УДК 81’33

Л.О. Васькина, А.Ю. Хоменко

СЛОВА-ИНТЕНСИФИКАТОРЫ КАК СПОСОБ ИДЕНТИФИКАЦИИ АВТОРА ПИСЬМЕННОГО ТЕКСТА

Аннотация. В работе говорится о процессе атрибуции текста, его взаимосвязи с идиостилем и идиолектом автора текста. Важной частью работы является категоризация лексико-семантической группы слов-интенсификаторов. Проверяется возможность использования слов-интенсификаторов как средства идентификации автора письменного текста.

Ключевые слова: слова-интенсификаторы, идиостиль, атрибуция текста, количественные методы.

L.O. Vaskina, A.Yu. Khomenko

INTENSIFIER WORDS AS A WAY OF WRITTEN TEXT AUTHOR IDENTIFYING

Abstract. In the paper the main steps of text attribution process are named, the relationship between attribution and individual written style is described. An important part of the work is the categorization of the semantic field of intensifier words. The possibility of using intensifier words as means of distinguishing individual styles of written text authors is tested.

Keywords: intensifier words, individual written styles, authorship attribution, quantitative text processing methods.

Проблема определения автора текста привлекла пристальное внимание ученых в XX в. Так, В.В. Виноградов в своих трудах указывал на значимость разработки теории атрибуции и перечислил условия, в которых обращение к методам атрибуции может быть необходимым [4]. На современном этапе развития научного направления задача идентификации автора текста зачастую стоит очень остро. В основном это актуально для таких сфер деятельности, как искусствоведение, юриспруденция, политика, коммерческая сфера и др. Существующие на сегодняшний день разработки в области авторизации не являются совершенными. Актуальным остается вопрос о создании методики с возможностью применения её на корпусе текстов любого объема и функциональной отнесенности.

Экспертиза авторства текста может быть проведена множеством различных способов. Фокус может быть направлен на разные

языковые уровни, всевозможные элементы языка и речи и отношения между ними, их анализ с применением множества разнящихся подходов.

В данной работе используется комплексная методика. При анализе применяются как качественные (квалификативные), так и количественные методы: метод компонентного анализа, метод контекстного анализа, количественный (в частности, статистический) метод.

Цель исследования заключается в определении эффективности использования слов-интенсификаторов в качестве инструмента для идентификации идиостиля автора письменного текста. В качестве объекта анализа был выбран индивидуальный авторский стиль, репрезентируемый в аспекте компонентов конкретной лексико-семантической группы.

В основу исследования положена гипотеза о том, что интенсификаторы, будучи компонентами языковой системы, могут выступать в роли идентификаторов языковой личности, а коэффициент корреляции Пирсона – метрикой для сравнения авторских индивидуальных стилей при авторизации текста.

Человек творчески использует единицы и приемы языка для выражения своих мыслей. В своей речи индивид отбирает языковые средства и употребляет их с определенной частотой. Другими словами, автор избирает определенный способ речевого самовыражения, который является уникальным, индивидуальным стилем автора, или *идиостилем*. При исследовании данного явления невозможно обойтись без понятия *языковая личность (ЯЛ)*. Интерес к его изучению возрос во второй половине XX в. после публикации трудов В.В. Виноградова об особенностях языка автора художественного текста [4]. На дальнейшее развитие и популяризацию понятия *языковая личность* в научной сфере повлиял Ю.Н. Караулов, определяющий ЯЛ как «многоуровневую совокупность языковых навыков, необходимых для осуществления речевой деятельности» [5, с. 104].

Индивидуальный язык личности характеризуется набором особенностей, свойственных речи этой личности, и называется *идиолектом*. Данные особенности могут быть обнаружены на всех языковых уровнях: фонетическом, морфологическом, лексическом, синтаксическом. Языковая личность не только избирает ту или иную единицу в ряду других подобных, но и отдает предпочтение определенному варианту сочетания различных единиц. Следует также

выделить термин *идиостиль,* который связан именно «с процессом формирования языковой личности» [6, с. 119]. Идиостиль, в отличие от идиолекта, включает в себя коммуникативные и прагматические намерения автора, а также способы реализации этих намерений.

Как уже было сказано, индивид сознательно отбирает те или иные языковые средства для формирования смыслов в речи. В последствии эти языковые средства анализируются исследователями для характеристики идиостиля автора текста и описания специфических особенностей его речи. Этот анализ, в свою очередь, может быть использован для решения задач *атрибуции*. В узком смысле, атрибуция – это установление авторства текста, в широком смысле, это понятие включает в себя приписывание тексту некоторых подходящих ему атрибутов, в том числе, имени автора, жанра, места, времени создания и пр.

На сегодняшний день существует большое количество методик авторизации текста. Они могут различаться, например, в зависимости от уровня языка, которому отдается предпочтение при анализе. Распространено мнение, что работу следует проводить совокупно сразу на нескольких или даже на всех языковых уровнях [8; 10]. В настоящее время большой популярностью пользуются количественные методы исследования, различные математические, в частности, статические методы. Они позволяют проводить анализ автоматически с использованием компьютерных инструментов. К примерам таких подходов можно отнести предложенную А.Н. Барановым во «Введении в прикладную лингвистику» [1, с. 25] методику количественного исследования квазисинонимичных лексем. Именно эта методика послужила прототипом для описываемой в настоящей работе. Основой методики А.Н. Баранова является анализ частоты употребления служебных слов в тексте.

На сегодняшний день универсальный метод, который бы указывал, какие именно данные и какое их количество необходимо для однозначного объективного определения автора текста, еще не разработан. Не вызывает сомнения тот факт, что идиолект, как «различительный, специфический выбор, делающийся индивидом при порождении текста» [9, с. 431] (перевод наш – *Авт.*), играет важную роль при его атрибуции. Он позволяет прослеживать предпочтения и способности автора в использовании функционально-формальных элементов языка.

Лексические средства выражения *интенсивности* признака являются одними из экспликатов авторского идиолекта. Наиболее характерным для русского языка средством выражения интенсивности того или иного признака служит группа так называемых *слов-интенсификаторов*. «Интенсификаторы – это лексические единицы языка, которые выполняют функцию модификаторов знаменательных элементов» [2, с. 47]. Иными словами, интенсификаторы указывают на количественное изменение компонента значения в сторону его увеличения или уменьшения. Наиболее многочисленными и характерными представителями лексико-семантической группы слов-интенсификаторов являются наречия, однако не стоит забывать, что они не единственные. Значение интенсивности может быть выражено при помощи прилагательных, например, *Здесь творится* ***настоящий*** *бардак!,* при помощи местоимений *(**Какая** замечательная погода!),* частиц *(**Вот так** лето!)*. Таким образом, для достижения цели настоящего исследования был создан аутентичный список слов-интенсификаторов из 93 единиц, в который наряду с наречиями вошли некоторые прилагательные, местоимения, частицы. Данный список был сформирован при помощи грамматических справочников, различных научных статей, словарей синонимов русского языка, исследования корпусов русского языка, в том числе собранных автоматически («Araneum Rusicum», «ruTenTen»), и метода интроспекции.

Работа проводилась на авторизованном материале художественных текстов для того, чтобы подтвердить или опровергнуть исследовательскую гипотезу: слова-интенсификаторы могут выступать в роли идентификаторов языковой личности при решении идентификационной задачи атрибуционного исследования. Было использовано три текста российских авторов: роман В.О. Пелевина «Омон Ра» и две повести В.Г. Сорокина – «День опричника» и «Метель». Повесть «День опричника» выступала в роли спорного произведения, автор которого якобы неизвестен; «Омон Ра» В. Пелевина и «Метель» В. Сорокина выступали в качестве текстов-образцов. Так, решается идентификационная задача атрибуционной лингвистики типа «сравнение по образцу» [1, с. 25].

Вышеназванные тексты прошли предварительную обработку для последующего изучения. Лексические единицы в них были приведены к начальной форме, то есть лемматизированы, с помощью

консольной программы MyStem, производящей морфологический анализ русскоязычных текстов.

В дальнейшем была проведена количественная обработка материала при помощи инструмента корпусного анализа AntConc. С его помощью были обработаны все лексические элементы из предварительно составленного списка слов-интенсификаторов. Число демонстрируемых на экране строк не было ограничено, были представлены все случаи употребления единиц в анализируемом тексте (Concordance Hits = N).

Затем посредством функций Concordance и Concordance Plot для соответствующего элемента методом контекстного анализа были выделены случаи употребления лексемы именно со значением интенсификации.

Для возможности последующего сопоставления величин каждая полученная абсолютная частота была преобразована в относительную, а именно – частоту IPM (instances per million), находящуюся по следующей формуле:

$$\text{IPM} = \frac{\text{абсолютная частота вхождений слова}}{\text{объём корпуса}} \cdot 1\,000\,000$$

Аналогичная работа была проведена для каждого текста, по результатам была составлена матрица, где количество строк n – число элементов анализа, количество столбцов m – число рассматриваемых текстов, для всех анализируемых слов (фрагмент матрицы см. в *таблице 1)*. Слова-интенсификаторы с частотой встречаемости равной нулю во всех трех корпусах в дальнейшем анализе не учитывались и для удобства были удалены из матрицы.

Сравнительный анализ проводился в трех направлениях. Во-первых, было проведено сопоставление спорного текста («День опричника») с двумя авторизированными текстами («Метель» и «Омон Ра»). Во-вторых, повесть «Метель», автор которой был идентифицирован и совпадал с автором спорного текста, была аналогичным образом сопоставлена с романом «Омон Ра». Полученные в результате количественного анализа IPM-частоты использовались для подсчета некоторых статистических критериев, в том числе для нахождения коэффициента корреляции Пирсона с целью определения степени взаимосвязи наборов данных между собой *(таблица 2)*. Коэффициент корреляции был вычислен для каждой из описанных выше пар.

Таблица 1

Фрагмент матрицы абсолютных и относительных частот для анализируемых текстов

	Спорное произведение – В. Сорокин «День опричника»		***В. Пелевин «Омон Ра»***		***В. Сорокин «Метель»***	
	Абс. частота	IPM	Абс. частота	IPM	Абс. частота	IPM
абсолютно	0	0	2	65	0	0
бесконечный	1	30	1	33	9	252
довольно	1	30	24	783	2	56
еле	4	121	2	65	7	196
немного	1	30	10	326	1	28
очень	8	243	65	2121	26	728
совсем	12	364	23	750	34	952
так	20	606	24	783	26	728
ужасно	0	0	0	0	2	56
чуть	3	91	20	653	2	56

Таблица 2

Коэффициент корреляции для анализируемых пар текстов

	«День опричника» – «Метель»	***«День опричника» – «Омон Ра»***	***«Метель» – «Омон Ра»***
Коэффициент корреляции	0,83	0,62	0,69

Согласно теории математической статистики [3, с. 251], значимым считается коэффициент корреляции, превышающий величину 0,60. Тем не менее следует учитывать, что при лингвистическом исследовании, величина коэффициента корреляции, указывающего на тесную связь между наблюдаемыми явлениями, должна быть выше. Точный уровень значимого коэффициента корреляции при авторизации текста находится в зоне, превышающей пороговую величину 0,82–0,87 [7, с. 164]. В данном случае значение коэффициента корреляции 0,83 подтверждает, что сравниваемые тексты принадлежат одному автору. Наряду с этим, величины коэффициента корреляции равные 0,62 и 0,69 доказывают отсутствие существенной взаимосвязи между текстом В.О. Пелевина и текстами В.Г. Сорокина.

Данная работа преследовала цель определить эффективность использования слов-интенсификаторов в качестве инструмента для идентификации автора письменного текста и проверить гипотезу о релевантности использования коэффициента корреляции Пирсона для сравнения авторских индивидуальных стилей. В процессе изучения научной литературы, которая составила теоретико-методологическую базу исследования, количественного анализа компонентов речи в отобранном для изучения материале, а также сравнительного анализа полученных данных были сделаны следующие выводы:

– лексико-семантическая группа слов-интенсификаторов может быть использована в качестве средства идентификации автора письменного текста, так как количественный анализ данной группы слов дает валидные результаты;
– коэффициент корреляции Пирсона, превышающий уровень 0,82–0,87, доказывает, что сравниваемые индивидуальные стили близки и автором сравниваемых текстов является одно лицо.

Библиографический список

1. *Баранов А.Н.* Введение в прикладную лингвистику: учебное пособие. М., 2001.
2. *Безрукова В.В.* Интенсификация и интенсификаторы в языке и речи (на материале английского языка): дис. … канд. филол. наук. Воронеж.: 2004.
3. *Гмурман В.Е.* Руководство к решению задач по теории вероятностей и математической статистике. М., 1979.
4. *Виноградов В.В.* Проблема авторства и теория стилей. М., 1961.
5. *Караулов Ю.Н.* Русский язык и языковая личность. М., 2010.
6. *Матвеева Т.В.* Полный словарь лингвистических терминов. Ростов н/Д., 2010.
7. *Радбиль Т.Б., Маркина М.В.* Вероятностно-статистические модели в производстве автороведческой экспертизы русскоязычных текстов // Политическая лингвистика. 2019. Вып. 2 (74). С. 157–167.
8. *Хоменко А.Ю. и др.* Автоматическая обработка текста и лингвистическое моделирование как способы решения проблем атрибуционной лингвистики // Политическая лингвистика. 2020. № 3 (81). С. 215–224.
9. *Coulthard M.* Author identification, idiolect, and linguistic uniqueness / M. Coulthard. Text: unmediated // Applied Linguistics. 2004. № 24 (4). P. 431–447.

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

Лидия Олеговна ВАСЬКИНА
студент
Высшая школа экономики, г. Москва

Анна Юрьевна ХОМЕНКО
старший преподаватель, стажер-исследователь лаборатории теории и практики систем поддержки принятия решений
Высшая школа экономики, г. Москва