about the world is a certain way of conceptualizing reality reflected through the prism of cultural and national characteristics inherent in a particular language community; an interpretation of the world according to national conceptual and structural canons reflection of reality in the minds of a group, which are absorbed by a person in the process of socialization. In this context, the concept of world modeling qualifies as: *Universal, Orderly, Sustainable, Systematic*. While the language model of the world is*: Fragmented, Mobile, Isomorphic to the dynamics of the environment*.

Both types of world modeling are realms of existence and functioning of linguistic units in the minds of the media and help in the reproduction of a coherent picture of the world.

It should be noted that the conditions of modern globalized civilization determine the expansion and refinement of paradigm views on theoretical premises of identifying the principles and characteristics of the consolidation of world order and its perception in the culture, collective social consciousness and natural language. Thus, the intellectualization of modern global culture defines a new approach to understanding the processes of the parallel development of human activity and cognitive (intellectual) experience.

The aforementioned ties into the emergence and methodological development of the concept of "noosphere". Noosphere is defined as the current stage of development of the biosphere, associated with the development of humanity, and is interpreted as a part of the planet and circa-planetary space with traces of human activity.

According to the theory of V. Vernadsky, the noosphere is the third in a sequence of major phases of the Earth as the formation of the geosphere (inanimate) and the biosphere (wildlife). Just as the biosphere is formed by the interaction of all organisms on Earth, the noosphere is composed by all minds interacting. Noosphere is identified as the unity of "nature" and culture (in the broadest interpretation of the latter involving technosphere as a component of cultural space [43]), especially from the moment when the spiritual culture reaches (by force of impact on the biosphere and geosphere) power of a certain 'geological force.

Given the definite unity of nature and culture (in their interaction) there are two stages determined in the development of the noosphere: 1) noosphere stage in its development, in the process of natural development, since the emergence of humans [40]; 2) noosphere that is consciously improving joint efforts of people in the interests of humanity as a whole and each individual separately [8; 41]. The digital dimension of linguistic interoperability of reality stems from its cognitive structure and content of noosphere components: ANTHROPOSPHERE - a set of people, their activities and achievements; SOCIOSPHERE - a set of social factors characteristic of society development and its interaction with nature; TECHNOSPHERE - a set of artificial objects created by man, and natural objects, altered as a result of human activity.

Given the context outlined transformation of initial position awareness of the principles and foundations of the universe integrated modeling, we note that at the turn of the XXI century modern cyberspace as part of the technosphere (and respectively - the noosphere) takes up more space in the public consciousness and functional activity of mankind. As an integral product of civilization, Computer reality (Cyberspace) is gradually separated into independent existential whole. Within its limits the digital media serve not only as a means of transmitting information or interaction, but manifest their own world-building, sense-building, and, consequently, linguogenerative potential.

Based on the conceptual identification system of onto-mental and linguistic-mental complex structures to determine reality constructs, Cyberspace and its innovative linguistic casing can be located within the set of the following philosophic ontological coordinates: A specific type of substance - material and ideal reality united in all forms of development – being [18]; Implementation environment for "post-humanistic" trends of anthropogenesis [38]; A segment of the noosphere (the technosphere); A system of hierarchal social relations [33] – sociosphere; A psychosomatic and emotional plane, the sphere of spiritual experience [37]; A worldview, semiotic model of the world.

Cyberspace, thus, is defined as ***a complex, multidimensional sphere of synthesis of reality, human experience and activity mediated by the digital and information technologies, a component of the technosphere of human existence.***

Digital Data Encoding as a Logosphere

The hypothesis of the study is that the typological characteristics of innovative logosphere of cyberspace as a macro-object of a phenomenological investigation determine the specificity of static configuration and dynamic interaction of formal and substantive constituents of its microstructure.

The philosophic universality of natural language in cyberspace is accessed through is the concept of the **logosphere**, synthetically perceived as 1) the plurality of language units, which are conditionally exhaustive phenomenological realizations of abstract and empirical elements of different spheres of life [2; 5; 14]; 2) the zone of integration of thought, speech, and experience continuums of cultures [3; 21]; 3) the plurality of culturally relevant universal meanings and signs - **semiosphere** [22]; 4) a plurality of transcendent spiritual meanings – **pneumatosphere** [9]. The logocentric approach to integrative research directions, mechanisms, and ways of Cyberspace structuring provides a generalized in-depth understanding of the phenomenological nature of meta-language encoding processes, categorization, mental mapping, meta-language reference, significative correlation, respectively.

Phenomenological lense [28] to the study of innovations in the cyberspace allows to efficiently investigate manifestation of cyberspace integrated ontology, to closely study the dimensions of cyberberspace as an outlook both generic and critical, to expose the phenomenological origin and upstream direction of

cyberspace dynamics as a comprehensive linguistic and communicative structure.

Parameterization principles of a concept of "logos" in the paradigm of the humanities in general, linguophilosophy, and linguistics - in particular, allow to identify the features of logosphere as a complex object system pertaining the following parameters: Ubiquity (inclusiveness); Ontocentricity; Integrativity; Automorphism; Normativity; Lingual substantiality; Phenomenology of thesaurus units; Information-capacity; Referential and semiotic isomorphism of the referent and meaning.

Note that through the fragmented set of qualitative features, logosphere of cyberspace is tangent to the concepts of complex system simulators of linguistic-mental outlook, such as: Model of the world / world view (inclusive, integrative, self-identity); Language picture of the world (phenomenology of linguistic constituents - the ability to summarize and signify objects of reality); Noosphere (onthotsentrism, info-capacity). For the listed set of features the integral notion of logosphere of cyberspace stands as a synthesis of these concepts.

The framework innovation of cyberspace logosphere (CSL) of cyberspace (a multidimensional, complex, dynamic system) is the most comprehensive quantitative and qualitative terms of language representation of the linguistic actualization of being, determined by a number of qualifying conditions of its emergence, existence and development, including: 1) exhaustive synchronization process of the object, phenomenological and anthropological field of computer being and development processes of the ICT meta-language; 2) exhaustive output of parameterization isomorphism of ontological (substance phenomenological), anthropic and digitized structures of reality; 3) flexibility, adaptability and dynamic potential of the vocabulary of the modern languages (heavily influenced by English hegemony) in correlation with the ICT sphere (that is fulfilled, in particular through info-capacity, sign hybridization, the evolution of the basic ontological and functional features of neologisms in relevant areas).

In view of the foregoing, the innovative cyberspace logosphere (ICSL) is defined as: a) a syncretic, consolidated within its semantic scope, plurality of verbal units that are the asymptotically (i.e. in unlimited approximation) exhaustive embodiments of substantive and factual elements of modern computer being; b) as a vertically integrated at the macro and micro levels plurality of ICT thesaurus, its typological specificity are relatively exhaustive phenomenological correlates of multi-substrat elements of computer being.

Given the features of logosphere as specific linguistic-onthological, phenomenological-linguistic and a linguistic-semiotic object, it is possible to distinguish the following typological characteristics of ICSL:

A) The ability to conditionally complete phenomenological realization of substantive identity of the cyberspace in significative characteristics of verbal units that constitute the relevant innovative logosphere. The following typological characteristics of ICL are to be phenomenologized, particularly at the level of the external form of discrete ICSL units. For example, paronymic unit elements of affixation paradigm based on formant dot- one that pertains to the Internet: dot-biz - legal body that implements its activities through Internet, dot-con - offender that performs fraud (con) through Internet (in these units is dot- verbal manifestation graphical point - [.] - as semiotic marker recording Interent protocol address). A meta-term innovation 404 – to be offline for a long while (404 – a semiotic representation of protocol error on the results of an unsuccessful search Internet page). On the internal form level of discrete ICL units: sextuple-u – a metaphtonymic conventional transcoding of an Internet protocol address: www (where: three-double-u - initial transcoding → 3x2-u =-u 6 – a metaphtonymic correlate); 888 in Japanese (pronounced as ぱちぱちぱち, the sound of snapping or clapping) – an online communication formula.Due to a combination of external and internal form configurations of discrete units ICL: for example, an innovation paradigm Web 2.0/Web.3.0/Web 4.0/Web. 5.0 - the newest visual and technological configuration of Internet space where the Web - Internet 2.0 (N.0) is an analogical representation of meaningful semiotic element "a new (improved) version" (operating system, software, software, etc.).

B) Structural density volume, uniformity and conditional completeness of innovative codification of multi-substrat configuration of Cyberspace Logosphere.

Cosequently, the innovative cyberspace logosphere (ICSL) is defined as ***a vertically integrated at the macro and micro levels plurality of innovations of natural languages, which in its typological specificity are a relatively exhaustive phenomenological correlates of the multi-substrat elements of cyberspace***.

From the above system of parametric characteristics of innovative cyberspace logosphere (ICSL) macrostructure it is evident that the principle of hierarchical abstraction correlation powers its integrative macrostructure within the conceptual dyad substance :: substrate. In this case, the substance is identified as an objective reality in terms of the internal unity of all forms of its manifestation and self-development [35; 39]. The term "substrate" in turn, denotes the simplest structure or formation [26], which remains stable, unchanged under any transformation of the object and determines its specific properties.

Thus, macrostructure of ICSL is defined within this study as comprehensive language body of neologisms in the systemic of reference semantic unity in correlation with substantive (ontological, epistemic, anthropological) measurements and computer being elements of comprehensive, innovative super-dense verbalization which determines the phenomenological originality of logosphere.

The dynamics of innovative cyberspace logosphere is defined as ways, directions and appropriate language implementation mechanisms of qualitative changes in the content area of the projection of the conceptual nucleus of the referred innovative logosphere.

The structure of the content of the innovative cyber-term, as a constituent of ICSL, is distributed across the several tiers of abstraction, consistent with the through-vertical ratio of philosophical categories of "essence" → "phenomenon": 1) - onthological denotatum (OD) - a set of meaningful elements of exhaustive degree of substance and epistemic abstraction (phenomenalogization attributes, parameters and properties of elements multi-substrat cyberspace) in the structure of the meaning of innovative cyber-term → 2) - conceptual denotatum (CD) - a set of meaningful elements of median level of abstraction, mediated by anthropogenic (subjective and collective) cognitive experience of speakers in the area of operation and use of computer technology, the projection area of conceptual ICSL nucleus → 3) - lingual denotatum (LD) - semantics of innovative cyber-term.

The degree of abstraction of these tiers structure is correlated with the degree of abstraction of cyberspace parametric features. Tier (1) "ontological denotatum" corresponds to the parametric feature "existential dimension", tier (2) "conceptual denotatum" - parametric feature "concept" and the parametric feature "notion", stage (3) "lingual denotatum" - parametric feature of a "language unit."

The highest index of representativeness within content the innovative cyberspace logosphere is determined to be the following ontological elements combination: |SUBSTANCE TYPE: CYBERSPACE|, |SUBSTANTIVE QUALITY: TECHNOGENESIS|, |SUBSTANTIVE DURATION: SPACE|, |SUBSTANT AFFILIATION: OBJECT OF CYBERSPACE|, |SUBSTANT AFFILIATION: SUBJECT OF CYBERSPACE|, |SUBSTANT AFFILIATION: SIMULACRUM OF CYBERSPACE|, |CYBERMORPHISM|.

AI-enhanced Data Encoding for LLMs Training in Digital Communication

The Cyber-speak is an ongoing electronic, multimodal lexicographic project that is based on the study of the late XX – current XXI century European and Asian languages vocabulary integral dynamics within the emergent digital technology framework.

A methodology of computer vocabulary innovative elements phenomenological features identification is introduced supplying the template for a new study field – phenomenological, AI-enhanced digital neology, neography and neosemiotics. Innovative Cyber-, AI And ICT Thesaurus is designed to define and categorize the key components of innovative cyber-vocabulary, instrumental to electronic communication environment construction and constitution [24].

All units, listed and clustered in the Thesaurus, are supplied with dominant and recessive phenomenological and conceptual markers or a combination thereof, indicative of unit allegiance to the corresponding ontological categories of cybercommunicative environment.

An inventory of innovative European computer logosphere microstructure constituents – EICT – static and dynamic qualities, featured through successive content levels, is shortlisted and used as lexicographic markers for the innovative cyber vocabulary semantic structures of different tiers.

The EICT static and dynamic qualities provide for the volume, boundaries and content of innovative European computer logosphere micro- and macro-dynamics assessment. The guidelines of innovative European computer logosphere both internal and external microstructure dynamic mobility are delineated by the structural and content patterns, inherent to this linguistic body.

Artificial intelligence, particularly machine learning and symbolic AI, has transformed corpus annotation through automated lexical extraction and multimodal encoding. The EU AI Act's classification of education as a high-risk domain underscores the necessity of robust AI governance models to ensure ethical corpus encoding practices.

The Cyber-Speak lexicographic database incorporates AI-augmented XML encoding to streamline neology processing. By training AI models on custom corpus data and phenomenological categorization, this approach enables automated extraction, tagging, and classification of innovative vocabulary within random and predefined corpora.

The foundational encoding model employs the following structure:

```
<entry lxnm:entryID='1243' xmlns:TYPE='TYPE' xmlns:QUALITY='QUALITY' xmlns:DURATION='DURATION' xmlns:lxnm='http://www.lexonomy.eu/'>
  <headword xml:space='preserve'>AI-pocalypse</headword>
  <partOfSpeech xml:space='preserve'>n</partOfSpeech>
  <sense taxon='|SUBSTANT TYPE: COMPUTER BEING|' hyperelement='|SUBSTANT AFILIATION|' hypoelement='|SUBSTANT QUALITY: ESCHATOLOGY|' xml:space='preserve'>|SUBSTANT DURATION: TIME/SPACE|</sense>
  <definition xml:space='preserve'>artificial intelligence + apocalypse</definition>
  <definition xml:space='preserve'>a disaster caused by an advanced artificial intelligence</definition>
</entry>
```

This approach enables NLP systems to autonomously recognize, extract, and tag neological formations across diverse transdiciplinary corpora, refining machine learning pipelines for real-time language evolution tracking**.**

This method provides automatic semantic categorization**,** improving NLP capabilities **in** predictive language modeling, sentiment analysis, and lexicographic data retrieval**.**

The integration of multimodal AI annotation expands corpus tagging capabilities by encoding syntactic-semantic matrices with dynamic phenomenological markers.

The BNC tagging model employs an interlinked matrix structure to encode phrasal verbs representing both synonymous and antonymic emotional expressions, categorizing them based on semantic intensity and sentiment polarity. On the other hand, tagging matrices reveal transdisciplinary fields, demonstrating how emotions such as Disgust, Anger, and Fear manifest through phrasal encoding (e.g., *shudder at*, *burn up*, *freak out*). These cognitive-tag submatrices allow AI-assisted predictive modeling, refining sentiment classification in multi-disciplinary linguistic corpora (legal, tech, educational, ecological etc.).

Example of corpus tagging integration:

```
<option value="shudder at">shudder at/BNC - Disgust emotion tagging</option>
<option value="burn up">burn up/BNC - Anger emotion tagging</option>
<option value="freak out">freak out/BNC - Fear emotion tagging</option>
```

By applying the structured annotation protocols of the BNC to global datasets, it is possible to: enhance sentiment tagging in multilingual corpora, refining AI-driven sentiment classification; establish cross-linguistic tagging models, ensuring NLP pipelines recognize universal emotion markers across languages; expand semantic tagging to domain-specific corpora**,** improving AI understanding in areas such as medical linguistics, legal discourse, and computational terminology**.**

When integrated into multilingual lexicons, BNC-style tagging enables cross-linguistic NLP models to recognize emotional verb structures in varied language environments, improving translation accuracy and cross-cultural sentiment detection.

Example of multilingual adaptation:

```
<option value="schaudern vor">schaudern vor/BNC adaptation - German corpus tagging for Disgust</option>
<option value="se fâcher">se fâcher/BNC adaptation - French corpus tagging for Anger</option>
<option value="驚く">驚く/BNC adaptation - Japanese corpus tagging for Amazement</option>
```

The tag-based markup of emotional states within corpora provides new avenues for domain interoperability, refining context-sensitive AI language processing.

As AI progressively refines lexicographic structuring, traditional disciplinary boundaries dissolve, leading to a post-disciplinary lexicon characterized by fluid, adaptive linguistic systems. The interaction between machine learning, symbolic AI, and corpus tagging fosters a new epistemological framework for AI-driven linguistic analytics. The post-human transformation of linguistic cognition demands a reconceptualization of knowledge transfer,

incorporating non-human agents into collaborative AI-mediated lexicons.

The matrix encoding model integrates semantic fracturing, defining linguistic innovation across virtuality, networked communication, and cyber-morphism. This ontological reconfiguration ensures adaptive AI models capable of multi-tiered lexical processing. By employing AI-enhanced matrix profiling, digital linguistics can automate lexical interpretation, ensuring greater semantic accuracy across multilingual corpora. By linking machine-readable lexicons with structured corpus tagging, LLMs gain the capability to:

- Predict semantic shifts based on usage trends.
- Improve contextual NLP modeling, refining sentiment algorithms.
- Enable real-time corpus annotation, ensuring LLM adaptability to linguistic transformations.

The protocol streamlined the digital lexicographic workflow and provided grounds for multi-faceted enhancement of AI-powered communication skills of the FLE graduate students focus-group. To improve AI literacy, structured annotation models must:

- Enhance semantic tagging accuracy for LLM lexicons.
- Enable multimodal NLP applications, refining voice-processing AI and conversational agents.
- Foster cross-disciplinary AI research, ensuring linguistic encoding methodologies benefit law, healthcare, education, and computational semantics.

3. CONCLUSIONS

The study findings make it possible to distinguish the following substantial characteristics of the linguistic sphere of cyberspace: 1) ability to synthesize features of ontological objects and phenomena and innovative verbal units, respectively; 2) ability for asymptotic (extremely close to exhaustive) embodiment of substantive and factual elements of cyberspace at the level of semiotic substance as a whole, and at the level of substantial characteristics of discrete verbal units; 3) semiotic density of embodiment of substantive and factual elements of cyberspace in the ontological, epistemological and anthropological planes.

Philosophical foundations of the study of innovative cyberspace logosphere (ICSL) as an integrated macro-and micro-entity are determined:

1) by the substantive definite features of logosphere as a macrostructure (including equifinality - ability to achieve states that do not depend on the initial conditions and specific parameters that are specific to innovative cyberspace logosphere; teleology - gnostic ability of innovative cyberspace logosphere to achieve the projected state);

2) by the phenomenological characteristics and properties of substrate microstructure of linguistic units of innovative cyberspace logosphere.

Qualitative and quantitative characteristics, features and properties of the integrative structure of the innovative cyberspace logosphere are informed by its dynamics on the macro and micro levels.

Transdisciplinary educational applications of these approaches to data encoding, subsequently, include: training AI-enhanced NLP models for transdisciplinary communication; developing standardized linguistic annotation protocols, ensuring interoperability across AI-driven lexicographic systems; integrating transdisciplinary discourse structures into machine-learning lexicons, refining AI adaptive language comprehension. These methodologies facilitate transdisciplinary AI protocols, refining computational lexicography, multimodal corpus analysis, and semantic-syntactic interoperability.

The fusion of AI-driven encoding with matrix corpus tagging enables higher-resolution semantic profiling**,** improving machine learning workflows for NLP**.** By synthesizing computational lexicographic methods**,** AI-enhanced tagging models can dynamically adjust to language evolution**,** ensuring scalable annotation across corpora and datasets.

4. ACKNOWLEDGEMENT

The paper has been reviewed by Nataliia Lazebna, Habilitated Doctor, Würzburg University, Germany, Ukraine. Empirical findings and theoretical procedures have been conducted under the auspices of Integrated Research framework of Romance and Germanic Philology Department of Borys Grinchenko Kyiv Metropolitan University Digital Transformative Linguistics and Cross-Cultural Communication (0123U102796), project COST Action CA21167 UniDive: Universality, Diversity and Idiosyncrasy in Language Technology, COST Action CA22126 ENEOLI: European Network on Lexical Innovation, COST Action CA23105 PLURILINGMEDIA: Language Plurality in Europe's Changing Media Sphere, COST Action CA22115 PhraConRep A Multilingual Repository of Phraseme Constructions in Central and Eastern European Languages. The authors extend special acknowledgement to the Armed Forces of Ukraine for providing safety to complete this work.

5. REFERENCES

[1] Aristotle, "Categories". **The Complete Works of Aristotle,** Princeton: Princeton University Press, 2014.

[2] Bakhtin M. **Aesthetics of verbal creativity**, M.: Art, 1979.

[3] Barthes R. **Elements of Semiology**, Hill and Wang, 1968.

[4] Callaos N., Marlowe T., "Inter-Disciplinary Communication Rigor". **Rigor and Inter-Disciplinary Communication: Intellectual Perspectives from Different Disciplinary and Inter-Disciplinary Fields**. TIDC, LLC, 2020, pp. 4-29.

[5] **Cambridge Dictionary,** CUP, Retrieved from: https://dictionary.cambridge.org, 2020.

[6] Chanyshev A. "Treaties on Non-Being", **Issues of Philosophy**, v.10, p. 160.

[7] Crystal D. **Language and the Internet.** Cambridge: CUP, 2001.

[8] Davis E. **Techgnosis: Myth, Magic and Mysticism in the Age of Information**. NY: New York Publishers, Inc., 2001.

[9] Florensky P. "Namehail as a philosophical proposition. On the name of God", **Studia Slavica Hung,** Budapest**,** Vol. 34/1–4, 1988. pp. 40–75.

[10] Fromm E. **The Forgotten Language: An Introduction to the Understanding of Dreams, Fairy Tales, and Myths.** Open Road Media, 2013**.**

[11] Gachev G. "Humanistic commentary to natural science", **Issues of Literature**, Issue 11, 1993, pp. 71–78.

[12] Gelernter D. **Virtual Realism**. Oxford: Oxford University Press, 1998.

[13] Kasavin I. "Perceiving the multitude of the Mind?", **Multiplicity of Academic knowledge**, Moscow, 1990, pp. 67–79.

[14] Kranz W. (ed.), **Die Fragmente der Vorsokratiker**, Zürich: Weidmann,1996.

[15] Kireev G. "O kartine mira. Kosmologicheskoe esse**"**, **Lybid**. v. 534, 2008, pp. 25–29.

[16] Heim M., **The Metaphysics of Virtual Reality.** LA: Westport Publishers, 1993.
[17] Hillis K. **Digital Sensations: Space, Identity, and Embodiment in Virtual Reality**. UM: University of Minnesota Press, 1999.
[18] Khoruzhy S. "Notes on Ontology of Virtuality". **Issues of Philosophy**, Vol. 6, 1997, pp. 53–58.
[19] Knight S. "Making authentic cultural and linguistic connections", **Hispania**, Vol. 77, 1994, pp. 289–294.
[20] Lazebna N. **English Language as Mediator of Human-Machine Communication**. Mysore, India: PhDians along with Ambishpere: Academic and Medical Publishers, Royal Book Publishing, 2021.
[21] Losev A. "Philosophy of the Name", **Being. Name. Cosmos**. Thought, 1993, pp. 613–801.
[22] Lotman, Yu. **Semiophere**. Art, 2000.
[23] Makhachashvili R., "Models and Digital Diagnostics Tools for the Innovative Polylingual Logosphere of Computer Being Dynamics", **Italian-Ukrainian Contrastive Studies: Linguistics, Literature, Translation. Monograph**. Peter Lang GmbH Internationaler Verlag der Wissenschaften, Berlin, 2020, pp. 99-124.
[24] Makhachashvili, R. et al. "AI-Enhanced Multilingual Lexicography for Digital Communication", **Proceedings of the 16th International Multi-Conference on Complexity, Informatics and Cybernetics: IMCIC**, 2025, 1. pp. 247-253.
[25] Makhachashvili, R., Semenist, I., "Linguistic philosophy of cyberspace". **Proceedings of the 25th World Multi-Conference on Systemics, Cybernetics and Informatics**, 2021, 2. pp. 24-29.
[26] Mamardashvili M., Pyatihorsky A. **Symbol and Mind. Metaphysical Rumination on the Mind, Symbolysm and Language.** Academia, 1997.
[27] Makhachashvili, Rusudan, "Cyber-speak Dictionary (ELEXIS)", **Slovenian language resource repository CLARIN.SI**, ISSN 2820-4042, 2020, http://hdl.handle.net/11356/1610
[28] Oke N. "Globalizing Time and Space: Temporal and Spacial Considerations in Discourses of Globalization", **International Political Sociology**, v. 3, 2009, pp. 310–326.
[29] Pigalev A. "Cultural Space". **Culturology of the 20th Century**, Libris, 1998, pp. 337–344.
[30] Pierce C.S. **Architectonics of Philosophy**. Philosophic Society, 2001.
[31] Price G. "Myths for Today, Hopes for Tomorrow", **Searcher**, v.1, 2000, pp. 3–5.
[32] Quinion M. "How words enter the language", **Information Concepts**, NC: NCU Press, 2003, pp. 41–43.
[33] Rheingold H. **The Virtual Community**. LA: California University Press, 1999.
[34] Ricoeur P. **Hermeneutics and the Human Sciences. Essays on language, action and interpretation**. Cambridge: University Press, 1981.
[35] Roelleke Th. **Information Retrieval Models: Foundations and Relationships**. Jefferson: UNC, 2013.
[36] Rosch E. "Principles of Categorization", **Cognition and categorization**. Hillsdale: Lawrence Erlbaum Ass., 1978.
[37] Semerikov, S., Mintii, I., Makhachashvili, R., "Digital Humanities Event Horizon", **Digital Humanities Workshop.** ACM International Conference Proceeding Series, 2021, pp. 1-28.
[38] Smolin L. **Scientific alternatives to the anthropic principle**. London: Weidenfeld & Nicolson, 2004.
[39] Spet G. **Phenomenon and Meaning (Phenomenology as a Science)**. Moscow: Academia, 2001.
[40] Schrijver L. "Designing the Networked Environments: Architectural Visions of Cybercommunities", **Posthumanity: Merger and Embodiment**, Oxford, UK: Inter-Disciplinary Press, 2010, pp. 110–114.
[41] Tapscott D. **Grown Up Digital: How Net Generation is Changing the World.** McGraw-Hill Education, 2008.
[42] Talmy L. "Force dynamics in language and cognition", **Concept Structuring Systems**, The MIT Press, v. 1, 2000, pp. 409–470.
[43] Vernadsky V. **Scientific thought as a planetary phenomenon.** Kyiv, 1991.