УДК 811:81’246.3

Н.А. Власенко, Н.Л. Кузьминская, А.А. Максименко

Текстометрические исследования многоязычных научных текстов

Описаны текстометрические исследования авторских научных публикаций и корпусов текстов по использованию новых информационных технологий в образовании на украинском, русском и английском языках с целью выявления особенностей авторского научного стиля.

Textometric researches of authors' scientific publications and texts packages on using new information technologies in education in Ukrainian, Russian and English languages are described to reveal the features of the author's scientific style.

Описано текстометричні дослідження авторських наукових публікацій та корпусів текстів з використання нових інформаційних технологій в освіті українською, російською та англійською мовами з метою виявлення особливостей авторського наукового стилю в освіті.

Введение. Термины *текстометрия* и *лексикометрия*, а также соответствующие исследования не получили должного распространения не только в украинской научной литературе, но и на постсоветском пространстве в целом. Наибольшей популярности эти исследования получили во Франции, где впервые указанные термины были введены в научный лексикон (*лексикометрия* в 60-х годах, а *текстометрия* в 90-х годах прошлого столетия). В Сорбонне даже создан текстометрический центр. У франкоязычной литературе под термином *текстометрический анализ* («*analyse textométrique*») понимают серию методов, предоставляющих возможность исследователю формально реорганизовывать тексты и проводить статистический анализ корпуса текстов. При этом под корпусом текстов понимается совокупность текстов, объединенных с целью сравнения и служащих базой для квантитативных исследований. При лексикометрическом анализе измеряются различные параметры лексико-семантической системы.

Нельзя сказать, что подобные исследования в Советском Союзе и на постсоветском пространстве не проводились и соответствующие программные системы не разрабатывались. В большинстве своем эти исследования относили к направлению *квантитативная лингвистика* и называли различными терминами *статистический анализ текста*, *нумерология текста*, *квантитативный* (*количественный*) *анализ текста* и др. По мнению авторов, термины *лексикометрия* и *текстометрия* точнее отражают суть исследуемого явления. Напрашивается сравнение с биометрией, представляющей собой совокупность автоматизированных методов и средств идентификации человека, основанных на его физиологической или поведенческой характеристике. Так же как биометрические технологии используются в задачах уникальной идентификации личности, текстометрические технологии могут использоваться в задачах определения авторства текста, или более широко – *атрибуции текста*. Под *атрибуцией текста* понимают соотнесение тексту соответствующих ему атрибутов: имя создателя, жанр произведения, время создания и др. Анализ существующих методов и методик по определению авторства текстов приведен в [1]. Вводится понятие *авторского инварианта*, под которым понимают количественную характеристику литературных текстов, однозначно характеризующую произведения одного автора и принимающую иные значения для произведений других авторов. В качестве такого авторского инварианта в [2] предлагается использовать частоту употребления 55-ти выделенных служебных слов (предлогов – 24, союзов – 14, частиц – 17).

История вопроса

Поиск методов, с помощью которых можно было бы определить авторство текста, интересовал исследователей давно, но наибольшую остроту и актуальность он приобрел в 70-е годы в связи со скандалом, вызванным сомнениями по поводу авторства М.А. Шолохова первых двух томов романа «Тихий Дон». Лауреата Нобелевской премии 1965 г. в области литературы М.А. Шолохова обвиняли в плагиате.

В 1984 г. норвежские ученые на основе статистических исследований подтвердили авторство М.А. Шолохова [3], а данные, полученные в [2] на основе анализа частоты употребления служебных слов, вновь подвергли его сомнению. Даже обнаруженные в 1999 г. рукописи романа не остановили споров.

Изложенное подтверждает тот факт, что определение авторства текста является задачей достаточно сложной, а полученные результаты бывают порой противоречивыми. Приведенный выше пример наглядно демонстрирует, что использовать какую-либо одну характеристику языка автора для задач атрибуции текста не вполне оправдано. Импонирует методика определения авторства, используемая в модели ЛингвоАнализатора [1].

Доказательство эффективности предлагаемых методов или методик разработчики, как правило, демонстрируют на примере художественных текстов, а разрабатываемые программные средства анализируют тексты на одном языке. С точки зрения авторов, если какая-либо методика определения атрибуции текста является истинной для одного языка, она должна быть истинной и для некоторых других языков. Задачи же атрибуции научного текста практически не рассматривались, поскольку научная работа, как правило, выполняется научным коллективом и поэтому статьи и монографии часто пишутся в соавторстве. Считается, что для научной статьи не столько важен стиль изложения, сколько структурированно представленные научные результаты.

Исходя из того факта, что наука воспринимается как универсальная, рациональная, безличная, такими же воспринимаются и научные тексты, хотя у каждого из них есть автор или группа авторов, создавших эти тексты. Такая кажущаяся «безликость» научного текста открывает путь в науку случайным людям, для которых она – своеобразный вид бизнеса, где результат можно получить быстро, тем более, что Интернет пестрит объявлениями о предоставлении услуг по написанию не только рефератов, но и научных статей, кандидатских и докторских диссертаций, а оспорить авторство научной публикации практически невозможно, если только она не содержит явный плагиат известной научной идеи или разработки. Нередко мы являемся свидетелями циничного плагиата, когда заимствуются не только научные результаты, но и их описания.

Особенности научного текста

Написание научной статьи требует не только знания истории проблемы, текущего состояния дел в изучаемой предметной области, осмысления полученного результата и его места среди подобных результатов, но и определенных навыков писательского мастерства. Для точного выражения научной мысли нужно найти языковые средства, адекватные этой мысли.

Научный текст отличается формально-логическим способом изложения, высокой стандартизированностью и насыщенностью специальными терминами, использованием специальных слов-организаторов научной мысли (*из сказанного следует*, *в заключение*, *в результате проведенных исследований* и др.), приглушенностью коннотативных (экспрессивно-эмоциональных, оценочных) оттенков и пр. Исчерпывающий анализ общенаучного лексикона приведен в [4].

Для анализа научных публикаций решили воспользоваться разработанной авторами программой *TextAnalyzer*[*], описанной в [5]. В качестве корпусов текстов были выбраны тексты тезисов докладов, представленных на Международную конференцию «*New information technologies in education for all*» (2006 – 2008 гг.), именуемые далее *ITEA*2006 [6], *ITEA*2007 [7] и *ITEA*2008 [8]. Поскольку доклады подавались на трех языках – украинском, русском или английском (на выбор автора), было сформировано девять корпусов текстов – по три на каждый язык. В корпус не включались название доклада, сведения об авторах, аннотации и литература. Уже первые попытки работы программы *TextAnalyzer* с научными текстами, показали следующую специфику научных публикаций, не учтенную разработчиками:

[*] Программа написана Кузьминским В.Н. и Кузьминской Н.Л.

1. В научных публикациях большой процент в сравнении с художественными текстами составляют сложные слова, пишущиеся через дефис. В предыдущих исследованиях авторы статьи использовали дефис в качестве разделителя слова, т.е. термин *информационно-образовательные технологии* разбивался на три слова *информационно*, *образовательные* и *технологии*. Следует учесть, что в сравнении с художественными текстами в научных публикациях таких сложных слов достаточно много и такое разбиение может повлиять на результат анализа.

2. В научных текстах, особенно при описании вклада в исследуемую проблему других ученых, используется запись с инициалами, например *С.П. Кудрявцева*. Программа разбивала эту запись на три слова *с*, *п*, *кудрявцева.* В результате при подсчете предлога ***с*** получали значительную погрешность, а в частотный словарь включалось несуществующее слово ***п*** и др.

Во избежание подобных ошибок, была создана новая версия программы *TextAnalyzer* версии 2.0, отличная от предыдущей не только некоторыми дополнительными функциями, но и интерфейсом.

Описание программы *TextAnalyzer* 2.0

Программа написана на объектно-ориентированном языке программирования *C#*, работающем в среде исполнения *.NET Framework* 2.0. Новая версия реализована в виде модулей, что позволяет добавлять новые функции к программе, не нарушая ее структуру.

Программа *TextAnalyzer* версии 2.0, в отличие от предыдущей, составляет не только частотный, но и реверсивный словарь для любого текстового файла, сохраненного в формате txt. Использование реверсивного словаря, в котором упорядочиваются слова не по начальным, а по конечным буквам, позволяет оценивать морфологические особенности языка автора.

Окно и все меню программы имеют вид (рис. 1).

В верхней части окна программы находятся иконки, с помощью которых можно управлять программой, а именно импортировать текстовый документ, создавать/пересчитывать словарь, загружать текст и словарь из файла, записывать текст и словарь в файл, изменять настройки языкового интерфейса, экспортировать словарь и статистику по нему в файл *Excel*, выходить из программы.

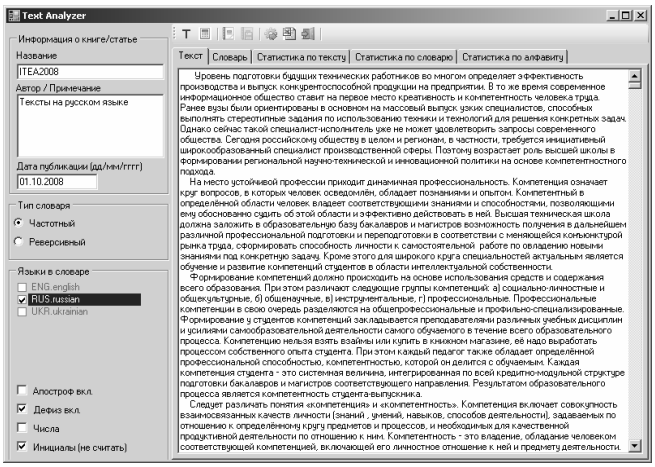

Рис. 1

До или после загрузки текста в окне программы нужно указать язык анализируемого текста, тип создаваемого словаря (частотный или реверсивный), кроме того в новой версии появилась возможность включать или выключать из рассмотрения дефис, инициалы, возможность анализа слов с апострофом и включение/исключение чисел сохранилась. Частотный/реверсивный словарь представлен в виде таблицы, состоящей из четырех столбцов: *слово*, *длина*, *частота*, *относительная частота*, аналогично предыдущей версии.

В новой версии в *Статистику по тексту* добавлен подсчет однократных слов и их относительной частоты (рис. 2). Это сделано для

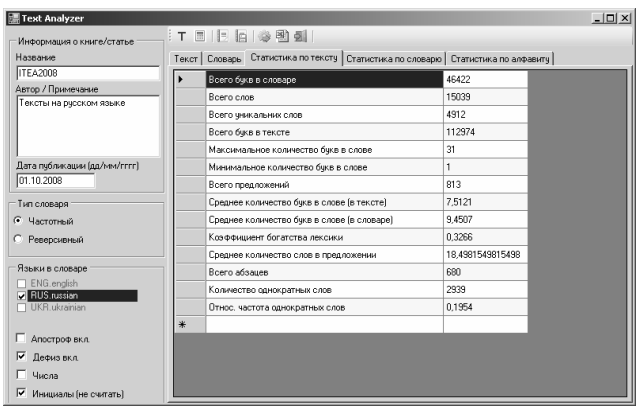

Рис. 2

подтверждения или опровержения предположения, выдвинутого в [9], что доля однократных лексем (встречающихся в тексте всего один раз) к общей массе лексем, тяготеет к золотому сечению (0,382) с небольшим отклонением в обе стороны. В новой версии добавлена полная «Статистика по алфавиту» и появилась возможность экспортировать данные в файл *Excel*.

Лексикометрические и текстометрические исследования авторского научного стиля

Для исследования выбраны тексты тезисов докладов, представленных одним и тем же автором (без соавторства)[**] на *ITEA*2006, *ITEA*2007 и *ITEA*2008, а также корпусов текстов всех тезисов докладов на трех языках, поданных на эти конференции.

Были выбраны следующие тексты тезисов докладов:

• на украинском языке – Артеменко В.Б. («Підтримка добування знань і навичок у системі моніторингу стійкості соціально-економічного розвитку регіонів» [6], «Моделювання взаємодії учасників *E*-навчання на засадах агентоорієнтованого підходу» [7], «Інституціональна підтримка дистанційних освітніх технологій у вищій школі» [8];

• на русском языке – Колос В.В. («Методики сравнительного и прогностического анализа телекоммуникационных информационно-образовательных сред», [6], «Таксономия телекоммуникационных информационно-образовательных сред» [7], «Информационный мониторинг телекоммуникационных информационно-образовательных сред» [8];

• на английском языке – *Ion Roceanu* «"*CAROL I*" *National Defense University`s Elearning Pilot Centre*» [6], «*E-education versus e-training*» [10], «*Virtual Learning Space Designed to Simulate Natural Disasters Scenarios and Citizens' education*» [8].

Таблица

Тексты	Конференция	Среднее количество букв в слове		Коэффициент богатства лексики	Среднее количество слов в предложении	Относительная частота однократных слов
		в тексте	в словаре			
Артеменко	*ITEA*2006	**7,1064**	**8,2654**	**0,5184**	**18,07692**	**0,3794**
	*ITEA*2007	**6,6798**	**7,9417**	**0,5268**	**19,70423**	**0,3724**
	*ITEA*2008	**6,7782**	**8,4063**	**0,4585**	**20,44444**	**0,3147**
Корпус текстов на укр.языке	*ITEA*2006	**6,9581**	**9,0776**	**0,2414**	**20,41085**	**0,1357**
	*ITEA*2007	**6,8822**	**9,1778**	**0,2315**	**20,58518**	**0,1315**
	*ITEA*2008	**6,9003**	**8,9848**	**0,23**	**19,18498**	**0,1259**
Колос	*ITEA*2006	**7,7608**	**9,0576**	**0,4637**	**17,86364**	**0,3149**
	*ITEA*2007	**7,9693**	**9,1349**	**0,4828**	**14,14474**	**0,3265**
	*ITEA*2008	**7,6429**	**9,1407**	**0,4398**	**22,54321**	**0,2985**
Корпус текстов на рус.языке	*ITEA*2006	**7,2645**	**9,6486**	**0,2083**	**17,91322**	**0,1109**
	*ITEA*2007	**7,4815**	**9,5571**	**0,3024**	**18,09293**	**0,1762**
	*ITEA*2008	**7,5121**	**9,4507**	**0,3266**	**18,49815**	**0,1954**
Roceanu	*ITEA*2006	**5,3539**	**6,7836**	**0,3139**	**28,0122**	**0,1855**
	ICVL 2007	**5,4645**	**7,0041**	**0,3613**	**24,30909**	**0,2274**
	*ITEA*2008	**5,7716**	**7,2202**	**0,3356**	**29,52727**	**0,2026**
Корпус текстов на анг.языке	*ITEA*2006	**5,4487**	**7,6712**	**0,1564**	**18,72736**	**0,0699**
	*ITEA*2007	**5,6786**	**7,6348**	**0,2395**	**18,53015**	**0,1336**
	*ITEA*2008	**5,4435**	**7,6503**	**0,1573**	**18,69821**	**0,067**

Некоторые из полученных статистических данных приведены в таблице.

Общие выводы по данным таблицы:

• Полученные данные подтверждают результаты как наших исследований[5, 8], так и исследований других авторов, состоящие в том, что независимо от анализируемых текстов (художественных, научных, корпусов текстов или авторских текстов) средняя длина слова английского слова как в тексте, так и словаре, построенном на базе этого текста, как правило, короче, чем у текстов на украинском и русском языках, а средняя длина украинских слов практически всегда короче русских в текстах одного функционального стиля.

• Соотношение золотого сечения можно наблюдать в авторских научных текстах, написанных на русском и украинском языках, и практически не возможно в авторских научных текстах на английском языке. На корпусах научных текстов по использованию новых информационных технологий в обучении такое соотношение невозможно для трех рассматриваемых языков.

[**] Исключение составляют тексты тезизов *Roceanu* – на *ITEA*2008 тезисы были поданы в соавторстве, а на *ITEA*2007 вообще не подавались и использовались тексты тезисов, представленных этим автором на *ICVL* 2007 [10].

• Коэффициент богатства лексики (отношение количества уникальных слов ко всем словам текста) авторских научных текстов выше, чем корпусов текстов аналогичной научной тематики.

Что касается особенностей авторского стиля, то здесь выделяются работы *Roceanu*, где намного превышаются значения среднего количества слов в предложении в сравнении с корпусом английских текстов. При рассмотрении текстов отмечено, что для указанного автора характерны длинные перечисления, что естественно влияет на длину предложения. Для текстов Колос В.В. характерно употребление длинных слов, при этом употребление сложных слов, пишущихся через дефис, минимально. Среди данных таблицы по текстам Артеменко В.Б. больших отклонений от данных корпусов украинских текстов не выявлено, но при анализе словаря отмечен большой процент слов на английском языке и сложных слов типа *агент-орієнтований*, *інформаційно-комунікаційні*, *навчально-методичні*, *науково-навчальний*, *науково-педагогічний* и др. Эти же особенности авторского стиля прослеживаются и в текстах автора, написанных на русском (см. статью Артеменко В.Б. в этом журнале).

В текстах Артеменко В.Б. и Колос В.В. проверена устойчивость такого параметра, как доля служебных слов из списка [2]. Полученные данные по статьям Артеменко – 0,018; 0,026; 0,014822 и Колос – 0,0197201; 0,0204841; 0,0169769. Такую нестабильность можно объяснять малым объемом тезисов, который существенно меньше задекларированного в [2] допустимого размера для получения достоверных результатов – 16 тыс. слов, поэтому невозможно подтвердить или опровергнуть достоверность этого параметра для научных текстов.

Заключение. К сожалению, сегодня нельзя говорить о том, что в текстометрии как художественных, так и научных текстов, есть параметр, способный дать достоверный результат авторства текста, аналогичный анализу ДНК или отпечатков пальцев в биометрии. Тем не менее доказать авторство как художественных, так и научных текстов возможно, проанализировав серию параметров. Для облегчения работы исследователям необходимо специализированное программное обеспечение, позволяющее проводить разноплановые измерения текста. Применение предложенной здесь программы *Text-Analyzer* версии 2.0 уже дает неплохие результаты. Готовится дополнительный модуль сравнения словарей, что поможет не только улучшить достоверность определения авторства, но и определить тенденции в развитии языка вообще и языка науки, в частности; отследить появление новых слов и терминов, а также слов и терминов, исчезающих из употребления.

1. *Хмелев Д.* Краткая история разработки методик определения авторского стиля. – http://www.rusf.ru/books/analysis/history.htm
2. *Фоменко В.П.*, *Фоменко Т.Г.* Авторский инвариант русских литературных текстов. Новая хронология Греции: Античность в средневековье. – М.: Изд-во МГУ, 1996. – Т. 2. – С. 768–820. – http://lib.ru/FOMENKOAT/greece.txt
3. *Кто* написал «Тихий Дон»? (Проблема авторства «Тихого Дона») / Г. Хьетсо, С. Густавссон, Б. Бекман и др. М.: Книга, 1989. – 186 с. – http://feb-web.ru/ feb/sholokh/critics/h89/h89-0162.htm
4. *Большакова Е.И.*, *Васильева Н.Э.*, *Морозов С.С.* Лексико-синтаксические шаблоны для автоматического анализа научно-технических текстов. – http://www.raai.org/resurs/papers/kii-2006/doklad/Bolshakova.doc
5. *Власенко Н.А.*, *Кузьминская Н.Л.*, *Максименко А.А.* Многоязычие в эпоху глобализации: исследование и примеры использования // УСиМ. – 2008. – №1. – С. 60–70.
6. *Сборник* трудов Первой междунар. конф. «Новые информационные технологии в образовании для всех», 29–31 мая 2006 года,. – К.: Академпериодика, 2006. – 530 с.
7. *Сборник* трудов Второй междунар. конф. «Новые информационные технологии в образовании для всех: состояние и перспективы развития», 21–23 нояб. 2007 года. – К.: Академпериодика, 2007. – 458 с.
8. *Сборник* трудов Третьей междунар. конф. «Новые информационные технологии в образовании для всех: система электронного образования», 1–3 окт. 2008 г. – К.: Академпериодика. – 2008. – 468 с.
9. *Мартыненко Г.Я.* Золотое сечение в нумерологии текста. – http://www.trinitas.ru/rus/doc/0232/004a/02321035.htm
10. *Roceanu Ion* E-education versus e-training Life Long Learning perspective // The 2nd Intern. Conf. on Virtual Learning, ICVL 2007. – http://www.cniv.ro/2007/disc2/icvl/documente/pdf/invited/invited3.pdf