



Исследовательская статья
УДК 811.134.2 [030+004.8]
<https://doi.org/10.24833/2541-8831-2022-4-24-67-79>

СОЗДАНИЕ КОРПУСОВ ИСПАНСКОГО ЯЗЫКА КАК ОДНА ИЗ ПРИОРИТЕТНЫХ ЗАДАЧ КОРОЛЕВСКОЙ АКАДЕМИИ ИСПАНСКОГО ЯЗЫКА В ЭПОХУ ЦИФРОВИЗАЦИИ

Илона Игоревна Горелая¹, Юлия Рашитовна Зиганшина²

^{1,2} МГИМО МИД России, Москва, Россия

¹ goilona@yandex.ru

<https://orcid.org/0000-0003-3903-1106>

² fishjul@mail.ru

<https://orcid.org/0000-0002-9006-2963>



Аннотация. Создание корпусов испанского языка становится одной из приоритетных задач Королевской академии испанского языка (RAE) в XXI в. Работа над составлением корпусов испанского языка является ответом на вызовы современной эпохи глобализации и основывается на новых компьютерных технологиях. Корпусы в лингвистике служат для изучения языка и представляют собой инструмент для обобщения данных о языке, имеющих разносторонний характер, начиная с частотности использования словоформ и заканчивая обучением искусственного интеллекта. Цель данной статьи — показать, как Королевская академия испанского языка (RAE), используя новые подходы в области корпусной лингвистики, расширяет области своей традиционной деятельности, расширяя испанский язык в цифровой сфере человеческой деятельности. Новизна данного исследования состоит в том, что все основные корпусы испанского языка, созданные Королевской академией испанского языка и Ассоциацией академий испанского языка, анализируются с точки зрения культурного вектора развития испаноговорящей цивилизации. Авторами определяется важность создания корпусов как способа развития данной языковой культуры в будущем, что подразумевает необходимость использования корпусов для обучения искусственного интеллекта. Поднимая тему важности цифровизации языка для общения машин с человеком, авторы считают составление корпусов необходимой основой для запуска этого

процесса. В статье, с одной стороны, анализируются существующие на данный момент академические корпусы, с другой, описываются возможности применения корпусов для языковых исследований и для обучения искусственного интеллекта, и в итоге обобщается значимость корпусов в цифровую эпоху.

Ключевые слова: корпусная лингвистика, корпус испанского языка, конкорданс, культурная идентичность, Королевская академия испанского языка, Международная ассоциация академий испанского языка, паниспанизм, обучение искусственного интеллекта

Для цитирования: Горелая И.И., Зиганшина Ю.Р. Создание корпусов испанского языка как одна из приоритетных задач RAE в эпоху цифровизации // Концепт: философия, религия, культура. — 2022. — Т. 6, № 4. — С. 67–79. <https://doi.org/10.24833/2541-8831-2022-4-24-67-79>

Research article

CREATION OF SPANISH LANGUAGE CORPORA AS ONE OF THE PRIORITIES OF RAE IN THE ERA OF DIGITALIZATION

Ilona I. Gorelaya¹, Yuliya R. Ziganshina²

^{1,2} MGIMO University, Moscow, Russia

¹ goilona@yandex.ru

² fishjul@mail.ru

<https://orcid.org/0000-0003-3903-1106>

<https://orcid.org/0000-0002-9006-2963>

Abstract. The creation of Spanish language corpora is becoming one of the priorities of the Royal Academy of the Spanish Language (RAE) in the 21st century. The work on the compilation of Spanish language corpora is a response to the challenges of the modern era of globalization and is based on new computer technologies. Corpora in linguistics are used for language research and are a tool for generalizing data about a language that have a versatile nature, starting with the frequency of use of word forms and ending with the training of artificial intelligence. The purpose of this article is to show how the Royal Academy of the Spanish Language (RAE), using new approaches in the field of corpus linguistics, expands the areas of its traditional activities, spreading the Spanish language in the digital sphere of human activity. The novelty of this study is that all the main Spanish language corpora created by the Royal Academy of the Spanish Language and the Association of Academies of the Spanish Language are analyzed from the point of view of the cultural vector of development of the Spanish-speaking civilization. The authors determine the importance of creating corpora as a way of developing this language culture in the future, which implies the need to use corpora for teaching artificial intelligence. Raising the topic of the importance of the digitalization of language for communication between machines and humans, the authors consider the compilation of corpora to be a necessary basis for launching this process. The article, on the one hand, analyzes the currently existing academic corpora, on the other, describes the possibilities of using corpora for language research and for teaching artificial intelligence, and as a result summarizes the importance of corpora in the digital age.

Keywords: corpus linguistics, corpus of the Spanish language, concordance, cultural identity, Royal Spanish Academy, Association of Academies of the Spanish Language, panhispanism, artificial intelligence training

For citation: Gorelaya, I. I., Ziganshina, Y. R. (2022) 'Creation of Spanish Language Corpora as One of the Priorities of RAE in the Era of Digitalization', *Concept: Philosophy, Religion, Culture*, 6(4), pp. 67–79. (In Russian). <https://doi.org/10.24833/2541-8831-2022-4-24-67-79>

Создание корпусов испанского языка становится в XXI в. одной из приоритетных задач Королевской академии испанского языка (RAE)¹. Отметим, что работа над составлением корпусов испанского языка во многом является ответом на вызовы современной эпохи глобализации, ведь она основывается на новых компьютерных технологиях. При этом корпусы в лингвистике, как известно, служат для изучения языка и одновременно представляют собой оригинальный инструмент сбора и обобщения данных о языке; современная корпусная лингвистика охватывает самые разные сферы языковой деятельности, начиная с частотности использования словоформ и заканчивая обучением искусственного интеллекта.

Напомним: корпусная лингвистика берёт своё начало в первых конкордансах², которые были составлены в Европе в период с XIII в. по 1960-е гг., и основывается на применении текстов, объединённых в корпусы согласно заранее установленным критериям: «В доэлектронную эпоху (т.е. до 60-х годов XX века) под понятием *корпус* в его лингвистическом значении признавалось отдельное религиозное или литературное произведение или собрание сочинений одного автора, к которому вручную составлялся конкорданс, формируемый преимущественно для теологических, литературоведческих и лексикографических исследований» [Солнышкина, Гатиятулина 2020: 132]. Современные лингвистические корпусы хранятся в электронном

виде, а для проведения исследовательской работы с корпусом текстов специалистам необходима соответствующая техническая база: сервер для хранения данных, поисковая система, система маркировки текстов, программа, с помощью которой можно посмотреть результаты проводимых исследований и т.д. В современной корпусной лингвистике есть множество определений понятия *корпус*, однако, по мнению ведущего отечественного специалиста в этой области В.П. Захарова, существует четыре основных характеристики любого языкового корпуса, как называет их автор *ипостаси*: «1) электронный; 2) репрезентативный; 3) размеченный; 4) прагматически ориентированный (создаваемый для определённых целей)» [Захаров 2016: 152].

Эти четыре принципа построения языковых корпусов легли в основу определения корпусной лингвистики и корпусов текстов: «Корпусная лингвистика — раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с использованием компьютерных технологий. Под названием *лингвистический*, или *языковой*, *корпус текстов* понимается большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач. В понятие *корпус текстов* входит также система управления

¹ RAE — La Real Academia Española (Королевская академия испанского языка) была основана в 1713 г. для работы над словарём и грамматикой испанского языка. RAE: сайт. — URL: <https://www.rae.es/la-institucion/historia>

² Конкорданс представлял собой своего рода словарь используемых слов в отдельно взятом произведении, например, в Библии. Указывалось значение слова и место, где оно применяется. Отечественный лингвист В.П. Захаров даёт следующее определение понятию «конкорданс»: «список всех употреблений данного слова в контексте со ссылками на источник» (Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. — Санкт-Петербург: Санкт-Петербургский государственный университет, 2005. — С. 48.).

текстовыми и лингвистическими данными, которую в последнее время чаще всего называют *корпусным менеджером* (или *корпус-менеджером*) (англ. *corpus manager*). Это специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме»³.

Как видно из определения, в корпусной лингвистике используются такие термины, как *языковой корпус*, *лингвистический корпус* и *корпус текстов*. По мнению М.С. Мальцевой, существует множество вариантов определений для этих понятий, однако в ходе своего исследования она приводит к выводу, что *языковой корпус* и *лингвистический корпус* являются синонимами, а *корпус текстов*, «как вид корпуса данных, близкого по определению к понятию *лингвистический* или *языковой корпус*, является частной формой проявления *лингвистического корпуса*». [Мальцева 2011: 257]. В настоящей статье вышеуказанные термины будут также использоваться как синонимичные понятия.

Следующим моментом, важным для изучения современных языковых корпусов, представляется их типология. Авторы статьи придерживаются принятой на сегодняшний момент в отечественной лингвистике типологии лингвистических корпусов: устные, письменные или смешанные; одноязычные (охватывают весь язык или создаются для определённой цели), двуязычные, многоязычные; корпусы с общим доступом и коммерческие; синхронные или диахронные; размеченные или не размеченные; свободно доступные, частично доступные или коммерческие [Козлова 2013: 86].

В то же время в современной испанской лингвистике приняты следующие подходы к корпусной лингвистике и корпусам текстов: 1) корпусная лингвистика не являет-

ся самостоятельной дисциплиной, а служит методологическим подходом, который можно адаптировать к различным дисциплинам; она отлична от методологий, основанных в основном на интроспекции (самоанализе языка); 2) корпусы текстов служат для иллюстрации реального узуса языка в его естественной форме существования. Заметим, что в испанской лингвистике термин «лингвистический корпус» имеет два различных значения в зависимости от области применения корпусов. Первое значение — собрание языкового материала с целью проведения конкретного исследования либо с целью обучения языку (в основном относится к прикладной лингвистике). Второе значение — репрезентативная коллекция текстов, собранная для того, чтобы проиллюстрировать реальное использование языка (относится к корпусной лингвистике)⁴. Приведённые типологии языковых корпусов дополняют друг друга, являются общепринятыми и могут быть использованы для характеристики корпусов, составленных на любом языке, в частности, на испанском. Корпусы испанского языка, в создании которых принимает участие Королевская академия испанского языка, могут быть классифицированы в соответствии с общепринятой типологией, а также с учётом типологии самой академии. В этом случае, как уже было сказано, классификация получается более расширенной. Например, особо выделяется цель создания корпуса и географическая принадлежность текстов. По внутренней систематизации академии корпусы классифицируются с точки зрения географической принадлежности текстов: географически маркированные и географически не маркированные. Для некоторых корпусов испанского языка, составленных Королевской академией, применяется определение «*de referencia*» (референсный), что означает «эталонный», «образцовый» «всеохватывающий». Королевская академия испан-

³ Захаров, 2005, с. 48.

⁴ Centro virtual Cervantes. Diccionario de términos clave de ELE. — URL: https://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/linguisticacorporus.htm

ского языка в принципе настаивает на том, что лингвистические корпуса испанского языка, которые состояются с её участием, являются нормативными, то есть «референсными».

Отправной точкой развития современных подходов корпусной лингвистики принято считать появление в 1963 г. так называемого *Брауновского корпуса* (*Brown Corpus*), авторами которого являются сотрудники американского университета Брауна У. Френсис (W. Francis) и Г. Кучера (H. Kucera). Именно в этот момент появляются первые вычислительные машины, способные хранить и быстро обрабатывать огромное количество текстов, слов, словоформ и т.д. В то же время, во второй половине XX в. в западной научной среде доминировал рациональный (хомскианский) подход, сторонники которого во главе с американским лингвистом Ноамом Хомским⁵ считали, что носителю языка быстрее самому придумать пример, иллюстрирующий какое-либо грамматическое явление, чем искать такой пример по многим миллионам случайных текстов [Майорова 2017: 44]. Последовательницей Хомского в Испании (и следовательно, противницей корпусной лингвистики) можно считать Марию Молинер⁶.

Позиция Королевской академии испанского языка по вопросу методики поиска примеров языковых явлений оказалась отличной от хомскианской, поскольку с момента основания академии составители словарей опирались на примеры авторитетных авторов⁷. Преимущества такого подхода представлялись очевидными — примеры были достоверными и не вызы-

вали сомнений. В то же время имелись и недостатки, вызывавшие критику: встречалась устаревшая, вышедшая из употребления лексика, и таким образом в словарь попадали архаизмы, а новые «живые» слова игнорировались. Учитывая ошибки прошлого, Королевская академия испанского языка при составлении лингвистических корпусов применяет принцип «живого языка», узуса, стремится отразить реальное состояние и фактическое использование лексем в письменной и устной речи, в разных видах дискурса и литературных жанрах на определенный период времени.

Накопленный опыт позволил академии воспринять идеи корпусной лингвистики: в 1995 г. RAE приняла решение изменить систему сбора материалов для составления грамматик и словарей, используя с этой целью лингвистические корпуса⁸ [Crónica de la lengua española, 2020: 220]. Первым был создан *Corpus de Referencia del español Actual* (CREA). Он содержал письменные и устные тексты с 1975 по 1999 гг. (см. таблицу). Работа с этим корпусом дала впечатляющие результаты, и вскоре был создан дополняющий его корпус *el Corpus Diacrónico del español* (CORDE), который содержал исключительно письменные тексты с момента появления первых письменных документов до 1974 г. Проект по составлению этих корпусов подразумевал, что CREA будет автоматически включать в себя только тексты, появившиеся за последние 25 лет, то есть с годами «устаревшие» тексты будут переноситься в CORDE.

В Хронике⁹ RAE 2020 г. указано, что в понимании академии корпус это — «собрание текстов в электронном формате, по-

⁵ Noam Chomsky (1928) — американский лингвист, автор теории генеративной лингвистики, направления лингвистики, в рамках которого изучаются правила порождения предложений в языке и строятся модели языка. Из моделей для отдельных языков делаются выводы об устройстве языка как такового. [Rojo, 2008].

⁶ María Moliner (1900–1981) — известный испанский лексикограф, создательница «Diccionario de uso del español» (1967). Все примеры словоупотреблений в словаре придуманы автором-носителем языка. (Moliner M. Diccionario de uso del español. Tercera edición. 2 vols. — Madrid: Editorial Gredos S.AU, 2007. — 3351 p.).

⁷ Первой работой Королевской академии испанского языка был *Diccionario de autoridades* (Словарь авторитетов), опубликованный в 1726 г. и в котором все примеры словоупотребления были взяты из трудов лучших испанских писателей Золотого века.

⁸ Книга в электронном формате в приложении tagus, для ссылок используется вариант разбивки на 983 страницы.

⁹ Хроника (Crónica) — ежегодный сборник статей Королевской Академии испанского языка и ассоциации академий, посвященный современному состоянию испанского языка и отчетам о работе академий в разных проектах. Выпускается с 2020 г.

добранное и составленное таким образом, чтобы можно было осуществлять научный анализ и находить в текстах необходимую информацию для проведения лексических и грамматических исследований.

Тексты, используемые в корпусе, должны быть естественного происхождения, иными словами, это могут быть романы, театральные пьесы, эссе, научные работы, новости из газет, репортажи, телевизионные новостные выпуски, лекции или разговор между друзьями. То есть эти тексты возникли в реальной ситуации, служат для коммуникации и не были специально написаны для того, чтобы проиллюстрировать некий лингвистический феномен или элемент, как это часто происходит с теми примерами, что встречаются в грамматиках или словарях. Кроме того необходимо, чтобы корпус был репрезентативным и сбалансированным. Иначе говоря, корпус должен содержать достаточное количество текстов, составленных на том языке или языковом варианте, который является предметом изысканий. Например, для исследований современного состояния испанского языка необходимо использовать корпус, состоящий из текстов, представляющих все испаноязычные страны»¹⁰ (*здесь и далее переводы наши — И.Г. и Ю.З.*).

Корпусы текстов широко применяются в настоящий момент, в том числе для составления словарей и грамматик, так как тексты, входящие в корпус, иллюстрируют узус слов в определённый исторический период, в том числе и на сегодняшний

день. Корпусы используются для фиксирования изменений, происходящих в языке — например, появление новых терминов, неологизмов и слов, заканчивающихся на *-mente*, в корпусе отмечается раньше, чем происходит закрепление этих слов в словарях. [*Crónica de la lengua española*, 2020: 484].

Языковые корпуса испанского языка, которые создаются при участии RAE, имеют следующие общие характеристики: 1) одноязычные (некоторые охватывают весь язык, другие же создаются с определёнными целями); 2) маркированные; 3) общедоступные (за некоторым исключением).

Вопросы репрезентативности являются ключевыми при работе над языковыми корпусами. Если корпус создается для изучения современного состояния испанского языка, соответственно требуется включить в него современные тексты, составленные на национальных вариантах испанского языка всех испаноговорящих стран, при этом должна быть возможность делать выборку либо по текстам той или иной страны, либо нескольких стран.

Очередной важный шаг в области создания языковых корпусов был сделан в 2007 г. На конгрессе ассоциации академий испанского языка (ASALE)¹¹, состоявшемся в городе Медельине (Колумбия), её члены поручили Королевской академии испанского языка создать корпус современного испанского языка *Corpus del Español del siglo XXI* (CORPES XXI)¹²; также было принято ре-

¹⁰ «El conjunto de textos en formato electrónico que ha sido diseñado y construido con el fin de que sea posible analizarlos científicamente y extraer de ellos la información necesaria para realizar estudios léxicos o gramaticales. Los textos que componen el corpus deben haber sido producidos en situaciones naturales. Es decir, esos textos fueron originariamente una novela, una obra de teatro, un ensayo, un texto de divulgación científica, una noticia de periódico, un reportaje, un noticiario televisivo, una clase o una conversación entre amigos. Por tanto, han sido generados con la intención real de comunicar algo, no construidos para ilustrar un determinado fenómeno o elemento lingüístico, como sucede con demasiada frecuencia con los ejemplos que figuran en las gramáticas o los diccionarios. En segundo término, el conjunto del corpus debe ser representativo y estar equilibrado. Dicho de otro modo, debe contener un número suficiente de textos que correspondan a la lengua o a la variedad lingüística cuyo estudio pretende facilitar. Por ejemplo, para que un corpus sea de utilidad en el estudio del español actual parece claro que tiene que estar compuesto por textos de todos los países del ámbito hispánico» [*Crónica de la lengua española*, 2020: 216].

¹¹ Asociación de Academias de la Lengua Española (ASALE) была создана в 1951 году для того, чтобы обеспечить целостность и единство испанского языка. ASALE. — URL: <https://www.asale.org>. Более подробно о проектах ASALE см. Раевская М.М. Испанский язык в современном мире: Испанский язык в междисциплинарной парадигме XXI века: социолингвистическое, политическое, экономическое, юридическое, дидактическое, библиографическое измерение. — Москва: Ленанд, 2021. — С. 270–274.

¹² RAE. — URL: <https://www.rae.es/banco-de-datos/corpes-xxi>

шение считать его совместным проектом, потому что к процессу создания корпусов привлекаются все академии испанского языка. Этот факт сотрудничества академий испанского языка разных стран является иллюстрацией изменений, произошедших в политике RAE в 2000 гг., когда системный характер начинает приобретать следование идеям *паниспанизма*. При этом происходит фактическое признание равенства всех национальных вариантов испанского языка; а соотношение текстов, составленных на национальных вариантах испанского языка, и включённых в языковые корпусы, демонстрирует явный перевес текстов не пиренейского происхождения.

Согласно Хронике испанского языка 2021 г. в настоящий момент существует восемь корпусов испанского языка, созданных или создаваемых академиями испанского языка [Crónica de la lengua española, 2021: 577–580], которые представлены в виде таблицы в приложении: см. таблицу.

Помимо международной ассоциации академий и Королевской академии испанского языка существует множество других организаций, занимающихся составлением корпусов испанского языка. Несмотря на масштабность проектов академий, на сегодняшний день самым большим из всех существующих корпусов, является **Corpus del Español**, созданный *Марком Дэвисом*¹³ в университете Иллинойса (США) в 2016 г. Этот корпус содержит нескольких десятков миллиардов слов из двадцати одной испаноговорящей страны. Среди других корпусов, составленных другими организациями, следует отметить три крупных корпуса:

- **Sketch Engine** — многоязычный корпус¹⁴.
- **Corpus de la web COW** — корпус создан Свободным Университетом Берлина¹⁵.
- **Corpus del español actual** — CEA¹⁶.

Данные, полученные при помощи корпусов, используются не только в различных областях традиционной теоретической и прикладной лингвистики, но и в компьютерной (математической) лингвистике, занимающейся моделированием интеллектуальных процессов человека при создании систем искусственного интеллекта, а следовательно эти данные могут быть использованы для обучения искусственного интеллекта.

Напомним: языковые корпусы используются для составления баз данных для программ машинного перевода и для обучения естественному языку искусственного интеллекта. Языковой материал для составления алгоритмов распознавания речи и обучения искусственного интеллекта ведению «диалога» с человеком-пользователем может быть взят из лингвистических корпусов. Согласно мнению З.Б. Долгих, именно «компьютерные лингвисты пытаются выявить и использовать статистические и лингвистические закономерности текстов для построения компьютерных моделей языка» [Долгих 2018: 25]. Чтобы машина и человек могли понимать друг друга, им нужен общий язык, а для этого необходимо создать технические устройства, способные к обучению, логическим выводам и рациональному поведению. И, самое главное, — к ведению диалога с человеком. С.А. Шумский, Директор научно-координационного совета Центра науки и технологий искусственного интеллекта МФТИ, президент Российской ассоциации Нейроинформатики, описывает ситуацию следующим образом: «Искусственный интеллект это, по сути, наделение компьютеров человеческими способностями, из которых владение языком — едва ли не важнейшая. Все знания человечества записываются и передаются на естественных языках. Поэтому не удивительно, что способность машин понимать чело-

¹³ Davies M. Corpus del español. — 2016. — URL: <https://www.corpusdelespanol.org/files/corpus-del-espanol.pdf>

¹⁴ Sketchengine. — URL: <https://www.sketchengine.eu/#blue>

¹⁵ Corpora from the Web. — URL: <https://www.fu-berlin.de/sites/digitale-lehre/ressourcen/phillegeist/Germanistik/corpora-web/index.html>

¹⁶ Corpus del Español Actual: powered by CQPweb. — URL: <http://cea.spanishfn.org/cea/>

веческий язык и общаться с людьми “почеловечески” всегда была в фокусе исследований по искусственному интеллекту»¹⁷.

Поскольку на сегодняшний день в науке нет единства по поводу перспектив так называемого «сильного» искусственного интеллекта, не представляется возможным достоверно судить о том, насколько искусственный интеллект в принципе способен воспринимать лексическое значение слов. В то же время существует консенсус по поводу того, что искусственный разум в его «слабой» версии «лишь» использует алгоритмы, подставляя слова в зависимости от того, что заложено в программу. Иными словами, принято считать: машина «не понимает» лексического значения слов, не понимает смысла сказанного. В то же время работа над обучением «сильного» ИИ с тем, чтобы появилась возможность полноценного диалога между человеком и искусственным разумом, идёт в настоящее время очень активно. Свидетельством тому служат научные разработки, направленные на преодоление конкретных трудностей «перевода», связанных с особенностями обработки естественного языка, его семантическим и стилистическим аспектами, передачей эмоциональной окрашенности речи и т.д.¹⁸

Идея энтузиастов в области изучения этой проблемы, к которым можно отнести и представителей RAE, состоит в том, что примеры использования слов и словоформ в человеческой речи, взятые из языкового корпуса, могут помочь искусственному интеллекту «научиться» говорить. В частности, одной из важных функций корпуса, применяемой для общения машины с человеком, является возможность быстрого

построения множества N-грамм, то есть цепочек из двух и более элементов, по модели слово X-искомое слово*-слово Y [Lingüística de corpus en español, 2022: 19]. Таким образом, машина может предугадать речь, заполнить пробелы в речи или сконструировать высказывание и проверить его правильность. В разных корпусах эта функция реализована по-разному.

Вопрос, на каком языке будет говорить искусственный интеллект, на сегодняшний день остаётся как известно открытым. Изначально шансы на победу в этой борьбе выше у английского языка. Однако на испанском языке, по утверждению академиков RAE, в настоящее время говорит больше машин, чем людей: «700 миллионов машин каждый день общаются с 580 миллионами людей» [Crónica de la lengua española, 2020: 480]. Также представители Королевской академии испанского языка прямо говорят о необходимости преемственности испанских культурных традиций, этики и морали в процессе обучения искусственного интеллекта; о стремлении закрепить использование «правильного», нормированного, испанского языка для развития цифровых технологий. Хорошо осознавая тот факт, что английский язык широко распространён в компьютерной сфере, Королевская академия испанского языка тем не менее запускает самостоятельный проект LEIA¹⁹, который находится на стыке корпусной лингвистики и новых цифровых технологий. Его цель — «...**обеспечивать правильное использование испанского языка машинами**, а с другой стороны, воспользоваться **искусственным интеллектом**, чтобы создавать инструменты, которые будут способствовать пра-

¹⁷ Шумский С.А. Машинная обработка естественного языка // Альманах «Искусственный интеллект». — 2019 — сентябрь. — С. 10-12. — URL: <https://singapore-academy.org/libcdo/100.pdf>

¹⁸ Согласно мнению И. Пионтовской, «учёные достигли большого прогресса в области генерации текстов, машинного перевода и извлечению информации, даже выраженной неявно. Однако до “разговорного искусственного интеллекта” ещё далеко. Текущие алгоритмы пасуют, когда дело доходит до логических связей, применения общих знаний о мире и человеческой эмпатии — понимания текущего поведения собеседника» (Пионтовская И. Обработка естественного языка, поиск и извлечение информации из текстов // Альманах «Искусственный интеллект». — 2019 — сентябрь. — С. 23. — URL: <https://singapore-academy.org/libcdo/100.pdf>)

¹⁹ LEIA — Lengua Española e Inteligencia Artificial. (Испанский язык и искусственный интеллект). RAE. — URL: <https://www.rae.es/leia-lengua-espanola-e-inteligencia-artificial>

вильному использованию испанского языка людьми»²⁰.

Интересно, что партнёрами данного проекта становятся западные гиганты компьютерных и коммуникационных технологий: *Telefónica, Google, Amazon, Microsoft* и др., а также фонд *Fundación Endesa*. Королевская академия испанского языка подписала с вышеуказанными организациями соглашения о совместной разработке и создании голосовых помощников, редакторов текстов, поисковиков, мессенджеров, социальных сетей и любых других ресурсов с тем, чтобы услуги и сервисы этих компаний были доступны пользователям на грамотном испанском языке. На официальном сайте RAE приводятся слова выступления Директора проекта LEIA и по совместительству председателя ASALE Сантьяго Муньоса Мачадо: «Мы переживаем такой же переломный момент, как пережили наши предки в XVIII в. (с людьми), и мы должны сделать то, что сделали они тогда: создать нормы для языка машин и искусственного интеллекта. Язык искусственного интеллекта имеет тенденцию к диверсификации... Искусственный интеллект в основном говорит на английском и нам следует озаботиться тем, чтобы испанский язык постепенно занял заметное положение не только в мире искусственного интеллекта, но и в мире социальных сетей»²¹. Задача научить искусственный интеллект говорить на правильном испанском языке была озвучена в эпическом видеоролике²², где сказано, что искусственный интеллект может стать «хранителем величия испанского языка» на новом витке истории.

Взаимодействию испанского языка и

искусственного интеллекта был также посвящён ряд мероприятий, проведённых университетом им. Менендеса Пелайо (UIMP) в июле 2022 г. Испанские лингвисты и специалисты в области искусственного интеллекта обсудили в ходе этого форума вопросы, представляющие интерес для всех его участников. И это только один из шагов, предпринимаемых Королевской академией испанского языка для продвижения испанской культуры через язык и закрепление его в будущем в качестве языка общения искусственного интеллекта с человеком.

Подтверждением той значимости, которую придаёт испанскому языку руководство Испании на государственном уровне, заботясь о его распространении в мире и закреплении в цифровой среде, служат слова государственного секретаря по вопросам Цифровизации и Искусственного Интеллекта Карме Артигас²³, который обратил внимание на следующее. Во-первых, испанский язык уже достаточно широко распространён в мире, где он помимо прочего имеет огромное экономическое значение — на нём говорят 600 миллионов человек. Во-вторых, очень важно, чтобы развитие искусственного интеллекта имело в качестве отправной точки Испанию и закрепляло «этику, наши (испанские) принципы и ценности». Карме Артигас настаивает на том, что испанцы хотят принимать участие в происходящих изменениях, а не быть пассивными зрителями.

Предпринятое нами исследование показало, что Королевская академия испанского языка целенаправленно проводит политику усиления влияния испанского

²⁰ «...**velar por el buen uso de la lengua española en las máquinas** y, por el otro, aprovechar la **inteligencia artificial** para crear herramientas que fomenten el **uso correcto del español en los seres humanos**» // RAE. — URL: <https://www.rae.es/noticia/la-rae-presenta-el-proyecto-lengua-espanola-e-inteligencia-artificial-leia-en-el-xvi>

²¹ «estamos en un momento crucial en el que tenemos que hacer algo que hicieron nuestros antecesores del siglo XVIII (con los humanos): normativizar la lengua de las máquinas y de la inteligencia artificial (IA). Su lengua tiende a diversificarse y hay que tomar medidas. La IA habla inglés, fundamentalmente, y tenemos que procurar que, poco a poco, el español coja una posición eminente en el mundo de la IA, pero también en el mundo general de las redes» // ASALE. — URL: <https://www.asale.org/noticia/la-rae-presenta-el-proyecto-lengua-espanola-e-inteligencia-artificial-leia-en-el-xvi-0>

²² El director de la RAE inaugura las jornadas sobre lengua española e inteligencia artificial en la UIMP // RAE. — 2022. — 6 jun. — URL: <https://www.rae.es/noticia/el-director-de-la-rae-inaugura-las-jornadas-sobre-lengua-espanola-e-inteligencia-artificial>

²³ Ibid.

языка и распространения испаноязычной культуры в мире. Для реализации этой амбициозной задачи RAE, декларируя преемственность научных традиций, отказывается от претензий на исключительную правильность пиренейского варианта испанского языка: а именно, при составлении корпусов опирается академия опирается на идею равноправия всех национальных вариантов испанского языка (в том числе, их равноправия в цифровом пространстве).

При этом начиная с 2000-е гг. академия демонстрирует стремление не только закрепить сложившиеся нормы испанского языка, но и зафиксировать его фактический узус, для чего включает в языковые корпусы новые тенденции (например, в словообразовании). Важно, что это делается именно путём обращения к приёмам корпусной лингвистики. Таким образом корпусная лингвистика, являясь инструментом изучения языка, становится в деятельности академии частью процесса объединения достижений лингвистики и цифровых технологий, что происходит за счёт 1) составления «референсных» языковых корпусов и 2) создания системы обучения искусственного интеллекта на испан-

ском языке. В последнем случае речь идёт о том, чтобы люди, пользующиеся компьютерными технологиями и работающие над развитием искусственного интеллекта, использовали испанский язык *правильно*. С этой целью продвигается проект LEIA, реализация которого должна обеспечить расширение влияния испанского языка в мире, попутно решая задачу популяризации испанской культуры в ходе цифровизации. Несмотря на то, что у испанского языка в сфере новейших цифровых технологий серьёзным соперником является английский язык, испанский при успешной и последовательной работе академий испанского языка имеет хорошие шансы занять достойную нишу в этой области, сохранив и даже усилив своё влияние и значимость в цифровую эпоху.

Таким образом, исходя из культуроцентричного постулата, согласно которому при исчезновении языка исчезает и культура народа, говорящего на этом языке, академия реализует проект по созданию корпусов испанского языка для укрепления его позиций в будущем.

Корпусы, созданные при содействии RAE и ASALE

	КОРПУС	ОПИСАНИЕ	РЕПРЕЗЕНТАТИВНОСТЬ	СООТНОШЕНИЕ ТЕКСТОВ ПО ГЕОГРАФИЧЕСКОМУ ПРИНЦИПУ: Пиренейский вариант/национальные варианты испанского языка	РАЗМЕР КОРПУСА (кол-во лексем)
1	CORPES XXI - <i>Corpus del Español del siglo XXI</i> Корпус испанского языка XXI века	Общедоступный, смешанный, синхронно-диахронический, референсный, географически маркированный. Цель создания: получение общей и целостной картины испанского языка на определенный момент времени - XXI век.	Широкий охват текстов: научные публикации, публицистика, художественная литература, театральные пьесы, сценарии к фильмам, новости из прессы, эссе, записи новостных радио- и телепередач, записи бесед, выступлений и т.д. В текстах корпуса зафиксирован реальный узус. Соотношение текстов: 90% письменные, 10% устные. Источники письменных текстов: 40% книги, 40% периодические издания, 7,5% интернет, 2,5% другое.	30/70	350 млн слов Версия beta 0.94 (06.07.2021) (Ежегодно корпус прирастает 25 миллионами форм, соответствующих одному году XXI столетия. Работа по пополнению корпуса продолжается). (https://www.rae.es/banco-de-datos/corpes-xxi)
2	CORDE - <i>Corpus Diacrónico del Español</i> Диахронический корпус испанского языка	Общедоступный, письменный, диахронический, географически маркированный. Цель создания: сохранение ценных исторических данных, с целью проведения исследований в области изменения значения слов.	Состоит исключительно из письменных текстов, охватывающих период с момента появления первых испанских текстов и до 1974 г. Тексты самой широкой тематики, есть возможность выбрать между поэзией и прозой. (тексты могут совпадать с текстами CREA)	74/25 (1% еврейско-испанский).	250 млн слов https://www.rae.es/banco-de-datos/corde

	КОРПУС	ОПИСАНИЕ	РЕПРЕЗЕНТАТИВНОСТЬ	СООТНОШЕНИЕ ТЕКСТОВ ПО ГЕОГРА- ФИЧЕСКОМУ ПРИН- ЦИПУ: Пиренейский вариант/нацио- нальные варианты испанского языка	РАЗМЕР КОРПУСА (кол-во лексем)
3	CREA - <i>Corpus de Referencia del Español Actual</i> Референсный кор- пус современного испанского языка CREA – 0.3 <i>Versión anotada</i> Расширенная версия корпуса CREA	Общедоступный, смешанный, диахрони- ческий, географически маркированный. Цель создания: составление рефе- ренсного корпуса испанского языка на определенный момент времени. Цель создания: обеспечение более усовершенствованной системы поиска, включающей расширенные параметры по сравнению с первой версией корпуса CREA.	Корпус делится на пятилетние периоды: 1975–1979; 1980–1984; 1985–1989; 1990–1994; 1995–1999, 2000–2004. Соотношение текстов: 90% письменные, 10% устные. Источники письменных текстов: 49% книги, 49% пресса, 2% разное (рекламные буклеты, проспекты, электронные письма, блоги и пр.) Тексты в CREA и CORDE могут совпадать.	50/50	160 млн слов https://www.rae.es/banco-de-datos/crea
4	FICHERO GENERAL Главная картотека	Общедоступный, письменный, диахронический, географически не маркированный. Цель создания: сохранение ценных исторических данных, для проведения исследований в области изменения значения слов, а также для составления электронного словаря <i>Nuevo diccionario histórico del español</i> (NDHE). (Новый исторический словарь испанского языка) (Общедоступный, письменный, диахронический, географически не маркированный). Цель создания: сохранение ценных исторических данных, для проведения исследований в области изменения значения слов, а также для составления электронного словаря <i>Nuevo diccionario histórico del español</i> (NDHE). (Новый исто- рический словарь испанского языка)	Оцифрованные карточки, содержащие лексему и цитату - пример использования слова. Примеры взяты из текстов самого разного характера и места происжде- ния, начиная с момента возникновения испанского письменного языка. Наибо- льшее количество карточек было собрано в период с 1930 по 1996.	—	10 млн карточек (на одну лексему может приходиться от одной до двух сотен карточек) https://www.rae.es/banco-de-datos/fichero-general
5	CDH - <i>Corpus del Diccionario Histórico de la Lengua Española</i> Корпус истори- ческого словаря испанского языка	Общедоступный, диахронический, гео- графически маркированный. Цель создания: для составления электронного исторического словаря испанского языка (DHE).	Тексты охватывают период с XII века по 2000. Корпус разделен на три временных модуля: 1) основной (частично совпадает с текстами CORDE и CREA); 2) с XII в. по 1975; 3) 1975–2000.	60/40	355 740 238 слов https://www.rae.es/banco-de-datos/cdh
6	CORPUS DE TEXTOS DE CARACTER CIENTÍFICO- TÉCNICO GENERADO A PARTIR DE RECURSOS DE LA RAE y FECYT Корпус на- учно-технических текстов, сформи- рованный на базе ресурсов RAE и FECYT (Испанский фонд науки и технологии)	Общедоступный, письменный, диахро- нический, географически маркироанный. Цель создания: более широкое использование испанского языка в науке, правильное использование на- учно-технических терминов в различных контекстах, популяризация науки.	Три источника текстов: 1) научно-технические тексты из CORPES XXI; 2) база SINC (Информационная служба научных новостей); 3) другие материалы FECYT. Тексты относятся к периоду 2000–2019.	—	112 млн слов https://enclavedeciencia. rae.es/contenidos/ sobre-edc

	КОРПУС	ОПИСАНИЕ	РЕПРЕЗЕНТАТИВНОСТЬ	СООТНОШЕНИЕ ТЕКСТОВ ПО ГЕОГРА- ФИЧЕСКОМУ ПРИН- ЦИПУ: Пиренейский вариант/нацио- нальные варианты испанского языка	РАЗМЕР КОРПУСА (кол-во лексем)
7	CAPITEL - <i>Corpus del Plan de Impulso a las Tecnologías del Lenguaje</i> Корпус Плана развития языко- вых технологий	С ограниченным доступом, синхронный. Цель создания: является частью госу- дарственной программы по созданию и продвижению языковых технологий, направленных на развитие машинного перевода с испанского на другие коо- фициальные языки Испании на основе обработки естественного языка.	Тексты новостного характера за период 2007–2019. Находится в процессе создания (начало 2020 г.)	—	243 млн слов https://www.sites.google.com/view/capitel2020
8	CORDIAM - <i>Corpus Diacrónico y Diatópico del Español de América</i> . Диахронический корпус испанско- го языка Северной и Южной Америки	Общедоступный, письменный, диахронический, географически маркированный. Цель создания: углубление знаний об испанском языке, об историческом раз- витии языка, сохранение исторической памяти, рассказ о жизни на Амери- канском континенте, изложенный на испанском языке.	Тексты охватывают период с 1494 по 1905 и делятся на три группы: 1) CORDIAM - Документы; 2) CORDIAM - литература; 3) CORDIAM- Пресса.	0/100	9 млн (www.cordiam.org)

Список литературы:

Долгих З.Б. Обзор ряда корпусных возможностей в сфере лингвистических исследований (на при-
мере анализа средств градуирования в португальском языке) // Вестник Московского государственного
лингвистического университета. Гуманитарные науки. — 2018. — № 5. — С. 21–32.

Захаров В.П. Прологомены к корпусной лингвистике // Вопросы психолингвистики. — 2016. —
№ 28. — С. 150–161.

Козлова Н.В. Лингвистические корпуса: определение основных понятий и типология // Вестник
Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация.
— 2013. — Т. 11, № 1. — С. 79–89.

Майорова А.Д. Корпусная лингвистика: исторический и лингводидактический аспекты // Междуна-
родный научно-исследовательский журнал. — 2017. — № 5–2. — С. 42–46. — <https://doi.org/10.23670/IRJ.2017.59.141>

Мальцева М.С. К определению термина "лингвистический корпус" и его вариантов // Социально-
экономические явления и процессы. — 2011. — № 7. — С. 255–258.

Солнышкина М.И., Гатиятуллина Г.М. История развития корпусной лингвистики (на примере ан-
глоязычных корпусов) // Вестник Томского государственного университета. Филология. — 2020. —
№ 63. — С. 132–160. <https://doi.org/10.17223/19986645/63/8>

Crónica de la lengua española 2020 / RAE, ASALE. — Barcelona: Editorial Planeta, 2020. — 983 p.

Crónica de la lengua española 2021 / RAE, ASALE. — Madrid: Espasa, 2021. — 1272 p.

Lingüística de corpus en español / G. Parodi, P. Cantos-Gómez, Ch. Howe et al. — London: Routledge,
2022. — 600 p. <https://doi.org/10.4324/9780429329296>

Rojó G. Lingüística de corpus y lingüística del español // Actas del XV Congreso Internacional de ALFAL.
— Montevideo, 2008. — URL: https://gramatica.usc.es/~grojo/Publicaciones/Lgca_corpus_lgca_espanol.pdf

References:

Dolguikh, Z. B. (2018) 'An Overview of Some Basic Possibilities of corpus in the Field of Linguistic
Research, (on the Example of the Analysis of Graduation Means in the Portuguese Language', *Vestnik of
Moscow State Linguistic University. Humanities*, (5), pp. 21–32. (In Russian).

- Kozlova, N. V. (2013) 'Linguistic Corpus: Typology and Terms', *Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriâ: lingvistika i mezhkul'turnaâ kommunikaciâ*, 11(1), pp. 79–89. (In Russian).
- Maioirova, A. D. (2017) 'Corpus Linguistics: Historical and Linguistic Diagnostic Aspects', *Research journal of international studies*, (5–2), pp. 42–46. (In Russian). <https://doi.org/10.23670/IRJ.2017.59.141>
- Maltseva, M. S. (2011) 'About the Definition of the Term Linguistic Corpus and its variation', *Social'no-èkonomičeskie âvleniâ i processy*, (7), pp. 255–258. (In Russian).
- Parodi, G. et al. (2022) *Lingüística de corpus en español*. London: Routledge. <https://doi.org/10.4324/9780429329296>
- RAE and ASALE (2020) *Crónica de la lengua española 2020*. Barcelona: Editorial Planeta.
- RAE and ASALE (2021) *Crónica de la lengua española 2021*. Madrid: Espasa.
- Rojo, G. (2008) 'Lingüística de corpus y lingüística del español', in *Actas del XV Congreso Internacional de ALFAL*. Montevideo. https://gramatica.usc.es/~grojo/Publicaciones/Lgca_corpus_lgca_espanol.pdf
- Solnyshkina, M. I. and Gatiyatullina, G. M. (2020) 'The History of Corpus Linguistics (On the Example of the English Language Corpora)', *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya*, (63), pp. 132–160. (In Russian). <https://doi.org/10.17223/19986645/63/8>
- Zakharov, V. P. (2016) 'Prolegomena to Corpus Linguistics', *Voprosy psiholingvistiki*, (28), pp. 150–161. (In Russian).

Информация об авторах

Илона Игоревна Горелая — старший преподаватель кафедры испанского языка МГИМО МИД России, 119454, Москва, проспект Вернадского, 76 (Россия)

Юлия Рашитовна Зиганшина — преподаватель кафедры испанского языка МГИМО МИД России, 119454, Москва, проспект Вернадского, 76 (Россия)

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Information about the authors

Ilona I. Gorelaya — Senior lecturer of the Department of Spanish Language, MGIMO University, 76, Prospect Vernadskogo, Moscow, Russia, 119454 (Russia)

Yuliya R. Ziganshina — Lecturer of the Department of Spanish Language, MGIMO University, 76, Prospect Vernadskogo, Moscow, Russia, 119454 (Russia)

Conflicts of interest. The authors declare absence of conflicts of interest.

Статья поступила в редакцию 06.10.2022; одобрена после рецензирования 23.11.2022; принята к публикации 10.12.2022.

The article was submitted 06.10.2022; approved after reviewing 23.11.2022; accepted for publication 10.12.2022.