Министерство просвещения Российской Федерации
ФГБОУ ВО «Ярославский государственный педагогический
университет им. К. Д. Ушинского

Бойчук Е.И., Лагутина Н.С., Лагутина К.В., Воронцова И.А.,
Шляхтина Е.В., Мишенькина Е.В., Беляева О.В.

АВТОМАТИЗИРОВАННЫЙ АНАЛИЗ ТЕКСТОВЫХ ХАРАКТЕРИСТИК

Учебное пособие

Ярославль, 2020

УДК 004
ББК 32
А22

Исследование выполнено при финансовой поддержке РФФИ
в рамках научного проекта №19–07–00243

Рецензенты:

В.А. Соколов – доктор физико-математических наук, профессор, заведующий кафедрой теоретической информатики ФГБОУ ВО «Ярославский государственный университет им. П. Г. Демидова».

Т.Н. Хомутова – доктор филологических наук, профессор, заведующий кафедрой лингвистики и перевода института лингвистики и международных коммуникаций ФГАОУ ВО «Южноуральский государственный университет (НИУ)».

Бойчук Е.И., Лагутина Н.С., Лагутина К.В., Воронцова И.А., Шляхтина Е.В., Мишенькина Е.В., Беляева О.В.

А22 **Автоматизированный анализ текстовых характеристик** : учебное пособие / Е.И. Бойчук, Н.С. Лагутина, К.В. Лагутина, И.А. Воронцова, Е.В. Шляхтина, Е.В. Мишенькина, О.В. Беляева. – Ярославль, 2020. – 88 с.

ISBN

Данное учебное пособие ориентировано на студентов филологических факультетов и факультетов информатики и вычислительной техники, увлекающихся автоматизированной обработкой текстов различных жанров. В пособии представлены некоторые формы работы с текстом с опорой на приложения и программы, облегчающие исследовательскую работу с материалом, его количественной и качественной обработкой. Представленные приложения и методики ориентированы на работу с русским языком, а также с различными иностранными языками, в частности английским, французским, испанским. Среди прочих представлено приложение, позволяющее проводить ритмический анализ текстов в четырех иностранных языках (ProseRhythmDetector). Данное приложение основано на понимании ритма как периодической повторяемости определенных языковых элементов в тексте.

Структура пособия включает пять глав. В первой раскрываются возможности компьютерной обработки текстов разных типов и жанров в родном и иностранном языках. Во второй главе представлен обзор различных корпусов текстов, использование которых необходимо в процессе обучения и при проведении научных исследований. В третьей главе раскрываются возможности инструмента PRD для анализа ритма прозы. В четвертой представлен комплексный анализ ритма текста с опорой на инструмент, а также представлены возможности проведения сравнительного анализа оригинала и перевода в русском и английском языках. В пятой главе описывается механизм проведения статистических экспериментов с использованием инструмента PRD.

Каждый раздел сопровождается заданиями и вопросами по темам.

УДК 004
ББК 32

ISBN

СОДЕРЖАНИЕ

Глава I.
ИНСТРУМЕНТЫ И ПРИЛОЖЕНИЯ ДЛЯ РАБОТЫ С РАЗЛИЧНЫМИ ТИПАМИ ТЕКСТОВ

Обработка естественного языка (*Natural Language Processing, NLP*) – подраздел информатики, посвященный компьютерному анализу естественных человеческих языков. Он изучает решение следующих задач:

- организация взаимодействия между компьютерами и человеком на естественном языке;
- разработка средств анализа больших объемов данных, представленных на естественном языке;
- распознавание речи и понимание естественного языка компьютерными системами;
- генерация естественного языка.

Считается, что начало развития данного направления было положено в 1950-х годах на основе идей и работ А. Тьюринга [Turing, 2009: 23-65]. На ранних этапах развития систем обработки естественного языка они строились, главным образом, на основе экспертных систем с запрограммированными экспертными правилами. С 1980-х годов началось активное использование машинного обучения и статистических методов [Bird, 2009].

С помощью машинного обучения автоматически происходит генерация правил на основе анализа больших массивов реальных текстов. При этом используются такие методы, как нейронные сети и деревья решений, а также кластеризация и др. [Большакова, 2011].

Современные инструменты работы с текстами автоматизируют достаточно широкий спектр задач компьютерной лингвистики, в частности анализ отдельных слов, фраз и их контекста, а также текста в целом. Тем не менее, каждый инструмент сам по себе имеет довольно узкую специализацию и ориентирован на конкретные задачи.

В данной главе мы представим лишь некоторые инструменты, которые будут интересны лингвистам с точки зрения выполнения различных операций с текстом. Это онлайн ресурсы для семантического анализа текстов, одна из наиболее ярких и современных разработок платформа #LancsBox и рифмовники различных типов.

§1 Онлайн ресурсы для семантического анализа текста

Семантический анализ позволяет определить самые важные ключевые слова, фразы, что помогает грамотно сформировать семантическое ядро. Данные качественного семантического анализа могут использоваться в торговле для анализа спроса на товары по полученным отзывам, в поисковиках, системах автоматического перевода и пр.

Семантический анализ текста оценивает количество слов или фраз, которые определяют смысл текста, то есть его семантическое ядро, и статистические показатели. Правильно сформированное семантическое ядро способно быстро продвигать статью в поисковой системе. Комбинируя слова, составляя грамотно фразы, можно создать текст, который будет эффективно воздействовать на читателя, побуждая его к тем действиям, в которых заинтересованы владельцы сайта.

Поисковые системы также выполняют семантический анализ, определяя смысл текста, впоследствии чего в ответ на запрос предлагают выбранные материалы.

К статистическим показателям относятся: количество символов с пробелами и без, количество слов, в том числе уникальных и значимых, стоп-слов, количество воды, грамматических ошибок, процент классической и академической тошноты, семантическое ядро. При подсчете учитывается число уникальных слов (без повторений), число значимых слов (существительных), стоп-слов (которые лишены своего смысла). Процент воды определяется путем деления числа значимых слов на общее количество. Количество воды нельзя считать показателем качества текста, но все же лучше, чтобы этот показатель не превышал 65%. Если в тексте обнаружено 75% воды и больше, стоит уменьшить число незначимых слов.

Классическая тошнота определяет, сколько раз повторяется в тексте одно и то же слово. Оптимальное значение классической тошноты – 7. Повышение данного показателя приводит к торможению продвижения сайта. Коэффициент академической тошноты указывает на повторение большого количества слов в тексте. Соответственно, увеличение плотности ключевых слов приводит к его повышению.

В настоящее время существует достаточно большое количество он-лайн ресурсов для семантического анализа текста на русском и на английском языка, например, www.istio.com, www.us-ingenglish.com ,www.seoscout.com, www.online-utility.org, www.roadtogrammar.com. Это бесплатные онлайн сервисы анализа текстов.

Одним из самых популярных инструментов является Анализатор текста https://www.texta-nalyzer.ru/ (Рис. 1).

Для проверки текста нужно ввести его в представленное поле и нажать кнопку «Анализировать». В результате, в таблицах дается подробная статистика текста (Рис.2). Здесь можно узнать общее количество символов, в том числе без пробелов; общее количество слов, уникальных слов, стоп-слов, среднюю длину слова и водность текста, а также общее количество предложений, минимальную, среднюю и максимальную длину предложения; часто встречаемые последовательности слов и частотный словарь.

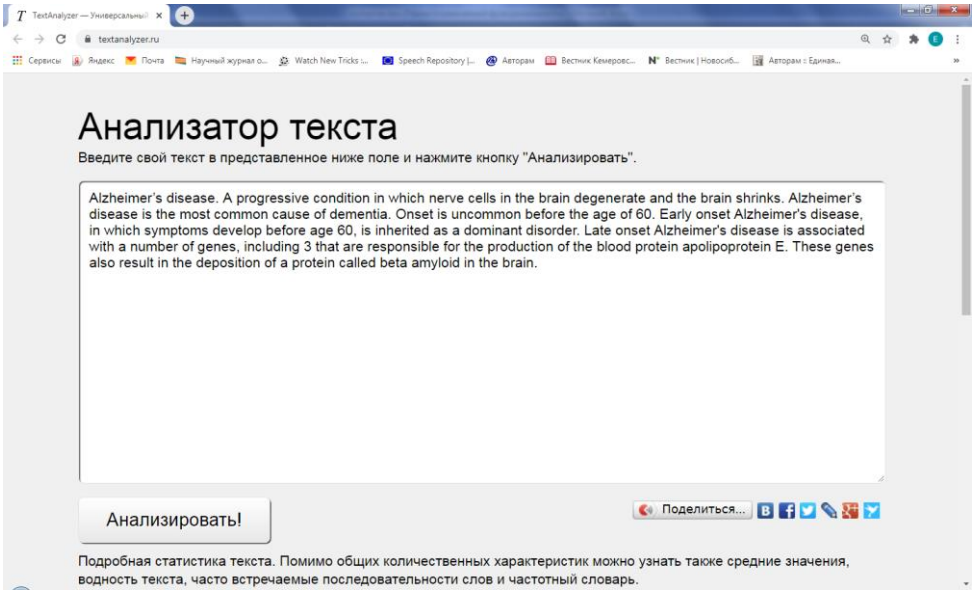


Рис. 1. Анализатор текста

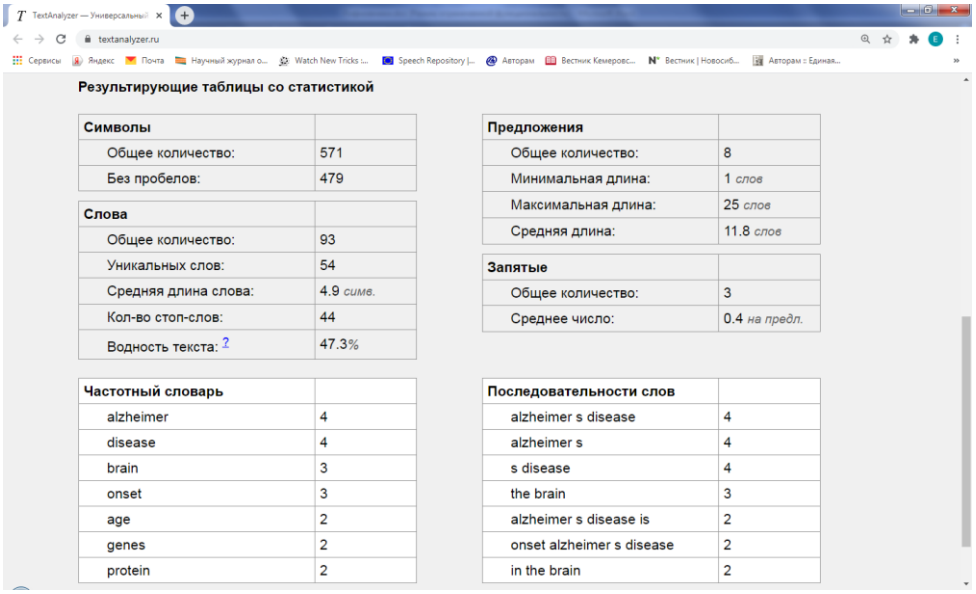


Рис. 2. Статистика по тексту

Другой популярный инструмент для проведения семантического анализа можно найти на сайте https://advego.com/text/seo/.

Анализ текста Адвего определяет:

- плотность ключевых слов, процент ключевых фраз;
- частотность слов;
- количество стоп-слов;
- объем текста: количество символов с пробелами и без пробелов;
- количество слов: уникальных, значимых, всего;
- водность, процент воды;
- тошноту текста, классическую и академическую;
- количество грамматических ошибок.

Данный онлайн сервис показывает семантическое ядро текста страницы — все значимые и ключевые слова, стоп-слова и грамматические ошибки в тексте (Рис. 3).

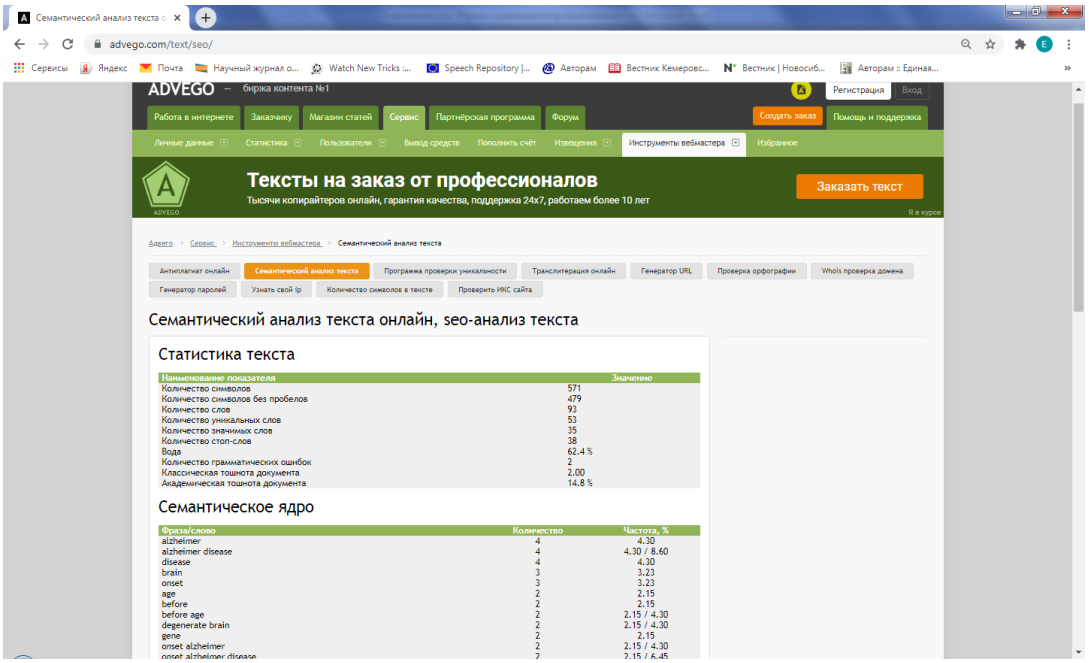


Рис. 3. Статистика текста. Семантическое ядро

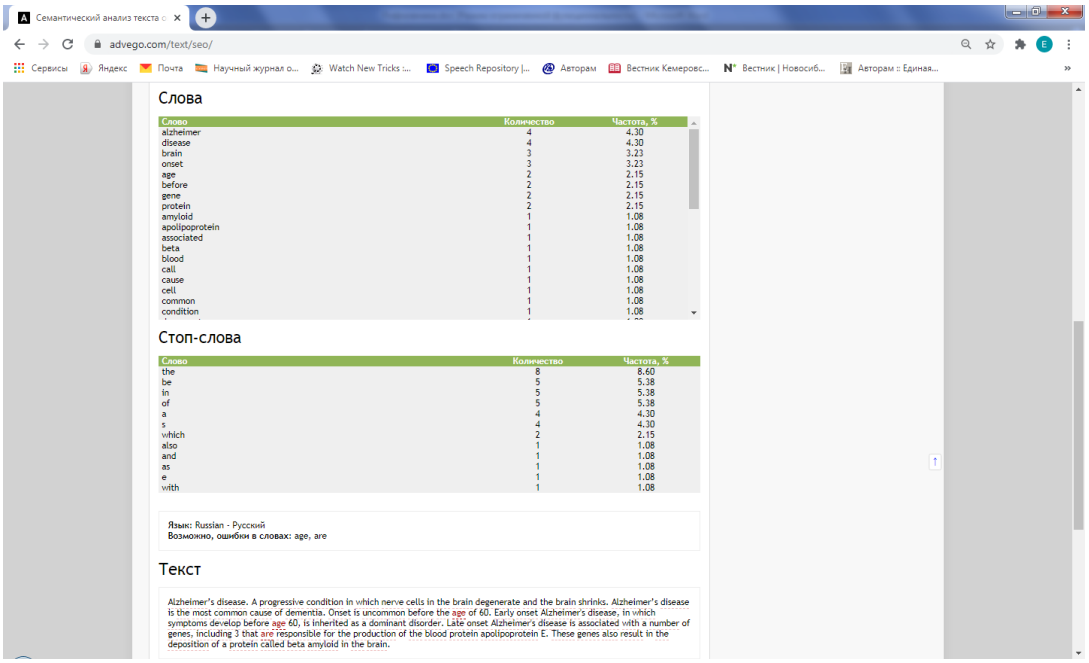


Рис. 4. Анализатор текста с использованием списка стоп-слов

Семантический анализ текста можно быстро выполнить и на других сайтах https://mir-atext.ru/seo_analiz_text, https://textis.ru/podschet-simvolov-onlayn/, https://text.ru/seo, https://seogift.ru/

tools/generator-klyuchevyh-slov-s-teksta/, https://voyant-tools.org/, https://www.online-utility.org/text/analyzer.jsp, https://seoscout.com/tools/keyword-analyzer. Проверить можно текст как на английском, так и на русском языке, за исключением последней ссылки.

При выполнении семантического анализа необходимо учитывать следующее: хотя программы и обладают стандартным алгоритмом, результаты могут немного отличаться.

🕮 Вопросы и задания

1. Зачем нужен семантический анализ текста?
2. Что такое тошнота текста? Какие бывают виды тошноты?
3. Что собой представляют стоп-слова?
4. Проанализируйте текст с семантической точки зрения и выявите ключевые слова:

Susan Davidson, the chief executive of the Zac Posen fashion house, remembers seeing great potential in the raw, cavernous space offered for sale in a co-op conversion at 82 Greene Street in the late 1980s.

The top-floor loft in a cast iron building in the SoHo-Cast Iron HistoricDistrict, for which she and her former husband paid nearly $460,000, was once part of a textile warehouse. What it lacked in walls, fixtures and finishes, it more than made up for in industrial charm.

“It was one big open space with plywood floors,” said Ms. Davidson, who has helped run other retailers besides Zac Posen, including Scoop NYC, Liz Claiborne and DKNY Jeans.

Today the impeccably revamped loft, now a triplex of about 3,200 square feet, can easily be described as rustic chic, retaining some of its early-1870s architectural detail alongside modern conveniences. Ms. Davidson and her current husband, Allen Miller, a corporate lawyer, spent around $4 million to painstakingly transform the space with the help of the designer Todd Klein, and contracted Tyler Horsley to landscape three terraces totaling around 1,100 square feet with plants, flowers and fruit-bearing trees.

But the couple now plan to move to an even older home – a late-1830s house on 30 acres in the Hudson Valley, where entire fruit orchards can be harvested. They are putting the loft, unit No. 5F, just off Spring Street, up for sale. The asking price is $11.75 million, according to Rebecca Edwardson and Bonnie Chajet of Warburg Realty, who are listing the apartment.

The monthly maintenance is an unusually low $1,592, reflecting the co-op building’s paid-off mortgage. (Ms. Davidson also noted that the five-story structure underwent an extensive renovation about five years ago that included new windows and a roof.)

The home has three bedrooms, three baths and six skylights. The first level – with dark-stained oak floors, 13-foot ceilings and original wood columns – is entered through double doors with a glass transom just off a semiprivate elevator landing where potted kumquat and lemon trees greet visitors.

The foyer flows into a gallery and a spacious dining room with a connected sitting area – a space where numerous parties have been held, including fund-raisers attended by Hillary Clinton and other dignitaries. The chef’s kitchen, separated with pocket doors, has marble countertops and high-end appliances, including two Miele dishwashers.

Nearby is a marble bath that leads to a sizable dressing area and a walk-in closet, where Ms. Davidson keeps her extensive collection of shoes, handbags and designer frocks, many from Zac Posen.

5. Проверьте текст из задания №4 на наличие воды и сократите ее количество.
6. Посчитайте в тексте из задания №4 количество уникальных и значимых слов.
7. Выявите последовательности слов в тексте из задания №4.

§2 Частотный анализ текста, оценка читабельности и фоносемантический анализ

Для проведения частотного анализа текста существует, например, бесплатный инструмент на www.abakbot.ru/online-5/97-freq-letter и аналогичный англоязычный ресурс на www.dcode.fr/frequency-analysis.

Частотный анализ – это один из методов криптоанализа, основывающийся на предположении о существовании нетривиального статистического распределения отдельных символов и их последовательностей как в открытом, так и в шифрованном тексте, которое с точностью до замены символов будет сохраняться в процессе шифрования и дешифрования.

Частотный анализ предполагает, что частота появления заданной буквы алфавита в достаточно длинных текстах одна и та же для разных текстов одного языка. При этом в случае моноалфавитного шифрования, если в шифрованном тексте будет символ с аналогичной вероятностью появления, то можно предположить, что он и является указанной зашифрованной буквой. Аналогичные рассуждения применяются к биграммам (двубуквенным последовательностям), триграммам в случае полиалфавитных шифров.

Данный вид анализа основывается на том, что текст состоит из слов, а слова из букв. Количество различных букв в каждом языке ограничено, и буквы могут быть просто перечислены (Рис. 5). Важными характеристиками текста являются повторяемость букв, пар букв (биграмм) и вообще N-ок (N-грамм), сочетаемость букв друг с другом, чередование гласных и согласных и некоторые другие.

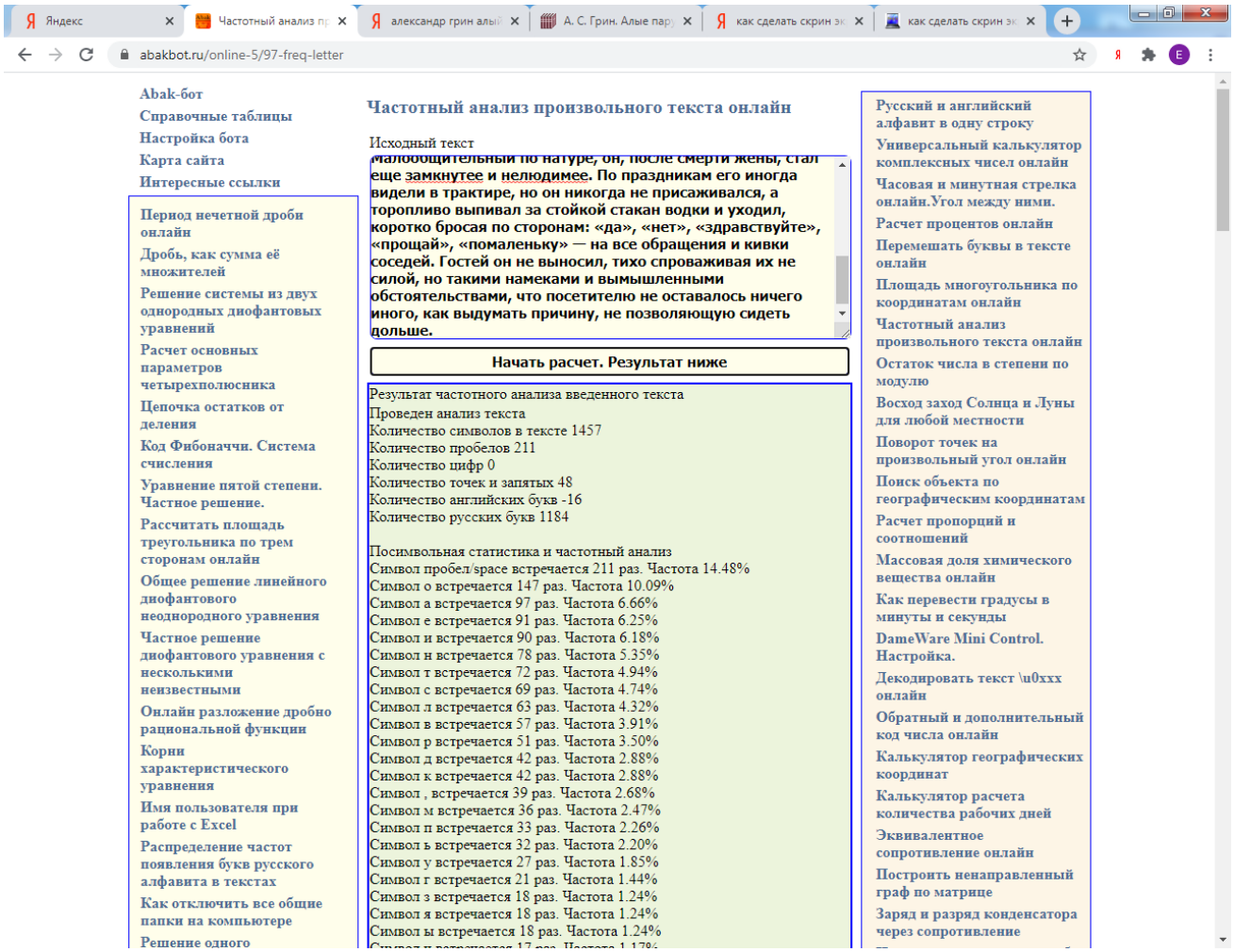


Рис. 5. Анализ сочетаемости букв и звуков

Существуют он-лайн инструменты для определения уровня читабельности текста. Проанализировать русскоязычный текст можно на бесплатном ресурсе ru.readability.io и англоязычный текст на ресурсе www.readable.com. Уровень читабельности текста определяется такими индексами как:

• Формула Flesch-Kincaid (Индекс удобочитаемости) – мера определения сложности восприятия текста читателем. Индекс удобочитаемости может вычисляться на основе нескольких параметров: длины предложений, слов, удельного количества наиболее частотных (или редких) слов и т. д.

• Индекс Колман-Лиау: индекс удобочитаемости, который наряду с индексом ARI может использоваться для определения сложности восприятия текста читателем путём аппроксимирования сложности текста к номеру класса в американской системе образования, ученикам которого данный текст будет понятен. Данный индекс был разработан Мэри Колман (Meri Coleman) и Т. Л. Лиау (T. L. Liau).

• Формула Дэйла-Чейла: это тест читаемости, который предоставляет числовой показатель сложности понимания, которую читатели испытывают при чтении текста. Он использует список из 3000 слов, которые группы американских студентов четвертого класса могли бы достоверно понять, считая любое слово, не включенное в этот список, сложным.

• Automatic Readability Index: мера определения сложности восприятия текста читателем, аппроксимирующая сложность текста к номеру класса в американской системе образования, ученикам которого данный текст будет понятен.

• SMOG: является мерой читаемости, которая оценивает период обучения, необходимый для понимания части текста. SMOG - аббревиатура от "Simple Measure of Gobbledygook" (Рис.6).

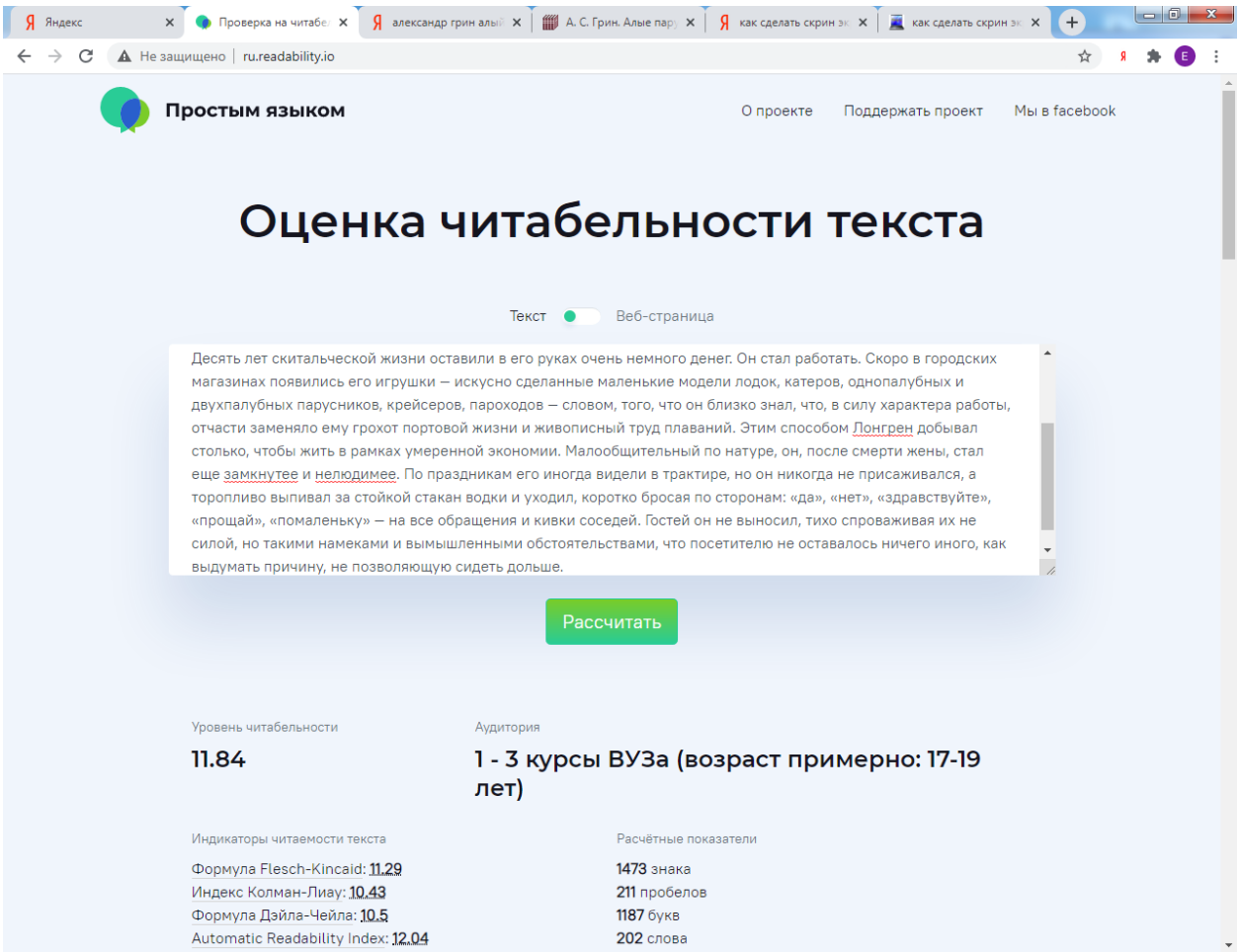


Рис. 6. Оценка читабельности текста

Современные он-лайн инструменты позволяют провести фоносемантический анализ текста, например, с помощью программы ВААЛ-мини - www.vaal.ru/prog/setupvm.exe. Программа является двуязычной, обрабатывает тексты на русском и украинском языках. Она позволяет оценивать фонетическое значение слов и текстов, кроме того, добавлена оценка звуко-цветового значения слов и текстов.Слова и тексты могут быть оценены по алгоритмам А.П.Журавлева и В.В.Левицкого. Программа распространяется одновременно в двух реализациях - в виде исполняемого файла **VaalMini.exe** и в виде библиотеки **VaalMini.dll**, которая подключается к редактору **Microsoft Word** и позволяет производить фонетическую оценку прямо в нем (Рис.7).

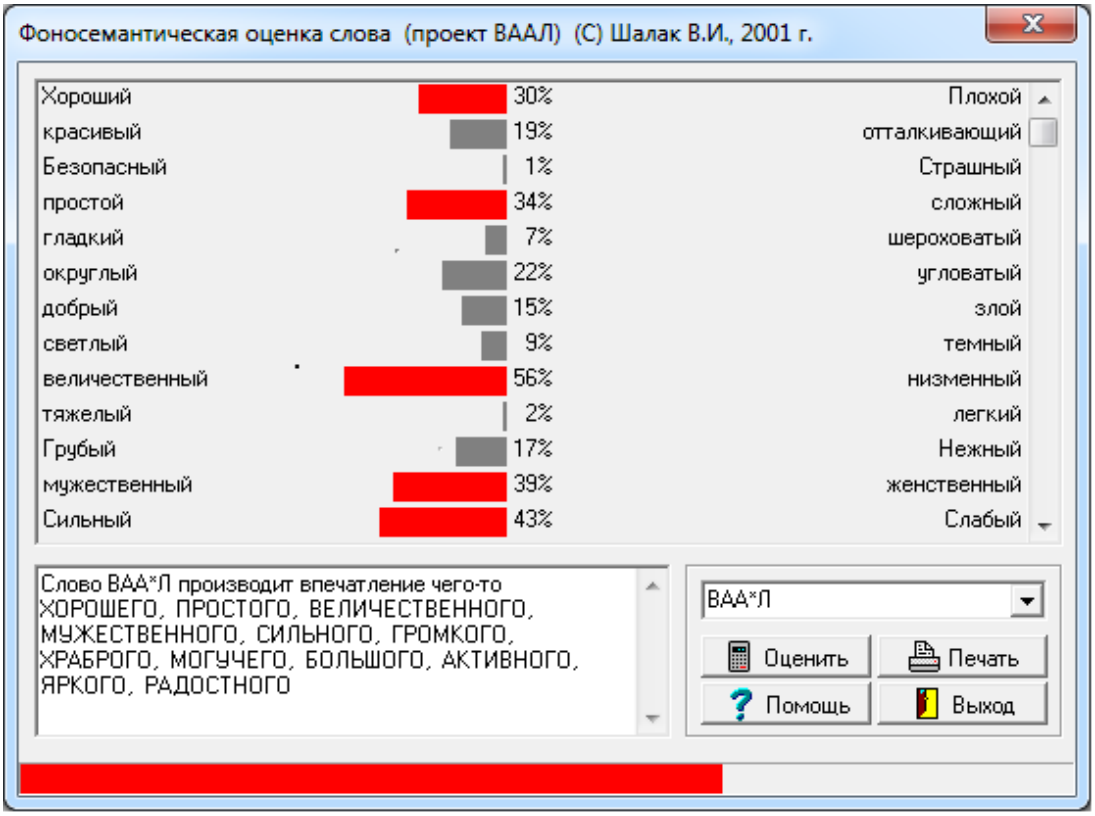


Рис. 7. Фоносемантическая оценка слова

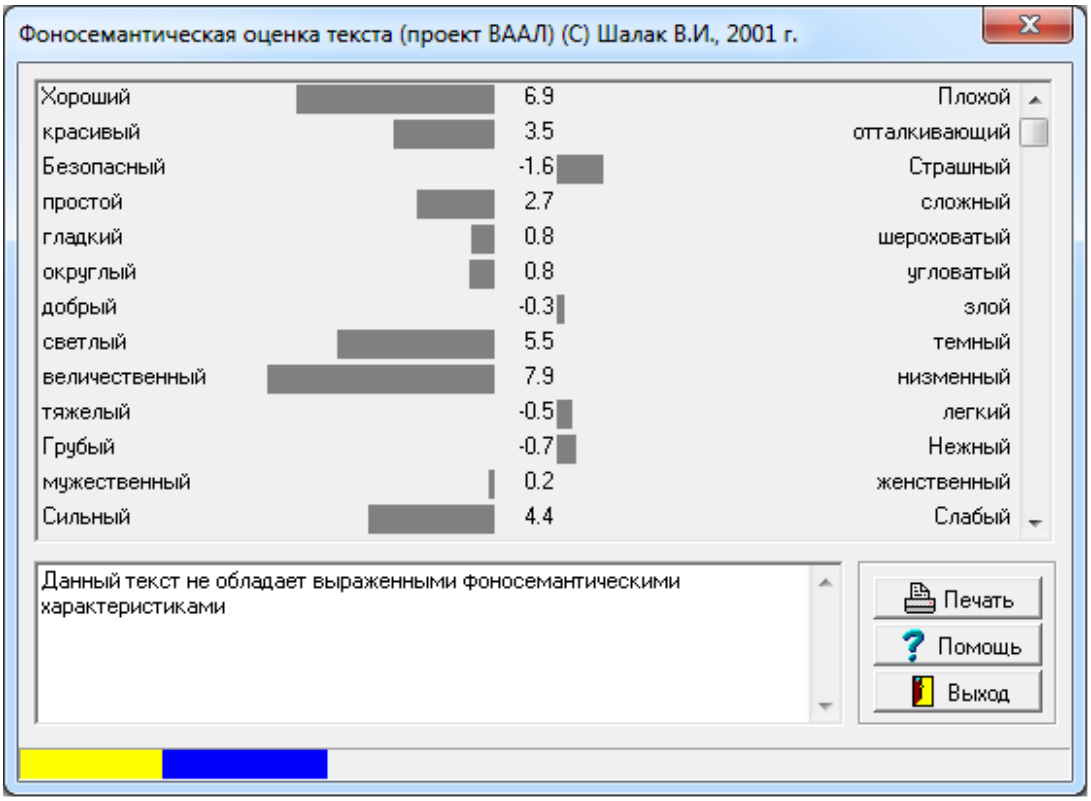


Рис. 8. Фоносемантическая оценка текста

Существуют он-лайн инструменты, с помощью которых можно проверить грамотность, но эти ресурсы являются платными, например, www.orfogrammka.ru. С помощью данной программы можно проверить грамотность – пунктуацию, орфографию, грамматику и стилистику; красоту – программа найдёт тавтологии и неблагозвучия, подберёт синонимы и эпитеты; проверит качество текста – оценит SEO-параметры: воду, частотные и неестественные сочетания.

🕮 Вопросы и задания

1. Используя инструменты для SEO анализа текста, проанализируйте отрывок художественного текста на языке оригинала и на языке перевода, сопоставьте полученные результаты. Определите отличительные черты текста в оригинале и при переводе.
2. Используя он-лайн инструменты для частотного анализа текста (https://abakbot.ru/online-5/97-freq-letter и https://www.dcode.fr/frequency-analysis), проанализируйте отрывок художественного текста на языке оригинала и в переводе на частотность гласных – о, а, у, а также согласных –п, к, ш. Определите изменения частотности данных букв.
3. Определите в художественном тексте Формулу Flesch-Kincaid (Индекс удобочитаемости) – мера определения сложности восприятия текста читателем. Индекс удобочитаемости может вычисляться на основе нескольких параметров: длины предложений, слов, удельного количества наиболее частотных (или редких) слов и т. д.
4. Определите у выбранного вами текста Формулу Дэйла-Чейла: это тест читаемости, который предоставляет числовой показатель сложности понимания, которую читатели испытывают при чтении текста. Он использует список из 3000 слов, которые группы американских студентов четвертого класса могли бы достоверно понять, считая любое слово, не включенное в этот список, сложным.
5. Определите у выбранного Вами текста индекс SMOG: который является мерой читаемости и оценивает период обучения, необходимый для понимания части текста. SMOG – аббревиатура от "Simple Measure of Gobbledygook".
6. В выбранном Вами рассказе с помощью он-лайн инструмента, определите количество слов с более чем 4-мя слогами и количеством слов до 4-х слогов включительно.
7. Оцените фоносемантическое значение слов в выбранном Вами художественном тексте.

§3 Рифмовники

Если Вы решили начать писать стихи или прозу и Вам необходимо придумать рифму, на помощь придут рифмовники и генераторы рифм. Таких инструментов в настоящее время существует большое количество. Большинство из них просты в обращении, поэтому не представляет труда подобрать рифму к слову.

Например, по адресу https://rifmovnik.ru/ Вы можете найти бесплатную программу **Rhymes** для Windows и iOS, которая позволяет легко подобрать рифму, синоним или эпитет, узнать лексическое значение и произношение слова, посмотреть примеры его употребления. Это универсальный словарный инструмент, который будет полезен всем, пишущим тексты на русском языке.

В ее состав входит:

• Большой словарь рифм А.А. Зализняка (5,4 млн. словоформ / 177 тыс. слов)

• Викисловарь (только для iOS), который содержит 430 тыс. статей: толкования, примеры, синонимы и антонимы и др.

• Дополнительные словари (только для Windows): а) Грамматический словарь (156 тыс. слов / 4,5 млн. словоформ); б) Орфоэпический словарь (98 тыс. слов); в) Большой толковый словарь (110 тыс. слов / 66 тыс. статей); г) Современный словарь синонимов (46 тыс. слов и выражений); д) Словарь русских синонимов и сходных по смыслу выражений Н. Абрамова (ретро-словарь (1915 г.) на 20 тыс слов и фраз); е) Словарь эпитетов (8700 эпитетов к 1300 опорным словам).

Программа обладает следующими возможностями:

– Вызов из других приложений (Карточка из буфера обмена. Копирование результатов, т.е. рифм и словоформ, в буфер обмена. На iPad с iOS 9, 10 поддержаны режимы Slide Over и Split View).

– Интеллектуальный словник (Список вариантов слов при последовательном наборе. Поиск статей по любой форме слова. Определения ударения слова (при подборе рифм) и др.).

– Фильтрация и сортировка результатов (В Словаре рифм и в Грамматическом словаре – по алфавиту, встречаемости, части речи, качеству (для рифм), количеству слогов.

– Качественная визуальная разметка статей (Расширенная навигация в карточках. Слова и фразы как ссылки и др.).

– История и Избранное (Возможность отобрать рифмы из списка результатов для дальнейшего использования. Отобранные слова сохраняются в Истории совместно с запросом. Статьи из Викисловаря, просмотренные ранее, доступны офлайн).

– Удобное управление (Основные операции доступны с клавиатуры. Можно почти не использовать мышь).

Кроме того, на сайте есть Генератор рифм в Разделе «Рифма онлайн» (Рис. 9). В настройках можно выбрать какую рифму вы хотите подобрать (точную, хорошую, среднюю, плохую, любую), указать ее часть речи, к какой лексике рифма будет относиться (базовая, частотная, обычная, редкая, уникальная) и количество слогов.

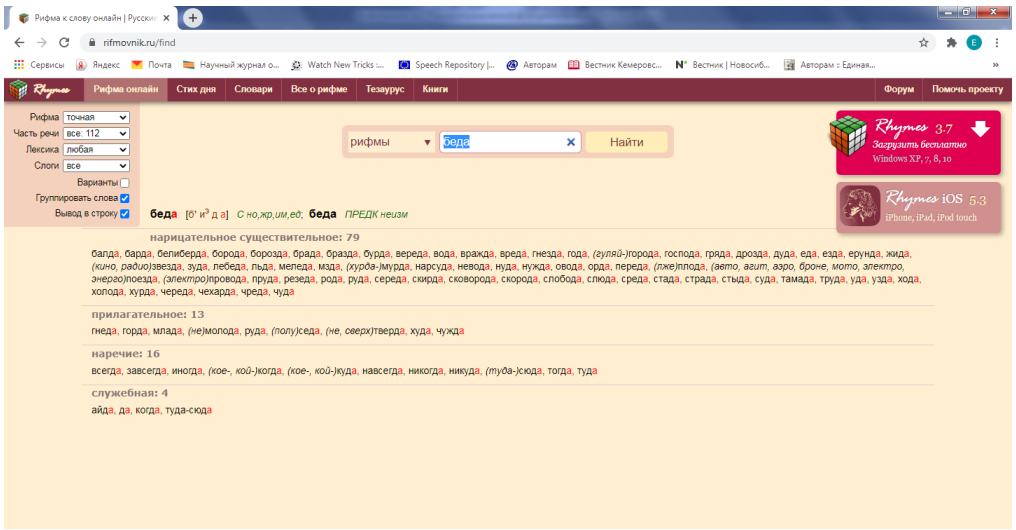


Рис. 9. Рифма онлайн

Раздел «Стих дня» представляет собой поэтический календарь: дается информация, кто из поэтов родился в тот или иной день и приводится несколько его или ее стихотворений.

В разделе «Все о рифме» можно ознакомиться с книгой В. Жирмунского «Рифма, ее история и теория» (Петроград, 1923 г.)

В разделе «Тезаурус» даются ссылки на тезаурусы и словари, которые могут помочь автору при создании рифм.

В разделе «Книги», даются ссылки на работы, встречающиеся на сайте. Также на сайте присутствует раздел «Форум» и «Помочь проекту».

Для постоянной работы предлагается использовать бесплатное приложение **Rhymes,** которое не требует доступа в Интернет. Его можно установить на Windows XP, 7, 8, 10 (Rhymes 3.7) и на iPhone, iPad, iPod touch (Rhymes iOS 5.3).

У программы есть и недостаток: не всегда может найти рифму даже к простым словам, типа *ночь*.

Далее, можно создать рифму и на сайте https://rifmik.net/. Здесь дается краткое описание, что собой представляет рифма, и какой она бывает. Можно посмотреть популярные слова, и что недавно искали другие пользователи.

Алгоритм поиска и создания рифмы прост: нужно ввести слово в окно поиска (рис. 10), при этом предлагается выбрать словоформу или производное слово. Практически сразу показывается количество найденных слов и предлагаются рифмы из 1, 2, 3 и более слогов (в зависимости от заданного слова). Также можно воспользоваться фильтрами (выбрать часть речи, упорядочить по силе рифмы и популярности слова и др.).

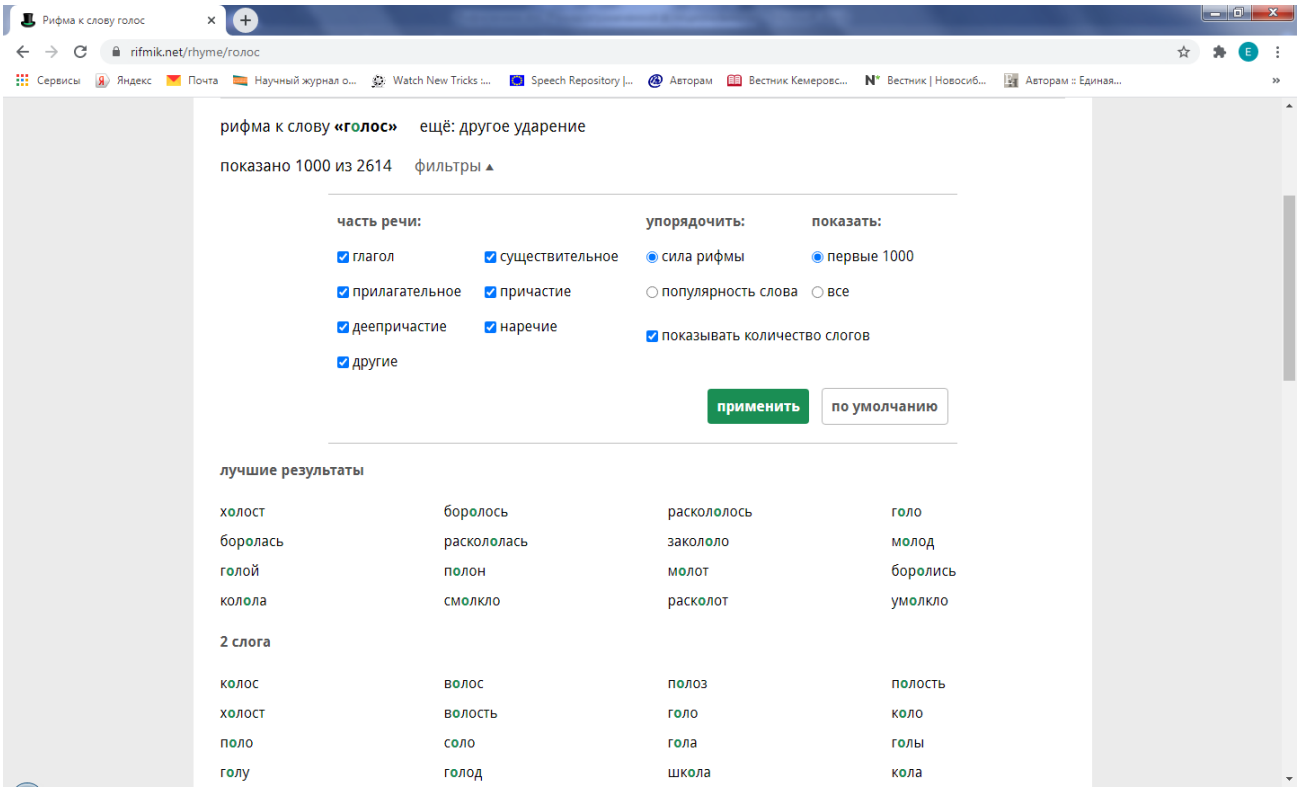


Рис. 10. Создание рифмы на сайте rifmik

Другие инструменты по созданию рифмы на русском языке можно найти на таких сайтах, как https://rifmus.net/, https://rifme.net/, https://stihi.ru/, https://ryfma.com, https://rifma-online.ru/, http://neogranka.ru/podbor_rifmy.html, https://poeziya.ru/rhyme/ и т.д.

Кроме того, существуют программы, которые создают целые стихотворения на русском языке: https://ultragenerator.com/stihov/, http://neogranka.ru/generator_stihov.html, https://stihizakazhu.ru/stihi-besplatno/kiberpoet.php.

Программы по созданию рифмы на английском языке работают по таким же принципам. Например, легко подобрать рифму с помощью инструмента **Rhymezone** (https://www.rhymezone.com/,) (Рис. 11).

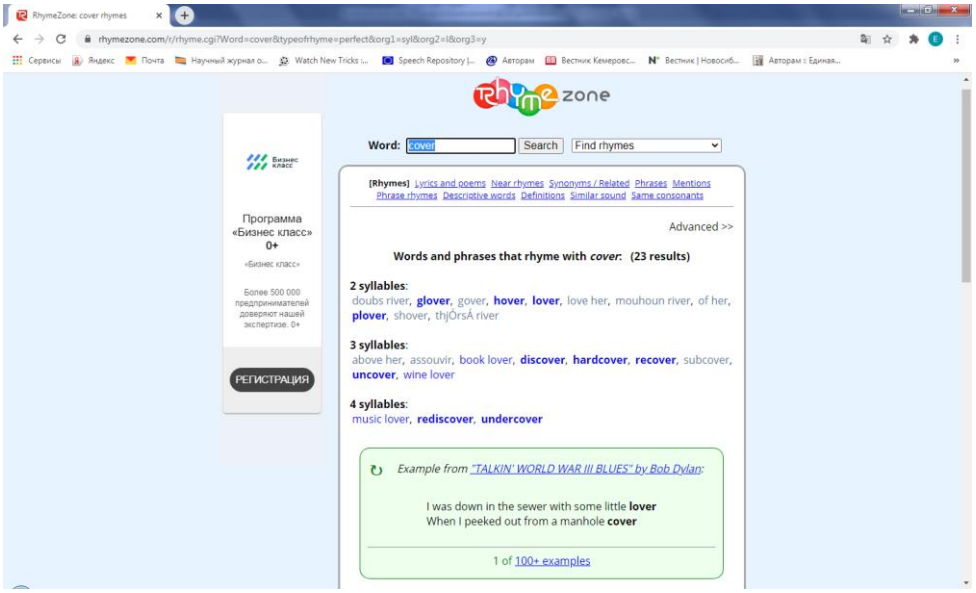


Рис. 11. Поиск рифмы на сайте rhymezone

К другим популярным генераторам рифмы можно отнести https://www.double-rhyme.com/, https://www.rhymer.com/, https://www.rhymes.net. Рифму можно подобрать и в так называемых словарях рифм на английском языке. Например, https://www.festisite.com/poems/rhyme/, http://www.b-rhymes.com/.

Как и в русском языке, существуют программы, которые создают на английском языке двустишия https://www.poem-generator.org.uk/rhyming-couplets/ и даже целые стихотворения https://www.poem-generator.org.uk/. На последнем сайте можно создать такие виды стихотворного творчества, как quick poem, free verse, haiku, didactic cinquain, rhyming couplets, sonnet, villanelle, limerick, acrostic, love poem, narrative poem, concrete, tanka.

Также существует онлайн калькуляторы, которые помогают посчитать количество слогов в слове и предложении. Такой инструмент можно найти на таких сайтах, как https://planetcalc.ru (считает количество гласных в строке), https://slogi.su (считает только слоги), https://slogislova.ru/ и https://syllables.ru/ (разбивает слово на слоги, прописывает, какие именно слоги входят в слово и расставляет переносы в соответствии с правилами русского языка) (Рис. 12).

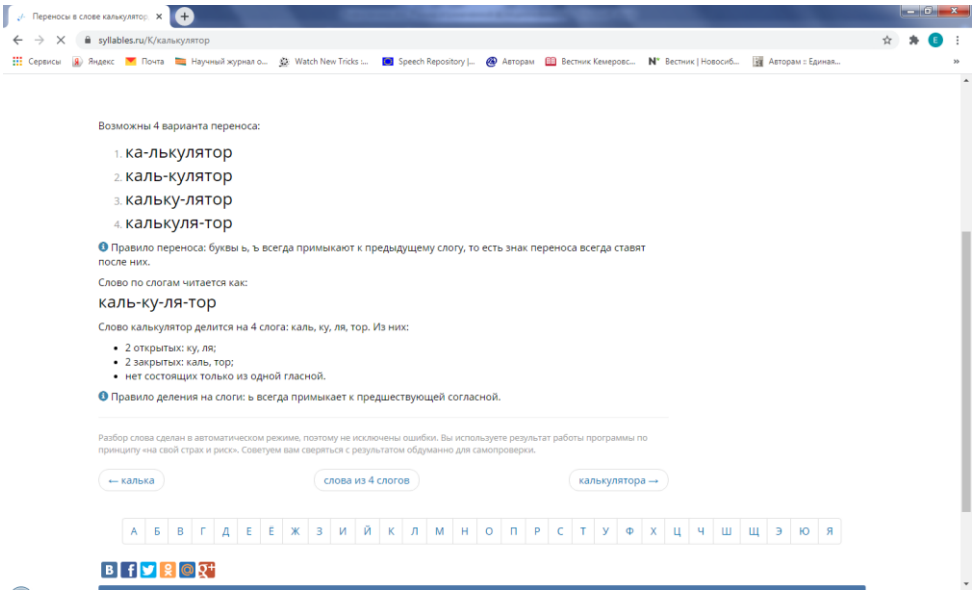


Рис. 12. Онлайн калькулятор для подсчета слогов

Для работы с английскими словами и текстами можно использовать калькуляторы на следующих сайтах https://syllablecounter.net/ (считает только слоги), https://www.wordcalc.com/ (считает количество слогов, слов, предложений и букв), https://syllablecounter.org/ (считает количество слогов,

слов и знаков), http://www.syllablecount.com/ (считает количество слогов, слов, знаков, количество каждой буквы), https://www.howmanysyllables.com (считает количество слогов, показывает деление слова на слоги, ударный слог, транскрипцию и дает возможность услышать, как слово произносится) (Рис. 13).

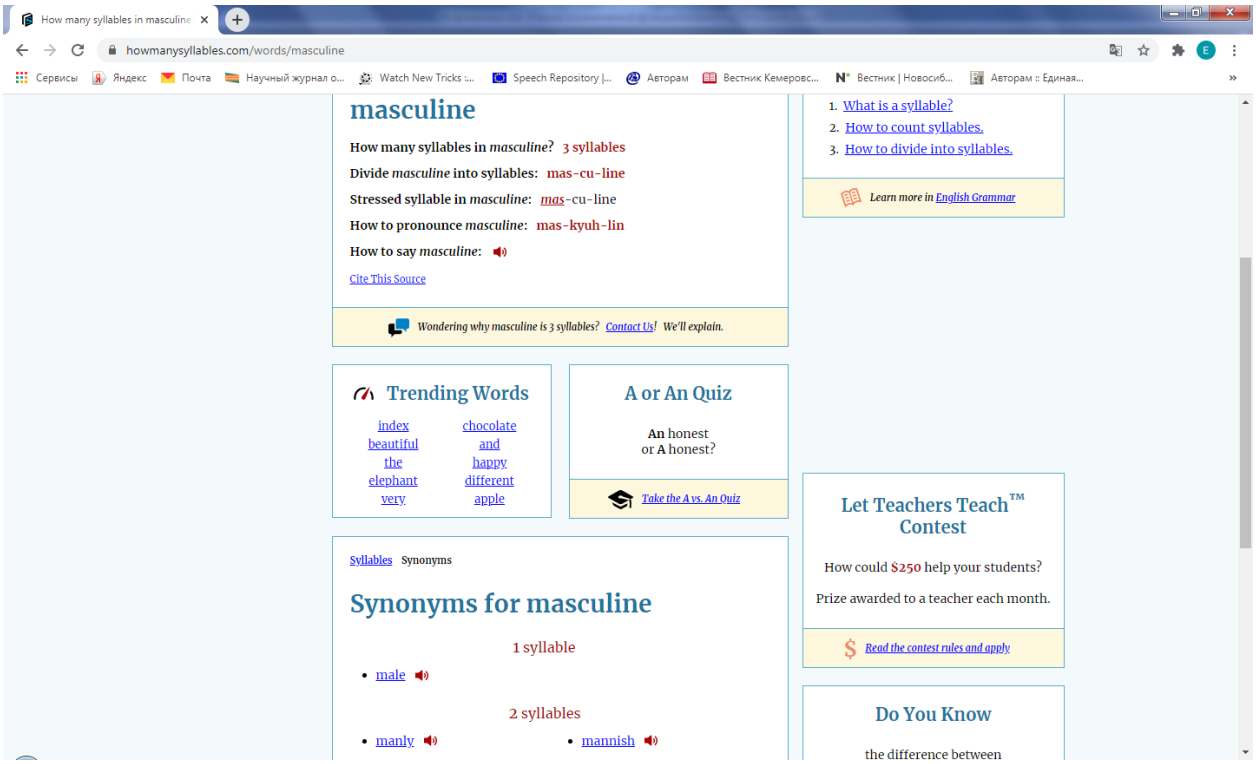


Рис. 13. Калькулятор слогов для англоязычных текстов

🕮 Вопросы и задания

1. Что отличает рифмовник Rhymes от других подобных инструментов?
2. Подберите рифму к следующим словам: колесо, дух, время, философия, контраст; date, program, waterfall, bicycle, mobile. Какие из предложенных вариантов вы считаете наиболее удачными и почему?
3. Используя следующие слова, создайте стихотворение на русском языке: вечер, дорога, жара.
4. Используя следующие слова, создайте стихотворение на английском языке: evening, road, heat.
5. Дайте перевод на русский язык и отличия следующих видов стихотворного творчества: quick poem, free verse, haiku, didactic cinquain, rhyming couplets, sonnet, villanelle, limerick, acrostic, love poem, narrative poem, concrete, tanka.
6. Посчитайте количество знаков, слогов и слов в следующих предложениях: а) One of the best things any household can do to protect their finances is to amass some savings so that they do not have to rely on credit in an emergency. б) Прекрасная квартира в импозантном викторианском здании с прямым выходом в закрытый коммунальный сад и всего в нескольких минутах ходьбы от станции метро.
7. Какой из предложенных калькуляторов вы считаете наиболее эффективным? Обоснуйте свой выбор.

§4 Описание инструмента #LancsBox 5.0

#LancsBox представляет собой систему программного обеспечения, разработанную сотрудниками Университета Ланкастер [http://corpora.lancs.ac.uk/lancsbox/docs/pdf/LancsBox_5.0_manual.pdf]. Данный инструмент используется для анализа больших по объему языковых данных и корпусов.

Основные характеристики #LancsBox:

- может быть использован для обработки ваших собственных данных и существующих корпусов;
- может быть использован широким кругом специалистов: лингвистами, преподавателями, историками, социологами и другими, кто так или иначе связан или интересуется исследованиями в области языка;
- может визуализировать языковые данные;
- может обрабатывать данные на любом языке;
- может автоматически аннотировать данные по частям речи;
- работает со всеми основными операционными системами (Windows, Mac, Linux).

1. Установка #LancsBox

#LancsBox является инструментом нового поколения для анализа корпусов текстов на разных языках. Прежде всего, Version 5 была разработана для 64-разрядных операционных систем (Windows 64-bit, Mac, Linux), что в свою очередь обеспечивает лучшую работу инструмента. #LancsBox может быть также использован и более старыми 32-разрядными системами, но тогда некоторые функции будут ограничены.

Скачать и загрузить инструмент можно по следующей ссылке: http://corpora.lancs.ac.uk/lancsbox/download.php.

2. Загрузка и импортирование данных

Данные могут быть загружены и импортированы в #LancsBox через вкладку ‘Corpora’ (Рис. 14). Эта вкладка открывается автоматически, когда вы запускаете #LancsBox. данный инструмент работает с корпусами в любом формате (.txt, .xml, .doc, .docx, .pdf, .odt, .xls, .xlsx, .zip etc.), а также и со списками слов (.csv). Существует два способа загрузки корпусов и списков слов: 1. загрузить (свои собственные) данные; 2. загрузить данные, которые представлены в самом инструменте #LancsBox.

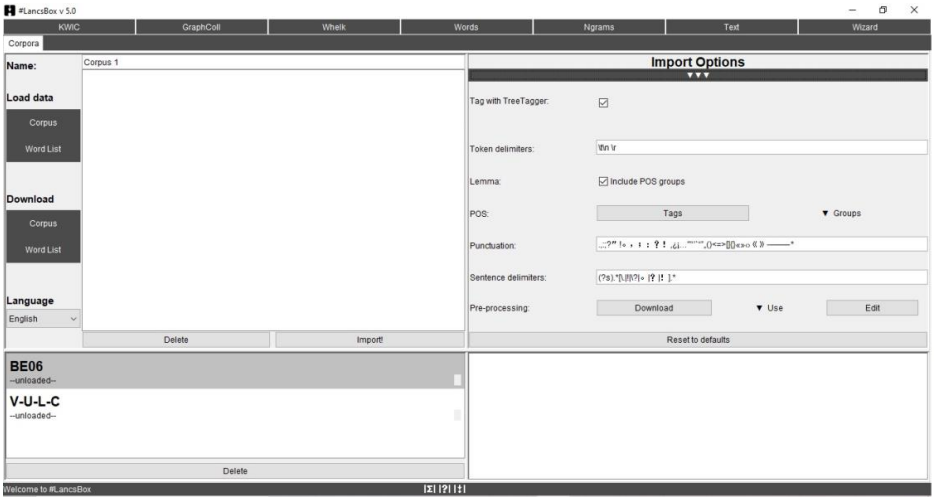


Рис. 14. Вкладка Corpora

В верхней панели вы можете:

- выбрать свой корпус или список слов для загрузки;
- загрузить корпус или список слов, представленные самим инструментом #LancsBox;
- выбрать язык;
- пересмотреть разметку по частям речи;
- пересмотреть знаки препинания и членение предложения;
- установить первичную обработку данных через скрипты, настраиваемые по вашим техническим заданиям.

В нижней панели вы можете:

- активировать или удалить импортированные корпуса или списки слов;
- пересмотреть размер корпуса и текста (лексемы, леммы, формы слов);
- предварительно просмотреть текст;
- сохранить обработанные корпуса с разметкой по частям речи (POS) и т.д.

Инструмент #LancsBox позволяет легко работать со своими собственными корпусами и списками слов. Это те корпуса, которые хранятся на вашем компьютере или на любом носителе информации или облаке.

Чтобы загрузить ваш корпус или список слов вам нужно:

1. Во вкладке Corpora, кликнуть левой клавишей мыши по ‘Corpus’ или ‘Word List’ под ‘Load data’, в зависимости от того, собираетесь вы загружать корпус или список слов.
2. Откроется окно, в котором вы можете выбрать папку, в которой хранится ваш корпус или список слов.
3. Вы можете выбрать конкретный файл или несколько файлов при нажатии клавиш Ctrl и левой кнопки мыши по выбранным файлам, либо вы также можете выбрать все файлы в папке, нажав комбинацию клавиш Ctrl + A.
4. Кликните левой кнопкой мыши по ‘Open’, чтобы загрузить ваши документы/файлы.
5. Выберите язык вашего корпуса или списка слов. #LancsBox автоматически поддерживает лемматизацию (автоматическое составление словарей) и частеречую разметку на многих языках. Данный процесс осуществляется с помощью разметки в виде дерева (Tree Tagger). Если ваш язык не представлен списке, выберите ‘Other’; в этом случае, автоматическая лемматизация и частеречная разметка будут отключены.
6. [По выбору: Вы можете пересмотреть/ изменить опции импорта, кликнув левый кнопокой мыши по строке с тремя треугольниками (▲▲▲). В большинстве случаев, вы можете использовать выбор по умолчанию.]
7. Кликните левой кнопкой мыши по ‘Import!’, чтобы импортировать ваш корпус в #LancsBox. По умолчанию, #LancsBox автоматически добавит разметку по частям речи в ваш корпус.

#LancsBox позволяет вам работать с существующими корпусами, к которым открыт доступ при наличии лицензии. Существует два вида доступа к корпусам: 1) открытый доступ; 2) ограниченный доступ. Разработчики постоянно добавляют новые корпуса к списку.

1. Во вкладке ‘Corpora’, кликните левой кнопкой мыши ‘Corpus’ или ‘Word List’ под ‘Download’;
2. Откроется окно, в котором вы можете выбрать корпуса и списки слов, размещенные в #LancsBox. Кликнув левой кнопкой мыши по корпусу, вы увидите дополнительную информацию о корпусе или списке слов, включая язык, дату, тип текста, лицензию и т.д.
3. Просмотрите и согласитесь с лицензией корпуса.
4. Кликните левой кнопкой мыши по ‘Download’, чтобы скачать выбранный корпус или список слов.
5. Кликните левой кнопкой мыши по ‘Import!’, чтобы импортировать ваш корпус в #LancsBox. По умолчанию, #LancsBox автоматически добавляет разметку по частям речи в ваш корпус.

Примечание: Чтобы переключаться между открытым и ограниченным доступом к корпусу, используйте кнопку ‘Switch access’ в нижнем левом углу. Корпуса с ограниченным доступом размещены как зашифрованные и имеют несколько ограничений касательно демонстрации и использования. Например: они не могут быть отображены в инструменте ‘Text’ или сохранены на ваш компьютер.

Все корпуса и списки слов, которые были импортированы в #LancsBox, отображаются в нижней панели во вкладке ‘Corpora’. Эта панель позволяет посмотреть корпуса, совершить предпросмотр файлов и быстро перезагрузить корпуса и списки слов, когда #LancsBox закрыт или открыт заново.

1. Если вы импортировали корпус или список слов, он появится в нижней панели вместе с другими корпусами и списком слов, который вы уже импортировали. Их можно удалить, кликнув левой клавишей мышки по ‘delete’. В разделе, который находится внизу справа, вы можете посмотреть структуру корпуса: отдельных текстовых документов, из которых и состоит корпус.
2. В нижней панели (окно внизу слева) может быть указан корпус по умолчанию. Корпус по умолчанию – это корпус, который предлагает #LancsBox как выбор по умолчанию в отдельных модулях. Корпус по умолчанию может быть указан с помощью двойного нажатия левой клавиши мышки на названии корпуса; рядом с названием корпуса по умолчанию появится заполненный прямоугольник.

3. Если #LancsBox закрыт, корпус и списки слов останутся импортированными, но будут выгружены. Чтобы активировать (загрузить снова) корпуса или списки слов для использования, нажмите два раза левой клавишей мыши по корпусам или спискам слов.

4. Вы можете также совершить предпросмотр документов, щелкнув правой клавишей мыши по ним. Они появятся в инструменте ‘Text’. Список документов, включая информацию об их размере, может быть скопирован (Ctrl/Command+C) и вставлен (Ctrl/Command+V) в электронную таблицу или текстовой документ.

5. Теперь корпуса готовы к анализу, используя пять модулей: KWIC, Whelk, GraphColl, Words и Text. Списки слов могут быть использованы в инструменте ‘Words’.

В качестве примера мы загрузили корпус текстов, размещенных в инструменте #LancsBox. Как видно ниже, на рисунке 15 представлена информация по автору текстов (Austen): язык (English), количество документов (8 files), количество лексем (777966 tokens), количество форм слов (20820 types) и количество лемм (18888 lemmas).

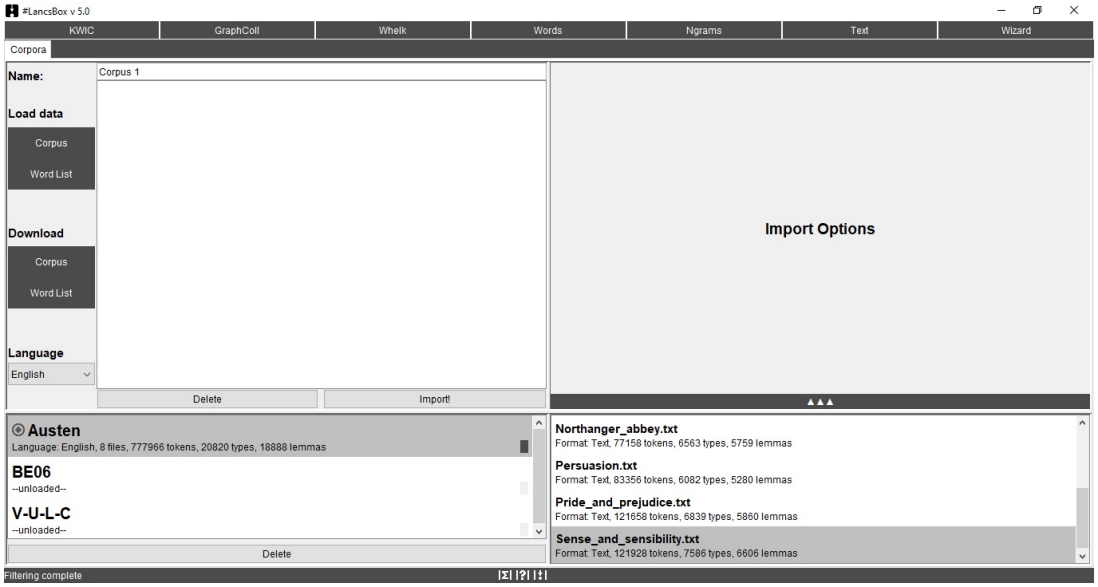


Рис. 15. Информация по тексту

Сохранение корпуса

#LancsBox сохраняет корпуса в горизонтальном или вертикальном формате.

1. Щелкните правой клавишей мыши по корпусу, который вы хотите сохранить.
2. Выберите соответствующие опции (Рис. 16).

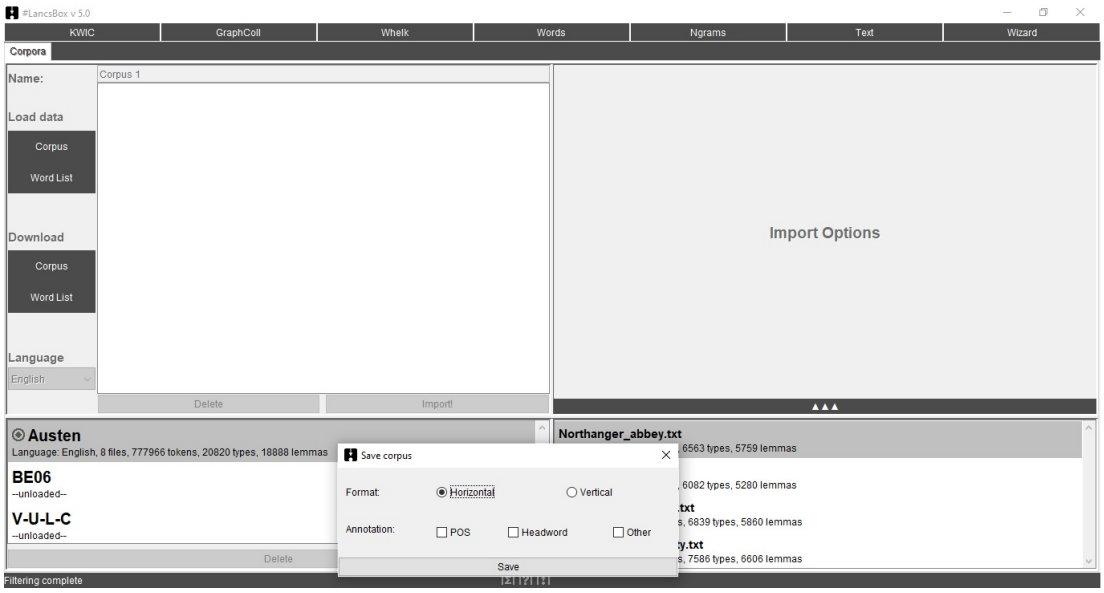


Рис. 16. Сохранение корпуса

3. Щелкните ‘Save’. Выберите папку на своем компьютере, куда хотите сохранить документ (Рис. 17).

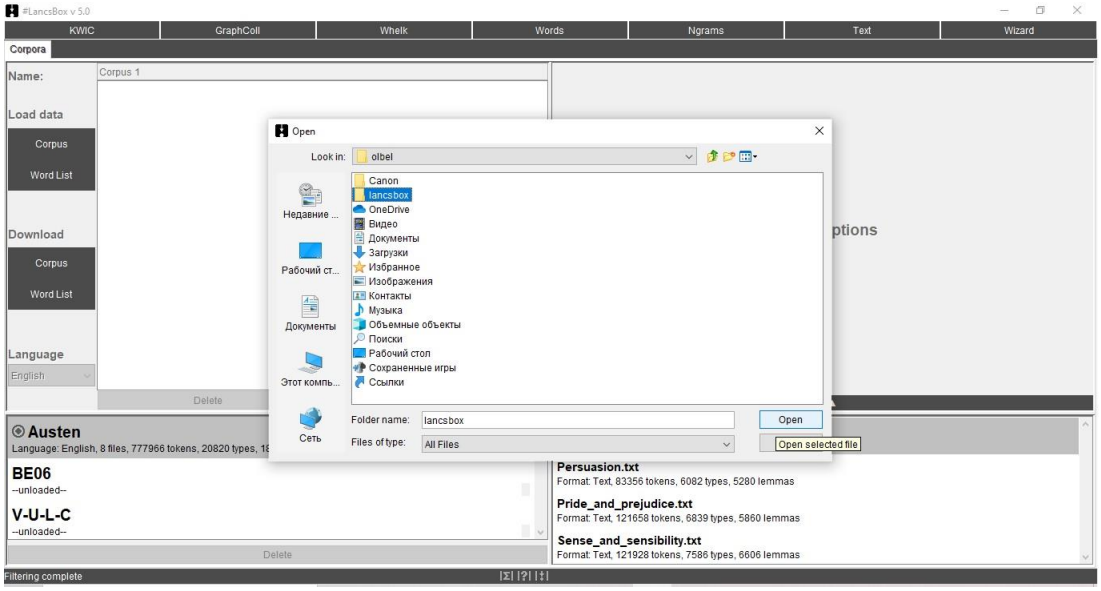
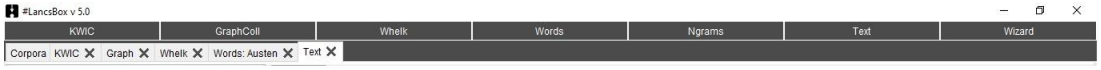


Рис. 17. Сохранение документа

Инструменты и вкладки

#LancsBox поддерживает осуществление нескольких анализов одновременно и работу с несколькими корпусами. Инструмент включает в себя пять основных модулей (инструментов): KWIC, Whelk, GraphColl, Words и Text. Каждый инструмент можно открывать в виде отдельных вкладок. Модули в #LancsBox взаимосвязаны: они могут быть запущены в виде всплывающих окон внутри модуля.

1. Рисунок ниже показывает верхнюю строку в #LancsBox с клавишами для открытия отдельных модулей и нескольких вкладок.



2. Модули в #LancsBox имеют несколько параметров функционирования:

- KWIC отвечает за конкорданс (указатель, связывающий словоупотребление с контекстом);
- Whelk показывает распределение элемента поиска в документах корпуса;
- GraphColl определяет и визуализирует коллокации;
- Words составляет списки слов, определяет и визуализирует ключевые слова.;
- Ngrams составляет список N-грамм, определяет и визуализирует ключевые N-граммы;
- Text отображает полностью контекст элемента поиска.

Инструмент KWIC (Key Word in Context)

Данный инструмент составляет список всех случаев использования элемента поиска в корпусе в виде конкорданса (Рис. 18). Он может быть использован:

- для установления частоты использования слова или словосочетания в корпусе;
- установления частоты использованиях разных частей речи (существительных, прилагательных, глаголов);
- установления сложных лингвистических конструкций (например, использование пассивного залога) с помощью ‘smart searches’;
- сортировки, отбора и перемешивания строк конкорданса;
- осуществления статистического анализа, сравнивая использование элемента поиска в двух корпусах.

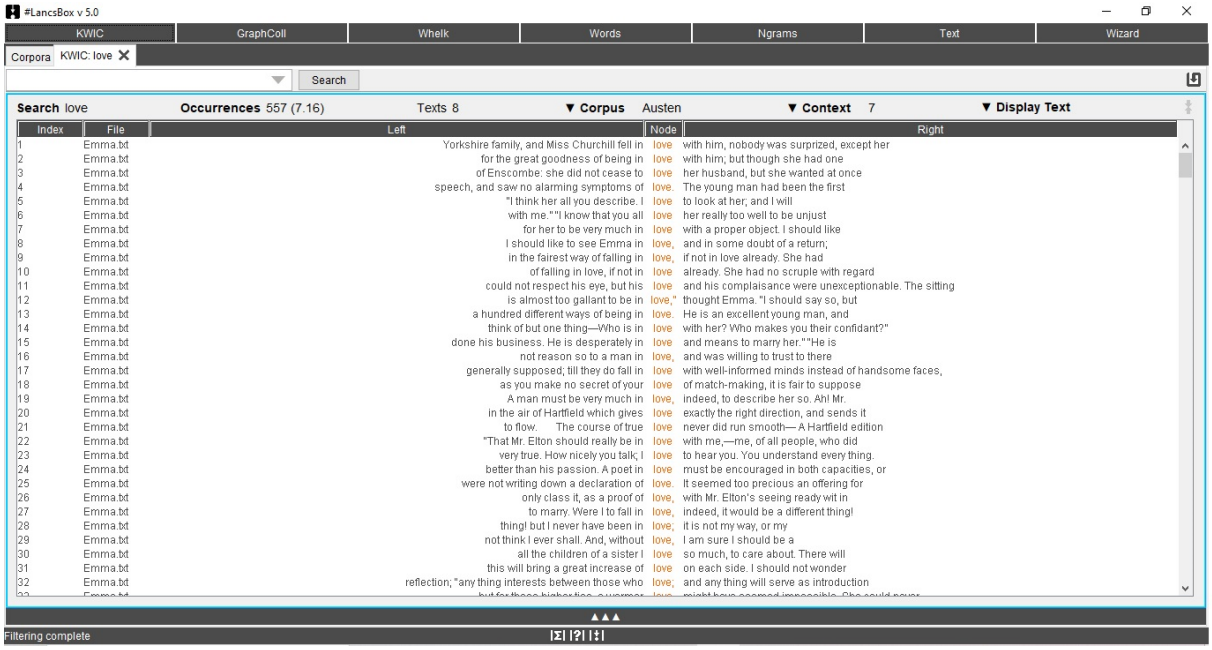


Рис. 18. Инструмент KWIC

Инструмент Whelk

Данный инструмент (Рис. 19) предоставляет информацию в отношении того, как элемент поиска распределен по всем документам корпуса. Он может быть использован для следующих целей:

- установления абсолютной и относительной частотности элемента поиска в документах корпуса;
- отбора результатов согласно разным критериям;
- сортировки документов в зависимости от абсолютной и относительной частотности элемента поиска.

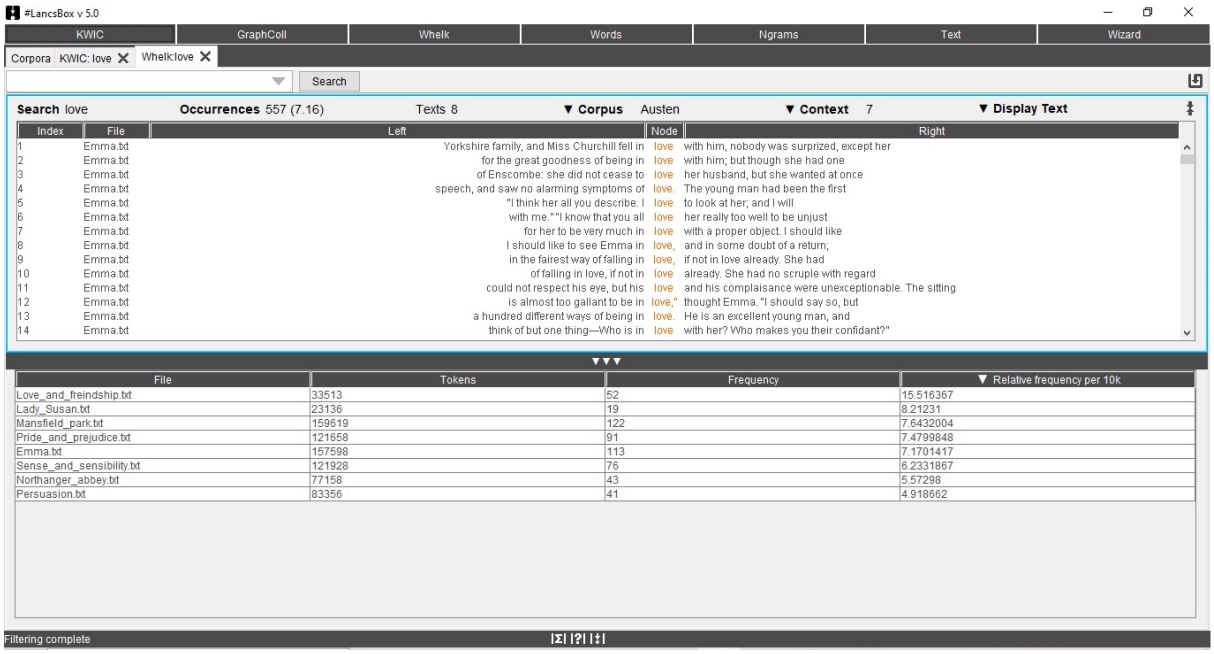


Рис. 19. Инструмент Whelk

Инструмент GraphColl

Данный инструмент определяет коллокации и отображает их в таблице, а также в виде диаграммы или сетки. GraphColl (Рис. 20) может быть использован для

- поиска коллокаций слов или словосочетаний;
- поиска коллигаций (совместное появление грамматических категорий в тексте);

• визуализации коллокаций и коллигаций;
• установления общих элементов коллокации слова или словосочетания;
• описания дискурса с точки зрения близости точек на диаграмме его коллокаций.

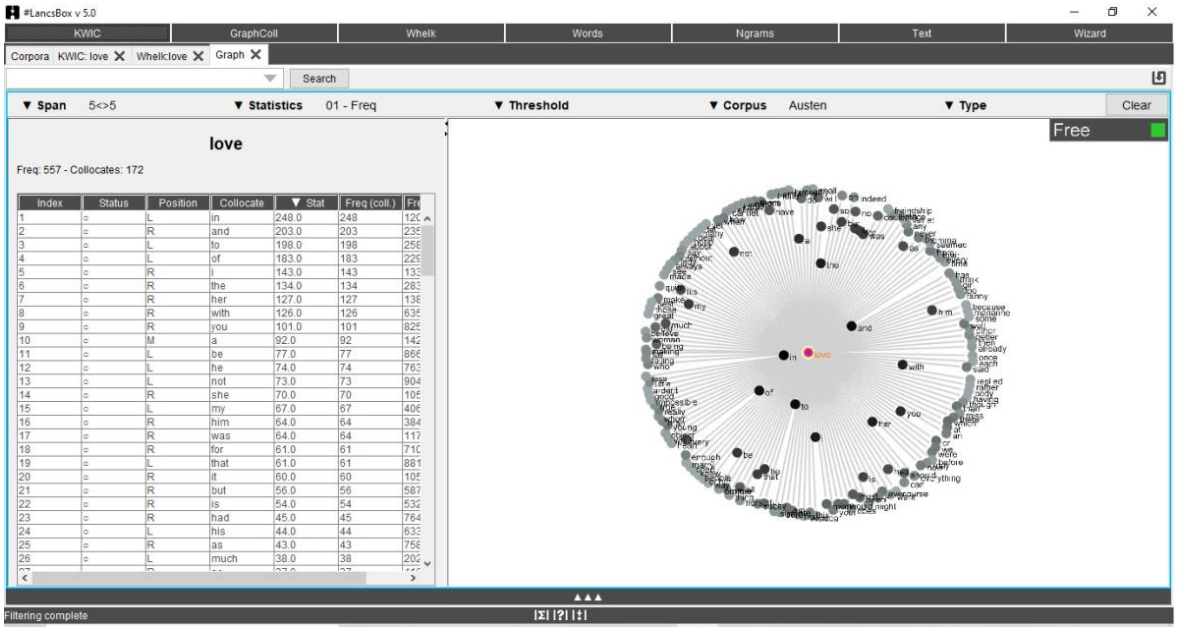


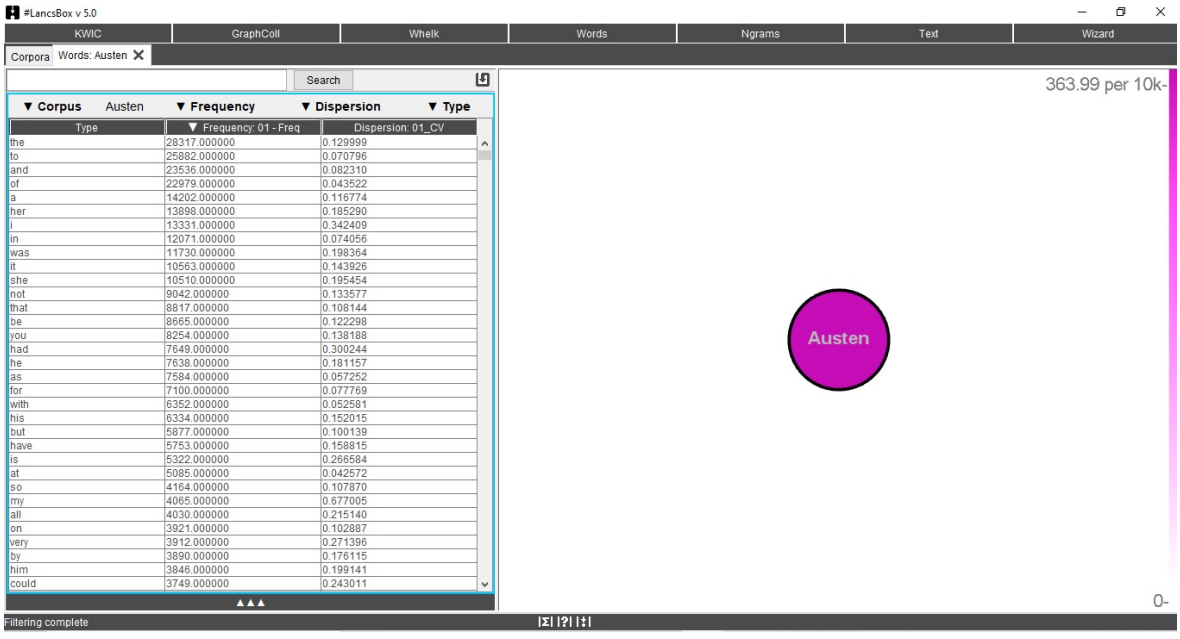
Рис. 20. Инструмент GraphColl

Инструмент Words

Данный инструмент позволяет проводить глубокий анализ частотности форм слов, лемм и частей речи, также как и сравнение корпусов, используя ключевые слова. Он может быть использован для того, чтобы

• подсчитать уровень частотности и дисперсии форм слов, лемм и частей речи;
• визуализировать частоту и дисперсию в корпусах;
• сравнить корпуса, используя ключевые слова;
• визуализировать ключевые слова.

Модуль Words отображает корпуса и документы корпуса (Рис. 21, 22). Он визуализирует частотность и дисперсию слов при помощи усиления цвета и положения отдельных документов, которые показаны в форме окружностей. Размер окружности указывает на относительный размер корпуса/ документа.



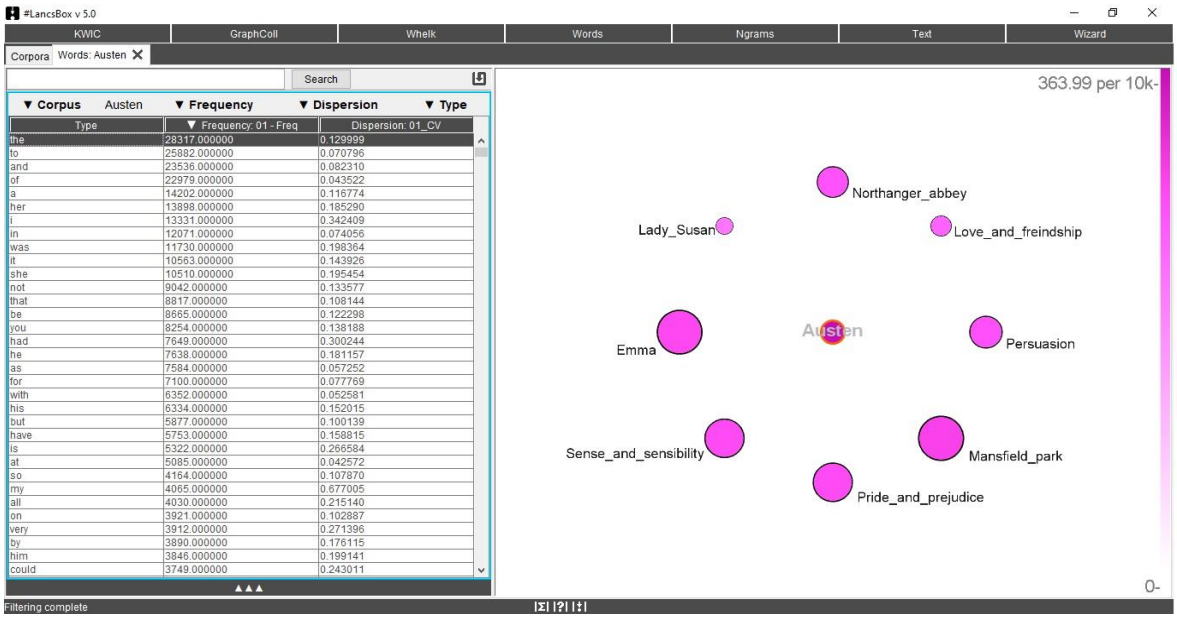


Рис. 21, 22. Инструмент Words

1. Чтобы визуализировать частотность элемента исследования, необходимо щелкнуть левой клавишей мыши по данному элементу в таблице частотности. Оттенок цвета корпуса поменяется в зависимости от значения частоты данного элемента.

2. Чтобы визуализировать дисперсию элемента в таблице, щелкните левой клавишей мыши дважды по корпусу (большая окружность). Корпус растянется, чтобы отобразить отдельные документы (маленькие окружности), из которых и состоит корпус. Размер каждой окружности пропорционален размеру подраздела корпуса. Оттенок цвета маленьких окружностей будет меняться в зависимости от значения частоты элемента поиска в списке частотности. Перечеркнутые окружности указывают на то, что элемент не встречается в данном документе корпуса. В дополнение к этому, документы корпуса располагаются в порядке относительной частоты использования элемента исследования. Документ с наибольшей относительной частотой будет располагаться по часовой стрелке, указывающей на 12 часов, в то время, как другие файлы будут располагаться по часовой стрелке в порядке снижения относительной частоты употребления данного элемента.

Данный модуль может также провести сравнение частотности между двумя корпусами или списками слов, используя специальную статистическую меру. Данный инструмент определяет и визуализирует положительные ключевые слова, отрицательные ключевые слова и стоп-слова.

1. Щелкните левой клавишей мыши по ▲▲▲, чтобы появилась нижняя панель.

2. В нижней панели выберите корпус для сравнения, в то время, как в верхней панели сохраняйте ваш корпус.

3. В панели визуализации (справа), переместите мышью окружности, которые представляют два корпуса вместе. Также, нажмите на клавишу пробела.

4. Получившаяся таблица будет отображать информацию касательно частоты и дисперсии по двум корпусам, также как и статистику по ключевым словам; графическое изображение отобразит 10 самых часто используемых положительных ключевых слов, 10 самых часто используемых отрицательных ключевых слов и 10 стоп слов.

5. В установках вы можете изменить 1) статистику ключевого слова и 2) порог. Статистика ключевого слова: это мера, которая сравнивает два списка частотности [по умолчанию: с константой k = 100]. Порог: значения порога для определения положительных ключевых слов, негативных ключевых слов и стоп-слов.

Модуль Words проводит базовый статистический анализ:

1) дает общую статистику

2) статистику, связанную с лексикой.

Для этого необходимо:

1. щелкнуть правой кнопкой мыши по корпусу (по окружности);

2. в таблице, которая появится на экране, можно переходить от общей статистики к статистике, связанной с лексикой.

Инструмент Ngram

Данный инструмент позволяет проводить глубокий анализ частотности N-грамм (биграмм и т.д.) (Рис. 23), которые могут быть определены как смежные сочетания форм слов, лемм и частей речи. Инструмент также определяет ключевые N-граммы путем сравнения двух корпусов, используя похожий прием с ключевыми словами. Он может быть использован для того, чтобы:

- определить N-граммы, устойчивые словосочетания и P-кадры (также пропуская N-граммы);
- подсчитать частотность и дисперсию N-грамм, форм слов, лемм и частей речи;
- визуализировать частотность и дисперсию N-грамм в корпусах;
- сравнить N-граммы в двух корпусах по ключевому слову;
- визуализировать ключевые N-граммы (Рис. 24).

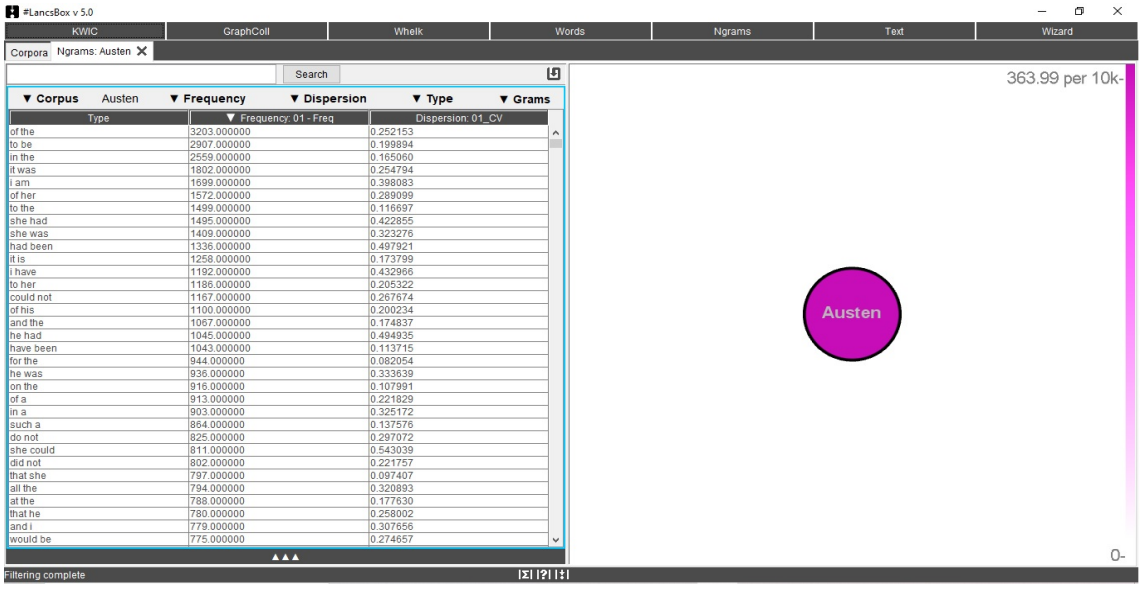


Рис. 23. Поиск n-грамм

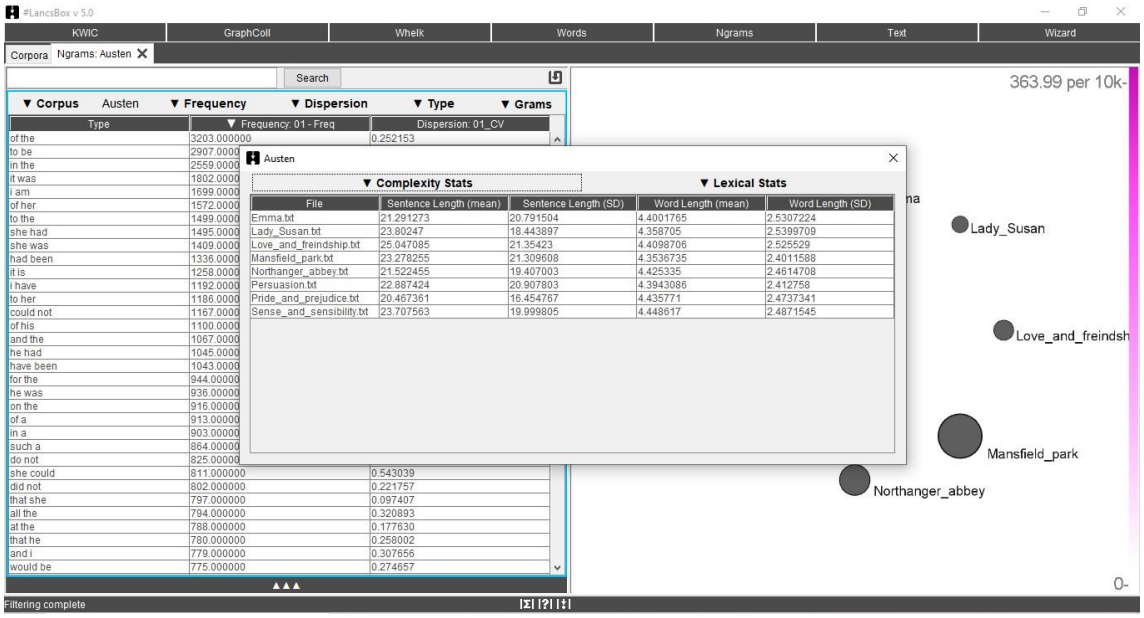


Рис. 24. Определение частотности n-грамм

Инструмент Text

Данный инструмент способствует глубокому пониманию контекста, в котором слово или выражение используются. Он может быть использован:

- для предпросмотра текста;
- предпросмотра корпуса в виде сплошного текста;
- проверки разных уровней аннотаций текста/ корпуса (Рис. 25).

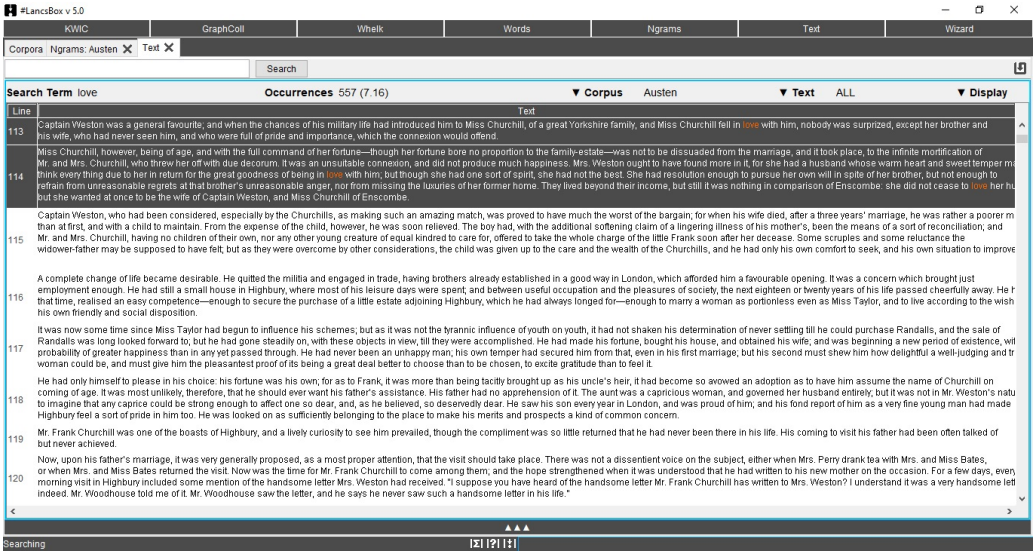


Рис. 25. Осуществление поиска в модуле Text

1. Введите слово поиска в строку поиска (вверху слева). Щелкните левой кнопкой мыши по ‘Search’.

2. Это позволит выделить серым все строки в тексте, в которых появляется слово поиска, которое в свою очередь будет выделено красным цветом. Чтобы передвигаться между выделенными строками вверх и вниз, используйте стрелки ↑ и ↓.

3. Информация о частотности (абсолютной и относительной на 10000 токенов) появится под ‘Occurrences’.

4. Отдельная строка может быть выделена путем нажатия левой клавишей мыши по строке. Чтобы выделить несколько строк, нажмите на Ctrl (Command) и щелкните левой кнопкой мыши по нужным строкам.

5. Выделенные строки могут быть скопированы (Ctrl/Command+C) и вставлены (Ctrl/Command+V) в текстовой редактор.

Инструмент Wizard

Данный модуль сочетает в себе все инструменты #LancsBox, ищет корпуса и генерирует отчет о проведенном исследовании для печати (docx) и в цифровом поле (htlm) (Рис. 26). Он может быть использован:

- для проведения простых и комплексных исследований;
- подготовки чернового варианта отчета;
- загрузки всех необходимых данных.

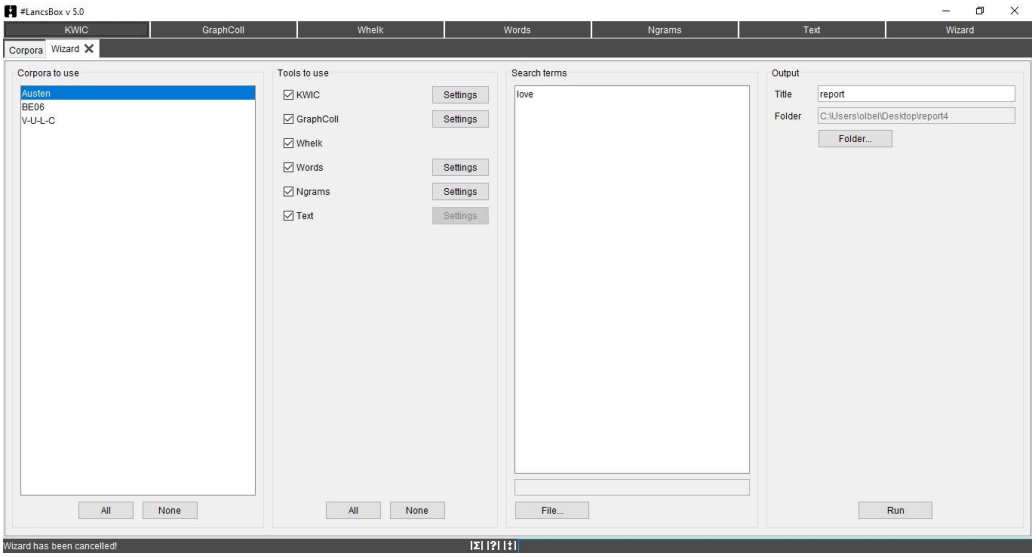


Рис. 26. Инструмент Wizard

🕮 Вопросы и задания

1. Каковы основные характеристики инструмента #LancsBox.
2. Как установить данный инструмент?
3. Каковы функции данного инструмента?
4. Загрузите свой корпус или корпус, размещенный в #LancsBox. Предоставьте информацию, касающуюся автора текстов, языка, количества документов, количества лексем, количества форм слов и количества лемм.
5. Перейдите в модуль KWIC в #LancsBox и найдите следующие слова/ выражения в корпусе VULC (предоставленный #LancsBox). Запишите их частотность и распределения по тексту.

Вид поиска	**Элемент поиска**	**Количество случаев использования на 10K**	**Количество текстов**
Simple	though		
Simple	and		
Phrase	apart from		
Smart search	NOMINALIZATIONS		
Regex	/however|but/		
Regex	state [as headword] V* [as POS]		

6. В этом же инструменте KWIC осуществите поиск следующих выражений и примените фильтры. Запишите их частотность и распределение по текстам.

Элемент поиска	**Фильтр**	**Количество случаев использования на 10K**	**Количество текстов**
VERBS	however [anywhere LEFT]		
might	be [in R1 position]		

7. Перейдите в модуль GraphColl. Постройте графическое изображение коллокации, проведя простой поиск по слову “project”. Что у вас получилось?

Глава II
ИСПОЛЬЗОВАНИЕ КОРПУСОВ ТЕКСТОВ В УЧЕБНЫХ И НАУЧНЫХ ЦЕЛЯХ

Активное внедрение корпусов в процесс обучения иностранным языкам, а также проведение исследований с опорой на корпуса текстов в настоящее время являются неотъемлемыми компонентами интегративного обучения в свете междисциплинарного подхода. Невозможно при изучении иностранного языка и при проведении научных исследований не воспользоваться предложенными данными и разработками, которые позволяют не только отследить современные тенденции в развитии языка, но и проводить глубокий анализ этих изменений.

§1 Национальные корпуса языка

При работе с текстами различных стилей и жанров полезными будут он-лайн ресурсы, на которых размещены национальные корпуса языка.

Первым большим компьютерным корпусом считается Брауновский корпус (**Brown University Standard Corpus of Present-Day American English (Brown Corpus)** [https://www.sketchengine.eu/brown-corpus/], который был создан в 1960 -е годы в Университете Брауна и содержал 500 фрагментов текстов по 2 тысячи слов в каждом, которые были опубликованы на английском языке в США в 1961 году.

Затем по модели близкой к БК в 1970-е годы был создан частотный словарь русского языка под редакцией Л.Н. Засориной [http://project.phil.spbu.ru/lib/data/slovari/zasorina/zasorina.html], построенный на основе корпуса текстов объёмом также в 1 миллион слов и включавший общественно-политические тексты, художественную литературу, научные и научно-популярные тексты из разных областей знания и драматургию.

В 1980-е годы был создан русский корпус при Университете Уппсалы в Швеции. В связи с ростом компьютерных мощностей, способных работать с большими объёмами текстов, в 1980-е годы в мире было предпринято несколько попыток создать корпусы большего размера. В Великобритании такими проектами были Банк Английского (Bank of English) в Бирмингемском университете и Британский национальный корпус (British National Corpus, BNC). В СССР таким проектом был Машинный фонд русского языка, создававшийся по инициативе А. П. Ершова.

Корпус современного американского английского языка (**COCA**), представленный на ресурсе www.english-corpora.org/coca/ – единственный большой, жанрово сбалансированный корпус американского английского языка. COCA, вероятно, является наиболее широко используемым корпусом английского языка, и он связан со многими другими существующими корпусами английского языка.

Корпус содержит более одного миллиарда слов текста (25 + миллионов слов каждый год в течение 1990-2019) из восьми жанров: разговорный, художественный, популярных журналов, газет, академических текстов, и (с обновлением в марте 2020 года): ТВ и фильмы субтитры, блоги и другие веб-страницы.

Данная программа позволяет проанализировать все тексты (например, студенческие сочинения или статью из интернет-газеты), а затем просмотреть подробную информацию о словах и фразах в тексте на основе данных из COCA. Программа позволяет получить обширную информацию о слове при нажатии на него -определение, перевод, этимология, произношение, изображения, видео, связанные слова, разговоры, темы, кластеры, линии соответствия и многое другое. Нажав на фразу в тексте, можно найти связанные фразы в COCA, что позволяет найти "только правильную/верную фразу" для данного жанра.

ACE - Asian Corpus of English (https://corpus.eduhk.hk/ace/), **Australian Corpus of English** (http://korpus.uib.no/icame/manuals/ACE/INDEX.HTM) – корпуса, которые позволяют осуществлять исследования на материале австралийского, британского, американского вариантов английского языка.

Национальный корпус русского языка (НКРЯ) представлен на ресурсе www.ruscorpora.ru. На этом сайте помещен корпус современного русского языка общим объёмом более 600 млн слов. Корпус русского языка – это информационно-справочная система, основанная на собрании русских текстов в электронной форме.

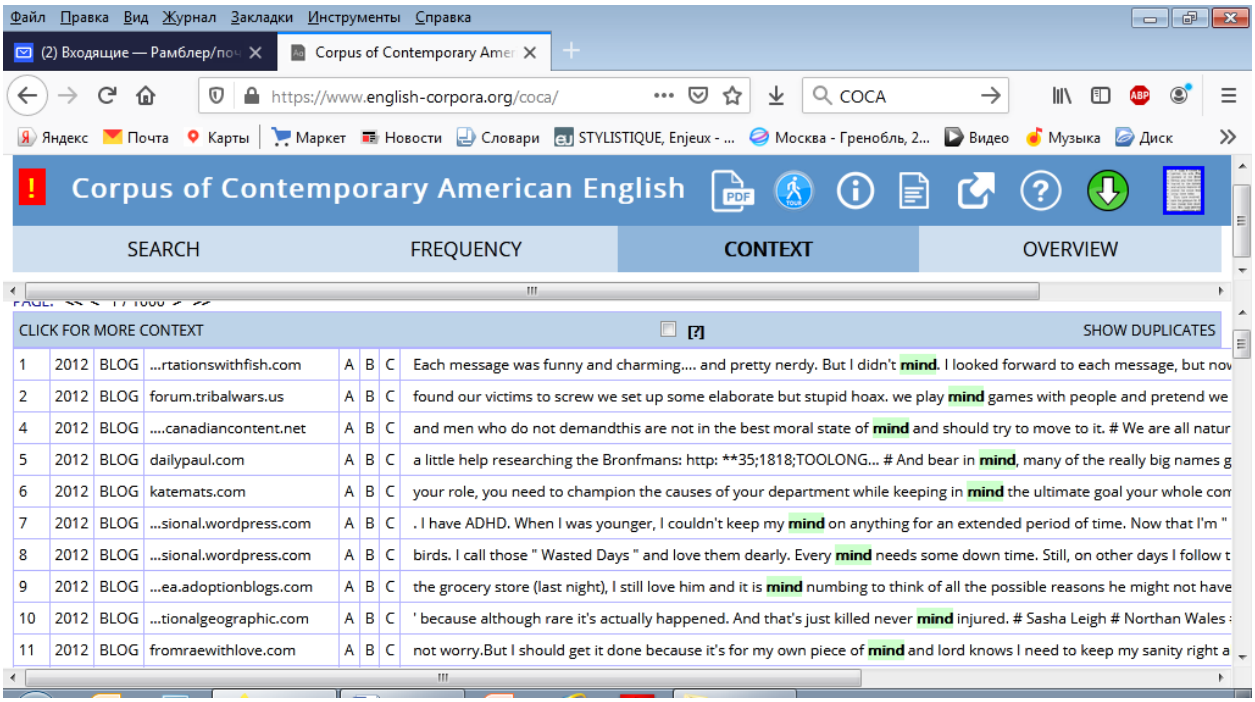
Корпус предназначен для всех, кто интересуется самыми разными вопросами, связанными с русским языком: профессиональных лингвистов, преподавателей языка, школьников и студентов, иностранцев, изучающих русский язык.

Развитие подкорпусов НКРЯ (основного, поэтического, параллельного, акцентологического, диалектного) в 2015 году осуществлялось при поддержке РГНФ, проекты № 15-04-12018 «Развитие специализированных модулей НКРЯ» и № 14-04-12012 «Корпус диалектных текстов Национального корпуса русского языка. Пополнение и разметка».

Наряду с представительными корпусами, которые охватывают большой набор жанров и функциональных стилей, в лингвистических исследованиях часто используются и оппортунистические коллекции текстов, например, газеты (часто The Wall Street Journal и The New York Times), новостные ленты (Рейтер), коллекции художественной литературы (**Библиотека Максима Мошкова** [http://lib.ru/] или **Проект «Гутенберг»** [https://web.eecs.umich.edu/~lahiri/gutenberg_dataset.html], электронная библиотека **Royallib** [https://royallib.com/], **Ebooks** [https://www.ebooksgratuits.com/] и многие другие).

🕮 Вопросы и задания

1. В процессе работы с корпусом COCA отследите употребление лексемы happy. Какова его частотность в текстах корпуса?
2. Проанализируйте сочетаемость лексемы mind в текстах корпуса COCA:



3. Определите контекст употребления данной лексемы и ее частотность в текстах.
4. Отследите контекст сочетания make laugh в корпусе COCA. Проанализируйте источники употребления данного сочетания. Как это может быть связано с низкой частотностью его употребления?
5. В Национальном корпусе русского языка найдите и проанализируйте употребление лексемы «мухобойка». Какова ее сочетаемость с другими частями речи, какова частотность употребления и характер текстов?
6. Может ли НКРЯ осуществлять поиск пословиц или поговорок?
7. Проверьте, встретится ли слово «синхронизация» в поэтическом корпусе НКРЯ.

§2 Обзор корпусов текстов, аудио- и видеоматериалов для учебных целей

Существуют различные типы корпусов, ориентированных, например, на применение их данных в процессе обучения FLE.

Лаборатория UNIL (Université de Lausanne) в лице Кристиана Сюркуфа (Christian SURCOUF) и Алена Озони (Alain AUSONI) при поддержке Фонда педагогических инноваций Университета в Лозанне (Fonds d'innovation pédagogique de l'Université de Lausanne, Suisse) предлагает разработку корпуса французского языка для изучения французского как иностранного **FLORALE** (https://florale.unil.ch/).

Корпус обеспечивает работу с произносительными навыками, а также лексико-грамматическими особенностями французского языка. В частности, в разделе Произношение (Prononciation) предлагаются подразделы с возможностью прослушивания сложных моментов разговорной речи, а именно сокращение местоимения tu или je (faut que je te [ʒtə] présente), особое произношение предлога de в потоке речи (Tout le monde était de super [tsypɛr] bonne humeur), непроизносимые согласные на конце слов (enfin bon c'était raisonnable [rɛzonab] quand même), особенности произношения некоторых сочетаний в потоке речи (en quinze jours j'avais déjà [dʒa] tout trouvé), некоторые правила связывания и особенности интонационного членения фразы. FLORALE предлагает более 5 ч. аудиотекстов для занятий FLE.

Также корпус рассматривает некоторые специфические случаи построения фразы (конструкции с отрицанием, употребление местоимений, построение вопросительного предложения, выражение степени и количества и т.д.), лексические особенности построения речи (использование вспомогательных слов для построения разговорной речи alors, du coup, bon, bref, donc etc.) , а также устойчивые слова и выражения для разговорной речи (ce qui fait que, sur le coup, se mettre en place etc.). Кроме того, корпус предлагает небольшую подборку наиболее употребительных аббревиатур.

Alignoscope (http://www.miaojun.net/alignoscope-intro/) – это текстометрический инструмент, предназначенный для изучения оригинальных и переведенных текстов на французском и китайском языках. Разработчиками данного инструмента являются Джун Миао *(*Jun MIAO*)* и Ким Гердес, преподаватель университета Paris III. Этот инструмент графически представляет распределение вхождений определенных элементов по всему тексту в обоих языках. Функции поиска предоставляют прямой доступ к отдельным абзацам в двух языках, отображающимся одновременно, что позволяет отследить особенности перевода тех или иных конструкций. В настоящее время инструмент работает с текстом романа Р. Роллана «Жан-Кристоф» (первый том оригинального текста с тремя различными китайскими переводами и вся книга с переводом Фу Лея).

Платформа **SimpleApprenant** (https://simpleapprenant.huma-num.fr/SimplifyYourFrench/accueil является результатом работы над проектом Idex SimpleApprenant (LiLPa, Университет Страсбурга). Студентам, изучающим французский, как иностранный, предлагается платформа, генерирующая упражнения, предназначенные для изучения идиоматических выражений, которые имеют переносное значение и специфические морфосинтаксические свойства и представляют трудности для изучающих FLE. В данной платформе выражения классифицируются по уровню CEFR (A1-C2) и являются адаптированными к уровню знаний учащегося.

Платформа предлагает тренировочные упражнения на заполнение пропусков, заучивание устойчивых сочетаний, на генерирование собственного текста с использованием данных выражений. Выражения выбираются в зависимости от уровня пользователя (Рис.27). Учащийся может перейти к письменным заданиям, что позволяет использовать в контексте выражения, произвольно извлеченные из списка, управляемого платформой. Еще одна предлагаемая функция, – это возможность проверять тексты, написанные на французском языке, и автоматически выявлять любые орфографические, лексические, синтаксические ошибки. Кроме того, возможен поиск необходимых пользователю выражений или отдельных слов.

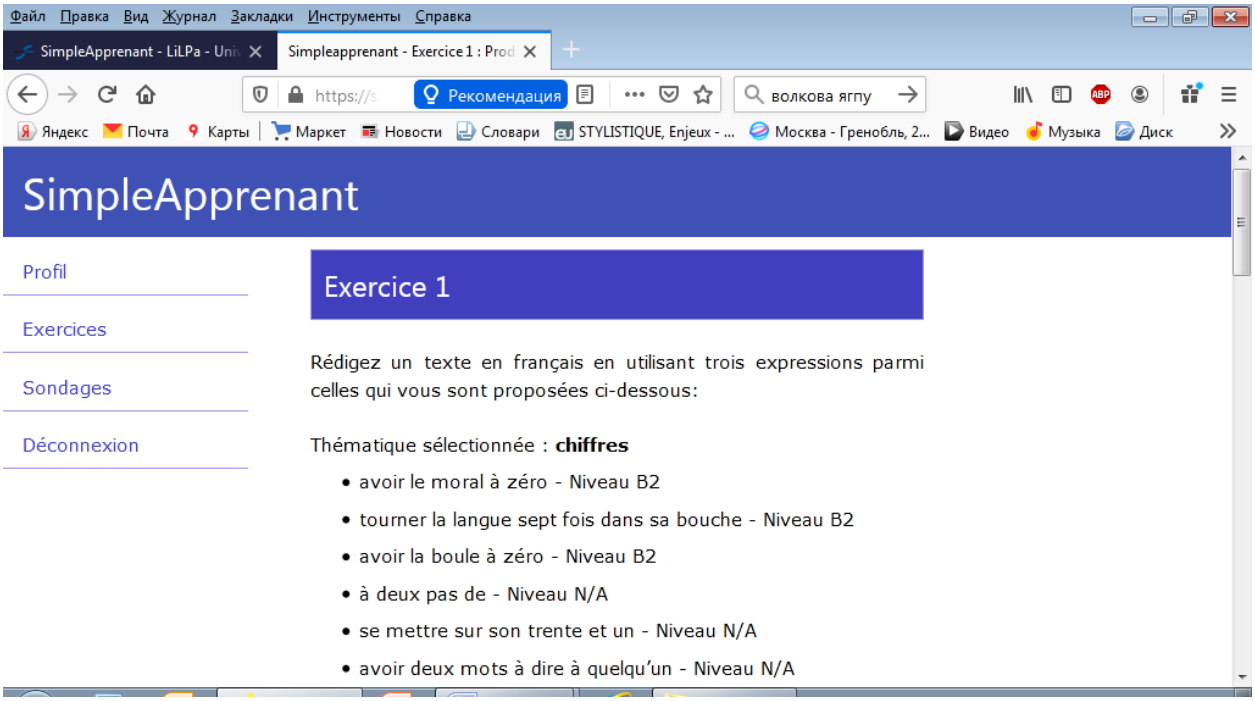


Рис. 27. SimpleApprenant

В рамках работы лаборатории Laboratoire Ligérien de Linguistique университета в Орлеане (Université d'Orléan) создан корпус аудио текстов corpus **ESLO** (http://eslo.huma-num.fr/) по различным тематикам, включающий два основных раздела: ESLO1 и ESLO2. В рамках первого предлагаются аудиотексты по следующим темам:

- Interviews sur questionnaire
- Opération sur le Vif- Contacts
- Opération sur le Vif- témoins en situation sociales ou professionnelles
- Communications téléphoniques
- Interviews avec des personnalités
- Conférences-débats
- Enregistrements divers
- Consultation CMPP

Второй предлагает следующую тематику:

- Entretiens
- Itinéraires
- Entretiens jeunes
- Cinéma
- Discours
- Repas
- Interviews Personnalités
- Ecole
- Entretiens Chercheurs
- Médias
- Boulangeries
- Livre pour enfants
- 24h
- AG
- Marché
- Commerces
- Guichets
- Soirées

Корпус содержит огромное количество транскрибированных диалогов по перечисленным выше темам. Перемещающееся выделение произносимой реплики позволяет отслеживать произносимые фразы и работать с ними. При этом диалоги записаны носителями в реальных условиях и представляют собой живую речь франкофонов (Рис.28).

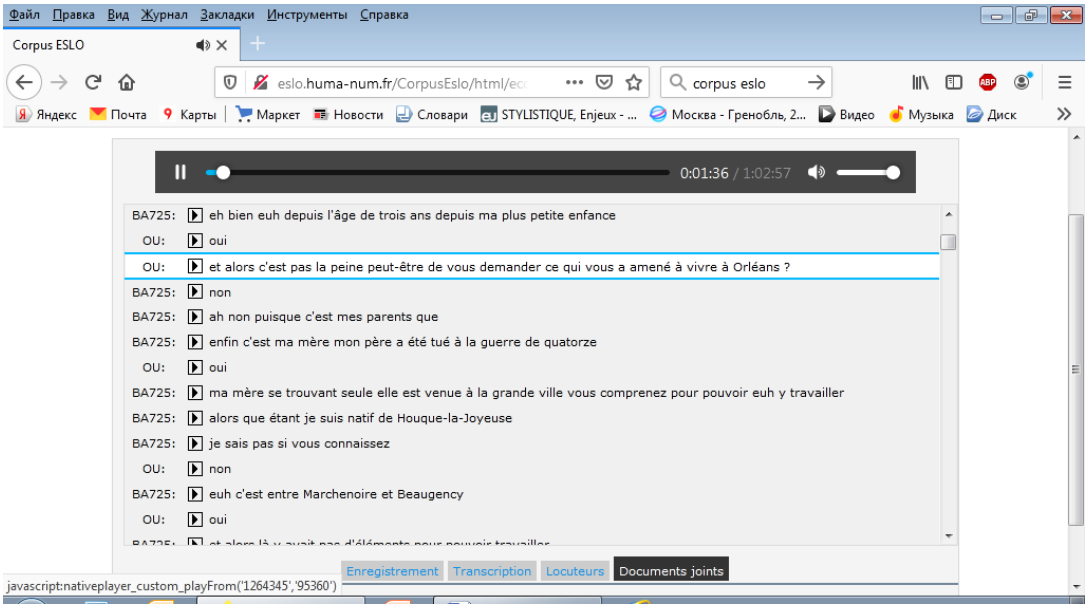


Рис. 28. ESLO

Корпус **Fleuron** (https://fleuron.atilf.fr/) предлагает аутентичные мультимедийные ресурсы, которые иллюстрируют ряд ситуаций из жизни студента во Франции. Ресурсы предлагают материалы по следующим микротемам: регистрация студента в административной службе, получение студенческого билета и других официальных документов, получение информации об административном и образовательном положении студента, общение со студентами на различные темы, обсуждение с преподавателем программы обучения.

Каждый аудиофайл сопровождается субтитрами, также предлагается транскрипция и сформирован необходимый для данной тематики глоссарий (Рис. 29).

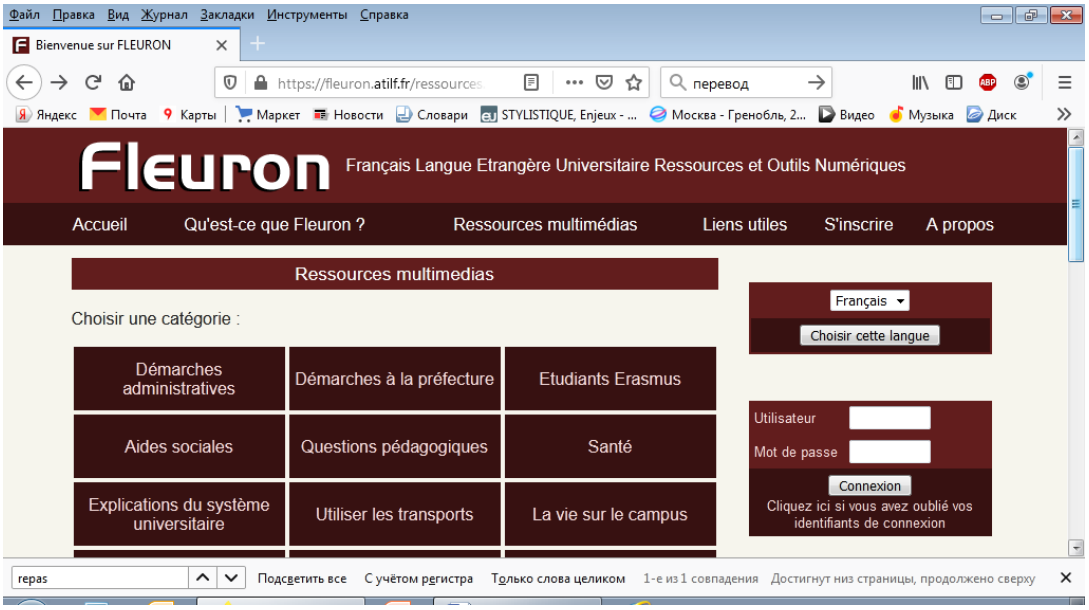


Рис. 29. Fleuron

Сайт национального центра текстовых ресурсов **CNRTL (Centre National des Ressources Textuelles et Lexicales)** (https://cnrtl.fr/) является платформой, принимающей несколько корпусов, работающих с лексикой и текстами в различных форматах. На сайте можно найти:

Frantext, представляющий собой корпус французских текстов различных авторов,
Corpus journalistique de l'Est Républicain, являющийся корпусом публицистических текстов,
Traitement de Corpus Oraux en Français (TCOF) - корпус аудиотекстов,
Корпус лингвистических публикаций из журнала "Sciences Humaines"
Корпус аннотированных текстов DEDE.

Кроме того, в разделе «Лексика» предлагается поиск значений слов в различных словарях, в том числе этимологическом (Рис. 30), словарях синонимов и антонимов. В разделе Словари предлагаются издания Le dictionnaire de l'Académie française, Le dictionnaire Oeconomique de Chomel, Le Dictionnaire du Moyen Français, Le Dictionnaire Électronique de Chrétien de Troyes, La troisième édition (1552) du Dictionarium latinogallicum de Robert Estienne и многие другие.

Рис. 30. CNRTL ORTOLANG (Open Resources and TOols for LANGuage) – www.ortolang.fr, https://hdl.handle.net/11403/democrat/v1.1

Корпус **CLAPI-FLE** (http://clapi.icar.cnrs.fr/FLE/) предлагает бесплатный доступ к видео и аудио ресурсам для аутентичного общения на французском языке. Предлагаемая тематика: профессиональная деятельность, рабочие встречи, обеды с друзьями или семьей, покупки или продажи в магазинах, медицинские консультации, игры, приглашения, экскурсии, телефонные звонки и многое другое. Проект ориентирован на применение в процессе преподавания FLE и был разработан командой LIS лаборатории ICAR в сотрудничестве с учителями FLE и лингвистами-переводчиками, на основе базы данных CLAPI.

В настоящее время CLAPI-FLE включает в себя 40 транскрибированных аудиофайлов, выстроенных в соответствии с различными уровнями сложности, комментарии к сложным элементам синтаксиса и многое другое (Рис. 31).

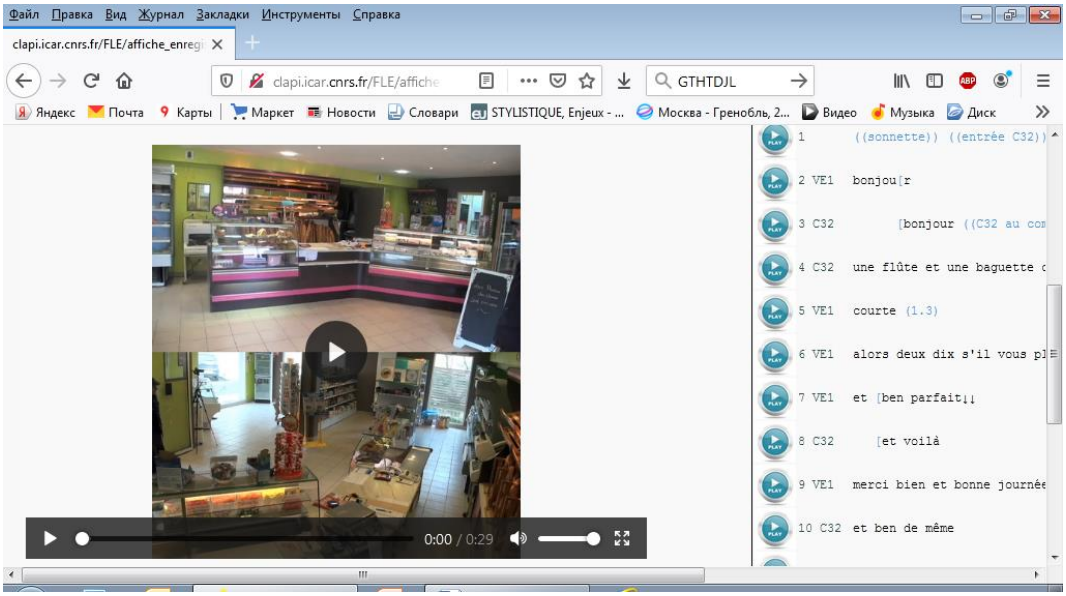


Рис. 31. CLAPI-FLE

Корпус **PFC-EF** (https://www.projet-pfc.net/le-projet-pfc-ef/ressources-linguistiques/corpus-thematique/) также предлагает множество аутентичных аудиотекстов по различным темам, представленных франкофонами из разных уголков мира (Рис. 32).

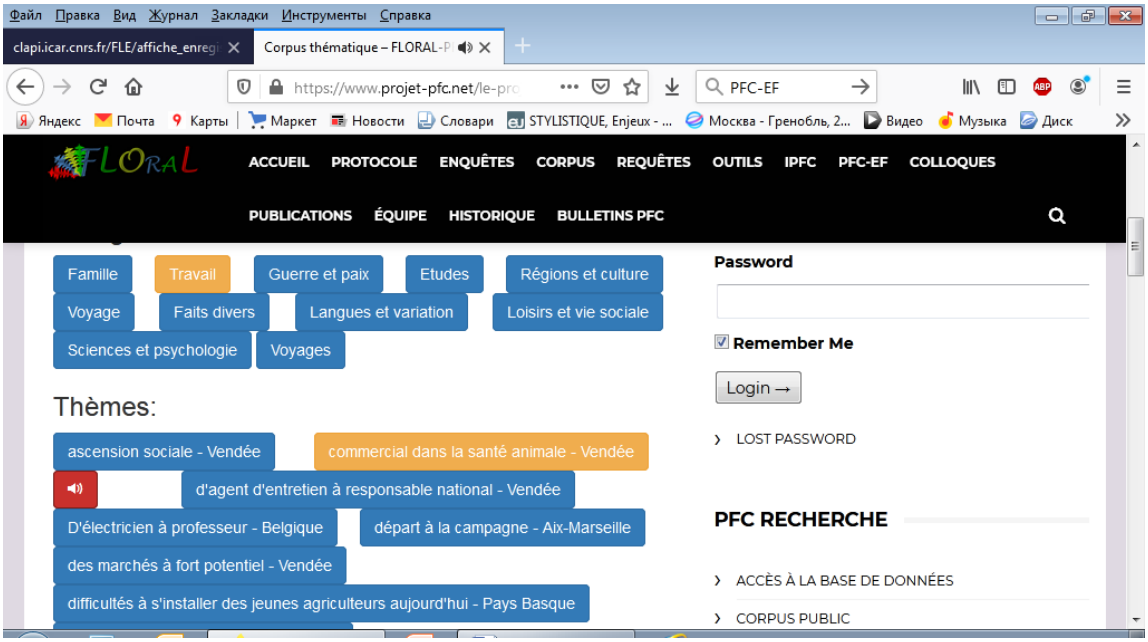


Рис. 32. PFC-EF

Корпус **RésolCo** (http://redac.univ-tlse2.fr/corpus/resolco.html) составляют тексты работ школьников и студентов с сохраненной орфографией (394 текста). Корпус составлен с целью осуществления анализа основных ошибок при изучении французского языка на всех уровнях его усвоения, а также с целью работы над повышением уровня грамотности носителей языка.

MANULEX - вебсайт, направленный на обучение детей чтению на французском языке. Manulex включает корпус из 1,9 миллиона слов, используемых во французских начальных школах с первых по пятые классы. Читатели охватывают целый ряд тематических областей, каждая из которых содержит значительный объем данных, поступающих из разных типов текстов (от романов до различных художественных произведений, от газетных репортажей до технических текстов и от поэзии до театральных пьес), написанных разными авторами из разных стран.

База данных содержит два словаря: словарь словоформ (48886 статей) и словарь лемм (23812 статей). Каждый содержит список слов в зависимости от уровня обучения, которые встречаются у читателей первого, второго и третьего-пятого классов (далее называемые уровнями G1, G2, G3-5 соответственно). Четвертый уровень (G1-5) был создан путем объединения текстов всех читателей.

🕮 Вопросы и задания

1. Какие корпуса и инструменты из перечисленных выше можно использовать для преподавания FLE?
2. Какие инструменты работают с сопоставлением особенностей перевода?
3. Выберите несколько упражнений на платформе SimpleApprenant и выполните их.
4. На сайте FLEURON выберите корпус для работы и проанализируйте выбранный вами текстовый материал на предмет частотности слов.
5. На платформе FLORAL выберите категорию Travail, тему Le métier d'enseignant и составьте алгоритм работы с диалогом для учащихся факультета иностранных языков университета.
6. На сайте CLAPI-FLE просмотрите видео, по тематике «Прием гостей» L'accueil et l'installation des invités. Воспроизведите диалог с соблюдением интонации : http://clapi.icar.cnrs.fr/FLE/liste_extraits.php.
7. На сайте CNRTL выберите вкладку dictionnaires, далее Dictionnaire de l'Académie française и найдите значение интересующего вас слова.

§3 Обзор корпусов текстов для научных целей

Проект **DEMOCRAT** (http://www.lattice.cnrs.fr/democrat/) объединяет исследователей из нескольких французских лабораторий, в частности, Lattice (Париж), LiLPa (Страсбург), ICAR и IHRIM (Лион). Это проект, который направлен на развитие исследования языка и текстового структурирования французского языка в корпусе текстов, написанных между 9-м и 21-м веками, а также нацелен на работу с различными текстовыми жанрами. Акроним DEMOCRAT означает: DEscription and Modeling of Reference Chains: инструменты для аннотации корпуса (в диахронии и синхронии) и автоматической обработки.

Корпус DEMOCRAT – это текстовый корпус, аннотированный ссылками. Ссылочные выражения идентифицируются и аннотируются идентификатором референта, что позволяет создавать ссылочные строки. Состав корпуса устанавливается с целью изучения вариаций эталонных цепочек в соответствии с дискурсивными жанрами и эпохами. Композиция устанавливается по трем критериям: эпоха, тип, жанр текста. Размер корпуса позволяет использовать приложения для автоматической языковой обработки.

FRANTEXT (https://www.frantext.fr/) – это база данных, содержащая 5340 ссылок (256 млн слов). Она позволяет осуществлять анализ форм, лемм и грамматических категорий. Корпус текстов довольно обширный (с IX до XXI вв.). В разделе «Корпуса текстов» (Corpus) вы можете загружать, создавать, визуализировать и составлять ваши собственные корпуса для работы согласно определенным параметрам, например, классифицируя по авторам, дате издания, литературному жанру.

В разделе «Исследования» (Recherche) вы можете осуществлять анализ согласования в предложении, частоты, сочетаемости лексем. Простой поиск позволит найти необходимые слова или сочетания слов в тексте, развернутый поиск позволит сочетать некоторые параметры при поиске (форма+лемма). При этом возможен количественный подсчет встречаемости слов, визуализация повторяющихся слов и их контекста слева и справа. Кроме того, в разделе «Списки слов» Listes des mots можно создавать свои собственные списки, необходимые для исследования.

В разделе «Грамматика» Grammaire содержатся уже сформулированные грамматические правила, однако платформа дает возможность пользователю самому сформулировать правила для своего исследования, включая правила сочетаемости слов.

GLOZZ (http://www.glozz.org/) – это платформа, предлагающая осуществлять ручное аннотирование, на основе которого можно проводить анализ дискурсивной структуры текста и сочетаемости элементов в предложении. Результаты анализа представляются в виде графиков и схем. Платформа разработана международной командой в составе Yann Mathet and Antoine Widlöcher from the Greyc.

DECLICS (Dispositif d'Etude Cliniques sur les Corpus Santé https://acte.uca.fr/declics-dispositifs-d-etudes-cliniques-sur-les-corpus-sante-71799.kjsp) является результатом работы четырех исследовательских лабораторий Оверни, в области социальных наук и медицины. Проект ориентирован на осуществление анализа устных консультаций во время приема врача с целью улучшения результатов врачебной практики, совершенствования общения между врачом и пациентом. Корпус текстов позволяет анализировать аудио данные, в частности синтаксис, прагматику, семантику речи специалиста, чтобы сделать результаты данного анализа доступными для других дисциплин (философии, нейропсихологии, психиатрии).

LEXICOSCOPE (http://phraseotext.univ-grenoble-alpes.fr/lexicoscope/index.php?errorAccess) – инструмент для изучения сочетаемости слов в тексте, например, в составе фразеологизмов. Этот инструмент позволяет исследовать особенности сочетаемости лексем в зависимости от жанра и эпохи произведения. В частности на данный момент Lexicoscope работает с текстами средневековых романов.

🕮 Вопросы и задания

1. Проанализируйте сочетаемость лексем с глаголом mettre в текстах корпуса Frantext.
2. Найдите значения полученных сочетаний.
3. Определите наиболее частотную лексику пациента на приеме врача в выбранном вами диалоге при помощи аудиотекстов на платформе DECLICS.

4. Что такое ручное аннотирование текстов?
5. Какое исследование можно проводить на основе данных аннотированного текста?
6. На сайте DECLICS отыщите последние публикации в области медицины.
7. Каковы основные направления работы лаборатории ACTé, представленной на сайте DECLICS?

Глава III
ИНСТРУМЕНТ ДЛЯ АНАЛИЗА РИТМА ПРОЗЫ PROSE RHYTHM DETECTOR (PRD)

§1 Предварительная обработка текста

Задачи обработки текста на естественном языке являются очень сложными. Методы их решения разнообразны. Однако есть этапы работы, которые во многом одинаковы для большинства задач. Таким этапом является предварительная обработка текста, состоящая из преобразования в формат, удобный для обработки, деление текста на отдельные части (слова и предложения) и маркировка, т. е. вычисление некоторых признаков отдельных частей на основе морфологического и синтаксического анализа. В этом разделе рассматриваются этапы предварительной обработки текста и существующие программные инструменты для этого.

Преобразование формата

Применение программных алгоритмов всегда подразумевает определённую структуру входных данных. Тексты на естественном языке редко имеют одинаковое устройство, даже если относятся к одному стилю: статьи, художественные произведения, электронные письма, короткие сообщения в социальных сетях.

Поэтому самым первым шагом является приведение всех текстов, отобранных для решения некоторой задачи к общему формату, который удобен для автоматической обработки. Следует отметить, что этот формат очень сильно зависит от поставленной задачи и особенностей входных данных. Из-за этого алгоритм преобразования практически всегда будет уникальным. Однако можно выделить некоторый набор подходов, на основании которого строятся такие алгоритмы.

В первую очередь имеет значение электронный формат используемых текстов. Так называемый «сырой» текст – это формат txt. Тексты либо находятся в этом формате, либо преобразуются к нему из других, таких как pdf или doc. Если используемый набор данных получен как база данных или корпус текстов, то это означает, что он уже предварительно преобразован к некоторому формату, в котором надо разобраться.

Вторым важным моментом является удаление шума, удаление символов и частей текста, которые могут помешать дальнейшему анализу. При этом очень важен учёт особенностей предметной области, в которой поставлена задача. Рассмотрим некоторые шаги форматирования текста в этом ключе.

Сначала обсудим удаление из текста символов определённого набора. Это могут быть непечатные символы, попавшие туда при оцифровке, разрывы строк или страниц, табуляция. В набор для удаления могут включаться символы пунктуации. Однако часто они необходимы для дальнейшего выделения предложений и фраз. При анализе художественных произведений может встречаться достаточно сложная пунктуация, например, при передаче прямой речи. Поэтому для разных задач могут быть лишними разные наборы символов пунктуации.

Числа и цифры, встречающиеся в тексте, могут нести информацию или, наоборот, быть шумом. Например, шумом часто являются номера страниц или разделов. Удаление чисел в этом случае зависит от контекста, в котором они встречаются. Может быть ситуация с текстами книг и статей, когда число на отдельной строке между абзацами является номером страницы и подлежит удалению.

Очень часто приходится удалять элементы стороннего форматирования, например теги html, и специальные наборы символов, например, *RT в ретвитах.* Может удаляться заголовок и предисловие или оглавление и приложения, если речь идёт о книгах или статьях.

В ряде задач качество обработки зависит от учёта или не учёта регистра букв. Для поиска информация о стране «Индия» и «индия» скорее всего равноценна. В тексте о языке программирования Java слова «System» и «system» будут существенно различаться, так как первое является часто используемым классом языка, а второе имеет отношение к общей терминологии программирования. В настоящий момент времени параметры учёта регистра часто используются при настройке

стандартных систем автоматической обработки текста. Эти системы, например, умеют распознавать именованные сущности с определённой степенью точности.

В любых ситуациях необходимо точно определить, какую информацию несут отбрасываемые символы, чтобы не потерять нужную информацию. Возможно какие-то символы или группы символов следует не удалить, а заменить на другие.

Таким образом, задача преобразования текстов к определённому формату для каждого конкретного случая решается по-своему. Обычно это осуществляется созданием небольших программных утилит или подбором подходящих среди существующих. Для того, чтобы лучше понимать программные технологии обработки текста, рассмотрим в следующем разделе механизм регулярных выражений, которые являются одной из основ этих технологий.

Регулярные выражения

Регулярное выражение (Regular expressions) – это шаблон последовательности символов, используемый для поиска и замены текста в строке. Одному шаблону может соответствовать много разных строк. Регулярные выражения используют два типа символов: обычные символы и специальные метасимволы для обозначения свойств строки.

В таблице приведено описание некоторых метасимволов и примеров регулярных выражений.

Таблица 1

Примеры регулярных выражений

Метасимвол	Описание	Пример использования	Пример подходящих строк
.	Один любой символ, кроме символа перехода на новую строку (\n).	к.р.ва	корова карова кор7ва
\d	Любая цифра	ИВТ\d\d	ИВТ12 ИВТ42 ИВТ00
\D	Любой символ, кроме цифры	12\D12	12a12 12+12 12:12
\s	Любой пробельный символ (пробел, табуляция, конец строки и т.п.)	кор\sова	кор ова кор ова
\S	Любой непробельный символ	да\Sда	да+да да-да даДда
\w	Любой символ, являющийся частью слова: буква, цифра и подчёркивание (_)	\w\w\w	Ааа а_а 1a_
\W	Любой символ, не являющийся частью слова	кто\W	Кто% Кто? кто№
[ ]	Один из символов в скобках, в том числе символ из диапазона, например a-z – все маленькие латинские буквы	[A-Z][0-9a-z]	Ff N2 Bp
[^]	Любой символ, кроме перечисленных в скобках	[^0-9]	A % #
{n}	Ровно n повторений предыдущего символа или выражения	[0-9]{4}	3752 4040 9999
{m,n}	От m до n повторений включительно предыдущего символа или выражения	\d{2,4}	12 123 9877
?	Ноль или одно вхождение предыдущего символа или выражения	абв?	абв аб
*	Ноль или более повторений предыдущего символа или выражения	ИВТ\d*	ИВТ ИВТ454 ИВТ0

Метасимвол	Описание	Пример использования	Пример подходящих строк
+	Одно или более повторений предыдущего символа или выражения	ИВТ\d+	ИВТ454 ИВТ0 ИВТ2222222222
( )	Группа символов	(123)+	123 123123123123 123123
\	Экранирование метасимвола для обозначения обычного символа	123\)+	123) 123))))))) 123))

Узнать больше о регулярных выражениях можно в статьях [Уроки по регулярным выражениям] и [Васильев, 2019].

Выделение единиц текста и их параметров

Автоматическая обработка текста базируется на представлении конкретного текста на естественном языке в виде математической модели. Модель обычно включает разбиение текста на части (единицы), определение параметров этих частей и связей между ними.

Наиболее важным этапом предварительной обработки является выделение единиц текста, часто это называют *токенизацией* (от английского tokenization), и определение их морфологических и синтаксических характеристик, т. е. маркировка.

Основная единица, с которой происходит работа, – это слово. В алгоритмах автоматической обработки текста используется более формальный термин «токен». Токеном может быть как привычное слово, так и, например, число цифрами. Токен обычно отделен от других токенов пробелами или знаками препинания. Знаки препинания могут рассматриваться как отдельные токены.

Следующие единицы – предложение и абзац. Предложения обычно отделяются определёнными знаками препинания, абзацы – знаками перехода на новую строку. Эти единицы связаны в первую очередь с контекстом определённого токена. В большинстве задач таким контекстом, который рассматривается при работе с моделью текста, выступает предложение, включающее исходный токен. Однако, для ряда случаев модель должна учитывать более широкий контекст, выходящий за рамки одного предложения. Это важно для определения ритмических характеристик текста. Ещё одна такая задача – построение диалоговых систем, где важно учитывать предыдущие вопросы и ответы.

Следующее действие – вычисление признаков каждого токена или маркировка. Признаки включают в себя часть речи или POS-тег (part of speech), начальная словоформа, позиция в тексте, роль в предложении. Для языков со сложной морфологией, например, русского, также важны морфологические признаки: падеж существительного, род прилагательного. Так как одно и то же слово может быть в разных формах, например, в тех же падежах важно распознавать разные слова и разные формы слова. Для этого есть алгоритмы лемматизации и стемминга. Приведение слов к начальным формам называется лемматизацией. Стемминг – это выделение основы слова, чаще всего корня.

Стемминг более грубый эвристический процесс, который отрезает приставки, суффиксы и окончания от корня слова, часто это приводит к потере словообразовательных элементов. Обычно в алгоритмах стемминга никак не учитывается контекст использования слова в тексте. Однако у стеммеров есть и свои преимущества: их проще внедрить и они работают быстрее. *Лемматизация* – это более сложный процесс, который использует словарь и морфологический анализ, чтобы в итоге привести слово к его канонической форме – лемме.

Синтаксический анализ – это процедура нахождения вариантов разбора фраз и предложений, соответствующих формальным грамматическим правилам. Вариант разбора представляет информацию о роли токена в предложении чаще всего в виде иерархической структуры (дерева). Это наиболее сложный этап обработки текста, но без него невозможно осуществлять извлечение фактов, автоматический перевод, аннотирование, построение вопросно-ответных систем. Синтаксический анализ основывается на формальных грамматиках, описывающих конструкции синтаксически верных предложений, т. е. как фраза или предложение может быть представлена в виде структуры более маленьких частей.

Последовательность разделения текста на отдельные компоненты и получение признаков текста и его отдельных сегментов называется *пайплайном NLP*. Рассмотрим четыре библиотеки обработки текста на языке Python, которые используются для выделения токенов и предложений и определения признаков токенов.

Stanza

Первый пакет анализа естественного языка, который мы подробно рассмотрим – это Stanza [Stanza – A Python NLP Package for Many Human Languages]. Он содержит инструменты, которые можно использовать для преобразования строки, содержащей текст на человеческом языке, в списки предложений и слов, для определения базовых форм этих слов, их частей речи и морфологических признаков. Эти функции работают для 66 языков.

Установка Stanza описана в руководстве [Installation & Getting Started// Stanza]. Начать работу с этой библиотекой можно со следующих команд для интерпретатора Python:

```
import stanza # подключение библиотеки
config = {
   'processors': 'tokenize,mwt,pos', # процессоры обработки языка
   'lang': 'en', # язык текста
}
nlp = stanza.Pipeline(**config) # инициализация процесса обработки текста
doc = nlp("Text. Sentence. Token.") # обработка текста
print(doc) # вывод результата работы предыдущей строки
```

Алгоритм работы заключается в следующем: необходимо задать конфигурацию процесса обработки текста, т. е. пайплайна. Здесь можно настраивать несколько параметров, перечисляя их набор в виде отдельной переменной. Инициализировать процесс методом Pipeline с указанием конфигурации (nlp = stanza.Pipeline('en')). Получить модель текста как результат выполнения заданного процесса (doc = nlp("Text. Sentence. Token.")). В таблице приведены основные настройки процесса обработки языка.

Таблица 2

Настройки процесса обработки языка библиотеки Stanza

Параметр конфигурации	Описание параметра	Значение параметра	Пояснения
processors	процессоры, которые используются в процессе обработки текста	tokenize	разделяет документ на предложения, каждое из которых содержит список токенов (Token); предложения доступны через свойство sentences; токены состоят из одного или нескольких слов, слова доступны через свойство words у sentences; обязательная первая часть процесса
		mwt	разделяет составные токены на отдельные слова (Word), работает только для нескольких языков; является вторым этапом процесса перед следующими
		pos	определяет лексические и грамматические свойства токенов: часть речи (свойство Word.upos), дополнительные морфологические параметры, например, род, тональность, степень сравнения (свойство Word.feats).
		lemma	выделяет условную основную форму слова (лемму), свойство Word.lemma
		depparse	определяет синтаксическую роль слова в предложении (свойства Word.head и Word.deprel)
		ner	определяет именованные сущности, т. е. слова или словосочетания, именующие конкретных людей, животных, предметов, явлений (свойство Token.ner)
lang	язык текста	'en','ru','fr' и т. д.	
dir	Каталог для хранения моделей Stanza.		По умолчанию Stanza хранит свои модели в папке в каталоге, где установлена.

Рассмотрим несколько примеров обработки текста. Первым рассмотрим процесс деления русскоязычного текста на предложения и слова. В этом примере nlp – это объект, обрабатывающий текст text_short, doc – результат обработки, модель текста. Из модели можно извлечь список предложений – doc.sentences, а из предложений – список токенов sentence.tokens. В данном случае каждый токен состоит из одного слова. token.id – номер токена в предложении, token.text – текстовое содержимое токена.

Пример 1.
Код

```
import stanza
PIPELINE_CFG = {"processors": 'tokenize'}
nlp = stanza.Pipeline(**PIPELINE_CFG, lang='ru')
text_short = """Пять тысяч они просят. Не миллион же! Я верну."""
doc = nlp(text_short)
for i, sentence in enumerate(doc.sentences):
  print(f'====== Sentence {i+1} =======')
  for token in sentence.tokens:
    print(*'id: {token.id}\t text: {token.text}\n')
```

Результат работы:

```
====== Sentence 1 =======
id: 1  text: Пять
id: 2  text: тысяч
id: 3  text: они
id: 4  text: просят
id: 5  text: .
====== Sentence 2 =======
id: 1  text: Не
id: 2  text: миллион
id: 3  text: же
id: 4  text: !
====== Sentence 3 =======
id: 1  text: Я
id: 2  text: верну
id: 3  text: .
```

Рассмотрим процесс определения частей речи и морфологических признаков слов. Список слов – sentence.words, часть речи – word.upos, морфологические признаки – word.feats.

Пример 2.
Код

```
import stanza
PIPELINE_CFG = {"processors": 'tokenize,pos'}
nlp = stanza.Pipeline(**PIPELINE_CFG, lang='ru')
text_short = """Пять тысяч они просят. Не миллион же! Я верну."""
doc = nlp(text_short)
for sentence in doc.sentences:
  for word in sentence.words:
    print('word: {}\tupos: {}\tfeats: {}'.format(word.text, word.upos, word.feats if word.feats else "_"))
```

Результат работы:

```
word: Пять       upos: NUM    feats: Case=Acc
word: тысяч  upos: NOUN   feats: Animacy=Inan|Case=Gen|Gender=Fem|Number=Plur
word: они   upos: PRON   feats: Case=Nom|Number=Plur|Person=3
word: просят upos: VERB   feats: Aspect=Imp|Mood=Ind|Number=Plur|Per-
son=3|Tense=Pres|VerbForm=Fin|Voice=Act
word: . upos: PUNCT  feats: _
word: Не    upos: PART   feats: _
```

```
word: миллион upos: NOUN  feats: Animacy=Inan|Case=Nom|Gender=Masc|Number=Sing
word: же    upos: PART   feats: _
word: !    upos: PUNCT  feats: _
word: Я   upos: PRON   feats: Case=Nom|Number=Sing|Person=1
word: верну  upos: VERB   feats: Aspect=Perf|Mood=Ind|Number=Sing|Per-
son=1|Tense=Fut|VerbForm=Fin|Voice=Act
word: .    upos: PUNCT   feats: _
```

В каждой строке выводится текст слова и его маркеры, определённые библиотекой. Сначала выводится часть речи в виде краткого обозначения. Список всех определяемых частей речи можно найти в [Universal POS tags]. Соответствующие значения можно использовать в коде программы для поиска и сравнения в ходе реализации алгоритмов. Затем идёт список морфологических характеристик. Список морфологических признаков приведён в [Universal features].

Рассмотрим процесс лемматизации. Лемма слова – word.lemma.

Пример 3.

Код

```
import stanza
PIPELINE_CFG = {"processors": 'tokenize,pos,lemma'}
nlp = stanza.Pipeline(**PIPELINE_CFG, lang='ru')
text_short = """Пять тысяч они просят. Не миллион же! Я верну."""
doc = nlp(text_short)
for sentence in doc.sentences:
  for word in sentence.words:
    print('word: {word.text+" "}\tlemma: {word.lemma}\n')
```

Результат работы:

```
word: Пять          lemma: пять
word: тысяч         lemma: тысяча
word: они           lemma: они
word: просят        lemma: просить
word: .             lemma: .
word: Не            lemma: не
word: миллион       lemma: миллион
word: же            lemma: же
word: !             lemma: !
word: Я             lemma: я
word: верну         lemma: вернуть
word: .             lemma: .
```

Рассмотрим процесс синтаксического разбора предложений и определения роли слова в предложении. Уровень слова в дереве синтаксического разбора предложения – word.head, роль слова – word.deprel.

Пример 4.

Код

```
import stanza
PIPELINE_CFG = {"processors": 'tokenize,pos,lemma,depparse'}
nlp = stanza.Pipeline(**PIPELINE_CFG, lang='ru')
text_short = """Пять тысяч они просят. Не миллион же! Я верну."""
doc = nlp(text_short)
for sentence in doc.sentences:
  for word in sentence.words:
    print('id: {}\tword: {}\thead id: {}\thead: {}\tdeprel: {}'.
format(word.id, word.text, word.head,                    sentence.words[word.head-1].text
if word.head > 0 else "root", word.deprel))
```

Результат работы:

```
id: 1  word: Пять   head id: 2   head: тысяч   deprel: nummod:gov
id: 2  word: тысяч  head id: 4   head: просят  deprel: obl
id: 3  word: они    head id: 4   head: просят  deprel: nsubj
id: 4  word: просят  head id: 0   head: root   deprel: root
id: 5  word: .      head id: 4   head: просят  deprel: punct
id: 1  word: Не     head id: 2   head: миллион  deprel: advmod
id: 2  word: миллион head id: 0   head: root   deprel: root
id: 3  word: же     head id: 2   head: миллион  deprel: advmod
id: 4  word: !      head id: 2   head: миллион  deprel: punct
id: 1  word: Я      head id: 2   head: верну   deprel: nsubj
id: 2  word: верну  head id: 0   head: root   deprel: root
id: 3  word: .      head id: 2   head: верну   deprel: punct
```

Рассмотрим непосредственно содержимое модели текста. Параметры каждого слова объединены фигурными скобками, предложения ограничены квадратными скобками.

Пример 5.

Код

```
import stanza
PIPELINE_CFG = {"processors": 'tokenize,pos,lemma,depparse'}
nlp = stanza.Pipeline(**PIPELINE_CFG, lang='ru')
text_short = """Пять тысяч они просят. Не миллион же! Я верну."""
doc = nlp(text_short)
for sentence in doc.sentences:
  print(sentence)
```

Результат работы:

```
[
 {
  "id": "1",
  "text": "Пять",
  "lemma": "пять",
  "upos": "NUM",
  "feats": "Case=Acc",
  "head": 2,
  "deprel": "nummod:gov",
  "misc": "start_char=0|end_char=4"
 },
 {
  "id": "2",
  "text": "тысяч",
  "lemma": "тысяча",
  "upos": "NOUN",
  "feats": "Animacy=Inan|Case=Gen|Gender=Fem|Number=Plur",
  "head": 4,
  "deprel": "obl",
  "misc": "start_char=5|end_char=10"
 },
 {
  "id": "3",
  "text": "они",
  "lemma": "они",
  "upos": "PRON",
  "feats": "Case=Nom|Number=Plur|Person=3",
  "head": 4,
  "deprel": "nsubj",
```

```
 "misc": "start_char=11|end_char=14"
},
{
 "id": "4",
 "text": "просят",
 "lemma": "просить",
 "upos": "VERB",
 "feats": "Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act",
 "head": 0,
 "deprel": "root",
 "misc": "start_char=15|end_char=21"
},
{
 "id": "5",
 "text": ".",
 "lemma": ".",
 "upos": "PUNCT",
 "head": 4,
 "deprel": "punct",
 "misc": "start_char=21|end_char=22"
}
]
[
{
 "id": "1",
 "text": "Не",
 "lemma": "не",
 "upos": "PART",
 "head": 2,
 "deprel": "advmod",
 "misc": "start_char=23|end_char=25"
},
{
 "id": "2",
 "text": "миллион",
 "lemma": "миллион",
 "upos": "NOUN",
 "feats": "Animacy=Inan|Case=Nom|Gender=Masc|Number=Sing",
 "head": 0,
 "deprel": "root",
 "misc": "start_char=26|end_char=33"
},
{
 "id": "3",
 "text": "же",
 "lemma": "же",
 "upos": "PART",
 "head": 2,
 "deprel": "advmod",
 "misc": "start_char=34|end_char=36"
},
{
 "id": "4",
 "text": "!",
 "lemma": "!",
 "upos": "PUNCT",
```

```
  "head": 2,
  "deprel": "punct",
  "misc": "start_char=36|end_char=37"
 }
]
[
 {
  "id": "1",
  "text": "Я",
  "lemma": "я",
  "upos": "PRON",
  "feats": "Case=Nom|Number=Sing|Person=1",
  "head": 2,
  "deprel": "nsubj",
  "misc": "start_char=38|end_char=39"
 },
 {
  "id": "2",
  "text": "верну",
  "lemma": "вернуть",
  "upos": "VERB",
  "feats": "Aspect=Perf|Mood=Ind|Number=Sing|Person=1|Tense=Fut|VerbForm=Fin|Voice=Act",
  "head": 0,
  "deprel": "root",
  "misc": "start_char=40|end_char=45"
 },
 {
  "id": "3",
  "text": ".",
  "lemma": ".",
  "upos": "PUNCT",
  "head": 2,
  "deprel": "punct",
  "misc": "start_char=45|end_char=46"
 }
]
```


Кроме уже встречавшихся в предыдущих примерах маркеров, в начале описания каждого токена указан его номер в предложении – id, а в конце – индексы местонахождения в тексте в целом посимвольно.

Дополнительно можно отметить возможность анализа тональности для трёх языков: английского, немецкого и китайского. Процессор sentiment добавляет метку настроения к каждому предложению. Метка обозначает отрицательные, нейтральные и положительные предложения, представленные цифрами 0, 1, 2 соответственно.

Подводя итог, можно отметить, что следующие **качества Stanza**:

1. Реализация на языке Python, требующая минимальных усилий для установки.
2. Полный нейронный сетевой конвейер для надежной аналитики текста, включая токенизацию, расширение многострочного токена (MWT), лемматизацию, тегирование частей речи (POS) и морфологических функций, разбор зависимостей и распознавание именованных объектов.
3. Предварительно обученные нейронные модели, поддерживающие 66 (человеческих) языков.

NLTK

Одной из универсальных библиотек для обработки естественного языка является NLTK (Natural Language Toolkit). – это важная библиотека, поддерживающая такие задачи, как деление текста на предложения и слова, стемминг, лемматизация, морфологическая разметка, синтаксический анализ. Она поддерживает работу со множеством языков, в том числе, с русским.

Для того, чтобы использовать NLTK её надо загрузить с официального сайта [Natural Language Toolkit], затем импортировать в код на Pyton. При этом мы получаем доступ к функционалу библиотеки через объект nltk. Библиотека работает довольно медленно, поэтому для оптимизации её использования можно импортировать отдельные модули и работать именно с ними для выполнения нужных действий. Учебник и описание библиотеки можно найти в книге [Bird, 2020].

Рассмотрим некоторые возможности NLTK. Разделение текста на предложения-компоненты осуществляется методом nltk.sent_tokenize. Результатом работы метода является список предложений.

Пример 1.
Код

```
import nltk
text_short = """Пять тысяч они просят. Не миллион же! Я верну."""
sentences = nltk.sent_tokenize(text_short)
for sentence in sentences:
  print(sentence)
```

Результат
Пять тысяч они просят.
Не миллион же!
Я верну.

Деление на отдельные слова осуществляется методом nltk.word_tokenize. Этот метод выделяет отдельные слова и не ориентирован на предложения. Результатом работы метода является список слов текста.

Пример 2.
Код

```
from nltk import word_tokenize
text_short = """Пять тысяч они просят. Не миллион же! Я верну."""
tokens = word_tokenize(text_short)
print(tokens)
```

Результат
['Пять', 'тысяч', 'они', 'просят', '.', 'Не', 'миллион', 'же', '!', 'Я', 'верну', '.']

Из результата видно, что в качестве отдельных токенов рассматриваются знаки препинания. Чтобы задать другой способ разбиения, можно использовать **регулярные выражения и RegexpTokenizer.**

Пример 3.
Код

```
from nltk.tokenize import RegexpTokenizer
text_short = """Пять тысяч они просят. Не миллион же! Я верну."""
tokenizer = RegexpTokenizer(r'\w+')
tokens = tokenizer.tokenize(text_short)
print(tokens)
```

Результат
['Пять', 'тысяч', 'они', 'просят', 'Не', 'миллион', 'же', 'Я', 'верну']

Чтобы определить часть речи для слов можно использовать nltk.pos_tag. Метод работает со списком токенов и возвращает список пар: токен и маркер. К сожалению алгоритм маркировки далеко не всегда работает правильно, как и в других библиотеках.

Приведём список некоторых маркеров

- CD цифра
- FW иностранное слово
- JJ прилагательное
- JJR прилагательное в сравнительной степени
- NN существительное в единственном числе
- NNS существительное во множественном числе
- NNP существительное имя собственное в единственном числе

- PRP личное местоимение
- RB наречие
- RBR сравнительное наречие
- RP частица
- UH междометие
- VB основная форма глагола
- VBD глагол в прошедшем времени
- VBP глагол в настоящем времени
- WP местоимение

Полный список можно найти в [NLTK Part of Speech Tagging Tutorial].

Пример 4.

Код

```
import nltk
from nltk import word_tokenize
text_short = """What were you doing behind the curtain?"""
tokens = word_tokenize(text_short)
tagged = nltk.pos_tag(tokens)
print(tagged)
```

Результат работы:

[('What', 'WP'), ('were', 'VBD'), ('you', 'PRP'), ('doing', 'VBG'), ('behind', 'IN'), ('the', 'DT'), ('curtain', 'NN'), ('?', '.')]

Для выделения корня слова, т. е. стемминга можно использовать разные алгоритмы. Они отличаются в частности наборами языков для которых работают. Приведём примеры их использования:

1.
```
from nltk.stem.porter import PorterStemmer
porter_stemmer = PorterStemmer()
print(porter_stemmer.stem("crying"))
```
1.
```
from nltk.stem.lancaster import LancasterStemmer
lancaster_stemmer = LancasterStemmer()
print(lancaster_stemmer.stem("crying"))
```
2.
```
from nltk.stem import SnowballStemmer
snowball_stemmer = SnowballStemmer("russian")
print(snowball_stemmer.stem("следующий"))
```

Русский язык поддерживает SnowballStemmer

Лемматизация осуществляется с помощью дополнительного инструмента WordNetLemmatizer (nltk.stem.WordNetLemmatizer()). Его метод lemmatize(слово) позволяет получить лемму указанного слова. Улучшить качество лемматизации можно уточнив маркер слова в методе lemmatize: lemmatizer.lemmatize("seen", wordnet.VERB) – здесь указывается, что необходимо найти начальную форму глагола "seen".

Если слова встречаются в тексте часто и не несут большой информативной нагрузки, то они называются стоп-словами. Библиотека NLTK также имеет список стоп-слов, который предварительно необходимо скачать. Это можно сделать следующим образом:

```
import nltk
nltk.download('stopwords')
```

После этого доступен список стоп-слов для разных языков, в том числе для русского языка:

```
from nltk.corpus import stopwords
stopwords.words("russian")
```

Поскольку stopwords.words возвращает список, то к нему можно добавить дополнительные слова или, наоборот, удалить из него те, которые важны для решаемой задачи. Например, если нужно дополнить стоп-слова из другого списка tokens:

```
stop_words = stopwords.words("russian")
for token in tokens:
  if token not in stop_words:
    stop_words.append(token)
```

Чтобы удалить стоп-слова из текста, обычно используют методы списков: текст преобразуется в список слов, из которого удаляют элементы списка стоп-слов.

Выделим важные особенности NLTK:

- Наиболее известная и многофункциональная библиотека для NLP;
- Поддерживается множество языков.
- Медленная;
- Сложная в изучении и использовании.

TextBlob

TextBlob [TextBlob: Simplified Text Processing] предоставляет простой интерфейс для решения стандартных задач по обработке текста. Кроме того, эта библиотека содержит методы, реализующие алгоритмы определения тональности, взаимодействия со словарём английского языка WordNet, поиска n-грамм, которые на сегодняшний день являются очень популярными параметрами методов NLP.

Рассмотрим сначала стандартные задачи токенизации и маркировки текста. Части речи маркируются аналогично NLTK.

Пример 1.

Код

```
from textblob import TextBlob
text_short = """What were you doing behind the curtain?"""
blob = TextBlob(text_short )
print(blob.tags)
```

Результат

```
[('What', 'WP'), ('were', 'VBD'), ('you', 'PRP'), ('doing', 'VBG'), ('behind', 'IN'), ('the', 'DT'), ('curtain', 'NN')]
```

Как и другие библиотеки TextBlob умеет выделять слова методом words в виде списка и предложения методом sentences

Пример 2.

Код

```
from textblob import TextBlob
text_short = """Пять тысяч они просят. Не миллион же! Я верну."""
blob = TextBlob(text_short )
print(blob.words)
print(blob.sentences)
```

Результат

```
['Пять', 'тысяч', 'они', 'просят', 'Не', 'миллион', 'же', 'Я', 'верну']
[Sentence("Пять тысяч они просят."), Sentence("Не миллион же!"), Sentence("Я верну.")]
```

Кроме отдельных предложений, можно выделить фразы. Как правило, они выделяются на основе пунктуации.

Пример 3.

Код

```
from textblob import TextBlob
text = '''Как только ударял в Киеве поутру довольно звонкий семинарский колокол, висевший у ворот Братского монастыря, то уже со всего города спешили толпами школьники и бурсаки.'''
blob = TextBlob(text)
print(blob.noun_phrases)
```

Результат

```
['как только ударял в киеве поутру довольно звонкий семинарский колокол', 'висевший у ворот братского монастыря', 'то уже со всего города спешили толпами школьники и бурсаки']
```

Слова, которые получаются в процессе токенизации являются объектами класса Word. В этом классе есть методы преобразующие слова ко множественному числу pluralize() и единичному

singularize(), а также лемматизирующие слово lemmatize(), lemmatize(pos), в последнем методе можно указать часть речи для повышения качества лемматизации.

Рассмотрим некоторые специальные возможности TextBlob. В библиотеке есть возможность определения тональности текста. Для любого текста определяется свойство sentiment с двумя параметрами: polarity (полярность), subjectivity (субъективность). Оценка полярности измеряется в диапазоне от -1,0 (негативная) до 1,0 (позитивная). Субъективность измеряется в диапазоне от 0,0 (очень объективная) до 1,0 (очень субъективная).

Пример 4.

Код

```
from textblob import TextBlob
text_short = """Пять тысяч они просят. Не миллион же! Я верну."""
blob = TextBlob(text_short)
for sentence in blob.sentences:
  print(sentence.sentiment.polarity)
```

Результат

```
0.0
0.0
0.0
```

TextBlob использует электронный словарь-тезаурус для английского языка WordNet. Из него могут быть извлечены множества синонимов (синсеты) для слова методом synsets и определения слов методом definitions. Метод correct() позволяет исправить орфографию.

Все библиотеки обработки текста содержат методы, возвращающие списки слов. Поэтому числовые характеристики, например, частоту встречаемости, как правило рассчитывают программными средствами, методами списков, коллекций. Однако TextBlob позволяет методом word_counts ("слово") находить частоту встречаемости слова в тексте.

Ещё один полезный метод возвращает список n-грамм текста – ngrams(n). N-граммы – это последовательности из заданного количества слов в тексте. Вычисление частоты n-грамм является частью многих методов обработки текстов, например, для определения авторского стиля. Каждая n-грамма представляет отдельный список WordList из заданного количества последовательных слов текста. Деление на предложения не учитывается. Если это важно, метод необходимо применять не ко всему тексту, а к предложениям по отдельности.

Пример 5.

Код

```
from textblob import TextBlob
text_short = """Пять тысяч они просят. Не миллион же! Я верну."""
blob = TextBlob(text_short)
print(blob.ngrams(n=3))
```

Результат

[WordList(['Пять', 'тысяч', 'они']), WordList(['тысяч', 'они', 'просят']), WordList(['они', 'просят', 'Не']), WordList(['просят', 'Не', 'миллион']), WordList(['Не', 'миллион', 'же']), WordList(['миллион', 'же', 'Я']), WordList(['же', 'Я', 'верну'])]

Выделим важные особенности TextBlob:

- Наиболее простая в использовании библиотека
- Вычисляет n-граммы
- Решает задачи по определению тональности
- Слишком медленная для обработки текстов большого объёма

SpaCy

SpaCy [SpaCy. Industrial-streng Natural Language Processing in Pyton] относительно новая библиотека, предназначенная для обработки текста в больших объёмах и промышленных приложениях. SpaCy предлагает самый быстрый синтаксический парсер, имеющийся сегодня на рынке. К сожалению, на сегодняшний день она поддерживает только семь языков и среди них нет русского. Однако

растущая популярность машинного обучения, NLP и spaCy как ключевой библиотеки означает, что этот инструмент может вскоре начать поддерживать больше языков.

Чтобы работать с этой библиотекой, её надо подключить (import spacy) и загрузить объект для обработки текста методом spacy.load. Основные части процесса обработки текста: токенизатор и маркер, определение синтаксических зависимостей, распознавание именованных сущностей. Рассмотрим небольшой пример.

Пример 1.

Код

```
import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp("There was no possibility of taking a walk that day.")
for token in doc:
  print(token.text, token.pos_, token.dep_, token.lemma_)
```

Результат:

```
There PRON expl there
was AUX ROOT be
no DET det no
possibility NOUN attr possibility
of ADP prep of
taking VERB pcomp take
a DET det a
walk NOUN dobj walk
that DET det that
day NOUN npadvmod day
. PUNCT punct
```

В этом примере объект doc содержит в себе модель текста, полученную в результате обработки. Из этой модели выводится информация о токене: текст (token.text), маркировка, в основном часть речи (token.pos_), роль в предложении, на основании дерева синтаксической зависимости (token.dep_), а также лемма (token.lemma_).

SpaCy предоставляет возможность поиска текста по определённому шаблону. Построение этого шаблона осуществляется с помощью специального объекта Matcher (from spacy.matcher import Matcher). Для поиска необходимо задать правило с набором параметров, по которому будет осуществляться поиск. Например, надо найти в тексте слово number в значении глагола "нумеровать". В этом случае надо задать правило поиска слова "number" с маркером "VERB" (глагол). Правило добавляется в **объект** matcher **на основе ш**аблона pattern. Шаблон определяется заданием последовательности текста и свойств этого текста.

Пример 2.

Код

```
import spacy
nlp = spacy.load('en_core_web_sm')
from spacy.matcher import Matcher
matcher = Matcher(nlp.vocab)
pattern = [{'TEXT': 'number', 'POS': 'VERB'}]
matcher.add('rule_1', None, pattern)
doc1 = nlp("We must number all the pages in the copy-book")
matches = matcher(doc1)
print(matches)
doc2 = nlp("Draw the relevant number for each picture")
matches = matcher(doc2)
print(matches)
```

Результат:

```
[(7604275899133490726, 2, 3)]
[]
```

Объект matches содержит результаты поиска текста, соответствующего правилу во всём документе, в том числе номер позиции токена. Если текста, соответствующего правилу, нет, объект пустой. В примере глагол «number» найден в doc1 и отсутствует в doc2.

Результат обработки текста можно визуализировать с помощью объекта displacy.

Пример 3.

Код

```
import spacy
from spacy import displacy
nlp = spacy.load("en_core_web_sm")
doc = nlp("There was no possibility of taking a walk that day.")
displacy.serve(doc, style='dep')
```

В результате получится изображение со схемой текста с указанием маркеров и зависимостей токенов (рис. 34).

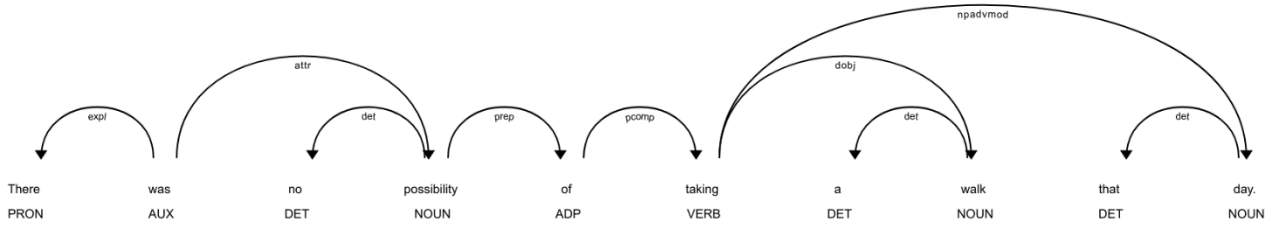


Рис 33. Схема текста

Библиотека SpaCy предоставляет возможность строить векторное представление слов и текстов. Вектора затем могут быть использованы в задачах обработки естественного языка и алгоритмах машинного обучения, например, для классификации текстов. Построение векторов осуществляется на основании модели word2vec. Идея векторного представления основывается на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами, будут иметь близкие векторы (подробнее с этой моделью можно познакомиться в статье [Zafar Ali, 2020]. SpaCy использует существующий корпус текстов языка, чтобы построить векторные представления слов.

Каждый документ и токен имеют метод similarity(), позволяющий сравнить его с другим объектом и определить сходство по шкале от 0 до 1. Метод токена vector возвращает вектор для лексемы соответствующего слова, а метод vector текста возвращает среднее значение векторов своих токенов.

Выделим важные особенности SpaCy:

- Самая быстрая библиотека для NLP;
- Простая в изучении и использовании;
- Есть встроенные вектора слов;
- Использует нейронные сети для тренировки моделей.
- Поддерживает маленькое количество языков.

В заключение следует отметить, что технологии обработки текстов на естественном языке могут предоставить широкий спектр ценных идей и решений для современных программных систем. Поэтому важно знать и использовать возможности инструментов обработки текста при решении задач как в сфере IT, так и в сфере классической лингвистики.

🕮 Вопросы и задания

1. Из каких частей может состоять предварительная обработка текста?
2. Назовите основные единицы текста, которые выделяются для его автоматической обработки?
3. Дайте определение понятиям стемминг и лемматизация.
4. Какие из четырёх описанных библиотек по обработке текста поддерживают работу с русским языком?
5. Установите одну из библиотек по обработке текста поддерживающую работу с русским языком. Создайте файл, содержащий текст рассказа или статьи. Например, выберите рассказ А. П. Чехова

из электронного ресурса (http://chehov-lit.ru/chehov/text/rasskazy.htm). С помощью языка программирования pyton решите следующие задачи:

а) определите количество слов текста;

б) определите количество предложений в тексте;

в) определите количество знаков пунктуации в тексте;

г) определите долю существительных и долю глаголов относительно всех слов текста;

д) введите слово и выведите список предложений, в которых оно встречается;

е) определите количество различных слов в тексте;

ж) выведите предложения, которые начинаются и заканчиваются на одно и тоже слово;

з) выведите пары предложений, которые начинаются одним и тем же словом или словосочетанием.

6. Напишите регулярное выражение, которое определяет строки, начинающиеся с заглавной буквы.

7. Установите вторую библиотеку по обработке текста, как для задания 5, и сравните ее с первой по скорости выполнения задач а-з.

§2 Ритмические средства для автоматизированного анализа

Оптимизация работы с ритмом русскоязычных англоязычных, франкоязычных и испаноязычных текстов с целью определения специфики ритма произведения привела к необходимости создать инструмент (ProseRhythmDetector - PRD, позволяющий осуществлять автоматизированный анализ некоторых ритмических средств. В их число вошли **анафора, эпифора, симплока, диакопа, эпизевксис, анадиплозис, эпаналепсис, полисиндетон**.

Выбор данных средств для анализа ритма, а именно для их автоматизированного поиска и количественной обработки обусловлен несколькими факторами. Во-первых, это наиболее частотные ритмические средства, употребляемые в прозаических текстах[1], и именно они выделяются в качестве ритмических средств на лексико-грамматическом уровне большинством лингвистов, проводящих исследования в области ритмизации текста. Во-вторых, эти средства в большей степени доступны с точки зрения автоматизированной обработки ритма. В-третьих, данные средства являются наиболее ощутимыми и показательными с точки зрения восприятия ритма реципиентом. В-четвертых, ритмические средства способствуют определению идиолекта автора, его творческого этоса.

Безусловно, данные средства в большей степени характеризуют языковые особенности стиля писателя, но они имеют очень тесную связь с образами и характерами, представленными в произведении, а, следовательно, с картиной мира писателя, его восприятием окружающей действительности. Ключевые слова, выделяемые инструментом, их формы и статистика позволяют очертить круг лексем, входящих в лексико-семантическое поле языка писателя.

Так, поиск **анафоры** сосредоточен на тавтологической анафоре, которая рассматривается как **повтор слов или словосочетаний в начале нескольких идущих подряд предложений**, а также в начале частей одного предложения (внутрифразовая анафора), например:

It is true she had neither strong feelings to overcome, nor tender feelings by which to be miserably pained. **It is true** likewise that she had an important avocation, a real business to fill her time, divert her thoughts, and divide her interest (Ch. Brontë «Villette»);

Неужто это тот самый голос?

Неужто передо мной лицо самого автора письма?

Неужто передо мной на темном чердаке – Джон Грэм, доктор Бреттон собственной персоной?Да, это был он.

Неужто это тот самый голос?

Неужто передо мной лицо самого автора письма? (Ш. Бронте «Городок»);

Le malheur est notre plus grand maître, **le malheur** lui apprendra la valeur de l'argent, celle des hommes et celle des femmes (O. de Balzac «Gobseck»).

Инструмент выделяет лексическую анафору, элементы которой в большей степени выражены местоимением (местоименная анафора) и существительным в английском и французском языках, а также местоимением, наречием, существительным и глаголом в русском языке.

При выделении внутрифразовой анафоры, как в последнем примере, встает вопрос о разграничении анафоры и эпаналепсиса, указывающего на повтор слова после промежуточных слов. В этом случае основным критерием разграничения является близость повторяющихся элементов друг к другу: чем они ближе, тем больше вероятность того, что перед нами эпаналепсис.

В качестве **эпифоры** инструмент рассматривает **повтор слова или сочетания слов в конце нескольких предложений или в конце нескольких компонентов предложений***:*

Laissez−moi tranquille avec votre hideuse **réalité !** Qu'est−ce que cela veut dire, la **réalité ?** (Flaubert «Education sentimentale»),

Cependant, objecta Frédéric, de longs cheveux **noirs,** avec de grands yeux **noirs**... (Flaubert «Education sentimentale»);

Понятие **симплоки** трактуется лингвистами как **фигура синтаксического параллелизма в смежных стихах**, у которых а) одинаковые начало и конец при разной середине и б) наоборот,

[1] Данное утверждение основывается на исследованиях, проведенных авторами на материале французских романов О. де Бальзака, Г. Флобера, Стендаля, Г. де Мопассана, Э. Золя, А. Моруа, М. Дюрас, Ф. Саган, А. Гавальда, Ф. Бегбедера, А. Нотомб, а также англоязычных романов Ч. Диккенса, Ш. Бронте, К. Аткинсон, Д. дю Морье, Дж. К. Роулинг, Дж. Остин, Э. Гаскелл, Дж. Джойс, А. Мердок, Ф. С. Фицджеральда, а также испаноязычных романов Г. Г. Маркеса.

разные начало и конец при одинаковой середине: **Во поле** березонька **стояла**, **Во поле** кудрявая **стояла**. [Квятковский, 1966].

Инструментом осуществляется поиск варианта симплоки, при котором определяются одинаковые начала и окончания нескольких фраз при разной середине:

Я поняла это тотчас и не стала смотреть дальше если б мне и хотелось смотреть, то просто времени не оставалось; уж было поздно; мы с Джиневрой собрались на улицу **Фоссет**.

Я поднялась и распрощалась с крестной и с мосье де Бассомпьером.То ли профессор Эманюель заметил, что я не поощряла веселости доктора Бреттона, то ли догадался, что мне горько и что вообще для легкомысленной мадемуазель Люси, охотницы до развлечений, вечер оказался не таким уж праздником, но когда я покидала залу, он встал и спросил, провожает ли меня кто-нибудь до улицы **Фоссет** (Ш. Бронте «Городок»).

Эпаналепсис рассматривается в отечественной и зарубежной стилистике как **риторическая фигура, созданная путем повторения одного и того же слова, оборота, фрагмента предложения** [Жеребило, 2010], а также как стилистическая фигура, заключающаяся в повторении одного и того же слова или выражения в длинной фразе или периоде; словесные повторы в начале и конце строфы или только в конце строф (в поэзии) [Квятковский, 1966]. В зарубежной стилистике данное средство рассматривается в качестве синонима эпифоры.

Инструмент осуществляет поиск эпаналепсиса как **повторений начального и конечного элементов целого предложения или компонента предложения:**

J'ai gagné une bataille, se dit-il aussitôt qu'il se vit dans les bois et loin du regard des hommes, **j'ai donc gagné une bataille** ! (Stendhal «Le Rouge et le Noir»)

Меня могут мучить боли или слабость, но болезнь или недомогание, вызванные любовными страданиями, никогда еще не одолевали **меня** (Там же).

Под **эпизевксисом** (epizeuxis, épizeuxe), понимается **повтор слова без промежуточных элементов,** который в зарубежной стилистике часто называется палилогией (palilogia, palilogie):

Allons, **allons**, la soupe est cuite. (Maupassant «Mon ami»)

Анадиплозис – это средство, при котором **повтор слов наблюдается на стыке частей предложения или на стыке предложений**:

A bold thought was sent to **my mind**; **my mind** was made strong to receive it (Brontë «Villette») – анадиплозис.

В качестве **диакопы** мы рассматриваем **повтор одного и того же слова, употребленного через промежуточные слова**, например: – **Peur** ?... reprit le commandant, oui, **peur**. J'ai toujours eu **peur** d'être fusillé comme un chien au détour d'un bois sans qu'on vous crie: Qui vive ! (Balzac «Les Chouans»). Это позволяет осуществить более четкий поиск пограничных средств, связанных с повтором слов в тексте.

Полисиндетон (многосоюзие) – стилистическая фигура, состоящая в намеренном увеличении количества союзов в предложении, обычно для связи однородных членов. Замедляя речь вынужденными паузами, многосоюзие подчёркивает роль каждого из слов, создавая единство перечисления и усиливая выразительность речи [Словарь литературоведечских терминов, 1974]. Инструментом осуществляется поиск предложений, количество союзов в которых не менее трех: The heaviest rain, **and** snow, **and** hail, **and** sleet, could boast of the advantage over him in only one respect (Dickens «A Christams Carol»).

Перспективными результатами работы с полученным инструментом на данном уровне могут быть: 1) результаты работы по сопоставлению специфики ритма англоязычного и франкоязычного текстов и их перевода на русский язык; 2) по расширению спектра ритмических средств при помощи добавления новых стилистических средств, включающих в свою структуру повтор; 3) автоматизации процесса определения коэффициента ритма для исследуемых текстов с целью выявления коэффициента ритмизации для того или иного автора по совокупности его произведений; 4) атрибуции текстов на основе авторского коэффициента ритмизации.

🕮 Вопросы и задания

1. Анализ каких ритмических средств осуществляется инструментом PRD?
2. Чем обусловлен их выбор для анализа ритма?

3. Дайте определение анафоры, эпифоры и симплоки.
4. В чем различия употребления эпизевксиса и анадиплозиса?
5. В чем заключаются различия употребления симплоки и эпаналепсиса?
6. Дайте определение диакопы.
7. Каковы перспективы использования инструмента PRD?

§3 Алгоритмы поиска ритмических средств

Для того, чтобы проанализировать ритм прозаического произведения и, например, сравнить его с ритмом перевода, был разработан комплекс алгоритмов, автоматически находящих в тексте ритмические средства, а именно лексические и синтаксические.

Входными данными для всех алгоритмов является необработанный текст. Текст разделяется на предложения, каждое предложение представляется как набор слов. Кроме того, для поиска средств алгоритмы используют стоп-слова (уникальные для каждого средства). Если итоговый аспект состоит только из стоп-слов, он исключается из списка. А алгоритмы поиска многосоюзия используют списки простых союзов, парных союзов и союзных наречий. В качестве выходных данных каждый алгоритм возвращает список аспектов и контекстов заданного ритмического средства.

Каждый аспект состоит из слова или словосочетания, повторяющегося в нескольких предложениях. Список этих предложений является контекстом для данного аспекта. По контекстам эксперт-лингвист может сравнить оригинал текста и его перевод, чтобы проанализировать, как именно переводчик отражает ритмические средства автора текста. Также по контексту можно определить, верно ли выделен аспект.

Приведём пример появления ритмического средства в тексте:

He does. He is fond of you. You are his favourite.

Здесь употреблен **анадиплозис.** Его появление, т.е. слово *you*, считается аспектом. Контекст – пара предложений *He is fond of you, You are his favourite*.

Рассмотрим подробнее алгоритмы поиска ритмических средств.

Список **анафор** составляется следующим образом. На первом шаге составляется список кандидатов в анафоры, где каждый кандидат состоит из слова и списка соседних предложений. Аспект для кандидата составляется циклически: на каждом шаге цикла добавляется очередное повторяющееся слово. Если таких слов оказывается несколько для разных предложений, берется аспект, соответствующий большему числу предложений. После добавления слова в аспект, фильтруется контекст, т. е. остаются предложения, начинающиеся с набора слов из аспекта, но не состоящие из него полностью. Как только не удается найти подходящее повторяющееся слово, цикл заканчивается.

Алгоритм поиска **эпифоры** аналогичен поиску анафоры.

Для алгоритма поиска **анадиплозиса** представим текст как набор слов и знаков препинания. Алгоритм ищет анадиплозис циклически, на каждом шаге выявляя аспекты с большим количеством повторяющихся слов. На первом шаге алгоритм перебирает текст и ищет в нем списки из двух одинаковых слов, разделённых одним знаком препинания. Затем ищутся пары, тройки одинаковых слов, разделенных одним знаком препинания и т. д.

Алгоритм поиска **диакопы** просматривает уникальные слова в данном предложении и для каждого из них ищет все их позиции, которые не являются смежными. Если слово повторяется два или более раз, оно образует диакопу. После этого алгоритм перебирает найденные диакопы и объединяет те из них, которые есть в соседних словах. Это позволяет находить средство с многократным повторением слов.

Алгоритм поиска **эпаналепсиса** проходит по первой половине данного предложения и проверяет, заканчивается ли предложение словами из начала предложения. Повторение с максимальной длиной образует эпаналепсис.

Поиск **эпизевксиса** осуществляется при помощи двух различных алгоритмов. Первый алгоритм ищет его в соседних предложениях, например: *Weak! Weak! Weak!* Он просматривает предложения и проверяет, совпадает ли данное предложение с предыдущим. Если это так, то эти предложения являются частью цепного повторения. Цепочка повторений образует эпизевксис.

Второй алгоритм ищет эпизевксис внутри предложения. Например: *Pretty, pretty good!* Алгоритм проходит первую половину данного предложения и проверяет, повторяется ли эта часть дважды. Если это так, то алгоритм проверяет оставшуюся часть предложения и ищет дополнительные повторы этой же части. Повторения образуют эпизевксис.

В алгоритме поиска **многосоюзия** (**полисиндетона**) для каждого союза или союзного наречия каждого предложения проверяется условие, что союз/наречие повторяется в предложении более одного раза, возможно, в начале или в конце. Если условие выполняется, в список аспектов добавляется аспект с этим союзом. Список стоп-слов для этого аспекта не применяется.

Алгоритм поиска **симплоки** ищет анафору и эпифору, контексты которых пересекаются. Эти фигуры, употребленные одновременно, образуют симплоку.

🕮 Вопросы и задания

1. Что такое аспект и контекст средства?
2. Какие лексические средства ищут описанные алгоритмы?
3. Какие типы эпизевксиса ищут описанные алгоритмы?
4. Какие части речи ищет алгоритм для многосоюзия?
5. Для поиска каких средств необходим список стоп-слов?
6. Вспомните другие ритмические средства. Нужны ли для их поиска стоп-слова?
7. Вспомните другие ритмические средства. Ограничиваются ли они заданным набором слов, как многосоюзие?

Глава IV
ЛИНГВИСТИЧЕСКИЕ ВОЗМОЖНОСТИ ИНСТРУМЕНТА PRD

Основной формой работы с инструментом PRD является анализ текста любого типа и жанра с точки зрения употребления ритмических средств, перечисленных в предыдущей главе. Инструмент работает с текстами на английском, русском, французском и испанском языках. Приведем несколько примеров анализа текста с опорой на результаты, полученные при помощи инструмента.

§1 Анализ ритма художественного текста

В данном разделе приведем пример комплексного анализа ритмической структуры современных текстов. Для этого нами были отобраны романы французских авторов:

G. Legardinier (L'Exil des anges, 2009; *Ça peut pas rater! 2014;* Et soudain tout change, 2013; Nous étions les hommes, 2014; Quelqu'un pour qui trembler, 2015; Le premier miracle, 2016; Demain j'arrête, 2017; Une fois dans ma vie, 2017);

M. Lévy (Et si c'était vrai, 2000; Vous revoir, 2005; Le premier jour, 2009; La Première nuit, 2009; Les enfants de la liberté, 2007; Le voleur d'ombres, 2010; Si c'était à refaire, 2012; Un sentiment plus fort que la peur, 2013);

G. Musso (7 ans après, 2012; Demain, 2013; Et après, 2004; Je reviens te chercher, 2008; La fille de papier, 2010; Que serai-je sans toi? 2009; Sauve-moi, 2005; Seras-tu là? 2006).

Результаты количественной обработки средств по произведениям отражены в таблице 3.

Таблица 3

Ритмические средства современных французских авторов

автор, произведение/ средство	анафора	эпифора	симплока	эпизевксис	анадиплозис	диакопа	эпаналепсис	полисиндетон
G. Legardinier								
L'Exil des anges	18	32	0	6	3	230	2	20
Ça peut pas rater !	14	34	0	7	1	312	0	11
Et soudain tout change	18	54	1	5	2	365	0	23
Nous étions les hommes	15	21	1	4	3	210	1	21
Quelqu'un pour qui trembler	13	26	0	8	6	243	1	13
Le premier miracle	24	33	1	8	3	267	1	15
Demain j'arrête	11	12	0	6	3	211	0	16
Une fois dans ma vie	12	35	0	5	2	322	0	12
Всего:	**125**	**247**	**3**	**39**	**23**	**2160**	**5**	**132**
M. Levy								
Et si c'était vrai	4	22	0	3	0	287	0	14

автор, произведение/ средство	анафора	эпифора	симплока	эпизевксис	анадиплозис	диакопа	эпаналепсис	полисиндетон
Vous revoir	1	12	0	4	0	114	1	11
Le Premier jour	14	24	0	2	2	123	0	5
La Première nuit	7	30	0	12	3	145	0	12
Les enfants de la liberté	14	9	0	10	7	321	1	17
Le voleur d'ombres	3	9	0	2	2	232	0	11
Si c'était à refaire	8	27	0	3	7	276	0	10
Un sentiment plus fort que la peur	6	26	0	3	3	172	1	12
Всего:	**54**	**159**	**0**	**39**	**24**	**1670**	**3**	**92**
G. Musso								
7 ans après	5	12	0	7	1	50	0	6
Demain	11	13	0	12	1	156	1	8
Et après	11	12	0	7	2	151	1	4
Je reviens te chercher	22	21	1	6	5	154	1	9
La fille de papier	8	21	0	8	3	121	1	4
Que serai-je sans toi ?	19	23	0	5	4	187	1	12
Sauve-moi	6	12	0	10	1	112	0	8
Seras-tu là ?	12	13	0	1	4	114	1	7
Всего:	**94**	**127**	**1**	**56**	**21**	**1045**	**6**	**62**

Очевидным является преобладание диакопы, как свободного повтора слова через определенные промежутки текста (в рамках одного предложения). Ввиду того, что повтор практически не ограничен позицией, мы наблюдаем наиболее частое его проявление в исследуемых текстах. Кроме того, наблюдается общая тенденция к такой расстановке средств по количеству в совокупности произведений одного автора (в порядке убывания): *диакопа, эпифора, полисиндетон, анафора, эпизевксис, анадиплозис, эпаналепсис, симплока*.

Для того, чтобы определить, является ли это тенденцией, характеризующей литературу XXI века или только лишь язык данных авторов, необходимо изучить достаточное количество текстов с той же целью, выявить наиболее частотные ритмические средства и определить специфику их употребления.

Поиск *диакопы* осуществляется инструментом в рамках одного предложения. Это позволяет «услышать» повтор, поскольку рассредоточение повторяющихся элементов в абзаце или в нескольких предложениях делает этот повтор незаметным, невоспринимаемым:

Le soleil était ***loin*** *d'avoir atteint son zénith, et ma punition* ***loin*** *d'être achevée. (M. Lévy "Le voleur d'ombres")*

– C'est vrai, admit April, mais je pense que, d'une certaine façon, tu te complais dans la ***douleur*** *et que tu l'entretiens, car ta* ***douleur*** *est le dernier lien qui te rattache encore à Kate et… (G. Musso "Demain")*

En ***bruit*** *de fond, on pouvait entendre des éclats de voix et des cris d'encouragement qui venaient de la fenêtre: sans doute le* ***bruit*** *des gosses jouant au basket sur le bitumen (G.Musso "Et après")*

Употребление данного средства в тексте позволяет возвратиться к определенной мысли, к образу или идее. Диакопа в большей степени проявляется в однократном повторе основного элемента. Это не позволяет тексту звучать навязчиво, но в то же время читатель улавливает то или иное ощущение, чувство, которое хочет передать автор.

Анафорический повтор в исследуемых текстах употребляется чаще в предложениях, где повторяется уже сказанное персонажем, чаще это реплики героев. Такой тип повтора в некотором роде схож с мимезисом:

*– À tes souhaits, mon **bébé.***

*– Je suis plus un **bébé** (G. Musso "Demain")*

*– Ça vous arrive d'écouter **de la vraie musique?***

*– C'est quoi pour vous, **de la «vraie musique»** (G.Musso "La fille de papier")*

Это средство выразительности за редким исключением передает авторские мысли, рассуждения, и совсем не встречается в повествовании или описании.

Что касается *полисиндетона*, то его довольно частое употребление характеризует современный стиль с той позиции, что многие произведения передают лишь некоторую последовательность действий, оставляя читателя без той мыслительной работы, к которой приучали нас авторы-классики. Рассуждения о смысле жизни, о чести, доблести, правде и любви, вероятно, утомляют современного читателя, поэтому проще описать последовательность действий героя, сопровождающуюся определенными его переживаниями:

*Il suffit de rencontrer une Cléa **pour que** plus aucun matin ne soit le même, **pour que** plus rien ne soit comme avant, **pour que** la solitude s'efface. (M. Lévy "Le voleur d'ombres")*

*Ça fait six jours **que** je ne dors pas, **que** je passe mes journées à te guetter, **que** je frôle ta porte. (G. Legardinier "Demain j'arrête")*

Наиболее частотным союзом, безусловно, является союз *que*, однако в некоторых случаях инструменту сложно определить, является повтор союза полисиндетоном или нет. В этом случае решение принимает исследователь. Это обусловлено сложностью идентификации инструментом однородных придаточных предложений и разграничения однородных предложений и придаточных, относящихся к разным антецедентам:

*– Écoute-moi bien, grande girafe ! (Si j'arrive un jour au bureau avec la tête d'un type) ((qui est resté coincé sur un escalier roulant pendant un mois)), (**que** je repars en colère ((alors que je ne perds jamais mon calme)), (**que** de la fenêtre tu me vois marcher sur le trottoir le bras en l'air à quatre-vingt-dix degrés à l'horizontale, puis ouvrir la portière de ma voiture à un passage) ((qui n'existe pas)), (**que** non content de l'effet provoqué je continue à parler en gesticulant dans la voiture), ((comme si je parlais à quelqu'un)), ((mais qu'il n'y a personne, vraiment personne)), et (**que** pour seule explication je te dis) ((**que** je viens de rencontrer un fantôme)),[j'espère] (**que** tu seras aussi inquiet pour moi) ((**que** je le suis pour toi en ce moment)). (M. Lévy "Et si c'était vrai").*

В приведенном примере к полисиндетону можно отнести только союзы, вводящие подчеркнутые однородные придаточные предложения. Остальные формы, находясь в постпозиции по отношению к главному предложению, хоть и вводятся союзом *que*, уже не включаются в эту последовательность.

Наиболее распространенным типом *анафоры* является местоименная анафора, это обусловлено спецификой грамматической структуры французского предложения. Однако в исследуемых текстах нередкими являются примеры, в которых в качестве повторяющегося элемента выступают имена собственные, вопросительные слова и синтаксические конструкции, которые состоят из подлежащего и части сказуемого (вспомогательного глагола при составном именном или глагольном сказуемом). При этом наблюдается не более 4-5 повторов в разных предложениях и редко в составе одного сложносочиненного предложения:

***J'aurais dû** deviner la supercherie, il n'y avait ni adresse, ni timbre. **J'aurais dû** me douter que mon père nous quitterait un jour. (M. Lévy "Le voleur d'ombres")*

***Pourquoi** s'était-il vu infliger ces «retrouvailles» avec son ex-femme? **Pourquoi** son fils enchaînait-il connerie sur connerie? **Pourquoi** sa fille de quinze ans se mettait-elle en tête de coucher avec des garçons? **Pourquoi** sa situation professionnelle menaçait-elle de s'effondrer (G. Musso "7 ans après")*

***MOSCOU** raccrocha et héla son chauffeur, la voiture arriva à sa hauteur, le garde du corps descendit lui ouvrir la portière. **MOSCOU** s'installa à l'arrière du véhicule qui repartit à vive allure.(M. Lévy "La dernière nuit ")*

В качестве *эпизевксиса* рассматривается повтор элементов в контактной позиции. Положение внутри предложения не оговаривается, однако в этом случае встает проблема разграничения эпизевксиса и анадиплозиса, для которого определены условия повтора на стыке частей предложения. В большей степени эпизевксис проявляется в начале предложения, повторяющиеся лексемы при этом чаще выражены либо глаголом в форме повелительного наклонения, либо существительным:

«***Merde, merde*** *!» Frank le saisit par les épaules et le serra fortement (M. Lévy "Et si c'était vrai ")*

*– **Allons, allons**, reprit celui qui se faisait appeler Giovanni, et l'hospitalité africaine, qu'en faites-vous? (M. Lévy "La première nuit").*

Анадиплозис в качестве повтора на стыке предложений употребляется с большей частотой, чем повтор на стыке частей предложения. При этом наиболее распространенной частью речи, в функции которой употреблено данное средство, является прилагательное или существительное:

*Elle captiva l'assemblée dès le début de son **exposé. Exposé** n'est pas le bon mot, c'était une histoire qu'elle nous racontait. (M. Lévy "Le premier jour")*

*Je n'ai rien dit, parce que j'étais **furieux, furieux** de l'avoir entraînée dans cette histoire, furieux de me sentir coupable de la perte de son travail, et incapable de l'éloigner des dangers que je pressentais. (M. Lévy Le premier jour).* В данном примере употребление анадиплозиса сопровождается диакопой - повтором лексемы *furieux* в качестве однородных членов в функции именной части составного именного сказуемого.

*– Vous n'êtes pas **contente**?*

***Contente**, le mot est faible. (G. Legardinier "Demain j'arrête")*

Последний пример демонстрирует частотное употребление анадиплозиса на стыке реплик персонажей (вопрос-ответ). Повтор слова здесь машинален, выступает как ответ на вопрос, который вызывает затруднение или сильную положительную или отрицательную эмоцию. Это в большом количестве случаев может быть переспрос.

Эпаналепсис (повтор слов и сочетаний в начале и в конце предложения по типу кольца) и *симплока* (повтор слов и сочетаний в начале и в конце двух смежных предложений, сочетание анафоры и эпифоры) являются довольно редкими средствами, поскольку в большей степени присущи поэтическим текстам. В составе исследуемых текстов в несколько большем количестве эти средства были обнаружены у Г. Мюссо:

*– **Bien sûr** je te ferai **mal.***

*– **Bien sûr** tu me feras **mal** (G. Musso "Je viendrai te chercher*) -симплока здесь сочетается с синтаксическим параллелизмом.

***Accepte** sa proposition, bon sang, **accepte** (G. Musso "Je viendrai te chercher");*

***Elliott** soignait ses malades à l'hôpital; Ilena s'inquiétait pour ses orques; Matt n'avait pas revu Tiffany, mais travaillait activement au démarrage de l'exploitation viticole achetée avec **Elliott** (G. Musso "Seras-tu là?")*

Таким образом, основываясь на проведенном анализе, можно определить конкретные тенденции в употреблении некоторых ритмических средств в современных французских текстах.

Во-первых, наиболее распространенным средством является диакопа (простой повтор через небольшие промежутки текста в рамках предложения). Преобладает в текстах Ж. Легардинье. Со смысловой точки зрения определяет повтор как возврат к определенной навязчивой мысли или эмоции.

Во-вторых, анафора, эпифора и полисиндетон, хоть и распространены в гораздо меньшей степени, но по сравнению с остальными средствами (анадиплозисом, эпаналепсисом, эпизевксисом и симплокой) употребляются довольно часто. В качестве повторяющихся элементов выступают имена существительные и прилагательные, часто имена собственные, вопросительные слова.

В-третьих, при употреблении анадиплозиса (повтора на стыке предложений или частей предложений) в диалогической речи часто используются конструкции, напоминающие мимезис, что объясняется выражением недопонимания или таких эмоций, как возмущение, негодование, удивление, реже радость.

В-четвертых, количество повторяющихся элементов для всех исследуемых средств за исключением анафоры (4-5 повторов) и полисиндетона (его употребление отслеживается для случаев с не менее, чем тремя повторами), сводится к одному, то есть имеется основной элемент и один его повтор.

В-пятых, в текстах редким является сочетание разных ритмических средств в рамках одного предложения, что обусловлено, в первую очередь небольшим объемом предложений, степенью их распространенности, а также довольно ограниченным на данный момент количеством исследуемых средств (восемь).

🕮 Вопросы и задания

1. Выберите фрагмент художественного текста, проведите его предварительную обработку, затем получите количественные данные по ритмическим средствам. Оформите статистику в виде таблицы.
2. Сопоставьте полученные результаты с бумажным текстом того же произведения. Определите процент, составляющий погрешность инструмента.
3. Чем с языковой и с технической точек зрения можно объяснить большое количество употреблений диакопы и полисиндетона?
4. Проанализируйте примеры употребления любого средства на выбор. Как повтор связан с сюжетно-образной структурой фрагмента или произведения?
5. Выберите для анализа поэтический текст. Определите наиболее частотные ритмические средства при помощи инструмента.
6. Определите связь ритмической структуры стихотворения и его содержания.
7. Выберите для анализа тексты других стилей, например, публицистического или научного. Проследите характер употребления ритмических средств и сравните с художественным.

§2 Анализ ритмической структуры текста и его перевода

Современная проблематика исследования ритма в переводе

Необходимость передачи ритмических характеристик при переводе текстов прозы – один из актуальнейших вопросов современной переводческой практики.

Текст, коммуникативная единица высшего уровня, обладает качеством смысловой завершенности как цельное литературное произведение, то есть законченное ***информационное*** и ***структурное*** целое. Агрегируя в себе разные виды информации (когнитивную, оперативную, эмоциональную, эстетическую и т.д.) [Алекссева, 2004], текст использует широкий репертуар языковых средств для их реализации. Ритмичность текста, во многом достигающуюся за счет апелляции к фигурам ритма, основанным на повторе, следует относить к области ***эмоционально-эстетической информации***, наряду с другими средствами образности (метафорой, метонимией, оксюмороном, средствами комического и др.). Передача всех видов информации, заложенной в тексте, тем более образующих его доминанту, является залогом успешного (в оптимальном сочетании эквивалентного и прагматически-ориентированного, адекватного) перевода. Воссоздание ритмического рисунка текста приобретает первостепенную важность при переводе ***примарно-эмоциональных*** и ***примарно-эстетических текстов*** [Алекссева, 2004], иными словами, текстов, специализирующихся на передаче ***эмоционально-эстетической информации – публицистических текстов*** с выраженной функцией воздействия, ***текстов художественной публицистики, собственно художественных текстов, поэтических текстов, текстов рекламы.***

Достоверное воссоздание ритма в переводе – одна из первостепенных задач для переводчика, хотя механизм влияния внешнего ритма на наше восприятие не вполне ясен. Существует теория, что «способность чутко реагировать на ритм закрепилась в ходе эволюции: прилив энергии, вызванный услышанным ритмом, и желание «пуститься в пляс» на самом деле являются рудиментами далекой эпохи, когда, заслышав топот ног соплеменников, наши предки должны были либо бежать от опасности вместе с ними, либо драться» [Мурзина, 2012]. Более современные исследования указывают на то, что восприятие ритма, в том числе ритма текста, слушателем является частным случаем фундаментального природного явления, называемого ***синхронизацией***. Мозг человека можно в определенном смысле моделировать сетью связанных и частично синхронизированных очагов активности, осцилляторов. Накопленный экспериментальный материал позволяет предположить, что спектр частот синхронизации влияет на способность мозга воспринимать и обрабатывать получаемую информацию [Мурзина, 2012].

Проблема передачи ***ритма художественного (прозаического) текста*** становится центральной для современных работ по проблеме ритмизации текста и особенностям межъязыковой трансляции ритма.

В художественном тексте все значимо, в том числе и ритм. Прозаический ритм заключает в себе множество смыслов, которые чувствуют писатели и читатели; они замечают его изменчивость, плавные и резкие переходы, которые вместе с лексикой и грамматикой создают сложный механизм эмоционального и эстетического воздействия [Иванова-Лукьянова, 2011: 302].

С точки зрения М.М. Бахтина, в художественном произведении всегда есть ритм, и это – ритм творца: «эстетический объект – это творение, включающее в себя творца» и «автор, как конститутивный момент формы, есть организованная, изнутри исходящая активность цельного человека, он нужен весь – дышащий (ритм), движущийся, видящий, слышащий, помнящий и понимающий»; «ритм, прикрепленный к материалу, выносится за его пределы и начинает проникать собою содержание как творческое отношение к нему, переводит его в новый ценностный план – эстетического бытия». Ритм обладает «формирующей силой»: «Из этого фокуса чувствуемой активности порождения, прежде всего, пробивается ритм (в самом широком смысле слова – стихотворный и прозаический) и вообще всякий порядок высказывания не предметного характера, порядок, возвращающий высказывающего к себе самому, к своему действующему, порождающему единству». М.М. Бахтин указывает: ритм «как форма упорядочения звукового материала, эмпирически воспринятого, слышимого и познаваемого, – ритм композиционен; эмоционально направленный, отнесенный к ценности внутреннего стремления и напряжения, которую он завершает, ритм архитектоничен» [Голубева-Монаткина, 2016:107-108].

На современном этапе преобладают частные исследования ритма в переводе, из чего следует, что комплексное изучение принципов и приемов передачи в переводе надречевых, надречево-

речевых и речевых форм ритма (композиционный ритм, ритм смыслов, ритм «языкового материала») является очень трудоемким. Среди причин отсутствия интегрированных исследований, посвященных ритму в художественном переводе, необходимо назвать чрезвычайную комплексность самого объекта исследования – ритма прозы, сложность и многоаспектность феномена художественного перевода, а также ограниченность математического инструментария и отсутствие эффективных автоматизированных систем оценки разноуровневых параметров ритма оригинального и переведенного текстов [Идиостиль и ритм текста 2019: 53].

Остановимся на некоторых исследованиях, дающих представление о тематике и проблематике исследования ритма в переводе.

Книговеды рассматривают книгу как «специфическую пространственную систему со своими законами ее внутренней организации», подчеркивают, что «книжное пространство не статично, а определенным образом организует наше движение в нем – перемещение взгляда и внимания читателя в процессе чтения» [Герчук, 1984: 10]. Плоскость страницы или разворота – это «особое двухмерное пространство», которое «строго организовано, построено»; у страницы есть «четкая система координат, образованная горизонталями строк и вертикалями столбцов […]. Тем самым это уже не просто бесформенная и бесконечная гладь, но некая двухмерная среда, на которой наш глаз ориентируется уверенно, отталкиваясь от зафиксированных на ней опорных линий и точек отсчета» [Герчук, 1984: 39].

Н.И. Голубева-Монаткина отмечает существование ритма самой книги: «Ритм в книге – форма организации ощущений, получаемых в процессе чтения. Эти ощущения приходят в динамике, в движении. Они связаны с воздействием на читателя не только смысла текста, его логики, понятий, образов, но и формы связей всех элементов книги. Верно трактованная в художественно-конструктивном плане книга обеспечивает управляемость ее восприятия читателем […]. Мы можем говорить о ритме книги в целом, когда рассматриваем закономерное чередование книжных элементов, о ритме на странице и развороте, когда рассматриваем закономерное чередование составляющих их элементов, о ритме в слове, строке, предложении и абзаце, где отмечаем форму и связь буквенных знаков и пробелов, и о ритме в иллюстрациях» [Голубева-Монаткина, 2016: 110].

Н.Н. Миронова обобщает методы исследования художественного текста, применение которых на этапе его предпереводческого анализа позволяет выявить и передать при переводе индивидуальные черты авторского стиля, в том числе средства ритмизации прозы: герменевтика, или глубокое проникновение в интенции автора; лингвистическая поэтика, т.е. структурный анализ текста (грамматика повествования, анализ языковых знаков); дискурс-анализ, или процесс структурирования действительности; критический дискурс-анализ – отражение в тексте общественных противоречий; эстетика рецепции, т.е. реконструкция восприятия текста читателем; интертекстуальность, или деконструкция семантики в текстах [Миронова, 2013: 80].

Л.В. Татару исследует позиционную структуру текста, фиксирующую его композиционно-симметрическую модель и создающую основу для восприятия композиционного ритма и его передаче при переводе. Автор предлагает модель анализа нарративного текста, включающую следующие уровни: анализ акцентно «выдвигаемой» семантической информации, располагающейся в начале текста, позволяющей уловить его первичный смысл и вводящей первичные пространственно-временные ориентиры; постепенное «погружение в текст» посредством пошаговой интерпретации новой информации в линейно следующих контекстах, создаваемой приращением новых (ритмообразующих) смыслов к уже известным значениям повторяемых (метрообразующих) единиц; концептуальный анализ ключевых слов в симметричных, позиционно сильных фрагментах текста (начальном и конечном), в их динамике; установление субъекта восприятия и речи (фокализатора) как коммуникативной позиции, выбираемой автором для выдвижения ключевых концептов. По мнению Л.В. Татару, такое целостное освоение информации о фрагменте мира, представленном в конкретном тексте, позволяет исследователю увидеть взаимозависимости между разными языковыми планами представления точки зрения и композиционной «логикой повествования», определяющей формирование глобальной ментальной репрезентации мира истории [Татару, 2008: 31].

Л.А. Мурзина, понимая под ритмом повторяющуюся вариацию длины и акцента в серии произносимых звуков, рассчитывает степень совпадения энергетического спектра на основе ударных и неударных слогов в поэтическом тексте на материале украинского и русского языков. Для расчета автором было выбрано представление ритмического сигнала строки стиха в виде последовательностей нулей и единиц: нули соответствуют безударным гласным (слогам), единицы – ударным

гласным (слогам). Энергетический спектр ритмического сигнала из N ударных/безударных слогов $xj = \{0, 1\}$ рассчитывается по стандартным формулам Фурье-преобразования [Мурзина, 2012].

Безусловно, при передаче ритмической структуры текста-оригинала структура и специфика обоих языков играют очень важную роль. Так, Т. Паркс указывает на то, что английский язык характеризуется большим количеством односложных слов по сравнению с другими европейскими языками, что вероятно также сказывается на переводе, в котором нарушается или видоизменяется силлабическая структура конструкции [Parks, 1998].

В работе Х. Пекканена ритм в переводе рассматривается на уровне соответствия (совпадения или несовпадения) синтаксических конструкций, позиции главного предложения по отношению к придаточному, передачи грамматических форм сказуемого [Pekkanen, 2014].

Н.И. Голубева-Монаткина акцентирует внимание на необходимости точного воспроизведения при переводе длины абзацев, а также зеркального отражения в переводе их начальных предложений («абзацных» фраз), если они являются воспроизводимыми и, следовательно, маркирующими авторский стиль (например, имена персонажей или даты в романе Гарсиа Маркеса «Сто лет одиночества») [Голубева-Монаткина, 2015: 62].

К. Скотт утверждает о важности для сохранения ритма текста передачи при переводе любого визуального, графического элемента: помимо метрической структуры при переводе современного текста должны быть переданы также «искусство шрифта и верстки, поэтика страницы» [Scott, 2002: 209-38].

Таким образом, частные исследования, посвященные межъязыковой трансляции ритма, выделяют проблему передачи ритма на уровне композиции и архитектоники текста, абзацирования текста, передачи ритма «языкового материала», в частности зеркального отражения синтаксических конструкций и силлабического ритма в переведенном тексте.

Ритм в переводе: проблема эквивалентности и коммуникативной ситуации

Постановка проблемы достижения эквивалентности в художественном переводе вообще и переводе текстовых фрагментов, содержащих фигуры ритма, в частности, требует обращения к понятию собственно переводческой эквивалентности.

В условиях существования широчайшей палитры текстовых типов и жанров определить понятие переводческой эквивалентности весьма затруднительно. Некогда постулированный тезис «об исчерпывающей передаче содержания оригинала не находит подтверждения в наблюдаемых фактах и его сторонники вынуждены прибегать к многочисленным оговоркам, которые фактически выхолащивают исходное определение» [Комиссаров, 2011]

Что же тогда следует понимать под эквивалентностью? Попробуем представить диахронический портрет данного понятия.

Одним из ключевых выводов отечественных теоретиков в области перевода стал вывод о необходимости различения ***потенциально достижимой эквивалентности***, под которой понималась максимальная общность содержания двух разноязычных текстов, допускаемая различиями языков, на которых созданы эти тексты, и ***переводческой эквивалентности*** – реальной смысловой близости текстов оригинала и перевода, достигаемой переводчиком в процессе перевода. ***Пределом переводческой эквивалентности*** было принято считать максимально возможную (лингвистическую) степень сохранения содержания оригинала при переводе, где в каждом отдельном переводе смысловая близость к оригиналу в разной степени и разными способами приближалась бы к максимальной [Комиссаров, 1990].

В связи с этим укажем на выделение понятия ***частичная эквивалентность***, когда неизменным для оригинала и перевода остается план содержания, при этом план выражения претерпевает изменения и даже потери [Ивлева, 2017: 71-77].

Еще один подход к решению проблемы переводческой эквивалентности заключается в попытке обнаружить в содержании оригинала какую-то инвариантную часть, сохранение которой необходимо и достаточно для достижения эквивалентности перевода. Наиболее часто на роль такого инварианта предлагается либо функция текста оригинала, либо описываемая в этом тексте ситуация. Иными словами, если перевод может выполнить ту же функцию или описывает ту же самую реальность, то он эквивалентен подлиннику [Ивлева, 2017: 71-77]. Однако какая бы часть содержания оригинала ни избиралась в качестве основы для достижения эквивалентности, всегда обнаруживается множество реально выполненных и обеспечивающих межъязыковую коммуникацию переводов,

в которых данная часть исходной информации не сохранена. И, наоборот, существуют переводы, где она сохранена, однако они не способны выполнить свою функцию в качестве эквивалентных оригиналу. В таких случаях мы оказываемся перед неприятным выбором: либо отказать подобным переводам в праве быть переводами, либо признать, что инвариантность данной части содержания не является обязательным признаком перевода.

Следующий подход к определению переводческой эквивалентности можно назвать эмпирическим. Суть его заключается в том, чтобы сопоставить большое число реально выполненных переводов с их оригиналами и выяснить, на чем основывается их эквивалентность. Проделав такой эксперимент, мы неизбежно придем к выводу, что степень смысловой близости к оригиналу у разных переводов неодинакова, и их эквивалентность основывается на сохранении разных частей содержания оригинала. Отсюда произрастают многоуровневые теории эквивалентности (Г. Егер, А.Д. Швейцер, В.Н. Комиссаров).

Существует подход, определяющий эквивалентность через классификацию речевых функций. Здесь цель коммуникации может быть истолкована как часть содержания высказывания, выражающая доминантную функцию самого высказывания. Сохранение цели коммуникации оказывается необходимым и достаточным условием эквивалентности перевода [Ивлева, 2017: 71-77].

Постепенно внимание переводоведов под воздействием скопос-теории, пражского структурализма, манипулятивной школы перевода смещается в область поиска межъязыковых различий. В центре внимания оказывается переводчик, именно ему «вменяется в обязанность преодолевать различия между оригиналом и переводом» [Ивлева, 2017: 71-77].

Эволюция подходов обнаружила движение к осознанию зависимости асимметрии языков от асимметрии культур. Автор оригинала утрачивает монополистическую роль в формировании смысла текста, читатель активно вовлечен в поле коммуникации, «оригинал возрождается контингентным переводом» [Ивлева, 2017: 71-77]. Переводчик играет чрезвычайно важную роль. Именно правильная оценка переводчиком коммуникативной ситуации, ее параметров, выработка стратегии перевода, в рамках которой применяются релевантные ситуации тактики, позволяют переводчику создать текст, эквивалентность которого является динамичной, изменчивой во времени, подчиненной прагматике текста оригинала, культурному фону реципиента и его информационно-прагматическим ожиданиям, то есть фактически вбирает в себя феномен адекватности и прагматической ориентированности.

Закономерно поднимается вопрос о необходимости разграничения эквивалентности в письменном и устном переводе. Затрагиваются вопросы репрезентации чужих культур с помощью перевода, конкуренции властей при воссоздании знаний и смыслов в отношении господствующих и подчиненных культур. Постколониальная теория позволила отказаться от европоцентристского понимания перевода и пересмотреть биполярную модель мышления. Логическим продолжением этих процессов стала утрата центральной позиции эквивалентностью. Теперь уже не эквивалентность между текстами и их элементами, а творческое различие между текстовыми и культурными элементами занимает доминирующее положение в культурном сознании индивида [Ивлева, 2017: 71-77]. Иными словами поиск переводческого решения определяется не эквивалентностью, а коммуникативной ситуацией. Переводчик является полноправным участником коммуникации, сотворцом культурного пространства, в котором осуществляется перевод.

Значит ли это, что отказ от категории эквивалентности целесообразен? Эквивалентность как инструмент оценки перевода, вероятно, все же имеет полезность. Речь, по всей видимости, должна идти о пересмотре границ и объема понятия, о включении «плавающих маркеров» эквивалентности, зависящих от прагматики оригинала, его предназначенности и пригодности для перевода, коммуникативной ситуации, в которой протекает перевод, от категории адресата переведенного текста, наконец, от языковой личности переводчика.

И. Левый, Л. Венути и представители этики идентичности в переводе считают эквивалентность центральным понятием, говоря о необходимости сохранения и миметическом воссоздании авторской и культурной подписи исходного текста. Постструктуралистская этика идентичности преследует не только цель сохранения самобытности чужой культуры, но и цель ознакомления реципиентов принимающей культуры с опытом этой чужой культуры и их последующего диалога [Ивлева, 2017: 71-77]. Именно этика идентичности позволяет принимающей культуре обогатиться за счет знакомства с другой культурой. Подход к эквивалентности в рамках этики идентичности имеет особую значимость именно для художественного перевода – благодаря ему внедряется «другость» чужой культуры [Ивлева, 2017: 71-77].

Принимая во внимание тот факт, что ритмический рисунок текста определенно является маркером авторского стиля, а в отдельных случаях может быть отнесен и к культурно-речевым традициям носителей разных языков, подход к толкованию эквивалентности в рамках этики идентичности представляется весьма продуктивным. Прагматический аспект перевода и участие переводчика в формировании смыслового и культурного пространства текста вновь заставляет обратиться к идее динамических или «плавающих» параметров эквивалентности.

В качестве системы координат при оценке воспроизводства ритма в художественном переводе, таким образом, может быть принята система (теория) динамической или функциональной эквивалентности, предполагающая установление эквивалентности исходного и переведенного текстов на основании равенства (в широком смысле) впечатлений их реципиентов.

Автор теории динамической эквивалентности – выдающийся представитель американской школы структуральной лингвистики Юджин Найда – считает задачей перевода создание на языке перевода «наиболее близкого ***естественного*** эквивалента» («the closest natural equivalent») тексту оригинала. Именно динамическая эквивалентность должна, по мнению Ю. Найды, обеспечить выполнение главной функции перевода – полноценной коммуникативной замены текста оригинала. При этом ориентация на реципиента неизбежно приобретает самодовлеющее значение. Наиболее четко эта ориентация выражена в часто повторяемом тезисе: традиционный вопрос – «верен ли перевод?» — нуждается в уточнении — «верен для кого?». Ориентированность перевода на коммуникативные ожидания реципиента вызывает к жизни идею о необходимости существенной культурной адаптации текста при переводе. Предполагается, что необходимое воздействие на реципиента можно обеспечить лишь при условии, если текст перевода не будет содержать чуждых для него культурно-этнических фактов или основанных на таких фактах образов или ассоциаций.

Это коррелирует с системой (категорией) адекватности (И.А. Афонина), предусматривающей учет цели и задач перевода, а также «выравнивание» прагматического намерения отправителя (автора) исходного текста и когнитивных ожиданий реципиентов текста на языке перевода. Р. Динг соглашается с приведенными тезисами, попутно отмечая необходимость комплексного восприятия и передачи ритма, включающего «рифму, грамматические повторы и риторические фигуры» [Ding, 2008].

Отдельное место в таких исследованиях занимает вопрос применения стратегии (в частных случаях, приема) прагматической и социокультурной адаптации при передаче ритма в переводе, например, за счет отказа от использования средств распевности и особого сказового порядка слов при переводе русских фольклорных сказок на французский язык (Н.А. Фененко).

Проблема передачи ритма текста при переводе, в особенности ритма, основанного на повторе, требует обращения к еще одному понятию теории Ю. Найды – понятию «информационной нагрузки». Как известно, надежность приема, то есть понимания сообщения реципиентом (слушающим или читающим) обеспечивается благодаря избыточности речи. Трудность понимания текста зависит от количества информации, измеряемого степенью неопределенности, неожиданности появления в тексте новых элементов. Увеличение количества информации путем использования в сообщении редких слов, необычного синтаксиса и других элементов, непривычных для реципиента (то есть обладающих большой неопределенностью), затрудняет прием или «декодирование» полученного сообщения. Поскольку перед переводчиком стоит задача обеспечить для реципиента надежный прием сообщения, содержащегося в исходном тексте, возникает проблема сохранения в переводе достаточной степени избыточности и легкости декодирования. Такая избыточность может относиться как к семантической составляющей высказывания (повтор смыслов), так и к формально-языковой (***повтор языковых знаков*** любого уровня системы). В этом отношении в задачи переводчика входит анализ степени избыточности, достаточной для легкого считывания сообщения, и трансформирование текста перевода таким образом, чтобы избыточность оказалась «сбалансированной» по сравнению с текстом оригинала. В том случае, если информационная нагрузка в переводе оказывается недостаточной в силу отсутствия у реципиента необходимых фоновых знаний (например, при передаче реалий, фразеологических оборотов, интертекстуализмов и других культурно-маркированных единиц) или при объективном расхождении грамматических моделей исходного и переводящего языка (например, избыточность маркеров грамматических категорий лица и числа выше в русском языке по сравнению с английским – см. анализ причин утраты анафоры в англо-русском переводе) переводчику необходимо «усилить» избыточность путем добавления информации (как на глубинно-смысловом, так и на поверхностно-вербальном уровне). Напротив, если информационная нагрузка

в переводе оказывается чрезмерной, переводчик регулирует ее за счет опущения информации или выбора альтернативных языковых средств с меньшей информационной нагрузкой.

Анализ ритмической структуры текста романа Айрис Мердок «The Black Prince» и его перевода на русский язык с применением автоматизированного инструмента ProseRhythmDetector

Прежде чем переходить к анализу ритмической структуры художественного текста, рассмотрим текст романа как продукт речемыслительной деятельности автора и выделим существенные признаки авторского стиля, создающие своего рода идиостилистический «генотип» Айрис Мердок.

Роман «Черный принц» – это вариация традиционной повествовательной формы – мистификации с «найденной рукописью», приближенная к форме «романа о романе». Рукопись публикуется после смерти ее автора (Брэдли Пирсона) и сопровождается предисловием и послесловием его друга, издателя Локсия, и четырьмя «постскриптумами».

Название книги многозначно. Оно ассоциируется и с образом Гамлета как великим символом подлинного искусства, и с безумной, но возвышенной страстью – «черным эросом», и с платоновским толкованием аполлоновского начала как сочетания любви и творческого озарения [Михальская, 1982: 156].

Композиционно-стилистическую организацию романа можно назвать полифонической. Основу организации повествовательной речи составляют авторская речь во всех ее разнообразных разновидностях и стилизованная речь, также в ее многочисленных формах.

Стилистика романа «Черный принц» сложна и многомерна. К образным доминантам идиостиля следует отнести авторские тропы и интертекстуальность, проявляющую себя через аллюзии и реминисценции из античной и современной западной философии, классической литературы и произведений постмодерна. Однако наиболее значимые интертекстуальные связи роман «Черный принц» образует с трагедией «Гамлет» Шекспира.

Выше мы указывали, что «Черный принц» – это роман в романе. Существует автор произведения – сама Айрис Мердок, которая создает Брэдли Пирсона – художника, написавшего книгу об одной «истории своей жизни». Но эта «история» творчески переосмыслена Брэдли-творцом; это события, которые Пирсон преподносит читателю через призму шекспировской трагедии [Филотенкова, 2013].

Брэдли Пирсон – человек, от чьего лица ведется повествование, немолодой интеллектуал – в конце своей жизни, сидя в тюрьме за преступление, виновника которого мы так и не узнаем (возможно, это сам Брэдли), создает произведение искусства. Развязка романа неоднозначна: убийство друга и соперника Брэдли – Арнольда Баффина – могло быть осуществлено как самим Брэдли, так и женой Арнольда, Рейчел [Филотенкова, 2013].

Если выдвинуть гипотезу о том, что виновен Брэдли Пирсон, то перед нами сложится схема семьи шекспировского Гамлета, но в своей трактовке Айрис Мердок переносит акценты с одних действующих лиц на другие. Теперь перед нами трагедия «Гамлет» глазами Клавдия, как если бы он жил в 20 веке и был художником, человеком по имени Брэдли Пирсон, которому суждено написать выдающееся objet d'art. Но если пьеса Шекспира начинается с трагического убийства короля, то роман Мердок этим убийством скорее завершается. Таким образом, роман «Черный принц» может восприниматься как предыстория сюжета легенды о принце Датском. Айрис Мердок рисует трагедию Клавдия, породившую, в свою очередь, трагедию Гамлета [Филотенкова, 2013].

Однако прочесть образ Брэдли Пирсона можно и иначе, допустив игру в перевоплощения, которую Айрис Мердок использует в качестве литературного приема и в других романах. В конце своей истории Брэдли Пирсон признается обществом сумасшедшим. Таким образом, к завершению романа он как будто сам превращается в Гамлета. Им обоим трудно смириться с тем обществом, в котором они жили. Их духу слишком тесно в этом мире. «Весь мир тюрьма» — говорит Шекспир, и Айрис Мердок вторит ему своей историей о Брэдли Пирсоне. Отсюда образ тюрьмы, который стал для Пирсона образом новой, неизведанной жизни. Его «наконец-то ждал его собственный, достаточно увесистый крест, и на нем значилось его имя», Брэдли Пирсон обрел свой путь, и только теперь жизнь для него стала полной [Филотенкова, 2013].

Особым признаком идиостиля являются и рекуррентные средства ритмизации текста, основанные на повторе, – эпаналепсис (3340 случаев), анафора (585 случаев), многосоюзие (219 случаев), эпифора (149 случаев), эпистрофа (78 случаев), анадиплозис (37 случаев), редупликация (23 случая), симплока (9 случаев) и др.

Вышеназванные средства ритмизации были выявлены с помощью компьютерной программы ProseRhythmDetector (описание инструмента см. выше). В ходе экспериментальной работы была, во-первых, установлена частотность появления и достоверность ритмических средств в оригинальном тексте романа, однако основные усилия были сосредоточены на анализе переноса средств ритмизации в текст перевода романа на русский язык, выполненный И. Бернштейн и А. Поливановой.

Порядок описания и интерпретации результатов сравнительно-сопоставительного исследования приведены ниже.

Анафора

1) Из 585 контекстов оригинальной анафоры 102 контекста (17,5%) воспроизведены с большей или меньшей количественной и позиционной точностью, а именно:

а. В 45 контекстах (7,7%) количество и позиция повторяющихся элементов соответствуют полностью, например, I have never tried to please at the expense of truth. I have known, for long periods, the torture of a life without self-expression. – Я никогда не стремился к приятности за счет правды. Я знал долгие мучительные полосы жизни без самовыражения;

б. В 22 контекстах (3,8%) позиция анафор сохранена, количество повторяющихся элементов меньше на 1-4 элемента из 4-6, например I waited. I tried to develop a new routine: monotony, out of which value springs. I waited, I listened. – Я ждал. Я снова постарался выработать упорядоченный образ жизни, создать монотонность, из которой рождаются всплески. Я выжидал, вслушивался;

в. В 31 контексте (5,3%) незначительно изменена позиция анафоры (перенос на 1 предложение вперед или назад) и в 18 из них изменено количество повторяющихся элементов, преимущественно в сторону их уменьшения, например, And I was suddenly deeply frightened by the possibility of having my sister on my hands. I simply did not love her enough to be of any use to her, and it seemed wiser to make this plain at once. I waited for about ten minutes, trying to calm and clear my mind, and then went back to the bedroom door. I did not really expect that Priscilla would have got dressed and be ready to leave. I did not know what to do. I felt fear and disgust at the idea of mental breakdown, the semi-deliberate refusal to go on organizing one's life which is regarded with such tolerance in these days. I peered into the room. – Я боялся, как бы сестра не оказалась вдруг у меня на руках. Я просто-напросто не настолько любил ее, чтобы она могла на меня рассчитывать, и, видимо, лучше всего было сказать ей об этом прямо. Переждав минут десять, пока успокоятся мои нервы и прояснится голова, я встал и подошел к двери в спальню. В сущности, я и не надеялся, что застану Присциллу одетой и готовой к уходу. Что мне делать, я не знал. Мне противна и страшна была сама мысль о «нервном расстройстве», этом наполовину сознательном уходе от упорядоченной жизни, к которому в наши дни принято относиться с такой терпимостью. Я заглянул в комнату; в 2 случаях в переводе число повторяющихся элементов больше, чем в оригинале (3 против 2 в оригинале);

г. 4 контекста (0,7%) содержат меньшее количество элементов, повторяющихся не подряд, из чего следует, что повтор перестает восприниматься как анафорический, а переходит в эпаналепсис, например, I went out of the room and closed the door quietly behind me. I heard a soft bound and then the key turning in the lock. I went down the stairs feeling very shaken and, yes, she had been right, disgusted. – Я спустился по лестнице с чувством растерянности и – да, она была права – отвращения. За это время стемнело, солнце больше не сияло на улице, и все в доме стало коричневым и холодным. Я вошел в гостиную, где сидели и беседовали Арнольд с Фрэнсисом.

2) В 483 контекстах оригинала (82,5%) анафорический повтор при переводе либо утрачен, либо заменен другим средством ритмизации, преимущественно эпаналепсисом, например, Arnold Baffin wrote too much, too fast. Arnold Baffin was really just a talented journalist. – Арнольд Баффин писал слишком много, слишком быстро. По существу, Арнольд Баффин был всего лишь талантливым журналистом.

3) Причины утраты анафоры:

а. Невозможность ее воспроизведения в силу объективного расхождения грамматических возможностей английского и русского языков (для проанализированных примеров это часто разница в употреблении залоговых форм, например, для русского языка исключена возможность употребления страдательного залога с непереходными глаголами, глаголами с предложным управлением, в то время как для английского языка это норма: I was upset and annoyed when my father once approached the subject, and although I could see that he had been made utterly miserable, I resolutely refused to discuss it. – Когда мой отец попытался однажды заговорить со мной о Присцилле, мне это было крайне неприятно, и я, хотя и видел, как он расстроен, решительно отказался обсуждать эту тему (соответственно, один из элементов анафорического повтора «I» теряется)).

б. Опасность нарушения узуса переводящего языка (He needed) – «ему был нужен», а не «он нуждался»);

в. Наличие формульных (клишированных) фраз (коллокаций) – I am 58 – Мне 58, I still dream about it at least once a week – Он и по сию пору снится мне примерно раз в неделю;

г. «Агрессивность» (назойливость) повтора в 5-6 элементов может приводить к нарушению стилистики речи. Кроме того, повтор местоимений-подлежащих (ими текст изобилует, особенно «I») является избыточным для русского языка, где значение грамматических категорий дублируется через спряжение глаголов, изменение корневых гласных, согласных, свойственное для местоимений определенных лиц и чисел и др. (I waited, I listened. – Я выжидал, вслушивался).

4) В оставшихся 244 контекстах перевода анафора содержится в предложениях (фразах), оригиналы которых анафоры не имеют. Данная переводческая стратегия может рассматриваться как попытка компенсировать утрату анафоры позиционно по сравнению с оригиналом.

Эпифора

1) На 149 случаев оригинальной эпифоры – 11 совпадений, найденных инструментом (7,4%).

2) В 24 контекстах (16,1%) оригинальная эпифора при переводе заменена другими видами повтора.

а. Самой частотной является замена эпифоры на эпаналепсис (13 случаев): You'd better get your own doctor tomorrow. Oh, I think I shall be better tomorrow. – Вам надо будет завтра самому вызвать доктора. – О, завтра, я думаю, мне будет лучше.

б. Далее следует замена на смысловой повтор (замена на синоним), иногда в сочетании с изменением порядка слов: – 5 случаев Roger has become a devil. Some sort of devil. – В Роджера просто дьявол вселился. Какой-то демон.

в. Обнаружено по 2 случая замены эпифоры на анадиплозис: I can't remember. Rachel, you must remember. – Не помню. – Вспомните, Рейчел.; градационный повтор: For the first time in my life I urgently wanted silence. […] I had always in a sense been a devotee of silence. – Но тогда я нуждался в самой настоящей, буквальной тишине; и сочетание лексического свертывания с парцелляцией: We're Jewish. At least we're partly Jewish. I don't mind your being Jewish. – Мы ведь евреи. Наполовину. – И на здоровье.

3) В переводе присутствуют контексты с эпифорой, которой нет в зеркальных оригинальных предложениях, что может рассматриваться как попытка частично компенсировать утрату эпифоры по сравнению с оригиналом.: «You must have thought Rachel and I were being ridiculously solemn this afternoon about very little.» «I see you're playing it differently now,» I said. - Так и знал, что вы появитесь здесь с этим торжествующим видом…– Откуда у меня может быть торжествующий вид? У меня нет никакого повода торжествовать.

4) Статистика воспроизведения (компенсации) эпифоры при переводе в соотношении 3/1 весьма прогнозируема. Утрата 2/3 контекстов с эпифорическим повтором ожидаема с точки зрения разницы в позиции ремы в английском и русском языках (например, безударные личные местоимения в конце предложения или фразы на английском языке, которые могут составлять эпифору, не несут смысловой нагрузки, в то время как финальное слово в предложении или фразе русского языка автоматически получает фразовое ударение и воспринимается как рема. Для сохранения правильного тема-рематического членения предложения необходимо прибегать к перестановке элементов предложения (фразы), что неизбежно ведет к утрате эпифоры; в значительном числе контекстов она компенсируется другими средствами ритмизации (в том числе другими видами повтора)).

Симплока

1) 1 из 9 случаев симплоки, найденных в оригинальном тексте (11,1%), имеет зеркальное совпадение в тексте перевода: I turned back and drove the other way, through the village, past the church. I even stopped and went into the church. - Я повернул назад и поехал через деревню мимо церкви. Я даже остановился и зашел в церковь. 3 случая симплоки, зафиксированные в тексте перевода, употреблены переводчиком самостоятельно (возможно, в порядке компенсации утраченных средств ритмизации).

2) Однако в переводе средства ритмизации не были утрачены, но подверглись стилистическим (ритмическим) преобразованиям. Таким образом, из 8 оставшихся случаев 5 раз симплока заменена на эпаналепсис (например, She kept drawing it out without telling me and buying clothes. She went mad over buying clothes. – А она втихомолку тянула фунт за фунтом и покупала себе тряпки. Она на этих тряпках просто помешалась – здесь следует отметить появление при переводе разговорного

«тряпки», что можно истолковать как добавочное средство эмфатизации при переводе), 2 раза заменена на эпифору (сочетание эпифоры и эпаналепсиса) (например, What did you say? What could I say? – А вы что ответили? – Что я могла ответить?), 1 раз сохранен синтаксический параллелизм (I write whether I feel like it or not. I complete things whether I think they're perfect or not. – Я пишу независимо от того, легко мне или трудно. И завершаю любую работу, удалась она мне или не удалась).

Анадиплозис

1) Из 37 случаев анадиплозиса 15 случаев (40,5%) идентифицированы некорректно, фактически смешаны с редупликацией: например, Only Bradley. Only Bradley. The voice, still almost inaudible, was firmer.

2) Из 22 случаев истинного анадиплозиса 2 (9%) точно воспроизведены в переводе, с точки зрения инструмента, например: Art is imagination. Imagination changes, fuses. – Искусство – это воображение! Воображение пресуществляет, плавит в своем горниле. «Ручной» поиск позволил выявить еще 1 дополнительный пример оригинального анадиплозиса, воспроизведенного в переводе: Priscilla suddenly started to scream quietly. «Scream quietly» may sound like an oxymoron, but I mean to indicate the curiously controlled rhythmic screaming which goes with a certain kind of hysterics. – Присцилла вдруг спокойно завыла. «Спокойно выть» – казалось бы, оксюморон, но этим термином я обозначил странно ритмичные, рассчитанные вопли, сопровождающие некоторые истерические состояния.

3) В приведенном выше контексте был также обнаружен пример анадиплозиса, отмеченный инструментом и представленный в переводе эпаналепсисом: …I mean to indicate the curiously controlled rhythmic screaming which goes with a certain kind of hysterics. Hysterics is terrifying because of its willed and yet not willed quality. – …этим термином я обозначил странно ритмичные, рассчитанные вопли, сопровождающие некоторые истерические состояния. Истерика пугает тем, что она произвольна и непроизвольна в одно и то же время.

4) В остальных 18 контекстах анадиплозис утрачен без какой-либо ритмической компенсации (фигур ритма) в составе данных предложений. Однако общее количество анадиплозисов в переведенном тексте (20 истинных – 4 полностью или частично соответствуют оригиналу, 16 употреблены переводчиком самостоятельно) позволяет говорить о равновесном употреблении данного ритмического средства в оригинале и переводе, пусть и в порядке позиционной компенсации.

Эпистрофа

В переведенном тексте инструмент не зафиксировал ни одного случая применения эпистрофы, поэтому сравнительная характеристика быть сделана не может.

Редупликация

Выявленные в оригинальном тексте случаи редупликации (23) являются достоверными. Совпадений при переводе – 2 (No, no – нет, нет; Well, well, well – хорошо, хорошо), другие 6 контекстов – это переданная редупликация, идентифицированная машиной как эпифора или анадиплозис.

Эпаналепсис

1) Сравнительно-сопоставительный анализ оригинальных и переведенных контекстов позволил сделать заключение о том, что эпаналепсис полностью (иногда с утратой 1-2 элементов, если количество повторяющихся элементов сравнительно велико, i.e. более 4) воспроизводится при переводе примерно в 25% случаев (834 контекста): I learnt later with abhorrence that he had set up in business as a self-styled psychoanalyst. Later still I heard he had taken to drink. Позже я с отвращением узнал, что он завел себе практику в качестве самозваного «психоаналитика». Еще позже я слышал, что он пьет.

2) Еще в 5% случаев (167 контекстов) эпаналепсис заменяется другим средством ритмизации, что, тем не менее, способствует сохранению ритмического рисунка авторского текста. В проанализированных примерах были выявлены случаи замены эпаналепсиса на смысловой повтор в сочетании с синтаксическим параллелизмом и позиционной компенсацией эпаналепсиса (например, He was stout (the raincoat failed to button) and not tall, with copious greyish longish frizzy hair and a round face and a slightly hooked nose and big very red lips and eyes set very close together. Он был толст (макинтош явно не застегивался), невысок ростом, волосы густые и курчавые, давно не стриженные, с проседью, лицо круглое, со слегка крючковатым носом, толстыми, очень красными губами и удивительно близко посаженными глазами.); замены на анадиплозис с позиционной компенсацией эпаналепсиса (например, Real bears, I believe, have eyes rather wide apart, but caricatured bears usually have close eyes, possibly to indicate bad temper or cunning. Он походил, как я потом подумал, на карикатурного медведя. Не на настоящего – у настоящих медведей глаза, по-моему, расставлены широко, а вот на карикатурах их рисуют с близко посаженными глазами – вероятно, для того, чтобы выразить их свирепость и коварство.).

3) В остальных случаях эпаналепсис при переводе утрачивается с последующей его компенсацией в предложениях, оригиналы которых эпаналепсиса в своем составе не имеют (чаще всего в составе этих предложений имеются другие ритмические фигуры – примеры см. выше, в отдельных случаях предложения не содержат повтора вообще). Среди основных причин утраты эпаналепсиса следует назвать невозможность с точки зрения узуса стопроцентного воспроизводства личного местоимения «I» в функции подлежащего, а также притяжательных местоимений, в особенности в препозиции к существительным, обозначающим части тела, предметы одежды, термины родства и др.

Многосоюзие (полисиндетон)

1) Полисиндетон вполне успешно воспроизводится при переводе: точная передача данного ритмического средства составила порядка 45% (98 контекстов). Следует отметить, что практически во всех переведенных контекстах появлялись дополнительные средства многосоюзия (ср. 219 примеров полисиндетона в оригинале и 366 в переводе) например, Both taxman and dentist only too readily image forth the deeper horrors of human life: that we must pay, perhaps ruinously, for our pleasures, that our resources are lent, not given, and that our most irreplaceable faculties decay even as they grow. – И зубной врач, и налоговый инспектор, естественно, символизируют для нас подспудные ужасы жизни; они говорят о том, что мы должны платить, даже если цена разорительная, за все наши удовольствия, что блага даются нам в долг, а не даруются, что наши самые невосполнимые богатства гниют уже в процессе роста.

2) В других контекстах полисиндетон компенсируется смысловым повтором (через употребление синонимичного союза) или утрачивается (см. выделение курсивом), например, The shop sold daily papers and magazines, *writing paper and so on*, and horrible gifts. – Они продавали газеты и журналы, *бумагу всевозможных сортов*, а также безобразные «подарки».

Апозиопеза

Обобщая приведенные результаты, констатируем, что наиболее высокую точность воспроизведения при переводе имеют диакопа, эпизевксис и полисиндетон (30-45%), наименьшую – эпифора и анадиплозис (7-9%). Среди причин «успешной» передачи фигур ритма при переводе следует назвать равные возможности английского и русского языков в формировании ритмического рисунка текста (внепозиционный повтор, редуплицирующий повтор). «Неудачи» в воспроизводстве отдельных типов ритмических средств, напротив, обусловлены объективными расхождениями в грамматической структуре английского и русского языков и более «агрессивным» поведением повтора в русском языке, обусловленным фоно-морфологическими причинами. Следует допустить и субъективное влияние творческого стиля переводчика. К основным стратегиям передачи фигур ритма следует отнести замену и позиционную компенсацию.

Обращаясь к теории динамической (функциональной) эквивалентности, констатируем, что формальное «статистическое» соответствие, которое обеспечило бы более высокий процент передачи повторяющихся элементов, стало бы губительным для текста перевода, как в смысловом отношении, так и в аспекте соблюдения узуса и «речевого баланса» на переводящем языке, что в свою очередь повлекло бы за собой коммуникативную неудачу, вызванную искажением авторского стиля (а, возможно, и авторского посыла) и нарушением коммуникативных ожиданий реципиента, носителя русского языка и русскоязычной коммуникативной культуры. Напротив, использованные переводчиком трансформации позволили создать более близкий ***естественный эквивалент*** тексту оригинала, сбалансировать информационную нагрузку и обеспечить выполнение главной функции перевода – полноценной коммуникативной замены текста оригинала. Еще раз подчеркнем и активную роль переводчика в анализе коммуникативной ситуации, в особенности анализе различий между текстовыми и культурными элементами на входе и выходе и повторим, что переводчик является полноправным участником коммуникации, сотворцом культурного и коммуникативного пространства, в котором осуществляется перевод.

🕮 Вопросы и задания.

1. Как классифицируются тексты с точки зрения доминирующей в них информации? В каких текстах проблема передачи ритма текста встает наиболее остро? Почему воссоздание ритма необходимо при переводе текстов прозы?
2. Какие исследования ритма в переводе преобладают на современном этапе и почему? Приведите примеры.

3. Как менялись представления об эквивалентности в истории переводоведения?

4. Соотнесите понятия коммуникативной ситуации, эквивалентности и адекватности перевода. Что такое переводческая прагматика?

5. Что принимается в качестве системы координат при оценке воспроизводства ритма в художественном переводе?

6. Кто автор теории динамической эквивалентности и в чем ее суть? Как теория динамической эквивалентности соотносится с теорией (категорией) адекватности перевода?

7. Что такое «информационная нагрузка»? Каким образом ритм текста связан с его информационной нагрузкой? Каковы общие стратегии в отношении сохранения информационной нагрузки в тексте перевода?

☑ Практические задания

1. Выберите оригинальный текст, относящийся к примарно-эмоциональному типу, на английском (французском / испанском) языке. Обработайте его с помощью инструмента **ProseRhythmDetector**. Произведите верификацию отобранных инструментом фигур ритма, основанных на повторе, составьте список фигур, обращая внимание на контекст.

2. Найдите перевод выбранного текста на русский язык. Обработайте его с помощью инструмента **ProseRhythmDetector**. Произведите верификацию отобранных инструментом фигур ритма, основанных на повторе, составьте список фигур, обращая внимание на контекст.

3. Соотнесите полученные списки и контексты, выведите статистику. Сделайте выводы относительно полученных результатов. Насколько полученная статистика совпадает с цифрами, приведенными в части 3 настоящего раздела? Какие выводы это позволяет сделать относительно возможности передачи фигур ритма, основанных на повторе, с иностранного на русский язык? Можно ли оценить влияние индивидуального стиля переводчика на принимаемые решения?

Глава V
ПРОВЕДЕНИЕ СТАТИСТИЧЕСКИХ ЭКСПЕРИМЕНТОВ С ИСПОЛЬЗОВАНИЕМ PRD

Чтобы получить значимые научные результаты в области компьютерной лингвистики, необходимо проводить эксперименты на значительном количестве текстов. Такой объём текстов невозможно обработать вручную за разумное время, поэтому подобную работу требуется автоматизировать. В данной главе описываются эксперименты, которые позволяют исследовать статистические/стилометрические характеристики текстов. Структура таких автоматизированных экспериментов обычно следующая:

1. Сбор корпуса текстов, на которых эксперименты будут проводиться.
2. Разметка корпуса текстов, т. е. указание для текстов лингвистических параметров, которые исследуются в экспериментах.
3. Оценка качества новой разметки.
4. Вычисление статистических характеристик текстов.
5. Визуализация статистических характеристик текстов для последующего анализа.

Рассмотрим реализацию этих шагов подробнее.

§1 Составление корпуса текстов

Корпус текстов может составляться как вручную, когда подходящие тексты выбирает и обрабатывает эксперт или группа экспертов, так и автоматически, когда корпус формируется с помощью специальной программы, находящей тексты и преобразующей их к заданному формату.

В настоящее время для научных исследований в области обработки естественного языка имеется немало корпусов текстов в открытом доступе. Их обзор для различных языков был дан во второй главе пособия. Они обычно имеют разметку, подходящую для решения конкретной задачи: разметку по темам, авторам, дате публикации и т. п. Чтобы найти корпус, можно обратиться к специализированному поисковому движку, например, Google Dataset Search (https://datasetsearch.research.google.com/), к сайту, который аккумулирует корпуса текстов (https://data.world/datasets/nlp, https://www.kaggle.com/tags/nlp) или ссылки на них (https://github.com/niderhoff/nlp-datasets). Кроме того, авторы научных статей в своих работах нередко дают ссылки на набор данных, который они использовали.

Если используется чужой корпус текстов, необходимо обратить внимание на лицензию, по которой он распространяется. Как правило, авторы корпуса разрешают использовать его для научных исследований, при этом они могут потребовать процитировать их статью. Если нужен корпус для коммерческой разработки, то круг возможных вариантов становится меньше: лицензии многих корпусов прямо запрещают их коммерческое использование.

Трудности с поиском подходящего корпуса текстов могут возникнуть при работе с национальными языками. Для английского языка разработано достаточно много корпусов, подходящих для различных задач анализа текстов на естественном языке, в то время как для национальных языков в открытом доступе находится гораздо меньше корпусов. Для наиболее популярных задач компьютерной лингвистики: определения авторства, тематической классификации, найти готовый корпус вполне возможно. Но если нужна специфическая разметка, например, ритмические средства внутри текстов, корпуса придётся собирать и размечать самостоятельно.

По структуре корпуса обычно состоят из текстов одного типа. Это могут быть художественные тексты, статьи, официальные документы или Интернет-тексты. Каждый из типов имеет свои специфические особенности.

Интернет-тексты – это, как правило, короткие документы, написанные в произвольном стиле, часто неформальным языком, например, блоги, отзывы, электронные письма, сообщения на форумах

и т. п. Примеры корпусов: корпус электронных писем Enron (https://www.cs.cmu.edu/~./enron/) и корпус отзывов на фильмы IMDb 62 (https://www.dropbox.com/s/np1u1hl343gd73m/imdb62.zip).

Статьи и официальные документы обычно больше по размеру, чем тексты из предыдущей категории, они написаны официальным языком и содержат больше терминологии. Это могут быть новостные, научные или журнальные статьи, юридические или медицинские документы и т. п. Примеры корпусов: новостные статьи Reuters-21578 (http://disi.unitn.eu/moschitti/corpora.htm), тексты судебных заседаний Judgement (https://umlt.infotech.monash.edu/?page_id=152).

Художественные тексты являются документами большого размера, написанными литературным языком. Очень известным в научной среде корпусом художественных текстов является проект Gutenberg [Welcome to Project Gutenberg].

Обычно тексты в корпусе имеют один и тот же формат представления. Это делается для удобства автоматической обработки, чтобы программа могла извлекать информацию из разных текстов с помощью одного алгоритма. Например, каждый текст может находиться в отдельном файле заданного формата (txt, html, json, epub, …), где каждый абзац располагается ровно на одной строке. Также файл может иметь название с чётко заданной структурой, например «год публикации-фамилия автора-название текста.txt». Кроме того, у каждого текста может быть разметка, то есть набор метаданных: ФИО автора, заголовок, дата публикации, ключевые слова, информация об издателе, тональность текста и т. п. Метаданные тоже следует хранить однотипным образом так, чтобы можно было однозначно определить, какой текст они описывают. Они могут быть вынесены в название файла, в отдельный файл с названием, похожим на название соответствующего текста, или содержаться в том же файле, что и текст (например, в виде json-атрибутов).

Лингвистическая разметка текста может быть достаточно разнообразной. Найденные ритмические средства также являются частью этой разметки. Информация о том, где именно в тексте находятся ритмические средства, может храниться разными способами: как непосредственно в тексте (например, при помощи разметки слов и предложений по цветам), так и отдельно от него. В частности, приложение ProseRhythmDetector хранит и текст, и его разметку по ритмическим средствам в общем файле с расширением. prd в json-формате.

Фрагмент prd-файла:

```
{
 "metadata": {
  "language": "ru"
 },
 "text": [
  "First paragraph",
  "Second paragraph",
  …
],
 "features": [
  {
   "type": "anaphora",
   "context": [51, 82],
   "words": [51, 52, 70, 71]
  },
  {
   "type": "anaphora",
   "context": [183, 219],
   "words": [183, 195]
  },
  …
 ]
}
```

Сам текст хранится в поле *text* как список абзацев (если описывать формат более точно, то и с разделением на главы, но это выходит за рамки учебника). Поле *metadata* содержит дополнительные

данные о тексте, в частности, язык, на котором текст написан. Поле *features* хранит лингвистическую разметку: список аспектов. Для каждого аспекта в *type* указывается средство (анафора – anaphora), *words* – номера повторяющихся слов, *context* – номера первого и последнего слова в контексте. Нумерация слов ведётся с 0 от начала текста.

Формат представления корпуса следует выбирать так, чтобы он был максимально удобен для использования. Если тексты будут обрабатываться только компьютерными программами, то предпочтительнее использовать удобные машиночитаемые форматы вроде json. Если же необходимо, чтобы тексты без предварительной обработки смотрел человек, следует выбрать формат, который удобен для стандартных программ (текстовых редакторов), например, txt или docx.

§2 Оценка качества разметки

Когда собственный корпус собран и размечен, качество этой разметки необходимо оценить. Это позволит определить, насколько хороши разработанные алгоритмы, если разметка выполнялась автоматически. Рассмотрим метрики качества более подробно.

Введём несколько обозначений. Пусть в тексте имеется N анафор, а алгоритм нашёл M анафор, причём среди них он нашёл M_t анафор правильно и M_f – неправильно. Отметим, что $M_t + M_f = M$. И пусть N_f – это количество анафор, не найденных алгоритмом. Тогда $M_t + N_f = N$ (анафора либо найдена алгоритмом, либо не найдена).

Для оценки качества алгоритмов, которые решают задачи обработки текстов, можно использовать следующие стандартные метрики:

точность (P, precision) – это доля правильно выбранных элементов среди всех результатов алгоритма, то есть $P = \frac{M_t}{M}$;

полнота (R, Recall) – это доля правильно выбранных элементов среди всех элементов, которые должны были быть выбраны, то есть $R = \frac{M_t}{N}$;

F-мера (F, F-measure) – это среднее гармоническое точности и полноты, $F = \frac{2PR}{P+R}$. Эта метрика вводится, чтобы сбалансировать точность и полноту и ввести один критерий качества вместо двух. В приведённой формуле точность и полнота учитываются одинаково. Такая F-мера называется сбалансированной или F_1. Если же одна метрика более значима, чем другая (например, полнота поиска важнее точности или наоборот), то можно в формулу ввести дополнительный коэффициент β. Формула для F-меры будет иметь следующий вид: $F = \frac{(\beta^2+1)2PR}{\beta^2 P+R}$. Если значения β берутся из диапазона $0 < \beta < 1$, то предпочтение отдаётся полноте. При β = 1 формула превращается в предыдущую. И если $\beta > 1$, то предпочтение отдаётся точности.

Значения всех трёх метрик находятся в диапазоне от 0 до 1. Чем ближе значение метрики к 1, тем лучше сработал алгоритм.

В дополнение к этим трём классическим метрикам можно посчитать ошибки алгоритма. Ошибки первого рода – это элементы, которые алгоритм нашёл неверно. Мы обозначили их количество как M_f. Ошибки второго рода – это элементы, которые алгоритм не нашёл, хотя должен был. Мы обозначили их количество как N_f.

Метрики для ошибок:

доля ошибок первого рода: $E_1 = \frac{M_f}{M} = 1 - P$;

доля ошибок второго рода: $E_2 = \frac{N_f}{N} = 1 - R$.

У этих метрик диапазон значений также лежит от 0 до 1, но, в отличие от предыдущих метрик, чем меньше их значение, тем лучше работает алгоритм, т. е. тем меньше ошибок совершает.

До сих пор разбирался случай, когда оценивается только одно найденное средство (анафора в примере). Как поступить в случае, когда средств несколько и нужно посчитать общую метрику?

Для точности, полноты и F-меры есть микро и макро-усреднения.

При микро-усреднении сначала считаются количества правильных и неправильных ответов для всех метрик (N, M, M_t для анафоры, эпифоры, диакопы и т. п.). Затем они суммируются, то есть находятся общее количество найденных средств, имеющихся средств и правильно найденных средств. Для этих чисел считаются единственные точность, полнота и F-мера по описанным формулам. Особенность этого подхода в том, что средства будут по-разному влиять на результат. В частности, диакопа, как одно из самых частотных средств, будет иметь самые большие значения N, M, M_t и внесёт наибольший вклад в результат. Тогда как эпизевкис, который может встречаться даже в большом тексте буквально несколько раз, практически не повлияет на итоговые значения метрик.

Когда необходимо оценивать качество поиска ритмических средств таким образом, чтобы каждое средство вносило равный вклад в итоговую оценку, следует использовать макро-усреднения. При этом точность и полнота рассчитываются для каждого средства отдельно, а затем итоговые точность и полнота считаются как средние арифметические значения по всем средствам. F-мера при этом может считаться и как среднее гармоническое итоговых точности и полноты, и как среднее арифметическое F-мер для каждого средства.

Конечно, список возможных метрик не ограничивается разобранными. Более подробно метрики качества можно изучить, прочитав классическую работу Соколовой и Лапалме [Sokolova, 2009: 427-437.].

Метрики оценки качества следует выбирать исходя из поставленной задачи: какую именно характеристику алгоритма требуется улучшить. С другой стороны, следует в первую очередь отдавать предпочтение стандартным метрикам, которые активно используются другими исследователями. Это позволит сравнить качество работы вашего алгоритма с уже имеющимися решениями.

§3 Статистические характеристики

Когда качественная разметка будет готова, на её основе можно вычислять стилометрические (статистические) характеристики, которые описывают ритм текста в целом и позволяют сравнивать тексты между собой. По стилометрическим характеристикам можно сопоставлять как отдельные тексты, например, тексты одного автора или переводы одного произведения, так и наборы текстов, например, литературные эпохи, века, десятилетия и т. п.

Стилометрические характеристики классифицируются на две крупные категории следующим образом:

Низкоуровневые характеристики описывают базовые элементы текста: слова и символы.

Высокоуровневые или лингвистические характеристики вычисляются при помощи знаний о структуре и особенностях языка.

Низкоуровневые характеристики в свою очередь делятся на два уровня:

Характеристики уровня символов, которые описывают отдельные символы в тексте: общее количество символов, доли конкретных символов среди всех символов текста, количества букв, знаков препинания, n-граммы букв и т. п.

Характеристики уровня слов, которые описывают лексику текста: общее количество слов, частоты встречаемости слов, n-граммы слов, богатство словаря, доля уникальных слов, количества слов заданной части речи и т. п.

Высокоуровневые характеристики более разнообразны:

- Синтаксические характеристики, которые описывают структуру предложений: количества простых и составных предложений, средние длины предложений в словах и символах, количества или доли полных и неполных предложений, предложений с прямой речью и т. п.
- Семантические характеристики, которые описывают семантические отношения между элементами текста, например, характеристики графа семантических отношений.
- Тематические характеристики, которые описывают тематическое содержание текста, например, ключевые слова и их статистические характеристики.
- Ритмические характеристики, которые описывают ритм текста, например, частоты встречаемости ритмических средств, ударных и безударных слогов и т. п.

В этом списке указаны только самые распространённые категории характеристик. Эту классификацию можно расширять и другими характеристиками стиля текста и автора. Чтобы изучить различные типы характеристик подробнее: насколько они распространены, как вычисляются и применяются, можно обратиться к обзору авторов учебника [Lagutina, 2019: 184–195] и научным работам, на которые авторы ссылаются.

Рассмотрим подробнее ритмические стилометрические характеристики. Они могут характеризовать как отдельные ритмические средства, так и их структуру.

Примеры ритмических характеристик:

- Общее количество ритмических средств в тексте.
- Количества отдельных ритмических средств в тексте: анафор, эпифор, аллитерации, ассонанса и т. д.
- Доли ритмических характеристик: количества конкретных средств, делённые на общее количество средств.
- «Плотность» средств – количество всех средств или конкретного средства, делённое на количество предложений в тексте.
- Количества конкретных частей речи, появляющихся в аспектах средств: существительных, прилагательных и т. д.
- Доли частей речи среди слов из аспектов, то есть количества конкретных частей речи среди слов из аспектов, делённые на общее количество слов из аспектов.
- Доля уникальных слов. Применительно к ритмическим средствам, основанным на повторении слов, это доля слов, которые повторяются только в одном аспекте, среди всех слов из аспектов.

Среди этих характеристик 1-я характеризует все ритмические средства в целом, 2-я, 3-я и 4-я характеризуют отдельные ритмические средства, 5-я, 6-я и 7-я – структуру средств.

1-я, 2-я и 5-я характеристики имеют значения в абсолютных числах, так как они представляют собой количества некоторых элементов текста. Недостаток данных характеристик в том, что они

зависят от размера текста: чем больше текст, тем больше их значения. При сравнении текстов различной длины использование таких характеристик создаст дополнительную зависимость результата от объёма текста. Избежать этой проблемы, можно двумя способами. Первый – вместо абсолютных значений взять частоты встречаемости или доли элементов среди аналогичных: доля анафор среди всех средств, доля существительных среди всех слов из аспектов и т. п. Второй – сделать поправку на размер текста напрямую, например, как в 4-й характеристике. Длину текста можно выражать не только в предложениях, но и в словах и символах.

Стилометрические характеристики, как правило, считаются для каждого текста в отдельности, что позволяет сравнивать по ним тексты между собой. Но если нужно сопоставить группы текстов, то характеристики текстов внутри группы необходимо агрегировать в общие характеристики группы.

Например, требуется сравнить по ритму два десятилетия: 2000-е и 2010-е года. Пусть для каждого десятилетия имеется по 10 текстов, для которых выделены ритмические средства: анафора, эпифора, симплока. Для сравнения этой пары десятилетий по ритму нужно выполнить следующие действия.

Посчитать ритмические характеристики для отдельных текстов. Например, количества анафор, эпифор и симплок, делённые на количество предложений в тексте, т. е. «плотности» анафор, эпифор и симплок. Так для каждого текста будет составлена математическая модель ритма в виде вектора из трёх характеристик.

Агрегировать характеристики текстов внутри десятилетия. Для этого для 10 текстов из 2000-х годов мы составляем общий вектор из тех же трёх характеристик: плотности анафор, эпифор и симплок. Для 10 плотностей анафор из данных 10 текстов находим среднее арифметическое значение. Это будет плотность анафор для 2000-х годов. Аналогично вычисляем плотности эпифор и симплок, а также такие же характеристики и второго десятилетия.

В результате мы имеем математические модели двух десятилетий, которые можно сравнивать между собой.

Агрегировать характеристики можно не только с помощью среднего значения, но и с помощью других статистических функций: медианы, минимального и максимального значения.

Ритмические характеристики, на основе которых можно рассуждать о стиле текста, не ограничиваются разобранным списком. Эта научная область сейчас активно развивается, так что можно предлагать собственные идеи, как смоделировать ритм текста математически.

§4 Визуализация результатов

Найденные стилометрические характеристики текстов (или других объектов, например, десятилетий) можно визуализировать несколькими способами:

- В виде тепловых карт, которые описывают близость текстов. Это квадратные тепловые карты, на осях которых расположены десятилетия, а оттенок в ячейке обозначает степень близости пары текстов: чем темнее оттенок, тем ближе объекты друг к другу. В качестве меры близости можно использовать популярные метрики: расстояние Чебышёва, коэффициент корреляции, расстояние Евклида, манхэттенское расстояние т другие.
- В виде тепловых карт, которые описывают диапазоны значений стилометрических характеристик. На горизонтальной оси располагаются названия конкретных характеристик, на вертикальной – тексты. Ячейки карты содержат значение характеристики, а также имеют цвет, оттенок которого обозначает величину значения относительно других. Самые большие значения обозначаются светлыми оттенками, самые маленькие – тёмными. Справа на карте отображается столбик с диапазоном значений и оттенками для разных значений.
- В виде графиков, которые демонстрируют изменения стилометрических характеристик текстов для разных периодов времени. Время публикации/написания текстов откладывается по горизонтали, а характеристика текста – по вертикали. На таком рисунке можно отображать несколько графиков, каждый из которых описывает свою характеристику.

Рассмотрим все эти способы подробнее.

Тепловая карта (heat map) – графическое представление данных, где значения отображаются при помощи цвета. Разные оттенки цвета помогают проанализировать градацию значений.

Самый простой способ понять тепловую карту – это подумать о таблице, которая содержит цвета вместо цифр. У цветового градиента по умолчанию имеются два цвета, которые обозначают противоположные или самые далёкие значения. Например, синий и красный, чёрный и белый. Промежуточные оттенки при этом обозначают промежуточные значения.

Тепловые карты хорошо подходят для визуализации больших объемов многомерных данных и могут использоваться для поиска кластеров объектов с одинаковыми значениями, поскольку они отображаются в виде областей одинакового цвета.

Например, при помощи карт можно сравнивать, насколько близки объекты (Рис. 34).

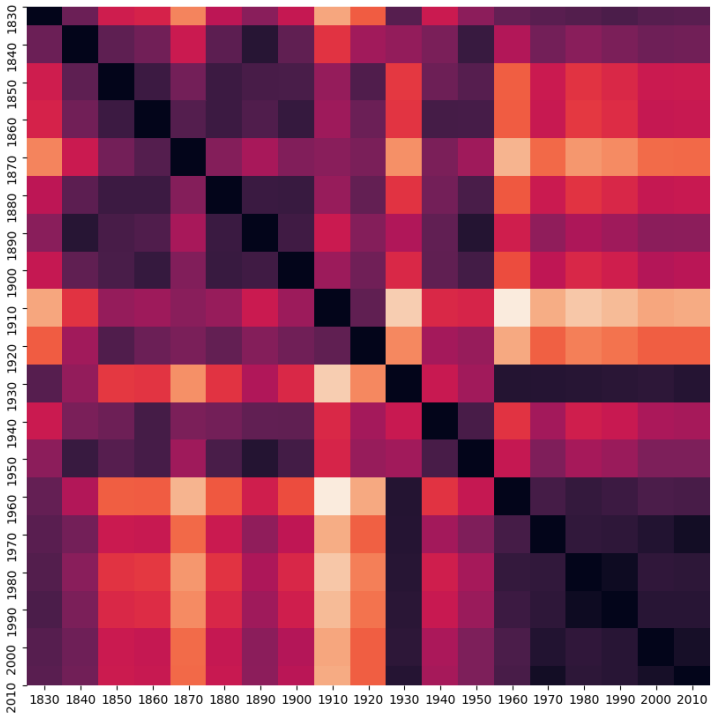


Рис 34. Тепловая карта близости текстов по годам

На рисунке представлена тепловая карта для корпуса из 150 русских текстов 19-21 веков. По вертикали и горизонтали отложены десятилетия. Для каждого десятилетия берутся средние значения числовых ритмических характеристик из текстов, таким образом, десятилетие представляется как вектор числовых ритмических характеристик. Каждая пара десятилетий сравнивается между собой с помощью метрики близости. Для этого рисунка использовалась метрика Чебышёва. Подробнее метрики близости векторов можно изучить в документации Python-библиотеки SciPy: docs.scipy.org/doc/scipy/reference/spatial.distance.html.

Каждая ячейка — это результат сравнения двух десятилетий. По диагонали располагаются чёрные квадраты — это десятилетия сравниваются сами с собой. Чёрный цвет обозначает максимальную близость. Противоположный, белый цвет, означает десятилетия,наиболее далёкие по ритму. Например, на этой карте получилось так, что 1910-е и 1920-е близки между собой, потому что их ячейка окрашена в тёмно-серый, близкий к чёрному. А 1910-е и 2010-е далеки, потому что их ячейка очень светлая.

На карте можно выделять кластеры близких объектов. Например, все ячейки для группы десятилетий 1960-2010 тёмные, так что о них можно сказать, что это кластер близких по ритму произведений.

Эта тепловая карта создана с помощью кода на языке Python:

```
import seaborn as sn
import pandas as pd
import matplotlib.pyplot as plt
from scipy.spatial.distance import cdist
df = pd.read_csv("rus_decades.csv", header=0, index_col=0)
df_array = df.to_numpy()
metric = 'chebyshev'
dist_mat = cdist(df_array, df_array, metric=metric)
dist_df = pd.DataFrame(dist_mat, columns=df.index, index=df.index)
sn.heatmap(dist_df, xticklabels=dist_df.columns, yticklabels=dist_df.columns, cbar=False)
plt.savefig('heatmap.png', fmt='png')
```

Первые строки этой программы описывают, какие библиотеки Python необходимы для работы основного кода: seaborn, pandas, numpy, matplotlib и scipy. pandas и scipy используются для загрузки и обработки данных, seaborn — для создания тепловой карты, а matplotlib — для работы с графиком, на котором располагается тепловая карта.

В начале программы с помощью pandas загружаются стилометрические характеристики десятилетий с 1800 по 2010-е из файла rus_decades.csv. Это файл в формате csv, в котором десятилетия описываются в виде таблицы:

```
decade,anaphora,epiphora,symploce,anadiplosis,diacope,epizeuxis,epanalepsis,polysyndeton,one_word,feat_per_sent,NOUN,ADJS,VERB,ADVB
1990,0.019192587690271344,0.01786896095301125,0.0006618133686300463,0.003970880211780278,0.7147584381204499,0.009927200529450696,0.0013236267372600929,0.12772998014559894,0.3589116478141241,0.8391793514228988,0.16600428003668602,0.04799755426475085,0.1274839498624274,0.07856924487924183
2000,0.02317339149400218,0.01772082878953108,0.0002726281352235551,0.0073609596510359875,0.8056161395856052,0.025354416575790618,0.0038167938931297713,0.14285714285714285,0.2769503176752421,1.0209923664122138,0.16560774919279242,0.05301531090511405,0.14738048119987 5,0.05228622018539735
…
```

То есть в первой строке находятся названия характеристик, в первом столбце — номера десятилетий, а в других строках — значения характеристик десятилетия.

Требуется сравнить эти строки между собой. Для этого они извлекаются из df — объекта таблицы DataFrame в числовую матрицу:

```
df_array = df.to_numpy()
```

Затем между каждой парой строк матрицы df_array находится расстояние по метрике Чебышева с помощью cdist из scipy:

```
metric = 'chebyshev'
dist_mat = cdist(df_array, df_array, metric=metric)
```

В результате получается числовую матрица расстояний между десятилетиями dist_mat. Она преобразовывается в таблицу DataFrame, где в строках и столбцах находятся десятилетия из исходной таблицы:

```
dist_df = pd.DataFrame(dist_mat, columns=df.index, index=df.index)
```

Далее для новой таблицы строится тепловая карта с подписанными десятилетиями:

```
sn.heatmap(dist_df, xticklabels=dist_df.columns, yticklabels=dist_df.columns, cbar=False)
```

И карта сохраняется в файл как картинка 'heatmap.png':

```
plt.savefig('heatmap.png', fmt='png')
```

Таким образом получается изображение с тепловой картой, которая визуализирует близость объектов.

Можно рисовать и другой тип тепловых карт, чтобы посмотреть, как варьируются характеристики. Рассмотрим рисунок 35.

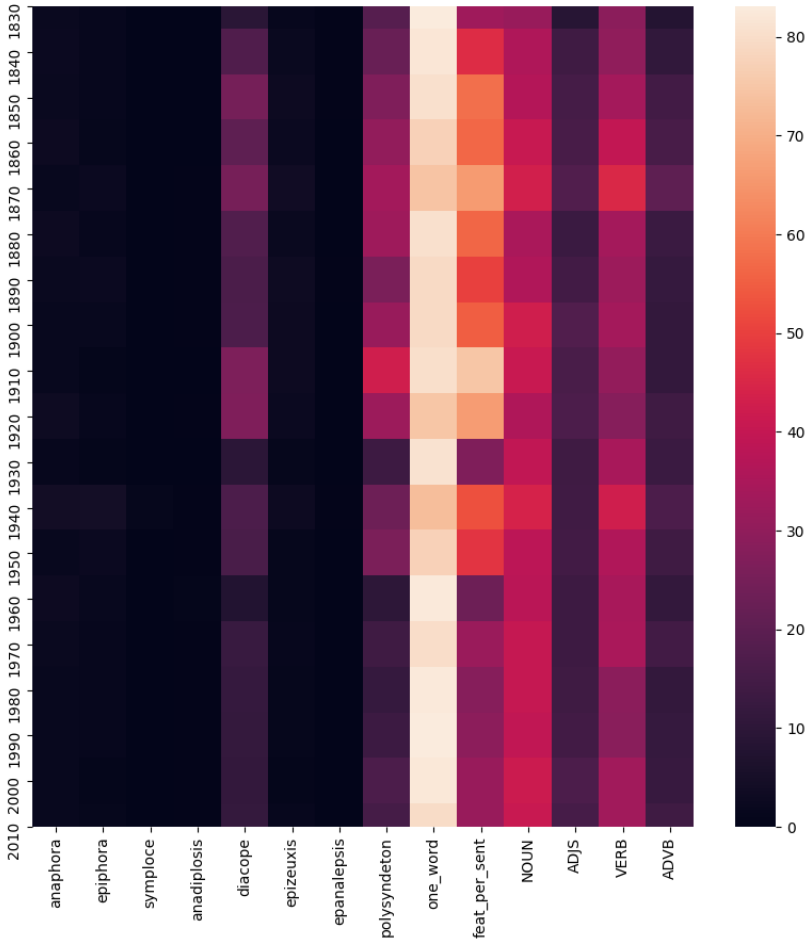


Рис 35. Тепловая карта частоты встречаемости ритмических характеристик

Здесь по вертикали отложены десятилетия, по горизонтали – характеристики. Характеристики с названиями ритмических средств – это среднее количество соответствующих средств в 100 предложениях, своеобразная плотность средств, feature_per_sent – это среднее количество всех ритмических средств в 100 предложениях. one_word – это процент уникальных слов в ритмических средствах. NOUN, ADJS, VERB, ADVB – процент существительных, прилагательных, глаголов и наречий в ритмических средствах. Вертикальная полоса справа – это градация оттенков и значения метрики для каждого оттенка.

Эта карта строится проще предыдущей:

```
import seaborn as sn
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("rus_decades.csv", header=0, index_col=0)
df = df*100
```

```
sn.heatmap(df, cbar=True, fmt="3.1f")
plt.savefig('heatmap.png', fmt='png')
```

Здесь считывается таблица из csv-файла. Параметры header=0 и index_col=0 говорят о том, что в csv-файле первая строка отведена под названия столбцов, а первый столбец — под названия строк.

При визуализации важно обращать внимание на диапазоны характеристик, чтобы отображать их в наглядном и удобном для интерпретации виде. Значения в csv-файле, который используется для примера, находятся в диапазоне от 0 до 1. Для визуализации хотелось бы перевести их в более наглядный диапазон: процентов или количества средств на 100 предложений. Для этого все значения в таблице умножаются на 100:

```
df = df*100
```

В конце снова строится тепловая карта для таблицы и сохраняется в файл:

```
sn.heatmap(df, cbar=True, fmt="3.1f")
plt.savefig('heatmap.png', fmt='png')
```

Параметр cbar=True означает, что справа от тепловой карты будет показан столбик с диапазоном значений и соответствующими им оттенками (рис. 36).

Можно прописать значения метрик напрямую в каждой ячейке, если добавить параметр annot=True:

```
sn.heatmap(data, cbar=True, annot=True)
```

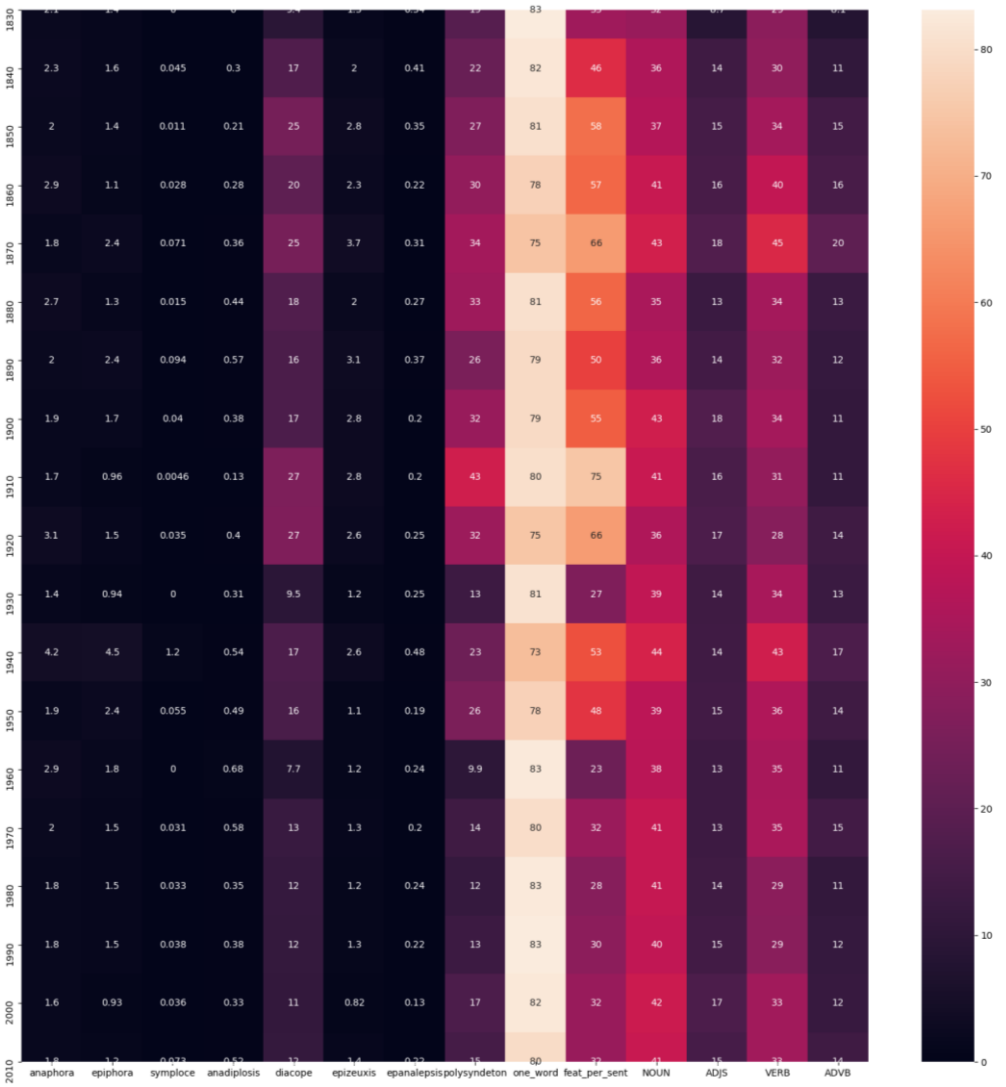


Рис 36. Тепловая карта с указанием значений частоты встречаемости ритмических характеристик

На тепловой карте можно посмотреть на компактную картинку в целом. На этой карте видно, что общее количество средств отличается от десятилетия к десятилетию: в этой колонке много

разных оттенков, причём в 19 веке больше светлых оттенков, а значит и больше средств, а в 21 — фиолетовых, то есть средства встречаются существенно реже.

Столбики, например, с анафорой, эпифорой и симплокой, тёмные. Значит, эти средства встречаются редко. Например, в 1970-х анафора встречается в среднем 2 раза на 100 предложений.

Эти же данные можно визуализировать не только на карте диапазонов, но и на графиках на рис 37.

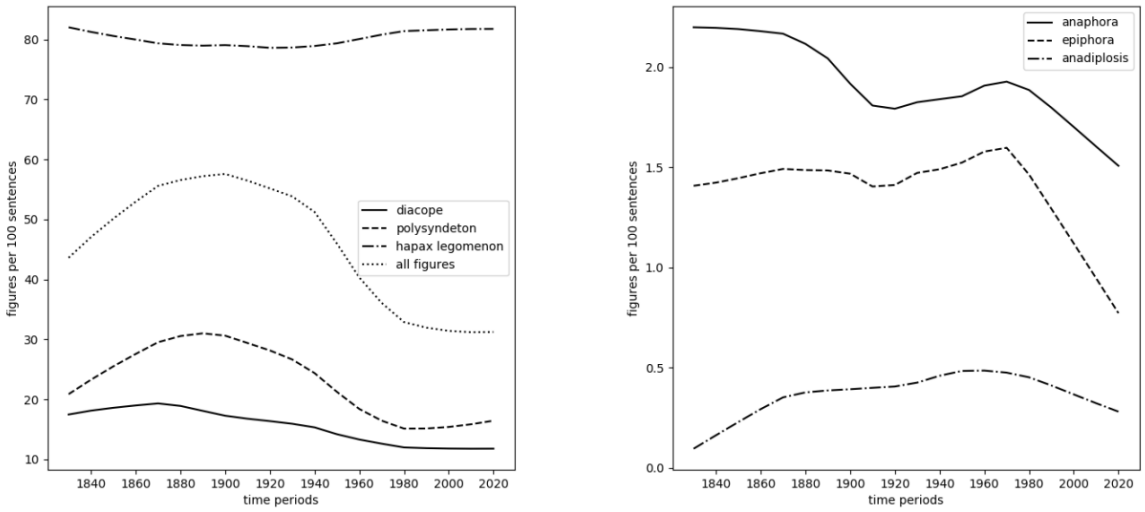


Рис 37. Графики частоты встречаемости ритмических характеристик

По горизонтали здесь отложены периоды публикации, по вертикали - значения характеристик в диапазоне от 0 до 100. Значения характеристик сглажены, чтобы получить аккуратный график с явно выраженными тенденциями. На левом рисунке отображены графики для количества диакоп на 100 предложений (diacope), многосоюзий (polysyndeton), всех найденных ритмических средств (all figures) и доля «уникальных» слов, т. е. тех, которые повторяются ровно один раз (hapax legomenon). На правом рисунке отображены средние количества на 100 предложений для анафоры, эпифоры и анадиплозиса.

Графики создаются при помощи следующего Python-кода:

```
import matplotlib.pyplot as plt
import pandas as pd
from statsmodels.nonparametric.smoothers_lowess import lowess
decades = [1810, 1820, 1830, 1840, 1850, 1860, 1870, 1880, 1890, 1900, 1910, 1920, 1930, 1940,
1950, 1960, 1970, 1980, 1990, 2000, 2010, 2020]
data = pd.read_csv("rus_decades.csv", header=0, index_col=0)
data = data * 100
data = data[['diacope', 'polysyndeton', 'all figures', 'hapax legomenon']]
features = data.to_numpy()
feature_names = list(data.columns.values)
plt.figure(figsize=(7, 7))
plt.locator_params(axis='x', nbins=len(decades))
plt.xlabel('time periods')
plt.ylabel('figures per 100 sentences')
line_types = ['-k', '--k', '-.k', ':k']
for i, name in enumerate(feature_names):
  feature = list(features[:, i])
  filtered = lowess(feature, decades, frac=1./2)
  plt.plot(decades, filtered[:, 1], line_types[i], label=name)
plt.legend()
plt.savefig('plot.png', fmt='png')
```

Здесь снова характеристики десятилетий считываются из csv-файла и умножаются на 100, чтобы они были приведены к диапазону от 0 до 100. Далее для графиков выбираются только 4 столбца из всех:

```
data = data[['diacope', 'polysyndeton', 'all figures', 'hapax legomenon']]
```

В отдельные переменные записываются числовые значения характеристик и их названия:

```
features = data.to_numpy()
feature_names = list(data.columns.values)
```

На следующем шаге описываются параметры графика.

Размер рисунка:

```
plt.figure(figsize=(7, 7))
```

Значения по горизонтальной оси X, т. е. номера десятилетий:

```
plt.locator_params(axis='x', nbins=len(decades))
```

Подписи к осям:

```
plt.xlabel('time periods')
plt.ylabel('figures per 100 sentences')
```

И заводятся типы линий графика, в том числе сплошная линия, линия из дефисов, линия из точек, линия из чередующихся дефисов и точек:

```
line_types = ['-k', '--k', '-.k', ':k']
```

k обозначает чёрный цвет, который также можно изменить на другой, если нужен цветной рисунок.

В цикле перебираются все четыре характеристики. Значения каждой сглаживаются, чтобы сделать линию плавной и выделить тенденции:

```
filtered = lowess(feature, decades, frac=1./2)
```

Затем линия добавляется на график с указанными типом и названием для легенды:

```
plt.plot(decades, filtered[:, 1], line_types[i], label=name)
```

В конце на график добавляется легенда, а сам график сохраняется как png-изображение:

```
plt.legend()
plt.savefig('plot.png', fmt='png')
```

Графики и тепловые карты диапазонов фактически отображают одни и те же данные и тенденции изменения этих данных. Достоинства тепловых карт в том, что они располагают данные более компактно, не накладывают их друг на друга и используют широкий диапазон цветов и оттенков. Достоинства графиков в том, что с ними удобнее проводить вычисления: находить значение характеристики в данной точке, рассчитывать, на сколько изменилась характеристика за заданный период и т. п.

Таким образом, рассмотренные способы визуализации достаточно наглядны и позволяют проанализировать как динамику изменения стилометрических характеристик, так и близость текстов или десятилетий друг к другу с точки зрения стиля. Конечно, способов визуализации стилометрических характеристик гораздо больше трёх. Например, можно отображать на графиках не линию, ведущую от десятилетия к десятилетию, а отдельные тексты как набор точек, чтобы оценить разброс значений в соседних годах. Или можно оценивать близость текстов/десятилетий с помощью дендрограммы — результата кластеризации. Подробнее о визуализации данных можно прочитать в документации Python-библиотек Matplotlob [Matplotlib: Visualization with Python], Seaborn [Seaborn: statistical data visualization] и других, а также изучить язык программирования R [What is R].

🕮 Вопросы и задания

1. Как структурированы статистические эксперименты?
2. На что нужно обращать внимание при подборе корпуса текстов?
3. Проанализируйте корпус Amazon Fine Food Reviews (https://www.kaggle.com/snap/amazon-fine-food-reviews). Какую разметку он уже имеет? Для каких задач компьютерной лингвистики его можно использовать?
4. Как оценить качество разметки корпуса? Какие метрики для этого можно использовать?
5. Дан фрагмент текста с разметкой по эпифорам (найденные правильно и неправильно эпифоры выделены зелёным):

Приходи ко мне завтра. Ты ведь придешь завтра?

Я к тебе завтра обязательно приду. Даже если будет дождь, все равно приду. Обязательно приду.

Была ли я благодарна ему? Была ли?

Оцените точность и полноту этой разметки.

6. Как можно классифицировать стилометрические характеристики текстов?

7. Какие способы визуализации стилометрических характеристик полезны для анализа динамики изменения стиля по временным эпохам?

БИБЛИОГРАФИЯ

1. *Алекссева И.С.* Введение в перевод введение: Учеб. пособие для студ. филол. и лингв, фак. высш. учеб. заведений. – СПб.: Филологический факультет СПбГУ; М.: Издательский центр «Академия», 2004. – 352 с.
2. *Большакова Е.И.* и др. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е. И., Клышинский Э. С., Ландэ Д. В., Носков А. А., Пескова О. В., Ягунова Е. В. – М.: МИЭМ, 2011. – 272 с.
3. *Васильев В.С.* Учебник: Регулярные выражения (regular expressions) // Блог программиста: сайт. 2019. URL: https://pro-prof.com/archives/5414 (дата обращения: 15.09.2020).
4. *Герчук Ю.И.* Художественная структура книги. М.: Книга, 1984. 208 с.39.
5. *Голубева-Монаткина Н.И.* О ритме художественной прозы в свете идей М.М. Бахтина и переводе. Русский язык и культура в зеркале перевода: VI Международная научная конференция; 13–17 мая, 2016 г., Афины, Греция: Материалы конференции. – М.: МАКС Пресс, 2016. С. 107-108.
6. *Жеребило Т.В.* Словарь лингвистических терминов, Изд. 5-е, испр-е и дополн. – Назрань: Изд-во "Пилигрим", 2010. - 486с.
7. *Иванова-Лукьянова Г.Н.* Ритм художественных прозаических текстов как отражение жизненных ритмов человека. Русский язык в современном мире… Материалы II международной научной конференции. М.: Изд. Высшая школа перевода МГУ. 2011. С. 302.
8. *Ивлева А.Ю., Свойкин К.Б.* Категория эквивалентности в отечественном и западном переводоведении. Вестник Самарского университета. История, педагогика, филология. 2017с. С. 71-77 – цит. с. 73.
9. *Квятковский А.П.* Поэтический словарь. – М.: Сов. Энцикл., 1966. – 376 с.
10. Комиссаров В.Н. Общая теория перевода (лингвистические аспекты): Учеб. для ин-тов и фак. иностр. яз. М.: Высш. шк., 1990. 253 с.
11. *Миронова Н.Н.* Когнитивные аспекты перевода художественной литературы // Вестник Московского университета. Сер. 22. Теория перевода. – 2013. – № 3. С. 77 – 83.
12. *Михальская Н.П., Г.В. Аникин* Английский роман XX века: Учеб. пособие для филол. специальностей. – М.: Высш. школа. – 1982. – 192с.
13. *Мурзина Л.А. Кравчук С.В.* Исследование передачи ритма при переводе литературных произведений на неродной язык [Электронный ресурс] / Л.А. Мурзина, С.В. Кравчук // Режим доступа: http://www.ayk.gov.tr/wp-content/uploads/2015/01/MURZ%C4%B0NA-L.-A.-.pdf].
14. Словарь литературоведческих терминов / Л. И. Тимофеев, С. В. Тураев. - М.: Просвещение, 1974.
15. *Татару Л.В.* Композиционный ритм и когнитивная логика нарративного текста (сборник Дж. Джойса «Дублинцы»). Известия Российского государственного педагогического университета им. А.И. Герцена. – 2008. Режим доступа: https://cyberleninka.ru/article/n/kompozitsionnyy-ritm-i-kognitivnaya-logika-narrativnogo-teksta-sbornik-dzh-dzhoysa-dublintsy.
16. Уроки по регулярным выражениям // Ravesli|Программирование для начинающих: сайт. URL: https://ravesli.com/uroki-po-regexp/ (дата обращения: 15.09.2020).
17. *Филотенкова Е.А.* Шекспировские аллюзии в романе Айрис Мердок «Черный принц», рубрика: Филология, лингвистика, опубликовано в «Молодой ученый № 8 (55)», дата публикации: 05.08.2013.
18. *Bird S., Klein E., Loper E.* Ntural language processing with Python: analyzing text with the natural language toolkit. – " O'Reilly Media, Inc.", 2009. pp. 465.
19. Ding R. Rhythm in translations [Электронный ресурс] / R. Ding // International Education Studies Vol1, № 4, 2008 Режим доступа: http://www.ccsenet.org/journal/index.php/ies/article/view/624.
20. Installation & Getting Started// Stanza: официальный сайт. URL: https://stanfordnlp.github.io/stanza/installation_usage.html (дата обращения: 15.09.2020).
21. *Lagutina K., Lagutina N., Boychuk E., Vorontsova I., Shliakhtina E., Belyaeva O., Paramonov I.* A Survey on Stylometric Text Features // Proceedings of the 25th Conference of Open Innovations Association FRUCT. – IEEE. 2019. – P. 184–195.

22. Matplotlib: Visualization with Python// Matplotlib: официальный сайт. URL: https://matplotlib.org/ (дата обращения: 15.09.2020).

23. Natural Language Toolkit// NLTK: официальный сайт. URL: https://www.nltk.org/ (дата обращения: 15.09.2020).

24. NLTK Part of Speech Tagging Tutorial// Python Programming: сайт. URL:https://pythonprogramming.net/natural-language-toolkit-nltk-part-speech-tagging/ (дата обращения: 15.09.2020)

25. *Parks T.* Translating Style [Электронный ресурс] / T. Parks. - St. Jerome Publishing Manchester, UK & Kinderhook (NY), USA, 2007. Режим доступа: https://apeiron.iulm.it/retrieve/handle/10808/3908/45891/Translating%20Style.pdf.

26. *Pekkanen H.* Who's got rhythm? Rhythm-related shifting in literary translation [Электронный ресурс] / H. Pekkanen // Palimpsestes № 27, 2014. Режим доступа: http://journals.openedition.org/palimpsestes/2072.

27. Seaborn: statistical data visualization// Seaborn: официальный сайт. URL: https://seaborn.pydata.org/ (дата обращения: 15.09.2020).

28. *Sokolova M., Lapalme G.* A systematic analysis of performance measures for classification tasks //Information processing & management. – 2009. – Т. 45. – №. 4. – P. 427-437.

29. *Sokolova M., Lapalme G.* A systematic analysis of performance measures for classification tasks //Information processing & management. – 2009. – Т. 45. – №. 4. – P. 427-437.

30. SpaCy. Industrial-streng Natural Language Processing in Pyton// SpaCy: официальный сайт. URL: https://spacy.io/ (дата обращения: 15.09.2020).

31. Stanza – A Python NLP Package for Many Human Languages// Stanza: официальный сайт. URL: https://stanfordnlp.github.io/stanza/ (дата обращения: 15.09.2020).

32. *Steven Bird, Ewan Klein, and Edward Loper.* Natural Language Processing with Python// NLTK: официальный сайт. URL: https://www.nltk.org/book/ (дата обращения: 15.09.2020).

33. TextBlob: Simplified Text Processing// TextBlob: официальный сайт. URL: https://textblob.readthedocs.io/en/dev/ (дата обращения: 15.09.2020).

34. *Turing A.M.* Computing machinery and intelligence //Parsing the turing test. – Springer, Dordrecht, 2009. – С. 23-65.

35. Universal features// Universal Dependencies: официальный сайт. URL: https://universaldependencies.org/u/feat/index.html (дата обращения: 15.09.2020).

36. Universal POS tags// Universal Dependencies: официальный сайт. URL:https://universaldependencies.org/u/pos/ (дата обращения: 15.09.2020).

37. Welcome to Project Gutenberg// Project Gutenberg: официальный сайт. URL: http://www.gutenberg.org/ (дата обращения: 15.09.2020).

38. What is R?// The R Project for Statistical Computing: официальный сайт. URL: https://www.r-project.org/about.html (дата обращения: 15.09.2020).

39. *Zafar Ali.* A simple Word2vec tutorial// Medium corporation: сайт. 2019. URL: https://medium.com/@zafaralibagh6/a-simple-word2vec-tutorial-61e64e38a6a1 (дата обращения: 15.09.2020).

Учебное пособие

Бойчук Е.И., Лагутина Н.С., Лагутина К.В., Воронцова И.А., Шляхтина Е.В., Мишенькина Е.В., Беляева О.В.

АВТОМАТИЗИРОВАННЫЙ АНАЛИЗ ТЕКСТОВЫХ ХАРАКТЕРИСТИК

Компьютерная верстку *С.С. Ламан*

Подписано в печать 28.11.2020 г.
Объем 11 п.л. Тираж 100 экз. Заказ № 5415.

Отпечатано с предоставленных оригинал-макетов
в типографии «Канцлер»
150008, г. Ярославль, ул. Полушкина роща, д. 16, стр. 66а.
Тел.: 8-4852-58-76-33, 8-4852-58-76-39
E-mail: kancler2007@yandex.ru