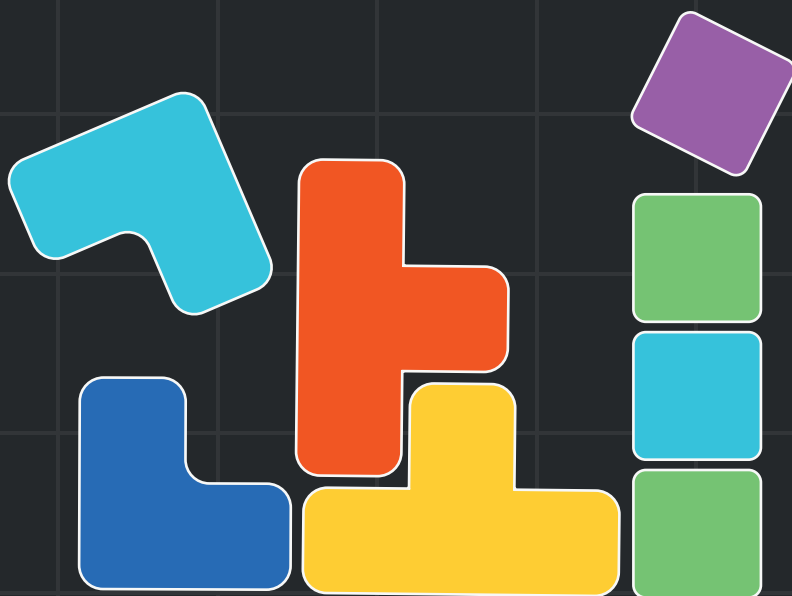


IDO / 2025  
CONF

# ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА (NLP) ДЛЯ ЛИНГВИСТИКИ, IT И БИЗНЕСА

Сборник материалов международной  
научно-практической конференции  
«Обработка естественного языка (NLP)  
для лингвистики, IT и бизнеса»



28 февраля — 1 марта 2025  
Томск

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ  
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

## **ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА ДЛЯ ЛИНГВИСТИКИ, IT И БИЗНЕСА**

**Сборник материалов  
Международной научно-практической конференции  
(28 февраля – 1 марта 2025 г.)**

Томск  
Издательство Томского государственного университета  
2025

**УДК 004.8**  
**ББК 32.813**  
**О13**

**Редакционная коллегия:**

З.И. Резанова (отв. ред.); А.В. Бочаров; К.П. Крапивина

**О13**      Обработка естественного языка для лингвистики, IT, бизнеса :  
сборник материалов Международного научно-практического  
семинара (30 февраля – 1 марта 2025 г.) / отв. ред.  
З.И. Резанова. – Томск : Издательство Томского  
государственного университета, 2025. – 58 с.

ISBN 978-5-907890-87-9

В сборнике представлены результаты научных исследований и прикладных проектов, реализуемых с использованием методов и технологий автоматической обработки естественного языка (NLP). Основные направления научных и прикладных разработок: создание и использование корпусов текстов, баз данных, чат-ботов; автоматизированный анализ текстов с применением когнитивных технологий; использование больших языковых моделей для генерации и обработки текстов.

Для исследователей и практиков, заинтересованных в разработке и применении технологий NLP и AI для решения широкого круга социально значимых проблем.

УДК 004.8  
ББК 32.813

**Рецензенты:**

*А.В. Колмогорова*, доктор филологических наук, профессор;  
*К.С. Шиляев*, кандидат филологических наук, доцент

ISBN 978-5-907890-87-9      © Томский государственный университет, 2025

– на третьем этапе осуществляется морфологический анализ вербальных единиц с помощью библиотеки Stanza, его ручная корректировка и запись этих данных в созданный на предыдущем этапе файл JSON;

– на четвертом этапе предполагается ручная разметка тематики нескольких текстов экспертами, полученный набор данных выступит материалом для обучения модели RoBERTa и дальнейшей автоматической разметки тематики текстов.

Результатом работы станет общедоступный корпус звучащей речи, состоящий из двух подкорпусов («Устная история Иркутска» и «Русские говоры Прибайкалья»). В корпусе предполагается возможность организации лексико-грамматического поиска и поиска по просодии, обеспечивающего возможность лингвистических исследований устной речи, и тематического поиска, который может представлять интерес для специалистов других областей гуманитарного знания. Созданием сайта и разработкой системы поиска занимаются программисты ИМИТ ИГУ, на сегодняшний день сайт существует в альфа-версии и доступен только для участников проекта.

## **АНАЛИТИКА УЧЕБНЫХ ТЕКСТОВ ДЛЯ ИЗУЧЕНИЯ РУССКОГО ЯЗЫКА КАК ИНОСТРАННОГО: ФУНКЦИОНАЛ ПЛАТФОРМЫ RULINGVA\***

**М.И. Андреева<sup>1</sup>, М.И. Солнышкина<sup>1</sup>, А.В. Данилов<sup>1</sup>**

*<sup>1</sup> Казанский (Приволжский) федеральный университет, Казань, Россия*

Профилирование и классификация текстов, контрастирование их лингвистических параметров с целью выявления референсных диапазонов для последующего структурирования больших

---

\* Исследование выполнено за счет гранта Российского научного фонда № 24-78-10129, <http://rscf.ru/project/24-78-10129/>.

языковых данных рассматривается в настоящее время как весьма актуальная задача в текстовой аналитике в частности и прикладной лингвистике в целом. Выполнение такого рода задач возможно только при помощи мощных текстовых анализаторов и профайлеров, широко используемых редакторами, разработчиками тестовых и учебных материалов, исследователями-лингвистами.

Представленное исследование осуществлено с целью выявления относительной сложности двух текстов уровня А2, используемых в практике преподавания русского языка как иностранного (далее РКИ), и выполнено с использованием текстового анализатора RuLingva ([rulingva.kpfu.ru/lexdiv](http://rulingva.kpfu.ru/lexdiv)). RuLingva рассчитывает метрики 120 параметров текстов на русском языке, включая уровень сложности (А1–С2) каждой лексемы и может выступать в качестве инструмента, способного прогнозировать трудность понимания текста для респондентов разных категорий. Материалом исследования выступили адаптированные для изучения русского языка как иностранного (далее РКИ) тексты «Злой мальчик» и «Месть» (по А.П. Чехову), объем которых составляет 341 и 576 словоформ соответственно. Тексты маркированы А2\_1 и А2\_2. В качестве источника использован Учебный корпус русского языка (Свидетельство о регистрации базы данных RU 2020622254).

Отдельную проблему в преподавании и тестировании РКИ представляет ранжирование текстов по уровням. В изучаемых текстах А2 доля лексики уровня А2 составляет только 3,23% (А2\_1) и 5,38% (А2\_2), а лексика уровня В1 – 8,50% и 11,98%. Около 6% лексики в каждом из текстов относятся к уровню В2. Параметрический анализ текстов также выявил 20 (7,33%) и 34 (5,90%) лексемы, отсутствующие в лексических минимумах РКИ. Большую часть данных слов составляют именованные сущности, а именно, имена, отчества и фамилии персонажей. Например, *Анна, Семёновна, Иванович, Лапкин, Саввич, Турманов, Дегтярёв и др.*

RuLingva представляет пользователям возможность осуществлять синонимические замены в тексте, если текст содержит лексические единицы более высоких уровней. Например, слово *отличный*, очевидно, не знакомое студентам уровня A2, предлагается заменить на *хороший, идеальный, прекрасный*, присутствующие в списках A2. Приведем примеры синонимов, использованных при адаптации оригинального текста «Месь» (Чехов А.П.) для уровня A2 РКИ. Определение уровня сложности лексических единиц автоматизировано на платформе RuLingva и указывается в скобках рядом с каждым словом. Пример, «Пожалуй (C2), пожалуй (C2) ...» (Оригинал) – «Конечно (A1), конечно (A1)» (A2\_РКИ); «густой (B2) сочный (B2) бас (B2)» (Оригинал) – «мужской (A1) голос (B1)» (A2\_РКИ).

#### Сложность лексики текстов

| Текст | Текст РКИ до RuLingva                                                                                                                                                                                                                                                        | Текст РКИ после замены синонимов на RuLingva                                                                                                                                                                                                                             |
|-------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| A2_1  | Иван (A1) Иванович (UNL) Лапкин (UNL), <u>приятный</u> (B1) молодой (A1) человек (A1), и (A1) Анна (A1) Семёновна (UNL), молодая (A1) девушка (A1), пришли (A1) на (A1) берег (A2) реки (A1), <u>спустились</u> (B2) к (A2) воде (A1) и (A1) сели (A2) на (A1) скамейку (B2) | Иван (A1) Иванович (UNL) Лапкин (UNL), <u>хороший</u> (A1) молодой (A1) человек (A1), и (A1) Анна (A1) Семёновна (UNL), молодая (A1) девушка (A1), пришли (A1) на (A1) берег (A2) реки (A1), <u>подошли</u> (B1) к (A2) воде (A1) и (A1) сели (A2) на (A1) скамейку (B2) |
| A2_2  | Но (A1) ему (A1) было (A1) очень (A1) <u>неприятно</u> (B2) и (A1) <u>обидно</u> (B2), что (A1) его (A1) <u>приятель</u> (B2) называет (A1) его (A1) старым (A1) толстым (B1) <u>дураком</u> (C1)                                                                            | Но (A1) ему (A1) было (A1) очень (A1) <u>плохо</u> (A1) и (A1) <u>грустно</u> (A2), что (A1) его (A1) <u>друг</u> (A1) называет (A1) его (A1) старым (A1) толстым (B1) <u>глупым</u> (A2) человеком (A1)                                                                 |

В качестве дополнительных функций RuLingva предлагает разбивку на предложения, их нумерацию, визуализацию различных метрик, частотность отдельных слов и др.

## **ОТ ПОЛЕВЫХ ДАННЫХ К АВТОМАТИЧЕСКОМУ АНАЛИЗУ: РАЗРАБОТКА МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА ДЛЯ НАУКАНСКОГО ЯЗЫКА\***

**Е.М. Будянская<sup>1</sup>, А.О. Бузанов<sup>1,2</sup>,  
Д.О. Жорник<sup>1</sup>, А.А. Пихтин<sup>1,2</sup>**

*<sup>1</sup> Институт языкознания РАН, Москва, Россия*

*<sup>2</sup> Высшая школа экономики, Москва, Россия*

В сфере компьютерной обработки естественного языка принято разделять языки на много- и малоресурсные. Многоресурсные языки, такие как английский, испанский или русский, характеризуются наличием обширных корпусов текстов, размеченных датасетов и развитых инструментов обработки. К малоресурсным относятся языки с ограниченным количеством цифровых ресурсов – такие как албанский, грузинский, исландский или хинди. Особую категорию составляют экстремально малоресурсные языки – преимущественно языки без государственного статуса, для которых электронные ресурсы практически отсутствуют. В эту категорию попадает большинство языков России.

В конце XX в. индустрия естественной обработки языка совершила прорыв, перейдя от правилых методов к статистиче-

---

\* Работа ведется в рамках проекта по документации науканского языка, поддержанного грантом РНФ № 23-18-00204 «Динамика языковой ситуации в сообществах Крайнего Севера: диахроническая документация науканского языка» (рук. В.А. Плунгян).