УДК 004.415.2; 81.13

Программный инструмент «Златоуст» для поиска и обработки анафорических повторов в текстах

© В.Л. Аршинский, Е.А. Осипова, А.С. Исаченков
Иркутский национальный исследовательский технический университет,
г. Иркутск, Российская Федерация

Аннотация. Статья посвящена описанию результатов разработки специального программного инструмента для поиска и статистической обработки анафорических повторов при проведении филологических исследований структуры текста в текстосимметрике. Рассматриваются основные функциональные возможности программного продукта и его пользовательский интерфейс. Данное приложение ориентировано на пользователей, не обладающих навыками программирования, то есть применение его на практике не требует профессиональных знаний и умений в области информационных технологий. Разрабатываемое приложение предназначено для автоматизации обработки текстов согласно методике позиционных срезов, включающей в себя построение схематической модели текста на основе пропорции золотого сечения. Данная методика была предложена Г.Г. Москальчук и экспериментально подтверждена последователями на большом количестве текстов русского языка разных стилей. Проведённые эксперименты доказывают, что, во-первых, при определении темы текста и заглавия реципиенты ориентируются в большей степени на такие единицы текста, как повторы. Во-вторых, управляющим параметром восприятия является различная концентрация этих единиц в тексте. В-третьих, позиционная структура текста выступает синхронизатором количественного распределения повторяющихся единиц и восприятия их неслучайной последовательности и локализации. Тем самым предлагаемый авторами программный инструмент «Златоуст» позволит расширить область применения вышеуказанной методики, что в конечном итоге будет способствовать пониманию текстов на уровне осознания композиционной структуры целого и уяснения глубинных связей между словами и предложениями.

Ключевые слова: программные средства, текстосимметрика, анафоры, схематическая модель текста

Zlatoust Software Tool for Searching and Processing Anaphoric Repetitions in Texts

© Vadim L. Arshinsky, Elizaveta A. Osipova, Alexander S. Isachenkov
Irkutsk National Research Technical University,
Irkutsk, Russian Federation

Abstract. The article describes the results of the development of a special software tool for the search and statistical processing of anaphoric repetitions when conducting philological studies of the text structure in textual symmetry. The article discusses the basic functionality of the software product and its user interface. This application is aimed at users who do not have programming skills, that is, its application in practice does not require professional knowledge and skills in the field of information technology. The developed application is designed to automate text processing according to the positional slicing technique, which includes the construction of a schematic text model based on the proportion of the golden ratio. This technique was proposed by G.G. Moskalchuk and experimentally confirmed by followers on a large number of Russian texts of different styles. The experiments carried out prove that, firstly, when determining the topic of the text and the title, the recipients are guided to a greater extent by such units of the text as repetitions. Secondly, the control parameter of perception is the different concentration of these units in the text. Thirdly, the positional structure of the text acts as a synchronizer of the quantitative distribution of repetitive units and the perception of their non-random sequence and localization. Thus, the Zlatoust software tool proposed by the authors will expand the scope of the above methodology, which will ultimately contribute to the understanding of texts at the level of comprehending the compositional structure of the whole and understanding the deep connections between words and sentences.

Keywords: software, textual symmetry, anaphora, schematic text model

Введение

Текстосимметрика[1] – это направление филологических исследований, целью которого является создание качественной и количественной теории формообразования текста. Предметом текстосимметрики выступает линейная форма текста, динамически структурированная по законам симметрии, а её объектами – текст-система и динамически изменяющий свою структурную роль повтор[2].

Одной из множества задач, решаемых исследователями-лингвистами в области текстосимметрики, является задача поиска и статистической обработки таких элементов симметрии текстов [1, 2], как анафорические повторы, а также моделирование структуры текста в привязке к сильным позициям. Объектом исследования являются оригинальные и переводные тексты на различных языках (например, древнерусская литература).

Под анафорой (или анафорическим повтором) в свою очередь понимается синтаксическое средство выразительности. Анафора – фигура речи, состоящая в повторении в начале фразы или стихотворной строки одних и тех же звуков, слов, словосочетаний и конструкций. Анафора – это не просто повторение: благодаря ей смысл определённых слов усиливается, они выделяются как наиболее важные, следовательно, мысль говорящего или пишущего актуализируется; слушатель или читатель невольно обращает на неё большее внимание[3] [3].

Анализ существующих программных средств для исследования повторов в тексте

В настоящее время исследователи вынуждены выполнять поиск повторов в тексте с помощью тривиальных программных средств (текстовых редакторов, «небольших» корпусных менеджеров) и обрабатывать результаты вручную.

Связано это прежде всего с тем, что, несмотря на то, что программные инструменты лингвостатистического анализа повторов в текстах существуют [4] (среди них есть и свободно распространяемые), все они имеют свои ограничения.

Например, для работы с корпусами небольших размеров часто используется свободно распространяемое мультиплатформенное средство AntConc [5, 6], с помощью которого можно производить такие операции, как построение частотного списка словоформ и/или лемм с указанием абсолютной частоты, выделение ключевых слов в тексте и построение графиков в формате штрихового кода, где видно позиции ключевых слов, и др. Ограничением является то, что исходный код этого корпусного менеджера является закрытым, что не позволяет расширять спектр его функциональных возможностей для решения задач текстосимметрики.

Иным подходом к разработке лингвистических программ, предназначенных для смыслового анализа текстов с применением частотного анализа и вероятностно-статистических методов, является использование готовых библиотек и моделей. В контексте данного исследования можно выделить модель Word2vec [7], которую в 2013 году предложила компания Google в качестве относительно нового подхода к обработке естественного языка. В настоящее время в рамках проекта Word2vec активно создаются C++ и Python программы, позволяющие с помощью векторной модели слов Word2vec с машинным обучением получить «словарь», однозначно характеризующий использование заданных слов в контексте определённых тем. Здесь для поиска семантически связанных документов применяются различные метрики их подобия, которые вычисляются с использованием числовых векторов для встречающихся слов. Эти векторы, в частности, содержат информацию о частоте встречаемости слова. На языке Python инструмент Word2vec нашёл свою реализацию в бесплатной библиотеке Gensim [8]. Ограничением для использования исследователями-филологами подобных библиотек является то, что для их применения требуются хорошие умения и навыки программирования на языках программирования высокого уровня (C++ или Python).

Использование неспециализированных программных средств и ручная обработка результатов существенно увеличива-

[1] Корбут А.Ю. Текстосимметрика как раздел общей теории текста: дис. д-ра филол. наук. Барнаул, 2005. 343 с.

[2] Корбут А.Ю. Повтор как средство структурной организации художественного прозаического текста (элементы симметрии): автореф. дис. канд. филол. наук. М., 1995. 21 с.

[3] Мизинина И.Н. Изобразительно-выразительные средства языка. (Тропы, лексические и синтаксические средства). Словарь-справочник. М.: Н-ПРО, 2016. 140 с.

ют время, затрачиваемое на проведение исследования, а также приводят к возникновению существенного числа ошибок в подсчётах. Это обуславливает необходимость создания специализированных программных средств для поддержки проведения исследований в области текстосимметрики.

Программный инструмент «Златоуст» для поиска анафорических повторов в тексте

В ходе анализа методов текстосимметрики[1] было принято решение о необходимости разработки программного инструмента с простым графическим интерфейсом, обладающего следующими функциональными возможностями:

— *поиск и статистическая обработка анафорических повторов;*

— *визуальное цветовое выделение найденных анафор в исследуемом тексте;*

— *визуализация координат найденных анафор с помощью схематической модели текста.*

Для удобства использования было принято решение реализовать программный инструмент в идеологии кроссплатформенного настольного приложения. В качестве средств реализации была выбрана платформа Java SE8.

В ходе постановки задачи были сформулированы следующие функциональные требования:

1. работа с текстом:
   1) загрузка текста из файла;
   2) редактирование текста;
   3) возможность работы с текстами на различных языках, использующих различные алфавиты, включая устаревшие символы (церковнославянские, древнерусские и т. п.);
   4) цветовая подсветка найденных анафор по тексту в текстовом редакторе;
   5) подсчёт общего количества слов и символов;
   6) добавление варианта анафоры путём выделения слова мышью;
   7) возможность настройки шрифта и размера текста для отображения в редакторе;
2. поиск повторяющихся слов/фраз:
   1) возможность задания ограничения на минимальную длину и количество повторов;
   2) просмотр координат повторов;
3. работа с анафорами:
   1) создание анафор и добавление вариантов;
   2) просмотр и редактирование списка анафор;
   3) удаление анафор;
   4) добавление вариантов анафор;
   5) просмотр координат повторов для каждой анафоры;
   6) возможность назначения уникального цвета для каждой анафоры;
4. построение графика анафорических повторов:
   1) выбор анафор для визуализации их координат;
   2) возможность настройки отображения графика;
   3) возможность одновременного отображения координат нескольких анафор;
   4) возможность экспорта графика в виде графического файла (PNG, JPG);
5. возможность сохранения результатов исследования в виде файла проекта (проект включает в себя исследуемый текст, список анафор и их свойства, а также установленные настройки редактора).

В процессе разработки приложения были выделены следующие группы классов:

— пакет **data** содержит классы, выделенные в результате объектной декомпозиции предметной области: Anathora (представляет анафору), AnaphoraDictionary (словарь анафор), Project (сущность-проект, включающий в себя словарь анафор и исследуемый текст);

— пакет **view** содержит классы, реализующие вывод окон приложения и логику пользовательского интерфейса;

— пакет **dialogs** содержит классы, которые необходимы для реализации нестандартного диалога выбора цвета для выделения анафоры в тексте;

— пакет **util** включает в себя вспомогательные классы-утилиты (например, класс WordsCalculator для выделения слов в тексте и подсчёта их координат, класс DuplicateSearcher для поиска повторяющихся фрагментов в тексте, класс Statistics для получения статистики по найденным анафо-

[4] Корбут А.Ю. Текстосимметрика как раздел общей теории текста: дис. … д-ра филол. наук. Барнаул, 2005. 343 с.

рам в тексте, класс IO для организации ввода/вывода в файл и т. п.);

— пакет **chart** содержит классы, которые необходимы для построения и отрисовки графиков.

Пользовательский интерфейс приложения включает в себя восемь окон: «Редактор» (главное окно), «Словарь анафор», «Поиск повторов в тексте», «График встреч», «Окно настроек отображения графика», «Окно выбора данных для графика», «Окно выбора цвета для выделения анафоры по тексту».

На рис. 1 представлено главное окно приложения. В рабочей области загружен фрагмент текста из книги святого Иоанна Златоуста[5]. Две найденные анафоры выделены по тексту при помощи цветных маркеров. В строке состояния отображено общее количество слов и символов в данном тексте.

Меню главного окна содержит следующие группы команд: «Файл» (включает команды «Создать новый проект», «Открыть файл», «Сохранить проект», благодаря которым осуществляется сохранение и загрузка уже изученного текста вместе с найденными анафорами); «Правка» (содержит такие стандартные команды для редактирования текста, как «Копировать», «Вырезать», «Вставить»); «Текст» (команды «Открыть окно просмотра списка созданных анафор», «Открыть окна для поиска повторяющихся фрагментов», «Открыть окна графика», «Установить или снять опцию подсветки анафор»); «?» (команды «Открыть окно информации о программе» и «Открыть справку»).

На панели инструментов, помимо продублированных пунктов меню в виде пиктограмм, размещена кнопка для создания анафоры из выделенного фрагмента текста и элементы для выбора начертания и размера шрифта отображаемого текста.

На рис. 2 представлена таблица повторов с возможностью просмотра координат их размещения, полученная в результате автоматизированного поиска повторяющихся фрагментов текста.

На рис. 3 а приведён пример созданного пользователем списка анафор, соответствующего тексту на рис. 1. По имеющимся в представленном окне кнопкам видно, что через него можно вызвать окно редактора анафор, которое позволяет добавлять новые и редактировать существующие анафоры (рис. 3 б), кроме этого, можно вызвать окно свойств каждой анафоры (рис. 4).

На рис. 5 показана схематическая модель исследуемого текста, построенная по методике позиционных срезов Г.Г. Москальчук[6,7] [1, 9, 10]. Она имеет вид отрезка, представляющего любой текст как линейное целое (единицу), в котором координаты (точки) *0; 0,146; 0,236; 0,618; 0,944; 1* признаны с*ильными позициями* (отмечены на графике на рис. 5). Согласно экспериментальным данным А.Ю. Корбут информация, размещённая в этих местах, запоминается лучше на 15,77 %. Темы квазитекстов, сформулированные испытуемыми после прочтения самых разных текстов, на 46,7 % совпадают с тематикой повторов, расположенных в их сильных позициях [9].

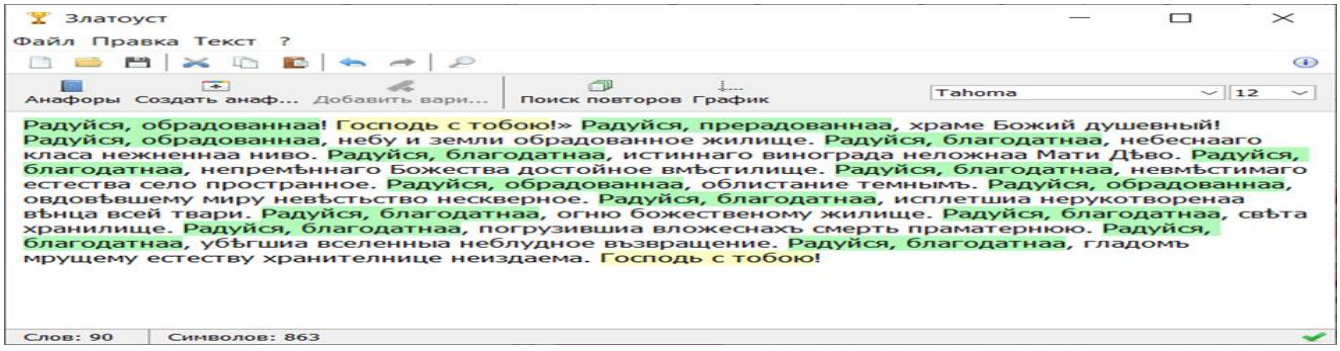


Рис. 1. Главное окно приложения

[5] Кни́га, глаго́лемая «Златоу́стъ», въ не́йже вся́ко ухищре́ніе Боже́ственнаго Писа́нія, истолко́вано святы́мъ вели́кимъ Іоа́нномъ Златоу́стомъ и про́чими святы́ми отцы́. М.: При Свя́то-Тро́ицко-Введе́нской це́ркви, въ тѵпогра́фіи единовѣ́рцевъ, 1910. С. 130–133.

[6] Москальчук Г.Г. Фразовый повтор в диалектной речи: автореф. дис. … канд. филол. наук. М., 1990. 16 с.

[7] Корбут А.Ю. Текстосимметрика как раздел общей теории текста: дис. … д-ра филол. наук. Барнаул, 2005. 343 с.

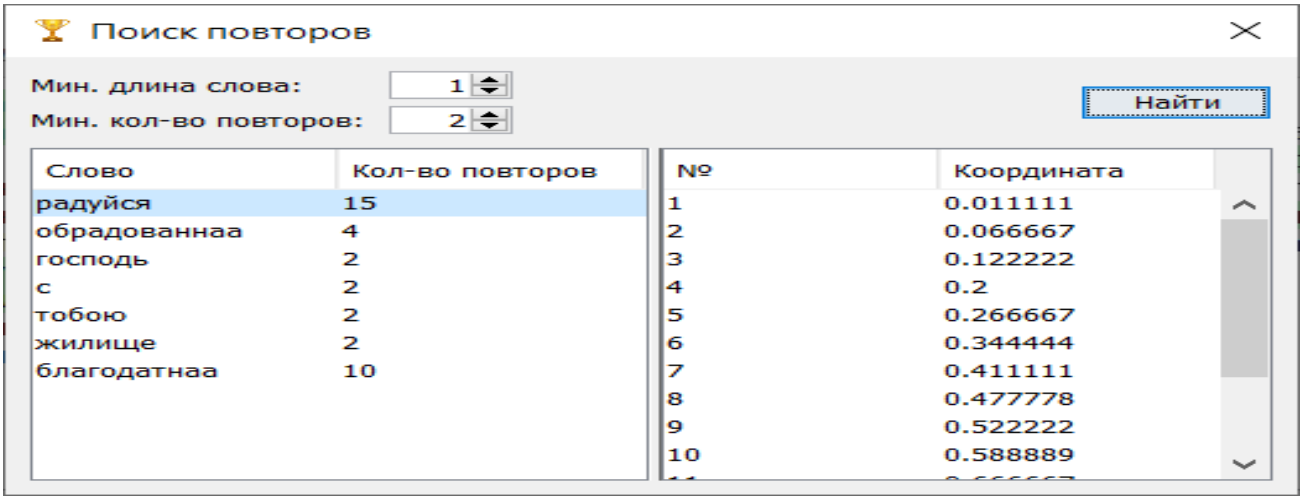
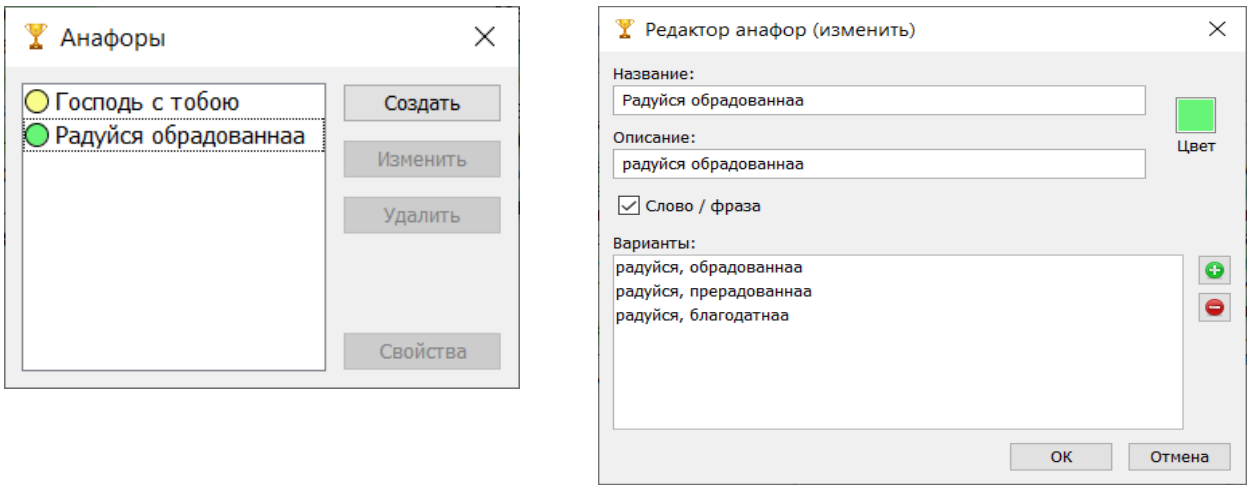


Рис. 2. Окно поиска повторов



а) б)

Рис. 3. а) окно просмотра списка имеющихся анафор; б) окно создания новой анафоры или редактирования существующей

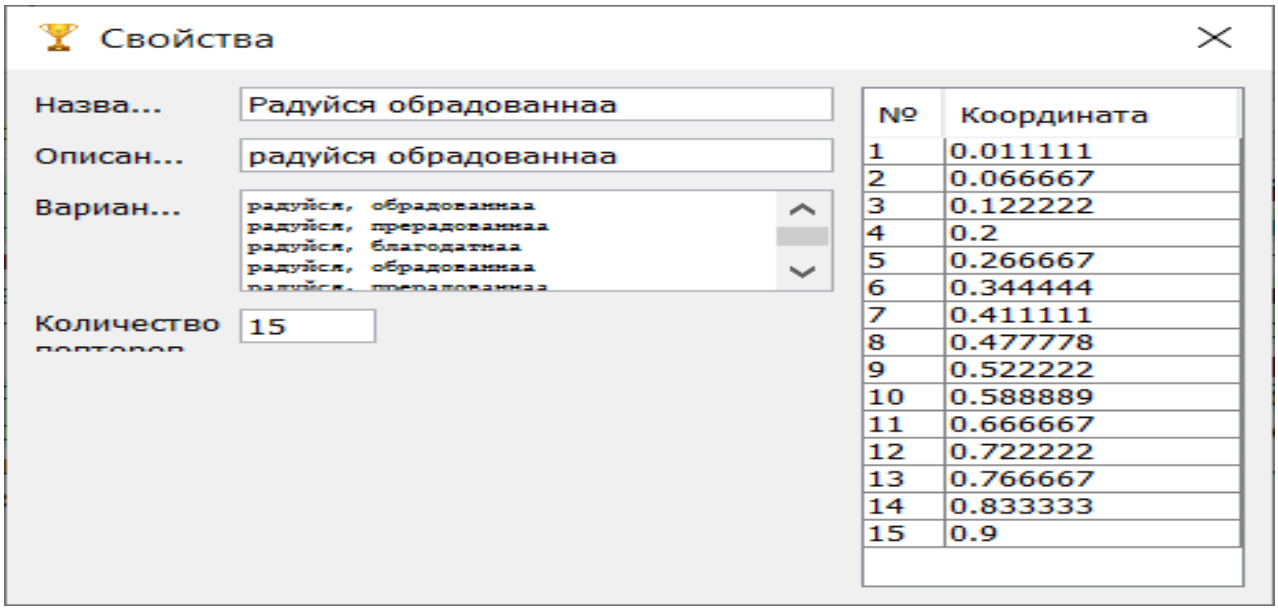


Рис. 4. Окно свойств анафоры

Схематическая модель текста позволяет сопоставлять пространственно-временные материальные структуры любых текстов в рамках отдельной отрасли лингвосинергетики – текстосимметрики.

С помощью окна выбора данных для графика можно выбрать анафоры, координаты которых надо отобразить на графике (рис. 6).

Через окно выбора параметров отображения графика (рис. 7) определяются:

– тип маркера (круг, ромб, квадрат, треугольник и т. д.);
– размер маркера;
– размер шрифта;
– режим отображения координат анафор (на оси или на разных уровнях);
– наличие стрелок на осях.

Рис. 5. Окно отображения графика

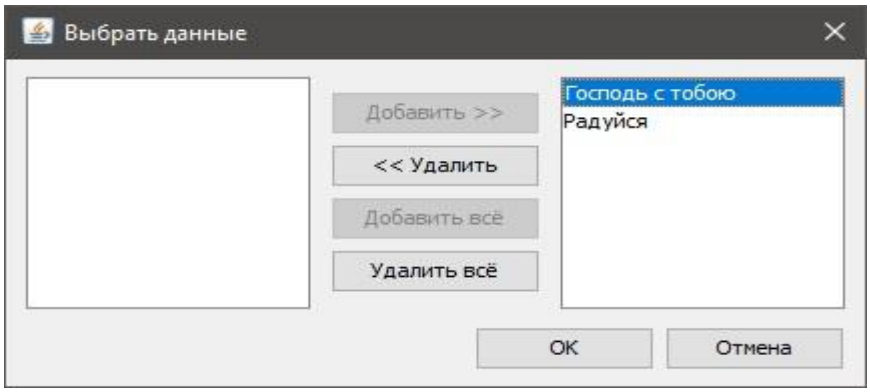


Рис. 6. Окно выбора анафор, отображаемых на графике

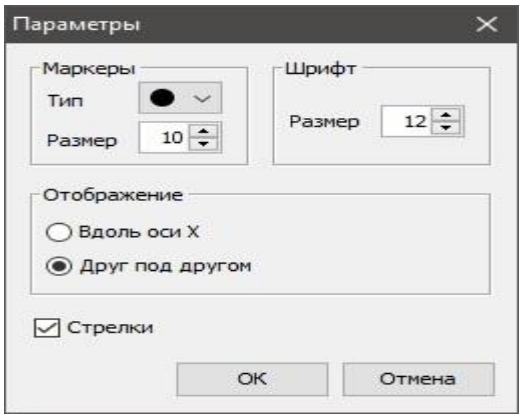


Рис. 7. Окно выбора параметров отображения графика

Заключение

В результате проделанной работы был реализован программный инструмент «Златоуст» для поддержки проведения исследований в области текстосимметрики. Программный инструмент обладает несложным графическим интерфейсом, схожим с типовым текстовым редактором, что делает его удобным в использовании филологами-исследователями, не обладающими навыками программирования. В статье описаны основные функциональные возможности разработанного программного продукта: поиск и статистическая обработка анафорических повторов, визуальное цветовое выделение найденных анафор в исследуемом тексте, визуализация координат повторов на основе схематической модели текста. Нами было получено авторское свидетельство на разработанный программный продукт[8]

[8] Аршинский В.Л., Исаченков А.С., Осипова Е.А. Программа для поиска и статистической обработки анафорических повторов в текстах («Златоуст»). Свидетельство о регистрации программы для ЭВМ RU 2019666323, 09.12.2019. Заявка № 2019665390 от 29.11.2019.

Библиографический список

1. Корбут А.Ю. Симметрия и информация в языке и тексте: монография. Иркутск: Изд-во ВСГАО, 2015. 328 с.
2. Корбут А.Ю. Принцип симметрии в научном тексте // Сибирский филологический форумКрасноярского государственного педагогического университета им. В.П. Астафьева. 2019. № 4 (8). С. 24–31.
3. Радченко И.И. Повтор как средство реализации лингвистических категории экспрессивности в тексте газетной статьи // Научная мысль Кавказа. 2013. № 1. С. 129–134.
4. Логичев С.В. Каталог лингвистических программ и ресурсов в Сети [Электронный ресурс]. URL: https://rvb.ru/soft/catalogue/c01.html (08.12.2020).
5. Общий обзор инструментария AntConc // Lektsii.org [Электронный ресурс]. URL: https://lektsii.org/5-72958.html (08.12.2020).
6. Прикладная и компьютерная лингвистика: коллективная монография / под ред. И.С. Николаева, О.В. Митрениной, Т.М. Ландо. М.: ЛЕНАНД, 2017. 320 с.
7. Проект компании Google word2vec // Google Code [Электронный ресурс]. URL: https://code.google.com/archive/p/word2vec/ (08.12.2020).
8. Вложения word2vec // Библиотека Gensim [Электронный ресурс]. URL: https://radimrehurek.com/gensim/models/word2vec.html (08.12.2020).
9. Москальчук Г.Г. Структура текста как синергетический процесс. М.: Едиториал УРСС, 2003. 296 с.
10. Дорофеева В.А. Композиционная значимость интервалов структуры в восприятии и понимании текста // Современные тенденции общественных наук: политология, социология, философия: материалы II Международной заочной науч.-практ. конф. (г. Новосибирск, 3 мая 2011 г.). Новосибирск, 2011.

Сведения об авторах / Information about the Authors

Аршинский Вадим Леонидович,
кандидат технических наук,
доцент «Центра программной инженерии»,
Институт информационных технологий и анализа данных,
Иркутский национальный исследовательский технический университет,
664074, г. Иркутск, ул. Лермонтова, 83, Российская Федерация,
e-mail: arshinskyv@mail.ru

Vadim L. Arshinsky,
Cand. Sci. (Technics),
Associate Professor at Software Engineering Center,
Institute of Information Technology and Data Analysis,
Irkutsk National Research Technical University,
83 Lermontov Str., Irkutsk, 664074, Russian Federation,
e-mail: arshinskyv@mail.ru

Осипова Елизавета Алексеевна,
кандидат технических наук,
доцент «Центра программной инженерии»,
Институт информационных технологий и анализа данных,
Иркутский национальный исследовательский технический университет,
664074, г. Иркутск, ул. Лермонтова, 83, Российская Федерация,
e-mail: osipovaelizaveta@yandex.ru

Elizaveta A. Osipova,
Cand. Sci. (Technics),
Associate Professor at Software Engineering Center,
Institute of Information Technology and Data Analysis,
Irkutsk National Research Technical University,
83 Lermontov Str., Irkutsk, 664074, Russian Federation,
e-mail: osipovaelizaveta@yandex.ru

Исаченков Александр Сергеевич,
студент,
Институт информационных технологий и анализа данных,
Иркутский национальный исследовательский технический университет,
664074, г. Иркутск, ул. Лермонтова, 83, Российская Федерация,
e-mail: isachenkoff.alexander@yandex.ru

Alexander S. Isachenkov,
Student,
Institute of Information Technology and Data Analysis,
Irkutsk National Research Technical University,
83 Lermontov Str., Irkutsk, 664074, Russian Federation,
e-mail: isachenkoff.alexander@yandex.ru