УДК 81’32
DOI 10.25587/2310-5453-2025-2-67-74

Original article

A diffusion-based model of language learning and interlingual distance

***Aleksandr V. Grigorev*[1]✉, *Zhenwei Guo*[2]**
[1]North-Eastern Federal University, Yakutsk, Russian Federation
[2]Liaocheng University, Liaocheng, P.R. China
✉ re5itsme@gmail.com

Abstract
Understanding the process of language learning and quantifying interlingual relationships are central challenges in linguistics, cognitive science, and language education. In this paper, we propose a novel framework that models second language acquisition as a diffusion process within a structured, multidimensional space of languages. We introduce a formal measure of interlingual distance, grounded in linguistic features, to quantify structural and functional differences between languages. Building on Barenblatt-type nonlinear diffusion models, we represent language learning as a multicontinua diffusion process, where distinct components of language – such as phonetics, grammar, vocabulary, and pragmatics – are treated as separate, interacting continua. Each continuum evolves independently according to its own diffusion dynamics, capturing the heterogeneous difficulty and pace of learning across linguistic subsystems. The interaction between these continua reflects the coupling between linguistic competencies in real-world acquisition. We can validate this model with empirical data on second language learning rates across various language pairs, demonstrating that diffusion distances in each continuum correlate with observed learning difficulties in the corresponding language domain. This approach not only offers a new theoretical lens on language learning but also provides a predictive framework for curriculum design, learner modeling, and applications in multilingual NLP and AI systems.

Keywords: second language acquisition, language distance, multicontinua diffusion, neural language embeddings, anisotropic diffusion modeling, finite element method.
For citation: Grigorev A.V., Guo Zh. A Diffusion-Based Model of Language Learning and Interlingual Distance. *Arctic XXI Century*. 2025, No 2. P. 67-74. DOI: 10.25587/2310-5453-2025-2-67-74

Оригинальная статья

Модель изучения языка, основанная на диффузии, и межъязыковая дистанция

А. В. Григорьев[1]✉, ***Ч. Го***[2]
[1]Северо-Восточный федеральный университет им. М.К. Аммосова, Якутск, Российская Федерация
[2]Университет Ляочэн, Ляочэн, КНР
✉ re5itsme@gmail.com

Аннотация
Понимание процесса изучения языка и количественная оценка межъязыковых взаимосвязей представляют собой ключевые задачи лингвистики, когнитивной науки и языковой педагогики. В данной работе предлагается новая концептуальная модель, описывающая освоение второго языка как процесс диффузии в структурированном многомерном языковом пространстве. Мы вводим формальную метрику межъязыковой дистанции, основанную на лингвистических признаках, для количественного измерения структурных и функциональных различий между языками. Опираясь на нелинейные модели диффузии Баренблатта, мы концептуализируем процесс изучения языка как многоконтинуальную диффузию, где различные языковые компоненты (фонетика, грамматика, лексика и прагматика) рассматриваются в качестве взаимодействующих, но самостоятельных континуумов. Каждый континуум эволюционирует согласно собственной динамике диффузии, что позволяет отразить вариативную сложность и скорость освоения различных языковых подсистем. Взаимодействие между континуумами моделирует взаимовлияние языковых компетенций в реальном процессе обучения. Предложенная модель верифицируется на эмпирических данных о скорости освоения второго языка для различных языковых пар. Результаты демонстрируют корреляцию между диффузионными расстояниями в каждом континууме и наблюдаемыми трудностями освоения соответствующих языковых аспектов. Данный подход не только предлагает новую теоретическую перспективу для исследования языкового обучения, но и создает прогностическую основу для разработки учебных программ, моделирования учащихся и применения в многоязычных NLP- и ИИ-системах.

Ключевые слова: освоение второго языка, языковое расстояние, мультиконтинуальная диффузия, нейронные языковые вложения, моделирование анизотропной диффузии, метод конечных элементов
Для цитирования: Григорьев А.В., Го Ч. Модель изучения языка, основанная на диффузии, и межъязыковая дистанция. *Арктика XXI век.* 2025, № 2. С. 67-74 (на англ.). DOI 10.25587/2310-5453-2025-2-67-74

Introduction

Second language acquisition (SLA) is a complex, multidimensional process influenced by structural properties of both native and target languages. Modeling this process quantitatively and systematically remains a major goal in cognitive science and language education [1; 2]. Traditional SLA models often treat language learning as a monolithic process or focus on isolated aspects. In contrast, we propose a unified mathematical model that treats SLA as a coupled diffusion process within multiple linguistic continua.

The present work draws on analogies from applied mathematics, particularly the modeling of transport phenomena in porous and biological media [3; 4]. Double-porosity models describe transport in media with distinct but interacting structures, such as fractures and pores [5]. Similarly, blood filtration in liver lobules can be modeled using a multicontinua framework, capturing both macroscopic vascular flows and microscopic capillary dynamics. These approaches inspire a novel representation of SLA dynamics as transport across multiple, interacting linguistic subsystems.

Theoretical Framework

Linguistic Continua

We represent the learner's knowledge state as a point in a high-dimensional space composed of few interacting continua (i.e. four continua):

Case 1

Oral language (α = 1): Speaking and Listening

Symbolic language (α = 2): Reading and Writing

Case 2

Phonetics (α = 1): Sound systems and articulation patterns

Syntax (α = 2): Grammatical structure

Semantics (α = 3): Meaning representation

Pragmatics (α = 4): Contextual and social language use

Each continuum α evolves over time according to its own diffusion equation, governed by a permeability tensor and interacts with other continua through exchange terms reflecting interdependence [3; 4].

Diffusion Model

To describe Second language acquisition process, we introduce multicontinua model such as Barenblatt et al. [5] introduced double porosity model for filtration process in ground soil based on the interrelation between the filtration flows in the fractures and those in the pore blocks. The mathematical model, called a multicontinua model, can be written in tensor form as follows:

$$w_i^{\alpha} = -\frac{k_{ij}^{\alpha}}{\mu} \nabla_j u^{\alpha}, \qquad (1)$$

$$\frac{\partial c^{\alpha}\rho}{\partial t} + \nabla_j \rho w_j^{\alpha} = q^{\alpha} \quad (2)$$

$q^{\alpha} = q^{\alpha}(p^{\propto}, \rho, \mu)$, $\rho^{\alpha} = \rho(p^{\alpha})$, $m^{\alpha} = m(p^{\alpha})$, α = 1, 2 or α = 1..4

where:

$u^{\alpha}(x,t)$ is the learner's proficiency in component α,

c^{α} is the capacity (difficulty weighting) for component α,

k_{ij}^{α} is the anisotropic diffusion tensor (capturing language-specific difficulty),

q^{α} is the coupling term between continua α and other continua.

Language Distance and Domain Construction

A key step in configuring the computational domain (Fig. 1) is the definition of interlingual distances. Determining distance (or proximity) between languages can be done in different ways – depending on which aspects of the language we are considering. In turn, language distance can be directly viewed as the length of the computational domain. For simplicity, we operationalize language distance using neural network-derived embeddings. Pretrained multilingual models such as mBERT [7], LASER [8], and XLM-R [9] offer fixed-size language representations that encode structural and functional similarities. For each language, a representative corpus (e.g., Universal Dependencies, Wikipedia) is embedded and projected into a shared vector space. Cosine similarity between embeddings provides a symmetric distance matrix, which we use to configure the diffusion domain geometry.

Numerical Implementation

For versatility, we discretize the spatial domain using finite element methods (FEM), with independent triangulations for each continuum. Time evolution is handled via an explicit-implicit scheme. The coupled variational problem is solved iteratively using GMRES methods [3]. Model parameters (capacities, diffusion tensors, coupling coefficients) are estimated from corpus and linguistic data [10]. For simplicity and generality, we repeat the results described in the article [4].

Consider a simple case of a two-dimensional model of double porosity for next computational geometry (Fig. 1).

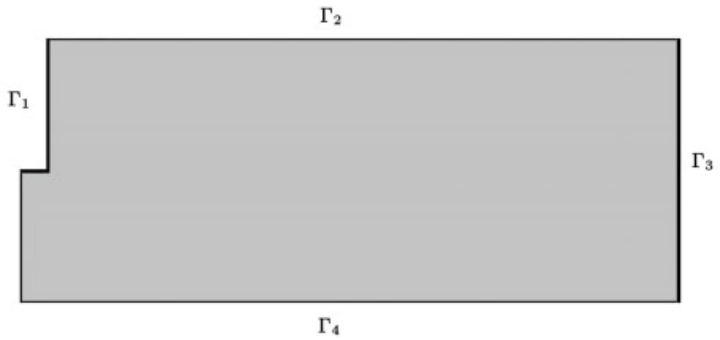

Fig. 1. Computational domain
(length of domain equal to distance between language)

Рис. 1. Вычислительная область
(длина области равна расстоянию между языками)

Results

We can directly inherit the results from the article [3], in this case we can talk about a different interpretation. As before, there is an interconnection of continua, that one aspect of language affects another aspect. The aspect that is better developed is the source for improvements in the other continuum (aspect of language).

Fig. 2. Language progress in aspect u1 at the moment t=1

Рис. 2. Прогресс языка в аспекте u1 в момент t=1

Fig. 3. Language progress in aspect u2 at the moment t=1

Рис. 3. Прогресс языка в аспекте u2 в момент t=1

Fig. 4. Language progress in aspect u1 at the moment t=2

Рис. 4. Прогресс языка в аспекте u1 в момент t=2

The main work has to be devoted to fine-tuning the model, and there is a significant difference in the interpretation of the results.

Discussion

This approach provides a generalizable, interpretable framework for modeling SLA. The multicontinua formalism allows explicit representation of linguistic heterogeneity, while the diffusion paradigm captures the gradual and interconnected nature of learning. The hierarchical design supports both theoretical insight and practical applications in educational technology and multilingual AI [1; 2; 10].

Language proficiency can be evaluated either at the right boundary of the computational domain – representing the final learning outcome – or as an integral value over the entire domain, capturing the cumulative level of acquisition. Additionally, specific language learning difficulties can be modeled as subregions with low permeability, effectively slowing down the diffusion process in those areas [3]. This allows us to simulate and observe the impact of obstacles in second language learning, as well as to analyze potential improvements when such barriers are removed.

Overall, the proposed framework supports the integration of hybrid intelligence technologies. It enables a synergistic combination of numerical methods for forecasting and neural networks for fine-tuning the model structure, configuring the computational domain, and optimizing key problem parameters. This fusion enhances both the predictive accuracy and adaptability of the system, making it suitable for personalized learning and intelligent language education platforms.

Conclusion

We introduced a novel diffusion-based model for second language acquisition that integrates ideas from porous media theory and neural networks applications. By treating linguistic domains as interacting continua and incorporating anisotropic diffusion and multicontinua coupling, we provide a structured and empirically grounded representation of SLA dynamics.

References / Литература

1. Dörnyei Z. *The Psychology of Second Language Acquisition*. Oxford: Oxford University Press. 2009.

2. Ellis NC. Selective attention and transfer phenomena in SLA: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*. 2006;27(2):164-194. DOI: 10.1093/applin/aml015

3. Vabishchevich PN, Grigoriev AV. Numerical modeling of fluid flow in anisotropic fractured porous media. *Numerical Analysis and Applications*. 2016;9(1):45-56. DOI: 10.1134/S1995423916010055

4. Grigorev AV, Vabishchevich PN. Two-level approach for numerical modeling of blood flow in the liver lobule. Journal of Numerical Analysis, Industrial and Applied Mathematics. 2022;16(1-2):15-28.

5. Barenblatt GI, Zheltov IP, Kochina IN. Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks. *Journal of Applied Mathematics and Mechanics*. 1960;24(5):1286-1303. DOI: 10.1016/0021-8928(60)90107-6

6. Rabinovich M, Ordan N, Wintner S. Found in translation: Reconstructing phylogenetic language trees from translated texts. *Transactions of the Association for Computational Linguistics*. 2017;5:169-182. DOI: 10.1162/tacl_a_00058

7. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*. 2019;4171-4186. DOI: 10.18653/v1/N19-1423

8. Artetxe M, Schwenk H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*. 2019;7:597-610. DOI: 10.1162/tacl_a_00288

9. Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL*. 2020;8440-8451. DOI: 10.18653/v1/2020.acl-main.747

10. Lake BM, Baroni M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *Proceedings of ICML*. 2018;80:2873-2882. Available at: https://proceedings.mlr.press/v80/lake18a.html

Об авторах

ГРИГОРЬЕВ Александр Виссарионович – кандидат физико-математических наук, доцент, научно-исследовательская кафедра «Вычислительные технологии», Институт математики и информатики, ФГАОУ ВО «Северо-Восточный федеральный университет имени М.К. Аммосова», Якутск, Российская Федерация, ORCID: 0000-0001-5231-8780, Researcher ID: H-7502-2016, Scopus Author ID: 57194029133, Elibrary AuthorID: 7855-8090, e-mail: re5itsme@gmail.com

ГО Ч. – кандидат физико-математических наук, преподаватель, Институт математических наук, Университет Ляочэн, Ляочэн, КНР, ORCID: 0000-0002-0165-9363, Researcher ID: GQO-9442-2022, Scopus Author ID: 57215305659, e-mail: guozhenweilcu@163.com

About the authors

Aleksandr V. GRIGOREV – Cand. Sci. (Physics and Mathematics), Associate Professor, Institute of Mathematics and Information Science, Scientific Research Department "Computing Technologies", Ammosov North-Eastern Federal University, Yakutsk, Russian Federation, ORCID: 0000-0001-5231-8780, Researcher ID: H-7502-2016, Scopus Author ID: 57194029133, Elibrary AuthorID: 7855-8090, e-mail: re5itsme@gmail.com

Zhenwei GUO – Cand. Sci. (Physics and Mathematics), Teacher, School of Mathematical Sciences, Liaocheng University, Liaocheng, P.R. China, ORCID: 0000-0002-0165-9363, Researcher ID: GQO-9442-2022, Scopus Author ID: 57215305659, e-mail: guozhenweilcu@163.com

Вклад авторов

Григорьев А.В. – разработка концепции, методология, программное обеспечение, верификация данных, проведение исследования, проведение статистического анализа, администрирование данных, создание черновика рукописи, редактирование рукописи, визуализация, руководство исследованием, администрирование проекта

Го Ч. – разработка концепции, методология, программное обеспечение, верификация данных, проведение исследования, администрирование данных, создание черновика рукописи, редактирование рукописи, визуализация

Authors' contribution

Grigorev A.V. – conceptualization, methodology, software, validation, investigation, formal analysis, data curation, writing - original draft, writing - review & editing, visualization, supervision, project administration

Guo Zh. – conceptualization, methodology, software, validation, investigation, data curation, writing - original draft, writing - review & editing, visualization

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов

Conflict of interests

The authors declare no conflict of interest

Поступила в редакцию / Submitted: 16.05.2025
Принята к публикации / Accepted: 30.05.2025