НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Факультет гуманитарных наук НИУ ВШЭ – Нижний Новгород

МОДЕЛИРОВАНИЕ ЯЗЫКОВОЙ ЛИЧНОСТИ АВТОРА ПИСЬМЕННОГО ТЕКСТА С ПОМОЩЬЮ МЕТОДОВ ТЕКСТОМАЙНИНГА И МОДЕЛЬНОЙ ЛИНГВИСТИКИ

Нижний Новгород, 2020

Грант РФФИ в рамках научного проекта № 19-31-27001 "Аспиранты" (№ 19-312-90022)

Атрибуция (от лат. attributio - приписывание) – это установление принадлежности анонимного художественного произведения определенному автору, местной или национальной художественной школе, а также определение времени его создания [Популярная художественная энциклопедия 1986, URL: https://rus-pictures-enc.slovaronline.com/285-Атрибуция (дата обращения 07.10.2018 г.)].

Автором текста считается «создатель какого-нибудь произведения» [Толковый словарь русского языка, 1994, URL: http://www.slovopedia.com/4/192/640227.html (дата обращения 07.10.2018 г.)]. Автор обладает основной для атрибуции именно его текстов особенностью – идиостилем, или индивидуальным стилем.

Идиостиль – «совокупность языковых и стилистико-текстовых особенностей, свойственных речи писателя, ученого, публициста, а также отдельных носителей данного языка» [Стилистический энциклопедический словарь русского языка, 2011: 95].

Грант РФФИ в рамках научного проекта № 19-31-27001 "Аспиранты" (№ 19-312-90022)

СОВРЕМЕННОЕ СОСТОЯНИЕ

Квантитативный подход	Качественный анализ текста
- программы и алгоритмы атрибуционного анализа: -по количественному анализу лексических составляющих текста; -по количественному анализу синтаксических составляющих текста; -по классическим статистическим параметрам текста (текстовые длины: слов, предложений; количества: слов, предложений и пр.);	- идиостиль (набор авторски маркированных речевых характеристик на всех уровнях языка); - речевые компетенции (авторский речевой потенциал: не только то, что автор сказал, но и то, что он может сказать); - психотипы (по Валерию Павловичу Белянинну: светлые, темные, активные и пр. тексты);

Грант РФФИ в рамках научного проекта № 19-31-27001 "Аспиранты" (№ 19-312-90022)

ПОЧЕМУ ИНТЕГРАЦИЯ?

стилеметрия

- Л. Кэмпбелл [Campbell 1867] и В. Лютославский [Lutoslawski 1897] на западе и Н.А. Морозова [Морозов 1916] в России

- [Mendenhall 1887, Mosteller, Wallace 1964, Захаров 2000, , Merriam 2003, Labbe, Labbe, 2001, Juola, Sofko, Brennan 2006, Мартыненко 2015, Litvinova, Seredin, Litvinova, etc., 2017, Wright 2017, Karlgren, Esposito, Gratton, etc. 2018, и пр.]

квалификативные исследования

- [Вул 1973, 2007; Горошко 2003, Комиссаров 2001, McMenamin 2002, Галяшина 2003, Coulthard 2004 и пр.]

- объяснение длины предложения [Степаненко 2017:19-20];
- объяснение **n**-грамм [Захаров, Хохлова 2008: 41-42];
- исследования глубинных синтаксических структур [Марусенко 1990; Родионова 2008 и пр.].

ПОЧЕМУ ИНТЕГРАЦИЯ?

- **Идея**

- интеграции анализа традиционных стилостатистических параметров (длин слов и предложений, наиболее частотных n-грамм, служебных слов и POS-tags) и анализа авторских идиосинкразем, в основном ошибок разного рода, предложенный [Koppel, Schler 2003]

- **Необходимость**

Междисциплинарность в науке

Нужды прикладных областей знания (судебное авторoведение):
законодательство [Приказ от 27 декабря 2012 года N 237; Федеральный закон от 31 мая 2001 г. N 73-ФЗ];
методика [Рубцова, Ермолаева, Безрукова и др. 2007]

НЕОБХОДИМОСТЬ - ИНТЕГРАЦИЯ КОЛИЧЕСТВЕННОГО И КАЧЕСТВЕННОГО ПОДХОДОВ

1) **Баранов** А.Н. Введение в прикладную лингвистику: Учебное пособие. — М.: Эдиториал УРСС, 2001. — 360 с.
2) **Koppel**, M., **Schler**, J. (2003). Exploiting Stylistic Idiosyncrasies for Authorship Attribution. Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, 69, 72-80.
3) Хоменко А. Ю. Алгоритм для автоматической идентификации автора письменного речевого произведения в судебном автороведении // Юрислингвистика. 2014. № 3. С. 83-93.
4) Хоменко А. Ю. Атрибуция текстов малого объёма. Статистические закономерности. Язык. Право. Общество: сб. ст. V Междунар. науч.-практ. конф. Пенза : ПГУ, 2018. С. 123-127.
5) Хоменко А. Ю. К вопросу об исследовании письменного речевого произведения в рамках автороведческой экспертизы на предмет его оригинальности // Политическая лингвистика. 2014. № 4. С. 306-312.
6) Хоменко А. Ю. Лингвистическая атрибуционная экспертиза в отечественной и зарубежной школах. перспективы развития автороведческих методик в России. Международная научная конференция «Современная теоретическая лингвистика и проблемы судебной экспертизы». М. : Государственный институт русского языка им. А.С. Пушкина, 2019. С. 536-550.
7) Хоменко А. Ю. Лингвистическое атрибуционное исследование коротких письменных текстов: качественные и количественные методы // Политическая лингвистика. 2019. № 2 (74). С. 177-187.
8) Хоменко А. Ю. Лингвистическое моделирование как инструмент выявления искажений речевых навыков автора письменного речевого произведения. Опыт практического исследования // Вопросы психолингвистики. 2018. № 2 (36). С. 209-226.
9) Хоменко А. Ю. Моделирование когнитивных структур на основе методик автороведческого анализа. Когнитивные исследования языка. Вып. XXXVII: Интегративные процессы в когнитивной лингвистике: материалы международного конгресса по когнитивной лингвистике / Отв. ред.: Т. В. Романова. Т. XXXVII: Интегративные процессы в когнитивной лингвистике: материалы международного конгресса по когнитивной лингвистике. Деком, 2019. С. 1069-1074.
10) **Pimonova** E., **Durandin** O., **Malafeev** A. Doc2vec or better interpretability? a method study for authorship attribution Paper presented at Dialogue 2020, Moscow, June 15–20, 2020. http://www.dialog-21.ru/media/4955/pimonovaeplusetal-132.pdf (2020) (дата обращения: 05.08.2020 г.).

Грант РФФИ в рамках научного проекта № 19-31-27001 "Аспиранты" (№ 19-312-90022)

ИНТЕГРАТИВНАЯ МЕТОДИКА, ИМЕЮЩАЯ СЛЕДУЮЩИЕ ИТЕРАЦИИ

1) автоматическое извлечение из текста параметров, описывающих идиостиль с точки зрения прагматикона, тезауруса и лексикона автора;
2) поиск традиционных стилеметрических текстовых данных;
3) присвоение веса каждому параметру;
4) построение математических моделей сравниваемых текстов;
5) сравнение математических моделей с целью выявления уровня их корреляции между собой.

Грант РФФИ в рамках научного проекта № 19-31-27001 "Аспиранты" (№ 19-312-90022)

МЕТОДИКО-МЕТОДОЛОГИЧЕСКАЯ БАЗА

1) методы когнитивной лингвистики и психолингвистики, как то: анализ языковой личности автора письменного текста по методике Ю.Н. Караулова, анализ компетенций языковой личности по методике С.М. Вула и Е.С. Горошко;

2) традиционный структуралистский подход к языку как уровневой системе (с применением морфологического, словообразовательного, лексического, синтаксического, семантического анализов);

3) компьютерные методы анализа речевых структур (преимущественно автоматическая обработка текстов с целью вычленения идентификационных параметров)

Основным **инструментом** исследования стал язык программирования «Python».

Грант РФФИ в рамках научного проекта № 19-31-27001 "Аспиранты" (№ 19-312-90022)

ГИПОТЕЗА ИССЛЕДОВАНИЯ

Интегративная атрибуционная модель, получаемая в результате ряда итераций, способна успешно решать идентификационную задачу атрибуционной лингвистики на текстах любого объема.

Грант РФФИ в рамках научного проекта № 19-31-27001 "Аспиранты" (№ 19-312-90022)

Методики атрибуции становятся необходимыми:

1) в филологических экспертизах при определении авторства известных художественных произведений (статьи Ф.М.Достоевского, работы М.А.Шолохова);
2) в судебных автороведческих экспертизах при решении диагностических и идентификационных задач, при анализе контента сети Интернет на предмет автоматического поиска содержания деликтной направленности;
3) при решении научных задач, связанных с идентификацией лиц по письменной речи и определением характеристик пола, возраста, социального статуса этих лиц.

Грант РФФИ в рамках научного проекта № 19-31-27001 "Аспиранты" (№ 19-312-90022)

ПЕРВЫЙ ЭТАП РАБОТЫ

[Campbell 1867], [Lutoslawski 1897], [Морозов 1916], [Mendenhall 1887, Mosteller, Wallace 1964, Захаров 2000, , Merriam 2003, Labbe, Labbe, 2001, Juola, Sofko, Brennan 2006, Мартыненко 2015, Litvinova, Seredin, Litvinova, etc., 2017, Wright 2017, Karlgren, Esposito, Gratton, etc. 2018, и пр.], [Вул 1973, 2007; Горошко 2003, Комиссаров 2001, McMenamin 2002, Галяшина 2003, Coulthard 2004 и пр.], [Степаненко 2017:19-20], [Захаров, Хохлова 2008: 41-42], [Марусенко 1990; Родионова 2008 и пр.], [Koppel, Schler 2003], [Баранов 2001: 43–52; Хоменко 2013; 2014а,б,в, 2018, 2019а,б,в, г, 2020; Pimonova, Durandin, Malafeev 2020] и пр.

Ежегодная международная конференция «Диалог» (URL: http://www.dialog-21.ru/); Международная конференция «AINL: Artificial Intelligence and Natural Language Conference» (URL: https://ainlconf.ru/); Международная конференция «International Conference on Analysis of Images, Social Networks and Texts» (URL: https://aistconf.org/); серия мероприятий «PAN» в рамках «Conference and Labs of the Evaluation Forum, или Cross-Language Evaluation Forum», (URL: https://pan.webis.de/):

[Bacciu, Morgia, La 2019], [Litvinova, Sboev, Panicheva 2018], [Custódio, Paraboni 2018], [Murauer, Tschuggnall, Specht 2018], [Muttenthaler, Lucas, Amann 2019], [Litvinova, Sboev, Panicheva 2018], [Custódio, Paraboni 2018], [Panicheva, Mirzagitova, Ledovaya 2018], [Gomzin, Laguta, Stroev 2018], [Korobov 2015], [Bacciu, Morgia, La 2019].

Грант РФФИ в рамках научного проекта № 19-31-27001 "Аспиранты" (№ 19-312-90022)

Архитектура модели

реализация прагматикона личности на синтаксическом уровне	описание тезауруса личности	вербально-семантический уровень авторского лексикона
вводные слова и конструкции, эксплицирующие субъективную модальность; целевые, выделительные и сравнительные обороты, репрезентирующие уровень освоения автором компетенций письменной речи и его отношение к действительности; синтаксические сращения, дающие представление в том числе о функциональной стилистической отнесенности текста; сравнительные придаточные, глагольные односоставные предложения, эксплицирующие репрезентацию действительности в текстовом материале; обращения;	наиболее частотные сочетания слов, которые описывают грамматико-семантические особенности текста; ключевые лексемы текста; экспликаты аксиологических текстовых доминант дихотомии «свой/чужой»; Пример формализованного правила: экспликаты субъективной модальности (вводные слова): 1)__,Prnt,__ 2)<начало предложения> Prnt,__ со списком вводных слов	частеречная отнесенность слов текста (количество глаголов, прилагательных, существительных и пр. частей речи), сложные слова полуслитного написания; модальные частицы, междометия, наличие/отсутствие модального постфикса «-то», предпочтительные слова-интенсификаторы.

Грант РФФИ в рамках научного проекта № 19-31-27001 "Аспиранты" (№ 19-312-90022)

Архитектура прототипа ПО

Стилостатистика

- Индекс удобочитаемости Флеша-Кинкейда
- Индекс туманности Ганнинга
- Средняя длина слова (в буквах)
- Средняя длина предложения (в словах)
- Количество предложений длиннее 8-ми слов
- Коэффициент предметности (Pr)
- Коэффициент качественности (Qu)
- Коэффициент активности (Ac)
- Коэффициент динамизма (Din)
- Коэффициент связности текста (Con)

Реализация прагматикона языковой личности

- Предложения с однородными рядами
- Предложения с обособленными приложениями
- Вводные слова и конструкции
- Целевые и выделительные обороты
- Синтаксические сращения
- Сравнительные придаточные
- Конструкции с сопоставительными союзами
- Вставные конструкции
- Сложные синтаксические конструкции
- Глагольные односоставные предложения
- Обращения

Описание тезауруса языковой личности

- Ключевые слова
- Наиболее частотные биграммы
- Наиболее частотные триграммы
- Дихотомия "свой/чужой"

Экспликация вербально-семантического уровня языковой личности

- Сложные слова полуслитного написания
- Модальные частицы
- Междометия
- Наличие/отсутствие модального постфикса «-то»
- Предпочтительные слова-интенсификаторы
- Количество слов несловарного написания

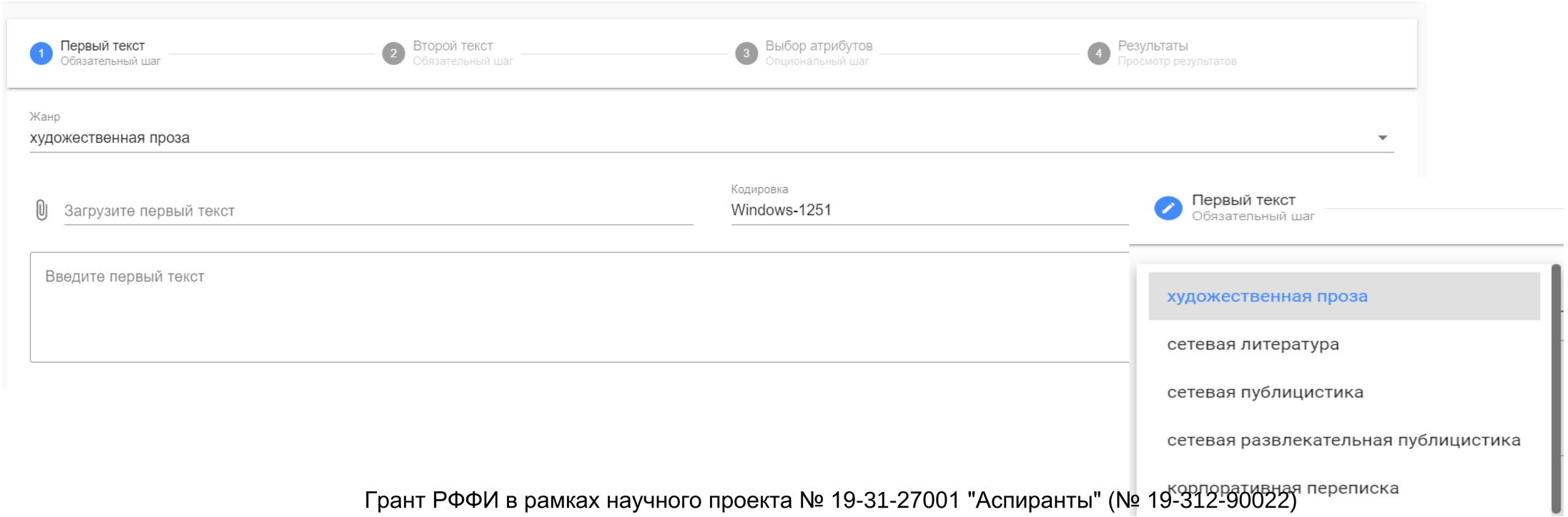
Набор предустановленных текстовых параметров ресурса: http://khorom-attribution.ru/#/

Грант РФФИ в рамках научного проекта № 19-31-27001 "Аспиранты" (№ 19-312-90022)

Описанная выше модель получила реализацию в виде прототипа программного продукта, размещенного в открытом доступе в сети Интернет, URL: http://khorom-attribution.ru/#/

Модуль пользователя:

Ввод данных: на вход подаются два текста A и B; пользователь имеет возможность ввести текст с клавиатуры или загрузить его с компьютер, а также выбрать жанр



Грант РФФИ в рамках научного проекта № 19-31-27001 "Аспиранты" (№ 19-312-90022)

Пользователь может строить модель не только на основе предустановленных параметров, но и имеет возможность выбирать те, которые считает наиболее релевантными для определенной пары текстов, те, которые, по его мнению, дают наибольший прирост информации (функция множественного выбора)

Описание тезауруса языковой личности

- [x] Ключевые слова
- [ ] Наиболее частотные биграммы
- [ ] Наиболее частотные триграммы
- [x] Дихотомия "свой/чужой"

Экспликация вербально-семантического уровня языковой личности

- [x] Сложные слова полуслитного написания
- [ ] Модальные частицы
- [x] Междометия
- [ ] Наличие/отсутствие модального постфикса «-то»
- [x] Предпочтительные слова-интенсификаторы
- [ ] Количество слов несловарного написания

Визуализация функции множественного выбора ресурса: http://khorom-attribution.ru/#/

Вывод данных: в качестве результата выводится значения коэффициента корреляции Пирсона, значение линейной регрессии, критерия Стьюдента для моделей двух сравниваемых текстов, а также значения метрик каждого параметра для двух текстов

Коэффициент корреляции Пирсона: 1

Линейная регрессия: p-value - 0, r-value - 1, stderr - 0.01

t-критерий Стьюдента: p-value - 0.99, statistic - 0.01

Корреляция по ключевым словам: -0.22

Корреляция по словам-интенсификаторам: -0.48

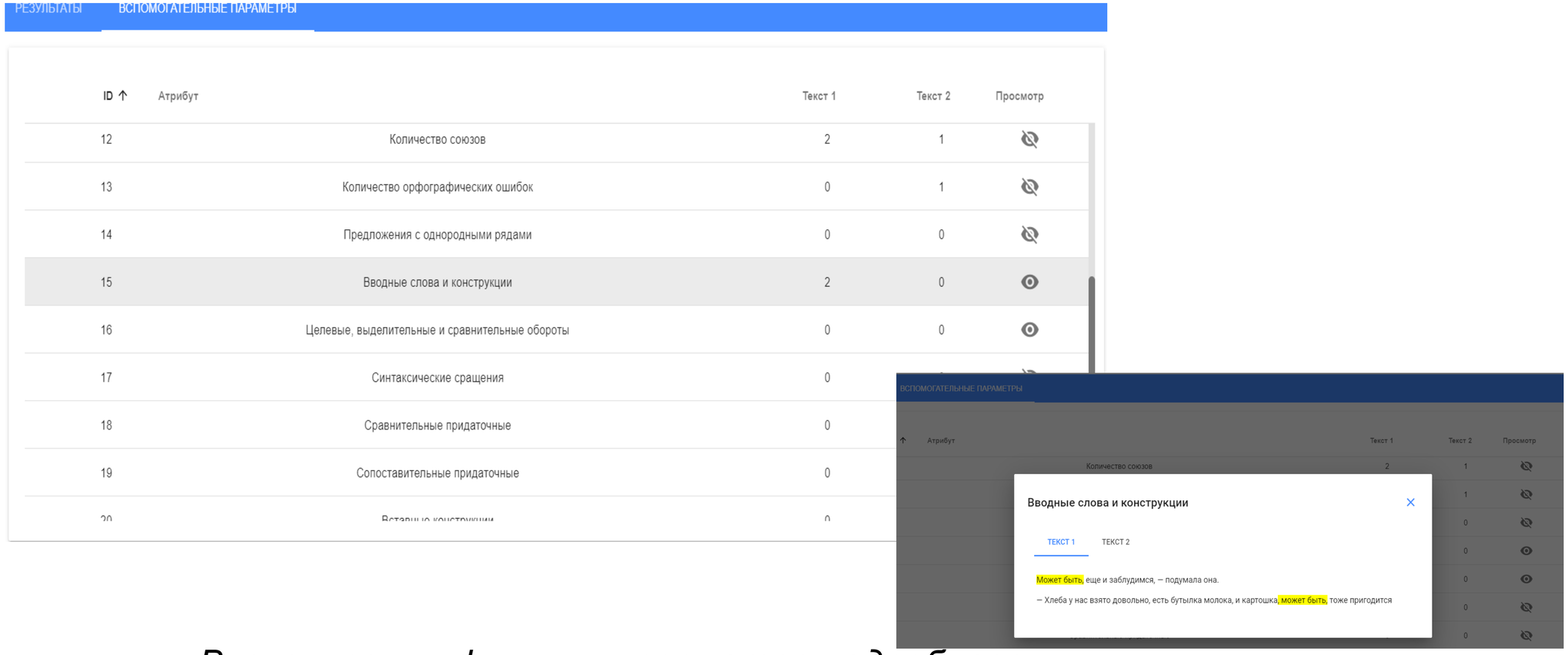
Корреляция по биграммам: -0.21

Корреляция по триграммам: -0.86

ID ↑	Атрибут	Текст 1	Текст 2
1	Индекс удобочитаемости Флеша-Кинкейда	12.7025	14.6661
2	Индекс туманности Ганнинга	15.8154	18.4731
3	Средняя длина слова (в буквах)	5.0144	5.4463
4	Средняя длина предложения (в словах)	9.9518	9.48
5	Количество предложений длиннее 8-ми слов	451754.386	442857.1429
6	Коэффициент предметности (Pr)	1.081	1.2082
7	Коэффициент качественности (Qu)	0.2957	0.3731
8	Коэффициент активности (Ac)	0.2234	0.2019

Визуализация вывода результата ресурса:
http://khorom-attribution.ru/#/

Вывод данных: для пользователя также запрограммирован модуль проверки: выбрав вкладку Вспомогательные параметры можно просмотреть все выделенные в тексте компоненты, для которых рассчитаны относительные частоты и иные метрики



Визуализация функции просмотра подробного результата исследования на ресурсе: http://khorom-attribution.ru/#/

Материал исследования

Ввод данных: на вход подаются два текста A и B; пользователь имеет возможность ввести текст с клавиатуры или загрузить его с компьютер, а также выбрать жанр

1) Подкорпус текстов сетевой газеты «The Village» (сетевая публицистика): включает в себя тексты 3 авторов-женщин, 3 авторов-мужчин; всего 600 текстов
2) Подкорпус текстов развлекательного портала «ЯПлакалъ» (развлекательная публицистика, комментарии, форум): включает в себя тексты 3 авторов-женщин, 3 авторов-мужчин; всего 600 текстов
3) Подкорпус текстов ресурса сетевой литературы «Книга фанфиков» (https://ficbook.net/): включает в себя тексты 3 авторов-женщин, 4 авторов-мужчин; всего 190 текстов
4) Подкорпус текстов художественной литературы (нежанровая проза), включающий тексты С.Д.Довлатова и В.П.Астафьева: 10 текстов (С.Д.Довлатова «Наши», «Чемодан», «Иностранка», «Заповедник», «Зона: Записки надзирателя», «Встретились, поговорили», В.П.Астафьева «Обертон», «Последний поклон», «Звездопад», «Так хочется жить»);
5) Подкорпус текстов корпоративной русскоязычной переписки: 2 автора-женщины, 2 автора-мужчины, всего 218 текстов.

Грант РФФИ в рамках научного проекта № 19-31-27001 "Аспиранты" (№ 19-312-90022)

Выводы

В настоящий момент происходит процесс финального тестирования и отладки ресурса. Данные текстовые коллекции позволяют разработчикам всесторонне проверить его работоспособность и валидность получаемых результатов.

Но уже на этом этапе можно говорить о том, что наибольший прирост информации при определении автора текста любого жанра и объема дают такие параметры, как:

- слова-интенсификаторы, количество слов несловарного написания на вербально-семантическом уровне языковой личности;
- ключевые слова, показатели дихотомии «свой/чужой», наиболее частотные биграммы на уровне тезауруса личности;
- сложные и осложненные синтаксические конструкции (вид конструкции зависит от жанра текстов: для публицистики, например, это предложения с однородными рядами, предложения с обособленными приложениями, конструкции с сопоставительными союзами, вставные конструкции; для художественных текстов это глагольные односоставные предложения, придаточные сравнительные, сложные синтаксические конструкции, вставные конструкции) на уровне прагматикона личности.

Грант РФФИ в рамках научного проекта № 19-31-27001 "Аспиранты" (№ 19-312-90022)

Спасибо за внимание!

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Хоменко Анна Юрьевна:
akhomenko@hse.ru
https://www.hse.ru/org/persons/65858472