Применение деревьев решений для анализа сильных позиций текста в задаче атрибуции произведений Ф. М. Достоевского

А.А. Рогов, А.А. Лебедев, Р.В. Абрамов, Н.Д. Москин, К.А. Кулаков

Петрозаводский государственный университет

rogov@petrsu.ru, perevodchik88@yandex.ru, monset008@gmail.com, moskin@petrsu.ru, kulakov@cs.karelia.ru

Аннотация

В работе рассматривается совокупность статей Ф. М. Достоевского и других авторов (М. М. Достоевский, Н. Н. Страхов, А. А. Головачев, И. Н. Шилль, А. Григорьев, А. У. Порецкий, Я. П. Полонский), опубликованных в журналах «Время» и «Эпоха» в период 1861-1865 гг. В текстах выделялись фрагменты размером 500, 700 и 1000 слов. При этом для увеличения объема выборки использовался шаг для отсчета начала следующего фрагмента: 100, 200 слов и т. п. На основе частеречного распределения фрагментов текстов были построены деревья решений, в узлах которых находятся условия ветвления, основанные на частоте встречаемости той или иной n-граммы (последовательности из *n* закодированных частей речи).
Анализ сильных позиций данных текстов (т. е. фрагментов, расположенных в начале или в конце текста) с помощью деревьев решений показывает возможность стилистической правки, которую вносил Ф. М. Достоевский в тексты изначальных авторов. Для проведения исследования использовалась информационная система СМАЛТ («Статистические методы анализа литературных текстов»), где была реализована автоматизированная разметка произведений с ручным контролем специалистов-филологов.

Ключевые слова: атрибуция текстов, корпусная лингвистика, Ф. М. Достоевский, сильные позиции текста, дерево решений, n-грамма, частеречное распределение

Библиографическая ссылка: Рогов А.А., Лебедев А.А., Абрамов Р.В., Москин Н.Д., Кулаков К.А. Применение деревьев решений для анализа сильных позиций текста в задаче атрибуции произведений Ф. М. Достоевского // Компьютерная лингвистика и вычислительные онтологии. Выпуск 4 (Труды XXIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2020, Санкт-Петербург, 17 – 20 июня 2020 г. Сборник научных статей). — СПб.: Университет ИТМО, 2020. С. 118-127. DOI: 10.17586/0000-0000-2020-4-118-127

Введение

Чтобы определить особенности произведений конкретного автора на фоне всего массива текстов художественной литературы, необходимо представить комплексный анализ его текстов, выделить ключевые особенности авторского восприятия мира, а также определить, как эти особенности воплощаются в произведениях. Вряд ли можно поставить под сомнение тот факт, что «каждый писатель подбирает языковые средства не только в соответствии с содержанием и замыслом, но и в зависимости от своего видения мира, обусловленного его мировосприятием, социальным положением, личностными качествами и психологическими особенностями» [1, с. 50]. Однако реализация подобранных автором языковых средств в тексте литературного произведения может быть самой разной.

К классификации авторского стиля как лингвистического явления существует несколько подходов; в зависимости от выбранного варианта исследователь направляет свою работу в нужное ему русло. В частности, можно выделить многоаспектный подход, предложенный в работе [2] и опирающийся на принципы коммуникации (рассмотрение отдельных компонентов художественной системы писателя, прежде всего, рассмотрения языковых средств (как правило, лексических) в сочетании с анализом различных структурных и смысловых форм организации языкового материала для выявления характера соотнесенности на фоне других идиостилей); анализ авторского стиля на основании экстралингвистических и интралингвистических факторов [3]; когнитивный аспект в анализе индивидуально-авторского стиля (истоки подобного подхода к идиостилю лежат в развитии теоретических положений А. А. Потебни, Л. В. Щербы, Г. О. Винокура, В. В. Виноградова, Б. А. Ларина), когда обращение к концептосфере писателя представляет собой возможность лучше понять художественный мир автора (см., например, [4; 5]).

Мы, вслед за В. В. Виноградовым, отмечавшим, что «изучение литературного стиля должно быть комплексным и системным» [6, с. 198] понимаем идиостиль как систему формальных и содержательных характеристик, которые присущи произведениям того или иного автора и отражают уникальный, авторский способ языкового выражения (подробнее см. [7]). Воплощение подобного авторского способа может проявляться на разных уровнях текста, в том числе и в анализе сильных его позиций.

Вопрос, связанный с выделением сильных позиций текста и определением их роли в понимании и восприятии произведений, не просто является объектом давнего интереса лингвистов [8; 9], но и в последнее десятилетие решается в аспекте анализа индивидуально-авторского стиля и определения особенностей творчества того или иного писателя [10; 11]. Под универсальными сильными позициями текста традиционно понимаются:

— заглавие текста;
— инициальная позиция текста (его первое предложение, первый абзац, первое сложное синтаксическое целое);
— конечная позиция текста (последнее предложение текста, последний его абзац, последнее сложное синтаксическое целое).

При этом в зависимости от жанра текста возможно появление и других сильных позиций (например, эпиграфа в художественных произведениях, рифмы – в стихотворных текстах, слогана – в рекламных текстах и т. п.), однако составляющие текста, перечисленные выше, следует признать универсальными – то есть, характерными для любого произведения. Именно в сильных позициях автор сосредотачивает важные для себя смысловые доминанты, передаваемые речевым произведением.

Поскольку сильные позиции текста играют наиболее значимую роль в выражении авторской идеи, то очевиден дополнительный интерес филологов именно к этим элементам текста при решении вопросов, связанных с атрибуцией текстов [12; 13], установлением авторства анонимных текстов [14], а также в ходе лингвистического анализа материалов, на которые, помимо автора, мог оказать непосредственное влияние другой человек.

В частности, анализируя тексты статей, представленных в журналах «Время» и «Эпоха» (1861-1865 гг.), мы обнаруживаем немалый список публиковавшихся там авторов (Ф. М. Достоевский, М. М. Достоевский, Н. Н. Страхов, А. А. Головачев, И. Н. Шилль, А. Григорьев, А. У. Порецкий, Я. П. Полонский), и, рассматривая морфологическую структуру некоторых материалов, казалось бы, не принадлежащих непосредственно Фёдору Михайловичу Достоевскому, мы можем говорить о непосредственном его влиянии на общее построение и смысловое наполнение текста (что находит отражение в том числе и в перечисленных выше сильных позициях).

Использованная в работе методика исследования предусматривает работу с большими фрагментами текста (500, 700, 1000 слов), что делает малозначимой позицию заглавия текста; однако инициальная и конечная позиции текста при таком подходе становятся объектом дополнительного интереса и могут привлекать внимание филолога в тех случаях, когда частеречное распределение в них отличается от распределения по частям речи во всем остальном тексте. Выбор в качестве признаков только частей речи трех первых позиций текста опирается на исследования, проведенные Г. Хетсо [15], который учитывает такие признаки, как общее распределение частей речи в первых двух позициях предложения и сочетание частей речи в первых двух позициях предложения. В данном исследовании методика Г. Хетсо была дополнена – под сильной позицией в рамках исследования понимаются три слова в началах каждого предложения текста.

Целью проводимого эксперимента была автоматизация выявления наиболее важных, фундаментальных отличительных характеристик текстов одного автора от текстов других авторов в аспекте распределения частей речи в сильных позициях текста (началах предложений) в сочетании с их последующей точечной лингвистической интерпретацией. В перспективе это должно помочь решить проблему проверки принадлежности перу Ф. М. Достоевского ряда статей, которые входят в раздел Dubia полного собрания сочинений писателя. Использование математических методов в этом случае позволяет лингвисту не просто избавиться от необходимости вручную выполнять подсчеты, но и помогает определить наиболее значимые отличия, которые могут быть не замечены или проигнорированы как нерелевантные при традиционном подсчете. Для решения задачи атрибуции текстов хорошо зарекомендовали себя следующие математические методы [12; 13; 14]: нейронные сети, деревья решений, машина опорных векторов (SVG), метод k-средних, метод QSUM, байесовский классификатор, марковские цепи, метод главных компонент, дискриминантный анализ, генетические алгоритмы, статистические критерии (хи-квадрат Пирсона, критерий Стьюдента, Колмогорова-Смирнова) и др. Среди этих методов интеллектуального анализа данных деревья решений выделяются тем, что они просты в понимании и интерпретации, а также не требуют специальной предварительной обработки данных. Построение дерева решений позволяет однозначно определить, какие из признаков являются наиболее существенными, а потому требующими дополнительной интерпретации в аспекте лингвостилистического анализа – они будут вынесены в вершину дерева решений.

1. Построение и анализ деревьев решений

В табл. 1 представлен список из 19 анализируемых текстов (среди авторов: Ф. М. Достоевский, М. М. Достоевский, Н. Н. Страхов, А. А. Головачев, И. Н. Шилль, А. Григорьев, А. У. Порецкий, Я. П. Полонский). 12 из них принадлежат Ф. М. Достоевскому, остальные – другим авторам. Выбор текстов был осуществлен случайным образом с учетом распределения публикаций этих авторов на страницах журналов «Время» и «Эпоха» (1861-1865 гг.). Грамматическая разметка данных текстов учитывала 14 частей речи (существительное, прилагательное, числительное, местоимение, наречие, категория состояния, глагол, причастие, деепричастие, предлог, союз, частица, модальное слово, междометие), а также позволяла выделять цитаты, иностранные слова, вводные слова, сокращенные слова и неязыковые символы.

Для проведения разметки использовалась информационная система СМАЛТ («Статистические методы анализа литературных текстов»), разработанная в Петрозаводском государственном университете [16]. В рамках системы выполнялся импорт текстовых произведений с автоматизированным разбиением на абзацы, предложения и слова; редактирование подобранного автоматически в ходе импорта морфологического разбора (часть речи и другие параметры) или создание нового разбора филологом; построение фрагментов текстов с заданным размером слов и отступом

и с описанием морфологического разбора для детального анализа. Таким образом, в системе СМАЛТ была выполнена автоматизированная разметка произведений с ручным контролем специалистов-филологов.

Таблица 1. Исходные тексты для анализа

Код	Название	Автор	Журнал	Год	№ журнала	Количество слов
2	Пожары	Ѳедор Достоевский	Время	1862	1	1943
11	Тарасъ Шевченко	Аполлон Григорьев	Время	1861	4	1724
13	Письмо к редактору	Полонский Я. П.	Время	1863	3	2303
34	Литературная истерика	Ѳедор Достоевский	Время	1861	7	2808
35	Молодое перо	Ѳедор Достоевский	Время	1863	2	1872
40	Подписка на 1863 годъ	Михаил Достоевский	Время	1863	1	2541
42	Ряд статей о русской литературе. Введение	Ѳедор Достоевский	Время	1861	1	12508
43	Славянофилы, черногорцы и западники	Ѳедор Достоевский	Время	1862	9	2058
75	Ряд статей... Г. -бов и вопрос об искусстве	Ѳедор Достоевский	Время	1861	2	11053
77	Книжность и грамотность. Статья вторая	Ѳедор Достоевский	Время	1861	8	14210
78	Последние литературные явления. Газета "День"	Ѳедор Достоевский	Время	1861	11	4323
82	Необходимое литературное объяснение, по поводу ра…	Ѳедор Достоевский	Время	1863	1	3680
86	Чтобы кончить. Послѣдніе объясненія съ "Современн…	Ѳедор Достоевский	Эпоха	1864	9	1378
87	Политическое обозрение	Головачев А.А.	Эпоха	1864	8	10309
89	Лермонтов и его направление. Статья вторая	Аполлон Григорьев	Время	1862	11	7480
92	Наши домашние дела	Порецкий А.У.	Эпоха	1864	12	8602
96	Голос за петербургского Дон-Кихота. (По поводу ст…	Ѳедор Достоевский	Время	1862	10	1334
97	Примечание	Ѳедор Достоевский	Эпоха	1864	9	1599
116	ДУРНЫЕ ПРИЗНАКИ	Страхов Н. Н.	Время	1862	11	6331

Далее текст разбивался на фрагменты размером 500, 700 и 1000 слов. Если заключительный фрагмент оказывался меньше указанного размера, то он отбрасывался. Для увеличения объема выборки применялся шаг для отсчета начала следующего фрагмента: 100, 200 слов и т. п. Поскольку на первом этапе было установлено, какие части речи встречаются во фрагментах и в каком порядке, на втором этапе можно перейти к подсчету частоты встречаемости n-грамм (последовательностей из *n* закодированных частей речи) [17].

Эти частоты легли в основу построения деревьев решений [18]. Каждая вершина дерева определяет условие ветвления по одному из признаков, что позволяет в дальнейшем

классифицировать тексты на основе их частеречного распределения. Подобные модели широко используются в интеллектуальном анализе данных, и, в частности, при анализе текстов. Например, их применяла А. Р. Дубовик для исследования принадлежности русскоязычных текстов к тому или иному функциональному стилю (научный, художественный, деловой или публицистический) с опорой на ряд статистических параметров, таких как средняя длина слова, средняя длина предложения, частота встречаемости в текстах определенных n-грамм и др. [19]. Также деревья решений применялись в задаче разграничения фольклорных текстов и текстов, стилизованных под фольклор [20].

На рисунке 1 показан фрагмент графа, полученного в результате расчетов с шагом 100 слов и размером фрагмента 1000 слов для биграмм. Достоинства данного метода заключается в том, что он прост в понимании и интерпретации, а также не требует специальной подготовки данных.

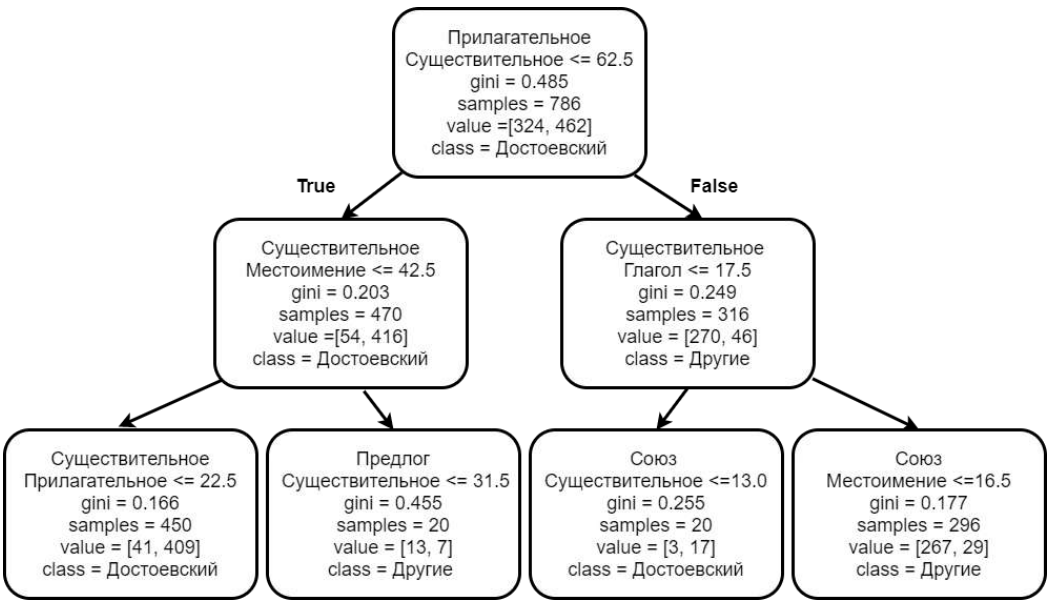

Рис. 1. Фрагмент дерева решений (биграмма, шаг 100 слов, фрагмент 1000 слов)

2. Результаты анализа и их интерпретация

Анализируя дерево решений, представленное на рисунке 1, можно заметить, что фрагменты текстов статьей Ф. М. Достоевского, как правило, содержат частоту встречаемости биграммы «Прилагательное + Существительное» меньше 62,5. Анализ этих фрагментов показывает, что среди неверно классифицированных фрагментов статей, которые принадлежат другим авторам (не Фёдору Михайловичу Достоевскому), есть и те, что расположены в начале или в конце текста, то есть в его сильных позициях. Это может быть свидетельством стилистической правки, которую вносил Ф. М. Достоевский в данные фрагменты, что сказывалось на частеречном распределении текстов и потому не соответствовало стилю изначального автора статьи.

В таблице 2 приведены тексты с указанием нетипичных фрагментов (значительная часть которых располагается именно в сильных позициях текста). Данные фрагменты, будучи правильно интерпретированными, могут стать объектом дополнительного интереса со стороны литературоведов в контексте определения вопроса авторства статей

и стилистического взаимовлияния Ф. М. Достоевского и других авторов «Времени» и «Эпохи».

Таблица 2. Аномалии в классификации фрагментов статей

Код	Название	Автор	Журнал	Год	№ журнала	Номера нетипичных фрагментов
2	Пожары	Ѳедор Достоевский	Время	1862	1	0-8
11	Тарасъ Шевченко	Аполлон Григорьев	Время	1861	4	0, 4
13	Письмо к редактору	Полонский Я.П.	Время	1863	3	10-12
35	Молодое перо	Ѳедор Достоевский	Время	1863	2	0-2
40	Подписка на 1863 годъ	Михаил Достоевский	Время	1863	1	0-14
92	Наши домашние дела	Порецкий А. У.	Эпоха	1864	12	77, 78
116	ДУРНЫЕ ПРИЗНАКИ	Страхов Н. Н.	Время	1862	11	51, 52

К примеру, в тексте «Наши домашние дела» с кодом 92 («Эпоха», 1864, № 12), принадлежащем А. У. Порецкому, именно два последних фрагмента текста имеют иное частеречное распределение в сравнении со всем остальным текстом статьи (в частности, реже встречаются пары «Прилагательное + Существительное» и «Существительное + Местоимение»), что в целом нехарактерно для всех остальных 76 фрагментов данной статьи. В тексте «Письмо к редактору», автором которого является Я. П. Полонский, («Время», 1863, № 3, код 13) последние три фрагмента имеют другое распределение. В тексте «Тарасъ Шевченко» Аполлона Григорьева («Время», 1861, № 4, код 11) обращает на себя внимание первый фрагмент и середина текста. В тексте «ДУРНЫЕ ПРИЗНАКИ» Н. Н. Страхова («Время», 1862, № 11, код 116) также явно выделяются два последних фрагмента.

Вероятно, не всегда причины подобного рода изменений в распределении частей речи следует искать именно в стилистической правке со стороны Ф. М. Достоевского. К примеру, в упомянутом выше тексте «Письмо к редактору» причиной снижения случаев встречаемости комбинации частей речи «Существительное» + «Местоимение» может быть переход автора от размещаемых в начале текста пространных теоретических размышлений о художественных достоинствах повести Л. Н. Толстого «Казаки», к описанию совершаемых в произведении действий, выраженных глаголами (*люди живутъ какъ живетъ природа, умираютъ, родятся, совокупляются, опять родятся, дерутся, пьютъ, ѣдятъ, радуются и опять умираютъ*), равно как и цитированию чужого текста, что сказывается на частеречном распределении.

В некоторых текстах появление определенных пар частей речи может объясняться тематикой произведения. К примеру, в статье «Тарас Шевченко» (автор: Аполлон Григорьев) активизация во многих фрагментах текста пары «Прилагательное + Существительное» может объясняться характером описываемого явления (для характеристики произведений Шевченко неоднократно используется словосочетание *малороссійская литература*), а сам Григорьев очень любит пользоваться эпитетом великий, совмещая его с разными существительными: *великому таланту, великими представителями, великаго поэта, великой литературы, великому кобзарю* и т. п.).

В то же время, даже предположить причину изменения частеречного распределения удается не всегда. К примеру, завершающие фрагменты текстов «Наши домашние дела» и «Дурные признаки», не отличаясь чем-то содержательно или тематически в сравнении с остальным содержанием статьи, классифицируются по-иному. Именно это несоответствие изначальному авторскому стилю может указывать на внесение стилистических правок неким иным лицом (возможно, самим Ф. М. Достоевским).

Заметим, что приведенный признак «частота биграммы «Прилагательное + Существительное» меньше 62,5» не является однозначным признаком стиля

Ф. М. Достоевского. Так, например, «Подписка на 1863 годъ» («Время», 1863, № 1, код 40) приписываемая М. М. Достоевскому целиком удовлетворяет этому параметру. Весь текст «Пожары» («Время», 1862, № 1, код 2) Ф. М. Достоевского не удовлетворяет этому признаку. Начало (три фрагмента) текста «Молодое перо» («Время», 1863, № 2, код 35) тоже не удовлетворяет этому параметру.

С учетом вышенаписанного текст статьи «Подписка на 1863 годъ» может стать объектом более пристального изучения литературоведов – весьма вероятно, что Фёдор Михайлович Достоевский мог оказать немалое влияние на Михаила Михайловича (как опосредованное, так и прямое – если предположить, что он составил какую-то часть текста вместо своего старшего брата). Не менее интересен и вопрос авторства текста «Пожары», ответ на который по мнению некоторых литературоведов не столь однозначен, как это может показаться – вполне возможно, что это был не Ф. М. Достоевский (см. статью Н. Г. Розенблюма «Петербургские пожары 1862 г. и Достоевский» [21]).

Выводы

Ф. М. Достоевский, будучи мастером работы не только с художественным, но и с публицистическим текстом, прекрасно представлял себе важность сильных позиций текста с точки зрения их воздействия на читателя, а потому мог уделять более пристальное внимание внесению правок в начальные и в конечные абзацы текстов чужих статей. Ведь именно эти элементы оказывают наибольшее воздействие на читателя – благодаря началу текста складывается первое впечатление о статье, а финал – подводит итоги и, как правило, хорошо откладывается в памяти читателя. Именно поэтому, комплексно решая вопросы, связанные с атрибуцией текстов в журналах «Время» и «Эпоха», в качестве одной из задач можно выделить специальный анализ данных элементов текста.

Приведенный метод предоставляет возможность выделять в тексте фрагменты, отличающиеся частеречными характеристиками. Анализ этих фрагментов позволяет исследователю ставить и решать различные задачи, начиная от авторства тех или иных фрагментов до автоматического реферирования текстов.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, грант № 18-012-90026.

Литература

[1] Гаспарян С.К., Князян А.Т. К вопросу об изучении индивидуального стиля автора // Филологические науки. 2004. № 4. С. 50 – 57.

[2] Болотнова Н.С. и др. Коммуникативная стилистика художественного текста: лексическая структура и идиостиль / Н.С. Болотнова, И.И. Бабенко, А.А. Васильева, С.М. Карпенко, О.В. Орлова, С.В. Сыпченко, Р.Я. Тюрина. Томск. 2001. 331 с.

[3] Шаркунова О.В. Идиостиль художественного текста как индивидуальное сочетание экстра- и интралингвистических параметров, основанных на референтных отношениях // Материалы международной заочной научно-практической конференции «Актуальные вопросы филологии, искусствоведения и культурологии». Новосибирск. 2011. С. 75 – 80.

[4] Тарасова И.А. Категории когнитивной лингвистики в исследовании идиостиля // Вестник СамГУ. 2004. № 1 (31). С. 163 – 169.

[5] Фокина Ю.М. Особенности репрезентации индивидуально-авторской концептосферы в англоязычной и русскоязычной прозе (на материале рассказов А.П. Чехова и Д. Джойса). Автореф. дис. … канд. филол. наук. Саратов, 2010. 184 с.

[6] Виноградов В.В. Проблема авторства и теория стилей // М.: Художественная литература. 1961. 613 с.

[7] Лебедев А.А. Идиостиль П. А. Вяземского: синтаксический аспект // Петрозаводск: Изд-во ПетрГУ. 2013. 134 с.
[8] Арнольд И.В. Значение сильной позиции для интерпретации художественного текста // Иностранные языки в школе. М. 1978. № 4. С. 23 – 31.
[9] Гальперин И.Р. Текст, как объект лингвистического исследования // М.: Наука, 1981. 139 с.
[10] Патроева Н.В., Лебедев А.А. Синтаксическая организация, размер и семантика инициальных предложений в лирике А. С. Пушкина // Вестник Томского государственного университета. Филология. Томск. 2018. № 53. С. 224 – 236.
[11] Петрова К.В. Сильные позиции текста в автобиографии Джанет Уинтерсон // Вестник Новгородского государственного университета. Серия «Гуманитарные науки». Великий Новгород. 2015. № 4 (87), часть 1. С. 73 – 76.
[12] Рогов А.А. и др. Математические методы атрибуции текстов / А.А. Рогов, А.В. Седов, Ю.В. Сидоров, Т.Г. Суровцова // Петрозаводск: Изд-во ПетрГУ. 2014. 95 с.
[13] Stamatatos E. A Survey of Modern Authorship Attribution Methods // Journal of the American Society for Information Science and Technology. 2009. Vol. 60, № 3. P. 538 – 556.
[14] Романов А.С. Методика и программный комплекс для идентификации автора неизвестного текста. Дис. ... канд. техн. наук. Томск, 2010. 149 с.
[15] Kjetsaa G. Attributed to Dostoevsky: The Problem of attributing to Dostoevsky anonymous articles in Time and Epoch. Oslo: Solum Forlag A. S., 1986.
[16] Рогов А.А., Кулаков К.А., Москин Н.Д. Программная поддержка в решении задачи атрибуции текстов // Программная инженерия. М.: Изд-во "Новые технологии", 2019. Т. 10, № 5. С. 234 – 240.
[17] Котов А.А. и др. Лингвистические корпусы / А.А. Котов, З.И. Минеева, А.А. Рогов, А.В. Седов, Ю.В. Сидоров // Петрозаводск: Изд-во ПетрГУ, 2014.
[18] Breiman L., etc. Classification and regression trees / L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone. Wadsworth, Belmont, Ca, 1984.
[19] Дубовик А.Р. Автоматическое определение стилистической принадлежности текстов по их статистическим параметрам // Компьютерная лингвистика и вычислительные онтологии. Выпуск 1. Труды XX Международной научной конференции «Интернет и современное общество», IMS-2017 (Санкт-Петербург, 21 – 23 июня 2017 г.). СПб.: Университет ИТМО, 2017. С. 29 – 45.
[20] Щеголева Л.В., Лебедев А.А., Москин Н.Д. Методы анализа данных в задаче разграничения фольклорных и авторских текстов // Вопросы языкознания. Москва, 2020. № 2. С. 61 – 74.
[21] Розенблюм Н.Г. Петербургские пожары 1862 г. и Достоевский (запрещенные цензурой статьи журнала «Время») // Достоевский Ф. М. Новые материалы и исследования. Литературное наследство. М., 1973. Т. 86. С. 16 – 54.

Application of Decision Trees for Analyzing the Strong Positions of the Text in the Problem of Attribution of Works by F. M. Dostoevsky

A. Rogov, A. Lebedev, R. Abramov, N. Moskin, K. Kulakov

Petrozavodsk State University

The paper considers a set of articles by F. M. Dostoevsky and other authors (M. M. Dostoevsky, N. N. Strakhov, A. A. Golovachev, I. N. Shill, A. Grigoriev, A. U. Poretsky, Y. P. Polonsky) published in the magazines "Time" and "Epoch" in the period 1861-1865. Fragments of 500, 700 and 1000 words were selected in the texts. At the same time a step (100, 200 words and so on) was used to count the beginning of the next fragment to increase the sample size. Based on the

distribution of parts of speech of text fragments, decision trees were constructed, whose nodes contain branching conditions based on the frequency of occurrence of a particular n-gram (a sequence of n encoded parts of speech).
Analysis of the strong positions of these texts (i.e. fragments located at the beginning or end of the text) with the help of decision trees shows the possibility of stylistic editing, which was made by F. M. Dostoevsky in the texts of the original authors. The SMALT information system ("Statistical Methods of Analysis of Literary Texts") was used to conduct the study, where automated markup of texts with manual control of specialists of philology was implemented.

Keywords: text attribution, corpus linguistics, F. M. Dostoevsky, strong positions of the text, decision tree, n-gram, partial distribution

Reference for citation: Rogov A., Lebedev A., Abramov R., Moskin N., Kulakov K. Application of decision trees for analyzing the strong positions of the text in the problem of attribution of works by F. M. Dostoevsky // Computer Linguistics and Computing Ontologies. Vol. 4 (Proceedings of the XXIII International Joint Scientific Conference «Internet and Modern Society», IMS-2020, St. Petersburg, June 17-20, 2020). - St. Petersburg: ITMO University, 2020. P. 118 – 127. DOI: 10.17586/0000-0000-2020-4-118-127

References

[1] Gasparyan S.K., Knyazyan A.T. K voprosu ob izuchenii individual'nogo stilya avtora // Filologicheskie nauki. 2004. № 4. S. 50 – 57. [In Russian].
[2] Bolotnova N.S. i dr. Kommunikativnaya stilistika hudozhestvennogo teksta: leksicheskaya struktura i idiostil' / N.S. Bolotnova, I.I. Babenko, A.A. Vasil'eva, S.M. Karpenko, O.V. Orlova, S.V. Sypchenko, R.YA. Tyurina. Tomsk. 2001. 331 s. [In Russian].
[3] SHarkunova O.V. Idiostil' hudozhestvennogo teksta kak individual'noe sochetanie ekstra- i intralingvisticheskih parametrov, osnovannyh na referentnyh otnosheniyah // Materialy mezhdunarodnoj zaochnoj nauchno-prakticheskoj konferencii «Aktual'nye voprosy filologii, iskusstvovedeniya i kul'turologii». Novosibirsk. 2011. S. 75 – 80. [In Russian].
[4] Tarasova I.A. Kategorii kognitivnoj lingvistiki v issledovanii idiostilya // Vestnik SamGU. 2004. № 1 (31). S. 163 – 169. [In Russian].
[5] Fokina YU.M. Osobennosti reprezentacii individual'no-avtorskoj konceptosfery v angloyazychnoj i russkoyazychnoj proze (na materiale rasskazov A.P. CHekhova i D. Dzhojsa). Avtoref. dis. … kand. filol. nauk. Saratov, 2010. 184 s. [In Russian].
[6] Vinogradov V.V. Problema avtorstva i teoriya stilej // M.: Hudozhestvennaya literatura. 1961. 613 s. [In Russian].
[7] Lebedev A.A. Idiostil' P. A. Vyazemskogo: sintaksicheskij aspekt // Petrozavodsk: Izd-vo PetrGU. 2013. 134 s. [In Russian].
[8] Arnol'd I.V. Znachenie sil'noj pozicii dlya interpretacii hudozhestvennogo teksta // Inostrannye yazyki v shkole. M. 1978. № 4. S. 23 – 31. [In Russian].
[9] Gal'perin I.R. Tekst, kak ob"ekt lingvisticheskogo issledovaniya // M.: Nauka, 1981. 139 s. [In Russian].
[10] Patroeva N.V., Lebedev A.A. Sintaksicheskaya organizaciya, razmer i semantika inicial'nyh predlozhenij v lirike A. S. Pushkina // Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya. Tomsk. 2018. № 53. S. 224 – 236. [In Russian].
[11] Petrova K.V. Sil'nye pozicii teksta v avtobiografii Dzhanet Uinterson // Vestnik Novgorodskogo gosudarstvennogo universiteta. Seriya «Gumanitarnye nauki». Velikij Novgorod. 2015. № 4 (87), chast' 1. S. 73 – 76. [In Russian].
[12] Rogov A.A. i dr. Matematicheskie metody atribucii tekstov / A.A. Rogov, A.V. Sedov, YU.V. Sidorov, T.G. Surovcova // Petrozavodsk: Izd-vo PetrGU. 2014. 95 s. [In Russian].

[13] Stamatatos E. A Survey of Modern Authorship Attribution Methods // Journal of the American Society for Information Science and Technology. 2009. Vol. 60, № 3. P. 538 – 556.
[14] Romanov A.S. Metodika i programmnyj kompleks dlya identifikacii avtora neizvestnogo teksta. Dis. ... kand. tekhn. nauk. Tomsk, 2010. 149 s. [In Russian].
[15] Kjetsaa G. Attributed to Dostoevsky: The Problem of attributing to Dostoevsky anonymous articles in Time and Epoch. Oslo: Solum Forlag A. S., 1986.
[16] Rogov A.A., Kulakov K.A., Moskin N.D. Programmnaya podderzhka v reshenii zadachi atribucii tekstov // Programmnaya inzheneriya. M.: Izd-vo "Novye tekhnologii", 2019. T. 10, № 5. S. 234 – 240. [In Russian].
[17] Kotov A.A. i dr. Lingvisticheskie korpusy / A.A. Kotov, Z.I. Mineeva, A.A. Rogov, A.V. Sedov, YU.V. Sidorov // Petrozavodsk: Izd-vo PetrGU, 2014. [In Russian].
[18] Breiman L., etc. Classification and regression trees / L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone. Wadsworth, Belmont, Ca, 1984.
[19] Dubovik A.R. Avtomaticheskoe opredelenie stilisticheskoj prinadlezhnosti tekstov po ih statisticheskim parametram // Komp'yuternaya lingvistika i vychislitel'nye ontologii. Vypusk 1. Trudy XX Mezhdunarodnoj nauchnoj konferencii «Internet i sovremennoe obshchestvo», IMS-2017 (Sankt-Peterburg, 21 – 23 iyunya 2017 g.). SPb.: Universitet ITMO, 2017. S. 29 – 45. [In Russian].
[20] SHCHegoleva L.V., Lebedev A.A., Moskin N.D. Metody analiza dannyh v zadache razgranicheniya fol'klornyh i avtorskih tekstov // Voprosy yazykoznaniya. Moskva, 2020. № 2. S. 61 – 74. [In Russian].
[21] Rozenblyum N.G. Peterburgskie pozhary 1862 g. i Dostoevskij (zapreshchennye cenzuroj stat'i zhurnala «Vremya») // Dostoevskij F. M. Novye materialy i issledovaniya. Literaturnoe nasledstvo. M., 1973. T. 86. S. 16 – 54. [In Russian].