Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2023 Vol. 27 No. 4 981–1004
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-35812
EDN: XMTURJ

Research article / Научная статья

Verb database: Structure, clusters and options

Nadezhda V. BUNTMAN[1], Anna S. BORISOVA[2] ✉ and Yulia A. DAROVSKIKH[1]

[1]Lomonosov Moscow State University, Moscow, Russia
[2]RUDN University, Moscow, Russia
✉borisova_as@pfur.ru

Abstract
The content and volume of language corpora provide an opportunity to obtain reliable information about the real use of a particular linguistic unit. Nowadays, there is a large number of corpora in different languages, their formation technologies are being improved. Nevertheless, some problems and limitations arise when using these resources in comparative studies. Corpora users need to work with annotated data submitted to tagging through annotation protocols. The article presents the structure and functionality of the supracorpora verb database (SVD)[1] developed on the basis of a parallel Russian-French subcorpus of the Russian National Corpus (RNC) and reveals the difference in their potentials. The described database is a pilot version of the final software, which is currently under development and is being tested. It consists of several clusters focused on solving such linguistic tasks as studying the grammatical semantics specifics and the distribution of verb forms in Russian and French; identifying the polysemantic structure in the two languages, which in turn verifies the understanding of the linguistic worldview of the speakers of Russian and French. It has been found that the mechanism of functioning of SVD cluster formations allows us to study both individual characteristics of verbs and the semantics of verbal lexemes and collocations. The manual annotation enables users to identify the systematic asymmetry of verb forms and cases of contextual and low-frequency asymmetry. Thus, SVD can be used in language pedagogy, teaching and studying discursive grammar, as well as the analysis of translation models variability.
Key words: *supracorpora verb database, clusters, manual annotation, comparative analysis, translation variant*

For citation:
Buntman, Nadezhda V., Anna S. Borisova & Yulia A. Darovskikh. 2023. Verb database: Structure, clusters and options. *Russian Journal of Linguistics* 27 (3). 981–1004. https://doi.org/10.22363/2687-0088-35812

[1] Supracorpora verb database, a system for annotating verb forms by clusters, was created by a team of Russian programmers, linguists, and translators in the laboratory of Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences in 2013.

Глагольная база данных: структура, кластеры, опции

Н.В. БУНТМАН[1], А.С. БОРИСОВА[2] ✉, Ю.А. ДАРОВСКИХ[1]

[1]*Московский государственный университет имени М.В. Ломоносова, Москва, Россия*
[2]*Российский университет дружбы народов, Москва, Россия*
✉borisova_as@pfur.ru

Аннотация
Содержание и объем лингвистических корпусов различного типа позволяет получать достоверную информацию о реальном функционировании той или иной языковой единицы. В настоящее время существует большое количество корпусов на различных языках, технологии их формирования постоянно совершенствуются. Однако при использовании данных ресурсов в сопоставительных исследованиях возникают некоторые проблемы и ограничения. В этой связи наблюдается необходимость работать с материалом, который был обработан с применением протоколов аннотирования и методов синтаксического анализа. Цель статьи – представить структуру и функционал надкорпусной глагольной базы данных (НГБД)[2], разработанной на основе параллельного русско-французского подкорпуса Национального корпуса русского языка (НКРЯ), а также показать разницу их потенциалов. Описываемая база данных представляет собой систему ручного аннотирования глагольных форм в соответствии с кластерами и является пилотной версией конечного программного обеспечения, которое в настоящее время находится в разработке и проходит апробацию. НГБД состоит из нескольких кластеров, ориентированных на решение ряда лингвистических задач: определить специфику контекстной грамматической семантики и распределения глагольных форм в русском и французском языках; выявить структуру полисеманта в двух языках, что в свою очередь позволяет верифицировать представления о языковых картинах мира носителей рассматриваемых языков. Результаты исследования показали, что механизм функционирования кластерных образований описываемого ресурса позволяет изучать как отдельные характеристики глаголов, так и семантику глагольных лексем и коллокаций. Проводимое ручное аннотирование предусматривает возможность выявить системную асимметрию глагольных форм, а также случаи контекстуальной и малочастотной асимметрии. Таким образом, НГБД может быть использована в лингводидактике, преподавании и изучении дискурсивной грамматики, а также в анализе вариативности моделей перевода.
Ключевые слова: *надкорпусная глагольная база данных, кластеры, ручное аннотирование, сопоставительный анализ, вариант перевода*

Для цитирования:
Бунтман Н.В., Борисова А.С., Даровских Ю.А. Глагольная база данных: структура, кластеры, опции. *Russian Journal of Linguistics.* 2023. Т. 27. № 4. С. 981–1004. https://doi.org/10.22363/2687-0088-35812

1. Введение

Актуальность корпусных исследований подтверждается растущим интересом отечественных и зарубежных ученых к методам обработки устной и письменной речи, использованию полученных данных в создании продуктов,

[2] НГБД была создана коллективом российских программистов, лингвистов и переводчиков в 2013 году в лаборатории Федерального исследовательского центра «Информатика и управление» Российской академии наук.

связанных с искусственным интеллектом, в исследованиях одного или нескольких языков, в обучении родному и иностранному языку.

С каждым годом современные лингвистические исследования все теснее переплетаются с компьютерными технологиями. Некоторые направления кажутся наиболее очевидными. Во-первых, это сближение с компьютерной лингвистикой, методы которой позволяют повышать точность расчетов лингвистического анализа благодаря способности искусственных нейронных сетей к обучению и возможности не только автоматизировать проводимый анализ, но и решать задачи отбора, модификации и сопоставления текстов различных типов и жанров (Solovyev, Solnyshkina & McNamara 2022, Sharoff 2022).

Во-вторых, это «широкое распространение корпусной лингвистики – как материала и как метода – на всю сферу гуманитраных исследований, в истории, социологии, литературоведении и т.д.» (Николаев 2017: 151). Современное корпусное исследование – это больше, чем просто методика анализа. Исследователи, работающие в данном дисциплинарном поле, не только используют корпусные методы или данные в своей работе, они разрабатывают и аннотируют различные корпусные ресурсы (Плунгян, Рахилина, Резникова 2022).

Выводы о функционировании той или иной языковой единицы, какого объема она бы ни была, – от отдельной лексемы до сверхфразового единства, особенно если речь идет о ее контекстуальной семантике или частотности употребления, – невозможны без опоры на большие массивы релевантных данных. Такой материал предоставляется, например, различными корпусами:

- моноязычными (Генеральный интернет корпус русского языка; корпус французского языка Frantext, постоянно дополняемый и насчитывающий 268 миллионов словоформ и др.);
- параллельными двуязычными (параллельные корпусы НКРЯ, представляющие наиболее крупную коллекцию параллельных текстов в русскоязычном сегменте корпусной лингвистики и др.);
- многоязычными (корпус Организации Объединенных Наций, где представлены документы ООН на шести языках: арабском, китайском, испанском, русском, английском и французском; Opus Parallel corpus, включающий тексты на более 100 языках; корпус DraCor, включающей тексты пьес в основном на европейских языках; параллельный корпус переводов «Слово о полку Игореве»; корпусный менеджер Sketch Engine и др.).

Надкорпусные базы данных строятся на материале корпуса, из которого автоматически извлекаются искомые языковые единицы, аннотируемые экспертами вручную сообразно с набором характеристик (признаков), необходимых для последующего поиска и обработки данных.

Описываемая надкорпусная глагольная база данных строится на материале текстов параллельного русско-французского корпуса Национального

корпуса русского языка (НКРЯ[3]) и представляет собой систему ручного аннотирования глагольных форм в соответствии с кластерами. НГБД была создана в 2013 году в лаборатории Федерального исследовательского центра «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН) коллективом российских программистов, лингвистов и переводчиков. Это пилотная версия конечного программного обеспечения, которое в настоящее время находится в разработке и проходит экспертизу. Демоверсия размещена на сайте ФИЦ ИУ РАН[4].

В процессе разработки также находится специальный интерфейс программы, предназначенный для работы студентов и преподавателей, однако его описание не входит в задачи данной статьи. Цель настоящей статьи – представить структуру и механизмы функционирования надкорпусной базы данных глагольных форм, созданной на основе параллельного русско-французского подкорпуса Национального корпуса русского языка и показать разницу их потенциалов.

Описываемая НГБД состоит из нескольких кластеров, ориентированных на решение следующих лингвистических задач:

- во-первых, изучение специфики контекстной грамматической семантики;
- во-вторых, распределение глагольных форм в русском и французском языках;
- в-третьих, выявление структуры полисеманта в языковой паре русский–французский, что в свою очередь позволяет верифицировать представления о языковой картине мира носителей рассматриваемых языков.

Функционал НГБД также позволяет решать широкий спектр практических задач современного переводоведения, в частности проводить контрастивный анализ особенностей передачи текстов художественной литературы и эссеистики с ИЯ на ПЯ, выполняемой профессиональными переводчиками.

Одна из задач исследования заключается в том, чтобы продемонстрировать некоторые опциональные отличия разработанной НГБД от параллельного русско-французского корпуса НКРЯ. Функционал описываемого ресурса позволяет его пользователям изучать как отдельные характеристики глагола (время, наклонение, аспектуальность, соотношение финитных и нефинитных форм), так и семантику глагольных лексем и коллокаций. Особый интерес также представляют возможности НГБД в плане работы с языковыми единицами, определяемыми как лингвоспецифичные. Например, русский глагол *собираться* и глаголы физического состояния *стоять, лежать* и др. выражены во французском языке широким спектром вариантов. Однако благодаря ручному методу обработки глагольных форм учитываются

[3] https://ruscorpora.ru/search?search=CiUqFwoICAAQChgyIAogADIFZ3JzdGRABXgBMgcIBRIDZnJhOgEBMAE%3D

[4] a179.frccsc.ru/PublicLingvoProjects/main.aspx

не только формальные характеристики глагола, но и ситуация высказывания (диалогическая реплика, отрицание, вопрос, восклицание). Соответственно, пользователи могут анализировать не только глагольные лексемы, но и их функционирование в предложении. Кроме того, разработанная глагольная база основана на текстах, взятых из французской и русской художественной литературы и эссеистики. Данный материал позволяет выявить системную асимметрию глагольных форм, описанную в теоретических исследованиях и практических пособиях, а также случаи контекстуальной и малочастотной асимметрии.

Таким образом, надкорпусная глагольная база данных может быть использована в лингводидактике, преподавании и изучении дискурсивной грамматики, анализе вариативности моделей перевода. По результатам представления параметров и опций НГБД в дальнейшем можно продолжить и углубить подобные исследования в области корпусной и сопоставительной лингвистики.

2. Теоретические основы исследования

Интерес к созданию корпусов и корпусным исследованиям возникает в 60-е годы XX столетия. Именно в этот период в Брауновском университете (США) создается первый лингвистический корпус английского языка – Brown University Standard Corpus of Present-Day American English, также известный как Брауновский корпус. Одним из наиболее масштабных европейских корпусных проектов является Уппсальский корпус русского языка, разработанный учеными Уппсальского университета (Швеция).

Первые корпуса состояли не из целых текстов, а из отрывков. В этом, на наш взгляд, заключается определенная необъективность текстового материала, потому что по субъективно выбранному отрывку сложно адекватно судить о контекстной семантике языковой единицы. С появлением устных и мультимедийных корпусов определение должно быть расширено и дополнено. На данный момент, учитывая информационное разнообразие видов материала, корпусом может считаться сбалансированная представительная (отличающаяся разнообразием жанров и хронологии) постоянно пополняемая коллекция размеченных (аннотированных) письменных, устных, видеодокументов, с возможностями поиска материала в зависимости от цели исследования.

Современные определения понятия «лингвистический корпус» отражают его основные свойства. Так, С.А. Шаров отмечает, что корпус – это не просто «коллекция текстов, собранная в соответствии с явно сформулированными принципами и возможно размеченная (annotated) на некотором уровне лингвистического анализа», а «*представительная* коллекция» (…), которая может адекватно представлять потенциально бесконечное множество текстов некоторого фиксированного типа в некотором диахроническом срезе» (Шаров 2003). Согласно В.П. Захарову и С.Ю. Богдановой корпус или

лингвистический корпус – это собрание текстов, представленное в электронном (машиночитаемом) формате, унифицированное, структурированное, размеченное и предназначенное для решения лингвистических задач (Захаров, Богданова 2020).

Однако на современном этапе развития корпусной лингвистики становится очевидным, что традиционные корпусные технологии (лингвистические корпуса и корпусные менеджеры: AntConc, Sketch Engine и др.) не способны обработать весь объем языковых данных и учесть все контекстуальные значения языковых единиц, а также их диахронические семантические трансформации. На ограниченность функционала современных корпусных технологий указывает и Ф. Растье, подчеркивая размытость языковой нормы как таковой, ее зависимость от типа и жанра дискурса (Rastier 2004). Аналогичная проблема касается продуктов, созданных на основе корпусов: компьютерных словарей, тезаурусов и др. В связи с этим внимание лингвистов сосредоточивается на сопоставлении словарных и корпусных данных (Чуйкова 2023). Возникает необходимость верификации грамматических норм в письменном дискурсе и типологизации узуса на корпусных данных (Кустова 2021, Letuchii 2018). Решение обозначенной проблемы видится в разработке надкорпусных баз данных, аннотированных вручную. Отметим, что корпусная лингвистика становится базовым инструментом для большинства исследователей именно благодаря появлению баз данных. Аннотирование текстовых корпусов приобретает все большую значимость, так как для повышения точности проводимых исследований пользователям необходим не только «сырой», но и аннотированный материал, т.е. размеченный с применением протоколов аннотирования и методов синтаксического анализа (Pons Bordería & Pascual 2021).

В отличие от корпусов надкорпусные базы данных позволяют хранить аннотированные данные. Иными словами, надкорпусная база данных – это определенным образом выстроенная коллекция структурированных данных. Кроме того, надкорпусные базы данных способны учесть фразеологические особенности языковых единиц, которые сложнее всего поддаются автоматической обработке. Целесообразно отметить, что до сих пор ведутся корпусные исследования в этой области, разрабатываются подходы и методология (Novakova & Siepmann 2020, Баранов, Добровольский 2021).

3. Материал и методология исследования

Надкорпусная база данных глагольных форм построена на параллельном русско-французском корпусе НКРЯ. Это открытый, постоянно пополняемый ресурс русских/французских полных оригинальных текстов, переведенных соответственно на французский/русский язык профессиональными переводчиками и опубликованные в известных издательствах. Для каждого текста указываются фамилии автора и переводчика, дата публикации оригинала и

перевода, сфера функционирования (учебно-научная, официально-деловая, художественная). При подборке текстов учитываются следующие критерии.

Во-первых, за редким исключением, отбираются переводы, сделанные во второй половине XX в., либо в начале XXI в., отражающие современные подходы к переводу и тенденции переводоведения. Во-вторых, чтобы отвечать критерию представительности, корпус должен быть сбалансирован. В описываемой надкорпусной глагольной базе данных данный критерий соблюдается. НГБД включает 61 текст (32 – на французском языке и 29 – на русском), общее количество слов – 7 123 534. Это сравнительно небольшой объем, особенно в сравнении с русско-английским параллельным корпусом в НКРЯ, в котором общее количество текстов –1189 текстов, слов – 44 477 958 (данные на 16.07.2023). Однако в процессе пополнения параллельного русско-французского корпуса в обязательном порядке учитывается жанровое разнообразие. Так, помимо беллетристики в состав корпуса входят русские и французские эссе по гуманитарным наукам: литературоведению (Gérard *Genette Figures I, II, III, Михаил Бахтин. «Вопросы литературы и эстетики»,* истории *Marie Pierre Rey 'L'Effroyable tragédie: Une nouvelle histoire de la campagne de Russie')*; культурологии (*Michel Pastoureau 'Bleu. Histoire d'une couleur');* антропологии (*Claude Lévi-Strauss 'L'Anthropologie face aux problèmes du monde moderne');* философии (*Николай Бердяев «О самоубийстве»).*

Художественные тексты на русском и французском языках включают в себя романы (детективы, автобиографии), рассказы, пьесы. Наличие в корпусе разных жанров – один из основных принципов его сбалансированности. Некоторые произведения приведены в нескольких переводах (Гоголь «Шинель» – 4 перевода, «Нос» – 4 перевода, Гончаров «Обломов» – 2 перевода, Толстой «Смерть Ивана Ильича» – 2 перевода, Jean Cocteau. «Difficulté d'être» – 2 перевода). Отметим, что при изучении языковых явлений и контекстного функционирования языковых единиц сравнение различных вариантов перевода (поливариантность) существенно обогащает исследование.

Для работы в параллельном корпусе или в базе данных оригинальный текст и его перевод предварительно выравниваются с целью совпадения смысловых и формальных блоков высказывания. Выравнивание выполняется с помощью компьютерных программ и обязательной последующей ручной редактурой, поскольку при переводе границы предложений и грамматических конструкций (клауз) могут не совпадать. Также необходимо учитывать, что при публикации перевода редко указываются выходные данные версии оригинала. Поскольку оригинальных изданий может быть несколько, параллельные тексты могут не совпадать.

Русские и французские тексты попадают в надкорпусную базу данных после автоматической морфологической разметки, выполненной анализатором MyStem компании Яндекс. При работе с глагольными формами особую сложность представляет собой неснятая омонимия некоторых лексем (*души,*

постели – повелительное наклонение 2 лица единственного числа), но уже начаты разработки по системному снятию омонимии в корпусе (Кустова и др. 2005).

Все тексты, входящие в основной корпус НКРЯ, содержат метатекстовую разметку (название текста, дата его создания, имя и год рождения автора, место и дата публикации, сфера функционирования, жанр и тип текста, хронотоп художественных произведений, мемуаров, целевая аудитория и др.) и лингвистическую разметку (морфологическую, словообразовательную, синтаксическую и семантическую). Морфологическая разметка в НКРЯ выполняется автоматически или вручную. Морфологическая разметка для параллельного русско-французского корпуса осуществляется с помощью специальных программ автоматического морфологического анализа и лемматизации. В НКРЯ возможно отобрать подкорпус по всем этим параметрам.

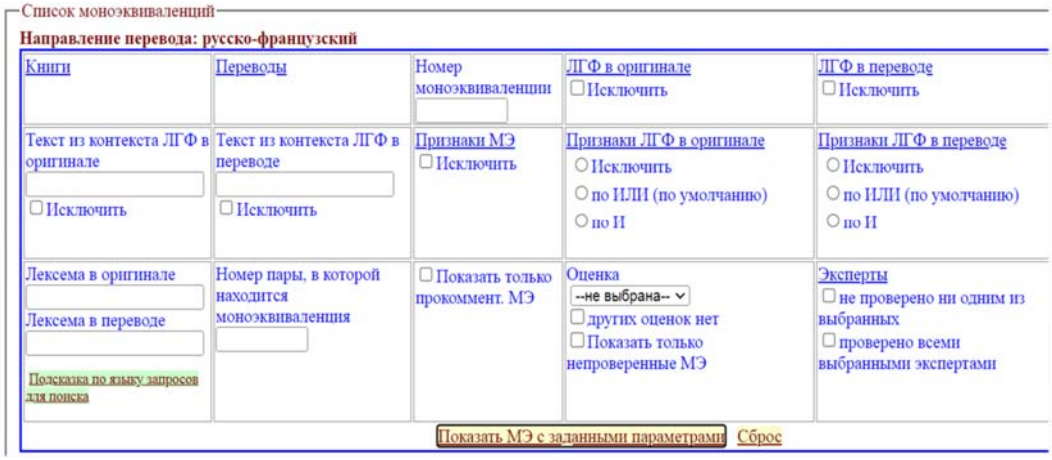

Рис. 1. Образец поискового запроса надкорпусной глагольной базы данных / Figure 1. Query by example in Supracorpora Verb Database
Направление перевода: русско-французский

Одно из первых описаний целей и задач, а также определение основных терминов: лексико-грамматическая форма (ЛГФ), моноэквиваленция (МЭ), полиэквиваленция (ПЭ) – были сформулированы в научных статьях, посвященных разработке этого ресурса. Лексико-грамматическая форма перевода, соответствующая некоторой ЛГФ оригинала, называется ее функционально эквивалентным фрагментом (сокращенно ФЭФ). Переводное соответствие, представляющее собой упорядоченную пару ЛГФ, ФЭФ, называется моноэквиваленцией (сокращенно МЭ) (Loiseau et al. 2013, Бунтман и др. 2014, Kruzhkov et al. 2014, Зализняк и др. 2015, 2016). Определение моноэквиваленции, а также подробное описание ее построения и аннотирования дается в статье А.А. Зализняк и М.Г. Кружкова на примере базы данных безличных глагольных конструкций (Зализняк, Кружков 2016). «Моноэквиваленции автоматически объединяются в полиэквиваленции в тех случаях,

когда в БД имеется несколько переводов одного и того же исходного текста; ценность полиэквиваленции как инструмента анализа состоит в том, что она показывает варианты перевода языковой единицы в одном и том же контексте» (Зализняк, Шмелев 2021:209).

В данной статье будет дано сравнение структуры тех кластеров НКРЯ и НГБД, которые непосредственно относятся к характеристикам глагола (данные по состоянию на 16.08.2023) Кроме того, будет подробно описана обоснованность создания кластеров и рубрик НГБД для поиска определенных языковых явлений на материале глаголов и глагольных конструкций.

4. Сравнение структуры параллельного французского корпуса НКРЯ и НГБД

Если необходимо найти точную глагольную форму, то в НКРЯ поиск точных форм и лексико-грамматический поиск возможны только в одном языке. Ниже на рис. 2 приведены результаты поискового запроса неопределенной формы глагола *собираться* во французском корпусе НКРЯ.

Left context	Center	Pun	Right context
должен был совет министров, которому предстояло	**собираться**		под председательством архиканцлера раз в неделю
прижал ночные кабаки, что тусовщикам осталось	**собираться**		только в барах, где танцы запрещены.
Мы любили	**собираться**		у меня всей компанией, и все
не производить шума, слушатели потихоньку стали	**собираться**	,	в храме началась тихая возня.
справки, посоветовалась с Буре и начала	**собираться**		в дорогу так, как если бы
Вечером стали	**собираться**		домой, но лошадей перекормили овсом, и
того, что она слишком рано начинает	**собираться**	.	

Рис. 2. Результат поиска инфинитива глагола *собираться* во французском корпусе НКРЯ / Figure 2. The result of the search for the infinitive of the verb *to gather* in the French corpus of the NCRL

Если в параллельном французском корпусе НКРЯ нужно найти варианты перевода в тексте оригинала, то пользователю предложена фраза в русском переводе с выделенной искомой лексико-грамматической формой и соответствующая фраза на французском языке. Переводной эквивалент не выделен. Отметим, что только в направлении французский → русский поисковый запрос в корпусе по лемме *собираться* выдает 325 примеров употребления этой лексемы со стимулом во французском и его переводе на русский язык. В этой связи А.А. Зализняк подчеркивает особое значение русского глагола *собираться* и выдвигает гипотезу о его «лингвоспецифичности», то есть сложности его перевода, и следовательно, вариативности функционально эквивалентных фрагментов. С помощью примеров из НКРЯ эта гипотеза не только подтверждается, но и обнаруживаются контекстуальные значения лексем, не зафиксированные в словарях (Зализняк 2006).

При работе с корпусом одновременно на двух языках имеет смысл поиск вариантов перевода той или иной морфосинтаксической категории либо

лингвоспецифической единицы. Поскольку предмет нашего исследования – глагольные формы и конструкции, рассмотрим кластеры и опции поиска в параллельном русско-французском корпусе НКРЯ и сравним их с теми же категориями в НГБД.

Таблица 1. **Контекст глагола *собираться* в оригинале и переводе во французском корпусе НКРЯ /** *Table 1.* **The context of the verb *to gather* (*s'agglutiner*) in the original and translation in the French corpus NCRL**

Full context	Para context 1	Lang 1
Эта мода залетела к нам из Нью-Йорка: тамошний мэр так прижал ночные кабаки, что тусовщикам осталось **собираться** только в барах, где танцы запрещены.	La mode vient de New York : là-bas, le maire a tellement restreint les autorisations de boîtes de nuit que tous les fêtards s'**agglutinent** dans des bars où il est interdit de danser.	**fra**

В НКРЯ лексико-грамматический поиск глагольных форм в направлении русский-французский эффективнее, поскольку механизм лемматизации и разметки более проработан. В направлении французский → русский возникают сложности, в основном связанные с неснятой омонимией, неточной и недостаточной лемматизацией. Так, при запросе кластера «время» обнаруживаются существительные *épaule, porte(s), monde, adresse, rue* . Также по тем же причинам фактически невозможен поиск конкретных времен. В кластере «время» есть лишь настоящее, будущее (без уточнения, какое именно), прошедшее (только *passé simple*) время с учётом наклонений.

Теперь обратимся к структуре надкорпусной базы данных глагольных форм. Прежде всего, надкорпусная база данных отличается от корпуса тем, что, помимо полученной автоматической лемматизации и разметки текста, происходит «ручное» аннотирование выровненной пары предложений и последующая лингвистическая экспертиза.

В современной версии НКРЯ представлена возможность просмотра и сопоставления нескольких переводческих версий одного и того же фрагмента оригинала. Так, если обратиться к четырем переводам повести Н.В. Гоголя «Нос», при рассмотрении нескольких переводов русского глагола состояния *лежать*, наблюдается вариативность глагольных форм во французском:

русский:
Чтобы я позволила у себя в комнате лежать отрезанному носу?
французский 1:
Que je permette qu'on laisse dans ma chambre un nez coupé ?
французский 2:
Que je permette, moi, de laisser traîner dans ma chambre un nez coupé !
французский 3:
Crois-tu, par hasard, que je vais garder ici un nez coupé ?
французский 4:
Que je permette à un nez coupé ***de rester*** *dans ma chambre ?*

В надкорпусных базах данных (на данный момент в ФИЦ РАН существуют и разрабатываются следующие ресурсы: БД глагольных форм, БД коннекторов, БД лингвоспецифичных единиц, БД безличных глагольных форм, БД анафоры, БД немецких модальных глаголов) предусмотрена возможность запроса полиэквиваленции, то есть различных вариантов словоформы оригинала. Ниже в табл. 2 мы видим, что русский глагол НСВ переводится на французский язык двумя различными способами, причем в одном из случаев – глаголом совершенного вида.

Таблица 2. **Результат запроса глаголов несовершенного вида в полиэквиваленциях /**
Table 2. **Query result for imperfective verbs in polyequivalences**

Контекст ЛГФ в оригинале	**ЛГФ в оригинале и ее признаки**	**ЛГФ в переводе**	
		Контекст ЛГФ в переводе	ЛГФ в переводе и ее признаки
Между титулярным советником и коллежским асессором **разверзалась** бездна,	Past-IPF	Un abîme **s'ouvrit** entre le conseiller honoraire et le conseiller de collège,	PasSim
		entre le conseiller titulaire et l'assesseur de collège, un abîme s'ouvrait,	Imparf

Основное отличие НГБД от НКРЯ заключается в том, что при аннотировании МЭ выделяются переводные соответствия, по которым впоследствии можно осуществлять поиск. В табл. 3 в одной версии глагол *лежать* заменяется глаголом *se vautrer (валяться*) со сменой стилистического регистра, во второй – глаголом *dormir (спать)* со сменой значения, поскольку в данном контексте важно, что главный персонаж романа – Обломов лежит целыми днями и не только во время сна.

Таблица 3. **Пример полиэквиваленции с выделенными функционально эквивалентными фрагментами (ФЭФ) /**
Table 3. **An example of polyequivalence with isolated functionally equivalent fragments**

что Обломов [...] только лежит да кушает	Oblomov ne faisait que se vautrer et s'empiffrer
	tous croyaient qu'il ne faisait que manger et dormir tout son soûl,

5. Кластеры и рубрики надкорпусной базы данных глагольных форм

Как ранее было уже указано, структура описываемой в статье НГБД состоит из нескольких кластеров и рубрик, ориентированных на исследование специфики контекстной грамматической семантики, распределение глагольных форм в русском и французском языках, а также на развитие теории соответствий и переводческих трансформаций.

Рассмотрим более подробно алгоритм кластеризации и рубрикации данных, используемый специалистами ФИЦ ИУ РАН. В кластер «особенности

МЭ» вошли общие характеристики МЭ, которые не относятся ни к базовому виду ЛГФ (в нашем случае форме конкретного глагола) русского или французского языка, ни к дополнительным признакам определенной ЛГФ, которые будут рассмотрены позже). Особенности МЭ затрагивают сложные переводческие лексико-синтаксические трансформации. Помета Nota bene! присваивается МЭ с неочевидным контекстуальным вариантом перевода глагольной формы, найденным в тексте.

Таблица 4. **Кластер «особенности моноэквиваленции» /**
Table 4. **Cluster "peculiarities of monoequivalence"**

Nota bene!	NB
Смена подлежащего	SubjCh
Paraphr	paraphrase
требуется экспертная оценка	Exp

Таблица 5. **Пример моноэквиваленции с пометой NB /**
Table 5. **An example of a monoequivalence with the label NB**

Врешь!	Pres-IPF <Exclam > <DialRepl >	Foutaises!	Subst ☐ <Exclam > ☐ <DialRepl >	☑ NB ☐ SubjCh ☐ paraphrase ☐ Exp

В примере из табл. 5 контекстное значение глагола *врать*, употребленного в краткой диалогической реплике, заменено во французском на существительное со значением *вранье*. Данная переводческая находка, не зафиксированная ни в одном из двуязычных русско-французских словарей, помечается тэгом Nota bene!

Для исследований в области теории и практики перевода, а также асимметрии тема-рематического членения высказывания особый интерес может представлять рубрика смены подлежащего SubjCh (Subject change). При этом можно, например, проследить, как безличная конструкция в русском оригинале заменяется на выраженного субъекта в переводе и наоборот (табл. 6).

Смена субъекта в ПЯ является проблемой общей теории перевода, но может быть рассмотрена в рамках частной теории перевода в языковой паре русский-французский. Контекстные переводческие варианты не только обогащают лексические соответствия, но и дополняют положения грамматики конструкций.

Проблема перифразирования подробно была рассмотрена Е.Л. Туницкой (Туницкая 2010). Отметим, что это не окончательно устоявшийся термин. В научных исследованиях встречаются следующие варианты: употребление перифразы (Есменская 2002, Бытева 2004) и перифраза (Сиривля 2004). В кластер общих характеристик МЭ введена рубрика paraphrase, позволяющая

отобрать и проанализировать случаи с перефразированием в переводе глагольных форм и конструкций.

Таблица 6. **Примеры моноэквиваленций с пометой SubjCh /**
Table 6. **Examples of monoequivalences labeled SubjCh**

что от пыли **заводится моль**?	Pres-IPF <SubCompl >	que **la poussière engendre** les mites?	Pres ☐ <SubCompl >	☐ NB ☑ SubjCh ☐ paraphrase ☐ Exp
Depuis six mois que tu es ici, **tu n'as pas eu** un seul ennui…	PasCom <TempDet > <Neg > <DialRepl >	С тех пор как ты здесь, **неприятностей у тебя не было**…	Past-IPF ☐ <TempDet > ☐ <Neg > ☐ <DialRepl >	☐ NB ☑ SubjCh ☐ paraphrase ☐ Exp

Таблица 7. **Примеры моноэквиваленций с пометой «paraphrase» /**
Table 7. **Examples of monoequivalences marked "paraphrase"**

Ai-je insisté?	PasCom <Interrog> <DialRepl>	Разве я не отказался от своих намерений?	Past-PF ☐ <ModDet > ☐ <Neg > ☐ <DialRepl >	☐ NB ☐ SubjCh ☑ paraphrase ☐ Exp
Только сегодня и надеюсь вздохнуть.	Pres-IPF <SubInf-PF >	Si je ne souffle pas un peu aujourd'hui, alors quand?	Pres ☐ <Neg> ☐ <SiCond>	☐ NB ☐ SubjCh ☑ paraphrase ☐ Exp

В НКРЯ список категорий лексико-грамматического поиска фактически идентичен для русского и французского языков, в частности, в кластере «наклонение» поиск сослагательного и условного возможен только для французского языка. В НГБД для ЛГФ русского языка созданы специальные рубрики для более точного поиска соответствующих глагольных форм.

Системная асимметрия форм условного наклонения принуждает французского переводчика к выбору одной из двух форм наклонения Conditionnel Présent или Conditionnel Passé, исходя из временных или дискурсивных маркеров узкого или широкого контекста языка оригинала. Анализ того, каким образом подобные маркеры обусловливают переводческое решение, представляет отдельную лингвистическую проблему. Так, в табл. 9 разница в значении финитной глагольной формы несовершенного вида с «бы» заключается

в семантике высказываний, что и ведет за собой употребление разных французских ЛГФ.

Таблица 8. **Кластеры русских глагольных форм, соответствующие французскому условному и сослагательному наклонениям /**
Table 8. **Clusters of Russian verb forms corresponding to the French conditional and subjunctive moods**

Форма с «бы» НСВ	Past-IPF+бы
Форма с «бы» СВ	Past-PF+бы
Форма с «было» НСВ	Past-IPF+было
Форма с «было» СВ	Past-PF+было
Форма с «если бы» НСВ	Past-IPF+если бы
Форма с «если бы» СВ	Past-PF+если бы
Форма с «чтобы» НСВ	Past-IPF+чтобы
Форма с «чтобы» СВ	Past-PF+чтобы

Таблица 9. **Примеры моноэквиваленций по запросу Past-IPF+бы /**
Table 9. **Examples of monoequivalences on request Past-IPF+бы**

Ведь и я бы мог все это…	Past-IPF+бы <DialRepl >	Moi aussi **j'aurais pu faire** tout ça,	CondPas ☐ <SubInf > ☐ <ModDet > ☐ <DialRepl >
И я бы тоже… хотел… – […] что-нибудь такое…	Past-IPF+бы <DialRepl >	Moi aussi... **je voudrais**... […]quelque chose de ce genre...	CondPr ☐ <ModDet> ☐ <DialRepl>

Рубрика «форма с *было* несовершенного и совершенного вида» была добавлена в кластер признаков ЛГФ глагольных форм русского языка, исходя из вариативности переводов и отсутствия переводных соответствий таких глагольных конструкций в словарях. Уже в начале поиска выяснилось, что данная конструкция представляет особую переводческую сложность. Например, при запросе Past-IPF+было в функционально-эквивалентном фрагменте (термин Д.О. Добровольского) обнаруживаются такие французские ЛГФ, как Imparfait, Conditionnel Passé, Passé Simple.

Сослагательное наклонение в кластере французских ЛГФ представлено четырьмя рубриками: Subjonctif Présent (SubjPres), Subjonctif Passé (SubjPas), Subjonctif Imparfait (SubjImparf) и Subjonctif plus-que-parfait (SubjPqParf). Для возможности двуязычного поиска в кластере ЛГФ русского языка есть две рубрики форм с «*чтобы*» несовершенного и совершенного вида.

В отличие от рубрики «будущее» кластера «время» русского языка НКРЯ, где помимо шума, вызванного неснятой омонимией, в составном будущем выделен лишь глагол «быть», а не значащий глагол, в НГБД в русском

языке представлены простое будущее (Fut-PF) и сложное будущее (Fut-IPF). Для французского языка в кластере «время» существуют четыре рубрики Futur Simple (Fut), Futur Immédiat (FutIm), Futur antérieur (FutAnt), Futur immédiat dans le passé (FutImPas).

Рубрики прошедшего времени в НКРЯ представлены прошедшим (ru), passé simple (fr), имперфектом (fr), причастием прошедшего времени (fr), то есть далеко неполным кластером. В НГБД помимо причастных форм настоящего Participe Présent (PartPr) и прошедшего времени (не всегда являющейся частью составного прошедшего времени). В НКРЯ поиск деепричастий возможен лишь в русском языке, в НГБД осуществляется поиск gérondif по французскому тексту. В ходе работы с НГБД в направлении русский → французский выяснилось, что при построении МЭ необходима рубрика Omission (Zero), то есть отсутствие (системное или контекстуальное) в переводе языковой единицы, соответствующей ЛГФ оригинала. Системное отсутствие ЛГФ перевода наблюдается в процессе трансформации русских глаголов говорения (verba dicendi) – формы *говорил, спросил, добавил, сказал, прибавил, перебил, возразил, продолжал* и проч. на французский язык.

Таблица 10. **Результат запроса моноэквиваленции, содержащей глагол, вводящий прямую речь /** *Table 10.* **The result of a request for a monoequivalence containing a verb that introduces direct speech**

«Какая у вас **пыль** везде!» —сказал он.	Past-PF <VerbDirSp >	Comme **c'est poussiéreux** chez vous !	Zero

Ранее речь шла о трудностях перевода с русского на французский глаголов состояния. Поиск по рубрике Zero позволяет подтвердить это предположение.

Таблица 11. **Результат сложного запроса моноэквиваленций по базовому виду ЛГФ Past-PF и Zero /** *Table 11.* **The result of a complex query of monoequivalences for the basic type of LGF Past-PF and Zero**

как сел.	Past-PF	dans cette même position.	Zero
на столе редкое утро [...] не валялись хлебные крошки.	Past-IPF <Neg >	rares étaient les matins où la table ne portait pas, [...] parmi des miettes de pain, un os rongé	Zero
Ведь я вот тут лежал [...]	Past-IPF <ModDet > <DialRepl >	C'est tout simplement que je réfléchis à la manière dont je puis sortir d'embarras...	Zero

Одним из распространенных видов языковых трансформаций является смена морфологического статуса лексемы (Гак 1998). В НГБД в направлении русский-французский более ста примеров перевода русской глагольной формы существительным.

Таблица 12. **Результат запроса перевода глагола в русском оригинале существительным во французском /**
Table 12. **The result of the request to translate the verb in the Russian original by the noun in French**

можно было бы подумать, что **тут никто не живет**!	Pres-IPF <SubInf-PF > <Neg > <SubCompl >	on aurait pu croire **la chambre** inhabitée,	Subst
что может **стоить** постройка собачьей конуры?	Pres-IPF <SubInf-IPF > <ModDet > <SubCompl >	à se renseigner sur **le coût** de la construction d'une niche.	Subst

Кластер «дополнительные признаки ЛГФ» в оригинале и переводе допускает поиск по видам предложений (вопросительное, восклицательное), видам придаточных, по видам подчинительных конструкций.

Таблица 13. **Кластер «дополнительные признаки ЛГФ в переводе в НГБД /**
Table 13. **Cluster "additional signs of LGF translated into NGBD"**

Accusativus cum infinitivo	**Acc.c.Inf**
Вопросительное предложение	Interrog
Восклицательное предложение	Exclam
Временной детерминант	TempDet
Глагол в изъяснительном придаточном	SubCompl
Глагол в определительном придаточном	SubAttr
Глагол, вводящий прямую речь	VerbDirSp
Диалогическая реплика	DialRepl
Модальный детерминант	ModDet
Отрицание	Neg
Подчиненный инфинитив	SubInf
Подчиненный инфинитив прош. времени	SubInfPas
Подчиняющий глагол	+SuperVerb
Придаточное в условном предложении	Si Cond
Прочие виды придаточных	Sub

Временные (*давно, долго, недавно, теперь, много лет, никогда, сначала*) и модальные (*вот, же, всё, уж, хоть бы, разве*) детерминанты позволяют уточнить функционирование глагольной формы в высказывании и его семантику. Таким образом, при передаче аспектуальности в переводе наличие определенного временного маркера определяет выбор глагольной формы. Если в НГБД сделать запрос TempDet в направлении французский → русский, то при употреблении в оригинале наречия *longtemps* и сохранении соответствующего временного детермината в переводе выявится частотная асимметрия PasSim/PasCom (fr) Past-IPF (ru). Однако при изменении семантического окружения *longtemps* в переводе может употребляться глагол совершенного вида (см. табл. 14).

Таблица 14. **Результаты запроса TempDet в направлении французский-русский /**
Table 14. **TempDet query results in French-Russian direction**

par laquelle je criai aussi longtemps que je le pus.	**PasSim** <TempDet > <SubAttr >	**стал кричать** из последних сил	**Past-PF** ☐ <SubInf-IPF >
Mais cela fait longtemps que Philippe **a tourné les talons**	**PasCom** <TempDet > <ModDet >	Филипп давно уже сбежал.	**Past-PF** ☐ <TempDet > ☐ <ModDet >

В целом, описанная в статье надкорпусная глагольная база обладает некоторыми функциональными преимуществами в сравнении с параллельным русско-французским корпусом НКРЯ. Например, пользователям НГБД доступна опция просмотра статистики вариантов перевода базовых видов глагольных ЛГФ: если НКРЯ выдаёт только параллельные тексты с выделенным поисковым словом в ИЯ, не выделяя при этом переводное соответствие в ПЯ, то НГБД не только показывает и выделяет переводные соответствия в ИЯ и ПЯ, но и позволяет увидеть наиболее и наименее частотные варианты перевода для той или иной лексической единицы. Данная опция представляет особый интерес для профессиональных переводчиков и специалистов в области переводоведения, поскольку позволяет обнаружить как частотные ФЭФ, так и их малочастотные варианты, и, следовательно, подобрать наиболее подходящий вариант перевода на основе статистических данных. Так, в ходе анализа переводов французских текстов художественной литературы и эссеистики (32 текста) было установлено, что русские глаголы несовершенного вида переводятся французскими глаголами в Passé Simple в 15,27 % случаев (445 МЭ) и Passé composé в 6,1 % случаев (178 МЭ) (данные на 14.08.2023).

Кроме того, НГБД обладает расширенной поисковой функцией, которая обеспечивает быстрый и эффективный поиск базовых видов ЛГФ глаголов, а именно предоставляет пользователям возможность: обнаружить точные формы глаголов в языковой паре русский – французский; выделить переводной эквивалент с учетом нескольких переводов оригинала; осуществлять более точный поиск соответствующих глагольных форм в кластерах «наклонение» и «время». Ниже в таблице 15 представлены сравнительные характеристики поисковых возможностей базовых видов ЛГФ глаголов в параллельном французском корпусе НКРЯ и надкорпусной базе данных глагольных форм (данные по состоянию на 16.08.2023).

Таблица 15. **Возможности поиска базовых видов ЛГФ глаголов в параллельном русско-французском корпусе НКРЯ и Надкорпусной базе данных глагольных форм /**
Table 15. **Possibilities of searching for the basic types of verbs' LHF in the parallel French corpus of NCRL and the Supracorpora database of verb forms**

Опции поиска	НКРЯ	НГБД
Поиск точных форм	Возможен в одном языке	Возможен в двух языках
Выделение переводного эквивалента	нет	Выделен переводной эквивалент
Поливариантность (несколько переводов оригинала)	есть	есть
Кластер «сослагательное наклонение» в русском	нет	Форма с «чтобы» НСВ и СВ
Кластер «сослагательное наклонение» во французском	Сослагательное (fr)	Subjonctif Présent Subjonctif Passé Subjonctif Imparfait Subjonctif plus-que-parfait
Кластер «условное наклонение» в русском	нет	Форма с «если» НСВ и СВ Форма с «если бы» НСВ и СВ
Кластер «условное наклонение» во французском	Условное (fr)	Conditionnel Présent Conditionnel Passé
Кластер «будущее время» во французском языке	будущее	Futur Simple Futur Immédiat Futur antérieur Futur immédiat dans le passé

5. Заключение

В данной статье мы ставили цель – представить структуру и функционал надкорпусной глагольной базы данных (НГБД), разработанной на основе параллельного русско-французского подкорпуса Национального корпуса русского языка (НКРЯ), а также раскрыть разницу их потенциалов.

В отличии от разработанной НГБД параллельный русско-французский подкорпус НКРЯ, как и большинство лингвистических ресурсов, в которых встречаются инструменты аннотирования, ориентирован на разметку текста только на одном языке исследуемой языковой пары. Его опции позволяют генерерировать и сохранять структурированную информацию об употреблении наименований объектов и явлений внеязыковой действительности. Однако этого недостаточно, если перед исследователем стоит задача не просто найти примеры интересующих его языковых единиц (в нашем случае глаголов и глагольных форм), но и оценить уровень их частотности в некоторой совокупности текстов корпуса или в корпусе в целом. Надкорпусная глагольная база данных, напротив, предоставляет более гибкий функционал пользователям, а именно предусматривает возможность изучения отдельных характеристик глаголов (времени, наклонения, аспектуальности, соотношения финитных и нефинитных форм), семантики глагольных лексем и коллокаций во всех отобранных текстах одного и того же параллельного корпуса.

В статье также показано, что наиболее значимым элементом разработанной надкорпусной базы данных глагольных форм на материале параллельного русско-французского корпуса НКРЯ является переводное соответствие, включающее глагольную форму оригинала и ее функционально-эквивалентный фрагмент в тексте одного или нескольких переводов. Двуязычный лексико-грамматический поисковый интерфейс позволяет искать соответствия тех или иных глагольных форм, а также получать данные о частотности этих соответствий. Благодаря тому, что языковой материал аннотируется вручную и тщательно выверяется, в построенных переводных соответствиях (моноэквиваленции) снята омонимия и нет шума. Аннотации формируются экспертами в результате последовательной обработки глагольных единиц. Обработка корпуса может одновременно осуществляется несколькими экспертами, что закономерно при проведении масштабных проектов. Таким образом, надкорпусные базы данных обеспечивают возможность интегрировать плоды их совместной работы, предоставить другим исследователям доступ и к полученным результатам, и ко всем использованным параллельным текстам. Это значительно упрощает процессы верификации полученных результатов и позволяет многократно использовать созданный информационный ресурс для решения многих научных и практических задач: лингводидактика и разработка учебных пособий, обучение устному и письменному переводу, обучение постредактированию машинного перевода, создание алгоритмов машинного обучения для тренировки нейросетей в машинном переводе, разработка терминологических и словарных баз данных, моноязычных и двуязычных словарей частотности.

Финансирование и благодарности

Работа выполнена в рамках проекта № 050738–0-000 системы грантовой поддержки научных проектов РУДН.

Acknowledgements

This paper has been supported by the RUDN University Scientific Projects Grant System, project No 050738–0-000.

References

Баранов А.Н., Добровольский Д.О. Об одном подходе к количественной оценке идиоматичности текста как характеристике авторского стиля // *Компьютерная лингвистика и интеллектуальные технологии: материалы ежегодной международной конференции «Диалог 2021»*. Т. 20. М.: РГГУ, 2021. С. 58–67. [Baranov, Anatoly N. & Dimitri O. Dobrovol'skij. 2021. Idiomaticity of a Text as a Matter of the Individual Style: A Quantitative Approach. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii (Computational Linguistics and Intellectual Technologies). Proceedings of the Annual International Conference 'Dialog 2021.'* Vol. 20. 58–67. Moscow: RSUHU Publ. (In Russ.)].

Баранов А.Н. Корпусный эксперимент в лингвистической экспертизе // *Компьютерная лингвистика и интеллектуальные технологии*: *материалы ежегодной международной конференции «Диалог 2022»*. Т. 21. М.: РГГУ, 2022. С. 42–49. [Baranov, Anatoly N. 2022. Corpus experiment in forensic linguistics. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii (Computational Linguistics and Intellectual Technologies). Proceedings of the Annual International Conference 'Dialog 2022*. Vol. 21. 42–49. Moscow: RSUHU Publ. (In Russ.)].

Богуславский И.М., Григорьев Н.В., Григорьева С.А., Иомдин Л.Л., Крейдлин Л.Г., Санников В.З., Фрид Н.Е. Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации // *Труды международного семинара «Диалог 2000».* [Boguslavskii, Igor' M., Nikolai V. Grigor'ev, Svetlana A. Grigor'eva, Leonid G. Kreidlin, Vladimir Z. Sannikov & Nina A. Frid. 2000. Annotirovannyi korpus russkikh tekstov: kontseptsiya, instrumenty razmetki, tipy informatsii (An annotated corpus of Russian texts: concept, markup tools, types of information.). *Proceedings of the International Seminar 'Dialog 2000'*. (In Russ.)].

Бунтман Н.В., Зализняк А.А., Зацман И.М., Кружков М.Г., Лощилова Е.Ю., Сичинава Д.В. Информационные технологии корпусных исследований: принципы построения кросс-лингвистических баз данных // *Информатика и ее применения*. 2014. Т. 8. № 2. С. 98–110. [Buntman, Nadezhda V., Anna A., Zaliznyak, IIgor' M. Zatsman, Mikhail G. Kruzhkov, G., Elena Yu. Loshchilova & Dmitrii V. Sichinava. 2014. Information technologies for corpus studies: Underpinnings for cross-linguistic database creation. *Informatics and Applications* 8 (2). 98–110. (In Russ.)].

Бытева Т.И. Основы лингвистической теории перифразы. Красноярск: КрасГУ, 2004. [Byteva, Tat'yana I. 2004. *Osnovy lingvisticheskoi teorii perifrazy (Fundamentals of the Linguistic Theory of Periphrase).* Krasnoyarsk: KraSGU Publ. (In Russ.)].

Гак В.Г. Языковые преобразования. М.: Школа «Языки русской культуры», 1998. [Gak, Vladimir G. 1998. *Yazykovye preobrazovaniya (Language Transformations).* Moscow: Shkola «Yazyki russkoi kul'tury». (In Russ.)].

Даровских Ю.А. Сопоставительный анализ семантики грамматических средств выражения аспектуальности в русском и французском языках // *Риторика – Лингвистика*. 2020. Т. 15. С. 76–89. [Darovskikh, Yuliya A. 2020. Comparative analysis of the semantics of grammatical aspect in Russian and French. *Ritorika – Lingvistika* 15. 76–89. (In Russ.)].

Добровольский Д.О., Кретов А.А., Шаров С.А. Корпус параллельных текстов: архитектура и возможности использования // *Национальный корпус русского языка: 2003–2005.* М.: Индрик, 2005. С. 263–296. [Dobrovol'skii, Dmitrii O., Aleksei A. Kretov & Sergei A. Sharov. 2005. Korpus parallel'nykh tekstov: arkhitektura i vozmozhnosti ispol'zovaniya (Corpus of parallel texts: Architecture and possibilities of use). *Natsional'nyi korpus russkogo yazyka: 2003–2005.* Moscow: Indrik. 263–296. (In Russ.)].

Есменская Н.А. Явление перифразы в аспекте смысловой связности текста // *Актуальные проблемы французской филологии*. *Сборник научных трудов.* Т.2. М.: 2002. С. 52–55. [Esmenskaya, Natal'ya A. 2002. Yavlenie perifrazy v aspekte smyslovoi svyaznosti teksta (The Phenomenon of Paraphrase in the Aspect of the Semantic Coherence of the Text). *Aktual'nye problemy frantsuzskoi filologii*. *Sbornik nauchnykh trudov*. Vol. 2. Moscow.: 52–55. (In Russ.)].

Зализняк А.А. Многозначность в языке и способы ее представления. М.: Языки славянской культуры, 2006. [Zalizniak, Anna A. 2006. *Mnogoznachnost' v yazyke i sposoby predstavleniya (Language Polysemy and Means of its Representation).* Moscow: Yazyki slavyanskoi kul'tury. (In Russ.)].

Зализняк А.А., Шмелев А.Д. Исследования по русской и компаративной семантике. М.: Издательский Дом. ЯСК, 2021. [Zalizniak, Anna A., Alexei D. Shmelev. 2021. Issledovaniya po russkoi i comporativnoi semantike (Studies on Russian and Comparative Semantics). Moscow: Izdatel'skii dom. Yask (In Rus.)].

Зализняк А.А., Зацман И.М., Инькова О.Ю., Кружков М.Г. Надкорпусные базы данных как лингвистический ресурс // *Труды международной конференции «Корпусная лингвистика-2015»*. СПб.: 2015. С. 211–218. [Zaliznyak, Anna A., Igor M., Zatsman, Olga U. Inkova & Mikhail G. Kruzhkov. 2015. Supracorpora databases as linguistic resource. *Proceedings of the Annual International Conference 'Corpus Linguistics-2015*. Saint Petersburg. 211–218. (In Russ.)].

Зализняк А.А., Кружков М.Г. База данных безличных глагольных конструкций русского языка // *Информатика и ее применения*, 2016. Т. 10. № 4. С. 132–141. [Zalizniak, Anna A. & Mikhail G. Kruzhkov. 2016. Database or Russian impersonal verbal constructions. *Informatics and Applications* 10 (4). 132–141. (In Russ.)].

Захаров В.П., Богданова С.Ю. Корпусная лингвистика. СПб.: Изд-во СПбГУ, 2020 [Zaharov, Viktor P. & Svetlana Yu. Bogdanova. 2020. *Korpusnaya lingvistika (Corpus Linguistics).* Saint Petersburg: Saint Petersburg University Publ. (In Russ.)].

Инькова О.Ю., Кружков М.Г. Надкорпусные русско-французские базы данных глагольных форм и коннекторов // *Славянские языки in comparatione*: *материалы IV Международной конференции по контрастивной лингвистике GELiTeC 2016*. Изд-во: Bergamo University Press, 2016. 365–392. [Inkova, Olga U. & Mikhail G. Kruzhkov. 2016. Nadkorpusnye russko-frantsuzskie bazy dannykh glagol'nykh form i konnektorov (Supracorpora Russian-French databases of verb forms and connectors). *Slavyanskie yazyki in comparatione (Slavic Languages in Contrast). Proceedings of the International Conference on Contrastive Llinguistics 'GELiTeC 2016'*. Bergamo University Press. 365–393. (In Russ.)].

Кружков М.Г. Информационные ресурсы контрастивных лингвистических исследований: электронные корпуса текстов // *Системы и средства информатики.* 2015. Т. 25. № 2. С. 140–159. [Kruzhkov, Mikhail G. 2015. Information resources for contrastive studies: Electronic text corpora. *Sistemy i Sredstva Informatiki* 25 (2). 140–159. (In Russ.)].

Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы / / Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: Индрик, 2005. С. 155–174. [Kustova, Galina I., Ol'ga N. Lyashevskaya, Elena V.Paducheva & Ekaterina V. Rakhilina. 2005. Semanticheskaya razmetka leksiki v Natsional'nom korpuse russkogo yazyka: printsipy, problemy, perspektivy (Semantic Markup of Vocabulary in the National Corpus of the Russian Language: Principles, Problems, Prospects). *Natsional'nyi korpus russkogo yazyka: 2003–2005.* Moscow: Indrik. 263–296. (In Russ.)].

Кустова Г.И. Типы инфинитивных конструкций с предикативами (по данным Национального корпуса русского языка) // *Компьютерная лингвистика и интеллектуальные технологии: материалы ежегодной международной конференции «Диалог».* Т. 20. Москва: РГГУ, 2021. С. 456–463. [Kustova, Galina I. 2021. The types of infinitive constructions with predicatives (according to the Russian National Corpus). *Komp'yuternaya lingvistika i intellektual'nye tekhnologii* (*Computational Linguistics and Intellectual Technologies). Proceedings of the Annual International Conference 'Dialog'*. Vol. 20. 456–463. Moscow: RSUHU Publ. (In Russ.)].

Кустова Г.И. Электронный семантический словарь глагольных прилагательных: структура и типы информации // *Компьютерная лингвистика и интеллектуальные технологии*: *материалы ежегодной международной конференции «Диалог-2009».* М.: 2009. С. 271–277. [Kustova, Galina I. 2009. The semantic database of verbal adjectives: Structure and types of information. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii (Computational Linguistics and Intellectual Technologies). Proceedings of the Annual International Conference 'Dialog-2009*. Moscow: RGGU. 271–277. (In Russ.)].

Кустова Г.И. Электронный словарь степенной сочетаемости на базе Национального корпуса русского языка // *Труды международной конференции «Корпусная лингвистика* – 2008». СПб.: 2008. С. 132–149. [Kustova, Galina I. 2008. Ehlektronnyi slovar' stepennoi sochetaemosti na baze Natsional'nogo korpusa russkogo yazyka (Electronic dictionary of power combination based on the national corpus of the Russian language). *Proceedings of the International Conference 'Corpus Linguistics – 2008*. Saint Petersburg. 132–149. (In Russ.)].

Прикладная и компьютерная лингвистика / под. ред. И.С. Николаева, О.В. Митрениной, Т.М. Ландо. М.: Ленинград, 2017. [Nikolaev, Il'ya S., Olga V. Mitrenina, Tat'yana M. Lando. (eds.). 2017. *Prikladnaya i komp'yuternaya lingvistika (Applied and Computer Linguistics)*. Moscow: Leningrad. (In Russ.)].

Сиривля М.А. Перифраз в современной лингвистике // *Теоретические и методологические аспекты языкознания: материалы международной научно-практической конференции*. Алматы: АГУ. 2004. С. 43–47. [Sirivlya, Madina A. 2004. Perifraz v sovremennoi lingvistike (Paraphrase in modern linguistics). *Teoreticheskie i metodologicheskie aspekty yazykoznaniya (Theoretical and Methodological Aspects of Linguistics). Proceedings of the International Research and Practice Conference*. Almaty: AGU. 2004. 43–47. (In Russ.)].

Сичинава Д.В. Параллельные тексты в составе Национального корпуса русского языка: Новые языки и новые задачи. // *Труды Института русского языка им. В.В. Виноградова*. 2019. № 21. С. 41–60. [Sitchinava, Dmitri V. 2019. On parallel texts within the Russian national corpus: New languages and new challenges. *Trudy Instituta Russkogo Yazyka imeni V. V. Vinogradova* 21. 41–60. (In Russ.)].

Туницкая Е.Л. Перефразирование в лингвопрагматическом аспекте на материале французского дискурса. М.: Издательский центр института всеобщей истории РАН, 2010. [Tunitskaya, Elena L. 2010. Perifrazirovanie v lingvopragmaticheskom aspekte na materiale frantsuzskogo diskursa (*Paraphrasing in the Linguo-pragmatic Aspect Based on French Discourse.*). Moscow: Izdatel'skii tsentr instituta vseobshchei istorii RAN. (In Russ.)].

Чуйкова О.Ю. Родительный партитивный в русском языке: словарные и корпусные данные // *Компьютерная лингвистика и интеллектуальные технологии: материалы ежегодной международной конференции «Диалог».* Т. 22. М.: РГГУ, 2023. С. 42–50. [Chuikova, Oksana Yu. 2023. Partitive genitive in Russian: Dictionary and corpus data. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. (Computational Linguistics and Intellectual Technologies). Proceedings of the Annual International Conference 'Dialog'*. Vol. 22. Moscow: RSUHU Publ. 42–50. (In Russ.)].

Шаров С.А. Представительный корпус русского языка в контексте мирового опыта // *Научно- техническая информация.* 2003. Т.2. № 6. С.12–16. [Sharov, Sergei A. 2003. Predstavitel'nyi korpus russkogo yazyka v kontekste mirovogo opyta (Representative corpus of the Russian language in the context of world experience). *Nauchno-tekhnicheskaya informatsiya* 2 (6). 12–16. (In Russ.)].

Kruzhkov, Mikhail, Nadezhda V. Buntman, Elena Yu. Loshchilova, Dmitri V. Sitchinava, Anna A. Zalizniak & Igor. M. Zatsman. 2014. The database of Russian verbal forms and their French translation equivalents. *Computational Linguistics and Intellectual Technologies*. Proceedings of the Annual International Conference 'Dialog-2014'. Moscow: RGGU. 275–287.

Letuchii, Alexandre B. 2018. Predicatives. Materials for the corpus grammar of the Russian language. No. III. *Parts of Speech and Lexical and Grammatical Classes.* Saint Petersburg: Nestor- Istoriya. 136–192.

Loiseau, Sébastien, Dmitri V. Sitchinava, Anna A. Zalizniak & Igor M. Zatsman. 2013. Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus. *Informatics and Applications* 7 (2). 100–109.

Novakova, Iva & Dirk Siepmann. 2020. *Phraseology and Style in Subgenres of the Novel: A Synthesis of Corpus and Literary Perspectives.* London: Palgrave Macmillan.

Pons Bordería, Salvador & Elena Pascual Aliaga. 2021. Inter-annotator agreement in spoken language annotation: Applying $_{u}\alpha$-family coefficients to discourse segmentation. *Russian Journal of Linguistics* 25 (2). 478–506. https://doi.org/10.22363/2687-0088-2021-25-2-478-506

Plungian, Vladimir, Ekaterina Rakhilina & Tatiana Reznikova. 2022. Perfective, performative and present: Some non-standard combinations in Slavic and beyond. *Russian Journal of Linguistics* 26 (4). 1012–1030. https://doi.org/10.22363/2687-0088-31252

Sharoff, Serge. 2022. What neural networks know about linguistic complexity. *Russian Journal of Linguistics* 26 (2). 371–390. https://doi.org/10.22363/2687-0088-30178

Solovyev, Valery, Marina Solnyshkina & Danielle McNamara. 2022. Computational linguistics and discourse complexology: Paradigms and research methods. *Russian Journal of Linguistics* 26 (2). 275–316. https://doi.org/10.22363/2687-0088-30161

Rastier, François. 2023. *Enjeux épistémologiques de la linguistique de corpus.* http://www.revue- texto.net/Inedits/Rastier/Rastier_Enjeux.html>. (accessed 12 July 2023).

Zatsman, Igor & Nadezhda Buntman. 2015. Outlining goals for discovering new knowledge and computerised tracing of emerging meanings. *Proceedings of the 16th European Conference on Knowledge Management. Reading: Academic Publishing International Limited*. 851–860.

Dictionaries

Большой толковый словарь русских глаголов: Идеографическое описание. Синонимы. Антонимы. Английские эквиваленты / под ред. проф. Л. Г. Бабенко. М., 2007. 576 с. (Сер. «Фундаментальные словари»). [Babenko, Lyudmila G. (eds.). 2007. *Big explanatory dictionary of Russian verbs: Ideographic description. Synonyms. Antonyms. English equivalents.* Moscow. (In Russ.)].

Article history:
Received: 31 August 2023
Accepted: 15 November 2023

Bionotes:

Nadezhda V. BUNTMAN is Doctor of Philology and Associate Professor of the Department of French at the Faculty of Foreign Languages and Regional Studies, Lomonosov Moscow State University. Her areas of research cover comparative and corpus linguistics, literary translation, stylistics of the French language and modern French literature. She is a translation award winner and a Cavalier of the French Order of Academic Palms.
e-mail: nabunt@hotmail.com
https://orcid.org/0009-0008-4945-1028

Anna S. BORISOVA is Doctor of Philology and Associate Professor of the Department of Foreign Languages, RUDN University. Her areas of research embrace translation studies, cognitive linguistics and discourse analysis.
e-mail: borisova-as@rudn.ru
https://orcid.org/0000-0002-7395-7028

Yuliya A. DAROVSKIKH is a PhD student and Lecturer of the Department of Foreign Languages at the Faculty of History, Lomonosov Moscow State University. Her research interests include comparative aspectology, corpus research and methods of teaching French.
e-mail: juliadarov@mail.ru
https://orcid.org/0009-0007-0606-1161

Сведения об авторах:

Надежда Валентиновна БУНТМАН – кандидат филологических наук, доцент кафедры французского языка факультета иностранных языков и регионоведения МГУ имени М.В. Ломоносова. Области ее исследований – сопоставительная и корпусная лингвистика, художественный перевод, стилистика французского языка, современная французская литература. Она является лауреатом переводческих премий и кавалером французского ордена «Академические пальмы».
e-mail: nabunt@hotmail.com
https://orcid.org/0009-0008-4945-1028

Анна Степановна БОРИСОВА – кандидат филологических наук, доцент кафедры иностранных языков филологического факультета РУДН. В сферу ее научных интересов входят теория и практика перевода, когнитивная лингвистика, дискурс-анализ.
e-mail: borisova-as@rudn.ru
https://orcid.org/0000-0002-7395-7028

Юлия Андреевна ДАРОВСКИХ – преподаватель кафедры иностранных языков исторического факультета МГУ имени М.В. Ломоносова, аспирант третьего года обучения. Ее научные интересы включают сопоставительную аспектологию, корпусные исследования и методику преподавания французского языка.
e-mail: juliadarov@mail.ru
https://orcid.org/0009-0007-0606-1161