Научный результат. Вопросы теоретической и прикладной лингвистики. Т. 11, №4 2025
Research result. Theoretical and Applied Linguistics, 11 (4). 2025

63

Maria I. Kiose[1] 🆔
Zhenishkul S. Khulkhachieva[2] 🆔
Veronika V. Izyumskaya-Kapitonova[3] 🆔
Artem V. Barmin[4] 🆔

Lexical-semantic clustering in the diagnostics of culture integration in discourse

[1] Moscow State Linguistic University,
38 Ostozhenka St., Moscow, 119034, Russia
Institute of Linguistics, RAS,
1 B. Kislovsky Ln., Moscow, 119019, Russia
E-mail: maria_kiose@mail.ru
ORCID: 0000-0001-7215-0604

[2] Moscow State Linguistic University,
38 Ostozhenka St., Moscow, 119034, Russia
E-mail: khulkhachieva@linguanet.ru
ORCID: 0000-0001-7234-2999

[3] Moscow State Linguistic University,
38 Ostozhenka St., Moscow, 119034, Russia
E-mail: v.v.linguist.ik@gmail.com
ORCID: 0000-0001-8460-8555

[4] Moscow State Linguistic University,
38 Ostozhenka St., Moscow, 119034, Russia
E-mail: art.barmin1@gmail.com
ORCID: 0000-0001-9616-4904

Abstract: Semantic clustering as a research method is commonly utilized to explore lexical systems transformations, meanwhile it can provide access to far more complicated processes of cognitive transformations in discourse or even culture. This work presents a cognitive-semantic approach to the study of cultural integration, based on semantic clustering. Two cognitive processes of cultural integration are identified: direct and indirect knowledge transfer. These are manifested in the direct and indirect translation of areas of knowledge from the donor culture to the recipient culture. The analysis procedure and results are demonstrated using a corpus of 864,005 examples of words borrowed from Russian into Kyrgyz from web discourse.
Objectives. The research objectives are to reveal the knowledge domains transferred directly from Russian to Kyrgyz culture through integrated loanwords in web discourse and to specify the roles of qualia in the indirect transfer of knowledge through loanword collocates.
Methods. While corpus analysis allows to identify the knowledge domains which are directly transferred from donor into recipient culture, semantic clustering helps specify the indirect culture integration since it explores the contexts mediated by the use of these domains.

Киосе М. И., Хулхачиева Ж. С., Изюмская-Капитонова В. В., Бармин А. В. Лексико-семантическая…
Kiose M. I., Khulkhachieva Zh. S., Izyumskaya-Kapitonova V.V., Barmin A.V. Lexical-semantic clustering…

64

Results. Based on the established lists of loanwords, the compiled database of their contexts and lexical-semantic clustering algorithm performed to identify their similarity in the context use of loanwords (accessed at https://osf.io/78u4m/), the study determined both the thematic attribution and qualia roles of the knowledge domains in direct and indirect knowledge transfer as a result of culture integration which is found in the context of Russian (as donor) and Kyrgyz (as recipient) cultures. Featuring 4,126 target words with their vector similarity weight identified, further analysis revealed their cognitive roles attributed to four qualia – formal, constitutive, telic, agentive, which were indirectly mediated by culture integration.

Conclusion. The study shows that, while direct knowledge transfer resulted in the penetration of novel thematic domains and the modification of existing ones, indirect knowledge transfer mostly resulted in the transformation of the formal role of encoding taxonomic information about the referent of the loanword. This suggests that the recipient culture adapted the knowledge domains within a wider domain matrix rather than adapting or transforming their components for more specific purposes. The results obtained provide evidence for novel applications of lexical-semantic clustering, as well as novel methods for exploring cultural integration and disintegration processes.

Keywords: Lexical-semantic clustering; Direct knowledge transfer; Indirect knowledge transfer; Loanword; Qualia-role; Culture integration

---

Киосе М. И.[1] ⓘ
Хулхачиева Ж. С.[2] ⓘ
Изюмская-Капитонова В. В.[3] ⓘ
Бармин А. В.[4] ⓘ

Лексико-семантическая кластеризация в диагностике процессов культурной интеграции в дискурсе

[1] Московский государственный лингвистический университет
ул. Остоженка, 38, стр. 1, Москва, 119034, Россия
Институт языкознания РАН
Большой Кисловский пер., 1, стр. 1, Москва, 119019, Россия
E-mail: maria_kiose@mail.ru
ORCID: 0000-0001-7215-0604

[2] Московский государственный лингвистический университет
ул. Остоженка, 38, стр. 1, Москва, 119034, Россия
E-mail: khulkhachieva@linguanet.ru
ORCID: 0000-0001-7234-2999

Научный результат. Вопросы теоретической и прикладной лингвистики. Т. 11, №4 2025
Research result. Theoretical and Applied Linguistics, 11 (4). 2025

65

3 Московский государственный лингвистический университет
ул. Остоженка, 38, стр. 1, Москва, 119034, Россия
E-mail: v.v.linguist.ik@gmail.com
ORCID: 0000-0001-8460-8555

4 Московский государственный лингвистический университет
ул. Остоженка, 38, стр. 1, Москва, 119034, Россия
E-mail: art.barmin1@gmail.com
ORCID: 0000-0001-9616-4904

Аннотация: Семантическая кластеризация как исследовательский метод традиционно используется для изучения преобразований лексических систем, однако данный метод может обеспечить доступ к гораздо более сложным когнитивным трансформациям, наблюдаемым в дискурсе и в культуре. Настоящая работа развивает когнитивно-семантический подход к изучению процесса интеграции культур с опорой на метод семантической кластеризации. Разграничиваются два когнитивных процесса интеграции культур, прямой и косвенный трансфер знания, проявляющиеся в непосредственной и опосредованной трансляции областей знания из культуры-донора в культуру-реципиент. Процедура анализа и результаты демонстрируются на материале сформированного корпуса примеров из веб-дискурса (всего 864,005 контекстов) с заимствованными словами из русского в киргизский язык.

Задачи. Задачи исследования заключаются в установлении сфер знания, интегрированных из русской в киргизскую культуру в ходе прямого трансфера знания в веб-дискурсе, и в определении квалиа-ролей, представляющих это знание, транслируемое косвенным образом через коллокаты заимствованных слов.

Методы. В то время как корпусный анализ позволяет установить области знания, которые транслируются из культуры-донора в культуру-реципиент напрямую, а также степень прямой интеграции культур, именно семантическая кластеризация помогает идентифицировать особенности косвенной интеграции культур с опорой на контексты, опосредованные использованием этих областей знания.

Результаты. На основе составленных списков заимствованных слов, сформированного корпуса их контекстов и семантического алгоритма кластеризации, реализованного для выявления их сходства в контекстах с заимствованными словами (материалы доступны по ссылке: https://osf.io/78u4m/), в исследовании определены как тематические области знания, так и квалиа-роли, участвующие в прямой и косвенной культурной интеграции русской (как культуры-донора) и киргизской (как культуры-реципиента) культур. Последующий анализ выявленных 4,126 таргетных слов с опорой на коэффициент их векторного сходства выявил их когнитивные роли, описываемые четырьмя квалиа – формальной, конститутивной, телической, агентивной, опосредованные интеграцией культур.

Заключение. Результаты показали, что прямой трансфер знания привел как к появлению новых областей знания, так и к трансформации областей знания,

Киосе М. И., Хулхачиева Ж. С., Изюмская-Капитонова В. В., Бармин А. В. Лексико-семантическая…
Kiose M. I., Khulkhachieva Zh. S., Izyumskaya-Kapitonova V.V., Barmin A.V. Lexical-semantic clustering…

66

уже существующих в культуре-реципиенте. Косвенный трансфер знания в основном привел к трансформации формальной (конструирование таксономической информации о референте заимствованного слова) роли, что свидетельствует о том, что культура-реципиент в основном адаптировала области знания в рамках более обширной матрицы областей знания, не трансформируя их компоненты для частных целей. Полученные результаты свидетельствуют о возможности новых применений метода лексико-семантической кластеризации для изучения процессов интеграции и дезинтеграции культур.

Ключевые слова: Лексико-семантическая кластеризация; Непосредственный трансфер знания; Опосредованный трансфер знания; Заимствованное слово; Квалиа-роль; Интеграция культур

1. Introduction

Semantic clustering is at present a robust natural language processing method (or rather a group of methods) employed in investigating semantic relationship through the shared meanings the words maintain in context (Kuhn et al., 2007; Witschard et al., 2022). It allows to cluster words and contexts through word embeddings (Word2Vec) or through transformer-based models (BERT) by calculating their similarity in the vector space and summarizing the data.

Meanwhile, we presume that apart from identifying the similarity between texts, semantic clustering integrated into cognitive framework can contribute to exploring more complicated research areas including discourse or even culture similarity. Cognitive framework allows model the relations of cognitive processes (Talmy, 2000; Langacker, 2007, Verhagen, 2007, Gallagher, Hutto, 2008); however, to proceed it requires collecting extensive data on how these processes are performed through experimental or corpus research. Since semantic clustering reveals the relationship of words, it can be attributed to knowledge domains and qualia roles – formal, constitutive, telic, and agentic (Pustejovsky, 1991, 2006) specifying discourse types and culture types. The current paper is the first attempt in exploring the cognitive process of knowledge transfer in culture integration through semantic clustering, which allows to identify the thematic attribution of knowledge domains and the qualia roles subjected to adapting the knowledge developed within a donor culture and transferred onto recipient culture. The found solution lies within distinguishing two types of knowledge transfer in culture integration, direct and indirect, exposed in immediate and mediated transfer of knowledge domains from donor into recipient culture.

Therefore, we propose a two-step research procedure which involves: 1) the corpus analysis allowing to identify the thematic attribution of knowledge domains which are directly transferred from donor into recipient culture and the degree of direct culture integration; 2) lexical-semantic clustering allowing to specify the qualia roles

Научный результат. Вопросы теоретической и прикладной лингвистики. Т. 11, №4 2025
Research result. Theoretical and Applied Linguistics, 11 (4). 2025

67

in indirect knowledge transfer since it explores the contexts mediated by the use of these domains. The procedure is verified featuring the lexical criterion applied to the loanwords contributing to culture integration with Russian language donor culture and Kyrgyz language recipient culture. The research objectives are consequently to reveal the thematic attribution of knowledge domains which were transferred from Russian into Kyrgyz culture directly through the integrated loanwords in web discourse, and to specify the qualia roles of the knowledge transferred indirectly through the loanword collocates. To proceed, a list of loanwords in the Kyrgyz language with Russian as etymon language was compiled, which was further applied to explore the direct and indirect knowledge transfer in culture integration.

The Theoretical Framework section provides an overview of the prerequisites for addressing cultural integration through the lexical criterion. Section Three presents the research methodology and two-step procedure of corpus analysis and lexical-semantic clustering for exploring direct and indirect transfer. Section Four demonstrates the results determining: 1) the knowledge domains representing the loanwords and the distribution of loanwords in the compiled web database as well as their thematic attribution, exposing the results of direct knowledge transfer, 2) the distribution of target words displaying high vector similarity with the loanwords in the compiled web database obtained through semantic clustering, and qualia roles representing the referents of these loanwords in the web contexts, exposing the results of indirect transfer. Section Four discusses the results in the context of semantic clustering framework and cognitive frameworks in attribution to culture integration. Section Five lists the study limitation. The Final Remarks section specifies the novel applications of lexical-semantic clustering in exploring culture integration and disintegration processes.

2. Theoretical Framework

2.1. Lexical-semantic clustering in exploring discourse and culture

Semantic clustering is generally referred to the practice of grouping elements into categories based on their meanings (Tinkham, 1993). So, the term "semantic cluster" is usually understood as a set of elements (words, concepts, etc.) that can be qualitatively or quantitatively analyzed. Guided by this general understanding of semantic clustering, various scientific theories develop specific methods for identifying and analyzing lexical-semantic clusters. For instance, cognitive network science (Siew et al., 2019) posits that lexical-semantic clusters can be represented as network structures – specifically, as sets of three nodes (concepts) interconnected by edges (semantic relations). Analyzing the overall clustering of a network reveals the efficiency of its structure for searching and retrieving information.

Troyer et al. (1997) defined a "semantic cluster" as a group of items from the same lexical-semantic subcategory (e.g., zebra, tiger and lion for "wild animals" subcategory) in a verbal fluency task. They measured two key indices: cluster size (the number of consecutively listed items from a single subcategory) and cluster switching frequency (the rate of alternation between various semantic subcategories). Kuhn et al. (2007) defined lexical-semantic clusters as groups of source artifacts (e.g., packages, classes or methods) that use similar vocabulary; and tried to identify them using Latent Semantic Indexing (Deerwester, 1990) and clustering (Jain et al., 1999). Lexical-semantic clustering is a versatile methodology with diverse research applications. For example, semantic clustering can be used to investigate the processes of foreign language acquisition (Pérez-Serrano et al., 2022), cognitive memory strategies in aging (Shaffer et al., 2024), improving the efficiency and robustness of large language models (LLMs) (Lee et al., 2025).

While these studies employ lexical-semantic clustering to explore the structure of knowledge domains, we found no ready

Киосе М. И., Хулхачиева Ж. С., Изюмская-Капитонова В. В., Бармин А. В. Лексико-семантическая…
Kiose M. I., Khulkhachieva Zh. S., Izyumskaya-Kapitonova V.V., Barmin A.V. Lexical-semantic clustering…

68

lexical-semantic clustering decisions to explore the extent of culture integration and the domains mediated by knowledge transfer. This study is the first attempt to develop this decision and to exemplify its efficacy in the context of determining the input of one language (donor language) into the second language (recipient language).

2.2. Culture integration exposed in direct and indirect knowledge transfer through loanwords

Following Myers-Scotton (2002, 2006) and Haspelmath (2009), in this study culture integration which is a complex phenomenon of forming shared cultural and historical knowledge, is addressed through lexical criterion, i.e., through the loanwords which fall into two types: "a) cultural borrowings which designate a new concept coming from outside, and b) core borrowings which duplicate meanings for which a native word already exists" (Haspelmath, 2009: 46). Whereas there is commonly no strong necessity to employ an alien loanword to construe a concept, it frequently penetrates the native language and culture due to a strong cultural convention in the community to use another language as a marker of cultural or ethnic identity. In most cases this is the social identity which stimulates the appearance of loanwords. Adapting this idea to exploring the distribution of loanwords in Russian and Kyrgyz, the official languages of two CIS countries, Russia and Kyrgyzstan, we expect to determine the knowledge domains of culture integration through their lexical systems integration. Importantly, in the study we do not classify these adaptation process into adoption or imposition (Winford, 2005) or reveal the reasons of borrowing, i.e., cultural pressure, loss of vitality, genealogical position, etc. (Thomason, 2001). To explore culture integration through lexical criterion of loanwords, we categorize the effect of a loanword from the donor language on the lexical stock of the recipient language as insertion without further classifying it into replacement and coexistence (Haspelmath,

2009), with the view of identifying the knowledge domains representing the loanwords in the recipient language and their distribution. To proceed, we adhere to the concept of cultural knowledge transfer following (Espagne, 1999; Feschenko et al., 2016) who view it as "culturally specific transfer of domains and conceptual modifications which appear in their importing and exporting from one culture into another" (Feschenko, Bochaver, 2016: 5). Knowledge transfer is commonly addressed through its techniques, technologies and routes (Postovalova, 2016; Demyankov, 2016; Iriskhanova, Kiose, 2016). The symptoms of culture integration are consequently the knowledge domains and their conceptual structure which are introduced or modulated through knowledge transfer. Consequently, culture integration is viewed through the integrated or modulated knowledge domains in the recipient culture development.

2.3. Qualia roles in identifying culture integration

To address knowledge domains and their structure, the cognitive framework developed a wide range of methods in exploring the cognitive processes of selection, perspective, abstraction, specificity, prominence, perspective, dynamicity, metaphoricity, integration, intersubjectivity (Talmy, 2000; Langacker, 2007). Meanwhile, these methods are seldom if ever used in computational linguistics since they are efficient in determining the principal regulations of construal processes, not their distribution (Verhagen, 2007, Gallagher, Hutto, 2008). To comply with the computational task, we address the method of qualia roles (Pustejovsky, 1991) developed within the generative lexicon framework.

A Quale is a term borrowed from philosophy and used in generative linguistics to refer to a single aspect of a word's meaning identified on the basis of the relation between the concept expressed by the word and another concept that the word evokes (Pustejovsky, Jezek, 2016). Four qualia roles

Научный результат. Вопросы теоретической и прикладной лингвистики. Т. 11, №4 2025
Research result. Theoretical and Applied Linguistics, 11 (4). 2025

69

are thus explored: 1) Formal encoding taxonomic information about the lexical item (the is-a relation, what kind of thing is it, what is its nature?); 2) Constitutive encoding information on the parts and constitution of an object (part-of or made-of relation, what is it made of, what are its constituents?); 3) Telic encoding information on purpose and function (the used-for or functions-as relation, what is it for, how does it function?); 4) Agentive encoding information about the origin of the object (the created-by relation, how did it come into being, what brought it about?) (Pustejovsky, 1991, 2006). As claimed, this is the process of syntagmatic co-composition (Pustejovsky, 1991; Copestake, Briscoe, 2000) which causes sense modulation; consequently, the qualia can be attributed in context only.

For instance, in the contexts such as: a. *This car weighs over 2,000 lbs.*; b. *We buy vehicles such as cars and buses*; c. *John started the car*, d. *You should warm your car up in winter*, e. *Did you lock the car?*, f. *The car screeched down the road*, two types of qualia are attributed to *car*, which are Formal qualia (sample b) and Constitutive qualia (samples a, c, d, e, in the construal of its engine, door, wheels). In the contexts such as a. *He owns a two-story house*, b. *Lock your house when you leave*, c. *We bought a comfortable house*, d. *The house is finally finished* (Pustejovsky, Jezek, 2016), all four qualia types are observed, which are Formal (sample a, house as an artifact), Constitutive (sample b, door as part of the house), Telic (sample c, purpose of the house), Agentive (sample d, origin of the house). Schematically, in (Pustejovsky, Jezek, 2016) these attributions are represented within formal semantic network in the following way (Figure 1).

Figure 1a. Qualia-roles attributed to *car* (Pustejovsky, Jezek, 2016: 8)

$$\left[\begin{array}{l} car \\ \text{QUALIA} = \left[\begin{array}{l} \text{F} = \textbf{vehicle} \\ \text{C} = \{ \textbf{engine, door, wheels, ...} \} \end{array} \right] \end{array} \right]$$

Figure 1b. Qualia-roles attributed to *house* (Pustejovsky, Jezek, 2016: 9)

$$\left[\begin{array}{l} house \\ \text{QUALIA} = \left[\begin{array}{l} \text{F} = \textbf{building} \\ \text{C} = \{ \textbf{door, rooms, ...} \} \\ \text{T} = \textbf{live\_in} \\ \text{A} = \textbf{build} \end{array} \right] \end{array} \right]$$

The qualia structure framework was further adopted in ontology studies (Pustejovsky et al. (2004), in computational linguistics, e.g., in Copestake and Briscoe (1995), Cimiano, Wenderoth (2005), in the design of lexical databases, e.g., in Jezek et al. (2009), in the Brandeis Semantic Ontology proposed in Pustejovsky et al. (2006), in cross-cultural semantic studies (Lee et al., 2010; Song, Qiu, 2013). These and other works confirm the efficiency of the qualia roles framework in "mapping out the range of possible conceptual transfers available and also motivating their existence" (Copestake and Briscoe, 1995: 15).

Following these tenets, we hypothesize that semantic clustering integrated into cognitive framework reveals the paths of knowledge transfer through the distribution of a) thematic attribution of knowledge domains and b) qualia roles in the web contexts with loanword collocates.

3. Methods and Research procedure

The research is conducted within the framework of cognitive linguistics, which uses computational methods to explore how knowledge domains are distributed. This framework is used in the present study to identify the knowledge domains affected by the culture integration through knowledge transfer displayed in the loanwords and their contexts. Following the studies exploiting the notion of knowledge transfer in application to cultural contacts (Espagne, 1999; Feschenko et al., 2016), we expect to reveal its effects onto the knowledge domains distribution due to Russian-Kyrgyz contacts. To determine the

Киосе М. И., Хулхачиева Ж. С., Изюмская-Капитонова В. В., Бармин А. В. Лексико-семантическая…
Kiose M. I., Khulkhachieva Zh. S., Izyumskaya-Kapitonova V.V., Barmin A.V. Lexical-semantic clustering…

70

integrated and modulated knowledge domains in the recipient culture, we consider two types of knowledge transfer in culture integration: 1) direct knowledge transfer which is a cognitive process exemplified in immediate incorporation of loanwords from the donor language into the recipient language appearing in the knowledge domains distribution; 2) indirect knowledge transfer which is a cognitive process exemplified in the knowledge domains distribution of target words as collocates of the loanwords. Indirect knowledge transfer displays the effectuated culture integration in the recipient language, since it reveals the knowledge domains which undergo the influence of loanwords.

Two types of knowledge transfer – direct and indirect – in the present study are explored in the web discourse database of the Kyrgyz language compiled and processed by the authors. Russian served as the lingua franca for Russian-Kyrgyz interactions, resulting in a large number of loanwords appearing in the Kyrgyz language. Meanwhile, their circulation within Kyrgyz culture reflects the norms and strategies of cross-cultural communication (Derbisheva, 2017, 2019; Kambaralieva, Sternin, 2019). At present due to the high rate of migration and at the same time the change in the vitality status of Russian in Kyrgyzstan (Shipilov, 2018), an efficient instrument for culture integration diagnostics is required. Since the web discourse reflects these social and cultural processes, we will use it to develop and validate the culture integration diagnostics instrument through revealing the knowledge domains affected by the knowledge transfer. Consequently, we apply corpus and lexical-semantic clustering analysis (Troyer et al., 1997; Kuhn et al., 2007, among others) to explore the direct and indirect knowledge transfer in this discourse type. Corpus analysis helps identify the distribution of loanwords and the thematic attribution of knowledge domains in the recipient language, while semantic clustering allows to reveal the distribution of target words displaying high vector similarity with

the loanwords in the compiled web database obtained through semantic clustering and qualia roles (Pustejovsky, 1991; Pustejovsky, Jezek, 2016) representing the referents of these loanwords in the web contexts, exposing the results of indirect knowledge transfer.

Two criteria for exploring knowledge domains are thus exploited. While considering the distribution of knowledge domains of loanwords, both cultural and core borrowings (Haspelmath, 2009), we apply the thematic criterion. Meanwhile, this criterion is not efficient in identifying the knowledge domains of the target words (the loanword collocates) since they appear in their contextual meanings, display varied POS attribution and are far more numerous. Consequently, to proceed we apply the qualia roles criterion developed in the computational research within the generative linguistics (generative lexicon) framework.

To comply with the research objectives, we developed a two-step procedure of exploring knowledge transfer in culture integration through lexical criterion of loanwords integrating two steps: 1) identifying the domains of direct knowledge transfer through the distribution of loanwords in the compiled web database; 2) identifying the domains of indirect knowledge transfer through the distribution of collocates of the loanwords in the database.

Step 1. At this stage, we identified the lexical domains of direct knowledge transfer and its indicators: a) the ratio of loanwords from Russian as donor language into the Kyrgyz to all loanwords and to all words in the language, b) the ratio of loanword samples in web databases to the overall number of samples.

Step 1.1. To compile a list of loanwords from Russian into the Kyrgyz language, we addressed the loanwords dictionaries which identified the direction of borrowing and the donor language. Two dictionaries were applied for the purpose: 1) "The Dictionary of Loanwords" compiled by Kh. Karasaev (1986) which contains 5,100 loanwords, where the loanwords with Russian as donor

Научный результат. Вопросы теоретической и прикладной лингвистики. Т. 11, №4 2025
Research result. Theoretical and Applied Linguistics, 11 (4). 2025

71

language are marked with "*op*" (or), 2) "The Dictionary of the Kyrgyz Language" compiled by K. Yudakhin (1985) which contains 100,000 words, where the loanwords with Russian as donor language are marked with "p" (r). Two lists of loanwords were further integrated, the words appearing in both lists were counted as one. All word forms of each loanword were next included into the final list of loanwords. This procedure allowed to identify the key knowledge domains and their thematic attribution in performing the transfer and the lexical indicator of direct transfer considering the ratio of loanwords from Russian to the total amount of loanwords and the total number of words in the Kyrgyz language.

Step 1.2. Next, two databases of web discourse were compiled to perform the corpus search of loanwords frequency. Compiling two databases, with the first (Kyrgyz Web text corpus (Kyrgyzstan)) featuring the language use up to 2019, and the second (Kyrgyz News Corpus) featuring the language use up to 2025, was necessitated by the absence of one resource covering a representational period of loanword use. The Kyrgyz Web Text Corpus (Kyrgyzstan) is part of the open-access Leipzig Corpora, which comprises 3,036,362 sentences and 36,581,130 tokens. The Kyrgyz News Corpus was created by AI Kyrgyz specialists (as part of The Cramer Project) launched in 2025, which contains 256,364 news items (as provided in the description) collected through web scraping from a range of online news portals. The dataset encompasses a wide thematic spectrum, including politics, economics, culture, and sports. Each entry consists of the full text of a news item along with metadata on its source (The Kyrgyz News Corpus dataset, 2025). While the Kyrgyz Web Text Corpus allowed for the search process, the Kyrgyz News Corpus did not; therefore, extensive pre-processing and processing were performed in Python.

The primary objective was the development of a structured dictionary of borrowed lexemes implemented in Python. In this context, a dictionary is understood not as a lexicographic product, but as a data structure — "a set of key-value pairs, with the requirement that the keys are unique (within one dictionary)" (Python Documentation, 2025). The keys were defined as the parameters of each lemma, its graphical variants, and its word forms. Such an approach, standard in natural language processing, not only facilitated the systematic structuring of lexemes but also enabled the subsequent extraction of examples containing the target units.

Manually extracted data from the Leipzig Corpora Collection were merged with borrowing examples from the Kyrgyz News Corpus, resulting in a consolidated dataset used for the construction of semantic networks. Finally, the parallel databases were integrated to further explore the use frequency of each loanword within the integrated database serving as an indicator of direct knowledge transfer featuring the lexical decision.

Step 2. At this stage, we identified the qualia roles that demonstrate indirect knowledge transfer and the indicators of indirect cultural integration. These are: a) the lexical system indicator, which shows the scaled contingency rate of collocates of loanwords from Russian as the etymon language in Kyrgyz; and b) the functional indicator, which shows the scaled average contingency rate of qualia roles that demonstrate indirect cultural integration. To proceed, we elaborated the semantic clustering procedural decision for corpus analysis. It helped identify: a) the distribution of loanword collocates, and b) the semantic clusters of loanwords within the integrated database.

We applied Word2vec – a word-embedding algorithm proposed in Mikolov et al. (2013), to vectorize our database data. This algorithm allows representing words in the text database as high-dimensional vectors (embeddings). These vectors are primarily used to measure the semantic similarity between words. Therefore, the greater the

Киосе М. И., Хулхачиева Ж. С., Изюмская-Капитонова В. В., Бармин А. В. Лексико-семантическая...
Kiose M. I., Khulkhachieva Zh. S., Izyumskaya-Kapitonova V.V., Barmin A.V. Lexical-semantic clustering...

72

similarity between word vectors, the more similar the words are semantically. This allows determining the closest neighbors of any word in the text database and analyze them as set words that form the context of a given word.

Next, the list with the neighbors for each of the loanwords was used to determine the qualia roles through semantic analysis of the collocates within the nominal groups (loanwords and collocates) and their distribution within the compiled database. We then determined the loanwords knowledge domains and their prevalence in the web database, which describes the direct knowledge transfer with Russian as a donor language and Kyrgyz as a recipient. Finally, we identified the qualia roles and their distribution of the target words as loanword collocates, which describes the indirect knowledge transfer.

4. Results and Discussion

In this Section we present the results which describe: 1) the knowledge domains representing the loanwords and the distribution of loanwords in the compiled web database, exposing the results of direct knowledge transfer: 2) the distribution of target words displaying high vector similarity with the loanwords in the compiled web database obtained through semantic clustering and qualia roles representing the referents of these loanwords in the web contexts, exposing the results of indirect knowledge transfer. The web database and the list of target words with their vector similarity indices as well as the list of target words with their qualia roles can be accessed at: https://osf.io/78u4m/ or https://doi.org/10.17605/OSF.IO/78U4M.

4.1. Direct knowledge transfer

On Step 1.1. a list of loanwords from Russian into the Kyrgyz language was compiled. The Dictionary of Loanwords edited by Kh. Karasaev (1986) contains 793 words borrowed from Russian, while the Dictionary of the Kyrgyz Language edited by K. Yudakhin (1985) contains extra 411 words, mostly graphological and phonological variants to the loanwords presented in

Karasaev (1986). Considering two input dictionaries, we further obtained a list of 1,204 loanwords from Russian as a donor language into Kyrgyz as recipient language. With the overall number of loanwords equal to 5,100 in Karasaev (1986), the ratio of loanwords from Russian (addressing only the items listed in this dictionary) is 15.55%. With the overall number of words in the Dictionary of the Kyrgyz Language edited by Yudakhin (1985) equal to 100,000, the ratio of words from Russian is 1.204%.

On Step 1.2. we performed the corpus search of loanwords (including their word forms) frequency in the two compiled databases, the Kyrgyz Web Text Corpus (Kyrgyzstan) featuring the language use up to 2019, and the Kyrgyz News Corpus. The analysis of loanwords frequency revealed considerable differences in their web representation; therefore, only 64 most frequent loanwords with frequency exceeding 800 in the base form in the Kyrgyz Web Text Corpus (Kyrgyzstan) were subjected to further analysis. To determine the borderline frequency in selecting these words, we adopted the following procedure. The overall frequency of 1,204 loanwords from Russian as a donor language into Kyrgyz as a recipient language in the Kyrgyz Web Text Corpus (Kyrgyzstan) in the base forms amounted to 197,827. The average word frequency was 208.24. Meanwhile, almost one third of the loanwords either did not appear in the corpus, e.g., *жаркеп* ('жаркое', 'roast'), *кечур* ('кучер', 'coachman'), *майыр* ('русский офицер', 'major'), *матике* ('дама', 'дамка', 'dame'), *нагирет* ('награда', 'award'), *обосол* ('новосел', 'new house dweller'), *памбаркут* ('панбархат', 'panne'), *селкебай* ('целковый', 'one rouble'), *тастаберне* ('удостоверение', 'certifying document'), *турдаден* ('трудодень', 'workday'), *тыпылыс* ('бязь тифлисская', 'calico'), *чомун* ('в темную', 'with eyes closed') *чыстай* ('чистый', 'аккуратный', 'clean'), *ыштан* ('штаны', 'панталоны', 'pantaloons'), etc., as seen mostly due to being the phonological and graphological

Научный результат. Вопросы теоретической и прикладной лингвистики. Т. 11, №4 2025
Research result. Theoretical and Applied Linguistics, 11 (4). 2025

73

variants. Additionally, a vast number of loanwords were infrequent (did not exceed 10), e.g., *арыстан* ('арестант', 'prisoner'), *агитатор* ('агитатор', 'agitator'), *аккорд* ('аккорд', 'chord'), *анархист* ('анархист', 'anarchist'), *балалайка* ('балалайка', 'balalaika'), *галифе* ('галифе', 'riding breeches'), *драп* ('драп', 'drape'), *колонизатор* ('колонизатор', 'colonizer'), *лексикография* ('лексикография', 'lexicography'), *материалист* ('материалист', 'materialist'), *пленум* ('пленум', 'plenum'), *шинель* ('шинель', 'overcoat'), etc. Three reasons for their infrequent use were identified: they either relate to the objects and notions which are not used in contemporary culture, not used in web discourse, or use other phonological and graphological word forms. Only 132 words out of 1,204 in the Kyrgyz Web Text Corpus (Kyrgyzstan) in the base form demonstrated the frequency which exceeded 400; with half of these being still infrequent, only the other half (disregarding the ones which were listed in the dictionary, however used only in their variants, e.g., *борум* ('форма', 'form')) was used for further analysis.

Next, we established the knowledge domains of 64 loanwords following the thematic criterion. Since we considered the loanwords lacking their context use, we adhered to their major dictionary meaning. These related to political, social, financial, cultural areas, as well as to calendar months, jobs, metrics, technology, evaluation spheres, establishments. Consequently, we can claim that both cultural and core borrowings (Haspelmath, 2009) appeared in this list. Next, we identified the distribution of these loanwords in the compiled web database, exposing the results of direct knowledge transfer. Their frequency in the Kyrgyz Web Text Corpus (Kyrgyzstan) in the base forms amounted to 119,476 (ranging from 800 to 13,500); in all forms – to 360,016. Their frequency in the Kyrgyz News Corpus in the base forms amounted to 173,399 (ranging from 411 to 18,034); in all forms – to 503,989. Considering the overall number of words in the Kyrgyz Web Text Corpus (Kyrgyzstan) equal to 36,581,130, we can claim that 13.108% of all words (tokens) functioning within web discourse appear from Russian as a donor language. Contrasted with the ratio of words from Russian listed in Yudakhin (1985), which is 1.204%, this number exceeding it more than 10 times shows that the borrowed words are highly active in contemporary discourse which proves that direct knowledge transfer has stimulated extensive conceptualization of the areas listed above.

Importantly, we ranged the frequency of these words to determine the key areas of direct knowledge transfer. As described above, the Kyrgyz Web Text Corpus (Kyrgyzstan) allowed for the search procedure, while pre-processing and processing was performed for the Kyrgyz News Corpus, given that the dataset itself was assembled through web scraping. Duplicates (880 items) were removed, and the data were cleaned of noise such as references to publishing outlets, tabulations, formatting artifacts, and copyright symbols. From the preprocessed dataset, web items containing at least one borrowed lexeme were extracted on the basis of the constructed dictionary. These texts were then segmented into sentences, and only sentences containing borrowings were retained in a new dataset. Each sentence was assigned a unique identifier, allowing the traceability of the extracted sentence back to its original news item in the main dataset. The resulting entries were stored in the form of id, lemma, word form, and the corresponding source sentence. All of the aforementioned steps, including the pre-processing of the raw dataset and the extraction of loanword-containing fragments, were implemented using Python and regular expressions.

Table 1 presents a list of loanwords extracted from two databases, showing the movement of loanwords from Russian (the donor culture) to Kyrgyz (the recipient culture). Their frequency (considering all word forms) in the compiled database is indicated.

Киосе М. И., Хулхачиева Ж. С., Изюмская-Капитонова В. В., Бармин А. В. Лексико-семантическая...
Kiose M. I., Khulkhachieva Zh. S., Izyumskaya-Kapitonova V.V., Barmin A.V. Lexical-semantic clustering...

74

Table 1. Frequency of loanwords (abs and ratio, %) in the compiled web database
Таблица 1. Частотность заимствованных слов в собранном веб-корпусе (абсолютные и относительные (в процентах) величины)

Loanword	Frequency abs (ratio)	Loanword	Frequency abs (ratio)	Loanword	Frequency abs (ratio)
Август	16906 (1.957)	Демократия	10005 (1.158)	Паспорт	8534 (0.988)
Автобус	6882 (0.797)	Депутат	66102 (7.651)	Пенсия	8380 (0.97)
Автор	9134 (1.057)	Директор	22297 (2.581)	План	11538 (1.335)
Актёр	5722 (0.662)	Документ	25391 (2.939)	Почта	2314 (0.268)
Актуалдуу	3593 (0.416)	Журналист	21734 (2.515)	Профессор	4320 (0.5)
Акция	10877 (1.259)	Интервью	1967 (0.228)	Радио	5907 (0.684)
Анализ	4442 (0.514)	Кадр	6026 (0.697)	Район	40096 (4.641)
Аппарат	9827 (1.137)	Канал	6793 (0.786)	Революция	3303 (0.382)
Апрель	22087 (2.556)	Картошка	2759 (0.319)	Сентябрь	18890 (2.186)
Архитектура	3299 (0.382)	Килограмм	6593 (0.763)	Такси	3600 (0.417)
Балл	5025 (0.582)	Кино	8626 (0.998)	Талант	5688 (0.658)
Банк	27020 (3.127)	Комитет	17549 (2.031)	Театр	5942 (0.688)
Батир	7085 (0.82)	Конкурс	3831 (0.443)	Телефон	17097 (1.979)
Белет	9774 (1.131)	Концерт	8634 (0.999)	Техника	21734 (2.515)
Бензин	2887 (0.334)	Машина	14133 (1.636)	Товар	9067 (1.049)
Бюджет	23872 (2.763)	Медаль	8859 (1.025)	Тонна	11012 (1.275)
Валюта	7190 (0.832)	Медицина	26357 (3.051)	Транспорт	15497 (1.794)
Врач	4183 (0.484)	Метр	19682 (2.278)	Факт	12692 (1.469)
Газ	12858 (1.488)	Митинг	8429 (0.976)	Февраль	15967 (1.848)
Гектар	6009 (0.695)	Операция	15714 (1.819)	Экономика	48761 (5.644)
Градус	17463 (2.021)	Парламент	41436 (4.796)		
Декабрь	6053 (0.701)	Партия	28561 (3.306)		

Table 1 shows that several loanwords demonstrate the highest frequency exceeding 2 per cent, which are *апрель* ('April'), *банк* ('bank'), *бюджет* ('budget'), *градус* ('degree'), *депутат* ('depute'), *директор* ('director'), *документ* ('document'), *журналист* ('journalist'), *комитет* ('committee'), *медицина* ('medicine'), *метр* ('meter'), *парламент* ('parliament'), *партия* ('party'), *район* ('region'), *сентябрь* ('September'), *техника* ('technics'), *экономика* ('economics'). These loanwords clearly demonstrate the main areas of knowledge that are transferred directly in web discourse: banking and finance; management and documentation; technology and economics; city management; and measurement and time control.

4.2. Indirect knowledge transfer

On Step 2, applying Word2vec – a word-embedding algorithm, we built word2vec model using *word2vec* R-package v.0.4.0. in RStudio v.2024.12.1+563 with R v.4.3.2. We set the following parameters of the model: type of word2vec algorithm (type) = skip-gram; dimensions of word vectors (dim) = 300; number of times a word should occur to be considered as part of the training vocabulary (min_count) = 1; the rest of the parameters were set as default. We then trained our word2vec model on the full database of web samples. We used the "predict" function in the *word2vec* R-package to determine the top five nearest neighbors for each of the 825 loanwords based on their vector similarity. Using the skip-gram word2vec model trained on the web database,

Научный результат. Вопросы теоретической и прикладной лингвистики. Т. 11, №4 2025
Research result. Theoretical and Applied Linguistics, 11 (4). 2025

75

we collected the five nearest neighbors for each of the 825 loanwords based on their vector similarity in the 300-dimensional space. This resulted in a total of 4,125 word neighbors. We also collected 4,125 numerical scores of vector similarity between each of the 825 loanwords and their five nearest neighbors. We then constructed a matrix consisting of three columns: source (loanwords), target (nearest neighbors of the loanwords), and weight (scores of vector similarity between the loanwords and their nearest word neighbors) – for further semantic networks construction.

To visualize our results as a semantic network, we uploaded the three-column matrix to Gephi v.0.10.1. – an open-source software for network visualization and analysis. The parameters of network were set as follows: undirected, with no specific edge merging strategy. Thus, the resulting network comprised a total of 3,397 nodes and 4,125 edges. We then used the Fruchterman-Reinhold layout algorithm (which places closely connected nodes closer to each other and poorly connected nodes further apart) to better optimize the aesthetic appearance of the resulting network.

For better network visualization, the words from the "source" column (loanwords) were colored red, while all other nodes in the network were gray. To better visualize the strength of the semantic connections between the nodes of the network, the edges were marked in color depending on their weight (vector similarity coefficients) values: so, the edges with a higher weight were colored purple, the edges with an average weight were green, and the edges with a lower weight were yellow. The resulting network and its fragment are shown in Figure 2.

Figure 2. Left: A semantic network representing words and their semantic relationships. Red nodes indicate words borrowed from Russian into Kyrgyz, while gray nodes show their nearest neighbors identified by Word2vec. Edge color indicates the degree of semantic similarity, with purple for strong and yellow for weak connections. Right: A fragment of this network.
Рисунок 2. Слева: Семантическая сеть, состоящая из слов и семантических связей между ними. Красные узлы означают заимствованные слова из русского в кыргызский язык, серые – их ближайших соседей, определенных с помощью Word2vec. Цвет ребер определяется степенью семантической близости: более сильные связи обозначаются фиолетовым цветом, более слабые – желтым. Справа: Фрагмент этой сети.
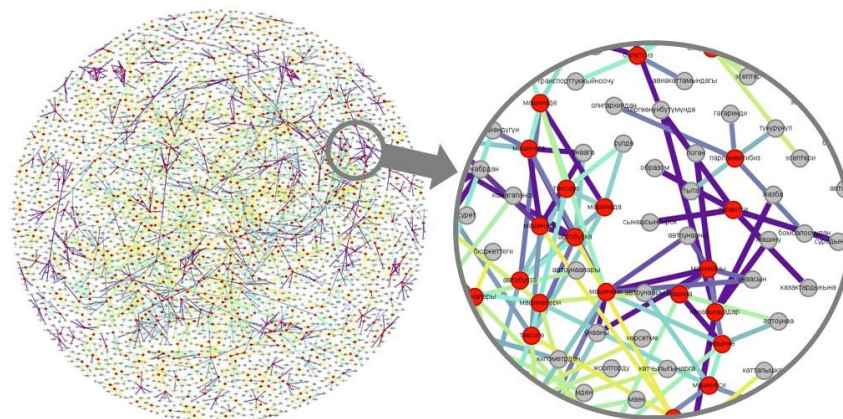
Next, the list with the nearest neighbors for each of the 825 loanwords was used to determine the qualia roles through semantic analysis of the collocates within the groups (loanwords and collocates) and their distribution within the compiled database.

Киосе М. И., Хулхачиева Ж. С., Изюмская-Капитонова В. В., Бармин А. В. Лексико-семантическая...
Kiose M. I., Khulkhachieva Zh. S., Izyumskaya-Kapitonova V.V., Barmin A.V. Lexical-semantic clustering...

76

Below, we specified the procedure which helped identify the qualia roles of collocates within the groups (loanwords and their collocates). Four key principles were observed. All word forms of each of the 64 highly frequent loanwords were considered to identify the vector similarity with the left and right collocates appearing in the same sample (the distance between the loanword and the collocates was not observed). To explore the distribution of qualia roles, only the collocates with vector similarity index exceeding 0.6 with 5 most frequent collocates for each word form were addressed. Overall, 4,126 collocates serving as target words were subjected to semantic analysis. In case the target word and the loanword were one and the same word, the target word was not considered. In most cases, each loanword form displayed contingency with several target words. For instance, the loanword *батир* ('flat') with *батир, батирге, батирде, батирден, батирди, батирдин, батири, батирим, батирлер, батирлери, батирлерибиз, батирлүү, батирсиз* demonstrated contingency with 55 target words. Among them the word form *батирлери* had 5 collocates, *батири* ('flat') (0.79), *мүлктөрү* ('property') (0.75), *мүлкү* ('property') (0.74), *эмереги* ('furniture') (0.73), *элиталык* ('elite') (0.73). Four target words were consequently addressed in identifying the qualia roles, where *мүлктөрү* ('property') and *мүлкү* ('property') were attributed to the Formal role displaying taxonomic information about the lexical item (Pustejovsky, 1991; Pustejovsky, Jezek, 2016), where in the hierarchy the property domain relates to superordinate category contrasted to the basic categorization of the domain of flat. Meanwhile, *эмереги* ('furniture') was attributed to the subordinate level, since in the hierarchy of semantic representation the furniture domain displays the part-whole relations contrasted to the domain of flat; consequently, we identify its qualia role as Constitutive (ibid.). The target word *элиталык* ('elite') encodes the information on its sphere of use (who will use it); therefore, we attribute the qualia role as Telic displaying purpose and function (ibid.). Consequently, the qualia structure of *батирлери* is as follows (Figure 3):

Figure 3. Qualia-roles attributed to *батирлери*
Рисунок 3. Квали-роли слова *батирлери*

$$\begin{bmatrix} батирлери \\ \text{QUALIA} = \begin{bmatrix} F = \text{property} \\ C = \text{furniture} \\ T = \text{elite} \end{bmatrix} \end{bmatrix}$$

The described procedure was employed to attribute the qualia roles of all target words. Finally, the distribution of qualia roles was established describing the indirect knowledge transfer. Semantic analysis of 4,126 target words revealed that 1,004 words serve to construe the qualia system of loanwords in the explored samples. The rest 3,122 words relate to one of the following groups: a) the word forms of the considered loanword; b) the words which are co-hyponyms to the loanword (e.g., *августка* ('August') and *сентябрга* ('September'); c) the words which are nominal components of the action chains with the loanword, e.g., *авторлор* ('author') and *окурмандар* ('readers'); d) the words which do not form the arguments of loanwords, e.g., *концертим* ('concert') and *пикиримде* ('in my opinion'); e) the words whose transition type cannot be unmistakably identified, e.g., *концертиң* ('concert') and *сунуштарын* ('their proposals'). As shown in (Pustejovsky, Jezek, 2016), qualia roles can be attributed to the sentence components in case a type of coercion (a semantic operation that converts an argument to the type that is

Научный результат. Вопросы теоретической и прикладной лингвистики. Т. 11, №4 2025
Research result. Theoretical and Applied Linguistics, 11 (4). 2025

77

expected by the function, where it would otherwise result in a type error) can be identified. We presumed that manual annotation of all the samples might extend and specify the number of coerced relations (specifically, agentic relations, see Limitations section below); meanwhile, in this paper we addressed only the ones which could be unambiguously attributed.

Further analysis showed that 1,004 target words which displayed some type of coercion to the loanwords had the following distribution of qualia. The Formal role distinguishing the referent of the loanword within a larger domain appeared most frequently presented (365 examples), e.g., in *августан* ('August') and *быйылтан* ('since this year'), *театрды* ('theatre') and *чыгармачылыгын* ('creation', 'creativity'), *товарда* ('product') and *сорту* ('type', 'kind'), *тоннада* ('tonne') and *татымалдар* ('spices'). Out of all functions of the Formal role identified in (Pustejovsky, Jezek, 2016), i.e., orientation, magnitude, shape, dimensionality, color, position, in the compiled database these were dimensionality and magnitude, since the target words often presented the components related metonymically to the loanwords construing the referent within the subdomain of action of the referent of the target word. The second frequent role was Agentive (290 examples) which displayed the relations involved in the origin or 'bringing about' of a referent, e.g., in *авторлоштуруу* ('author') and *каттатылат* ('will be registered'), *акцияга* ('actions') and *нааразылыкка* ('discontent'), *медалдары* ('medal') and *сыйланган* ('awarded'). Considering the functions of the Agentive role (creator, artifact, natural kind, causal chain) identified in (Pustejovsky, Jezek, 2016), we found that artifact and natural kind functions appeared most frequent. Artifact function was observed when coercion was performed by a transition event, e.g., in *парламентке* ('parliament') and *бийликке* ('to power') which is created by the transition event of taking control. Natural kind function is observed when a verb takes a

loanword as a complement, e.g., in *пландаштыруу* ('plan') and *өркүндөтүү* ('improve'). The third frequent role is Constitutive (284 examples). Within the functions of this role considering the relations of the referent and its constituents or parts Pustejovsky, Jezek (2016) specified three functions of material, weight, parts and component elements. In the compiled database, parts and component elements were mostly found, e.g., in *каналыңыздар* ('channel') and *жазба* ('record') or in *кинодо* ('cinema') and *образдарды* ('images'). The least frequent role was Telic (65 examples); out of two key functions specified in (Pustejovsky, Jezek, 2016), which are purpose that an agent has in performing an act and built-in function or aim that specifies certain activities, only *автор* ('author') and *чыгарма* ('book').

4.3. Discussion

The results obtained allow us to validate the potential of the lexical-semantic clustering method when applied to the exploration of discourse and culture. As this method helped to identify key knowledge domains representing indirect knowledge transfer in web discourse, we can conclude that, in addition to exploring semantic relations (Troyer et al., 1997; Kuhn et al., 2007), it is effective in exploring cultural relations based on a lexical criterion. The results prove that the lexical clustering decision is efficient in terms of searching and retrieving the knowledge domains demonstrating the areas of culture integration. In this case apart from exploring the clustering decision in terms of the cluster size, the types of alternating elements from different subcategories in the frequency of cluster switching (Troyer et al., 1997), additional methods should be elaborated to reveal the extent and knowledge domains of knowledge transfer in culture integration. As shown in the definition of the clustering method, it helps reveal the semantic subcategories represented by the sets of items as source artifacts (Kuhn et al., 2007); meanwhile, the compiled sets do demonstrate

Киосе М. И., Хулхачиева Ж. С., Изюмская-Капитонова В. В., Бармин А. В. Лексико-семантическая…
Kiose M. I., Khulkhachieva Zh. S., Izyumskaya-Kapitonova V.V., Barmin A.V. Lexical-semantic clustering…

78

the distribution of collocates of the loanwords, but do not demonstrate either the knowledge domains which must be additionally identified, or the extent of their representation. Therefore, in line with other frameworks, e.g., in cognitive psychology (Shaffer et al., 2024) or foreign language acquisition (Pérez-Serrano et al., 2022), we integrate other methods, here the method of qualia roles, to reveal the specifics of cultural integration through knowledge transfer. The results prove that the notion of cultural knowledge transfer is potent in revealing the culture integration by means of quantitative methods apart from exploring it qualitatively (Espagne, 1999). Additionally, they show that the types or ways of knowledge transfer (Postovalova, 2016; Iriskhanova, Kiose, 2016), here direct and indirect, reveal the attribution of knowledge domains mediated by culture integration.

It is noteworthy that the method of qualia roles was not previously adopted to exploring cultural knowledge transfer. Being a method of generative linguistics (generative lexicon) framework and developed by Pustejovsky (1991, 2006), it is commonly used to explore the semantics of events in clauses. It specifies the distribution of four qualia roles, formal, constitutive, telic and agentive, in the construal of referents within the clause; and further on the qualia structure framework was advanced in computational and ontology studies (Cimiano, Wenderoth, 2005; Pustejovsky et al., 2006, Jezek et al. 2009). Meanwhile, the framework has recently been applied to cross-cultural semantic studies (Lee et al., 2010; Song and Qiu, 2013). The current study builds on this to address the issue of cultural integration.

To develop this methodological decision, following Myers-Scotton (2002, 2006) and Haspelmath (2009), we explored culture integration through lexical systems integration found in the frequency and thematic areas of loanwords from the donor language (Russian) into the recipient language (Kyrgyz). As opposed to the traditional approaches considering loanwords within the

processes of adoption or imposition, replacement or coexistence (Winford, 2005; Haspelmath, 2009), determining them as either cultural or core borrowings (Haspelmath, 2009), revealing cultural, genealogical and linguistic reasons of borrowing (Thomason, 2001), in this study we address only one process of insertion since we do not explore the types or distribution of loanwords, but the knowledge domains which underlie culture integration.

The paper develops and advances a novel framework to exploring culture integration through lexical systems integration in loanwords which considers not only direct knowledge transfer proposed in (Thomason, 2001; Winford, 2005; Haspelmath, 2009), but indirect knowledge transfer which is identified through the knowledge domains and distribution of the collocates of the loanwords in discourse. Two knowledge transfer types complement each other in attaining the overall specificity of culture integration in reference to the two languages contrasted. Out of 1,204 identified loanwords from Russian as a donor language into Kyrgyz as a recipient language, 64 loanwords were selected for further analysis due to their extensive use in the compiled databases in web discourse. The results show that direct knowledge transfer is manifested in major knowledge domains in web discourse – banking and finance, managing and documentation, technics and economics, city management, measurement and time control. Therefore, the results specify the areas of Russian – Kyrgyz cross-cultural communication addressed in (Derbisheva, 2017, 2019; Kambaralieva, Sternin, 2019). Importantly, the results prove that despite the changing vitality status of Russian in Kyrgyzstan and high migration rate (Shipilov, 2018), the extent of direct knowledge transfer is rather high. As shown above, more than 13 per cent of all words functioning within web discourse appear from Russian as a donor language, although the dictionary rate of loanwords is only 1.2 per cent. Additionally, the results show that both cultural and core

Научный результат. Вопросы теоретической и прикладной лингвистики. Т. 11, №4 2025
Research result. Theoretical and Applied Linguistics, 11 (4). 2025

79

borrowings (Haspelmath, 2009) appear in the list of most frequent loanwords. Overall, the results describing the direct knowledge transfer suffice to claim that culture integration (from Russian into Kyrgyz) has stimulated the development of several areas which are currently extensively circulating in web discourse.

While direct knowledge transfer previously received attention in the studies, exploring indirect knowledge transfer is a novel research direction introduced in this paper, although the fact that it can be addressed through the knowledge domains mediated by the immediately inserted or imposed domains is discussed in the studies (Winford, 2005; Haspelmath, 2009). The problem of identifying knowledge domains in indirect knowledge transfer is addressed through semantic clustering and the implementation of the qualia roles framework. This analysis examined 4,126 target words alongside their vector similarity index as loanword collocates. Further analysis revealed that 1,004 target words displayed some form of coercion in relation to the loanwords exposed in the distribution of qualia. These words had a highly represented formal role, a sufficiently represented agentic role and a poorly represented telic role. The results prove that in conceptualizing the newly introduced referent, it is mostly construed within a larger knowledge domain attesting its dimensionality and magnitude where its artifact and natural kind functions are foregrounded with reference to parts and component elements. Consequently, lexical-semantic clustering, integrated into a cognitive framework using the qualia roles method of the generative lexicon, helps to determine knowledge domains and qualia role distribution in web contexts with loanwords. The detected way of indirect knowledge transfer serves to shape the cognitive process of adoption and interiorization of an alien culture into the native culture. Although we cannot presume that this process has a universal character, we still can claim that this was the way the donor culture of Russia affected the recipient culture of Kyrgyzstan through lexical systems integration, at least within web discourse.

5. Limitations

In line with all studies adopting a complex procedure and identifying the regulations in the use of corpus data, the present study has its limitations.

The first limitation relates to the dates on which the etymological dictionaries used for the study were collected and published. As the only authoritative Kyrgyz dictionaries specifying the language of origin of loanwords are those compiled by Karasaev (1986) and Yudakhin (1985), we submitted the loanwords selected from these dictionaries for further corpus and clustering analysis. However, we acknowledge that loanwords from Russian may have entered the Kyrgyz language since the publication of these dictionaries. Although several articles and theses have identified some of these loanwords, we could not rely on these data as the loanwords were not systematically presented. Therefore, no unified procedure for identifying their etymology could be established.

The second limitation relates to how the corpus data is presented. While the conventional standard for presenting corpus data is to specify IPM indices, these could not be subjected to lexical-semantic clustering. Following the example of other studies that adopted the aforementioned quantitative approach, we opted for a more suitable standard of data presentation. Additionally, as data from two different-sized corpora were initially combined to form a common database, we did not consider it appropriate to present the unified data in IPM indices. For these reasons, absolute and ratio data are presented in Table 1.

The third limitation relates to the procedure and results of identifying knowledge domains, specifically the thematic attribution of domains in direct transfer and qualia roles in indirect transfer. In the first case, thematic attribution was carried out by

Киосе М. И., Хулхачиева Ж. С., Изюмская-Капитонова В. В., Бармин А. В. Лексико-семантическая…
Kiose M. I., Khulkhachieva Zh. S., Izyumskaya-Kapitonova V.V., Barmin A.V. Lexical-semantic clustering…

80

considering only the dictionary meaning of loanwords, so only the first meaning listed was taken into account. In the second case, due to the huge number of context samples in the compiled database, there was only a slight chance of identifying the qualia roles in context. Therefore, we performed the procedure considering the coercion relation of two words, with the first being the loanword form and the second the collocate form. While the formal, constitutive and telic roles were easily identified in most cases, this was not the case for the agentive role, since agency was frequently found in the semantics of nouns and adjectives. It was decided to include these cases in the agentive sample stock. A further manual, systemic search of the entire database, which consists of 864,005 samples, might have clarified the situation; however, we abstained from doing so at this stage.

6. Final remarks

The research on exploring direct and indirect knowledge transfer in culture integration through semantic clustering extends the application of this research method. As opposed to its customary use in identifying the semantic similarity of words through their distribution in context, in this study it serves to reveal the knowledge domains affected by cultural knowledge transfer; therefore, it serves the needs of cognitive and cultural areas.

Additionally, the study contributes to developing the procedures of cognitive and computational semantics in exploring the cultural knowledge transfer through the qualia roles method. Being developed and integrated within generative lexicon, it can be efficient in exploring cultural semantics, at least in the areas where referent construal is involved.

The research perspectives are seen primarily in contrasting the distribution of knowledge domains mediated by direct and indirect transfer in different languages and different databases with the aim of identifying universal and culturally specific regulations in culture integration processes exposed in

lexicon. Meanwhile, in attribution to the languages which do not have thoroughly developed databases (like the Kyrgyz language in the present study), the preliminary research procedure should also include the database balancing decisions. As shown, the existing corpus description, which was submitted for analysis, did not provide data on how balanced it was in terms of the number and scope of web resources, although each elicited context was attributed to a resource. Secondly, in this study, we only analysed highly frequent loanwords; however, less frequent loanwords can also contribute to the distribution of thematic domains and qualia roles in knowledge transfer. The developed method may be less efficient in obtaining these data since the number of contexts will be too limited for corpus and word2vec clustering analysis. Therefore, incorporating other methods, including clustering decisions for smaller datasets, is necessary to explore both frequent and infrequent loanword distribution.

Overall, the study makes a contribution to the area of culture integration, since it advances a method of determining the extent of discourse system mediation in the recipient culture by that of the donor culture through a lexical criterion of loanwords. Further research can help range the mediation effects of different criteria in the process of culture integration considering among other factors the factor of transfer directionality.

References
Cimiano, P., Wenderoth, J. (2005). Automatically Learning Qualia Structures from the Web. In T. Baldwin, A. Korhonen, A. Villavicencio (Eds.), *Proceedings of the ACL workshop on Deep Lexical Acquisition*. Ann Arbor, Michigan, 28–37. *(In English)*

Copestake, A., Briscoe, T. (1995). Semi-Productive Polysemy and Sense Extension, *Journal of Semantics*, 12 (1), 15–67. *(In English)*

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990). Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, *41*(6), 391–407. *(In English)*

Научный результат. Вопросы теоретической и прикладной лингвистики. Т. 11, №4 2025
Research result. Theoretical and Applied Linguistics, 11 (4). 2025

81

Demyankov, V. Z. (2016). Language techniques of knowledge transfer, in Feschenko, V. V. (ed.) *Linguistics and semiotics of cultural transfers: methods, principles, technologies*. Moscow, Cultural revolution, 61–85. *(In Russian)*

Derbisheva, Z. K. (2017). *Language and ethnos*. Moscow, Flinta. *(In Russian)*

Derbisheva, Z. K. (2019). *The basics of linguocognitive exploring languages in contrast*. Moscow, Flinta. *(In Russian)*

Feschenko, V. V., Bochaver, S. Yu. (2016). Theory of cultural transfers: from translation studies – through cultural studies – to the theory of language, in Feschenko, V. V. (ed.) *Linguistics and semiotics of cultural transfers: methods, principles, technologies*. Moscow, Cultural revolution, 5–35. *(In Russian)*

Gallagher, S., Hutto, D. D. (2008). Understanding others through primary interaction and narrativepractice, in J. Zlatev, T. Racine, C. Sinha and E. Itkonen (eds.), *The shared mind: Perspectives of intersubjectivity*. Amsterdam, John Benjamins, 17–38. *(In English)*

Haspelmath, M. (2009). Lexical borrowings: Concepts and issues, in M. Haspelmath, U. Tadmor (eds.). *Loanwords in the world's languages*. Berlin, De Gruyter, 35–54. *(In English)*

Iriskhanova, O. K., Kiose, M. I. (2016). Technologies of interdisciplinary terms transfer into language studies, in Feschenko, V. V. (ed.) *Linguistics and semiotics of cultural transfers: methods, principles, technologies*. Moscow, Cultural revolution, 151–180. *(In Russian)*

Jain, A. K., Murty, M. N., Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys* (CSUR), 31(3), 264–323. *(In English)*

Jezek, E., Quochi, V., Calzolari, N. (2009). Relevance of Qualia Relations in Coercive Contexts: Conference paper, 5th International Conference on Generative Approaches to the Lexicon. URL: https://www.researchgate.net/publication/228960165_Relevance_of_Qualia_Relations_in_Coercive_Contexts. (Accessed 18 August 2015). *(In English)*

Kambaralieva, U. D., Sternin, I. A. (2021). *Russian and Kyrgyz communicative behaviour*. Voronezh, Ritm. *(In Russian)*

Kuhn, A., Ducasse, S., Gîrba, T. (2007). Semantic clustering: Identifying topics in source code, *Information and Software Technology*, 49 (3), 230–243. URL: https://doi.org/10.1016/j.infsof.2006.10.017 (Accessed 25 July 2015). *(In English)*

Langacker, R. W. (2007). Cognitive grammar, in D. Geeraerts, H. Cuyckens (eds.), *The Oxford handbook of cognitive linguistics*. Oxford, Oxford University Press, 421–507. *(In English)*

Lee, C., Chang, C., Hsu, W., Hsieh, S. (2010). Qualia modification in noun-noun compounds: A cross language survey, *Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING-2010)*, 379–390. *(In English)*

Lee, S., Kim, H., Hwang, J., Park, E. and Ok, J. (2025). Efficient Latent Semantic Clustering for Scaling Test-Time Computation of LLMs. URL: https://doi.org/arXiv preprint arXiv:2506.00344. *(In English)*

Mikolov, T., Chen, K., Corrado, G, Dean, J. (2013). Efficient estimation of word representations in vector space. URL: https://doi.org/arXiv preprint arXiv:1301.3781. *(In English)*

Myers-Scotton, C. (2002). *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford, Oxford University Press. *(In English)*

Myers-Scotton, C. (2006). *Multiple voices: An introduction to bilingualism*. Malden, MA, Blackwell. *(In English)*

Pérez-Serrano, M., Nogueroles-López, M. and Duñabeitia, J. A. (2022). Effects of semantic clustering and repetition on incidental vocabulary learning. *Frontiers in Psychology*, 13, 997951. *(In English)*

Postovalova, V. I. Routes and principles of knowledge transfer in the humanities, in Feschenko, V. V. (ed.) *Linguistics and semiotics of cultural transfers: methods, principles, technologies*. Moscow, Cultural revolution, 36–60. *(In Russian)*

Pustejovsky, J. (1991) The generative lexicon, *Computational Linguistics*, 17 (4), 409–441. *(In English)*

Pustejovsky, J. (2006). Generative lexicon, *Encyclopedia of language and linguistics*. Amsterdam, Elsevier, 138–147. *(In English)*

Pustejovsky, J., Hanks, P., Rumshisky, A. (2004). Automated induction of sense in context, *COLING*, 924–931. *(In English)*

Pustejovsky, J., Jezek, E. (2016). A Guide to Generative Lexicon Theory. Oxford, Oxford University Press. *(In English)*

Киосе М. И., Хулхачиева Ж. С., Изюмская-Капитонова В. В., Бармин А. В. Лексико-семантическая...
Kiose M. I., Khulkhachieva Zh. S., Izyumskaya-Kapitonova V.V., Barmin A.V. Lexical-semantic clustering...

82

Shaffer, C., Andreano, J. M., Touroutoglou, A., Barrett, L. F., Dickerson, B. C. and Wong, B. (2024). Semantic clustering during verbal episodic memory encoding and retrieval in older adults: One cognitive mechanism of superaging. Brain Sciences, 14 (2), 171. *(In English)*

Shipilov, A. V. (2018). *Russian – Kyrgyzstan: the historical experience of cross-cultural discourse construal (second half of XIX – XXI c.).* Bishkek. *(In Russian)*

Siew, C. S., Wulff, D. U., Beckage, N. M., Kenett, Y. N. (2019). Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*, 2019 (1), 2108423. *(In English)*

Song, Z. and Qiu, L. (2013). Qualia relations in Chinese nominal compounds containing verbal elements, *International Journal of Knowledge and Language Processing*, 4 (1), 1–15. *(In English)*

Talmy, L. (2000). Toward a cognitive semantics. Vol. 2, *Typology and process in concept structuring*. Cambridge, MA, MIT Press. *(In English)*

Tinkham, T. (1993). The effects of semantic clustering on the learning of second language vocabulary. *System*, 21, 371–380. *(In English)*

Thomason, S. G. (2001). Language contact. Washington, D.C.: Georgetown University Press. *(In English)*

Troyer, A. K., Moscovitch, M., Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology,* 11 (1), 138–146. *(In English)*

Verhagen, A. (2007). Construal and perpectivization, in D. Geeraerts, H. Cuyckens (Eds.), *The Oxford handbook of cognitive linguistics*. Oxford, Oxford University Press, 48–80. *(In English)*

Winford, D. (2005). Contact-induced changes: Classification and processes. *Diachronica*, 22 (2), 373–427. *(In English)*

Witschard, D., Jusufi, I., Martins, R. M., Kucher, K., Kerren, A. (2022). Interactive optimization of embedding-based text similarity calculations, *Information Visualization*, 21 (4), 335–353. URL: https://doi.org/10.1177/14738716221114372 (Accessed 25 September 2025). *(In English)*

Corpus materials

Karasaev, Kh. K. (1986). The Dictionary of Loanwords: 5100 words. Frunze: Kyrgyz Sovet. *(In Kyrgyz)*

Kyrgyz Web Text Corpus. Leipzig Corpora, available at: https://wortschatz.uni-leipzig.de/en (Accessed 6 September 2025). *(In English)*

Python Documentation, 5. Data Structures (2025). Python Software Foundation. URL: https://docs.python.org/3/tutorial/datastructures.html (Accessed 25 September 2025). *(In English)*

The Kyrgyz News Corpus dataset (2025). Hugging Face AI community. URL: https://huggingface.co/datasets/the-cramer-project/Kyrgyz_News_Corpus#kyrgyz_news_corpus (Accessed 6 September 2025). *(In English)*

Yudakhin, K. K. (1985). Kyrgyz-Russian Dictionary. Frunze: Glavnaya redaktsiya Kirgyzskoy Sovetskoy Entsiklopedii. *(In Russian)*

Список литературы

Cimiano P., Wenderoth J. Automatically Learning Qualia Structures from the Web / T. Baldwin, A. Korhonen, A. Villavicencio (Eds.) // Proceedings of the ACL workshop on Deep Lexical Acquisition. 2005. Pp. 28–37.

Copestake A., Briscoe T. Semi-Productive Polysemy and Sense Extension // Journal of Semantics. 1995. 12 (1). Pp. 15–67.

Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K. and Harshman R. Indexing by latent semantic analysis // Journal of the American society for information science. 1990. Vol. 41 (6). Pp. 391–407.

Демьянков В. З. Языковые техники «трансфера знаний» // Лингвистика и семиотика культурных трансферов: методы, принципы, технологии / Отв. ред. В. В. Фещенко. М.: Культурная революция, 2016. С. 61–85.

Дербишева З. К. Язык и этнос. М.: Флинта, 2017.

Дербишева З. К. Основы лингвокогнитивного сравнения языков. М.: Флинта, 2019.

Фещенко В. В., Бочавер С. Ю. Теория культурных трансферов: от переводоведения – через cultural studies – к теоретической лингвистике // Лингвистика и семиотика культурных трансферов: методы, принципы, технологии / Отв. ред. В. В. Фещенко. М.: Культурная революция, 2016. С. 5–35.

Gallagher S., Hutto D. D. Understanding others through primary interaction and

Научный результат. Вопросы теоретической и прикладной лингвистики. Т. 11, №4 2025
Research result. Theoretical and Applied Linguistics, 11 (4). 2025

83

narrativepractice / J. Zlatev, T. Racine, C. Sinha and E. Itkonen (Eds.) // The shared mind: Perspectives of intersubjectivity. Amsterdam: John Benjamins, 2008. Pp. 17–38.

Haspelmath M. Lexical borrowings: Concepts and issues / M. Haspelmath, U. Tadmor (Eds.). Loanwords in the world's languages. Berlin: De Gruyter, 2009. Pp. 35–54.

Ирисханова О. К., Киосе М. И. Технологии трансфера междисциплинарных терминов в лингвистику // Лингвистика и семиотика культурных трансферов: методы, принципы, технологии / Отв. ред. В. В. Фещенко. М.: Культурная революция, 2016. С. 151–180.

Jain A. K., Murty M. N., Flynn P. J. Data clustering: a review // ACM computing surveys (CSUR). 1999. Vol. 31 (3). Pp. 264–323.

Jezek E., Quochi V., Calzolari N. Relevance of Qualia Relations in Coercive Contexts: Conference paper // 5th International Conference on Generative Approaches to the Lexicon. 2009. URL: https://www.researchgate.net/publication/228960165_Relevance_of_Qualia_Relations_in_Coercive_Contexts. (Accessed 18 August 2025)

Камбаралиева У. Д., Стернин И. А. Русское и киргизское коммуникативное поведение. Воронеж: Ритм, 2021.

Kuhn A., Ducasse S., Gîrba T. Semantic clustering: Identifying topics in source code // Information and Software Technology. 2007. Vol. 49 (3). Pp. 230–243. URL: https://doi.org/10.1016/j.infsof.2006.10.017 (access date: 25.07.2025)

Langacker R. W. Cognitive grammar / D. Geeraerts, H. Cuyckens (Eds.) // The Oxford handbook of cognitive linguistics. Oxford: Oxford University Press, 2007. Pp. 421–507.

Lee C., Chang C., Hsu W., Hsieh S. Qualia modification in noun-noun compounds: A cross language survey // Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING-2010). 2010. Pp. 379-390.

Lee S., Kim H., Hwang J., Park E. and Ok J. Efficient Latent Semantic Clustering for Scaling Test-Time Computation of LLMs. 2025. URL: https://doi.org/arXiv preprint arXiv:2506.00344 (access date: 25.07.2025).

Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. URL: https://doi.org/arXiv preprint arXiv:1301.3781. 2013 (Accessed 25 July 2025).

Myers-Scotton C. Contact linguistics: Bilingual encounters and grammatical outcomes. Oxford: Oxford University Press, 2002.

Myers-Scotton C. Multiple voices: An introduction to bilingualism. Malden, MA: Blackwell, 2006.

Pérez-Serrano M., Nogueroles-López M. and Duñabeitia J. A. Effects of semantic clustering and repetition on incidental vocabulary learning // Frontiers in Psychology. 2022. Vol. 13. Article number: 997951.

Постовалова В. И. Пути и принципы трансферизации знания в гуманитарных науках // Лингвистика и семиотика культурных трансферов: методы, принципы, технологии / Отв. ред. В. В. Фещенко. М.: Культурная революция, 2016. С. 36–60.

Pustejovsky J. The generative lexicon // Computational Linguistics. 1991. Vol. 17 (4). Pp. 409–441.

Pustejovsky J. Generative lexicon // Encyclopedia of language and linguistics. Amsterdam: Elsevier, 2006. Pp. 138–147.

Pustejovsky J., Hanks P., Rumshisky A. Automated induction of sense in context // COLING. 2004. Pp. 924–931.

Pustejovsky J., Jezek E. A Guide to Generative Lexicon Theory. Oxford: Oxford University Press, 2016.

Shaffer C., Andreano J. M., Touroutoglou A., Barrett L. F., Dickerson B. C. and Wong B. Semantic clustering during verbal episodic memory encoding and retrieval in older adults: One cognitive mechanism of superaging // Brain Sciences. 2024. Vol. 14 (2). Article number: 171.

Шипилов А. В. Россия – Кыргызстан: исторический опыт формирования межкультурного дискурса (вторая половина XIX – XXI в.). дис. … д-ра истор. наук. Бишкек: Кыргызско-Российский Славянский университет, 2018.

Siew C. S., Wulff D. U., Beckage N. M., Kenett Y. N. Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics // Complexity. 2019. Vol. 1. Article number: 2108423.

Song Z. and Qiu L. Qualia relations in Chinese nominal compounds containing verbal elements // International Journal of Knowledge and Language Processing. 2013. Vol. 4 (1). Pp. 1–15.

Киосе М. И., Хулхачиева Ж. С., Изюмская-Капитонова В. В., Бармин А. В. Лексико-семантическая...
Kiose M. I., Khulkhachieva Zh. S., Izyumskaya-Kapitonova V.V., Barmin A.V. Lexical-semantic clustering...

84

Talmy L. Toward a cognitive semantics. Vol. 2. Typology and process in concept structuring. Cambridge, MA: MIT Press, 2000.

Tinkham T. The effects of semantic clustering on the learning of second language vocabulary // System. 1993. Vol. 21. Pp. 371–380.

Thomason S. G. Language contact. Washington D.C.: Georgetown University Press, 2001.

Troyer A. K., Moscovitch M., Winocur G. Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults // Neuropsychology. 1997. Vol. 11 (1). Pp. 138–146.

Verhagen A. Construal and perpectivization / D. Geeraerts, H. Cuyckens (Eds.) // The Oxford handbook of cognitive linguistics. Oxford: Oxford University Press, 2007. Pp. 48–80.

Winford D. Contact-induced changes: Classification and processes // Diachronica. 2005. Vol. 22 (2). Pp. 373–427.

Witschard D., Jusufi I., Martins R. M., Kucher K., Kerren A. Interactive optimization of embedding-based text similarity calculations // Information Visualization. 2022. Vol. 21 (4). Pp. 335–353. URL: https://doi.org/10.1177/14738716221114372 (Accessed 25 September 2015)

Материалы исследования

Карасаев Х. К. Словарь заимствованных слов: 5100 слов. Фрунзе: Киргизский Совет, 1986.

Юдахин К. К. Киргизско-русский словарь. Фрунзе: Главная редакция Киргизской Советской Энциклопедии, 1985.

Kyrgyz Web text corpus. Leipzig Corpora. URL: https://wortschatz.uni-leipzig.de/en (access date: 6.09.2015)

Python Documentation, 5. Data Structures. Python Software Foundation. 2025. URL: https://docs.python.org/3/tutorial/datastructures.html (access date: 25.09.2025)

The Kyrgyz News Corpus dataset. Hugging Face AI community. 2025. URL: https://huggingface.co/datasets/the-cramer-project/Kyrgyz_News_Corpus#kyrgyz_news_corpus (access date: 6.09.2025)

All authors have read and approved the final manuscript.
Все авторы прочитали и одобрили окончательный вариант рукописи.

Maria I. Kiose, Doctor of Philology, Associate Professor, Leading Researcher at the Centre for Socio-Cognitive Discourse Studies (SCoDis), Moscow State Linguistic University, Leading Researcher at the Laboratory of Multi-Channel Communication, Institute of Linguistics of the Russian Academy of Sciences, Moscow, Russia.
Мария Ивановна Киосе, доктор филологических наук, доцент, ведущий научный сотрудник Центра Социокогнитивных исследований (СКоДис) Московского государственного лингвистического университета, ведущий научный сотрудник Лаборатории мультиканальной коммуникации Института языкознания РАН, Москва, Россия.

Zhenishkul S. Khulkhachieva, PhD in Philology, Head of the Department of Languages and Cultures of CIS and FSU countries, Head of the Ch. Aitmatov Centre for Kyrgyz language and culture, Moscow State Linguistic University, Moscow, Russia.
Женишкуль Саматовна Хулхачиева, кандидат филологических наук, заведующий кафедрой языков и культур стран СНГ и ближнего зарубежья ИМОиСПН МГЛУ, директор центра киргизского языка и культуры им. Ч. Айтматова Московского государственного лингвистического университета, Москва, Россия.

Veronika V. Izyumskaya-Kapitonova, Junior Researcher at the Laboratory of Corpus Linguistics, Moscow State Linguistic University, Moscow, Russia.
Вероника Викторовна Изюмская-Капитонова, младший научный сотрудник лаборатории корпусной лингвистики Московского государственного лингвистического университета, Москва, Россия.

Artem V. Barmin, Junior Researcher, Laboratory for Cognitive Studies of Communication, Moscow State Linguistic University, Moscow, Russia.
Артем Вячеславович Бармин, младший научный сотрудник, Лаборатория когнитивных исследований основ коммуникации Московского государственного лингвистического университета, Москва, Россия.