УДК 372.881.161.1' 373.423.2. DOI: 10.26170/2071-2405-2026-31-1-174-184.
ББК Ч426.819=411.2,9+Ш105.315.1.
ГРНТИ 14.25.07. Код ВАК 5.8.2

MODELING RUSSIAN-SLOVAK INTERLINGUAL HOMONYMY VIA EMBEDDINGS

Lukas Gajarsky
University of Ss. Cyril and Methodius in Trnava (Trnava, Slovakia)
ORCID ID: https://orcid.org/0000-0001-8090-6977

Mikhail Kipchatov
University of Ss. Cyril and Methodius in Trnava (Trnava, Slovakia)
ORCID ID: https://orcid.org/0000-0003-3021-6390

A b s t r a c t . Interlingual lexical interference has long been an object of scholarly research. It is well documented that when learning a foreign language genetically related to the learner's native language, increased levels of cross-linguistic interference can occur. From this perspective, interlingual homonyms represent particularly high-risk lexical items, as their formal similarity may lead to incorrect mapping between form and meaning. Embedding models capture distributional differences in language use and broader semantic associations. In the present study, static embedding models (fastText and MUSE) and a deep embedding model (OpenAI) were applied to a dataset of Russian-Slovak lexical pairs consisting of interlingual homonyms and translation equivalents (150 items in each category). Based on similarity patterns observed across the models, a five-level typology of interlingual homonymy was proposed (evident-risk, hidden-risk, medium-risk, conceptual-risk, and asymmetric-risk). The predictive potential of the model was tested on a sample of 46 Slovak learners of Russian as a foreign language. A consistent correspondence between model-based risk predictions and learner performance was observed. Lexical items classified as high-risk produced significantly higher error rates among learners, and an asymmetry between productive and receptive tasks was also observed. The results suggest that embedding models may serve as an empirically grounded tool for supporting vocabulary learning in closely related languages.

K e y w o r d s : interlingual homonymy; vector word presentations; predictive didactics; lexica; interference; lexical units; language modeling; embeddings; lexical pairs; interlingual homonyms; Russian language; Russian lexicology; Slovak language; Slovak lexicology; Russian as a foreign language; methods of teaching Russian; Slovak students

A c k n o w l e d g e m e n t : For language editing and stylistic refinement, an AI-based language model (ChatGPT, OpenAI) was used as a supportive tool under the authors' supervision. All conceptual, methodological, and interpretative decisions remained the responsibility of the authors.

F o r c i t a t i o n : Gajarsky, L., Kipchatov, M. (2026). Modeling Russian-Slovak Interlingual Homonymy Via Embeddings. In *Philological Class*. Vol. 31. No. 1, pp. 174–184. DOI: 10.26170/2071-2405-2026-31-1-174-184.

МОДЕЛИРОВАНИЕ РУССКО-СЛОВАЦКОЙ МЕЖЪЯЗЫКОВОЙ ОМОНИМИИ С ПОМОЩЬЮ ЭМБЕДДИНГОВ

Гаярски Л.
Университет им. Св. Кирилла и Мефодия в Трнаве (Трнава, Словакия)
ORCID ID: https://orcid.org/0000-0001-8090-6977

Кипчатов М.
Университет им. Св. Кирилла и Мефодия в Трнаве (Трнава, Словакия)
ORCID ID: https://orcid.org/0000-0003-3021-6390

А н н о т а ц и я . Межъязыковая лексическая интерференция на протяжении длительного времени является предметом научных исследований. Хорошо известно, что при изучении иностранного языка, близкородственного родному, может наблюдаться повышенный уровень межъязыковой интерференции. С этой точки зрения межъязыковые омонимы представляют собой лексические единицы повышенного риска, поскольку их формальное сходство может приводить к неправильному соотнесению формы и значения. Модели эмбеддингов отражают как распределительные различия в употреблении языковых единиц, так и более широкие семантические ассоциации. В настоящем исследовании статические модели эмбеддингов (fastText и MUSE), а также глубокая модель эмбеддингов (OpenAI) были применены к корпусу русско-словацких лексических пар, включающему межъязыковые омонимы и переводные эквиваленты (по 150 единиц в каждой категории). На основе выявленных моделей сходства была предложена пятиуровневая типология межъязыковой омонимии (evident-risk, hidden-risk, medium-risk, conceptual-risk и asymmetric-risk). Предиктивный потенциал модели был проверен на выборке из 46 словацких учащихся, изучающих русский язык как иностранный. Было выявлено устойчивое соответствие между прогнозируемым моделью уровнем риска и результатами учащихся. Лексические единицы, классифицированные моделью как высокорисковые, вызывали значительно более высокий уровень ошибок. Кроме того, была зафиксирована асимметрия между продуктивными и рецептивными заданиями. Полученные результаты позволяют рассматривать модели эмбеддингов как эмпирически обоснованный инструмент поддержки освоения лексики при изучении близкородственных языков.

К л ю ч е в ы е с л о в а : межъязыковая омонимия; векторные представления слов; предиктивная дидактика; лексическая интерференция; лексические единицы; языковое моделирование; эмбеддинги; лексические пары; межъязыковые омонимы; русский язык; лексикология русского языка; словацкий язык; лексикология словацкого языка; РКИ; русский язык как иностранный; методика преподавания русского языка; словацкие школьники

Благодарности: для языкового редактирования и стилистической доработки в качестве вспомогательного инструмента под руководством авторов использовалась языковая модель на основе искусственного интеллекта (ChatGPT, OpenAI). Все концептуальные, методологические и интерпретационные решения оставались за авторами.

Для цитирования: Гаярски, Л. Моделирование русско-словацкой межъязыковой омонимии с помощью эмбеддингов / Л. Гаярски, М. Кипчатов // Филологический класс. – 2026. – Т. 31, № 1. – С. 174–184. – DOI: 10.26170/2071-2405-2026-31-1-174-184.

Introduction

Interference plays an important role in foreign-language instruction. Positive interference facilitates the correct association of form and meaning, whereas negative interference arises when learners associate formally similar units that differ in lexical meaning. This phenomenon poses significant challenges for both teachers and students. In genetically related languages such as Russian and Slovak, the issue becomes even more pronounced, as learners tend to assign meaning to unfamiliar lexical items on the basis of graphic or phonological similarity. Interlingual homonymy thus represents a source of systematic difficulty in foreign-language learning and carries a high potential for communicative error.

In classroom practice, systematic problems in vocabulary acquisition are particularly evident in lexical items that exhibit a high degree of formal similarity. Some interference-induced errors can be eliminated relatively easily; others require repeated explanation and clarification in multiple contexts; and some appear resistant even at higher levels of language proficiency. Certain items also tend to display irregular or unpredictable patterns of use. Moreover, an asymmetry between productive and receptive skills can be observed. Although learners are often able to correctly associate content with form in receptive tasks, significantly higher error rates tend to occur in productive performance.

Embedding models capture a distributional approximation of meaning, including the distributional behaviour of lexemes in corpora, degrees of contextual similarity, and collocational patterns. At the same time, they are capable of modelling conceptual relatedness, semantic similarity at more abstract levels, and cross-linguistic semantic mapping. In the subsequent analysis, distributional distance and conceptual proximity are treated as two analytically distinct dimensions. Embedding models thus hold substantial potential for linguistic research, as they enable the quantification of meaning and make it possible to conduct analyses that would otherwise be difficult or impossible to perform manually. They may therefore represent a valuable tool in the field of predictive didactics. In the present study, predictive didactics refers to the identification of potential learning difficulty prior to instruction on the basis of model-based lexical similarity patterns. Embeddings represent words, phrases, or entire texts as vectors in a high-dimensional space, where the distance between vectors reflects the degree of distributional or semantic similarity [cf. Devlin et al. 2016]. From this perspective, Russian and Slovak constitute a highly suitable language pair, as both languages contain hundreds of lexemes that are formally similar yet differ in meaning. The key question is whether embedding-based predictions can be validated using real learner data.

Based on these considerations, four research objectives were formulated. First, it was examined whether modern embedding models can reliably distinguish between semantic equivalents and interlingual homonyms. Second, the extent to which these models can serve as a foundation for predictive didactics was assessed, particularly in diagnosing lexical items that require increased instructional attention in foreign-language acquisition. Third, it was investigated whether the combination of static and deep embedding models can identify structured levels of interference risk and thereby support the development of a multi-level diagnostic framework. Finally, an empirical validation was conducted to determine whether lexical items predicted as high-risk by the proposed model correspond to increased error rates in learner performance.

To address these objectives, three embedding architectures – fastText, MUSE, and OpenAI text-embedding-3-large – were employed and their ability to model distributional and conceptual similarity between Russian and Slovak lexemes was compared. The resulting risk predictions were subsequently validated through a learner experiment involving Slovak students of Russian as a foreign language.

Unlike traditional approaches that rely on retrospective error analysis or contrastive descriptions of false friends, the present study adopts a predictive didactic perspective. Lexical risk is not inferred from previously observed learner errors but is operationalised prior to instruction through embedding-based similarity patterns. The learner experiment therefore functions not as an exploratory error study, but as an external validation of model-based risk predictions.

Theoretical Background

Interlingual homonymy and its didactic relevance. Interlingual homonymy is a phenomenon frequently encountered by translators, foreign-language teachers, and learners. It arises when two or more linguistic units from different languages – whether genetically related or unrelated – coincide fully or partially in their graphic or phonological form but differ in meaning. According to D. Kollár, interlingual homonymy refers to the acoustic or graphic similarity, or both, of words with different meanings across two or more language systems [1987: 223]. The current relevance of this topic is reflected in the large number of studies and dictionaries of “false friends,” a common umbrella term for such misleading lexical pairs.

Interlingual homonymy has various sources. In related languages, formally identical or analogous lexical units often share a common historical origin, such as roots derived from Proto-Slavic in the case of Slavic languages. In the present study, interlingual homonymy is understood as a set of phenomena characterised by formal similarity or analogy combined with semantic non-equivalence or partial divergence, including differences in connotations. However, as M. Pančíková [2003] notes, from a language-didactic perspective, practical methodologies for teaching and mastering interlingual homonyms remain insufficiently developed in Slavic languages.

In the scholarly literature on false friends and interlingual homonymy, classifications of homonym types and analyses of errors occurring either in teaching contexts or in translation are commonly encountered. Across the SLA literature, it is widely acknowledged that lexical interference is one of the most frequent forms of transfer in genetically related languages. The greater the formal similarity between lexical units, the higher the likelihood of incorrect associations between form and meaning. It is also well established that lexical units with a high degree of formal similarity may lead to errors in foreign-language production. Research further indicates the non-uniform nature of interference, suggesting that not all lexical items have the same potential to cause errors. Stronger interlingual interference effects are observed in lexical units characterised by formal similarity and differing frequency profiles in the first and foreign language [Kootstra, Dijkstra, Starren 2015]. From a cognitive perspective, "false friends" manifest themselves in specific cross-linguistic situations that may result in incorrect meaning mapping [O'Neill, Casanovas 1997]. Empirical studies also demonstrate an asymmetry between production and reception in lexical units marked by high similarity. This implies that the elimination of interference-related errors does not guarantee the absence of errors in foreign-language production [Odlin 1989]. Receptive abilities thus often precede productive competence. Given this variability and asymmetry, there is a need for approaches that allow potential lexical risk to be identified in advance. Modern tools such as quantitative models are capable of identifying high-risk lexemes in advance, which is emerging as one of the key challenges for contemporary didactic research [AlKhuzaey et al. 2023].

The combination of formal similarity and semantic divergence represents a measurable challenge for learners and, from a computational perspective, constitutes a particularly suitable and relevant phenomenon for distributional modelling.

Distributional and deep embedding models in the service of predictive didactics. In modeling lexical, semantic, and distributional relations in natural language processing (NLP), embedding models are widely adopted as a standard approach. Embeddings provide vector-based representations of words in a multidimensional space, and their distance correlates with the degree of similarity in their distributional and semantic properties.

In static embedding models, each word is assigned a single vector that does not change and is unaffected by context. A major strength of these models is their high sensitivity to syntagmatic and collocational patterns, which makes it possible to compare different languages, as they are represented within a shared vector space. From the perspective of cross-linguistic interference and contrastive lexicology, this aspect is essential, since it allows the measurement of cross-linguistic distances. Another advantage of these models lies in the relative simplicity of their architecture and the consistency and interpretability of their outputs.

The modern generation of embeddings – such as OpenAI's text-embedding-3-large – is based on the architecture of large language models (LLMs), which in turn rely on transformer networks. Deep models generate contextual representations [Brown et al. 2020]: a word therefore no longer has a single fixed vector; rather, its representation is dynamically conditioned by the surrounding context, with representations arising in the higher layers of the neural network. This enables such models to capture not only distributional patterns but also more abstract semantic regularities [Peters et al. 2018].

From a didactic perspective, this makes it possible to operationalise lexical similarity in quantifiable terms. In other words, these language models represent a tool that allows the identification of potential sources of cross-linguistic interference. Distributional models enable the examination of distributional distance, while deep embedding models capture conceptual proximity; in the present study, these are understood as two analytically distinct dimensions of lexical relatedness.

Embedding models do not model learner cognition directly; rather, they capture structured regularities of language use encoded in large-scale corpora.

Embeddings as a tool for identifying interference-risk lexemes. Embedding models make it possible to quantify semantic distance between formally similar lexemes and to compare cross-linguistic similarity profiles in a systematic and reproducible manner. Interference risk arises from specific configurations of distributional distance and conceptual proximity across models.

From a didactic perspective, the interaction between formal similarity and embedding-based similarity patterns is crucial. Lexical items combining strong formal overlap with particular cross-model similarity configurations are more likely to trigger systematic learner difficulties. The aim is therefore not to construct static lists of so-called "false friends," but to identify structured risk profiles grounded in semantic representations derived from complementary embedding architectures.

Previous research on Slavic languages has primarily focused on the technical evaluation of bilingual embeddings (e.g., fastText, MUSE). In the present study, these models are employed as diagnostic instruments for assessing interference risk within a predictive didactic framework. By integrating static and deep embeddings, a multi-level typology of interlingual homonymy is constructed, capturing qualitatively distinct similarity configurations rather than a simple linear scale of difficulty.

The proposed approach is subsequently validated through learner performance data.

Methodology

Data and lexical material. For this study, a dataset was created to enable both linguistic comparison of lexemes and the identification of items posing an increased risk of interference in vocabulary acquisition in Slavic languages. The dataset comprises 150 Russian-Slovak translation equivalents and 150 Russian-Slovak interlingual homonyms and is publicly available on the author's GitHub (https://github.com/Lukas-Gajarsky/russian-slovak-homonymy-dataset). The lexical material was selected to systematically cover both formally similar and semantically divergent items, allowing controlled contrasts between predicted high-risk and low-risk categories.

Translation equivalents were selected as stable, high-frequency lexical pairs with unambiguous semantic correspondence across languages (e.g., автор – autor, škola – школа, múzeum – музей), forming a low-risk reference baseline for comparison of distributional behaviour. Interlingual homonyms were defined as formally similar lexemes exhibiting partial or full semantic divergence (e.g., палец – palec, болото – blato, электричка – električka), corresponding to items with heightened interference potential. From a didactic perspective, these lexemes represent the group with the highest interference risk.

The list was compiled from multiple complementary sources:

– existing Russian-Slovak lexicographic literature;

– the author's dictionary of Russian-Slovak interlingual homonyms;

– previous contrastive and interference-oriented studies;

– manual verification of semantic distinctness and usage patterns.

The dataset was deliberately designed to be didactically relevant, covering common vocabulary found in Russian and Slovak foreign-language curricula and representing frequent sources of interference errors. Each item was labeled by type (translation equivalent / interlingual homonym) to enable reproducible risk profiling. The dataset was not intended to be lexicographically exhaustive but diagnostically representative, prioritizing internal consistency and pedagogical relevance over full coverage.

The dataset further served as the basis for constructing a five-level interference-risk typology presented in the Discussion section. This typology derives from model-based similarity patterns across multiple embedding architectures and is independently supported by learner performance data, providing external validation of the dataset design and mitigating potential concerns regarding selection subjectivity.

Embedding models and similarity computation. In this study, three embedding models were employed: the static distributional models fastText and MUSE, and a deep neural OpenAI embedding model. The static models were selected for their sensitivity to collocational and distributional variation, allowing the identification of lexical items that occupy distinct contextual environments across languages. The OpenAI model, operating within a shared multilingual vector space, enables direct cross-linguistic comparison and captures higher-level semantic associations that extend beyond immediate collocational patterns.

The inclusion of both static and deep embeddings was methodologically motivated. The study requires analytically separable similarity dimensions: distributional distance and conceptual proximity. Static embeddings alone tend to represent both fine-grained distributional contrasts and broader semantic associations along a single similarity continuum, limiting their independent examination. The deep embedding model thus provides a complementary similarity dimension that cannot be derived from static architectures alone.

The models were therefore employed in a complementary manner to minimise model-specific bias. By combining similarity scores across architectures, stable cross-model patterns of lexical proximity were identified and used as the empirical basis for constructing the interference-risk typology presented in the Discussion section.

Construction of the diagnostic typology. Semantic similarity between Russian and Slovak lexemes was operationalized using cosine similarity within the respective embedding spaces. Static distributional models (fastText, MUSE) were interpreted as reflecting distributional distance between lexemes, whereas the OpenAI embedding model was treated as capturing a complementary dimension of conceptual proximity within a shared multilingual vector space. The strong correlation between fastText and MUSE and their only moderate correlation with the OpenAI model empirically supported this analytical distinction.

Each lexical pair was thus represented through similarity scores that were analytically interpreted along two separable dimensions: distributional distance and conceptual proximity. Rather than constituting a strict geometric two-dimensional space, these dimensions represent complementary similarity perspectives derived from different embedding architectures. This distinction is consistent with well-documented sources of cross-linguistic interference described in the literature: lexical items may exhibit high formal similarity yet differ in their usage patterns (distributional divergence), or they may display intuitive category-based equivalence despite distributional mismatch (conceptual proximity).

This framework initially provided the basis for a quantitative similarity continuum. However, recurrent cross-dimensional patterns were observed across lexical pairs. The resulting five-level typology (evident-risk, hidden-risk, medium-risk, conceptual-risk, and asymmetric-risk homonyms) therefore reflects qualitatively distinct cognitive constellations rather than merely different degrees of similarity. Interference risk is thus interpreted not as a linear function of similarity, but as the outcome of specific interactions between distributional experience and conceptual structuring.

Statistical validation and model-based contrast. The underlying assumption was that translation equivalents would exhibit higher levels of embedding similarity than interlingual homonyms. The objective at this stage was therefore to verify the basic separability of these two lexical groups. A non-parametric Mann-Whitney U test was employed for this purpose. As the hypothesis was directional, a one-sided test was applied to similarity scores obtained from all embedding models. The results confirmed a statistically significant difference between the two groups of lexemes. The test thus served as a validation framework for the subsequent stages of the research.

Visualizations as diagnostic tools. A set of complementary visualization techniques such as boxplots, KDE, PCA, correlation analysis was employed to facilitate interpretation of similarity patterns. These were used for diagnostic support rather than inferential testing.

Reproducibility and technical environment. All computational experiments were conducted in a standard Python environment using established libraries for vector processing and similarity computation, including Gensim and NumPy. Cosine similarity was applied consistently across all embedding models. The fastText and MUSE embeddings were obtained from the official Meta AI repositories, while the OpenAI text-embedding-3 model was accessed via the OpenAI API.

To ensure transparency and reproducibility, the lexical dataset used in the study is publicly available, and all similarity computations can be replicated using standard embedding workflows. In addition, the anonymized binary learner-response data, including the scoring scheme used for analysis, are provided in spreadsheet format in the same GitHub repository.

Learner-based empirical validation. To externally validate the predictive potential of the proposed embedding-based diagnostic typology, a learner-based experiment was conducted. The aim was to examine whether lexical items classified as high-risk by the model indeed lead to increased error rates in learner performance.

The participant group consisted of 46 learners enrolled in a Russian-Slovak bilingual secondary school program in Slovakia. All participants were native speakers of Slovak with B1–B2 proficiency in Russian, ensuring a relatively homogeneous learner profile.

A targeted lexical test was constructed directly on the basis of the interference-risk typology. The test comprised 24 items, equally divided into productive (translation) and receptive (multiple-choice) tasks. Items represented both high-risk and low-risk categories, enabling a controlled comparison between model-based predictions and observed learner behavior.

Responses were scored using a strict binary scheme. Only the most direct and contextually appropriate translation equivalent was coded as correct (1); all alternative solutions were coded as incorrect (0) in order to preserve lexical precision.

Statistical analysis. Learner performance data were analysed using non-parametric methods due to the binary nature of the responses (correct = 1, incorrect = 0). Lexical items were classified as low-risk or high-risk prior to statistical analysis on the basis of the embedding-based typology. Analyses were conducted at both the item level and the participant level. At the item level, mean accuracy rates per lexical item were compared across risk categories. Group differences were tested using the Mann-Whitney U test, and effect sizes were calculated using r. At the participant level, individual learners' mean accuracy scores for high-risk and low-risk items were contrasted using a Wilcoxon signed-rank test, as the design involved paired (dependent) measurements. Effect sizes were calculated using r.

Results

The results are presented in two complementary parts. First, similarity patterns generated by the embedding models are reported and statistically evaluated. Second, learner performance data are analysed to empirically validate the predictive potential of the proposed embedding-based risk classification.

Descriptive statistics of embedding similarity. Across all three embedding models (fastText, MUSE, OpenAI), translation equivalents consistently exhibited higher cosine similarity values than interlingual homonyms. Mean similarity scores for both groups are summarized in Table 1.

Table 1

Model	*mean(eq)*	*mean(hom)*
fastText	0,501	0,272
MUSE	0,534	0,303
OpenAI	0,618	0,565

The static embedding models fastText and MUSE show a clearer separation between translation equivalents and interlingual homonyms, whereas the OpenAI model yields higher absolute similarity values with reduced variability. Boxplot visualizations (Figure 1) confirm a consistent upward shift in similarity distributions for translation equivalents across all models.

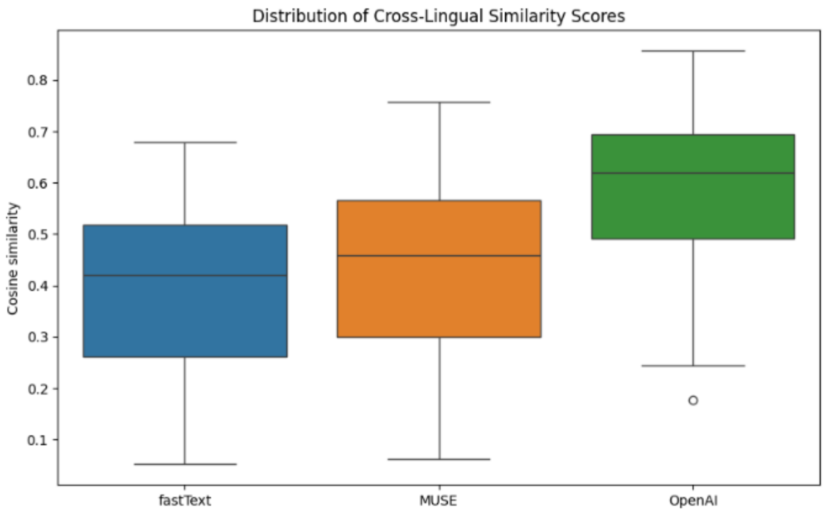

Figure 1. Boxplot of cosine similarity distributions between Russian and Slovak lexical units in the fastText, MUSE, and OpenAI models

Distributional patterns across models. KDE distributions provide a more fine-grained view of similarity patterns across embedding models. In the static models (fastText and MUSE), a broader spread of similarity values can be observed, whereas the OpenAI model exhibits a more right-shifted and comparatively compressed distribution. This pattern suggests that the OpenAI deep embedding model assigns generally higher similarity scores and displays reduced dispersion relative to the static models.

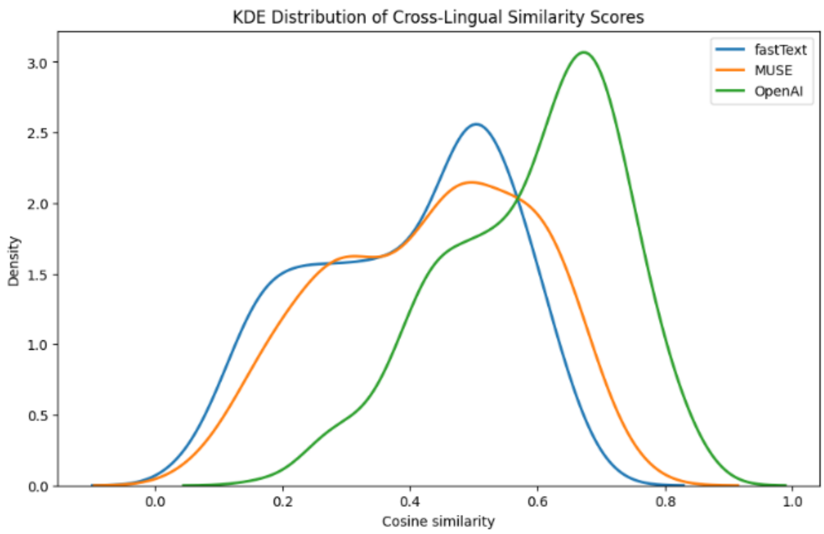

Figure 2. Kernel density curves (KDE) of cosine similarity for the fastText, MUSE, and OpenAI models

PCA-based representation of the OpenAI embedding space. The PCA projection reveals that Russian and Slovak lexemes tend to cluster predominantly according to language membership along the first principal component, while still exhibiting areas of overlap. This overlap corresponds to lexemes that occupy relatively proximate positions in the similarity space derived from the OpenAI embeddings. Such proximity is consistent with elevated cross-linguistic similarity scores and may signal increased confusability between certain lexical items.

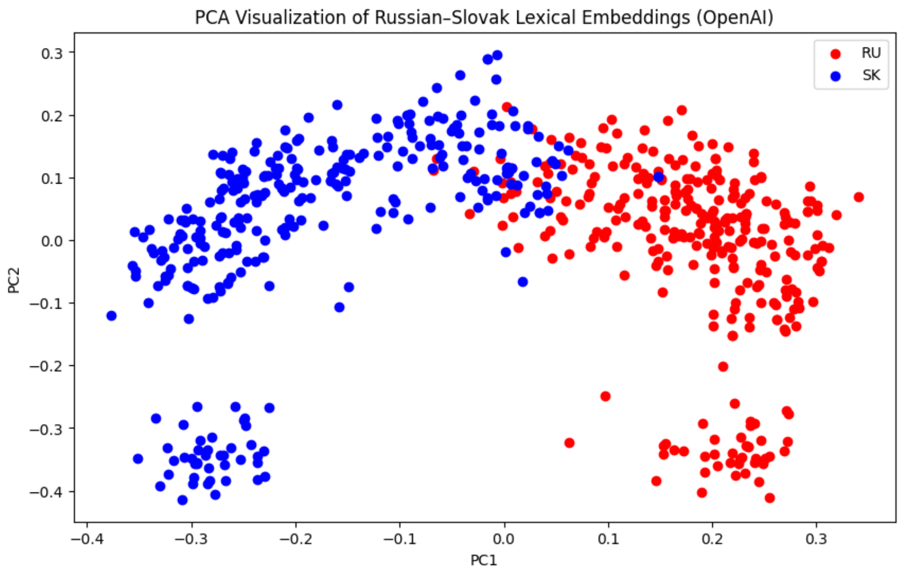

Figure 3. PCA projection of Russian and Slovak lexemes from the OpenAI text-embedding-3-large model

Cross-model correlation analysis. The moderate correlations between the static models and the OpenAI model (r = 0,42 and r = 0,40) indicate partial overlap alongside divergence in similarity structure. In contrast, the very strong correlation between fastText and MUSE (r = 0,94) confirms that the two static models encode highly aligned distributional patterns. This pattern supports the interpretation that the OpenAI embeddings capture similarity relations not fully reducible to distributional proximity.

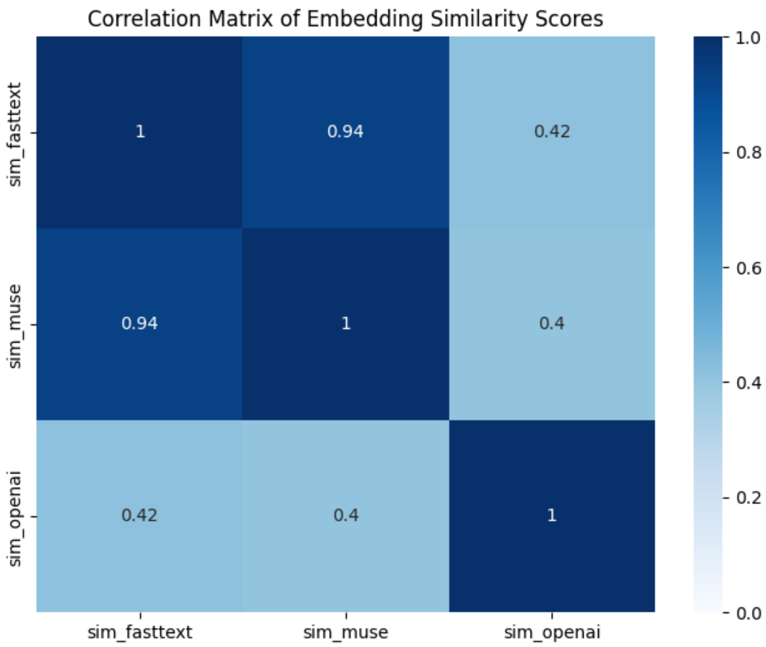

Figure 4. Correlation matrix

Statistical confirmation of model-based contrast. To statistically verify the observed group differences, a one-sided Mann-Whitney U test was applied to the similarity scores from each embedding model as the hypothesis was explicitly directional (translation equivalents were expected to exhibit higher similarity than interlingual homonyms). In all models, translation equivalents achieved significantly higher similarity values:

fastText: $U = 18{,}364.5$, $p = 1{,}55 \times 10^{-35}$

MUSE: $U = 16{,}144.5$, $p = 2{,}77 \times 10^{-32}$

OpenAI: $U = 14{,}032.5$, $p = 1{,}06 \times 10^{-4}$.

These results provide statistical confirmation that translation equivalents and interlingual homonyms are systematically distinguishable within the embedding space across model architectures. This made it possible to establish the foundation for the interference-risk typology.

Learner-based empirical validation results. In the next step, the results were externally validated using learner data. The aim was to examine whether lexical items classified by the model as high-risk would indeed be associated with higher error rates among learners. The data were analysed at two levels: the item level and the participant level.

At the item level, a non-parametric Mann-Whitney U test was applied. Mean accuracy rates of individual items were compared across risk categories. A statistically significant difference was observed between low-risk and high-risk items, with a moderate effect size. High-risk items yielded significantly higher error rates.

At the participant level (paired design), a Wilcoxon signed-rank test was employed, as the measurements were dependent. For each participant, mean accuracy in the low-risk and high-risk lexical categories was compared. The difference was highly significant ($W = 0.0$, $p = 3{,}30 \times 10^{-9}$). Nearly all participants achieved better performance on low-risk items, as illustrated by the paired comparison plot (Figure 5).

The expected asymmetry between receptive and productive performance was also observed: error rates were more pronounced in production tasks, which constitute a more sensitive indicator of lexical interference.

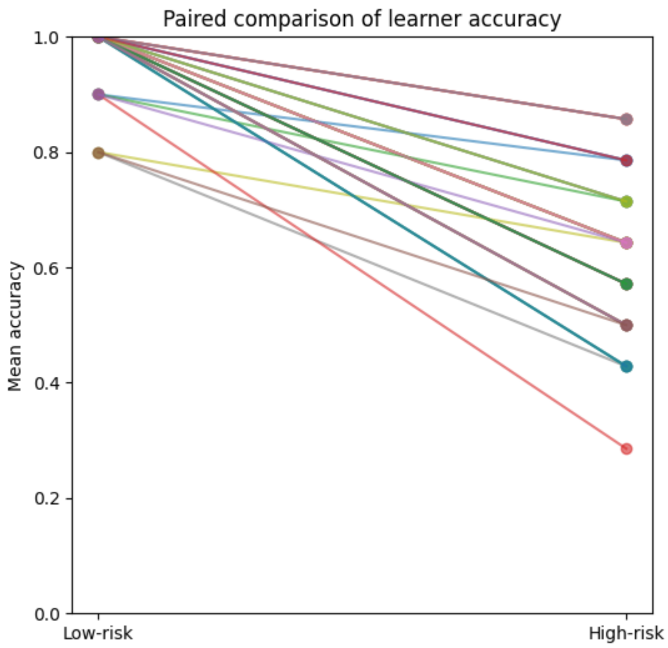

Figure 5. Paired line plot comparing mean learner accuracy for low-risk and high-risk lexical items. Each line represents an individual learner's paired mean accuracy across the two risk categories

Discussion

Model-based prediction of lexical interference. The aim of this study was to examine whether embedding systems, understood as computational representations of lexical similarity, can be used to predict interference risk – that is, to anticipate potential learner errors before they actually occur – rather than merely describing them ex post. The analysis was conducted using a dataset of Russian-Slovak interlingual homonyms.

The results showed that both static and deep embedding models reliably distinguish between interlingual homonyms, representing high-risk lexical items, and translation equivalents, which constitute low-risk items. In the next stage, a learner-based experiment was conducted to provide external validation. The experiment involved 46 students forming a relatively homogeneous participant group.

The results confirmed that lexical items classified as high-risk indeed produced higher error rates among learners. Embedding models capture both distributional signals and conceptual relations; they therefore function not only as descriptive tools but may also serve as indicators of potential difficulties in vocabulary acquisition.

Distributional and conceptual similarity as complementary interference mechanisms. The study distinguishes two dimensions of cross-linguistic interference: a distributional dimension, reflecting differences in lexical usage, and a conceptual dimension, related to mental categories. The distributional component – distributional and collocational similarity – is captured by static embedding models, in our case fastText and MUSE. From this perspective, these models effectively separate high-risk items (interlingual homonyms) from low-risk items (translation equivalents), thereby identifying overt interference risks.

However, relying solely on static embeddings would reduce the description of interference to a single continuum of similarity. To capture the conceptual dimension of interference – conceptual associations between lexical items – a deep embedding model (OpenAI) was therefore employed. This model may indicate latent interference risks that are not visible in purely distributional patterns.

The interaction of these two dimensions reflects different types of lexical similarity. Lexical interference thus emerges from the interaction between two factors: distributional divergence and conceptual proximity.

Five-level typology of interlingual homonymy. On the basis of the outputs of static and deep embedding models, a five-level typology of interlingual homonymy was proposed, allowing interference risk to be interpreted not only in terms of its intensity, but also in terms of its cognitive source. The typology does not constitute a continuous scale of difficulty; rather, it reflects distinct configurations of formal overlap, distributional divergence, and conceptual similarity. From this perspective, interference risk can be understood not merely as a matter of degree, but as the outcome of qualitatively different similarity patterns. The proposed categories are conceived as prototypical interference profiles rather than as mutually exclusive classes, reflecting the continuous nature of similarity relations in embedding spaces and the gradient character of lexical interference.

Table 2 provides an overview of the proposed five-level typology, summarising the characteristic similarity configurations and risk profiles associated with each homonym type.

Table 2

Homonym type	*Similarity configuration*	*Dominant interference mechanism*	*Risk profile*	*Representative example*
Evident-risk	low static × low deep	formal similarity	clear semantic divergence; predictable form-based error	*брак – brak*
Hidden-risk	low static × high deep	conceptual proximity	intuitive sense of equivalence despite divergent usage	*электричка – električka*
Medium-risk	moderate similarity across models	partial semantic overlap	correct category, incorrect specification of meaning	*лекция – lekcia*
Conceptual-risk	very low static × very high deep	conceptual illusion of equivalence	category-based mapping overriding usage	*палец – palec*
Asymmetric-risk	maximal model conflict	conflicting similarity signals	unstable lexical representation	*канон – kanón*

Evident-risk homonyms: clear semantic divergence. These pairs are characterised by very low similarity values across all models, indicating that their semantic and collocational profiles are fully separated despite strong formal resemblance. Learners therefore tend to infer meaning incorrectly.

A representative example is брак – brak ('defect' vs. 'waste material'). Although the two forms are nearly identical, the Russian lexeme is associated with craftsmanship and evaluative terminology, whereas the Slovak item denotes industrial residue. Similar patterns occur in захват – záchvat ('capture' vs. 'seizure') and заказник – zákazník ('nature reserve' vs. 'customer'). Because these items predictably trigger form-based misinterpretation, they require explicit warning and contrastive presentation, and after explicit instruction the error typically disappears.

Hidden-risk homonyms: high conceptual similarity despite distributional divergence. These pairs receive low similarity in static embeddings but high similarity in the OpenAI model. Although their distributional profiles differ substantially, they share conceptual domains that prompt a strong sense of equivalence in learners.

The clearest example is электричка – električka ('suburban train' vs. 'tram'). While static models correctly capture distinct collocational frames, the OpenAI model groups the items together under the broader category of urban transport. This mirrors the learner's intuitive categorisation and explains frequent confusion even among advanced speakers. Such items benefit from contextual and situational clarification.

Medium-risk homonyms: partial semantic overlap. Medium-risk pairs show moderate similarity across models and typically involve asymmetric polysemy or partial overlap in semantic fields. Learners often identify the general conceptual domain correctly but misinterpret the specific meaning.

Examples include лекция – lekcia ('lecture' vs. 'lesson'), болото – blato ('swamp' vs. 'mud'), and жаль – žiaľ ('pity' vs. 'sorrow'). Here, interference results not from a fundamental misunderstanding but from subtle semantic drift. Prototype clarification, collocational contrasts, and contextual examples help reduce error rates in this category.

Conceptual-risk homonyms: high-level conceptual affinity. These homonyms attain high similarity in the OpenAI model but very low similarity in static embeddings. In the semantic space of an LLM, they appear conceptually close even though their linguistic usage differs markedly. This therefore constitutes a conceptual "illusion" of equivalence. From a cognitive perspective, this involves mapping a foreign-language lexeme primarily at the level of mental categories rather than on the basis of its contextual and collocational behaviour. In this case, the interference error does not arise from a lack of knowledge of meaning, but from faulty abstraction: the lexemes are perceived as equivalent because they activate a similar conceptual frame, even though their linguistic usage differs systematically.

For instance, палец – palec ('finger' vs. 'thumb') belong to the same anatomical category, yet each language assigns them a different referential scope. In канон – kanón ('canon' vs. 'cannon'), conceptual clustering emerges from shared orthography and high-level categorical associations, despite striking semantic divergence. Such lexemes require discourse-anchored explanation and explicit differentiation of usage domains.

Asymmetric-risk homonyms: maximal conflict between distributional and conceptual signals. Asymmetric-risk pairs exhibit the strongest divergence between models: static embeddings classify them as unrelated, while the OpenAI model interprets them as conceptually similar. This produces conflicting cues and leads to persistent interference. The absence of a coherent similarity profile results in particularly unstable lexical representations and persistent production errors.

This pattern is again exemplified by канон – kanón. Distributionally, the Russian item belongs to normative and religious discourse, whereas the Slovak lexeme refers to a weapon system. Conceptually, however, the OpenAI model groups them together on the basis of formal similarity and broad categorical associations. Such items require explicit correction of intuitive but misleading semantic assumptions.

Taken together, the proposed typology shows that interlingual homonymy cannot be treated as a homogeneous phenomenon. Rather, individual homonym types activate different cognitive mechanisms and, as a consequence, call for different pedagogical responses.

Learner performance: perception vs production. The external validation test showed that high-risk items produced significantly higher error rates, and nearly all participants achieved better scores on low-risk items. SLA research has long documented an asymmetry between receptive and productive tasks, which was also observed in the present test. Although many lexical items were correctly recognised by learners, errors frequently occurred when they were required to produce them in translation tasks.

The most frequent difficulties were observed in the groups of hidden-risk, conceptual-risk, and asymmetric-risk homonyms. This pattern may be explained by the differing cognitive demands of receptive and productive tasks. In receptive tasks, learners only need to recognise the lexical item, whereas production requires active retrieval of the word and the inhibition of competing forms, which increases the likelihood of interference.

Production errors tended to occur with conceptually related lexical items. This pattern is consistent with the predictions of the embedding models: lexemes that occupy regions of conceptual proximity in the embedding space tended to trigger production errors more often. These findings suggest that productive performance constitutes a more sensitive indicator of lexical interference than receptive recognition.

Implications for predictive didactics. In the scholarly literature, analyses typically focus on errors that learners have already produced. In contrast to this retrospective perspective, predictive didactics aims to anticipate potential vocabulary difficulties before they occur. Embedding models and their measures of semantic similarity therefore offer promising possibilities for application within predictive didactics. The typology of interlingual homonymy proposed in this study, based on the dominant interference mechanism, makes it possible to apply differentiated pedagogical strategies. Predictive didactics cannot replace pedagogical judgment, but it may serve as a useful tool for identifying potential problem areas in vocabulary acquisition. In this sense, embedding-based analysis may function as an analytical bridge between computational modelling of lexical similarity and practical foreign-language pedagogy.

Conclusion

This study has shown that the potential of embedding-based similarity measures can be used not only as descriptive tools, but also as relevant means for predicting lexical interference in foreign vocabulary acquisition. By combining static embedding models designed to capture distributional divergence with deep embeddings reflecting different levels of semantic organization, it was possible to identify complementary mechanisms of interference. Moderate correlations between static and deep embeddings further indicate that these models encode distinct and non-redundant types of similarity.

The proposed five-level typology of interlingual homonymy, capturing qualitatively different mechanisms of lexical interference, suggests that lexical interference is a heterogeneous phenomenon. Validation on learner data confirmed that lexical items predicted by the model as high-risk were associated with higher error rates among learners, particularly in productive tasks.

Overall, the results of this study situate predictive didactics as an empirically grounded extension of contrastive linguistics. Embedding models may thus be viewed as offering potential for further application across different language pairs and learning contexts.

Литература

A comparative evaluation and analysis of three generations of Distributional Semantic Models / A. Lenci, M. Sahlgren, P. Jeuniaux [et al.] // Language Resources and Evaluation. – 2022. – Vol. 56, no. 4. – P. 1269–1313. – DOI: 10.1007/s10579-021-09575-z. – EDN YOSRAP.

BERT: Pre-training of deep bidirectional transformers for language understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). – Association for Computational Linguistics, 2016. – P. 4171–4186.

Boleda, G. Distributional Semantics and Linguistic Theory / G. Boleda // Annual Review of Linguistics. – 2020. – Vol. 6, no. 1. – P. 213–234. – DOI: 10.1146/annurev-linguistics-011619-030303. – EDN BYOZBQ.

Csiriková, M. Zrádná slova v ruštině: Slovník rusko-českých homonym / M. Csiriková, N. Koníčková. – LEDA, 2015.

Deep contextualized word representations / M. E. Peters, M. Neumann, M. Iyyer [et al.] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). – Association for Computational Linguistics, 2018. – P. 2227–2237.

Gajarský, L. Učebnica rusko-slovenských homonym / L. Gajarský, T. Grigorjanová. – Tribun EU, 2020.

Grigorjanová, T. Slovník rusko-slovenských medzijazykových homonym / T. Grigorjanová, L. Gajarský. – Tribun EU, 2019.

Iermachkova, O. Význam interkultúrnej komunikácie vo vzdelávaní študentov – rusistov / O. Iermachkova, A. Spišiaková // Inovatívne metódy vo výučbe cudzích jazykov z aspektu interkultúrnej komunikácie II : Medzinárodný vedecký zborník recenzovaných štúdií / ed. by M. Vojteková, Z. Slobodová. – Vydavateľstvo Prešovskej university, 2024. – P. 68–76.

Kollár, D. Medzijazyková homonymia / D. Kollár // Studia Academica Slovaca. – 1987. – Vol. 11. – P. 229–233.

Kootstra, G. J. Second language acquisition / G. J. Kootstra, T. Dijkstra, M. Starren // International encyclopedia of the social & behavioral sciences / ed. by J. D. Wright. – 2nd edition. – Elsevier, 2015. – P. 349–359. – DOI: 10.1016/B978-0-08-097086-8.53025-6.

Language models are few-shot learners / T. B. Brown, B. Mann, N. Ryder [et al.] // Advances in Neural Information Processing Systems. Vol. 33 / ed. by H. Larochelle, M. Ranzato, R. Hadsell [et al.]. – Curran Associates, Inc., 2020. – P. 1877–1901. – DOI: abs/10.5555/3495724.3495883.

O'Neill, M. False friends: A historical perspective and present implications for lexical acquisition / M. O'Neill, M. Casanovas Catalá // Bells: Barcelona English Language and Literature Studies. – 1997. – Vol. 8. – P. 103–115.

Odlin, T. Language transfer: Cross-linguistic influence in language learning / T. Odlin. – Cambridge University Press, 1989.

Pančíková, M. Niekoľko príkladov medzijazykovej homonymie – slovinčina, slovenčina / M. Pančíková // Obdobja: Metode in zvrsti v slovenskem jeziku, literatúre in kulturi. – Ljubljana : Center za slovenščino kot drugi/tuji jezik, Filozofska fakulteta, 2003. – P. 495–500.

Text-based Question Difficulty Prediction: A Systematic Review of Automatic Approaches / S. Alkhuzaey, F. Grasso, T. R. Payne, V. Tamma // International Journal of Artificial Intelligence in Education. – 2023. – DOI: 10.1007/s40593-023-00362-1. – EDN WYPOSM.

Uban, A. S. Automatically building a multilingual lexicon of false friends with no supervision / A. S. Uban, L. P. Dinu // Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020). – European Language Resources Association, 2020. – P. 3001–3007.

References

Alkhuzaey, S., Grasso, F., Payne, T. R., Tamma, V. (2023). Text-based Question Difficulty Prediction: A Systematic Review of Automatic Approaches. *International Journal of Artificial Intelligence in Education*. DOI: 10.1007/s40593-023-00362-1. EDN WYPOSM.

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*, 6(1), 213–234. DOI: 10.1146/annurev-linguistics-011619-030303. EDN BYOZBQ.

Brown, T. B., Mann, B., Ryder, N. et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems (vol. 33)*, 1877–1901. Curran Associates, Inc. DOI: abs/10.5555/3495724.3495883.

Csiriková, M., Koníčková, N. (2015). Zrádná slova v ruštině: Slovník rusko-českých homonym. LEDA.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2016). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 4171–4186. Association for Computational Linguistics.

Gajarský, L., Grigorjanová, T. (2020). Učebnica rusko-slovenských homonym. Tribun EU.

Grigorjanová, T., Gajarský, L. (2019). Slovník rusko-slovenských medzijazykových homonym. Tribun EU.

Iermachkova, O., Spišiaková, A. (2024). Význam interkultúrnej komunikácie vo vzdelávaní študentov – rusistov. *Inovatívne metódy vo výučbe cudzích jazykov z aspektu interkultúrnej komunikácie II*, 68–76. Vydavateľstvo Prešovskej university.

Kollár, D. (1987). Medzijazyková homonymia. *Studia Academica Slovaca*, *11*, 229–233.

Kootstra, G. J., Dijkstra, T., Starren, M. (2015). Second language acquisition. *International encyclopedia of the social & behavioral sciences (2nd edition)*, 349–359. Elsevier. DOI: 10.1016/B978-0-08-097086-8.53025-6.

Lenci, A., Sahlgren, M., Jeuniaux, P. et al. (2022). A comparative evaluation and analysis of three generations of Distributional Semantic Models. *Language Resources and Evaluation*, *56*(4), 1269–1313. DOI: 10.1007/s10579-021-09575-z. EDN YOSRAP.

O'Neill, M., Casanovas Catalá, M. (1997). False friends: A historical perspective and present implications for lexical acquisition. *Bells: Barcelona English Language and Literature Studies*, *8*, 103–115.

Odlin, T. (1989). Language transfer: Cross-linguistic influence in language learning. Cambridge University Press.

Pančíková, M.(2003). Niekoľko príkladov medzijazykovej homonymie – slovinčina, slovenčina. *Obdobja: Metode in zvrsti v slovenskem jeziku, literatúre in kulturi*, 495–500. Ljubljana: Center za slovenščino kot drugi/tuji jezik, Filozofska fakulteta.

Peters, M. E., Neumann, M., Iyyer, M. et al. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2227–2237. Association for Computational Linguistics.

Uban, A. S., Dinu, L. P. (2020). Automatically building a multilingual lexicon of false friends with no supervision. *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, 3001–3007. European Language Resources Association.

Данные об авторах

Гаярски Лукаш – доцент, старший преподаватель кафедры русистики, Университет им. Св. Кирилла и Мефодия, Словакия, г. Трнава, lukas.gajarsky@ucm.sk

Кипчатов Михаил – аспирант, Университет им. Св. Кирилла и Мефодия, Словакия, г. Трнава, kipchatov1@ucm.sk

Authors' information

Gajarsky Lukas – Associate Professor, Senior Lecturer of Department of Russian Studies, University of Ss. Cyril and Methodius in Trnava, Slovakia, Trnava

Kipchatov Mikhail – Postgraduate Student, University of Ss. Cyril and Methodius in Trnava, Slovakia, Trnava

Дата поступления: 11.02.2026; дата публикации: 31.03.2026

Date of receipt: 11.02.2026; date of publication: 31.03.2026