УДК 811:81’246.3

Н.А. Власенко

Текстометрия: художественный текст *VS* научный текст

Описаны текстометрические исследования текстов художественных произведений известных русских и украинских писателей в сравнении с научными текстами на русском и украинском языках.

Textometric analysis of artistic texts of famous Russian and Ukrainian writers in comparison with scientific texts in Russian and Ukrainian is described.

Описано текстометричні дослідження текстів художніх творів відомих російських та українських письменників у порівнянні з науковими текстами на російській та українській мовах.

Введение. Тексты различных функциональных стилей, условно разделяемые на разговорные, научные, официально-деловые, публицистические, литературно-художественные, различаются по всем основным параметрам — лексическим, грамматическим, фонетическим. Эти основные параметры можно исследовать не только описательно, но и точно измерить для каждого отдельного языка.

Для исследования были выбраны тексты научного стиля, в котором доминирует информативная функция, имеющая когнитивный характер, и тексты литературно-художественного стиля на русском и украинском языках. Из параметров в основном рассматривалась лексика и фонетика.

Предлагаемые исследования относятся к текстометрическим исследованиям, хотя, на первый взгляд, более уместным было бы употребление устоявшегося термина *стилеметрия*, введенного в научный обиход еще в 1880 году немецким филологом В. Диттенбергером, исследовавшим авторство диалогов Платона с помощью подсчета частот слов, в основном, служебных. По мнению автора, предпочтительнее использовать более широкий современный термин *текстометрия*, введенный в научный обиход французскими лингвистами только в 90-х годах прошлого века, поскольку полученные данные можно использовать не только для распознавания стилистических особенностей, но и для характеристики текста и языка в целом.

Статистические исследования текстов

Не всегда статистическим результатам можно безоговорочно доверять. Как пример, в таблице средних длин словоформ ста языков мира [1], дается средняя длина русской словоформы – 4,701 букв, украинской – 5,156, английской – 3,042, немецкой – 5,448, но не указано, на которой генеральной совокупности текстов были получены эти данные, которые по нашим наблюдениям, не всегда соответствуют действительности. На текстах различных функциональных стилей были получены отличные от приведенных результаты [2, 3].

На первый взгляд, заслуживают большего доверия данные [4], когда проводились статистические исследования на больших корпусах художественных текстов (около 10 млн буквенных символов) и были получены следующие значения: среднее значение длины слова для немецких текстов составляет 5,07 буквенных символа, английских – 4,24, а для русских – 5,13. Но теперь посмотрим, на каких текстах были получены эти результаты (для русского языка):

1. Скотт В. Айвенго.
2. Набоков В.В. Лолита.
3. Толстой Л.Н. Анна Каренина.
4. Дефо Д. Робинзон Крузо.
5. Диккенс Ч. Приключения Оливера Твиста.

Из них только два романа были написаны русскими авторами на русском языке, и то с некоторыми оговорками (в текстах Л. Толстого обычно бывает представлена не только русская, но и французская лексика). Три остальные – переводные, а значит стилистически связанные с оригиналом, уже не говоря о возможных ошибках переводчиков. Исследователи не указывают, какое программное обеспечение использовалось, и как проводился подсчет.

Казалось бы, элементарный подсчет средней длины слова можно проводить по-разному, исходя из того, *что* понимать под словом или с учетом того, *для каких целей* проводится этот подсчет. Приведем небольшой перечень вопросов, на которые исследователь должен обратить внимание:

• является ли число, записанное цифрами в тексте, или формула словом? Если подсчитывается средняя длина слова определенного языка, например, русского, то, наверное, число или формулу лучше не включать в подсчет, но если интересно знать среднюю длину слова в тексте (без особого акцента на язык написания), то, скорее всего, следует включать.

• Как считать сложные слова, пишущиеся через дефис, – одним словом или несколькими отдельными словами? Так, в романе Анна Каренина встречается большое количество сложных слов. Есть даже такие, в составе которых несколько дефисов (тюлево-ленто-кружевно-цветной, ивановско-штраусовско-ренановское, буро-красно-загорелое и др.). Очевидно, для некоторых задач (например, выявление лексических особенностей авторского стиля или при построении частотных авторских словарей такие слова лучше считать одним словом. Но тогда возникает следующая задача – включать ли дефис в подсчет длины слова, т.е. ивановско-штраусовско-ренановское – это слово длиною 33 знака или 30 букв?

• Учитывать ли слова, написанные на иностранных языках, в подсчет средней длины слова на исследуемом языке?

• Считать ли инициалы отдельными словами, либо приплюсовывать их к фамилии, либо вообще опускать?

Таких вопросов возникает множество. Поэтому при подсчете каких-либо параметров, необходимо указывать, исходя из каких ограничений проводился подсчет, чтобы этим данным можно было доверять.

В предлагаемом исследовании измерение проводилось с помощью программы *Text Analyzer*, разработанной в Международном научно-учебном центре информационных технологий и систем [3, 5]. Особенность этой программы – настройка статистических исследований на различные классы лингвистических задач и возможность исследовать тексты, написанные на любых языках (в основе которых лежит алфавит), включая редкие языки, как, например, гагаузский. Отмечая или убирая в блоках отметки апостроф, число, дефис, инициалы, фактически задаются рамки (условия) проведения исследований.

Изначально мы определяем идентичные для всех анализируемых текстов условия подсчета (исключаем из подсчета числа, формулы, инициалы, а сложные слова, пишущиеся через дефис, считаем одним словом).

Текстометрические исследования художественных текстов

Для исследования были взяты тексты известных русских писателей XIX –XX веков – А.С. Пушкина (1 – «*Полтава*», 2 – «*Барышня-крестьянка*», 3 – «*Сказка о царе Салтане*»), А.П. Чехова (1 – « *Дама с собачкой*», 2 – «*Анна на шее*», 3 – «*Душечка*»), Н.В. Гоголя (1 – «*Тарас Бульба*», 2 – «*Ночь перед Рождеством*»), а также тексты известных украинских писателей О. Кобылянской (1 – «*В неділю рано зілля копала*»), Л. Украинки (1 – «*Лісова пісня*», 2 – «*Давня казка*»), И. Франко (1 – «*Абу-Касимові капці*», 2 – «*Борислав сміється*», 3 – «*Захар Беркут*»), М. Коцюбинского (1 – «*Тіні збутих предків*»), Ю. Покальчука (1 – «*Заборонені ігри*»). Сводные таблицы базовых текстометрических показателей, полученных с использованием программы, приведены в табл. 1 на русском языке, а в табл. 2 – на украинском.

Исходя из данных таблиц, определим интервал (минимальное и максимальное среднее количество букв) для художественных текстов на русском языке: среднее количество букв в слове в тексте (4,6101; 5,1347); среднее количество букв в слове в словаре (5,9992; 7,1755). И то же самое для художественных текстов на украинском языке: среднее количество букв в слове в тексте (4,1901; 4,9706); среднее количество букв в словаре (6,0421; 7,4833), т.е. заявленные в [1, 4] средние длины русской словоформы – 4,701 и 5,13 букв, соответственно, попадают в наш интервал, а значит, этим данным можно

Таблица 1

Параметры	Пушкин (1)	Пушкин (2)	Пушкин (3)	Чехов (1)	Чехов (2)	Чехов (3)	Гоголь (1)	Гоголь (2)
Всего букв в словаре	20201	17009	7889	14080	12229	11423	84247	31331
Всего слов	6422	5427	4009	5113	3978	3851	38549	13155
Всего уникальных слов	3148	2379	1315	2052	1781	1708	11741	4455
Всего букв в тексте	31546	27866	18482	24336	19534	18491	186830	64735
Максимальное количество букв в слове	16	19	13	17	19	18	19	17
Всего предложений	715	462	279	365	254	288	2763	1252
Среднее количество букв в слове (в тексте)	4,9122	5,1347	4,6101	4,7596	4,9105	4,8016	4,8466	4,9209
Среднее количество букв в слове (в словаре)	6,4171	7,1496	5,9992	6,8616	6,8664	6,6879	7,1755	7,0328
Среднее количество слов в предложении	8,981818	11,74675	14,36918	14,00822	15,66142	13,37153	13,95186	10,50719
Относительная частота однократных слов	0,3712	0,3285	0,1988	0,2961	0,3396	0,329	0,2134	0,2357

доверять, а данные по средней длине украинской словоформы из [1] – 5,156, выходят за пределы интервала, а значит, нуждаются в дополнительной проверке. Из данных таблиц видно, что среднее слово поэтических произведений одного автора существенно короче его прозаических произведений. Это видно как на произведениях А.С. Пушкина, так и на произведениях И.Я. Франко.

Таблица 2

Параметры	О. Кобилянська (1)	Л. Українка (1)	Л. Українка (2)	І. Франко (1)	І. Франко (2)	І. Франко (3)	М. Коцюбинс. (1)	Ю. Покальчук (1)
Всего букв в словаре	73411	20584	9329	19320	121013	90569	38171	47559
Всего слов	54630	9002	3409	7554	77365	49832	15855	26530
Всего уникальных слов	9919	3267	1544	3145	16171	12321	5802	6647
Всего букв в тексте	251311	37719	15534	33667	366143	247696	75298	119077
Максимальное количество букв в слове	21	22	17	20	26	21	14	39
Всего предложений	7234	1740	326	675	6518	3805	1810	2414
Среднее количество букв в слове (в тексте)	4,6002	4,1901	4,5568	4,4568	4,7327	4,9706	4,7492	4,4884
Среднее количество букв в слове (в словаре)	7,401	6,3006	6,0421	6,1431	7,4833	7,3508	6,5789	7,155
Среднее количество слов в предложении	7,551839	5,173563	10,45706	11,19111	11,86944	13,09645	8,759669	10,99006
Относительная частота однократных слов	0,1071	0,2688	0,3256	0,3106	0,1348	0,1582	0,2516	0,1684

Текстометрические исследования научных текстов

В качестве научных текстов были выбраны тексты докладов, представленных на Международную конференцию «Новые информационные технологии в образовании для всех» (2006 – 2009 гг.), именуемые далее *ITEA*2006 [6], *ITEA* 2007 [7], *ITEA*2008 [8]и *ITEA*2009 [9]. Доклады подавались на эту конференцию на трех языках – украинском, русском или английском (на выбор автора). Из текстов на русском и украинском языках было сформировано восемь корпусов текстов – по четыре на каждый язык. В корпус не включались название доклада, сведения об авторах, аннотации и список литературы. Опыт использования материалов *ITEA* 2006, *ITEA*2007 и *ITEA*2008 описан в [5].

Подсчет проводился по тем же параметрам, что и для художественных текстов. Сводные таблицы базовых текстометрических показателей для научных текстов приведены в табл. 3 на русском языке, а в табл. 4 – на украинском.

Как видим, средняя длина слова как в художественных текстах, так и в научных в украинском языке меньше аналогичной в русском.

Как для художественных текстов, так и для научных подсчитывалась *Относительная частота однократных слов* с целью проверки

Таблица 3

Параметры	*ITEA*2006	*ITEA*2007	*ITEA*2008	*ITEA*2009
Всего букв в словаре	107022	54007	46422	86492
Всего слов	53256	18690	15039	35882
Всего уникальных слов	11092	5651	4912	9099
Всего букв в тексте	386878	139829	112974	260961
Максимальное количество букв в слове	31	30	31	31
Всего предложений	2973	1033	813	1871
Среднее количество букв в слове (в тексте)	7,2645	7,4815	7,5121	7,2728
Среднее количество букв в слове (в словаре)	9,6486	9,5571	9,4507	9,5057
Среднее количество слов в предложении	17,91322	18,09293	18,49815	19,17798
Относительная частота однократных слов	0,1109	0,1762	0,1954	0,146

Таблица 4

Параметры	*ITEA*2006	*ITEA*2007	*ITEA*2008	*ITEA*2009
Всего букв в словаре	100737	90411	70569	100037
Всего слов	53848	42551	34013	52784
Всего уникальных слов	10941	9849	7853	10992
Всего букв в тексте	372891	294013	235904	361376
Максимальное количество букв в слове	32	38	32	31
Всего предложений	2616	2078	1784	2762
Среднее количество букв в слове (в тексте)	6,9249	6,9097	6,9357	6,8463
Среднее количество букв в слове (в словаре)	9,2073	9,1797	8,9862	9,1009
Среднее количество слов в предложении	20,5841	20,4769	19,06558	19,11079
Относительная частота однократных слов	0,1083	0,1312	0,1262	0,1135

предположения из [10], что доля однократных лексем (встречающихся в тексте всего один раз) к общей массе лексем, тяготеет к золотому сечению (0,382). Как в наших предыдущих исследованиях [5], так и в настоящих, указанное предположение не находит подтверждения.

Фонетические различия

Интуитивно чувствуется, что частота употребления не только определенных слов, но букв в научных и художественных текстах не будут совпадать. Фоносемантические исследования показывают [11], что каждая звукобуква имеет не только определенный цвет, но может быть хорошей или плохой, большой или маленькой, нежной или грубой, сильной или слабой, светлой или темной и пр. Естественно, что тексты различных функциональных стилей должны иметь различную частоту употребления некоторых букв. В наших исследованиях не учитывается ударность/безударность и твердость/мягкость.

Была проведена сортировка (по убыванию) относительно произведения А.П. Чехова «Дама с собачкой» и построено распределение букв в произведениях русских писателей (рис. 1).

Как видим на табл. 1 объемы произведений Пушкина и Чехова, выбранных для анализа, существенно не отличаются, поэтому кривые практически совпадают. Что же касается произведений Гоголя, то они существенно больше объемом и в них заметно увеличение употребления букв *т*, *с*, *в*, *р* и *ы* в тексте «Тараса Бульбы», что с точки зрения фоносемантики говорит о характере этого произведения. Естественно, что звук осмысливается только в готовых словах, тем не менее, на показателях увеличения или

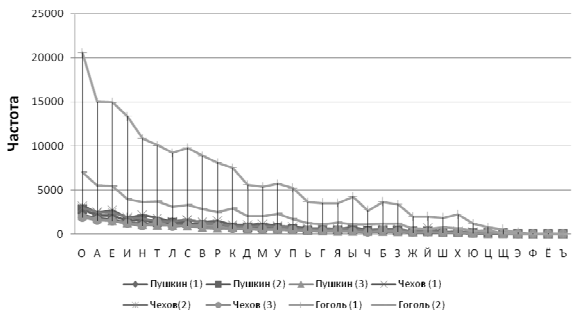


Рис. 1. Распределение букв в художественных текстах (русский язык)

уменьшения в тексте каких-то букв, может построить гипотезу о характере этого произведения. Для доказательства этого предположения определим относительную частоту встречаемости буквы «*л*» в исследуемых текстах. Так, в исследуемых произведениях эта частота будет следующей: А.С. Пушкин (1 – 0,04352; 2 – 0,05253; 3 – 0,04713), А.П. Чехов (1 – 0,05346; 2 – 0,05462; 3 – 0,05473), Н.В. Гоголь (1 – 0,04926; 2–0,04868). Как видим, в произведениях А.П. Чехова эта частота – самая высокая и практически остается стабильной для трех

рассматриваемых произведений, что придает этим произведениям особую мягкость и лиричность.

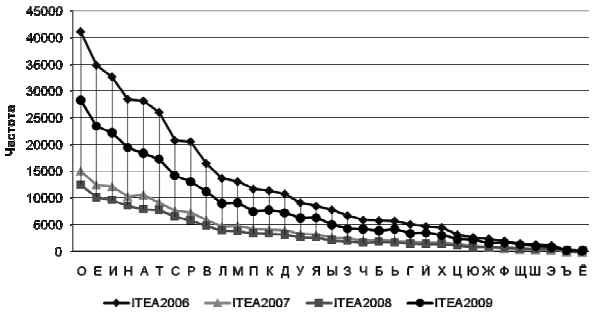


Рис. 2. Распределение букв в научных текстах (русский язык)

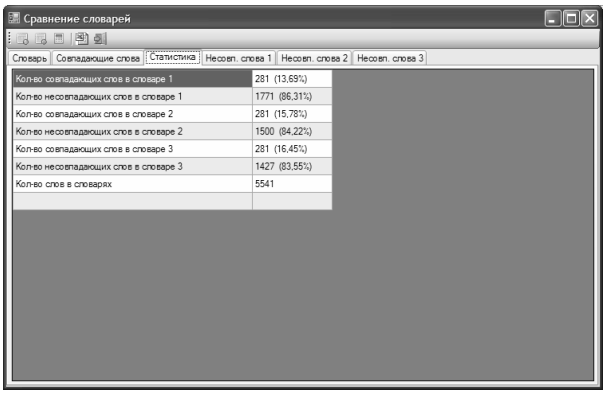
Поскольку средняя длина слова художественных произведений меньше, чем в научных текстах, то первые четыре позиции занимают гласные (О, А, Е, И), в научных же текстах на четвертой позиции уже появляется согласная Н, а порядок следования частот гласных тоже отличается. Так, частота употребления звука А (красного, активного, патетического) значительно ниже в научных текстах (четвертая или пятая позиция) по сравнению со второй либо третьей позицией в художественных текстах.

Сравнение словарей

В новой версии программы *Text Analyzer* добавлен модуль сравнения словарей. Для демонстрации остановимся только на сравнении словарей трех анализируемых произведений А.П. Чехова.

Для совпадающих слов указывается частота использования этого слова в каждом из текстов.

Как видим, фонетически (частота использования определенных букв) произведения А.П. Чехова практически идентичны, а процент совпадения лексики в трех произведениях этого автора незначителен и сдержит, в основном общеупотребительную лексику. Это характерно как для художественных текстов, так и для научных.



Заключение. В настоящем исследовании мы остановились лишь на нескольких аспектах использования программы *Text Analyzer*, возможности которой гораздо шире. Так, анализируя реверсивные словари, можно исследовать грамматические особенности тех или иных текстов или различных языков, а используя информацию по распределению слов определенной длины в тексте и словаре можно изучать особенности текстов тех или иных функциональных стилей различных языков.

Данные, полученные с программы *Text Analyzer*, подтверждают:

- средняя длина слова поэтических текстов обычно меньше средней длины прозаических текстов, независимо от языка;
- средняя длина слова научных текстов обычно больше средней длины художественных текстов, как в русском, так и в украинском языках;
- средняя длина слова как художественных, так и научных текстов на русском языке обычно больше аналогичных показателей на украинском;
- научные тексты отличаются от художественных как лексически, так и фонетически.

Естественно, эти выводы нуждаются в дальнейшей проверке с привлечением больших объемов текстов и расширением круга рассматриваемых языков.

1. *Пиотровский Р.Г.*, *Бектаев К.Б.*, *Пиотровская А.А.* Математическая лингвистика. Учеб. пособие для пед. ин-тов – М.: Высш. школа, 1977. – 383 с.
2. *Козачук М.В.* О лексическом богатстве русского и английского языков // НТИ. Сер. 2, Информ. процессы и системы – 2005. – № 3. – С. 28–31.
3. *Власенко Н.А.*, *Кузьминская Н.Л.*, *Максименко А.А.* Текстометрические исследования многоязычных научных текстов // УСиМ. – 2009. – № 2. – С.43–47.
4. *Бойков В.В.*, *Жукова Н.А.*, *Романова Л.А.* Распределение длины слов в русских, английских и немецких текстах. – http://tverlingua.by.ru/archive/ 001/01_ 1-006.htm
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. *Первая* междунар. конф. «Новые информационные технологии в образовании для всех», 29–31 мая 2006 г. Сб. тр. – К.: Академпериодика, 2006. – 530 с.
7. *Вторая* междунар. конф. «Новые информационные технологии в образовании для всех», 21–23 ноября 2007 г. Сб. тр. – К.: Академпериодика, 2007. – 458 с.
8. *Третья* междунар. конф. «Новые информационные технологии в образовании для всех», 1–3 окт. 2008 г. Сб. тр. – К.: Академпериодика, 2008. – 468 с.
9. *Четвертая* междунар. конф. «Новые информационные технологии в образовании для всех», 24–26 ноября 2009 г. Сб. тр. – К.: Академпериодика, 2009. – 568 с.
10. *Мартыненко Г.Я.* Золотое сечение в нумерологии текста. – http://www.trinitas.ru/rus/doc/0232/004a/ 02321035.htm
11. *Журавлев А.*П. Звук и смысл. – http://wmsafe.ru/buy.php?id=105486