ЛАВРЕНОВ А. Н.,
старший научный сотрудник,
кандидат физико-математических наук

Немного статистики о писателях

Когда впервые наука об управлении (кибернетика) заявила о себе, многие советские философы восприняли это научное направление в штыки. Термин лженаука – одно из немногих самых мягких поношений в область кибернетики. Нам вспомнилась эта ситуация перед написанием этой статьи, ибо мы также ожидаем элементы непонимания или даже яростного протеста в наш адрес. Но истина неумолима, нельзя отрицать то, что есть на самом деле. Поэтому мы решили познакомить наших уважаемых литераторов с результатами наших исследований в рамках нового научного направления компьютерной лингвистики.

Каждый автор одну и ту же мысль выражает по-своему. Это связано прежде всего с его восприятием окружающей действительности, настроением, психическим состоянием, знанием языка отображения, индивидуальным словарным запасом и любимыми подмножествами синонимов или антонимов слов языка, наконец, тем мало объяснимым аппаратом донесения мысли, который мы называем талантом, если этот эффект выражения мысли поражает большинство из нас.

Кирпичиками любых слов является алфавит языка выражения мысли. Если взять достаточно большое число описанных мыслей на определенном языке, то количество используемых букв-кирпичиков на очень большом объеме и приведенное к объему будет статистически вполне определенным и постоянным. Для корректного математического выражения данного факта введем понятие о нормированной или просто частотной характеристике буквы, например, в какой-то книге как частоту появления буквы в ней, деленной на длину или объем данной книги. Для полной картины нам необходимо учитывать совокупность частотных характеристик используемых букв (в пределе всего алфавита). Поэтому определим функцию распределения данной книги через произведение частотных характеристик используемых букв в ней. Эта функция отобразит нам конкретное статистическое постоянство применительно к данной книге. Специально подчеркнем, что аналогично мы можем поступить при изучении речи конкретного человека или других характеристик.

Также отметим, что для углубления рассмотрения может понадобиться использовать в качестве кирпичика нашего анализа не букву, а комбинацию букв (лексему). В дальнейшем одновременно будут рассматриваться два случая, а именно – когда анализ будет сделан для односимвольной и двухсимвольной лексемы.

Все вышесказанное позволяет говорить о закономерности таких оценок и может служить мерой данного языка или языка конкретного человека, использующего язык как таковой. Близость оценок языка конкретной пишущей личности к оценкам языка как такового может говорить о полноте его знаний и ис-

пользования этих знаний в его произведениях. Более того, если известны такие характеристики для конкретно пишущей личности, то можно идентифицировать произведение неизвестного автора в пользу той или иной кандидатуры с достаточно высокой степенью вероятности.

Понятно, что обработку такого огромного материала, как литературные произведения, можно сделать только на современной компьютерной технике. Мы воспользовались компьютерной библиотекой Мошкова, которая насчитывает достаточно много авторов.

О богатстве русского языка и близости к нему

Вначале кажется естественным поставить вопрос о русском языке как таковом вообще с описанной точки зрения. Определим нормированную функцию распределения русского языка через среднюю величину от нормированных функций распределения для выбранных (желательно бы, конечно, от всех существующих) произведений. Она является объективной, то есть не зависящей от субъективных влияний кого-либо, и математически корректной величиной. Ее явный вид приведен на рисунке. В нем вполне однозначно заключено словарное богатство русского языка, его многообразие для выражения различных ощущений и мыслей человека.

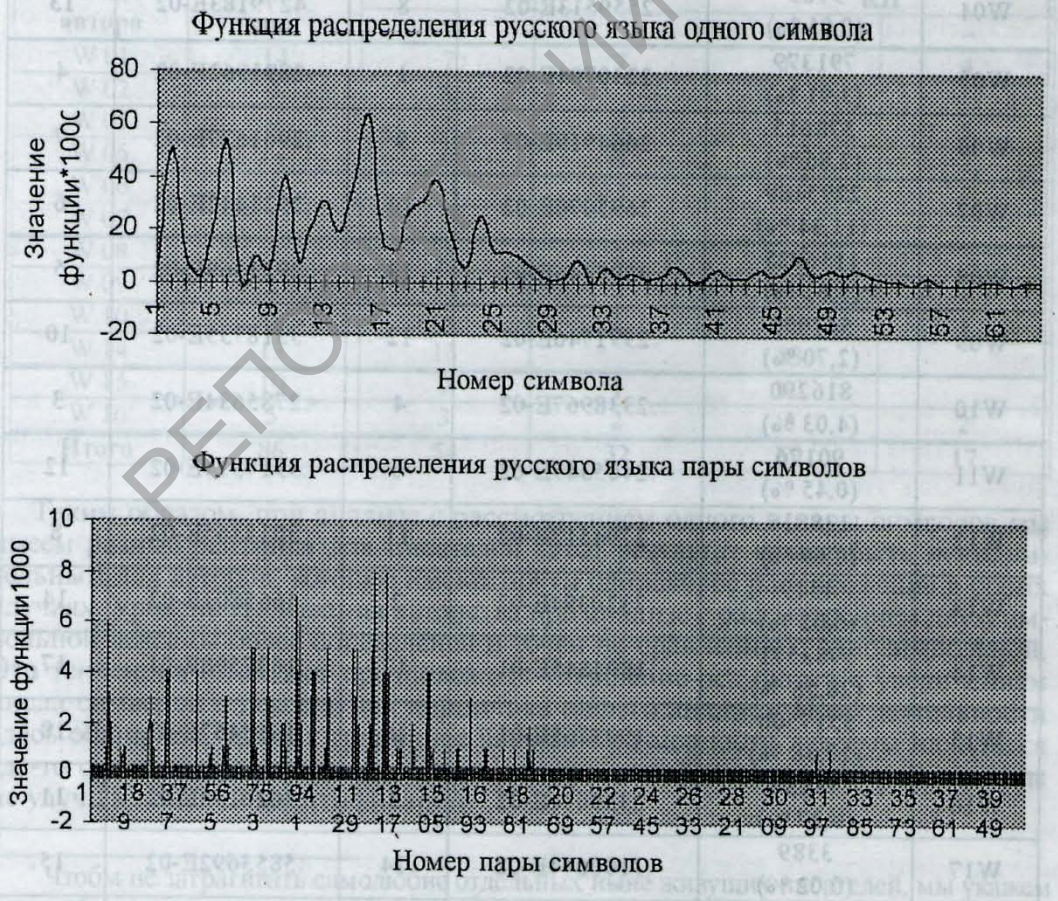

Рис. 1. Функция распределения русского языка

Однако, как мы знаем, в мире есть хорошее и плохое, яркое и посредственное. Другими словами, люди почему-то выделяют одного или нескольких авторов, предпочитая их другим. Они обычно в таких случаях говорят: "Этот автор пишет как англичанин (немец и т. п.) – сухо и скупо. А ты читал автора X? Какие у него приятные пассажи". Как выделить объективность в таких суждениях? Сразу оговоримся, что мы не навязываем свои какие-либо литературные вкусы в данном вопросе. Мы ищем и находим ту единственную объективную, существующую в независящим от нас окружающем мире реалию. В качестве возможного яблока раздора, а может, и дара свыше от природы, мы приводим в некотором смысле рейтинг или упорядоченность по близости определенного автора к богатству выражения ощущений и чувств посредством русского языка в виде табл. 1.

Таблица 1.

Рейтинг писателей по величинам QI и QII *

Автор	Объем	QI	Рейтинг	QII	Рейтинг
W01	2108714 (10,42 %)	.2538994E-02	7	.2613889E-02	2
W02	579859 (2,87 %)	.3283014E-02	15	.3299362E-02	9
W03	1778017 (8,79 %)	.2344059E-02	5	.2578495E-02	1
W04	9763 (0,05 %)	.2559513E-02	8	.4279183E-02	13
W05	791379 (3,91 %)	.2230704E-02	1	.2891962E-02	4
W06	578615 (2,86 %)	.2664919E-02	9	.2961677E-02	7
W07	2295189 (11,34 %)	.2695070E-02	10	.2921442E-02	6
W08	1136925 (5,62 %)	.3095346E-02	13	.2911369E-02	5
W09	547058 (2,70 %)	.2991746E-02	12	.3318735E-02	10
W10	816290 (4,03 %)	.2338967E-02	4	.2785634E-02	3
W11	90176 (0,45 %)	.2405347E-02	6	.3917314E-02	12
W12	138816 (0,69 %)	.2794512E-02	11	.3015365E-02	8
W13	4549 (0,02 %)	.2320394E-02	2	.4665801E-02	14
W14	5783571 (28,58 %)	.1894845E-01	17	.1607598E-01	17
W15	3517500 (17,38 %)	.1903765E-01	18	.1618467E-01	18
W16	53886 (0,27 %)	.2333634E-02	3	.3626133E-02	11
W17	3389 (0,02 %)	.3100644E-02	14	.5853692E-02	15
W18	1002 (0,00 %)	.3287372E-02	16	.7933305E-02	16
Итого	**20234698**				

В качестве такой меры близости мы приводим значение величины **QI (QII)**, которую определяем как усредненное значение суммы модулей отклонений по всем (парам символов) символам русского алфавита функции распределения конкретного автора от функции распределения русского языка. Функция распределения конкретного автора есть средняя величина от функций распределений для выбранных произведений данного автора.

Кто сказал "мяу"?

Следующей задачей, требующей пристального рассмотрения, является установление авторства неизвестного писателя. Мы ставим небольшой эксперимент: из имеющейся совокупности произведений мы должны угадать настоящего автора, исключив авторов, имеющих одно произведение (W 04, W 11, W 12, W 13, W 17, W 18). Результаты приведены в табл. 2. В ней знак +(–) означает (не) угаданное авторство произведения.

Самый простой критерий установления авторства следующий: если сумма квадратов отклонений значений функции распределения неизвестного произведения от функции распределения конкретного автора минимальна, то он получает с нашей точки зрения в свой литературный багаж это произведение.

Таблица 2.

Идентификация авторов по функции распределения односимвольной RI и двухсимвольной RII лексемы

Код автора	Всего книг	RI		RII	
		+	–	+	–
W 01	14	7	7	10	4
W 02	3	3		3	
W 03	9	3	6	4	5
W 05	14	6	8	12	2
W 06	3		3	1	2
W 07	3	3		3	
W 08	4	2	2	4	
W 09	2		2	1	1
W 10	6	5	1	5	1
W 14	13	13		13	
W 15	10	9	1	10	
W 16	5	3	2	3	2
Итого	**86**	**54**	**32**	**69**	**17**

Таким образом, при анализе с рассмотрением одного и пары символов мы имеем разные рейтинги для писателей. Этот результат может быть неутешительным для авторов, которые имеют почти одинаковое расположение в обоих случаях. Уровень угадывания авторства при анализе с рассмотрением двухсимвольной лексемы более определенен и точен по сравнению с односимвольной. Эта тенденция позволяет предположить улучшение результата с увеличением числа символов в лексеме как кирпичика нашего анализа. Поле деятельности здесь обширное, так как граничное значение размерности лексемы находится где-то около цифры тридцать. Отметим, что случай трехсимвольной лексемы не улучшил оценки идентификации.

---

* Чтобы не затрагивать самолюбие отдельных ныне живущих писателей, мы укажем только фамилии первых трех:

QI – Довлатов, Лермонтов, Пушкин;

QII – Булгаков, Аксенов, Коваль.