Политическая лингвистика. 2022. № 3 (93).
Political Linguistics. 2022. No 3 (93).

УДК 81.139:81'42
ББК Ш105.51 *ГСНТИ 16.31.21* *Код ВАК 10.02.19 (5.9.8)*

Анна Юрьевна Хоменко
Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород, Россия,
khomenko.anna.1989@gmail.com, https://orcid.org/0000-0003-3564-6293

Лингвистическое моделирование как основа для создания полуавтоматического атрибуционного алгоритма

АННОТАЦИЯ. *В статье речь идет об апробации интегративного атрибуционного алгоритма. Он основан на анализе идиостиля автора письменного текста методами интерпретативной лингвистики с последующей объективацией полученных данных с помощью математической статистики. Алгоритм решает идентификационную проблему атрибуции. Выбор параметров, описывающих индивидуальный стиль автора, основан на рассмотрении текста как продукта аутентичной языковой личности. Языковая личность описывается с использованием психолингвистических (Ю. Н. Караулов), социолингвистических и судебно-лингвистических (С. М. Вул, M. Coulthard, R. W. Shuy) методов. Для проверки гипотезы о том, что именно интегративная методика является наиболее эффективной при решении идентификационной задачи атрибуции, было создано электронное приложение «ХоРом», кумулирующее в себе описанные выше подходы к анализу языковой личности: http://khorom-attribution.ru/#/. С помощью ресурса можно сравнить две модели языковой личности и определить уровень их сходства посредством следующих метрик: коэффициента корреляции Пирсона, коэффициента детерминации линейной регрессии и t-критерия Стьюдента. Важно, что приложение также отображает интерпретируемую модель языковой личности, давая пользователю информацию о значении показателей каждого параметра. Система имеет обширный функционал, включая выбор параметров, просмотр реализации параметров в тексте документа и внесение изменений в окончательный список реализаций параметров (в случае неточности программы пользователь имеет возможность исправить ее работу вручную). Созданное программное обеспечение является лишь частью атрибуционного алгоритма. Полученные данные математической статистики необходимо анализировать экспертным путем с помощью разработанных для алгоритма методических рекомендаций. Эффективность методики доказана посредством ее апробации на текстах разного объема и жанровой отнесенности: был проанализирован ряд текстов художественного, публицистического, официально-делового, обиходно-бытового стилей. Для текстов всех дискурсов, кроме обиходно-бытового, разработанный алгоритм показал высокий уровень точности (F-мера от 0,8 до 1). Для улучшения работы алгоритма на текстах обиходно-бытового стиля автором исследования разработан ряд улучшений, планирующихся к внесению в алгоритм.*

КЛЮЧЕВЫЕ СЛОВА: *атрибуция, языковая личность, автоматическая обработка текста, лингвистические модели, математические модели, атрибутивное программное обеспечение.*

БЛАГОДАРНОСТИ: исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-312-90022.

ИНФОРМАЦИЯ ОБ АВТОРЕ: *Хоменко Анна Юрьевна, кандидат филологических наук, старший преподаватель, департамент прикладной лингвистики и иностранных языков, Национальный исследовательский университет «Высшая школа экономики» (Нижний Новгород, Россия); 603155, Россия, Нижний Новгород, ул. Большая Печерская, 25.*

ДЛЯ ЦИТИРОВАНИЯ: *Хоменко, А. Ю.* Лингвистическое моделирование как основа для создания полуавтоматического атрибуционного алгоритма / А. Ю. Хоменко. — Текст : непосредственный // Политическая лингвистика. — 2022. — № 3 (93). — С. 90-100.

Anna Yu. Khomenko
National Research University “Higher School of Economics” (Nizhny Novgorod, Russia), Nizhny Novgorod, Russia,
khomenko.anna.1989@gmail.com, https://orcid.org/0000-0003-3564-6293

Linguistic Modeling as a Basis for Creating Half-Automatic Authorship Attribution Algorithm

ABSTRACT. *This paper discusses the testing procedure and results of an integrative attribution analysis algorithm. It is based on the analysis of the individual style of an author of a written text via the methods of interpretative linguistics and further objectification of the data received through the usage of mathematical statistics methods. The algorithm solves the identification problem of authorship attribution. The choice of the parameters describing the author’s individual style is based on the interpretation of the text as a product of an authentic linguistic personality. The linguistic personality is described using psycholinguistic (Yu. N. Karaulov), sociolinguistic and forensic linguistic (S. M. Vul, M. Coulthard, R. W. Shuy) methods. To test the hypothesis that it is the integrative method that is the most effective way of performing the identification task of attribution, the author has created the electronic software “KhoRom” accumulating the approaches to the analysis of a linguistic personality described above: http://khorom-attribution.ru/#/ . With the help of this program, one can compare two models of linguistic personality and determine the level of their similarity through the following metric values: Pearson's*

correlation coefficient, coefficient of determination of linear regression, and Student's t-test. It is important that the application also reflects the model of the linguistic personality under interpretation, providing the user with information about the values of the indicators of each parameter. The system has extensive functionality including a multiple choice of parameters, an opportunity to view the realization of parameters in the text of the document and make changes in the final list of parameter realizations (in case of program inaccuracy, a user has the opportunity to correct its work manually). The software is only a part of the attribution algorithm. The mathematical statistics obtained should be analyzed by experts in accordance with the user manual developed for the algorithm. The effectiveness of the methodology was proved by testing it on texts of different volume and genres: fiction, journalistic, official, and colloquial styles were analyzed. The algorithm showed high level accuracy (F-score from 0.8 to 1) for texts of all kinds of discourse, except colloquial. To improve the work of the algorithm with colloquial texts, the author of the study has developed a number of improvements that are planned to be introduced into the algorithm.

KEYWORDS: *attribution, linguistic personality, automatic text procession, linguistic models, mathematical models, attribute software.*

ACKNOWLEDGMENTS: research is accomplished with financial support of the Russian Foundation for Basic Research (RFBR), project number 19-312-90022.

AUTHOR'S INFORMATION: *Khomenko Anna Yur'evna, Candidate of Philology, Senior Lecturer of Department of Applied Linguistics and Foreign Languages, National Research University "Higher School of Economics" (Nizhny Novgorod, Russia), Nizhny Novgorod, Russia*.

FOR CITATION: *Khomenko A. Yu*. (2022). Linguistic Modeling as a Basis for Creating Half-Automatic Authorship Attribution Algorithm. In *Political Linguistics*. No 3 (93), pp. 90-100. (In Russ.).

1. ВВЕДЕНИЕ

Настоящее исследование посвящено разработке проблем текстовой атрибуции на основе постулатов лингвистики моделей. Автороведение кумулирует в себе две основные задачи: идентификационную (собственно определение авторства: например, [Juola 2006]) и диагностическую (определение социально маркированных авторских характеристик: гендера [Хазова 2020; Степаненко 2017], возраста, социального статуса автора [Shuy 2005]). В работе речь пойдет о решении прежде всего идентификационной задачи закрытого класса (с ограниченным количеством авторов) при попарном сравнении письменных текстов.

На современном этапе развития текстовой атрибуции координация между стилеметрией и квалификативным подходом происходит в основном посредством объяснения стилеметрических данных с точки зрения традиционной интерпретативной лингвистики: объяснение длины предложения как отражения уровня компетенций автора в письменной речи [Степаненко 2017: 19—20], объяснение n-грамм как косвенной экспликации грамматических текстовых реалий [Захаров, Хохлова 2008: 41—42].

Следует отметить, что многие из квантитативных походов продуктивны [Korobov 2015; Murauer, Tschuggnall, Specht 2018; Muttenthaler, Lucas, Amann 2019; Litvinova, Sboev, Panicheva 2018; Custódio, Paraboni 2018; Gomzin, Laguta, Stroev 2018; Panicheva, Mirzagitova, Ledovaya 2018; Bacciu, Morgia, La 2019], но все они рассматривают индивидуальный стиль как череду языковых вероятностей, а не как продукт формирования речевой способности индивида. Тем не менее с помощью только квантитативных подходов, основанных на сборе данных о неких традиционных стилеметрических параметрах, пусть даже и в большом их количестве [Bhargava, Mehndiratta, Asawa 2013], невозможно создать полную, адекватно отражающую оригинал модель идиостиля, являющуюся экспликацией языковой личности автора в материальном эквиваленте.

Представляется логичным путь исследования глубинных синтаксических структур как базы для сравнения моделей индивидуальных авторских стилей. Разработкой данного направления занимается санкт-петербуржская школа прикладной лингвистики [Мартыненко 1988; Марусенко 1990; Родионова 2008 и др.]. Этот подход, безусловно, работоспособен, но его реализация возможна, с одной стороны, только на объемных текстах, с другой — она очень трудоемка и сложна в техническом отношении.

Более просто реализуемым выглядит подход, основанный на интеграции анализа традиционных стилостатистических параметров (длин слов и предложений, наиболее частотных n-грамм, служебных слов и POS-tags) и анализа авторских идиосинкразем [Koppel, Schler 2003], а также характеристик, относящихся к групповым признакам [Абрамкина 2019].

2. АУТЕНТИЧНЫЙ АТРИБУЦИОННЫЙ АЛГОРИТМ

В настоящем исследовании предлагается интегративный подход к решению задач текстовой атрибуции. Методы интерпретативной лингвистики выявляют информацию об авторских компетенциях в традиционном понимании (тезаурус личности, ее прагматикон и лексикон), а стилостатистика дает воз-

можность сделать результаты интерпретативного анализа объективными. Такой подход к анализу текста в теории должен быть универсальным и решать задачи атрибуции как в научных, так и в прагматических целях, в том числе в целях судебного автороведения, где полная автоматизация процесса невозможна, а субъективизм недопустим. Одновременно алгоритм должен решать проблему атрибуции текстов разного объема и жанровой отнесенности.

Предлагаемая методика реализуется по следующему алгоритму: 1) автоматическое извлечение из текста параметров, описывающих прагматикон, тезаурус и лексикон автора; 2) поиск традиционных стилеметрических данных; 3) присвоение веса каждому параметру; 4) построение математических моделей сравниваемых текстов; 5) сравнение математических моделей; 6) экспертный анализ статистических данных. Важно, что речь идет не об аутентичном пути автоматического определения авторства, а о концепции интегративной методики, которая соединяет два подхода, объективируя интерпретацию статистикой с последующим анализом статистических данных.

В основе формализации уровневой структуры языковой личности в исследовании лежат постулаты теории Ю.Н. Караулова [Караулов 2010]. Собственно процесс формализации строится на принципах семантического синтаксиса [Падучева 1974] и в соответствии с правилами грамматики русского языка [Русская грамматика, URL: http://rusgram.narod.ru/index.html].

Структура языковой личности рассматривается как совокупность трех уровней: вербально-семантического, лингвокогнитивного, мотивационного [Караулов 2010].

Языковая личность понимается как результат ее формирования в определенной социальной среде (автобиографический, социолингвистический и юрислингвистический подходы [Виноградов 1961; Coulthard 2004; Shuy 2005; Вул 2007]).

На основании эмпирического исследования 10 текстовых блоков общим объемом около 116 тыс. слов был определен ряд параметров языковой личности, которые заведомо являются важными компонентами авторского идиостиля, материального экспликатора языковой личности пишущего, и одновременно могут быть извлечены из текста автоматически с минимальным препроцессингом. Для извлечения компьютерным способом все формальные правила были запрограммированы и интегрированы в лингвистический ресурс «ХоРом»: *http://khorom-attribution.ru/#/*.

В результате эмпирического исследования на вербально-семантическом уровне были запрограммированы для поиска такие параметры, как частеречная отнесенность слов (количество знаменательных частей речи, соотношение разных частей речи – индекс удобочитаемости, коэффициент предметности и пр.), средние длины слов, наличие/отсутствие сложных слов полуслитного написания; модальные частицы, междометия, наличие/отсутствие модального постфикса «-то», предпочтительные слова-интенсификаторы. Формализованный поиск единиц этого уровня осуществляется в соответствии с морфологической аннотацией текста, то есть посредством присвоения каждому слову тега части речи и тегов всех грамматических категорий, которые присущи этой части речи. Например, поиск элементов с модальным постфиксом «-то» будет осуществляться в соответствии со следующим алгоритмом:

1) + Prnt-то

2) – SPRO, nom / gen / dat / acc/ ins / loc / voc / gen2 / acc2 / loc2, sin / pl

3) – APRO, nom / gen / dat / acc/ ins / loc / voc / gen2 / acc2 / loc2, sin / pl[1].

Так, схему можно прочитать следующим образом: осуществляется поиск любой части речи, имеющей модальный постфикс «-то», кроме местоимений-существительных и местоимений-прилагательных, в любом падеже множественного или единственного числа.

Под словом-интенсификатором подразумевается лексема, используемая для определения степени семантической категории интенсивности. Чаще всего говорят о наречиях-интенсификаторах, круг их хоть и велик, но ограничен (*очень, сильно, адски* — из современного дискурса). Тем не менее категория интенсивности не исчерпывается исключительно наречным наполнением, например: *Какая красота!*, — в данном случае интенсификатором служит местоимение *какая*. Так, в исследовании был создан свод правил для поиска структур с интенсификаторами; в список интенсификаторов входят как наречия с некоторыми грамматическими ограничениями (не осуществляется поиск структур, где наречие не эксплицирует категорию интенсивности, например, является частью составного именного сказуемого: *Он чувствует себя хорошо*), так и некоторые

[1] Здесь и далее используется номенклатура, соответствующая частеречному тегированию в Национальном корпусе русского языка: *https://ruscorpora.ru/new/corpora-morph.html*.

«/» — обозначение «или», «+» — наличие нескольких элементов в конструкции; A — прилагательное, N — существительное.

прилагательные и местоимения в соответствующих грамматических конструкциях, как то: A «настоящий», nom / acc, sin / pl + N: *настоящий бардак.*

Всего для поиска параметров вербально-семантического уровня было создано 107 аутентичных правил для вычленения из текста 11 различных конструкций. Поиск выбранных в данной работе параметров этого уровня (уровня идиолекта в соответствии с концепцией [Литвинова 2019]) легко формализуем, поскольку вербально-семантический уровень имеет «„более формальные“ языковые характеристики, которые априори считаются стабильными, хотя специально вопрос об их стабильности никак не исследуется» [Литвинова 2019: 2].

Для репрезентации фрагмента тезауруса личности были выбраны такие параметры, как ключевые лексемы, наиболее частотные словные триграммы и биграммы, экспликаторы аксиологических текстовых доминант дихотомии «свой/чужой».

Ключевые лексемы определяются с помощью алгоритма логарифмического правдоподобия при сравнении интересующего текста с референтным корпусом большого объема (использовался корпус «Opencorpora», URL: http://opencorpora.org, дата обращения: 08.02.2020, объемом на дату обращения 1 540 034 слова). В результате для каждого текста получаем список ключевых слов с числовой экспликацией значения меры логарифмического правдоподобия (loglikelihood score, или LL). В конечный список включаются лишь слова со значением LL более 50.

Поиск словных биграмм и триграмм основан на абсолютной частотности встречаемости слов рядом друг с другом и осуществляется с помощью функций выбранного языка программирования. Установление наиболее частотных сочетаний слов для текстов осуществляется после описанного выше препроцессинга, при подсчете также учитывается отсутствие слова в списке стоп-слов, кириллическое написание и длина слова более 2 символов. В результате при сравнении двух текстов для каждого формируется список наиболее частотных сочетаний слов.

При анализе ключевых лексем и наиболее частотных сочетаний слов из полученных списков удаляются сочетания с именами собственными, поскольку данные лексемы маркируют не собственно особенности авторских идиостилей, а тематическую отнесенность текстов.

Под экспликаторами аксиологических текстовых доминант групп «свой/чужой» в настоящем исследовании понимается дисперсия местоимений «я/мы-группы», «ты/они-группы», то есть «ведется подсчет местоимений всех разрядов в прямых и косвенных падежах по соответствующим группам» [Степаненко 2017].

Тезаурусный уровень наиболее труден для формализации. Можно автоматически создать материальную экспликацию авторского тезауруса [Бессмертный, Нугуманова 2012], тем не менее определить, как лексемы в нем «выстраиваются в упорядоченную, достаточно строгую иерархическую систему, в какой-то степени (непрямой) отражающую структуру мира» [Караулов 2010: 52], крайне сложно. Этот уровень репрезентирован наименьшим количеством параметров (3 стандартных стилеметрических алгоритма и 1 аутентичное правило) именно в силу стремления не просто формализовать некоторые компоненты языковой личности с целью ее компьютерной репрезентации, но и сделать конечную модель интерпретируемой.

Прагматикон языковой личности формализован посредством следующего набора параметров: вводные слова и конструкции, эксплицирующие субъективную модальность; целевые, выделительные и сравнительные обороты, репрезентирующие уровень освоения автором компетенций письменной речи и его коммуникативные стратегии и тактики; синтаксические сращения, дающие представление в том числе об авторских предпочтениях в функционально-стилистической отнесенности текста; сравнительные придаточные, глагольные односоставные предложения, эксплицирующие функциональный тип повествования; наличие/отсутствие и виды обращений как контактоустанавливающего элемента. Всего было использовано 10 стандартных стилеметрических алгоритмов и 32 уникальных правила.

В модели для указанного уровня заданы не собственно единицы прагматикона («коммуникативная сеть: сферы, ситуации, роли» [Караулов 2010: 61]), а их косвенные репрезентанты, компоненты синтаксического уровня языка. В том числе поэтому речь идет о том, что разработанный алгоритм не может быть реализован без привлечения экспертной оценки. То есть компетенции, готовности автора на прагматическом уровне следует восстанавливать из получаемой статистико-синтаксической информации интерпретативным путем. В качестве материала для иллюстрации этого процесса используем сборник рассказов Сергея Довлатова «Наши». Для сборника «Наши» с помощью «ХоРом» удается извлечь 171 вводную конструкцию, среди которых большинство яв-

ляются вводно-союзными компонентами (*кроме того, более того, значит* и пр.), создающими анафорические связи текста. Так, С. Довлатов реализует компетенцию создания когезии письменной речи, «готовность соотносить интенции, мотивы, запрограммированные смыслы со способами их объективации в тексте» [Там же]. Выявленное значение параметров также позволяет говорить о том, что эмоциональность речи («готовность использовать стилистические средства того или иного подъязыка» [Там же]) создается в основном за счет средств, отличных от вводных компонентов. Ведущим приемом создания эмоциональности в тексте становится образ, что доказывается сопоставлением синтаксических осложнителей: в тексте сравнительных конструкций намного больше, чем, например, целевых оборотов: относительная частота встречаемости первых — 2669,85, вторых — 715,14 [Хоменко 2021].

Для анализа синтаксических структур были прописаны правила, основанные на pos-tags, а также на том, какие синтаксические отношения имеют место в предложении [Падучева 1974] и какую грамматическую конструкцию реализуют те или иные его компоненты [Лингвистика конструкций 2010]. Например, для вычленения из текста вводных слов формализованное правило (алгоритм поиска) будет выглядеть таким образом:

– для машинного представления создается словарь из всех возможных вводных слов русского языка;

– прописывается грамматико-пунктуационное правило, которое позволяет выделить из текста именно вводную конструкцию, а не омонимичную ей:

1) __, Prnt,__

2) <начало предложения> Prnt,

где *Prnt* – любая часть речи; __ — некоторая часть предложения, *<начало предложения>* — обозначение начала предложения.

Поиск глагольных односоставных предложений, например, определенно-личных осуществляется в соответствии со следующим алгоритмом:

1) **+** V, 1per / 2per, sg / pl, praes / fut, indic

2) **+** V, sg / pl, imper

3) – N / SPRO, nom, sg / pl

4) – NUM, nomn _+ N в gen/ gen2, pl

5) – много/мало/несколько/немного/немало _ + N в gen/ gen2, pl.

Правило для поиска целевых оборотов основано на понятии семантической валентности [Падучева 1974: 44] и грамматики предложных конструкций с двойными предлогами [Лингвистические конструкции 2010]. Так, составные предлоги «с целью/из расчета» требуют инфинитива (условие валентности) при реализации целевого оборота, значит, формализованное правило для поиска таких структур будет выглядеть следующим образом: «с целью/из расчета» + INF, где обозначение *INF* использовано для инфинитива.

После извлечения всех параметров, связанных со словесными структурами, реализуется подсчет ipm (instance per million). Для синтаксических параметров количество каждого параметра делится на количество предложений в тексте. Разработать правила для автоматического поиска выбранных в данной работе структур вербально-семантического и мотивационного уровней несложно. Точность их работы высока: F-мера для всех параметров варьируется от 0,89 до 1.

В качестве результата работы алгоритма выводятся значения коэффициента корреляции Пирсона, значение линейной регрессии (оценивать следует коэффициент детерминации), критерия Стьюдента для моделей двух сравниваемых текстов, а также значения метрик каждого параметра для двух текстов, метрик, доказывающих или опровергающих гипотезу H_0 о том, что автором двух сравниваемых текстов является одно лицо.

Важно, что данный блок не является конечным шагом в разработанной методике. Как уже было сказано, текстовую статистику необходимо интерпретировать. Так, для традиционной математической статистики значимым считается коэффициент корреляции более 65 %, в случае работы программы говорить о сходстве моделей следует при коэффициенте корреляции 86 % и выше [Радбиль, Маркина 2019]. Программное обеспечение намеренно не выдает результат в виде выводного знания формата «*Автором двух сравниваемых текстов является одно лицо / Авторами двух сравниваемых текстов являются разные лица*», поскольку в разработанной методике именно эксперт, основываясь в том числе на статистических данных, принимает конечное решение об атрибуции текста, используя рейтерские таблицы [Khomenko, Baranova, Romanov, Zadvornov 2021], подготовленные по результатам исследования (таблица 1), и свой экспертный опыт.

Таблица 1

Пример рейтерской таблицы для оценки результатов работы атрибуционной модели

Тип дискурса	коэффициент корреляции Пирсона	коэффициент детерминации линейной регрессии	t-критерий Стьюдента (p-value)	Автором сравниваемых текстов, вероятно*, является одно лицо	Авторами сравниваемых текстов, вероятно, не является одно лицо	комментарий
Сетевая публицистика	достигает 1,00	достигает 1,00	обычно около 0, 95; не ниже 0,93	+	–	Для публицистики p-value t- критерия Стьюдента — значительно менее релевантная метрика, чем для остальных дискурсов. Если в публицистике значения КК и КД достигают единицы, можно говорить о едином авторстве сравниваемых текстов даже при условии не очень высокого значения p-value t-критерия Стьюдента. С другой стороны, p-value t-критерия Стьюдента может казаться высоким, но в случае низких или не очень высоких значений других метрик следует подходить комплексно и анализировать всю информацию.
Сетевая публицистика	обычно около 0,88 – 0,89	обычно около 0,71, но может достигать и 0,77	может быть как низкой (0,60), так и достаточно высокой: 0,85	–	+	
Сетевая публицистика	не очень высока: около 0,71	низкое значение: около 0, 50	может быть очень высокой: 0,98	-	+	

* Вероятностный характер вывода связан с тем, что в каждом конкретном случае в соответствии с разработанной методикой решение о конечном авторстве принимает исследователь.

Для создания подобных таблиц автор исследования использовал ряд текстовых коллекций (описание коллекций приводится в разделе 3 настоящей статьи). 40 % текстов каждой из них анализировалось с помощью ресурса «ХоРом» в соответствии со схемами *Автор А = Автору Б* (оба текста принадлежат одному автору) и *Автор А ≠ Автору Б* (тексты принадлежат разным авторам) в равных или примерно равных долях (20 % к 20 %) с целью наблюдения за «поведением» статистики в разных случаях. На основе этого исследования были созданы рейтерские таблицы для каждого текстового жанра (художественная нежанровая проза, сетевая беллетристика, сетевая публицистика, развлекательная публицистика, корпоративная переписка).

Оценка результатов работы методики проводилась с двух точек зрения: с одной стороны, полученные модели языковых личностей рассматривались с точки зрения теоретической оценки моделей [Bloomfield 1926; Ельмслев 2005; Лосев 2004; Апресян 1966; Штофф 1966; Ревзин 1977; Белоусов 2010 и др.] наряду со сводом критериев для определения типа лингвистической модели (модели речевой деятельности, исследовательские модели, метамодели и пр.).

Так, можно говорить о том, что с теоретической точки зрения интегративная атрибуционная модель, включающая в свой состав параметры трех языковых уровней, квантитативно объективированные и квалификативно оцененные экспертным путем,

является достаточно полной, всесторонне имитирующей оригинал и одновременно объективной. Речь идет о том, что сформированный пул параметров способен отражать достаточно необходимых сведений для идентификации авторства информации (полнота), структура модели объемно передает объект-оригинал, идиостиль автора, поскольку включает характеристики всех трех уровней языковой личности (всесторонность имитации оригинала), в структуре модели также отсутствуют личностные оценки и суждения исследователя (объективность).

Все это позволяет разработанной модели успешно решать на практике идентификационную задачу закрытого класса (с ограниченным количеством авторов) при попарном сравнении письменных текстов разного объема и жанровой отнесенности.

3. АПРОБАЦИЯ АТРИБУЦИОННОГО АЛГОРИТМА

Тестирование и апробация созданного алгоритма проходили на следующих текстовых коллекциях:

1) коллекция текстов художественной литературы, включающая тексты С. Д. Довлатова «Наши», «Чемодан», «Иностранка», «Заповедник», «Зона: Записки надзирателя», «Встретились, поговорили» и В. П. Астафьева «Обертон», «Последний поклон», «Звездопад», «Так хочется жить»; всего 10 текстов. Доля правильных ответов алгоритма (accuracy), точность (precision) и полнота (recall) равны 100 %, F-мера — 1 (здесь и далее значения метрик указаны в связи с интерпретацией статистических данных с помощью методических рекомендаций и рейтер-ских таблиц, разработанных для целей анализа);

2) коллекция текстов сетевой беллетристики (портал «Книга фанфиков»), включающая тексты 3 авторов-женщин, 4 авторов-мужчин; всего 190 текстов. Доля правильных ответов алгоритма (accuracy) — 83 %, точность (precision) — 67 %, полнота (recall) — 100 %, F-мера — 0,8;

3) коллекция текстов сетевой публицистики (газета «The Village» — ресурс заблокирован в РФ), включающая тексты 3 авторов-женщин, 3 авторов-мужчин; всего 600 текстов. Доля правильных ответов алгоритма (accuracy), точность (precision) и полнота (recall) равны 100 %, F-мера — 1;

4) коллекция текстов развлекательной публицистики (портал «ЯПлакалъ»), включающая тексты 3 авторов-женщин, 3 авторов-мужчин; всего 600 текстов. Доля правильных ответов алгоритма (accuracy) — 40 %, точность (precision) — 0, полнота (recall) — 0, F-мера — 0;

5) коллекция текстов корпоративной русскоязычной переписки, включающая тексты 2 авторов-женщин, 2 авторов-мужчин, всего 218 текстов. Доля правильных ответов алгоритма (accuracy) — 83 %, точность (precision) — 67 %, полнота (recall) — 100 %, F-мера — 0,8.

Часть каждой текстовой коллекции (около 60 %) исследовалась автором с помощью инструмента «ХоРом» в соответствии со схемами *Автор А = Автору Б* и *Автор А ≠ ≠ Автору Б* в равных или примерно равных долях с целью поиска истинно положительных (True Positive, TP), ложно положительных (False Positive, FP), ложно отрицательных (False Negative, FN), истинно отрицательных (True Negative, TN) результатов работы алгоритма. Результат представлялся в таблицах вида, проиллюстрированного таблицей 2.

Так, если для пары текстов А. Яковлев, «Подставные знакомства» — А. Яковлев, «Как встречают Новый год в плацкарте, самолете и на трассе» алгоритм «ХоРом» выдает следующую статистику: коэффициент корреляции Пирсона: 1; коэффициент детерминации линейной регрессии: 1; t-критерий Стьюдента: p-value: 0.94, — то, судя по рейтерской таблице (таблица 1), исследователь делает вывод, что «автором сравниваемых текстов, вероятно, является одно лицо». Данный вывод соответствует действительности, значит, в таблице 2 следует выбрать графу True Positive (TP).

Таблица 2

Пример вычисления оценочных мер при определении эффективности работы алгоритма

Текстовая пара		TP	FN	FP	TN
1	А. Яковлев, «Подставные знакомства» — А. Яковлев, «Как встречают Новый год в плацкарте, самолете и на трассе» (тексты одного жанра и одного автора-мужчины)	+	–	–	–
2	О. Карасева, «Где дешевле зимовать — на Бали или Шри-Ланке» — О. Карасева, «На что живут журналисты федеральных каналов» (тексты одного жанра и одного автора-женщины)	+	–	–	–
3	А. Яковлев, «Лучшие советские мозаики в Москве» — К. Руков, «Выживут только спекулянты: Как русский трейдер заработал миллионы на обвале американской биржи» (тексты одного жанра (тематика не учитывается) и разных авторов-мужчин)	–	–	–	+
4	О. Карасева, «Как сейчас поехать на дачу» — А. Дергачева, «Рабочие снова опустошают запасы бобров на Яузе» (тексты одного жанра (тематика не учитывается) и разных авторов-женщин)	–	–	–	+
и пр.					

5. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

В результате анализа удалось сделать некоторые выводы и получить следующие результаты: методика может быть применена для атрибуции текстов разных дискурсов при условии верной параметризации моделей и правильной интерпретации статистики для каждого. В ходе работы было установлено:

1) что для дискурса художественной прозы (как прозы признанных авторов, так и беллетристики) наиболее информативным является t-статистика Стьюдента;

2) для жанра современной беллетристики неинформативным является стилостатистический пул, поскольку, по экспериментальным данным, значения стилостатистических параметров близки для всех обследованных текстов;

3) для определения автора публицистического текста, чтобы признать гипотезу H_0 верной, значения коэффициентов корреляции и детерминации должны достигать единицы (необходимость такого высокого уровня значений связана с объемом текстового материала и его спецификой). Важно, что для публицистического дискурса следует признать значительно менее релевантной t-статистику, которая для художественного дискурса является наиболее информативным показателем. Что касается гендерной дифференциации материала, стоит заметить, что «женские» публицистические тексты более коррелируют с «женскими», равно как и «мужские» с «мужскими»; наибольшие корреляционные различия наблюдаются в индивидуальных стилях языковых личностей разной гендерной принадлежности;

4) для коротких текстовых сообщений: корпоративная переписка, комментарии в сети Интернет, — необходимо создание репрезентативной выборки из совокупности текстов объемом не менее 500 слов. Ограничение в 100 слов, выведенное еще С. М. Вулом и имеющее место до сих пор в судебном автороведении [Рубцова, Ермолаева, Безрукова и др. 2007] как объем, необходимый для определения авторства текстов, при встраивании в анализ статистической информации должен быть увеличен. Для улучшения работы алгоритма на данном материале в настоящий момент разрабатываются дополнительные параметры для построения моделей идиостиля как репрезентации языковой личности пишущего, они связаны с так называемым дигитальным почерком:

- графический литуратив;
- графическая гибридизация;
- обыгрывание архаичных аффиксов;
- использование элементов текста, написанных заглавными буквами;
- эмотиконы и прочие графические символы, выражающие эмоциональность речи;

5) разножанровые произведения тоже можно валидно обследовать с помощью разработанной интегративной методики (можно, например, сравнить текст электронного сообщения с публицистической статьей): доля правильных ответов алгоритма (accuracy) — 83 %, точность (precision) — 67 %, полнота (recall) — 100 %, F-мера — 0,8.

При использовании методики самыми ценными становятся не выводные данные автоматизированного алгоритма, а собственно модели идиостилей как репрезентации языковых личностей пишущих, созданные с его помощью. Эти модели являются понятными и простыми, легко интерпретируемыми экспертным путем, с одной стороны, и достаточно полными и адекватно имитирующими объект-оригинал — с другой.

Функциональность рассматриваемого алгоритма и разработанного электронного ресурса много шире изначально заложенных возможностей. Методику можно использовать не только для решения идентификационной задачи атрибуционной лингвистики, но и для исследования языковых личностей писателей, журналистов, политиков и прочих, при проведении диагностики языковой личности конкретного человека для решения задач психолингвистики, психологии, для обследования обобщенной языковой личности той или иной социальной группы, субкультуры и другого в целях решения задач социолингвистики, социологии. Важно, что при использовании разработанной методики в любом из представленных выше случаев модель языковой личности будет отвечать теоретическим принципам полноты, простоты, адекватности, технически точного и объективного описания оригинала, она будет экспланаторной, коммуникативной и интерпретируемой.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Абрамкина, Е. Е. Идентификационные признаки протокола допроса и методика автороведческого анализа / Е. Е. Абрамкина. — Текст : непосредственный // Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. — 2019. — Т. 17, № 3. — С. 97—108. — DOI 10.25205/ 1818-7935-2019-17-3-97-108.
2. Апресян, Ю. Д. Идеи и методы современной структурной лингвистики / Ю. Д. Апресян. — Москва : [б. и.], 1966. — 302 с. — Текст : непосредственный.
3. Белоусов, К. И. Модельная лингвистика и проблемы моделирования языковой реальности / К. И. Белоусов. — Текст : непосредственный // Вестник Оренбургского государственного университета. — 2010. — № 11 (117). — С. 94—97.
4. Бессмертный, И. А. Метод автоматического построения тезаурусов на основе статистической обработки текстов на естественном языке / И. А. Бессмертный, А. Б. Нугуманова. — Текст : непосредственный // Известия Томского политехнического университета. — 2012. — № 5. — С. 125—130.
5. Виноградов, В. В. Проблема авторства и теория стилей / В. В. Виноградов. — Москва : Гослитиздат, 1961. — 614 с. — Текст : непосредственный.
6. Вул, С. М. Судебно-автороведческая идентификационная экспертиза: методические основы / С. М. Вул. — Харьков : ХНИИСЭ, 2007. — 64 с. — Текст : непосредственный.
7. Ельмслев, Л. Пролегомены к теории языка / Л. Ельмслев ; пер. с англ. [В. А. Звегинцев и др.]. — Москва : URSS, 2005. — 243 с. — Пер.: Hjelmslev, Louis Prolegomena to a theory of language.
8. Захаров, В. П. Статистический метод выявления коллокаций / В. П. Захаров, М. В. Хохлова. — Текст : непосредственный // Языковая инженерия: в поиске смыслов : доклады семинара «Лингвистические информационные технологии в Интернете»: XI Всероссийская объединенная конференция «Интернет и современное общество». — Санкт-Петербург : Изд-во С.-Петерб. ун-та, 2008. — С. 40—54.
9. Караулов, Ю. Н. Русский язык и языковая личность / Ю. Н. Караулов. — Москва : ЛКИ, 2010. — 264 с. — Текст : непосредственный.
10. Лингвистика конструкций / отв. ред. Е. В. Рахилина. — Москва : Издательский центр «Азбуковник», 2010. — 584 с. — Текст : непосредственный.
11. Литвинова, Т. А. Идиолект как объект корпусной идиолектологии: к становлению нового лингвистического направления / Т. А. Литвинова. — Текст : непосредственный // Ученые записки Новгородского государственного университета имени Ярослава Мудрого. — 2019, N 7 (25). — С. 1—5.
12. Лосев, А. Ф. Введение в общую теорию языковых моделей / А. Ф. Лосев. — Москва : Едиториал УРСС, 2004. — 293 с. — Текст : непосредственный.
13. Мартыненко, Г. Я. Основы стилеметрии / Г. Я. Мартыненко. — Ленинград : Изд-во Ленингр. ун-та, 1988. — 173 с. — Текст : непосредственный.
14. Марусенко, М. А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов / М. А. Марусенко. — Ленинград : Изд-во Ленингр. ун-та, 1990. — 164 с. — Текст : непосредственный.
15. Падучева, Е. В. О семантике синтаксиса / Е. В. Падучева. — Москва : Наука, 1974. — 291 с. — Текст : непосредственный.
16. Радбиль, Т. Б. Вероятностно-статистические модели в производстве автороведческой экспертизы русскоязычных текстов / Т. Б. Радбиль, М. В. Маркина. — Текст : непосредственный // Политическая лингвистика. — 2019. — № 2 (74). — С. 156—166.
17. Ревзин, И. И. Современная структурная лингвистика: проблемы и методы / И. И. Ревзин ; АН СССР, Ин-т славяноведения и балканистики. — Москва : Наука, 1977. — 263 с. — Текст : непосредственный.
18. Родионова, Е. С. Методы атрибуции художественных текстов / Е. С. Родионова. — Текст : непосредственный // Структурная и прикладная лингвистика : межвуз. сборник. Вып. 7 / под ред. А. С. Герда. — Санкт-Петербург : Изд-во С.-Петерб. ун-та, 2008. — С. 118—127.
19. Рубцова, И. И. Комплексная методика производства автороведческих экспертиз : методические рекомендации / И. И. Рубцова, Е. И. Ермолаева, А. И. Безрукова [и др.]. — Москва : ЭКУ МВД России, 2007. — 192 с. — Текст : непосредственный.
20. Русская грамматика : в 2 т. / Н. Ю. Шведова (гл. ред.). — Москва : Институт русского языка им. В. В. Виноградова РАН, 2005. — URL: http://rusgram.narod.ru/index.html. — Текст : электронный.
21. Степаненко, А. А. Гендерная атрибуция текстов компьютерной коммуникации: статистический анализ использования местоимений / А. А. Степаненко. — Текст : непосредственный // Вестник Томского государственного университета. — 2017. — № 415. — С. 17—25. — DOI 10.17223/15617793/415/3.
22. Хазова, А. Б. Автоматическое определение половой принадлежности автора текста: феномен русской женской прозы / А. Б. Хазова. — Текст : непосредственный // Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. — 2020. — Т. 18, № 1. — С. 22—32. — DOI 10.25205/1818-7935-2020-18-1-22-32.
23. Хоменко, А. Ю. Возможности и сложности формализации структуры языковой личности для компьютерного представления / А. Ю. Хоменко. — Текст : непосредственный // Вопросы когнитивной лингвистики. — 2021. — № 2. — С. 111—117.
24. Штофф, В. А. Моделирование и философия / В. А. Штофф. — Москва ; Ленинград : Наука, 1966. — 304 с. — Текст : непосредственный.
25. Bacciu, A. CrossDomain Authorship Attribution Combining Instance-Based and Profile-Based Features: Notebook for PAN at CLEF 2019, 2019 / A. Bacciu, M. La Morgia, A. Mei, E. N. Nemmi, V. Neri, J. Stefa. — URL: http://ceur-ws.org/Vol2380/paper_220.pdf. — Text : electronic.

26. Bhargava, M. Stylometric Analysis for Authorship Attribution on Twitter Author's copy / M. Bhargava, P. Mehndiratta, K. Asawa ; Jaypee Institute of Information Technology, 2013. — URL: https://www.researchgate.net/publication/299669552_Stylometric_Analysis_for_Authorship_Attribution_on_Twitter. — Text : electronic.

27. Bloomfield, L. A set of postulates for the science of language / L. Bloomfield. — Text : unmediated // Language. — 1926. — № 2 (3). — P. 153—164.

28. Coulthard, M. Author identification, idiolect, and linguistic uniqueness / M. Coulthard. — Text : unmediated // Applied Linguistics. — 2004. — No 24 (4). — P. 431—447.

29. Custódio, J. E. EACH-USP Ensemble Cross-domain Authorship Attribution: Notebook for PAN at CLEF 2018, 2018 / J. E. Custódio, I. Paraboni. — URL: http://ceur-ws. org/Vol-2125/paper_76.pdf. — Text : electronic.

30. Gomzin, A. Detection of author's educational level and age based on comments analysis : Paper presented at Dialogue 2018. Moscow, 30 May — 2 June 2018, 2018 / A. Gomzin, A. Laguta, V. Stroev, D. Turdakov. — URL: http://www.dialog-21.ru/media/4279/gomzin_turdakov.pdf. — Text : electronic.

31. Juola, P. Authorship Attribution / P. Juola. — Text : unmediated // Foundations and Trends in Information Retrieval. — 2006. — Vol. 1, No. 3. — P. 233—334.

32. Khomenko, A. Linguistic Modeling as a Basis for Creating Authorship Attribution Software / A. Khomenko, Y. Baranova, A. Romanov, K. Zadvornov. — Text : unmediated // Компьютерная лингвистика и интеллектуальные технологии. — 2021. — Вып. 20 (27) : дополнительный том. — P. 1063—1074.

33. Koppel, M. Exploiting Stylistic Idiosyncrasies for Authorship Attribution / M. Koppel, J. Schler. — Text : unmediated // Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis. — 2003. — No. 69. — P. 72—80.

34. Korobov, M. Morphological analyzer and generator for Russian and Ukrainian languages / M. Korobov. — Text : electronic // AIST 2015 / eds.: M. Y. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov, V. G. Labunets (eds.). — CCIS / Springer, Cham, 2015. — Vol. 542. — P. 320—332. — URL: https://doi.org/10.1007/978-3-319-26123-2_31.

35. Litvinova, T. Profiling the Age of Russian Bloggers / T. Litvinova, A. Sboev, P. Panicheva. — Text : unmediated // Proceedings of the 7th International Conference, AINL 2018. — St. Petersburg, 2018. — P. 167—177.

36. Murauer, B. Dynamic Parameter Search for Cross-Domain Authorship Attribution: Notebook for PAN at CLEF 2018 / B. Murauer, M. Tschuggnall, G. Specht. — 2018. — URL: http://ceur-ws.org/Vol-2125/paper_84.pdf. — Text : electronic.

37. Muttenthaler, L. Authorship Attribution in Fan-Fictional Texts given variable length Character and Word N-Grams. Notebook for PAN at CLEF 2019 / L. Muttenthaler, G. Lucas, J. Amann. — 2019. — URL: http://ceur-ws.org/Vol-2380/paper_49.pdf. — Text : electronic.

38. Panicheva, P. Semantic feature aggregation for gender identification in Russian Facebook / P. Panicheva, A. Mirzagitova, Y. Ledovaya. — Text : electronic // AINL 2017. CCIS / eds.: A. Filchenkov, L. Pivovarova, J. Žižka. — Springer, Cham, 2018. — Vol. 789. — P. 3—15. — URL: https://doi.org/10.1007/978-3-319-71746-3_1.

39. Pimonova, E. Doc2vec or better interpretability? A method study for authorship attribution. Paper presented at Dialogue 2020, 2020. Moscow, June 15–20 / E. Pimonova, O. Durandin, A. Malafeev. — DOI 10.28995/2075-7182-2020-19-606-614. — Text : unmediated.

40. Shuy, R. W. Creating language crimes: How law enforcement uses (and misuses) language / R. W. Shuy. — New York : Oxford University Press, 2005. — 194 p. — Text : unmediated.

REFERENCES

1. Abramkina, E. E. (2019). Identifikatsionnye priznaki protokola doprosa i metodika avtorovedcheskogo analiza [Identification Features of the Minutes of Interrogation and Ways of Authorship Examination]. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, *17*(3), 97—108. DOI 10.25205/1818-7935-2019-17-3-97-108. (In Russ.)

2. Apresyan, Yu. D. (1966). *Idei i metody sovremennoy strukturnoy lingvistiki* [Ideas and methods of modern structural linguistics]. Moscow, 302 p. (In Russ.)

3. Belousov, K. I. (2010). Model'naya lingvistika i problemy modelirovaniya yazykovoy real'nosti [Model linguistics and the problem of language reality modelling]. *Vestnik of the Orenburg State University*, *11*(117), 94—97. (In Russ.)

4. Bessmertny, I. A., & Nugumanova, A. B. (2012). Metod avtomaticheskogo postroeniya tezaurusov na osnove statisticheskoy obrabotki tekstov na estestvennom yazyke [The method of automatic construction of thesauri on the basis of texts in natural language statistical processing]. *Bulletin of the Tomsk Polytechnic University*, *5*, 125—130. (In Russ.)

5. Vinogradov, V. V. (1961). *Problema avtorstva i teoriya stiley* [The problem of authorship and the theory of styles]. Moscow: Goslitizdat, 614 p. (In Russ.)

6. Vul, C. M. (2007). *Sudebno-avtorovedcheskaya identifikatsionnaya ekspertiza: metodicheskie osnovy* [Forensic authorship identification expertise: methodological foundations]. Kharkiv, 64 p. (In Russ.)

7. Hjelmslev, L. (2005). *Prolegomeny k teorii yazyka* [Prolegomena to a theory of language (Transl from English, V. A. Zvegintsev and others)]. Moscow: URSS, 243 p. (In Russ.)

8. Zakharov, V. P., & Khokhlova, M. V. (2008). Statisticheskiy metod vyyavleniya kollokatsiy [Statistical method for collocation detecting]. In *Language Engineering: in Search of Meanings: Reports of the Seminar "Linguistic Information Technologies on the Internet": XI Conference "Internet and Modern Society"* (pp. 40—54). Publishing House of St. Petersburg University. (In Russ.)

9. Karaulov, Yu. N. (2010). *Russkiy yazyk i yazykovaya lichnost'* [Russian language and language personality]. Moscow, 264 p. (In Russ.)

10. Rakhilina, E. V. (Ed.). *Lingvistika konstruktsiy* [Linguistics of constructions]. Moscow: Publishing center "Azbukovnik", 2010. 584 p. (In Russ.)

11. Litvinova, T. A. (2019). Idiolekt kak ob"ekt korpusnoy idiolektologii: k stanovleniyu novogo lingvisticheskogo napravleniya [Idiolect as an object of corpus idiolectology: to the formation of a new direction in linguistics]. *The Memoirs of NovSU*, *7*(25), 1—5. (In Russ.)

12. Losev, A. F. (2004). *Vvedenie v obshchuyu teoriyu yazykovykh modeley* [An introduction to the general theory of language models]. Moscow: Editorial URSS, 293 p. (In Russ.)

13. Martynenko, G. Ya. (1988). *Osnovy stilemetrii* [Basics of stylometry]. Leningrad: Publishing house Leningrad University, 173 p. (In Russ.)

14. Marusenko, M. A. (1990). *Atributsiya anonimnykh i psevdonimnykh literaturnykh proizvedeniy metodami raspoznavaniya obrazov* [Attribution of anonymous and pseudonymous literary works by image recognition methods]. Leningrad: Publishing house Leningrad. University, 164 p. (In Russ.)

15. Paducheva, E. V. (1974). *O semantike sintaksisa* [About syntax semantics]. Moscow: Nauka, 291 p. (In Russ.)

16. Radbil, T. B., & Markina, M. V. (2019). Veroyatnostno-statisticheskie modeli v proizvodstve avtorovedcheskoy ekspertizy russkoyazychnykh tekstov [Probabilistic-Statistical Models in Conducting Authoring Expertise of Russian Texts]. *Political Linguistics*, *2*(74), 156—166. (In Russ.)

17. Revzin, I. I. (1977). *Sovremennaya strukturnaya lingvistika: problemy i metody* [Modern structural linguistics: Problems and methods]. Moscow: Nauka, Academy of Sciences of the USSR, Institute of Slavic and Balkan Studies, 263 p. (In Russ.)

18. Rodionova, E. S. (2008). Metody atributsii khudozhestvennykh tekstov [Methods of attribution of literary texts]. In A. S. Gerda (Ed.), *Structural and Applied Linguistics* (Interuniversity Collection of Scientific Papers, Iss. 7, pp. 118—127). Saint Petersburg: Publishing house of St. Petersburg University. (In Russ.)

19. Rubtsova, I. I., Ermolaeva, E. I., Bezrukova, A. I. et al. (2007). *Kompleksnaya metodika proizvodstva avtorovedcheskikh ekspertiz : metodicheskie rekomendatsii* [Methodology for the production of forensic attribution expertise]. Moscow: Ministry of Internal Affairs of Russia, 192 p. (In Russ.)

20. Shvedova, N. Yu. (Ed.) (2005). *Russkaya grammatika* [Russian grammar] (In 2 volumes). Moscow: Institute of the Russian language V. V. Vinogradov RAS. Retrieved from http://rusgram.narod.ru/index.html. (In Russ.)

21. Stepanenko, A. A. (2017). Gendernaya atributsiya tekstov komp'yuternoy kommunikatsii: statisticheskiy analiz ispol'zovaniya mestoimeniy [Gender attribution in social network commu-

nication: the statistical analysis of pronouns frequency]. *Tomsk State University Journal*, *415*, 17—25. DOI: 10.17223/15617793/415/3. (In Russ.)

22. Khazova, A. B. (2020). Avtomaticheskoe opredelenie polovoy prinadlezhnosti avtora teksta: fenomen russkoy zhenskoy prozy [Automatic Detection of Gender Identity: The Phenomenon of Russian Women's Prose]. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, *18*(1), 22—32. DOI: 10.25205/1818-7935-2020-18-1-22-32. (In Russ.)

23. Khomenko, A. Yu. (2021). Vozmozhnosti i slozhnosti formalizatsii struktury yazykovoy lichnosti dlya komp'yuternogo predstavleniya [Abilities and difficulties in language personality structure formalization for computer representation]. *Voprosy Kognitivnoy Lingvistiki*, *2*, 111—117. (In Russ.)

24. Shtoff, V. A. (1966). *Modelirovanie i filosofiya* [Modeling and philosophy]. Moscow, Leningrad: Nauka, 304 p. (In Russ.)

25. Bacciu, A., Morgia, M. La, Mei, A., Nemmi, E. N., Neri, V., & Stefa, J. (2019). *CrossDomain Authorship Attribution Combining Instance-Based and Profile-Based Features* (Notebook for PAN at CLEF 2019). Retrieved from http://ceur-ws.org/Vol2380/paper_220.pdf.

26. Bhargava, M., Mehndiratta, P., & Asawa, K. (2013). *Stylometric Analysis for Authorship Attribution on Twitter Author's copy*. Jaypee Institute of Information Technology. Retrieved from https://www.researchgate.net/publication/299669552_Stylometric_Analysis_for_Authorship_Attribution_on_Twitter.

27. Bloomfield, L. (1926). A set of postulates for the science of language. *Language*, *2*(3), 153—164.

28. Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, *24*(4), 431—447.

29. Custódio, J. E., & Paraboni, I. (2018). *EACH-USP Ensemble Cross-domain Authorship Attribution* (Notebook for PAN at CLEF 2018). Retrieved from http://ceur-ws. org/Vol-2125/paper_76.pdf.

30. Gomzin, A., Laguta, A., Stroev, V., & Turdakov, D. (2018). *Detection of author's educational level and age based on comments analysis* (Paper presented at Dialogue 2018. Moscow, 30 May–2 June 2018). Retrieved from http://www.dialog-21.ru/media/4279/gomzin_turdakov.pdf.

31. Juola, P. Authorship Attribution. *Foundations and Trends in Information Retrieval*, *1*(3), 233—334.

32. Khomenko, A., Baranova, Y., Romanov, A., & Zadvornov, K. (2021). Linguistic Modeling as a Basis for Creating Authorship Attribution Software. In *Computational linguistics and intelligent technologies*, *20*(27, Supplementary volume, pp. 1063—1074). Publishing house of the Russian State Humanitarian University.

33. 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.

34. Korobov, M. (2015). Morphological analyzer and generator for Russian and Ukrainian languages. In M. Y. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov & V. G. Labunets (Eds.), *AIST 2015* (Vol. 542, pp. 320—332). Springer, Cham, CCIS. Retrieved from https://doi.org/10.1007/978-3-319-26123-2_31.

35. Litvinova, T., Sboev, A., & Panicheva, P. (2018). Profiling the Age of Russian Bloggers. In *Proceedings of the 7th International Conference. AINL 2018* (pp. 167—177). St. Petersburg.

36. Murauer, B., Tschuggnall, M., & Specht, G. (2018). *Dynamic Parameter Search for Cross-Domain Authorship Attribution* (Notebook for PAN at CLEF 2018). Retrieved from http://ceur-ws.org/Vol-2125/paper_84.pdf.

37. Muttenthaler, L., Lucas, G., Amann, J. (2019). *Authorship Attribution in Fan-Fictional Texts given variable length Character and Word N-Grams* (Notebook for PAN at CLEF 2019). Retrieved from http://ceur-ws.org/Vol-2380/paper_49.pdf.

38. Panicheva, P., Mirzagitova, A., Ledovaya, Y. (2018). Semantic feature aggregation for gender identification in Russian Facebook. In A. Filchenkov, L. Pivovarova, & J. Žižka (Eds.), *AINL 2017* (Vol. 789, pp. 3—15). Springer, Cham, CCIS. Retrieved from https://doi.org/10.1007/978-3-319-71746-3_1.

39. Pimonova, E., Durandin, O., & Malafeev, A. (2020). *Doc2vec or better interpretability? A method study for authorship attribution* (Paper presented at Dialogue 2020, Moscow, June 15—20). DOI: 10.28995/2075-7182-2020-19-606-614.

40. Shuy, R. W. (2005). *Creating language crimes: How law enforcement uses (and misuses) language*. New York: Oxford University Press, 194 p.