Khomenko Anna
HSE University

Baranova Yulia
CarrierX

Romanov Alexander
Tomsk State University of Control Systems and Radioelectronics

Zadvornov Konstantin
HSE University

LINGUISTIC MODELLING AS A BASIS FOR CREATING AUTHORSHIP ATTRIBUTION SOFTWARE

2021

The reported study was funded by RFBR, project number 19-31-27001

INTEGRATION → UNDERSTANDING OF THE INDIVIDUAL STYLE

a combination of language probabilities

- author's competencies in the traditional sense (the personality's thesaurus, pragmaticon, grammar and lexicon)

a specific language personality representation

- stylometry allows to objectify the interpretative analysis results

MODELLING of a language personality →
for texts of different genres and lengths

THEORETICAL BASIS

KARAULOV, YU. N. (1987). THE RUSSIAN LANGUAGE AND THE LANGUAGE PERSONALITY. MOSCOW, NAUKA, 264 P.

- pragmaticon level (speech strategies and tactics)
- thesaurus level (cognitive worldview)
- lexicon level (lexical and grammatical competence)

STRUCTURE MODELLING. FINDING EXPLICATORS OF SUBJECTIVE MODALITY

- a dictionary of subjective modality explicators is created;
- a punctuation rule, which allows to overcome homonymy is prescribed:

1) __, Prnt, __

2) <beginning of a sentence> Prnt, __

where *Prnt* is any part of speech; __ - some part of a sentence, < *beginning of a sentence* > - designation of a sentence beginning, «,» - comma, corresponding punctuation mark.

STRUCTURE MODELLING. PURPOSE SYNTACTIC STRUCTURE

‘с целью/из расчёта’ (for the purpose of/in order to) + INFN, where *INFN* – infinitive

STRUCTURE MODELLING. UNITS OF AUTHOR'S LEXICON

Modal postfix '-то' rule is the following: POST-то, other than NPRO or APRO in any case in plural or single form, where *POST* – any part of speech, *NPRO/APRO* – pronoun with noun/adjective semantics and syntactic function

STRUCTURE MODELLING. INTENSIFIER RULE

(1) *Какая красота!*
What a beauty!
— in this case, the pronoun *какая* (what) serves as an intensifier.

ADJ in direct cases in singular or plural form + NOUN, where *ADJ* – adjective:

(2) *Настоящий бардак*.
Real mess.

APPLICATION

KhoRom

http://khorom-attribution.ru/#/

APPLICATION ARCHITECTURE: KHOROM

1. Input

2. Automatic extraction of parameters describing author's individual style

2.1. Preprocessing

2.1.1. Sentence-splitting

2.1.2. Tokenization

2.1.3. Morphological parsing

2.2. Processing

2.2.1. Stylometry block

2.2.1.1. Calculation of basic metrics (number of words, sentences)

2.2.1.2. Search for traditional stylometric textual data (n-grams, indices)

2.2.2. Cognitive block

2.2.2.1. Search for parameters by preset rules

2.2.2.2. Assigning weight to each parameter

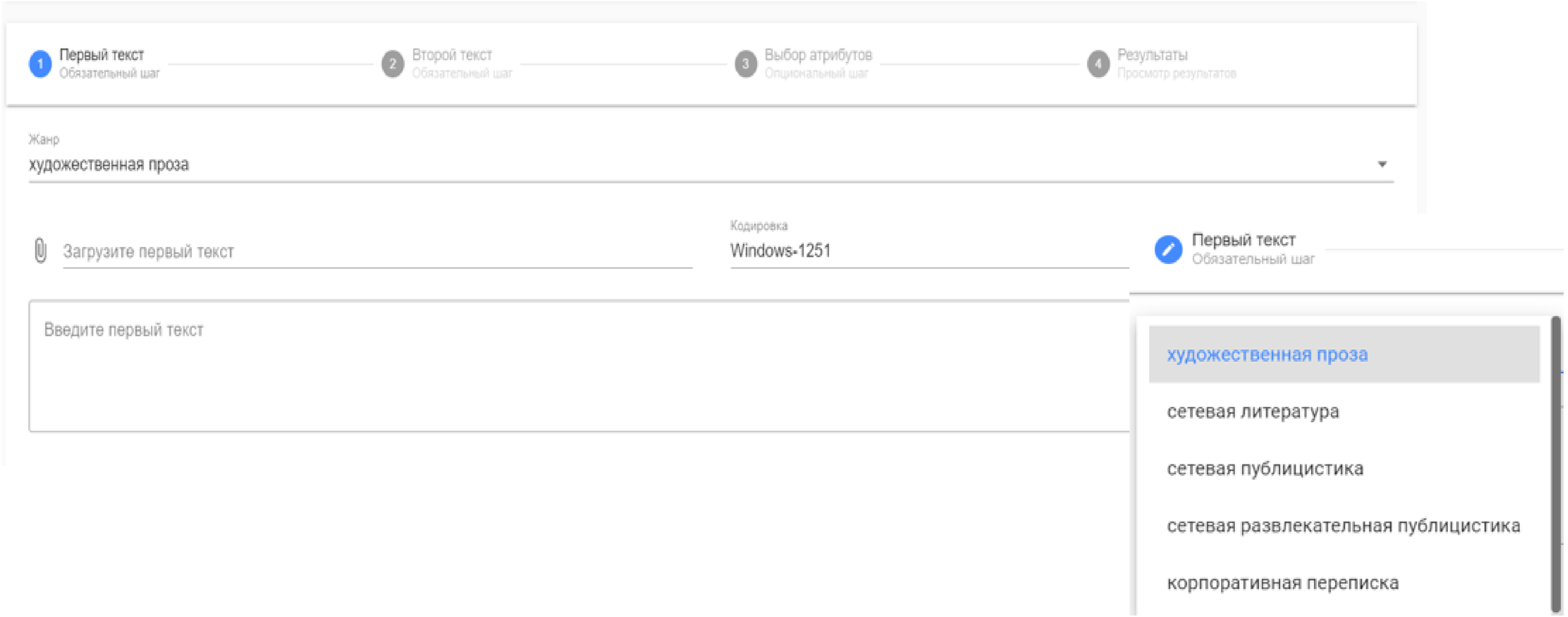
3. Building of mathematical models of the texts being compared: attribute presentation as a sequence of numerical features

4. Comparing the mathematical models

5. Sending the results to a client

Python + Javascript

THE USER MODULE. INPUT



THE USER MODULE. PRESET PARAMETERS

Стилостатистика

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

Реализация прагматикона языковой личности

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

Описание тезауруса языковой личности

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

Экспликация вербально-семантического уровня языковой личности

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

THE USER MODULE. THE MULTIPLE CHOICE FUNCTION

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

THE USER MODULE. OUTPUT

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

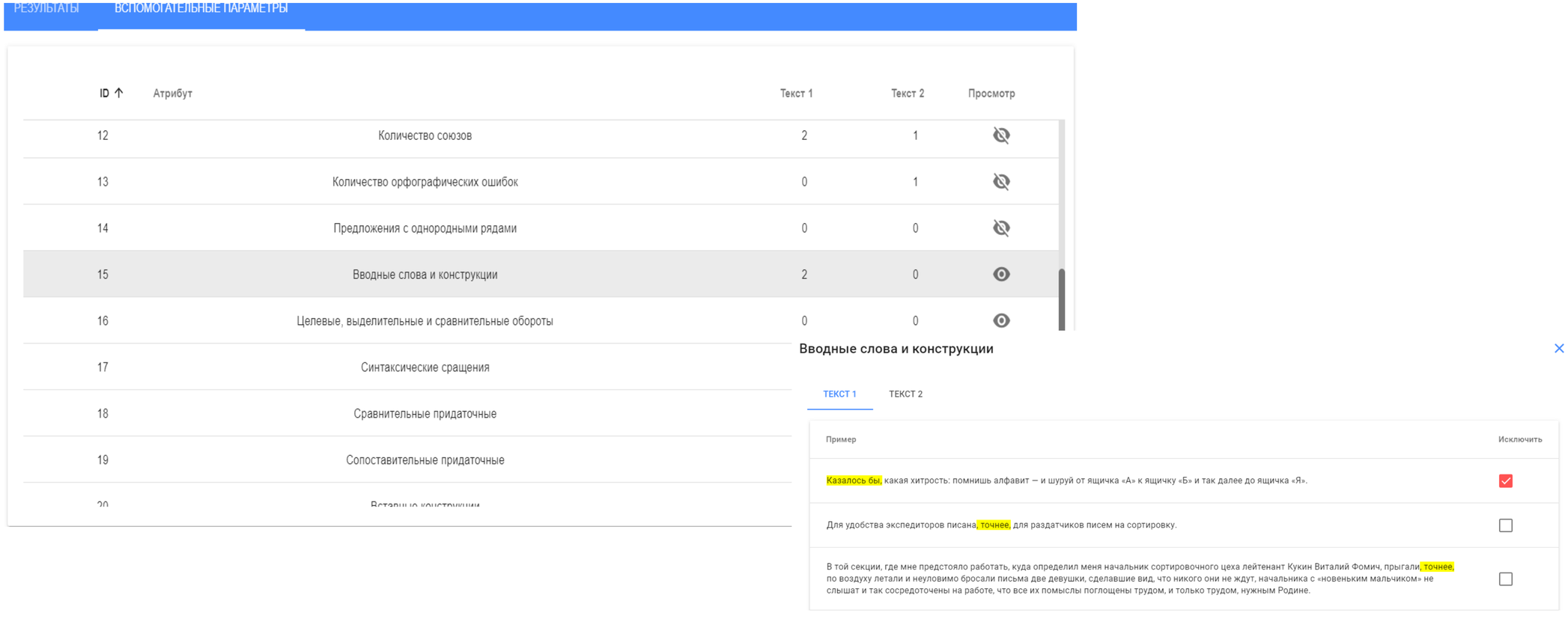
Корреляция по словам-интенсификаторам: -0.48

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

THE USER MODULE.
SOFTWARE WORK VERIFICATION BY THE USER



THE RESULTS OF THE ALGORITHM WORK

1) a collection of **fiction texts** (non-genre prose, famous fiction): includes 10 texts by S. Dovlatov and V. Astafiev: (the average text length is about 20,000 words). Accuracy, precision and recall is 100%, **F-score 1**.

2) a collection of texts from ‘Kniga Fanficov’ (**modern Internet fiction**), URL: https://ficbook.net/ (the average text length is 1,500 to 40,000 words): includes texts of 3 female and 4 male authors; the total of 190 texts. Accuracy – 83%, precision – 67% and recall –100%, **F-score 0,8**;

3) a collection of texts from ‘The Village’ (**online journalism**), URL: https://www.the-village.ru/ (the average text length is 500 to 1,500 words): includes texts of 3 female and 3 male authors; the total of 600 texts. Accuracy, precision and recall is 100%**, F-score 1**;

4) a collection of texts from entertainment portal ‘YaPlakal’ (**entertaining journalism, e-comments**), URL: https://www.yaplakal.com/ (the average text length is 50 to 100 words): includes texts of 3 female and 3 male authors; the total of 600 texts. **Accuracy – 40%,** precision – 0 and recall – 0.

5) a collection of **Russian business e-correspondence texts**: 2 female and 2 male authors; the total of 218 texts; the average text length is 50 to 500 words. Accuracy – 80%, precision – 67% and recall – 100%, **F-score 0,8**

FINDINGS

1) t-test is the most informative indicator for fictional discourse and e-correspondence;

2) t-test is significantly less relevant for journalistic texts;

3) to determine the author of a paper text the values of the correlation and determination coefficients must reach 1;

4) for modern fiction, the stylometry pool (lengths, indices) is uninformative;

5) for short text messages it is necessary to create a representative sample of at least 500 words.

6) the texts of different genres could be examined using the integrative technique (accuracy – 80%, precision –100% and recall – 67%? F-score 0,8): journalistic text could be compared with e-correspondence, for example.

CONCLUSION

The attribution algorithm based on integration of statistically objectified interpretative methods is rather effective:

forensic purposes

studying language personalities of writers, journalists, politicians

examining general language personalities of social groups, subcultures

Thank you very much for your attention!

The reported study was funded by RFBR, project number 19-31-27001