Вестн. Моск. ун-та. Сер. 19. Лингвистика и межкультурная коммуникация. 2024. Т. 27. № 4. С. 147–158
Lomonosov Linguistics and Intercultural Communication Journal, 2024, vol. 27, no. 4, pp. 147–158

ЛИНГВИСТИКА И УРОВНИ ЯЗЫКА

В.А. Нуриев, С.Д. Игнатова

НАДКОРПУСНАЯ БАЗА ДАННЫХ КАК ИНСТРУМЕНТ ИЗУЧЕНИЯ ПУНКТУАЦИИ

Федеральный исследовательский центр «Информатика и управление» Российской Академии Наук, Москва, Россия; nurievv@mail.ru, ignatova_sophia@mail.ru

Аннотация: В статье рассматриваются возможности таких современных информационных ресурсов, как надкорпусные базы данных, для многоаспектного изучения пунктуации. С одной стороны, в разных естественных языках при общем совпадении репертуара знаков препинания и их письменного обозначения могут обнаруживаться зоны функционального расхождения, в следствие чего правила расстановки одного и того же знака будут разниться от языка к языку. Знание этих межъязыковых расхождений принципиально важно для человека-переводчика и для обучения систем машинного перевода, в противном случае перевод может существенно нарушить смысловое содержание исходного текста. Некоторые такие различия зафиксированы в докорпусную эпоху. Еще больше межъязыковых пунктуационных дифференциаций позволяют выявить надкорпусные базы данных — информационные инструменты, возникшие благодаря консолидированным усилиям информатики, компьютерной лингвистики и корпусного переводоведения: они помогают верифицировать уже имеющиеся знания на больших текстовых массивах и дополнять их. С другой стороны, пунктуация традиционно считается областью языка, достаточно хорошо изученной, жестко регламентированной и потому наименее подверженной изменениям и инновациям. Однако надкорпусные базы данных предоставляют возможность выявить новые (еще не закрепленные в нормирующей литературе) функционально-семантические особенности употребления отдельно взятых знаков препинания. Всестороннее изучение функционально-семантической нагрузки пунктуационных знаков приобретает сейчас особое значение в связи с развитием информационных технологий на базе искусственного интеллекта, а именно: голосовых ассистентов. В статье на примере восклицательного знака в русском и французском

Нуриев Виталий Александрович — доктор филологических наук, ведущий научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук; nurievv@mail.ru.

Игнатова София Дмитриевна — инженер Федерального исследовательского центра «Информатика и управление» Российской академии наук; ignatova_sophia@mail.ru.

языках показано, какие возможности надкорпусные базы данных открывают для изучения пунктуации.

Ключевые слова: корпусные информационные ресурсы; аннотирование; пунктуация; контрастивные исследования; перевод; межъязыковая асимметрия; корпусное переводоведение; база данных

doi: 10.55959/MSU-2074-1588-19-27-4-11

Финансирование: Работа выполнена в Федеральном исследовательском центре «Информатика и управление» Российской академии наук с использованием ЦКП «Информатика» ФИЦ ИУ РАН.

Для цитирования: Нуриева В.А., Игнатова С.Д. Надкорпусная база данных как инструмент изучения пунктуации // Вестн. Моск. ун-та. Сер. 19. Лингвистика и межкультурная коммуникация. 2024. Т. 27. № 4. С. 147–158.

Введение

Развитие современных технологий приводит к тому, что внимание разного рода специалистов привлекают области языкознания, долго находившиеся на периферии научного исследования. Так, в последнее время возрос интерес к изучению пунктуации. Во-первых, это связано с необходимостью улучшить работу систем машинного перевода и повысить качество производимых ими текстов. В разных естественных языках репертуар знаков препинания и правила их расстановки могут не совпадать. Знание этих межъязыковых расхождений принципиально важно для обучения нейросетей, используемых в архитектуре автоматизированных переводных систем: пунктуационный компонент играет существенную роль в организации письменного текста, его деформация при переводе может привести к серьезным смысловым искажениям. Во-вторых, это обусловлено развитием искусственного интеллекта, а именно: появлением голосовых ассистентов нового поколения. В технологический стек голосовых ассистентов входят два основных алгоритма: автоматическое распознавание речи и преобразование текста в речь. При автоматическом распознавании речи нерешенной задачей остается правильное пунктуирование, т.е. в сгенерированном тексте некоторые знаки препинания либо отсутствуют вовсе, либо расставлены неверно [Zhou, Tan, Qian, 2022; Nguyen et al., 2019; Nozaki et al., 2022]. Определяющую роль играет пунктуация на этапе преобразования текста в звучащую речь, где главной целью является привнесение разных модальностей в речь голосового ассистента. Модальность реконструируется на основе семантически насыщенных знаков препинания (точки, восклицательного и вопросительного знаков, многоточия). В настоящее время предпринима-

ются попытки создания голосовых трансформеров, способных выражать модальность высказывания: AudioPALM, SeamlessM4T, PromptTTS2, Next-GPT [Barrault et al., 2023; Rubenstein et al., 2023].

Изучение такого феномена языковой реальности, как пунктуационные знаки (в рамках одного естественного языка и в межъязыковом ракурсе), предполагает обработку представительного массива речевых образцов, чтобы обеспечить достоверные результаты. Необходимым исследовательским потенциалом в этом случае обладают современные информационные инструменты, объединяющие новейшие специализированные разработки информатики, компьютерной лингвистики и корпусного переводоведения. Примером подобных инструментов служат надкорпусные базы данных, разрабатываемые в Федеральном исследовательском центре «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН). Эти базы данных работают на основе параллельных русско-французских и французско-русских текстов общим объемом около 7,5 млн словоупотреблений, полученных из Национального корпуса русского языка (НКРЯ).

В статье показано, каким образом надкорпусные базы данных можно применять для изучения пунктуации и каких результатов при этом можно достичь. Первая ее часть посвящена обзору работ, которые вносят вклад в формирование методологии пунктуационного исследования. Особое внимание уделяется корпусному методу и межъязыковому сопоставлению. Во второй части представлен эксперимент и его результаты. Эмпирический материал составили речевые образцы — фрагменты параллельных текстов с восклицательным знаком на русском и французском языках в обоих переводных направлениях.

Корпусный метод и межъязыковое исследование пунктуации

Методология межъязыкового изучения пунктуации формируется, начиная с первой четверти XX в. Значительный вклад в ее разработку внесли отечественные специалисты, и в силу своего прямого выхода в практику ее развитие было тесно связано с переводческой деятельностью. См. хотя бы инструкции переводчикам К.И. Чуковского в брошюре «Принципы художественного перевода» 1919 г., подготовка которой началась практически сразу же после образования РСФСР [Чуковский, 1919: 23]. Одной из первых работ, где представлен обстоятельный анализ межъязыковых пунктуационных расхождений, стала статья литературного критика и переводчика М.П. Столярова, посвященная разбору недавно опубликованных переводов с французского языка. Основополагающим в ней считается принцип переводной точности, в том числе в отношении знаков

препинания, призванных облегчить восприятие синтаксической структуры сообщения на письме. Вне зависимости от положения конкретного знака (внутри предложений или на их границе) перед переводчиком ставится первостепенная задача выяснить роль пунктуирования в переводимом произведении и у переводимого автора вообще, чтобы потом добиться того же эффекта, но средствами своего родного языка. При этом важно принимать в расчет, что особенно медиальные, находящиеся внутри предложения, знаки препинания подпадают под жесткие регуляторные правила переводящего языка [Столяров, 1937: 252].

Несколько примечательных работ, где рассматриваются знаки препинания (в том числе в межъязыковом ракурсе), вышло в 1990-е годы [Malmkjær, 1997; May, 1994], в них ограниченно использован корпусный материал, но о целостной методологии речи еще не идет. В этом смысле прорывной становится статья, опубликованная в 2007 г. [Bystrova-McIntyre, 2007]. Ее цель — сравнить употребление знаков препинания в русских и английских передовицах (в «Известиях» и “New York Times”) за 2005 г. Объем корпуса составил 20 тыс. словоупотреблений для каждого языка, он сформирован вручную, затем обработан посредством встроенных поисковых запросов программы Word. По результатам запятая, двоеточие и тире имели бóльшую частотность в русскоязычных текстах. Эти данные верифицированы в корпусе художественных текстов, который так же компилировался вручную и обрабатывался в программе Word. В русских текстах частотность изучаемых знаков снова оказалась выше. Между тем и в русском газетном корпусе, и в английском их продуктивность была ниже. В заключительной части статьи автор на примере самого частотного знака (двоеточия) анализирует причины межъязыковой асимметрии и обозначает возможные стратегии его перевода с русского на английский [Bystrova-McIntyre, 2007: 137–162]. Несмотря на ряд очевидных недостатков — небольшой объем задействованных корпусов и их структура, в этой работе используется именно корпусный метод, позволяющий существенно расширить эмпирическую выборку и уточнить наблюдения, полученные ранее — на речевом материале гораздо меньшего объема.

Корпусная методология в исследованиях пунктуации значительное развитие получила в последние десять лет (см. [Nádvorníková, 2020; Wollin, 2018; Youdale, 2020: 121–150]). Одним из новейших корпусных информационных инструментов, зарекомендовавших себя при изучении пунктуации, являются базы данных, которые разрабатываются в ФИЦ ИУ РАН с 2013 г. (подробнее см. [Нуриев, Кружков, 2023; Нуриев, Карпов, 2023]). Используя тексты параллельных подкорпусов НКРЯ, они позволяют автоматизированным об-

разом получать большие текстовые данные и при этом обеспечивают возможность гибкого изменения объекта исследования. Обработка полученного материала в них производится посредством лингвистического аннотирования [Гончаров et al., 2019], которое предполагает отбор характерологических свойств-признаков изучаемого языкового явления и последующее описание этого явления с применением отобранных признаков. Признаки могут объединяться в одну или несколько классификаций. Далее в статье показано, как формируется экспериментальный массив текстовых фрагментов (на примере французских и русских параллельных текстов), на котором тестируются отобранные для аннотирования признаки.

Экспериментальные данные

Первый этап выявления классификационных признаков для лингвистического аннотирования в надкорпусных базах данных связан с анализом специализированной литературы, где рассматривается изучаемый объект, в данном случае — восклицательный знак.

Традиционно за пунктуационными знаками закрепляются три функции: просодическая (обозначение декламационно-интонационного контура, ритма и темпа речи), синтаксическая (делимитация на границе и внутри предложения) и семантическая (передача определенного значения) [Catach, 1996; Riegel et al., 2014]. Восклицательный знак эти три функции в себе совмещает.

Несмотря на очевидную просодическую функцию, этот знак не всегда указывает на повышение тона и интенсивности голоса: предложения, оканчивающиеся восклицательным знаком, могут произноситься без восходящей интонации и усиленного напряжения в голосе [Шапиро, 1974: 124–125]. Тем не менее такие предложения отличаются от повествовательных ритмом и тембром.

Отвечая за делимитацию, восклицательный знак может быть концевым (располагаться в конце предложения) и медиальным (находиться внутри предложения) [Валгина, 1979: 68–70]. Его частотность обнаруживает корреляцию со структурой оформляемого им предложения: он чаще употребляется в простых предложениях, чем в сложных, что объясняется большей востребованностью восклицательного знака в разговорной и диалогической речи, которая обычно представлена простыми и односложными конструкциями [Валгина, 1979: 68]. Это же определяет и другое синтаксическое свойство восклицательного знака — его продуктивность после междометий и восклицательных частиц (см., напр.: [Drillon, 1991: 351–355]).

Что касается третьей, семантической, функции, то здесь ни у отечественных, ни у французских специалистов единого мнения нет.

Так, Н. Каташ считает, что восклицательный знак сигнализирует о положительных и отрицательных эмоциях говорящего, однако конкретных примеров эмоций и чувств, выражаемых восклицательным знаком, она не называет [Catach, 1996: 63]. Ф. Дрийон в «Трактате о французской пунктуации» 1991 г. перечисляет 20 контекстов с восклицательным знаком [Drillon, 1991: 350–365]. Перечень носит хаотичный характер: в нем сводятся воедино эмоциональные состояния, обозначаемые восклицательным знаком (ирония[1], сомнение, страх, удивление), речевые акты, оформляемые этим знаком на письме (запрет, мольба, оскорбление, пожелание, приказ, проклятие, просьба, распоряжение, ругательство, совет, упрек), его синтактико-позиционные возможности (постановка после междометия, восклицательных частиц, обращения) и просодические особенности его употребления (крик, возглас одобрения). А.Б. Шапиро в монографии «Основы русской пунктуации» 1955 г. выделяет 12 эмоциональных состояний (возмущение, восхищение, досада, ирония, недовольство, обида, опасение, разочарование, сожаление, сомнение, удивление, удовольствие), на которые указывает восклицательный знак [Шапиро, 1974: 130–137]. В более поздних исследованиях [Валгина, 2000; Dugas, 2004] также предпринимаются попытки выделить значения восклицательного знака, однако работа в данном направлении не закончена. С одной стороны, это обусловлено тем, что выявление, разграничение и анализ психофизиологических состояний, передаваемых восклицательным знаком в письменной речи, сопряжены с определенными трудностями. С другой стороны, в докорпусную эпоху не было информационных инструментов, позволяющих исследователю автоматически обрабатывать большие массивы речевых данных. А их применение, как показано ниже, открывает принципиально иные возможности для обнаружения нового научного знания, в том числе о функционировании восклицательного знака.

Второй, не менее важный, этап отбора классификационных признаков для аннотирования — это обработка параллельных контекстов с восклицательным знаком (см. примеры 1–5), полученных с помощью поискового запроса в базе данных. Структура запроса на поиск в надкорпусной базе данных ФИЦ ИУ РАН имеет следующий вид: тильда + искомый знак препинания. Поиск может осуществляться одновременно и в тексте оригинала, и в переводе. На данном этапе исследования всего было обработано 3000 параллельных контекстов с восклицательным знаком: по 1500 в каждом переводном направлении (русский-французский и французский-русский).

[1] Ирония в специализированной литературе может считаться речевым актом [Падучева, 2010: 303] и причисляться к переживаемым человеком чувствам [Рубинштейн, 2009: 576].

Анализ корпусного материала не только подтвердил необходимость учета значений восклицательного знака, выявленных в докорпусную эпоху, но и позволил зафиксировать новые особенности функционирования изучаемого объекта в сопоставляемых языках (см. примеры 1–5). Полученные данные свидетельствуют о том, что восклицательный знак оформляет письменные высказывания, семантическое наполнение которых в полной мере не учитывается имеющимися исследованиями. Так, наряду с указанными в [Шапиро, 1974; Drillon, 1991] смысловыми оттенками этот знак препинания может сигнализировать о высокомерии, пренебрежении и раздражении, как в (1); передавать оценку с оттенком неодобрения, зависти и раздражения, как в (2); служить для выражения настойчивого намерения, как в (3). Еще восклицательный знак применяется для оформления тостов (4), благодарности (5) и некоторых других речевых актов, которые в этой связи ранее не рассматривались. Принципиально важно, что все обнаруженные контексты позволяют постановку восклицательного знака и в русском, и во французском, указывая на сближение функциональной нагрузки этого знака в двух языках, что ранее в научной литературе не освещалось.

(1)

Цехновицер старался уличить их в тупости, хамстве, цинизме, достигая, естественно, противоположных результатов. <…> Если с Цехновицером дружески заговаривали, он вскидывал брови: — Предпочитаю слушать тишину! С. Довлатов. Иностранка (1986).

Il essayait de les coincer pour mettre en évidence leur bêtise, leur goujaterie, leur cynisme et obtenait, naturellement, le résultat contraire. <...> Si quelqu'un lui parlait amicalement, il haussait les sourcils: "Je préfère écouter le silence!" Tr. J. Michaut-Paterno (2001).

(2)

Вдруг из отцовских сорока сделал тысяч триста капиталу, и в службе за надворного перевалился, и ученый… теперь вон еще путешествует! Пострел везде поспел! Разве настоящий-то хороший русский человек станет все это делать? Русский человек выберет что-нибудь одно, да и то еще не спеша, потихоньку да полегоньку, кое-как, а то на-ко, поди! И.А. Гончаров. Обломов (1848–1859).

Avec les quarante mille du père il a fait tout d'un coup un capital de trois cents mille, et dans sa carrière il est allé au-delà de conseiller, sans parler de son instruction. Maintenant en plus il voyage! Partout on ne voit que lui! Est-ce qu'un bon vrai Russe ferait tout cela? Un Russe choisirait une seule chose, et encore, il la ferait pas à pas, en douceur, tranquillement, tandis que là, regarde-moi ça! Tr. L. Jurgenson (1988).

(3)

J'ai maigri, je crois. Envie de choses douces et sucrées. Un sucre d'orge, tiens! Au prochain village je ferai toutes les boutiques s'il le faut et je m'en offrirai un! Jean-Claude Mourlevat. La rivière à l'envers 2: Hannah (2000).

Кажется, я похудела. Хочется сладкого. Карамельки — вот чего мне хочется! В следующей деревне хоть все лавки обойду, а куплю себе карамелек! Пер. Н. Шаховская (2020).

(4)

Чурилин неожиданно поднялся:

– Да здравствуют трудовые резервы! И достал из кармана вторую бутылку. С. Довлатов. Чемодан (1986).

Tchouriline se leva brusquement:

– Vive les réserves de main-d'œuvre! Et il sortit de sa poche une deuxième bouteille. Tr. J. Michaut-Paterno (2001).

(5)

Та, отчаянно улыбаясь, только вскрикивала:

– Ах, покорнейше вас благодарю! Мерси! Мерси! М. Булгаков. Мастер и Маргарита (ч. 2) (1929–1940).

Avec un sourire éperdu, Annouchka s'écria:

– Ah! je vous remercie mille fois! Merci! Merci! Tr. Claude Ligny (1968).

В целом полученные данные свидетельствуют о высокой степени семантической синкретичности восклицательного знака — его способности обыкновенно совмещать два и более значений. Это серьезно осложняет следующий, третий, этап, который представляет собой оформление выявленных признаков в классификационную схему — структурную основу будущих аннотаций. Важно учитывать, что перечень признаков в классификации не является окончательным: в нем предусмотрены возможные изменения (как правило, дополняющего характера) в соответствии с новыми обнаруженными фактами языковой реальности. Формирование такого набора признаков определяет дальнейшее направление данного исследования.

Заключение

В статье показано, какие возможности открывают современные надкорпусные базы данных для изучения пунктуации, позволяя автоматизированным способом обрабатывать большие текстовые данные и тем самым обеспечивая высокую достоверность получаемых выводов. Эмпирическим материалом послужили текстовые данные по употреблению восклицательного знака при переводе с русского на французский и с французского на русский. Полученные результаты носят промежуточный характер и указывают на необходимость дальнейшего изучения пунктуации с применением кор-

пусных инструментов. Однако уже на этом этапе анализ параллельных текстовых фрагментов с восклицательным знаком помог выявить ряд ранее не зафиксированных в специализированной литературе особенностей его употребления в обоих сопоставляемых языках. Следующим этапом данного исследования станет формирование набора признаков для лингвистического аннотирования восклицательного знака. Дальнейшие перспективы исследования будут направлены также на определение устойчивых межъязыковых дифференциаций в употреблении изучаемого знака препинания и соответственно на всестороннее описание его функционального потенциала в русском и французском языках.

СПИСОК ЛИТЕРАТУРЫ

1. *Валгина Н.С.* Русская пунктуация: принципы и назначение. М., 1979.
2. *Валгина Н.С., Светлышева В.Н.* Русский язык: орфография и пунктуация. Правила и упражнения. М., 2000.
3. *Гончаров А.А., Инькова О.Ю., Кружков М.Г.* Методология аннотирования в надкорпусных базах данных // Системы и средства информатики. 2019. Т. 29. Вып. 2. С. 148–160.
4. *Нуриев В.А., Карпов В.И.* Методология корпусно-ориентированного исследования в области контрастивной пунктуации // Информатика и ее применения. 2023. Т. 17. № 2. С. 90–95.
5. *Нуриев В.А., Кружков М.Г.* Корпусные данные при контрастивном изучении пунктуации. Системы и средства информатики. 2023. Т. 33. Вып. 1. С. 14–23.
6. *Падучева Е.В.* Семантические исследования: Семантика времени и вида в русском языке; Семантика нарратива. 2-е изд., испр. и доп. М., 2010.
7. *Рубинштейн С.Л.* Основы общей психологии. СПб., 2009.
8. *Столяров М.* Искусство перевода художественной прозы // Литературный критик. 1937. № 5–6. С. 242–254.
9. *Чуковский К.* Переводы прозаические // Принципы художественного перевода. СПб., 1919. С. 7–24.
10. *Шапиро А.Б.* Современный русский язык. Пунктуация. М., 1974.
11. *Barrault L. et al.* SeamlessM4T-Massively Multilingual & Multimodal Machine Translation // arxiv preprint arxiv:2308.11596. 2023. URL: https://arxiv.org/abs/2308.11596 (accessed: 05.03.2024).
12. *Bystrova-McIntyre T.* Looking at the Overlooked: A Corpora Study of Punctuation Use in Russian and English // TIS. 2007. Vol. 2. Iss. 1. P. 137–162.
13. *Catach N.* La ponctuation. Histoire et système. P., 1996.
14. *Drillon F.* Traité de la ponctuation française. P., 1991.
15. *Dugas A.* Guide de la ponctuation. Montreal, 2004.
16. *Malmkjær K.* Punctuation in Hans Christian Andersen's stories and in their translations into English // Nonverbal communication and translation: New perspectives and challenges in literature, interpretation and the media / Ed. by F. Poyatos. Amsterdam; Philadelphia, 1997. P. 151–162.
17. *May R.* The Translator in the Text: On Reading Russian Literature in English. Evanston, 1994.

18. *Nádvorníková O.* The use of English, Czech and French punctuation marks in reference, parallel and comparable web corpora: a question of methodology // Linguistica Pragensia. 2020. Vol. 30. Iss. 2. P. 30–50.
19. *Nguyen B. et al.* Fast and accurate capitalization and punctuation for automatic speech recognition using transformer and chunk merging // 2019 22nd conference of the oriental COCOSDA international committee for the co-ordination and standardisation of speech databases and assessment techniques (O-COCOSDA). Cebu, 2019. P. 1–5.
20. *Nozaki J. et al.* End-to-end Speech-to-Punctuated-Text Recognition // arXiv preprint arXiv:2207.03169. 2022. URL: https://arxiv.org/abs/2207.03169 (accessed: 02.03.2024).
21. *Riegel M., Pellat J.-Ch., Rioul R.* Grammaire méthodique du français. 5e éd. P., 2014.
22. *Rubenstein P.K. et al.* AudioPaLM: A Large Language Model That Can Speak and Listen // arXiv preprint arXiv:2306.12925. 2023. URL: https://arxiv.org/abs/2306.12925 (accessed: 05.03.2024).
23. *Wollin L.* Punctuation: Providing the Setting for Translation? // Studia Neophilologica. 2018. Vol. 90. № S1. P. 37–49.
24. *Youdale R.* Using computers in the translation of literary style: Challenges and opportunities. N.Y.; L., 2020.
25. *Zhou Z., Tan T., Qian Y.* Punctuation Prediction for Streaming On-Device Speech Recognition // ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapor, 2022. P. 7277–7281.

Vitaly A. Nuriev, Sofia D. Ignatova

SUPRACORPORA DATABASE AS A TOOL FOR STUDYING PUNCTUATION

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia; nurievv@mail.ru; ignatova_sophia@mail.ru

Abstract: This paper explores the potential of modern information resources such as supracorpora databases for the multidimensional study of punctuation. On the one hand, in different natural languages, while the repertoire of punctuation marks and their graphic representations tend to coincide, there may be zones of functional divergence, so that the rules of placement of the same punctuation mark will differ from one language to another. To know these interlingual discrepancies is fundamentally important for a human translator and for training machine translation systems; otherwise, the translation may significantly distort the semantic content of its source text. Some of these differences were recorded in the pre-corpus era. More of them can be revealed with the aid of supracorpora databases, modern information resources created through the joint efforts of computer science, computational linguistics, and corpus-based translation studies; they not only help to verify the existing knowledge on a wide scale of texts but also to amplify it. On the other hand, punctuation has traditionally been regarded as an area of language that is fairly well-studied, tightly regulated, and therefore least susceptible to change and innovation. However, supracorpora databases provide an opportunity to identify new (not yet found in the normative literature) functional-semantic features of the use of a given punctuation mark. Nowadays, the development of artificial intelligence-

based technologies, namely voice assistants, makes it particularly important to thoroughly research the functional semantics of punctuation marks. The paper shows the opportunities that supracorpora databases provide for punctuation studies, using the example of the exclamation mark in Russian and French.

Keywords: corpus resources; annotation; punctuation; contrastive studies; translation; asymmetry between languages; corpus-based translation studies; database

For citation: Nuriev V.A., Ignatova S.D. (2024) Supracorpora Database as a Tool for Studying Punctuation. *Lomonosov Linguistics and Intercultural Communication Journal*, vol. 27, no. 4, pp. 147–158. (In Russ.)

Funding: The work was performed at the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences using the CKP "Informatics" of FRC CSC RAS.

About the authors: Vitaly A. Nuriev — Dr. Habil in Philology, Lead Researcher, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; nurievv@mail.ru; *Sofia D. Ignatova* — Engineer, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; ignatova_sophia@mail.ru.

REFERENCES

1. Valgina N.S. 1979. *Russkaya punktuatsiya: printsipy i naznachenie* [Russian punctuation: principles and purpose]. Moscow, Prosveshchenie Publs. (In Russ.)
2. Valgina N.S., Svetlysheva V.N. 2000. *Russkii yazyk: orfografiya i punktuatsiya. Pravila i uprazhneniya* [Russian language: spelling and punctuation. Rules and exercises]. Moscow, Neolit Publs. (In Russ.)
3. Goncharov A.A., Inkova O.Yu., Kruzhkov M.G. 2019. Metodologiya annotirovaniya v nadkorpusnykh bazakh dannykh [Annotation methodology of supracorpora databases]. *Sistemy i Sredstva Informatiki*, vol. 29, no. 2, pp. 148–160. (In Russ.)
4. Nuriev V.A., Karpov V.I. 2023. Metodologiya korpusno-orientirovannogo issledovaniya v oblasti kontrastivnoi punktuatsii [The methodology of the corpus-based studies in the field of contrastive punctuation]. *Informatika i ee Primeneniya*, vol. 17, no. 2, pp. 90–95. (In Russ.)
5. Nuriev V.A., Kruzhkov M.G. 2023. Korpusnye dannye pri kontrastivnom izuchenii punktuatsii [The parallel corpora perspective on studying contrastive punctuation]. *Sistemy i Sredstva Informatiki*, vol. 33, no. 1, pp. 14–23. (In Russ.)
6. Paducheva E.V. 2010. *Semanticheskie issledovaniya: Semantika vremeni i vida v russkom yazyke; Semantika narrative* [Semantic studies: Semantics of time and aspect in Russian; Semantics of narrative]. Moscow, Yazyki slavyanskoi kul'tury Publs. (In Russ.)
7. Rubinshtein S.L. 2009. *Osnovy obshchei psikhologii* [Fundamentals of General Psychology]. Saint-Petersburg, Piter Publs. (In Russ.)
8. Stolyarov M. 1937. Iskusstvo perevoda khudozhestvennoi prozy [The art of prose translation]. *Literaturnyi kritik*, no. 5–6, pp. 242–254. (In Russ.)
9. Chukovskii K. 1919. Perevody prozaicheskie [Prosaic translations]. *Principles of literary translation*. Petersburg, Vsemirnaya literatura Publs, pp. 7–24. (In Russ.)
10. Shapiro A.B. 1974. *Sovremennyi russkii yazyk. Punktuatsiya* [The modern Russian language. Punctuation]. Moscow, Prosveshchenie Publs. (In Russ.)

11. Barrault L. et al. 2023. SeamlessM4T-Massively Multilingual & Multimodal Machine Translation. *arXiv preprint arXiv:2308.11596*. URL: https://arxiv.org/abs/2308.11596 (accessed: 05.03.2024).
12. Bystrova-McIntyre T. 2007. Looking at the Overlooked: A Corpora Study of Punctuation Use in Russian and English. *TIS*, vol. 2, no. 1, pp. 137–162.
13. Catach N. 1996. *La ponctuation. Histoire et système*. Paris, PUF.
14. Drillon F. 1991. *Traité de la ponctuation française*. Paris, Gallimard.
15. Dugas A. 2004. *Guide de la ponctuation*. Montréal, Éditions Logiques.
16. Malmkjær K. 1997. Punctuation in Hans Christian Andersen's stories and in their translations into English. *Nonverbal communication and translation: New perspectives and challenges in literature, interpretation and the media*. Ed. F. Poyatos. Amsterdam, Philadelphia, John Benjamins Publishing Company, pp. 151–162.
17. May R. 1994. *The Translator in the Text: On Reading Russian Literature in English*. Evanston, IL, Northwestern University Press.
18. Nádvorníková O. 2020. The use of English, Czech and French punctuation marks in reference, parallel and comparable web corpora: a question of methodology. *Linguistica Pragensia*. vol. 30, no. 2, pp. 30–50.
19. Nguyen B. et al. 2019. Fast and accurate capitalization and punctuation for automatic speech recognition using transformer and chunk merging. *2019 22nd conference of the oriental COCOSDA international committee for the co-ordination and standardisation of speech databases and assessment techniques (O-COCOSDA)*, IEEE, pp. 1–5.
20. Nozaki J. et al. 2022. End-to-end Speech-to-Punctuated-Text Recognition. *arXiv preprint arXiv:2207.03169*. URL: https://arxiv.org/abs/2207.03169 (accessed: 02.03.2024).
21. Riegel M., Pellat J.-Ch., Rioul R. 2014. *Grammaire méthodique du français*. 5e éd. Paris, PUF.
22. Rubenstein P.K. et al. 2023. AudioPaLM: A Large Language Model That Can Speak and Listen. *arXiv preprint arXiv:2306.1292*. URL: https://arxiv.org/abs/2306.12925 (accessed: 05.03.2024).
23. Wollin L. 2018. Punctuation: Providing the Setting for Translation? *Studia Neophilologica*, vol. 90, no. S1, pp. 37–49.
24. Youdale R. 2020. *Using computers in the translation of literary style: Challenges and opportunities.* London, UK; New York, NY, USA, Routledge.
25. Zhou Z., Tan T., Qian Y. 2022. Punctuation Prediction for Streaming On-Device Speech Recognition. *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 7277–7281.

Статья поступила в редакцию 29.03.2024;
одобрена после рецензирования 10.04.2024;
принята к публикации 28.06.2024;

The article was submitted 29.03.2024;
approved after reviewing 10.04.2024;
accepted for publication 28.06.2024.