http://dx.doi.org/10.18623/rvd.v23.5126

ARTIFICIAL INTELLIGENCE IN LINGUISTICS: MODELING UNIVERSAL PHONOLOGICAL SYSTEMS FOR SUSTAINABLE COMMUNICATION

INTELIGÊNCIA ARTIFICIAL NA LINGUÍSTICA: MODELAGEM DE SISTEMAS FONOLÓGICOS UNIVERSAL PARA COMUNICAÇÃO SUSTENTÁVEL

Article received on: 11/3/2025
Article accepted on: 2/2/2026

Larisa Micallef*
*Financial University under the Government of the Russian Federation, Moscow, Russia
Orcid: https://orcid.org/0000-0001-9651-4353
larisa.olegovna@gmail.com

The authors declare that there is no conflict of interest

Abstract
Objective: this study examines how artificial intelligence (AI) can be combined with empirical linguistic data to develop models of universal phonological and orthographic systems. The broader aim is to contribute to more sustainable and inclusive tools for cross-linguistic communication. The work focuses on a central challenge in contemporary linguistics: the lack of reproducible AI-driven methods that link computational modeling with theoretical analysis and that ensure fair and accessible use of digital language technologies. Method: A mixed-method framework has been adopted, in which corpus-driven linguistic analysis has been integrated with neural-network modeling. The empirical data have been drawn from two open-access resources: PHOIBLE (Phonetics Information Base and Lexicon) and the r12a database (r12a.github.io). After standardization and tokenization, the datasets have been processed using Python-based AI modules to extract frequency distributions, identify clusters and detect structural patterns. The analytical workflow has followed a clear, reproducible sequence of steps informed by PRISMA principles, ensuring transparency and methodological rigor. Originality/Relevance: the paper brings together corpus linguistics, interlinguistics and artificial intelligence to propose a data-driven approach for identifying shared phonological and orthographic patterns across languages. By combining extensive linguistic datasets with computational techniques, the study demonstrates the potential of AI to support the creation of sustainable knowledge infrastructures and to promote more inclusive forms of digital communication — domains that are becoming central to innovation and strategic growth in the humanities. Main conclusions: the analysis revealed a relatively

Resumo
Objetivo: este estudo examina como a inteligência artificial (IA) pode ser combinada com dados linguísticos empíricos para desenvolver modelos de sistemas fonológicos e ortográficos universais. O objetivo mais amplo é contribuir para ferramentas mais sustentáveis e inclusivas para a comunicação interlinguística. O trabalho se concentra em um desafio central da linguística contemporânea: a falta de métodos reproduzíveis baseados em IA que liguem a modelagem computacional à análise teórica e que garantam o uso justo e acessível das tecnologias digitais de linguagem. Método: Foi adotada uma estrutura de método misto, na qual a análise linguística baseada em corpus foi integrada à modelagem de redes neurais. Os dados empíricos foram extraídos de dois recursos de acesso aberto: PHOIBLE (Phonetics Information Base and Lexicon) e o banco de dados r12a (r12a.github.io). Após padronização e tokenização, os conjuntos de dados foram processados usando módulos de IA baseados em Python para extrair distribuições de frequência, identificar clusters e detectar padrões estruturais. O fluxo de trabalho analítico seguiu uma sequência clara e reproduzível de etapas informadas pelos princípios PRISMA, garantindo transparência e rigor metodológico. Originalidade/Relevância: o artigo reúne linguística de corpus, interlinguística e inteligência artificial para propor uma abordagem baseada em dados para identificar padrões fonológicos e ortográficos comuns entre as línguas. Ao combinar extensos conjuntos de dados linguísticos com técnicas computacionais, o estudo demonstra o potencial da IA para apoiar a criação de infraestruturas de conhecimento sustentáveis e promover formas mais inclusivas de comunicação digital — domínios que estão se tornando centrais para

small set of phonemes and grapheme correspondences that recur across a wide range of the world's languages. These results offer empirical support for developing streamlined, accessible alphabetic systems and for designing universal auxiliary language models. The study further shows that AI-supported modeling can improve linguistic inclusivity and analytical precision, especially in low-resource and multilingual settings, while still relying on the interpretive judgement of human specialists. Theoretical/methodological contributions: the research contributes to interlinguistics by bringing together the concept of language universals and contemporary AI techniques. It outlines a reproducible pathway for connecting empirical linguistic data with computational tools and theoretical interpretation. In doing so, the study supports the sustainable development of language technologies and enriches our understanding of how human expertise and artificial intelligence can work together to strengthen global communication. Practical implications: identifying a universal phoneme core and stable sound-script correspondences can streamline multilingual analytical workflows, lessen structural biases toward non-Latin scripts and lower the overall costs of integrating low-resource languages into sustainable and reproducible knowledge systems.

Keywords: Linguistic Typology. Interlinguistics. Universal Phonology. Orthography. Multilingual Fairness. NLP.

a inovação e o crescimento estratégico nas ciências humanas. Principais conclusões: a análise revelou um conjunto relativamente pequeno de correspondências entre fonemas e grafemas que se repetem em uma ampla variedade de línguas do mundo. Esses resultados oferecem suporte empírico para o desenvolvimento de sistemas alfabéticos simplificados e acessíveis e para a concepção de modelos de línguas auxiliares universais. O estudo mostra ainda que a modelagem apoiada por IA pode melhorar a inclusão linguística e a precisão analítica, especialmente em ambientes com poucos recursos e multilíngues, sem deixar de contar com o julgamento interpretativo de especialistas humanos. Contribuições teóricas/metodológicas: a pesquisa contribui para a interlinguística ao reunir o conceito de universais linguísticos e técnicas contemporâneas de IA. Ela descreve um caminho reproduzível para conectar dados linguísticos empíricos com ferramentas computacionais e interpretação teórica. Ao fazer isso, o estudo apoia o desenvolvimento sustentável de tecnologias linguísticas e enriquece nossa compreensão de como a expertise humana e a inteligência artificial podem trabalhar juntas para fortalecer a comunicação global. Implicações práticas: identificar um núcleo fonêmico universal e correspondências estáveis entre sons e escritas pode otimizar os fluxos de trabalho analíticos multilíngues, diminuir os vieses estruturais em relação às escritas não latinas e reduzir os custos gerais de integração de línguas com poucos recursos em sistemas de conhecimento sustentáveis e reproduzíveis.

Palavras-chave: *Tipologia Linguística. Interlinguística. Fonologia Universal. Ortografia. Equidade Multilíngue. PLN.*

1 INTRODUCTION

By 2025, the rapid expansion of artificial intelligence (AI) has begun to influence not only technological domains but also the methodological foundations of research in disciplines such as linguistics. A wide range of tasks that once demanded substantial human effort such as processing large datasets, structuring information, translating texts and even producing written material are now carried out routinely by AI systems. This

development enables researchers to redirect their efforts toward interpretation, critical analysis and ethical reflection. With the onset of the modern AI period in the early 2020s, researchers have increasingly turned to hybrid frameworks that integrate AI not merely as a supportive technology but as a substantive collaborator, reshaping long-standing theory-based approaches.

Linguistics, concerned with language, cognition and communication, has felt these changes more sharply than many other disciplines. The field can now be viewed as moving through two distinct stages. The first, rooted in traditional scholarship, depended on close observation, introspection and theory-building grounded in human judgment. The subsequent phase, emerging alongside AI developments, makes use of computational models that reproduce linguistic patterns, propose additional lines of inquiry and contribute to the refinement of established theories. Together, these developments signal the emergence of a research environment in which human inquiry and intelligent systems intersect in new ways within the study of language. Despite major progress in natural language processing (NLP), machine learning (ML) and large language models (LLMs), the field still lacks coherent and reproducible frameworks that link AI methods to linguistic theory. This lack of integration undermines both the interpretability and the long-term viability of AI-driven linguistic research. As Alaqlobi *et al*. (2024) and Groenewald *et al*. (2024) point out, linguistic data are increasingly handled by automated systems, yet the theoretical consequences of this shift have received comparatively little attention. Lammers and Lasch (2023) add that public confidence and scholarly credibility hinge on interpretability – that is, on AI models that clarify rather than obscure linguistic phenomena. Taken together, these observations highlight a central challenge: developing AI-supported linguistic frameworks that deliver technological efficiency while remaining theoretically transparent and sustainable on a global scale. Meeting this challenge is crucial for ensuring the sustainable growth of digital communication and educational infrastructures. Recent studies in the Journal of Sustainable Competitive Intelligence highlight transparent data use, reproducible methodologies and equitable access as core components of durable knowledge systems (Hair & Sabol, 2025; Pussaignolli de Paula *et al.*, 2024). These principles extend directly to linguistic research: any sustainable linguistic framework must draw on open data, ethically grounded AI design and applicability across diverse linguistic and cultural settings. However, many existing

language technologies remain tailored to dominant languages, which perpetuates disparities in access to information. Creating AI tools that accommodate a wider range of tasks that used to require substantial human effort, e.g. handling data or translating texts, are now routinely delegated to AI system.

The present study advances this objective by introducing a methodological framework that links corpus-based linguistic analysis with neural network modeling to explore universal phonological and orthographic structures. Using open-access resources such as PHOIBLE (Phonetics Information Base and Lexicon) and the r12a database, it examines cross-linguistic sound–script correspondences through AI-assisted computational techniques. This combined approach illustrates how linguistic datasets can be converted into sustainable and reproducible models that foster multilingual innovation and inclusive digital communication.

Accordingly, the study is guided by the following central question:

How can AI-supported linguistic models help create sustainable, universally applicable structures for cross-linguistic communication and digital knowledge?

The research is structured around the following goals:

1) to analyze extensive phonological and orthographic data with the support of AI-based modeling;
2) to detect repeated patterns that point to universal sound-script relationships;
3) to explore how these insights can contribute to developing language technologies that are transparent, sustainable and inclusive on a global scale.

The article is organized in the following way. The *Methods* section explains the combined corpus-analytic and neural network methodology. The *Results and Discussion* sections detail the computational outcomes and explore what they mean for developing universal linguistic models. Finally, the *Conclusion* summarizes the broader theoretical and practical insights gained from bringing AI and interlinguistic concepts together to strengthen sustainable communication and digital knowledge systems.

2 METHODOLOGY AND DATA SOURCES

This research employs a mixed-method approach that combines corpus-based linguistic analysis with contemporary artificial intelligence techniques. Through the combined use of computational modeling, neural-network techniques and data-driven phonological analysis, the study examines the ways AI is reshaping contemporary linguistic research and contributing to sustainable knowledge-building in the communication sciences. The methodological framework draws on recent advances in neural-network linguistics, an interdisciplinary area situated at the crossroads of linguistics, cognitive science and computer science (Micallef, 2025). This perspective allows for a closer examination of how AI systems handle natural language processing and generation, and it provides a basis for assessing whether their outputs are linguistically reliable and interpretively sound.

The study examines how transformer-based neural networks, such as GPT models, operate by analyzing their two central phases: pre-training and fine-tuning. In the pre-training stage, the system is exposed to extensive multilingual text collections and acquires statistical regularities of language by predicting missing tokens. Fine-tuning then modifies the model's behavior through supervised human feedback so that its outputs better match contextual and communicative expectations. These architectures depend on several core components – tokenization, vector embeddings and attention mechanisms, which together enable the encoding of semantic relationships and the maintenance of coherence across complex linguistic inputs. Understanding these processes helps clarify how AI systems now perform tasks traditionally reserved for human specialists, including translation, speech recognition and the generation of written text.

Nonetheless, neural networks function as statistical models of language rather than as true agents of meaning. For this reason, human linguistic expertise is still required to interpret, assess and contextualize their outputs.

To provide an empirical foundation for the analysis, the study drew on two openly accessible linguistic databases:

1. PHOIBLE (Phonetics Information Base and Lexicon) is a large open-access resource that compiles phoneme inventories for 2,186 languages worldwide (Moran & McCloy, 2019). Created at the Max Planck Institute for the Science of

Human History, it consolidates information from several major phonological databases, including UPSID, StressTyp2 and PHONETICS. The integration of these sources results in a harmonized dataset containing more than 3,000 phonological systems. Due to its consistent metadata and unified formatting, PHOIBLE is well suited for computational linguistics and machine-learning applications, where standardized input is essential for modeling cross-linguistic phonological variation. Recent assessments of global phoneme datasets highlight how crucial careful harmonization and transparency are for producing reliable typological research. Anderson *et al*. (2023), for instance, compared leading phoneme-inventory collections and showed that methodological alignment, especially in how marginal and allophonic segments are treated, substantially improves the reproducibility of cross-linguistic analyses. Following these insights, the present study uses PHOIBLE as a stable, standardized and machine-readable empirical foundation.

2. r12a.github.io, created by Richard Ishida for the World Wide Web Consortium (W3C), is an interactive online resource that brings together typological information on more than 150 writing systems used worldwide (Ishida, n.d.). The platform details the structural, typographic and Unicode properties of each script, enabling systematic comparison of alphabetic, syllabic and logographic systems within a common analytical framework. In addition to its descriptive role, r12a also serves educational and technological purposes, offering tools that are valuable for AI-driven tasks such as text rendering, character encoding and multilingual processing.

Both resources were chosen for their complementary roles: PHOIBLE documents the structure of sound systems, whereas the Scripts App provides detailed information on their written forms. Taken together, they make it possible to examine sound–script correspondences, a key interface for speech technologies and multilingual NLP.

The datasets were elaborated through a structured sequence of computational and analytical steps in order to maintain transparency, consistency and reproducibility. As a first step, the PHOIBLE and r12a files were filtered and standardized to resolve discrepancies in phoneme notation and script categorization. This procedure yielded a coherent and uniform dataset, providing a reliable foundation for subsequent large-scale

computational analysis. Once the data had been cleaned and converted into a consistent format, the datasets were then analyzed using Python-based tools and AI components designed to identify clusters, extract frequency patterns, and detect structural regularities. Each phoneme was tokenized and assigned to categories based on articulatory traits, frequency levels and its spread across language families and geographic regions. Contemporary surveys in the field point to the need for comprehensive multilingual mappings, advanced deep-learning architectures and rigorously harmonized datasets in grapheme-to-phoneme research (Cheng *et al*., 2024). Drawing on these insights, the present study combines graphemic parsing, phoneme embeddings and attention-based modeling to trace sound–letter relationships across a wide array of writing systems. To make the trends more visible, the study generated statistical summaries and a series of graphs, which brought recurring phonological patterns into focus and pointed to possible connections between sound structures and orthographic practices.

The analysis was carried out using a structured and reproducible workflow modeled on the PRISMA principles of data selection and validation. The process unfolded in four stages: choosing datasets according to their linguistic coverage, cleaning and standardizing the data, identifying and modeling patterns through AI algorithms and interpreting the outcomes in light of established linguistic criteria. This design ensured methodological rigor and provided a reliable empirical foundation for the study.

Table 1

Steps in the AI-supported workflow for linguistic data processing and analysis

Stage	Description	Criteria / Tools	Sources
Identification	Initial search and listing of open-access linguistic resources relevant to phonology and orthography	PHOIBLE database (n = 2186 languages); r12a (n = 150 scripts)	2 primary sources identified; Initial global source of information: Cross-Linguistic Linked Data (clld.org).
Selecting datasets	Among the sources found, PHOIBLE and r12a were selected	These are the most comprehensive and open databases	PHOIBLE и r12a
Cleaning and normalizing the data	The data were standardized according to the IPA (International Phonetic Alphabet). Elements lacking sufficient information for conversion into this	The IPA description standard was chosen as the primary one because it is the most widely used, and ready-made software libraries exist for working with this standard.	IPA on Wikipedia

	format (insufficiently described phonemes) were excluded, as well as languages with insufficient descriptions (the bottom 10% of languages when ranked by the number of phonemes)		
Detecting and modeling patterns with AI algorithms	Data processing and pattern detection were carried out by developing software in the Python programming language in the form of so-called "Python notebooks"	The criteria for determining the frequency of phonetic data are described later in this article. The goal is to identify the most universal phonemes. Tools: Python, AI, and statistical tools and software libraries.	PHOIBLE и r12a
Interpreting the results	Summary of the most frequent phonemes and their typical Latin-based representations identified through AI-assisted analysis of global linguistic data	Data from PHOIBLE and R12a databases analyzed in Python using AI and statistical methods; phonemes ranked by global frequency and cross-linguistic consistency	PHOIBLE и r12a

The integration of AI-driven analytical methods made it possible to detect nuanced relationships among languages that are not readily observable through manual procedures. The combined use of computational modeling and expert linguistic interpretation yielded a more comprehensive account of interactions between phonological and orthographic systems, thereby informing future research in speech technology, multilingual NLP, and digital communication. This study's methodological approach is designed to bring together the rich empirical base of corpus linguistics and the analytical precision offered by AI. By applying neural-network techniques to extensive linguistic datasets, the research moves past conventional descriptive work and presents a more structured, data-oriented way of modeling linguistic patterns and variation. By working with standardized and machine-readable datasets like PHOIBLE and the r12a, the study ensures that cross-linguistic comparisons remain consistent and that the resulting AI models can be reproduced. These datasets supply the empirical grounding needed to analyze variation in sound and writing systems and to explore the universal patterns that shape language.

Bringing together computational accuracy with linguistic insight demonstrates that AI can serve not only as an analytical instrument but also as a way of viewing linguistic problems. This approach makes it possible to construct scalable models capable of tackling long-standing issues in phonology, typology, and multilingual communication. On a wider level, it shows how data-oriented linguistics can help build sustainable knowledge systems, connecting technological development with educational and communicative progress in a rapidly digitalizing world.

3 RESULTS AND DISCUSSION

This study demonstrates that combining linguistic databases such as PHOIBLE and the r12a with AI-assisted analysis strengthens research in typology and in modeling universal phonological and orthographic systems. Using these resources, a custom analytical tool was developed to examine languages from two complementary angles, sound systems and writing systems, allowing consistent cross-linguistic comparison and supporting the reconstruction of linguistic universals, including the idea of a potential universal auxiliary language. The tool classifies similar phonemes and identifies their common graphemic representations across languages. By processing large sets of phonological and orthographic data, it isolates the most frequent sounds and their typical letter forms in different scripts. This frequency-driven method reflects a basic linguistic principle: efficient communication tends to rely on the most widespread and easily recognized elements. The results make it possible to outline a minimal, functionally sufficient set of phonemes and graphemes that could form the basis for an international auxiliary language (Micallef & Yasnenko, 2024).

The idea of a universal language has a long intellectual history, but it became a subject of systematic linguistic study only in the late nineteenth century. Much of the interest arose from the difficulty and inconsistency of widely used international languages such as English, French and Spanish, which create learning obstacles for non-native speakers. In response, linguists and language reformers proposed a series of such planned languages as Esperanto, Interlingua and Ido, intended to simplify global communication through regular grammar and transparent word-formation. Foundational ideas about linguistic universals and the structure of constructed languages were shaped by the work

of Baudouin de Courtenay (1963) viewed artificial languages as deliberately engineered systems, connecting their structure to psychological aspects of speech. Jespersen (1928) approached the topic from another angle, focusing on economy and simplicity as engines of linguistic efficiency. Meillet (1918) added a more social perspective, emphasizing the role of historical and cultural context. Building on functionalist theory, Martinet (1967) formulated the concepts of double articulation and linguistic economy, demonstrating how languages streamline structure for clarity and effectiveness. All the represented ideas continue to influence contemporary work on the design of constructed languages.

While early twentieth-century linguists established the conceptual basis for studying universal and constructed languages, recent computational work has extended these ideas through quantitative and neural methods. Large-scale data analysis and AI modeling now allow researchers to test hypotheses about universals that were previously supported only by theoretical argument. For instance, Doucette *et al*. (2024) apply Bayesian regression to phonotactic data from 107 Northern Eurasian languages and uncover a consistent pattern: consonant similarity is generally avoided (consistent with the Obligatory Contour Principle), whereas vowel similarity often promotes harmony rather than avoidance. Their results show a clear asymmetry between consonants and vowels, with consonant systems more tightly constrained by similarity effects. This type of probabilistic modeling offers broad empirical backing for universality claims that have long been central to phonological theory and that also inform the simplification strategies used in planned and artificial languages. Complementing such typological studies, Wu, Ponti and Cotterell (2021) propose a Differentiable Generative Phonology framework that reinterprets the rule-based architecture of The Sound Pattern of English as an end-to-end neural model. Their system learns both underlying and surface representations as continuous vectors, effectively connecting theoretical phonology with computational learning. This approach demonstrates that deep-learning models can recover key linguistic generalizations, such as the need for abstract underlying forms and the separation of phonological levels, without relying on manually written rules.

Although these earlier studies remain important for establishing the theoretical foundations of interlinguistics, they represent only the initial phase of the field's development. Work published in recent years (2020–2025) has expanded these ideas through AI-based language modeling and digital typology, examining universal

phonological systems, sound–symbol relationships, and large cross-linguistic datasets such as PHOIBLE and the r12a. This newer research shows how classical linguistic principles can be tested, refined and sometimes reformulated within neural and probabilistic frameworks, effectively connecting traditional interlinguistic theory with contemporary AI methods.

A major theoretical contribution in this area was made by the Swiss linguist René de Saussure (1918). Drawing on his analysis of Esperanto as well as natural languages like French and German, he proposed core principles for designing an international auxiliary language. Central to his approach was the notion of phonological writing: a strict one-to-one relationship between sounds and letters to maximize clarity and ease of acquisition. In *La structure logique des mots dans les langues naturelles, considérée au point de vue de son application aux langues artificielles* (1918), he advocated simplifying the Latin alphabet by eliminating ambiguous graphemes such as *x*, *q* and *ph* and replacing them with the more transparent forms *ks*, *k* and *f*. He intended to create a writing system that was both practical and widely intelligible, preserving familiar elements while improving accessibility for learners.

Extending Saussure's insights, this study examines a much larger cross-linguistic dataset using PHOIBLE and the r12a. The AI-based analysis identifies a compact core of phonemes that recur in most languages worldwide. It also finds that Latin-script writing systems tend to map these shared sounds onto similar graphemic shapes. These patterns suggest that it is feasible to design a simplified and widely comprehensible alphabet grounded in common phonological and orthographic features. These observations are consistent with recent PHOIBLE-based statistical analyses confirming that a small number of consonantal and vocalic categories recur across genetically and geographically diverse languages. Yang (2025) demonstrated that frequency distributions of phonemes tend to follow highly stable probabilistic patterns, which align closely with the core inventory identified in the present study. This convergence underscores the empirical robustness of frequency-driven models for identifying universal phonological tendencies.

Table 2 below summarizes the most frequent phonemes and their typical Latin-based representations, providing empirical support for the development of future universal auxiliary language models.

Table 2

Frequent sounds and common Latin-based representations in world languages

IPA Symbol	Type of Sound	Common Latin Letter	Frequency in World Languages	Suitable for Universal Alphabet
/a/	Vowel	a	Found in over 90%	Recommended
/i/	Vowel	i	Found in over 90%	Recommended
/u/	Vowel	u	Appears in approximately 85%	Recommended
/p/	Consonant	p	Present in about 80%	Recommended
/t/	Consonant	t	Present in over 90%	Recommended
/k/	Consonant	k	Present in about 85%	Recommended
/m/	Consonant	m	Present in about 85%	Recommended
/n/	Consonant	n	Found in over 90%	Recommended
/s/	Consonant	s	Appears in approximately 85%	Recommended

The findings indicate that the phonological system of a universal auxiliary language could be based on a compact and stable inventory of highly frequent sounds that are already familiar to most speakers. These phonemes are usually written with letters that are widely recognizable across many scripts, making such an alphabet relatively intuitive and easy to learn. This observation aligns with Saussure's insistence on clear sound-letter mapping and reinforces his argument that universal linguistic patterns can serve as a foundation for fair and accessible communication. By bringing together traditional linguistic insights and modern AI-based analysis, this study shows how universal phonological patterns can be both empirically detected and computationally modeled. This integration strengthens interlinguistics as a theoretical and applied field, offering new tools for approaching global multilingualism through inclusive and culturally attuned language technologies.

4 CONCLUSIONS

The research given demonstrates that combining linguistic data analysis with artificial intelligence provides an effective framework for examining language structure and linguistic universals. Combining classical linguistic methods with computational techniques, the study shows that AI can enhance the reach and precision of linguistic inquiry without replacing the need for human analysis and theoretical reflection, which is still necessary. Drawing on the PHOIBLE and r12a datasets, the study identifies recurring

phonological and orthographic patterns found across a wide range of languages. The results offer empirical grounding for designing simplified, inclusive writing systems and informing the development of universal auxiliary language models. Applied within AI tools, such findings can enhance linguistic coverage and accuracy, especially for multilingual and low-resource language contexts.

The outcomes also reaffirm the value of interlinguistics as a field that connects linguistic theory with technological practice. The combined use of data-driven methods and linguistic principles supports the creation of systems that promote mutual understanding and more equitable global communication. Overall, the research suggests a direction for future language research, which is empirical, interdisciplinary and increasingly supported by intelligent technologies.

REFERENCES

Alaqlobi, O., Alduais, A., Qasem, F., & Alasmari, M. (2024). Artificial intelligence in applied linguistics: A content analysis and future prospects. *Cogent Arts & Humanities, 11*(1), 2382422. https://doi.org/10.1080/23311983.2024.2382422

Anderson, C., Tresoldi, T., Greenhill, S. J., Forkel, R., Gray, R., & List, J.-M. (2023). Variation in phoneme inventories: Quantifying the problem and improving comparability. *Journal of Language Evolution, 8*(2), 149-168. https://doi.org/10.1093/jole/lzad011

Baudouin de Courtenay, I. A. (1963). Vspomogatel'nyi mezhdunarodnyi yazyk [International auxiliary language]. In I. A. Baudouin de Courtenay, *Izbrannye trudy po obshchemu yazykoznaniyu v 2 tomakh* [Selected works on general linguistics in 2 volumes] (Vol. 2, pp. 144-160). Moscow: Izd-vo AN SSSR. (In Russian)

Cheng, S., Zhu, P., Liu, J., & Wang, Z. (2024). A survey of grapheme-to-phoneme conversion methods. *Applied Sciences, 14*(24), 11790. https://doi.org/10.3390/app142411790

Doucette, A., O'Donnell, T. J., Sonderegger, M., & Goad, H. (2024). Investigating the universality of consonant and vowel co-occurrence restrictions. *Glossa: A Journal of General Linguistics, 9*(1), 1-39. https://doi.org/10.16995/glossa.9373

Groenewald, E. S., Pallavi, P., Rani, S., Singla, P., Howard, E. M., & Groenewald, C. A. (2024). Artificial intelligence in linguistics research: Applications in language acquisition and analysis. *Naturalista Campano, 28*(1), 1253-1262. https://www.researchgate.net/publication/379239839

Hair, J. F., & Sabol, M. (2024). *L*everaging artificial intelligence (AI) in competitive intelligence (CI) research. *Journal of Sustainable Competitive Intelligence, 15*(00), e0469. https://doi.org/10.24883/eagleSustainable.v15i.469

Ishida, R. (Ed.). (n.d.). *r12a Scripts & Writing Systems App*. World Wide Web Consortium (W3C). Available at: https://r12a.github.io/scripts/switch.html

Jespersen, O. (1928). *An international language*. London: Allen and Unwin.

Lammers, S., & Lasch, A. (2023). Linguistic framing of artificial intelligence: What language to use when talking about artificial intelligence. *Chemie Ingenieur Technik, 95*(7), 1012-1017. https://doi.org/10.1002/cite.202200226

Martinet, A. (1967). *Les langues dans le monde de demain*. Paris: Presses Universitaires de France.

Meillet, A. (1918). *Les langues dans l'Europe nouvelle*. Paris: Payot.

Micallef, L. O. (2025). Lingvistika neyrosetey kak paradigma sovremennoy nauki o yazyke [Neural network linguistics as a paradigm of modern language science]. *World of Science, Culture, and Education, 1*(110), 467-473. https://doi.org/10.24412/1991-5497-2025-1110-467-469

Micallef, L. O., & Yasnenko, I. P. (2024). Principles of international auxiliary languages creation on the base of essential and artificial languages. *Macrosociolinguistics and Minority Languages*, *2*(1), 50-65. https://doi.org/10.22363/2949-5997-2024-2-1-50-65

Moran, S., & McCloy, D. (Eds.). (2019). *PHOIBLE: Phonetics Information Base and Lexicon*. Jena: Max Planck Institute for the Science of Human History. Available at: https://phoible.org

Pussaignolli de Paula, M., Noronha, M., Garcia Valente, U., Inacio Domingues, B. R., & Jahn Souza, L. (2024). Mapping of artificial intelligence and robotics technologies applied to offshore wind Energy. *Journal of Sustainable Competitive Intelligence, 15*(00), e0474. https://doi.org/10.24883/eagleSustainable.v15i.474

Saussure, R. de. (1918). *La structure logique des mots dans les langues naturelles, considérée au point de vue de son application aux langues artificielles*. Berne: Büchler.

Wu, S., Ponti, E. M., & Cotterell, R. (2021). Differentiable generative phonology [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2102.05717

Yang, B. (2025). Frequency distributions and phoneme associations in PHOIBLE. *Proceedings of Speech Sciences, 17*(3), 23-37.

Authors' Contribution

All authors contributed equally to the development of this article.

Data availability

All datasets relevant to this study's findings are fully available within the article.

How to cite this article (APA)

Micallef, L. (2026). ARTIFICIAL INTELLIGENCE IN LINGUISTICS: MODELING UNIVERSAL PHONOLOGICAL SYSTEMS FOR SUSTAINABLE COMMUNICATION. *Veredas Do Direito*, *23*, e235126. https://doi.org/10.18623/rvd.v23.5126