STUDIA PHILOLOGICA

Памяти
Владислава Митрофановича
Андрющенко

РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИНСТИТУТ РУССКОГО ЯЗЫКА им. В. В. ВИНОГРАДОВА

А. Я. ШАЙКЕВИЧ

СРАВНЕНИЕ ЧАСТОТНЫХ СЛОВАРЕЙ

Издательский Дом ЯСК
Москва 2026

УДК 811.161.1
ББК 81.2 Рус
Ш 17

Р е ц е н з е н т ы:
д. ф. н. С. А. Крылов,
д. ф. н. Л. Л. Шестакова

Утверждено к печати ученым советом
Института русского языка им. В. В. Виноградова РАН

Шайкевич А. Я.

Ш 17 Сравнение частотных словарей. — М.: Издательский Дом ЯСК, 2026. — 568 с. — (Studia philologica.)

ISBN 978-5-907498-92-1

Две формулы сравнения тестируются на широком материале самых разных частотных словарей. В первой части объектом сравнения служат опубликованные словари словацкого, английского, французского, русского языков и специально созданного для этой цели словаря на базе электронного корпуса китайского языка. Во второй части сравниваются жанры в авторском корпусе (Шекспир, Байрон, Пушкин, Достоевский, Чехов, Андрей Белый). В третьей части сравниваются разные подкорпусы прозы разного объема — от основных жанров до частей отдельных текстов. Приложение книги — Частотный словарь языка Пушкина. В электронную версию тома включены большие таблицы и Частотный словарь к текстам Байрона.

Two statistics are tested on a variety of frequency dictionaries, both published and specially constructed for the purpose. Part 1 deals with Slovak, American English, Chinese, French and Russian material. In Part 2 genres in an author's corpus are compared (Shakespeare, Byron, Pushkin, Dostoevski, Chehov and Andrey Belyy). Part 3 is a comparison of prose subcorpora of different dimension — from main genres to parts of a single text. The supplement to the book is A Frequency dictionary of Pushkin's language. The electronic version includes greater tables and A Frequency dictionary to Byron's works.

УДК 811.161.1
ББК 81.2 Рус

Дополнительные материалы к данной книге можно скачать по адресу http://lrc-press.ru/?m=159

ISBN 978-5-907498-92-1
9 785907 498921 >

Введение

Частотные словари, начиная со словаря Кединга[1], создавались обычно с какой-то практической целью. Хороший обзор частотных словарей мы найдем в книге Мистрика[2]. Сразу же определилась и форма представления данных в подобных словарях: частота слова (**f**), подсчитанная на каком-то собрании текстов, и порядковый номер слова (его ранг — **r**) в списке слов, упорядоченных по убыванию частоты. Первым, кто обратил внимание на математическое соотношение f и r в частотном словаре, был Zipf (в русской транслитерации — Ципф)[3]. Его вывод f·r = Const получил название «Закон Ципфа».

С течением времени кристаллизовались два способа общего представления материала. Список слов, упорядоченных по убыванию частоты, иначе — ранговый словарь. Именно ранговый словарь привлекал внимание математиков, видевших в нем новый и малоисследованный тип статистических распределений. У некоторых лингвистов понятия частотного словаря и рангового словаря совпадают. Но у рангового словаря есть один очень важный недостаток — в нем читатель не сможет найти нужное слово.

К счастью, остается второй способ — алфавитное упорядочение слов, где при каждом слове дается его частота. Во всех известных нам частотных словарях такой способ применяется. Так, в словаре Мистрика ранговый словарь (frekvenčný slovník) занимает 144 страницы и содержит 10 000 слов, а алфавитный словарь (abecedný slovník) на 383 страницах дает 21 283 слова с полной информацией по пяти подкорпусам. В нашем словаре Достоевского (далее — ССЯД)[4] Таблица 1 «Распределение лексем по основным жанрам» занимает 478 страниц и включает 43 577 лемм.

Таблица 3 дает верхушку рангового словаря (100 слов), вот как выглядят 40 самых частых слов этой таблицы:

r	f		r	f		r	f		r	f	
1	133761	и	11	33368	с	21	15733	ты	31	11750	один
2	83046	я	12	30160	вы	22	14724	мой	32	11105	только
3	73352	в	13	25704	а	23	14693	так	33	10856	еще
4	67165	не	14	24558	но	24	14336	что (мест.)	34	10603	знать
5	51828	он	15	23739	как	25	14087	они	35	10465	мочь
6	48107	что (союз)	16	23433	она	26	13810	о	36	10415	даже
7	44456	этот	17	20370	же	27	13568	у	37	10298	который
8	39233	быть	18	20065	тот	28	13504	свой	38	10242	себя
9	34852	весь	19	16853	к	29	12527	мы	39	10052	из
10	33992	на	20	16338	бы	30	12513	за	40	9913	от

Такие же ранговые словарики даются для каждого жанра: 1) художественных произведений, 2) критики и публицистики, 3) писем.

Наконец, одна страница отведена частотному спектру следующего вида:

r	f	ZZf		r	f	ZZf	
1	133761	133761	0,0470	500	607	1950209	0,6855
2	83046	217043	0,0762	1000	303	2161597	0,7598
5	51828	409002	0,1437	2000	143	2367811	0,8325
10	33992	603469	0,2121	5000	44	2603727	0,9151
20	16338	834820	0,2937	10000	15	2728488	0,9592
50	8900	1178623	0,4113	20000	4	2805533	0,9864
100	3378	1422612	0,4980	43577	1	2844630	1,0000
200	1732	1653649	0,5813				

В такой таблице уже нет места для конкретных слов языка.

Компьютер позволяет производить любые преобразования исходных частот, в том числе и такие, которые стали бы основой рангового словаря взамен частоты. В 1960-х гг. популярность приобрел способ, предложенный А. Жюйяном. В его частотном словаре румынского языка[5] наряду с частотой (frequency) вводится

---

[1] *Kaeding F. W.* Häufigkeitswörterbuch der deutschen Sprache. B., 1898.

[2] *Mistrík J.* Frekvencia slov v Slovenčine. Bratislava, 1969.

[3] *Zipf G. K.* The Psycho-Biology of Language. L., 1936.

[4] *Шайкевич А. Я., Андрющенко В. М.. Ребецкая Н. А*. Статистический словарь языка Достоевского. М., 2003.

[5] *Juilland A.* Frequency dictionary of Rumanian words. Hague, 1965.

показатель *usage*, учитывающий межжанровые расхождения. В словаре даются два ранговых списка для обоих показателей. В самом начале списка у некоторых слов ранги могут точно совпадать, но затем различия нарастают:

		r	frequency	usage			r	frequency	r	usage
de		1	19748	17938	tu	ты	21	3058	33	1298
un		6	10673	8395	voi	вы	84	535	99	363
pe	на	12	5574	5015	ochi	глаз	85	534	107	338
ce	что	19	3195	2489	vorbi	говорить	100	455	128	290
cum	как	40	1196	927	cap	голова	120	396	151	252
zi	день	60	752	616	dumneata	Вы	122	394	329	109

В частотных словарях 1960-х гг. неизменно даются не только итоговые частоты слов, но и частоты в подкорпусах (жанрах), а иногда и число выборок (текстов), в которых встретилось слово. Примером могут служить следующие строки из словацкого словаря Мистрика:

		ВСЕГО	драма	проза	поэзия	пресса	наука
ako	как	8150-5-59	865- 9	3107-14	1168-12	823-9	2187-15
cesta	дорога	785-5-55	57-10	294-14	152-10	155-8	127-13
hlava	голова	939-5-47	101- 9	570-14	216-13	22-5	20- 6
prax	практика	163-5-26	3- 3	5- 3	4- 3	49-7	102-10
sekretár		4-3- 4	1- 1			1-1	2- 2

Таким образом, общий словацкий частотный словарь включает пять (жанровых) частотных словарей. Именно частотные словари подкорпусов и будут основными объектами сравнения в данной книге. При этом ставится задача сравнивать частотные словари не как кривые частотных спектров, а как собрания конкретных слов языка. Этой задаче соответствует и главный лозунг настоящего исследования:

ОТ ЧАСТОТНОГО СЛОВАРЯ К СЛОВУ

Этот лозунг в начале книги звучит загадочно, его смысл будет раскрываться по ходу изложения.

Начнем с термина СЛОВО.

Будем называть последовательность знаков алфавита между двумя пробелами (или знаками препинания) ГРАФИЧЕСКИМ СЛОВОМ. Добавив к знакам латиницы знаки **’** и **-**, мы получим 32 графических слова в следующем отрывке из «Мадам Бовари».

Le Proviseur nous fit signe de nous rasseoir; puis, se tournant vers le maître d’études: — Monsieur Roger, lui dit-il à demi-voix, voici un élève que je vous recommande, il entre en cinquième.

Программными средствами можно легко и просто получить статистику графических слов, но лингвисту обычно нужно большее — графические слова должны быть сведены в осмысленные лингвистические единицы. Графическое слово d’études будет разделено на d’ и études, dit-il распадется на dit и il, слово же demi-voix останется неразделенным. Далее последует процесс лемматизации: de объединится с d’, études сведется к étude, tournant — к tourner (а может быть — к se tourner). Последний случай особенно интересен. Отдельной строкой в частотном словаре может стать и словосочетание. В ССЯД находим: ДРУГ ДРУГА f = 608, ДО СВИДАНИЯ f = 383 (при СВИДАНИЕ f = 288), ДО СИХ ПОР f = 1030 (при СЕЙ f = 980 и ПОРА f = 317), ПО КРАЙНЕЙ МЕРЕ f = 1219 (при МЕРА f = 343 и КРАЙНИЙ f = 219), МИЛОСТИВЫЙ ГОСУДАРЬ = 372 (при ГОСУДАРЬ = 92 и МИЛОСТИВЫЙ = 39). Подобные сочетания слов должны выявляться на этапе постредактирования исходного текста, что, конечно, увеличивает трудозатраты. Впрочем, обнаружить такие словосочетания можно и программными средствами.

Сложнее избежать ручного постредактирования при попытке разделить разные значения слова, ср. примеры из словаря Жюйяна: lume f = 344 «мир, свет» и lume f = 103 «народ», lucru f = 320 «вещь» и lucru f = 85 «работа», cap f = 396 «голова» и cap = 20 «руководитель», corp f = 93 «тело» и corp f = 26 «корпус». В ССЯД такие примеры исчисляются десятками: БАТЮШКА f = 151, БАТЮШКА (обращение) f = 255 и БАТЮШКА (священник) f = 54; БРАК f = 266 и БРАК (дефект) f = 2; ВОЛЯ f = 391 и ВОЛЯ (свобода) f = 53; ГЛАВА (книги) f = 512, ГЛАВА (руководитель) f = 125 и ГЛАВА (голова) f = 7; ГОЛОВА f = 2136 и ГОЛОВА (адм.) f = 6; ГУБКА f = 60 и ГУБКА (туалетн.) f = 2; ДВОР f = 315 и ДВОР (монарха) f = 15; ДОЛГ f = 243 и ДОЛГ (фин.) f = 418; ДОЛЖЕН f = 1754 и ДОЛЖЕН (деньги) f = 70; ДУМА f = 46 и ДУМА (учреждение) f = 32; ДУХ f = 876, ДУХ (сверхъестеств.) f = 32 и ДУХ (запах) f = 61; ДУША f = 1726 и ДУША (крепостная) f = 51.

Изредка в частотных словарях даются частоты словоформ слова, таков словарь Жюйяна, где, например, у слова glas «голос» указывается общая частота леммы (f = 26) и частоты составляющих ее словоформ (glas, glasul, glasului, glasuri, glasurile). При этом, естественно, растет объем печатного словаря. На русском материале известен лишь один частотный словарь такого типа, это интересный словарь Штейнфельдт, предназначенный для школьников Эстонии[1]. При изучении русского языка как иностранного такая информация особенно важна. Интересен этот словарь и для лингвиста, ведь здесь через статистику падежей становятся очевидными синтаксические предпочтения имен.

	И	Р	Д	В	Т	П
ребята, друг, начальник, директор, мама	64	14	10	5	6	1
товарищ, мальчик, девушка, девочка, мать, брат, женщина, парень, ученик, председатель, комсомолец, отец, секретарь	62	19	6	7	5	1
рука, глаза, нога, плечо, нос, спина	16	8	2	35	26	13
лета, час, минута, пора, месяц, утро, метр	12	67	2	17	1	2
вода, хлеб, молоко, металл	18	40	4	21	12	5
народ, совет, молодежь, партия, комсомол, армия, союз, организация, республика	18	60	4	8	5	5
школа, дом, класс, страна, завод, мир, район, колхоз, цех, фабрика, комбинат	14	40	4	16	2	24
место, комната, квартира, пол, кухня, кабинет	14	16	4	35	2	29
поле, гора, зал, клуб, площадь, село, этаж, лагерь, сцена, столовая, степь, остров, больница, дворец, ферма, стадион	17	20	5	22	5	31

В выборе той или иной стратегии лемматизации составитель частотного словаря руководствуется техническими соображениями (трудозатратами при обработке исходного текста), общепринятыми грамматическими взглядами (особенно — представлением о частях речи), наконец, главной целью создаваемого словаря.

Ясно, что при сравнении ранговых словарей разных языков принцип лемматизации будет иметь очень большое значение. Вот каковы списки десяти самых частых слов в трех европейских языках с артиклями:

r	английский[2]	французский[3]	немецкий[4]
1	the	le, la, l', les	die
2	and	de, du, del', des	der
3	of	à, au, aux	und
4	a, an	être	in
5	to	et	zu
6	be	un, une	den
7	that	je	das
8	have	avoir	nicht
9	in	il, ils	von
10	it	que	sie

В настоящем исследовании принятие единой стратегии лемматизации подорвало бы весь проект в самом начале. В главе первой принималась лемматизация источника. В других главах в русском материале, как правило, разделялись части речи, но некоторые словоформы становились самостоятельными леммами, ср.: ГОТОВ(-а,-ы), ПРАВ(-а,-ы), ПОЛОЖИМ, ПРИДЕТСЯ (пришлось), КАЖЕТСЯ (казалось), ПРОЩАЙ(те), ПОМИЛУЙ(те), СТУПАЙ. Принципиальное значение придавалось выделению в качестве особых лемм форм глагола БЫТЬ: БЫЛ(-а,-и,-о), БУДЕТ(-у,-ешь,-ем,-ете,-ут) и ЕСТЬ.

В английском материале иногда разводились омонимы, ср.: art «искусство» и «еси» (арх.), ball «бал» и «мяч», bank «берег» и «банк», bark «барк» и «лаять», bear «нести» и «медведь». Многие грамматико-лексические омонимы оставались неразведенными, да и нужды в их различении не было, ср.: attack, call, care, dance, force, help, hope, jest, kiss, love и т. п.

---

[1] *Штейнфельдт Э. А.* Частотный словарь современного русского литературного языка. Таллин, 1963.

[2] *Thorndike E. I., Lorge I.* The Teacher's Word Book of 30 000 Words. N. Y., 1944; *West M.* General Service List of English Words. N. Y., 1953.

[3] *Vander Beke G. E.* French Word Book. N. Y., 1929.

[4] *Kaeding F. W., Meier H.* Deutsche Sprachstatistik. Hildesheim, 1964.

Настоящее исследование проводится в ту эпоху, когда стали доступны, с одной стороны, громадные массивы текстов в электронной форме, а с другой — современный персональный компьютер, позволяющий обрабатывать эти тексты, с тем чтобы в результате возникал частотный словарь. Лингвостатистика в прошлом часто сводилась к усложнению применяемых формул с проверкой их на ограниченном материале. Свою задачу автор видит в проверке двух очень простых формул на широком материале существующих и создаваемых частотных словарей. В ходе работы он надеется установить применимость этих формул и одновременно выявить те ограничения, которые неизбежны в любом статистическом исследовании. Таким образом, данный проект относится к сфере эмпирической лингвистики[1].

[1] Заимствую этот термин у Дж. Сэмпсона (*Sampson G.* Empirical Linguistics. L.; N. Y., 2001), чья книга направлена против хомскианской лингвистики, но сам термин может найти более широкое применение.

Часть первая

МЕРЫ СРАВНЕНИЯ ЧАСТОТНЫХ СЛОВАРЕЙ

В этой части предлагаются две формулы, которые затем будут применяться на протяжении всего исследования. Материалом для проверки послужат пять частотных словарей, четыре из которых были опубликованы в печатном виде, а пятый получен из исходного корпуса текстов.

В качестве первого примера будет использован уже упоминавшийся словарь Мистрика. Объем этого словаря — N = 1 000 000 словоупотреблений (occur ences), в том числе пять подкорпусов (жанров):

a) dialógy — драма (1922–1963) 10 текстов, N = 105 266, или 10,5 % общего корпуса текстов;
b) umelecka próza — (1961–1966) 14 текстов, N = 301 674 (30,2 %);
c) poézia — (1932–1965) 12 текстов, N = 132 222 (13,2 %);
d) žurnalistika — (1964–1966) 9 текстов, N = 145 827 (14,6 %);
e) náučna próza — (1947–1966) 15 текстов (в том числе 6 учебников), N = 315 011 (31,5 %).

Частотные словари подкорпусов и будут сравниваться друг с другом.

1.1. Лексическое сходство частотных словарей

Попытку сравнения двух частотных словарей и измерения их сходства в свое время предпринял М. В. Арапов[1]. Работая в рамках ранговой лингвостатистики, он построил очень сложный математический аппарат получения результирующей величины «бета» — мерой сходства двух частотных словарей.

Следующая матрица дает величину «бета» для четырех жанров русского частотного словаря[2].

	Проза	Драма	Наука	Газета
Проза	-	0,68	0,48	0,51
Драма		-	0,48	0,52
Наука			-	0,68
Газета				-

В данной книге будет использоваться очень простая формула лексического сходства (lexical similarity)[3].

$$LS = \Sigma \min \{px_i, py_i\} \qquad [1].$$

Этот показатель приобретает значение 1 при сравнении частотных словарей одного и того же текста; он близок к 0 при сравнении словарей, использующих разную графику.

Само собой разумеется, при работе с частотными словарями вместо вероятностей используются относительные частоты.

Поскольку лексическое сходство частотных словарей может сильно различаться в зависимости от частоты слов, LS будет подсчитываться отдельно по трем ранговым зонам общего корпуса словаря: первая зона (r 1–10), вторая зона (r 11–100) и третья зона (r 101–1000).

Таблица 1.1.1
Самые частые слова словаря Мистрика

a, byť, ja, na, on/a/o/i/y, sa, ten, v, z, že

aby, aj, ak, ako, aký, ale, alebo, ani, by, čas, či, číslo, človek, čo, dať, do, dobre, dobrý, druhý, dva, ešte, chcieť, iba, iný, isť, jeden, jeho/jej, k, kde, keď, kto, ktorý, lebo, len, mať, medzi, môcť, môj, musieť, my, náš, nič, nie, no, o, od, oko, pod, potom, povedať, pre, preto, pri, prísť, prvý, rok, ruka, sám, slovenský, starý, strana, svet, svoj, tak, taký, tam, teda, tento, teraz, tu, ty, už, vec, veľký, vidieť, voda, však, všetok, vy, za, život

---

[1] *Арапов М. В.* Квантитативная лингвистика. М., 1988. С. 61–80.

[2] Частотный словарь русского языка / Ред. Л. Н. Засорина. М., 1977. М. В. Арапов использовал этот словарь, хотя он не очень надежен — в буквах А, Б и частично В последовательно идет двойной счет. Меня не оставляет чувство стыда: я заметил этот дефект только после публикации тома, а не тогда, когда знакомился с рукописью.

[3] Впервые доложено в: *Шайкевич А. Я.* Меры лексического сходства частотных словарей // Труды международной конференции «Корпусная лингвистика — 2015». СПб., 2015. С. 434–442.

Значения LS подкорпусов того или иного частотного словаря будут представлены в виде двухчастной матрицы. В верхней части (1/2) даются результаты по первой зоне (справа от главной диагонали) и по второй зоне (слева от нее). В нижней части (3/1–3) даются результаты по третьей зоне (справа от главной диагонали) и по совокупности всех трех зон (слева от главной диагонали).

Таблица 1.1.2
Лексическое сходство словацких жанров

a) драма, b) проза, c) поэзия, d) журналистика, e) наука

	a	b	c	d	e	1/2
a		179	137	130	136	
b	222		150	144	155	
c	177	180		138	136	
d	155	156	139		150	
e	152	152	134	169		

	a	b	c	d	e	3/1-3
a		165	130	115	100	
b	566		165	113	100	
c	444	495		101	87	
d	400	413	378		173	
e	389	407	357	492		

В первой зоне (среднее LS = 146) жанры расположились следующим образом:

Рис. 1

Во второй зоне (среднее LS = 164) жанры расположились следующим образом:

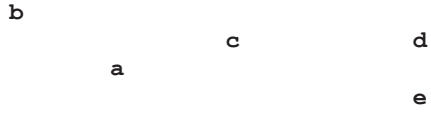

Рис. 2

Усилилось противопоставление художественных и прочих жанров. Картина стала еще ярче в третьей зоне (среднее LS = 125).

1.2. Лексические маркеры

Предложенная выше мера лексического сходства кажется очень естественной, ею можно воспользоваться при частотных словарях самого разного происхождения. Однако у этой меры есть один существенный недостаток (как, впрочем, и у любой другой попытки представить словарь в виде одной цифры). Недостаток этот — обезличенность. За единым показателем скрываются сотни и тысячи реальных различий.

В какой-то степени преодолеть этот недостаток могут другие меры. Рассмотрим подход, при котором сравниваемые частотные словари относятся к какому-то уже существующему (или специально конструируемому на этот случай) корпусу по крайней мере с двумя подкорпусами. Для каждого из подкорпусов можно статистическими средствами выделить лексические маркеры при помощи следующей формулы:

$$S = (f - m - 1)/\sqrt{m} \qquad [2],$$

где f — абсолютная частота слова, m — математическое ожидание этой частоты в рамках данной нулевой гипотезы.

Основную часть этой формулы я заимствовал из опередивших свой век работ Пьера Гиро[1]. Я добавил -1 в числителе, блокируя тем самым слова, появившиеся в тексте всего один раз.

За этой формулой стоит распределение Пуассона, где дисперсия равна средней[2]. При S = 2 мы ожидаем, что отклонение от средней может считаться значимым с вероятностью 0,95. На всем протяжении книги будет указываться значение S как принимаемого в данном конкретном случае порога. Во всех случаях S будет уменьшаться до целого числа[3].

Как назвать слово, преодолевшее заданный порог S? Французский термин Гиро mot-clef и английское соответствие keyword в русском переводе (ключевое слово) звучит очень непривычно для уха русского филолога, возбуждая почти философские ассоциации. Взамен я предлагаю нейтральный термин ЛЕКСИЧЕСКИЙ МАРКЕР.

Покажем применение формулы [2] на примере двух очень частых словацких слов.

		ВСЕГО	ДРАМА	ПРОЗА	ПОЭЗИЯ	ЖУРНАЛИСТИКА	НАУКА
		100%	10,5%	30,0%	13,2%	14,6%	31,5%
ja	f	8854	2342	4438	1753	188	137
	m		930	2656	1182	1296	2789
	f - m		1412	1782	571	-1108	-2652
	S		46	34	17		
ktorý	f	7758	290	1451	671	1565	3781
	m		815	2327	1050	1133	2441
	f - m		-525	-875	-379	432	1340
	S					13	30

Расстановка жанров, полученная формулой [1], подтверждается теперь лексчески: совершенно так же, как ja, приурочено к жанрам и môj 16:3:38, my 14:7:12, vy 37:9:6, čo «что» 30:12:7, okno 5:5:5, tam 5:10:7, tu «тут» 11:8:4, večer 5:6:3.

Аналогично ktorý ведут себя ak «если» 8:6, najmä «особенно» 7:13, o 13:8, okrem «кроме» 5:8, pre «для» 9:8, tento «этот» 7:23, vzťah «протяжение» 7:9; hlavný 8:7, nový 10:8, prvý 7:11;

výrobok «продукт» 5:7, výsledok «результат» 8:8, úloha «задача» 10:8, autor 8:7, filozofia 6:10, literatúra 10:7, spisovateľ «писатель» 5:7, oblasť 14:10, organizacia 8:8, podnik «предприятие» 14:6, člen 6:7;

а также «идеологические» маркеры:

národný 7:8, rozvoj «развитие» 9:8, socialistický 20:7, zjazd 6:10.

У драмы и прозы находим 46 общих маркеров:

on,-a,-i 9:53, pán 15:9,

ale «но» 13:30, ani «ни» 5:22, keby «если» 7:11, len «только» 8:11, prečo «почему» 18:10;

местоименные (в широком смысле) слова:
nič «ничего» 16:17, tak 18:9, ten «этот» 41:16, teraz «теперь» 12:13, všetok «всякий» 9:10;
глаголы:
byť 11:18, čakať «ждать» 6:7, chcieť «хотеть» 21:13, isť «идти» 20:12, myslieť «думать» 20:8, odisť «отойти» 5:8, povedať «сказать» 12:24, prísť «прийти» 12:13, robiť «делать» 7:9, rozumieť «понимать» 10:5, hádam «думаю», vedieť «знать» 25:17, vravierť «говорить» 10:13, vziať 7:9;
приметы повествования:
ešte «еще» 21:10, už 19:11;

[1] *Guiraud P*. Les caractères statistiques du vocabulaire, P. 1954; *Guiraud P.* Problèmes et méthodes de la statistique linguistique. Dordrecht, 1959. Недавно я обнаружил, что S = (f – m)/√m или Pearson residual есть составная часть широко известного критерия хи-квадрат, предложенного Пирсоном еще в 1900 г. (см.: *Upton G., Cook I.* A Dictionary of statistics. Oxford, 2002. P. 77).

[2] *Юл Дж. Э., Кендэл М. Дж.* Теория статистики. М., 1960. С. 231.

[3] В своем первом статистическом опыте (*Шайкевич А. Я.* Распределение слов в тексте и выделение семантических полей. «Иностранные языки в высшей школе». Вып. 2. М., 1963) я обращался к книге: Я. Янко (Математико-статистические таблицы. М., 1961. С. 81–83) и лишь позднее перешел к формуле [2]. Широчайшее применение последняя формула нашла в ССЯД.

приметы разговорной речи:
aha 5:5, počuť «послушай» 9:9, naozaj «на самом деле» 11:6, nuž 13:5, veď 21:7, veru «действительно» 7:7,
выразители эмоций:
rád 8:13, báť sa «бояться» 9:8, usmiať sa «улыбаться» 5:7,
родственники:
mať «мать» 16:11, otec 11:10, syn 5:10,
а также dobre 10:5, dobrý 8:6, doma 8:5, dvere 9:7, večer 5:6.

У драмы и поэзии всего 9 общих маркеров:

ach 8:16, dnes «сегодня» 8:7, kto 18:6, ľúbiť 7:8, tvoj 8:37, ty 51:31, umrieť 5:6, vaš 10:11, zas «опять».

Прозу и поэзию объединяют 32 маркера:

či «или», do 10:10, keď «когда» 9:15, nik «никто» 7:5, pod 5:13; cítiť «чувствовать» 10:5, hľiadeť 10:7, spať 6:11, spievať 5:11;

noc «ночь» 6:18, ohen 7:7, ráno «утро» 5:11, drahý 5:10, hlas 11:12, hrdlo «горло» 5:6, hora 5:11, slovo 5:11, strach 5:5, žena 8:8, vták «птица 5:12, záhrada «сад» 6:5, biely 8:8, čierný 5:5;
hlava 16:7, noha 12:7, oko 17:15, prst «палец» 8:9,
tvár «лицо» 14:13, ústa «рот» 9:15.

Только восемь маркеров пересекают границу, выявленную в 1.1 и отделяющую художественную литературу от прочих жанров. Это

предлог v 10:27:10 (поэзия - журналистика - наука), sloboda «свобода» (поэзия - журналистика), voda (поэзия - наука)

и **слова nejaký «какой-то», aby «чтобы», scena, týždeň «неделя» (драма - журналистика), môcť «мочь» (драма - наука).**

Наша формула выявила 1767 маркеров (при S * > 5) в том числе

Драма	**214**	**Журналистика**	**309**
Проза	**236**	**Научная проза**	**517**
Поэзия	**635**		

Полностью они представлены в Таблице I электронного издания. Ниже приводятся важнейшие примеры для каждого жанра.

ДРАМА

nie, no, ano да (част.), že ли, iba лишь, jednako, predsa все-таки, najsamprav прежде всего, vlastne именно, vždy ведь;
aký какой, kedy когда;
sem сюда, tamten вон тот, zajtra завтра;
niečo что-то, niekto кто-то, voľačo что-либо;
by, azda может быть, možno, musieť, pripadať придется, vari пожалуй;
okamih тотчас, poriadno правильно, pravda, rýchle быстро;
Ježuš, Kristus, boh, čert, preboha ради бога;
dakovať благодарить, dovoliť позволить, chvála, kvôli ради,
laskavo, odpustiť простить, prepáčiť извинить, prikývnuť, prisahať клясться,
prosiť, rozkaz приказ, uprimne искренне, vina, vitať приветствовать;
dať, chapať, mať иметь, ostať остаться, uvidieť, uznať, veriť, žiť;
behať, baviť веселиться, blazniť сходить с ума;
svedomie совесть, hanba стыд, nerád неохотно, ublížiť обидеть, vzrušiť взволновать;
človek, dievča, dievka, slečna барышня;
doktor, doktorka, lekár, plukovnik, tajomnik секретарь, veľvyslanec посол;
svadba, manželstvo брак, sobáš брак, oženiť sa;
súdruh товарищ, náhle неожиданно,
poznať познакомиться, pauza, kulisa,

ПРОЗА

a и, na, okolo, sa, von наружу, ako как, taký,
akosi как-то, akýsi какой-то, čosi, kdesi, ktosi;
dlho, dolu, odrazu вдруг, pomaly медленно, potom;
držať, hodiť бросить, chodiť, chýtiť, jesť есть, letieť, ležať, nosiť, pozrieť посмотреть,
prestať перестать, stať, vstať, sadnúť сесть, sedieť, ťahať, vidieť, začať,
zdať sa казаться; smiať sa,
deň, chvíľa минута, raz, zima;
ľudia, muž, mama, mater, otecko папа, babička, babka, brat, dietýa, chlap парень,
chlapec мальчик, kmotor кум, nevesta, sestra;
brada, chrbát спина, koleno, nos, plece, ucho, zub;
koleso, voz čiapka, žltý,
pekný хороший, zle плохо; um, hlupý, prazdny пустой;
kameň, dolina, kopec холм, potok;
dom, dvor, izba, stena, chalupa изба, hradska шоссе, chodnik тротуар;
učiteľ, šef, cigán, cudzinec;
zviera, vrana, kohút петух, kôň, krava, medved, ryba;
peniaze деньги, kúpiť, stratiť;
posteľ, stôl.

ПОЭЗИЯ[1]

k, z, jak как, cez через, čí чей, hoc хоть, kde, nad, svoj;
duch, anjel, chram, osud судьба, raj, slava, svätý, svet мир, zvon колокол, život;
krása, krásný, laska любовь, milý, milovať, nádhera красота, náruč объятие, objatie, neha, panna,
spanilý прелестный, útly нежный;
bežať, biť, hnáť, liať лить, niesť, piť, dýchať, znať;
bledý, čistý, chladný, bosý, holý, nahý, krivý, opustený одинокий, prostý, prudký резкий, pustý,
mocný, divý дикий, dravý буйный, hluchý, nemý;
svetlo, jas свет, blesk, lúč, hmla, mrak, stin тень, temný, tieň, tma;
zlaty, ružový, rudý красный, modrý синий, sivý седой, blankyt лазурь;
most, brana ворота, dlažba мостовая, hradba стена, strecha крыша, ulica;
pieseň, spev пение, hudba музыка, bubon барабан, struna, tanec, tón;
mdza межа, brázda, klas, pluh, socha, pole, studňa колодец, žito, džbán кувшин, hus,
chlieb, chyža хижина, svieca свеча;
rieka, prameň источик, breh, blato, dno, háj роща, hoľa луг, luka луг, jama, piesok,
prach пыль, priepasť, púšť пустыня;
strom дерево, kvet цветок, ruža роза, mak, breza, javorec клен, konár сук, jabloň, tráva;
sen сон, driemať, budiť;
smerť, mrtvy, hrob могила, rov могила, cintorin кладбище;
ó, duša, cit чувство, des ужас, hnev, hrôza ужас, plakať, slza, chvierť sa дрожать,
ľúto сожаление, nenávisť, pycha гордость, smútok грусть, stesk тоска, úzkosť страх, bolesť,
muka, žiaľ печаль;
štastie, smiech, snivať грезить, túha упование, útecha, veselie, víno;
čas время, kým пока, mladý, starý, vek, večný;
lice, čelo лоб, dlaň ладонь, kosť, ňadrá грудь (ж.), pera губа, srdce, prsia грудь, šija,
telo, vlas;
nebo, zem, slnce, hviezda звезда, luna, obloha небосвод, oblak, vôňa аромат, obzor горизонт,
zora заря;
dcera дочь, deva, detstvo, koliska колыбель;
sneh, dážď, dúha радуга, dym, hrom, vietor, vanok ветерок, vichor;
let полет, kŕdeľ стая, krídlo крыло, hniezdo, drozd, havran ворон, motýľ бабочка;
horieť, plameň, popol, požiar, teplý;
jar весна, jeseň осень, leto, máj;
král король, víla фея, žezlo; krv кровь; zrak взор;
more, ocean, pena, skala;

[1] Аномально большое число маркеров этого жанра вызывает некоторые подозрения: не вызвано ли оно какой-то ошибкой при составлении словаря? Не имея доступа к исходным текстам, не могу развеять эти сомнения. Косвенным аргументом в пользу таких сомнений служит исключительная частота (и высокое значение S) некоторых слов, ср.:

dyka f = 235 S = 41 кинжал, lež f = 243 S = 30 ложь, padať f = 174 S = 16, prameň f = 112 S = 21 источник,
rosa f = 65 S = 15 роса, sen f = 119 S = 40 сон, ruža f = 115 S = 20 роза, sladký f = 120 S = 15,
slava f = 111 S = 20, slza f = 137 S = 20, smrť f = 158 S = 15, smútok f = 111 S = 22 грусть,
srdce f = 469 S = 36, strom f = 113 S = 11 дерево, vlna f = 163 S = 23, zem f = 387 S = 22,
zrak f = 114 S = 17 взор.

nôž, dyka кинжал, meč, oceľ сталь, prapor знамя, rana, šabľa, šip стрела, štit щит, veža башня,
vrah, zakliatý, zbraň оружие;
šľahať стегать, hriva, cval галоп;
verš, báseň стих, básnik поэт, balada, muza, poezia, sonet.

ЖУРНАЛИСТИКА

aj также, medzi между, nemožno, predovšetkým прежде всего, takmer почти, totiž то есть;
dielo, generálny, praktický, problém, prostredie среда, situacia, úroveň;
americký, britski, československý, francuzský, nemecký, Práha, sovieyský, zahraničný;
mesto город;
kancelária, komisia, krajina, minister, odbor отдел, okres район, opatrenie мероприятие,
osobný личный, otázka вопрос, predseda, prezident, rada совет, republika, riešiť,
schódza собрание, služba, obec общество, správa управление, štatný, zástupca,
vedúci заведующий, vláda правительство, výbor комитет, zväz ассоциация;
brigáda, inžinier, kolektiv, materiál, stavebny строительный, stroj машина, technika;
budúci, cieľ, dnešný, etapa, iniciativa, návrh проект, plán, príprava подготовка, program;
ekonomický, finančný, koruna крона, majetok имущество, obchod торговля,
priemysel промышленность, výlkon выполнение, rolnický крестьянский, výroba производство;
rok год, hodina час, minuta, včera, apríl, september;
mladež, čestný, moralka,
chyba ошибка, diskusia, nedostatok,
tlač пресса, čitateľ, noviny, oznámió сообщить, verejný публичный;
škola, dekan, fakulta, skúška экзамен, študent;
kilometr, kus штука, miera, milion, percento;
divadlo театр, dráma, estetický, festival, film, hala зал, kino, herec актер, hra, hráč музыкант,
hudobný музыкальный, kritika, kultura, opera, próza, režisér, sezóna, tvorba творчество,
umelec художник, umenie искусство;
družstvo, ideológia, komunista, konferencia, ľudový, národ, oslobodenie, povstanie,
politika, práca труд, prejav выступление, spoločnosť, srdcovy сердечный, zápas борьба,
názor воззрение;
šport, štadión, futbal, Slavia, Sparta, súťaž состязание, štartovať, záver финал.

НАУЧНАЯ ПРОЗА

od, po, pri, u, však но, alebo, avšak однако, druhý, iný, jednak, kedže так как,
napriklad например, niektorý, preto поэтому, síce хотя, takže так что;

absolutný, bod точка, časopis журнал, časť, čislo, daný, definícia, dokázať, doklad,
držba обладание, empirický, figura, forma, funkcia, hľadisko точка зрения, hnutie движение,
jav явление, hodnota величина, charakter, idea, jestvovať существовать, konkretný, kvalita,
metoda, množstvo количество, objektivny, obrázok рисунок, obsah содержание,
osobiotný оригинальный, paragraf, pevný твердый, plocha поверхность, počet число,
pohýb движение, poloha положение, pisať, pokus попытка, potreba необходимость, povaha характер,
prax, použiť использовать, pozorovať наблюдать, pričina, proces, teória, rovnať сравнивать,
rozdiel разница, skúmať изучать, stred центр, smer направление, spôsob, sústava, systém,
štruktúra, tvar фигура, určiť определить, útvar форма, uvedený указанный, veda наука,
vyplývať следовать, zákon, zložka компонент;

centimetr, gram, kilogram, meranie, meter, rovina уровень, váha вес, rozsah размер, veličina;

fyzika, chémia, atóm, deliť, elektrón, element, elipsa, energija, interval, jadro,
jednotka единица, kladný положительный, kocka куб, konštanta, koreň, masa, matematika,
hmota материя, množina, sila, molekula, organický, rovnica уравнение, rychlosť скорость, soľ,
súčet сумма, súčin произведение, úhlík углерод, usadenina осадок, záporný отрицательный,
zlomok дробь, zloženie;

príroda, teleso тело, rastlina, bakteria, bunka клетка, poda почва;

cibuľa лук, cukor сахар, mlieko, maslo, múka, paprika, smotana, variť, zelenina овощи,
zemiak картошка;

nerast минерал, banský горный, kryštaľ, vápenatý известковый, vrstva слой, žula гранит;

abstrakcia, Aristoteles, bytie, dialektika, Engels, materializmus, idealizmus, kategoria,
kauzalita, logika, metodológia, myslenie, podstata сущность, pojem понятие, poznanie,
poznatok знание, skutočnosť действительность, substancia, výrok изречение, vývin развитие;

dejiny история, boj борьба, buržoázia, demokracia, epocha, šľacta, historický, kapitalizmus, kriza, Lenin, ľud, obdobie период, štrajk, parlament, postup прогресс, proletariát, revolúcia, robotník, Rusko, Uhorsko, vlastnosť собственность, zmluva договор.

При огромном числе лексических маркеров (S * > 5) находятся все же частые слова, не ставшие маркерами ни одного из пяти жанров:

	f		f		f
s	7005	vtedy	743	koniec	380
cely	1740	volať «звать»	448	kniha	328
každý	1446	vojna	444	červený «красный»	315
hovoriť	1170	jediny	381	dedina «деревня»	305
práve	802	ukazať	381	mnoho	214

Впрочем, если мы опустим порог до S = 3, даже предлог s станет маркером драмы. В этом случае в поэзии появятся маркеры bez f = 1047, les f = 247, mesiac f = 308, в научной прозе — dostať f = 658, skoro f = 400, veľmi f = 987. Cesta «путь» f = 785 будет маркером у прозы, поэзии и журналистики, než «чем» f = 347 у поэзии и науки, слово treba f = 860 — у драмы и науки.

1.3. Жанры в частотном словаре английского языка

В 1963–1964 гг. усилиями Н. Фрэнсиса и его сотрудников в Университете Брауна был создан электронный корпус современного американского английского языка. Тем самым было положено начало ныне процветающей традиции создания языковых корпусов. В том же университете к проекту присоединился Г. Кучера, через три года вышел в свет соответствующий частотный словарь[1]. Корпус составлен из 500 выборок (ср. 60 выборок в [Mistrik] при том же объеме) по 2000 словоупотреблений, распределенных по 15 жанрам (genres):

	N (тыс.)
A. Press: Reportage	88
B. Press: Editorial	54
C. Press: Reviews	34
D. Religion	34
E. Skills and Hobbies	72
F. Popular Lore	96
G. Belles Lettres, Biography, etc.	150
H. Miscellaneous	60
J. Learned and Scientific Writings	160
K. Fiction: General	58
L. Fiction: Mystery and Detective	48
M. Fiction: Science	12
N. Fiction: Adventure and Western	58
P. Fiction: Romance and Love Story	58
R. Humor	18

В KuFran даются частоты графических слов (лемматизация не проводится). Ранговый и алфавитный списки занимают по 133 страницы, при каждом слове даются три цифры: 1) частота, 2) число жанров, в которых встретилось слово, и 3) число выборок, в которых встретилось слово, ср.:

A	23237-15-500
ABEL	20-03-004
ABILITY	74-12-054
ABOUT	1815-15-426
ABSTRACT	34-09-020
ABSTRACTS	4-02-003
ACCORDION	1-01-001

[1] *Kučera H., Francis W. N.* Computational Analysis of present-day American English. Providence, 1967. Далее — KuFran.

Распределение частот по 15 перечисленным жанрам дается в особой таблице только для ста самых частых слов:

a, and, he, in, is, of, that, the, to, was;

about, after, all, also, an, any, are, as, at, back, be, been, before, but, by, can, could, did, do, down, even, first, for, from, had, has, have, her, him, his, I, if, into, it, its, like, made, man, many, may, me, more, most, much, must, my, new, no, not, now, on, one, only, or, other, our, out, over, said, she, so, some, such, than, their, them, then, there, these, they, this, through, time, to, two, up, way, we, well, were, what, when, where, which, who, will, with, would, years, you, your.

Каждое слово сопровождается тремя показателями: частотой (f), математическим ожиданием (m) и относительной частотой (p)[1], что дает возможность применить обе наши формулы [1] и [2], например:

		A	B	C	D	E	F	G	H	J	K	L	M	N	P	R
but	f	283	293	170	175	221	391	772	119	501	297	266	89	316	387	101
	m	383	236	153	149	314	420	658	270	700	252	209	52	253	254	79
	p	319	537	480	506	304	402	507	190	309	508	551	738	540	659	552

Таблица 1.3.1
Лексическое сходство американских жанров

	a	b	c	d	e	f	g	h	j	k	l	m	n	p	r
a		227	219	223	215	231	230	214	224	205	192	198	200	186	208
b	173		226	236	217	234	236	223	232	203	200	198	195	185	206
c	169	187		230	222	234	233	218	227	205	189	192	201	189	210
d	170	210	186		212	239	246	232	243	204	219	197	196	186	210
e	177	193	178	202		223	223	215	218	200	184	188	194	182	204
f	178	199	189	200	192		247	228	234	213	199	201	206	196	217
g	173	206	193	213	187	209		232	239	216	201	205	208	198	220
h	156	165	158	177	178	170	168		234	194	177	185	186	178	195
j	162	184	176	187	182	191	187	173		200	186	193	193	182	203
k	159	172	167	182	164	184	193	137	163		220	212	226	214	212
l	160	166	166	154	166	183	192	136	161	242		206	214	208	205
m	165	190	174	199	178	197	205	149	176	239	241		204	202	207
n	157	168	157	178	161	191	188	134	160	245	247	236		207	207
p	159	175	167	184	166	183	193	156	157	241	253	238	241		204
r	170	188	175	197	179	193	204	125	150	216	223	225	222	226	

Среднее значение LS равно 210 в первой зоне и 183 — во второй. (В словацком материале во второй зоне LS возрастало.) Если в первой зоне среднее значение LS мало колеблется вокруг цифры 210, во второй зоне отмечено падение до 160 у жанра h (miscellaneous — прочее). В первой зоне вырисовывается кластер жанров от A до J, во второй зоне явно выделяется кластер художественной литературы от K до R.

В нижеследующей таблице представлены слова, ставшие маркерами четырех и более жанров (при S * > 3)[2].

Таблица 1.3.2
Лексические маркеры американских жанров

	a	b	c	d	e	f	g	h	j	k	l	m	n	p	r
a			3		4	4									4
also	3			3	3				4						
are		3	3	3	12			8	11						
at	7									4	4			6	
back										6	16		14	9	
be		4			3			10	10						
but							4				4	5	3	8	
could										7	7	6	6	10	
did										7		6	3	8	
down										10	6		14	10	
had										25	17	10	15	17	5

[1] Относительная частота здесь показана на 100 000 словоупотреблений.

[2] В Таблице II электронного издания помещен полный перечень лексических маркеров среди ста самых частых слов.

	a	b	c	d	e	f	g	h	j	k	l	m	n	p	r
has	5	11	11		4										
have		4		4	5						2	2		2	
he										32	29	7	31	22	
her										8	13	7	21	37	
him										18	17	4	21	15	
his							11			19	12	21	9		
I										11	21	4	20	37	14
in	3					3	3		11						
into										4	6		6	3	
is		8	8	10	8		7								
it										4	10	5	5	9	4
like										8	8		7	13	
man				4			3			4	6		11	3	
may				4	4			9	7						
me										8	7		17	14	7
my				4						6	6	3	12	11	11
new	7	4	4	5	3			4							
no										3	8	6	4		
not		3		6			5					5			
now										6	6		5	4	
of				7		3	12	16	21						
on	4									3	6		4		
or					8	5		8	8						
our		7		6			8	9							
out										9	11		3	9	
said	17									8	10	4	15	20	8
she										20	12	4	13	43	3
the	3	3				3		4	11						
them										6	3	5	5	3	
then										4	8		6	7	
they										6		3	6		4
this		3		6	3			6	5						
up										7	13		10	9	
was										22	16	7	16	18	7
we		7		13			5		4						
were									4	3		4	4		3
what										7	5	3	3	5	
which				7			6	3	10						
who	4	5	5	3		3	6								
will	13	10			7			8							
would										11	4	6		7	
you					12					6	20	9	18	26	11
your					25						4			8	4

Таблица 1.3.3
Число общих лексических маркеров американских жанров

	a	b	c	d	e	f	g	h	j	k	l	m	n	p	r
a		5	3	3	5	3	4	3	3	3	3	1	2	2	1
b			5	9	7	2	6	9	6		1	2	1		
c				4	5	2	2	4	2			1		2	1
d					5	2	10	10	11	1	2	1	1	2	1
e						4	2	8	7	1	2	1	2	2	3
f							4	3	6						1
g								4	5	2	3	3	3	2	2
h									11					1	
j										1		1	1		1
k											34	20	32	31	10
l												18	32	34	12
m													21	19	10
n														28	9
p															10

Из этой таблицы непреложно следует факт существования кластера художественной речи с жанрами K–R, для которого характерны такие маркеры: личные и притяжательные местоимения (кроме we), приглагольные наречия (back, down, out, over, up), союз but, наречия времени (now, then), формы прошедшего времени (was, could, did, had), вопросительные what и when, а также слова like, no, said, would.

Едва проглядывает жанр, объединяющий жанр B, D, E, H, J («книжная речь»?) с такими группами маркеров: артикли и указательные местоимения (a, the, these, this), относительные местоимения (who, which), предлоги of и in (ср. [Mistrik]), местоимение we (+ our), слова also, may, new, or, will. Специфичны маркеры жанра A (репортаж) — предлоги at и on, а также слово said (!). Пока загадочным кажется маркер you (+ your) в жанре E (skills and hobbies), но см. ниже в 1.4.

К сожалению, в KuFran ранговый словарь «съел» половину печатного пространства. Если бы от рангового словаря осталась одностраничная таблица частотного спектра, на освободившемся месте можно было дать жанровую информацию о 10 тыс. слов.

Как бы то ни было, из ста самых частых слов рассматриваемого словаря можно сделать вывод о пользе и пределах лемматизации. В Таблице 1.3.4 показаны изменения в значениях S при лемматизации четырех слов.

Таблица 1.3.4
Лексические маркеры в зависимости от лемматизации

	a	b	c	d	e	f	g	h	j	k	l	m	n	p	r
I										11	21	4	20	37	14
I(+me,my)										15	23	6	27	35	17
he										32	29	7	31	22	
his							11			19	12	21	9		
HE(+him,his)							9			41	35	8	34	28	
is		8	8	10	8		7								
was										22	16	7	16	18	7
BE*				3					13						
had										25	17	10	15	17	5
HAVES*		3								4	8	7		12	

У личных местоимений лемматизация ведет к повышению значений S, но не меняет кардинального противопоставления двух кластеров. У глагола be значения S резко падает, а с ним и число маркеров. В меньшей степени это же наблюдается у глагола have.

Частые слова, рассмотренные в данном параграфе, заставляют вспомнить интереснейшую работу Дагласа Байбера[1], применившего технику факторного анализа к британскому корпусу LOB[2], построенному в точности по модели корпуса Университета Брауна (того же объема и с тем же набором 15 жанров).

К корпусу LOB были добавлены тексты из корпуса London-Lund[3] и 16 текстов двух жанров: Personal letters и Professional letters.

Байбер использовал 67 переменных (существительные, предлоги, наречия, местоимения 1 лица, местоимения 2 лица, инфинитивы, predictive modals (will, would, shall) и множество других, предполагающих серьезную и вдумчивую разметку исходного текста. Грамматические явления, исследуемые Байбером, конечно, далеко выходят за рамки нашей книги, тем не менее полезно рассмотреть некоторые из них в связи с изучаемыми жанрами.

[1] *Biber D.* Variation across Speech and Writing. Cambridge, 1988.

[2] *Johansson S., Leech G. N., Goodluck H.* Manual of information to accompany the Lancaster-Oslo-Bergen Corpus of British English… Oslo, 1978.

[3] *Svartvik J., Quirk R.* (eds.) A Corpus of English Conversation. Lund, 1980.