***М. В. Копотев, О. В. Кисселев, А. А. Климов
Kopotev, Mikhail, Kisselev, Olesya, Klimov, Aleksandr***

**CAT&Kittens: корпус русских академических текстов и основанные на нём инструменты анализа студенческих работ**[[1]](#footnote-1)

**CAT&Kittens: An academic corpus and corpus-based tools for analysis of Russian academic writing**

**Аннотация**. Проект, направлен на создание Корпуса академических текстов на русском языке (CAT) и связанного с ним сервиса исправления ошибок в студенческих текстах (*kittens*). Платформа позволяет искать примеры в большом академическом корпусе и проверять собственный текст на наличие неакадемических вариантов употребления.

**Ключевые слова.** *корпус, академическое письмо, корпусной анализ*

**Abstract.** The project aims to create the Corpus of Academic Russian Texts (CAT) a corpus-based service that provides searches for non-standard language usage in novice writers’ texts (*kittens*). The platform allows for simple searching in a large academic corpus and provides sophisticated evaluation of a novice writer’s text against the large corpus data.

**Keywords.** *corpus, academic writing, corpus-based analysis*

**1. Введение**

Изучение академического языка в значительной степени обусловлено необходимостью научить студентов, осваивающих этот жанр, устоявшимся практикам и языковым шаблонам. В течение последних нескольких десятилетий эта область исследований находилась под влиянием двух взаимосвязанных подходов: компьютерного обучения языку (англ. Computer-Assisted Language Learning, CALL) и компьютерной лингвистики, включая подходы и инструменты корпусной лингвистики.

Методы корпусной лингвистики широко используются для создания учебных языковых ресурсов [Crossley и др., 2017]. Особенно заметный вклад они внесли в анализ английского академического дискурса [Ackermann, Chen, 2013; Biber, Conrad, Cortes, 2004; Durrant, Mathews-Aydınlı, 2011; Gray, Biber, 2013].

Несмотря на то, что компьютерные подходы бурно развивались в последнее десятилетие, практических приложений CALL для изучения русского языка немного. Мы можем назвать проекты Revita [Katinskaia, Nouri, Yangarber, 2018], Текстометр [Laposhina и др., 2018], Visualizing Russian [Clancy и др., 2019], Русский конструктикон [Endresen и др., 2020], RuSkell [Apresjan и др., 2016], CoCoCo [Копотев, 2020].

Важно отметить, что существуют корпусные ресурсы, содержащие академические тексты. Например, НКРЯ включает тексты академического жанра (около 27,4 млн токенов). Однако подкорпус не позволяет выделять и сортировать различные поджанры и не сбалансирован с точки зрения представленности академических дисциплин. Наконец, НКРЯ невозможно выгрузить для более сложной языковой обработки. Еще один проект, который дает возможность анализировать академический стиль, — это подкорпус RU-AC в составе проекта IntelliText, возглавляемого C. Шаровым, [corpus.leeds.ac.uk/itweb](http://corpus.leeds.ac.uk/itweb). RU-AC относительно невелик (пять миллионов токенов), состоит из студенческих работ разного качества и не может считаться репрезентативным.

 Проект CAT&kittens заполняет эту лакуну в двух направлениях. Мы надеемся, во-первых, внести свой скромный вклад в изучение современного русского академического языка и выявить, хотя бы частично, типичные ошибки, что позволит использовать пропедевтические подходы для их предупреждения у следующих поколений студентов. Во-вторых, мы планируем создать платформу, которая предоставит пользователю возможность анализировать и редактировать собственный текст на основе сравнения с большим корпусом.

**2. Структура корпуса**

СAT задуман как двухчастный корпус академического русского языка, охватывающий несколько научных областей. Эталонный корпус, собственно CAT, включает 3600 научных статей, взятых из авторитетных рецензируемых российских академических журналов; статьи подкорпуса представляют шесть академических областей: экономика, образование и психология, юридические тексты, лингвистика, история и социология. Все статьи опубликованы в период с 2010 по 2018 год. Размер корпуса CAT — около 14 миллионов токенов, что достаточно для реализации множества задач, однако не всех. Для решения проблем, связанных прежде всего с нейросетевым моделированием, этого объема оказалось недостаточно, поэтому был собран дополнительный корпус CyberCAT, основанный на научной онлайн-библиотеке «Киберленинка», cyberleninka.org. Объем этого корпуса — около 155 миллионов токенов. Принципы создания и аннотирования обоих корпусов совпадают, однако в обучающей платформе будет использоваться только эталонный корпус CAT.

*Таблица 1.* Распределение текстов в CAT и CyberCAT

| КорпусДисциплина  | Эталонный корпус CAT | Корпус для машинного обучения CyberCAT |
| --- | --- | --- |
| Экономика | 2 494 422 | 16 548 007 |
| Педагогика и психология | 1 880 004 | 32773116 |
| Юриспруденция | 2 636 290 | 17361237 |
| Лингвистика | 2 691 363 | 11985639 |
| История | 2 808 313 | 55516958 |
| Социология | 1 500 196 | 21127403 |
|  **Всего** | **14 010 588** | **155 312 360** |

**3. Обработка текстов**

Все тексты в кодировке UTF-8 аннотированы с помощью анализатора UDpipe [Straka, Straková, 2017], который объединяет инструменты токенизации, лемматизации, морфологической разметки и синтаксического анализа в терминах грамматики зависимостей в формате CoNLL-U. Использование этого инструмента позволяет получать не только базовую информацию, например, о частоте токенов или лемм, но и более детальную информацию о грамматических характеристиках токенов и синтаксических зависимостях между ними.

Перед применением UDpipe все тексты были предварительно обработаны: ссылки, таблицы, рисунки, имена авторов, а также названия университетов или журналов и номера страниц в колонтитулах были убраны. Тексты были разделены на отдельные предложения; знаки препинания, обозначающие конец предложения, заменены на точки; остальные знаки препинания (например, запятые, точки с запятой и т. д.) были удалены. Все числа в текстах были заменены токеном NUM. В корпус были включены только основные разделы статей (введение, обзор литературы, анализ и результаты, обсуждение и заключение и т.п.); все таблицы, рисунки, уравнения и цитаты в тексте, а также библиография были автоматически удалены с помощью скриптов Python, доступных по адресу: github.com/kopotev/CATandkittens.

*Рис 1.* Пример разметки в формате UDpipe.

**4. Инструменты проверки текста**

На основе созданных корпусов создаётся онлайн-сервис CAT&kittens — многофункциональная платформа с различными инструментами поиска и анализа. Помимо стандартных функций поиска по корпусу платформа предоставляет возможности для оценки качества текста.

**Первый уровень проверки** — это общая оценка текста. С помощью визуализации различных параметров текста система дает автору представление об уровне сложности текста (*readability*), для которой используются тесты *Flesch Reading Ease* и *Flesch-Kincaid Grade*, адаптированные для русских академических текстов [Solovyev, Ivanov, Solnyshkina, 2018] и согласованные с общеевропейской шкалой языковой компетенции CEFR [Little, 2007]. Кроме этого, сервис рассчитывает среднюю длину предложений, которые сравниваются с соответствующими значениями в CAT с целью продемонстрировать, насколько анализируемый текст соотносится с общим уровнем сложности, принятым в академическом дискурсе. Слишком длинные или слишком короткие предложения выделяются в тексте для дальнейшей правки. На этом этапе мы проверяем также возможные повторы фрагментов в тексте.

**Второй уровень проверки** касается лексической вариативности и коллокаций. Эта часть включает прежде всего показатель sTTR (соотношение тип/токен), дающий общее представление о лексическом богатстве текста. На более детальном уровне система анализирует низкочастотные слова, в том числе *hapax legomen*, проверяет чрезмерное или недостаточное использование лексем из выбранной предметной области. Помимо простого определения отклонений от академического стандарта, система предлагает альтернативные варианты, подобранные на основе семантической близости слов, полученных на предобученной на данных CyberCAT модели.

Семантическая близость используется и в алгоритме поиска нестандартных коллокаций, который на первом этапе выявляет в тексте n-граммы, не найденные в большом корпусе, затем подбирает семантически близкие замены и на последнем этапе проверяет полученные коллокации на наличие в CAT, то есть в эталонном академическом дискурсе.

**Третий уровень** — проверка грамматики. В отличие от доступных средств проверки, например, Орфо, проверка в нашей системе ориентирована на обнаружение отклонений от стандартов академического письма. Эта часть алгоритма сравнивает пару «токен–тегсет» в тексте с аналогичными парами в корпусе и указывает на возможные отклонения от стандартного употребления морфологических форм. Таковым оказываются, например, императивы, нехарактерные для академических текстов, или неграмматичные формы, которым не находится соответствия в корпусе CAT. В этой же части анализируются более сложные морфосинтаксические отклонения от языкового стандарта, например, генитивные цепочки (набор вложенных в друг друга генитивных групп) или пропуск прономинального субъекта (*prodrop*), характерный скорее для разговорных регистров.

**5. Заключение**

В нашем докладе мы кратко описали создаваемую платформу CAT&kittens, основу которой составляет большой корпус академических текстов. На данном этапе платформа еще не завершена, однако уроки, извлеченные из ее разработки, имеют некоторое значение для методологии использования корпусных инструментов анализа. Один из уроков — это сочетание автоматической проверки и человеческой компетенции при создании инструментов анализа. Цель создания подобной системы состоит не в замене квалифицированного эксперта, не в полной автоматизации процесса, а в том, чтобы облегчить поиски подозрительных фрагментов и предложить автору возможные варианты. Окончательный выбор можно сделать либо обратившись к эксперту, либо, в соответствии с корпусной парадигмой, с помощью поиска в CAT — глубоко аннотированном репрезентативном корпусе русских академических текстов, который предоставляет как учащимся, так и исследователям, изучающим академические жанры, необходимые инструменты анализа.

**6. Литература**

1. Копотев М. В. О самом сложном: изучение сочетаемости слов онлайн // Русский язык за рубежом. 2020. № 6. С. 36–43.

2. Ackermann K., Chen Y.-H. Developing the Academic Collocation List (ACL)–A corpus-driven and expert-judged approach // Journal of English for Academic Purposes. 2013. Т. 12. № 4. С. 235–247.

3. Apresjan V. и др. RuSkELL: Online Language Learning Tool for Russian // Proceedings of the XVII EURALEX International Congress. Lexicography and Linguistic Diversity (6 – 10 September, 2016). Tbilisi, 2016.

4. Biber D., Conrad S., Cortes V. If you look at…: Lexical bundles in university teaching and textbooks // Applied linguistics. 2004. Т. 25. № 3. С. 371–405.

5. Clancy S. и др. Foundations of Russian: A cognitive and constructional approach to teaching Russian enriched by frequency data // SCLC-2019. Book of Abstracts. Cambridge, MA, 2019. С. 15.

6. Crossley S. и др. Applying natural language processing tools to a student academic writing corpus: How large are disciplinary differences across science and engineering fields? // Journal of writing analytics. 2017. Т. 1.

7. Durrant P., Mathews-Aydinli J. A function-first approach to identifying formulaic language in academic writing // English for Specific Purposes. 2011. Т. 30. № 1. С. 58–72.

8. Endresen A. и др. Русский конструктикон: новый лингвистический ресурс, его устройство и специфика // Компьютерная лингвистика и интеллектуальные технологии. Москва: РГГУ, 2020.

9. Gray B., Biber D. Lexical frames in academic prose and conversation // International Journal of Corpus Linguistics. 2013. Т. 18. № 1. С. 109–136.

10. Katinskaia A., Nouri J., Yangarber R. Revita: a language-learning platform at the intersection of its and call // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). : European Language Resources Association (ELRA), 2018.

11. Klimov A., Kisselev O., Kopotev M. Towards intelligent correction of collocational errors in Russian novice academic texts in the CAT&kittens writing support platform // Russian Language Journal. 2021. Т. 71. № 1. С. (в печати).

12. Laposhina N. и др. Automated text readability assessment for Russian second language learners // Компьютерная лингвистика и интеллектуальные технологии. Москва: РГГУ, 2018. С. 403–413.

13. Little D. The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy // The Modern Language Journal. 2007. Т. 91. № 4. С. 645–655.

14. Solovyev V., Ivanov V., Solnyshkina M. Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics // Journal of intelligent & fuzzy systems. 2018. Т. 34. № 5. С. 3049–3058.

15. Straka M., Straková J. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe // Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. 2017. С. 88–99.

Михаил Вячеславович Копотев

НИУ ВШЭ (Санкт-Петербург) / Хельсинкский университет

mkopotev@hse.ru

Олеся Викторовна Кисcелев

Университет Техаса в Сан-Антонио

Александр Антонович Климов / Хельсинкский университет

НИУ ВШЭ (Санкт-Петербург)

Kopotev, Mikhail

HSE University (Saint Petersburg) / University of Helsinki

mkopotev@hse.ru

Kisselev, Olesya

The University of Texas at San Antonio

Klimov, Aleksandr

HSE University (Saint Petersburg) / University of Helsinki

1. Статья частично основана на нашей публикации [Klimov, Kisselev, Kopotev, 2021]. Мы благодарим студентов НИУ ВШЭ А. Грилланди, М. Килину, Е. Носову, А. Сидорову и студента Портлендского университета Ивза (Наму) Винеке, без которых эта статья никогда бы не появилась. [↑](#footnote-ref-1)