А. Я. Шайкевич
Институт русского языка им. В. В. Виноградова РАН
(Россия, Москва)
irlras@mail.ru

ЛЕКСИЧЕСКИЕ МАРКЕРЫ ПОДКОРПУСОВ ТЕКСТОВ

В лингвостатистических исследованиях время от времени возникает задача выявить лексическую специфику того или иного подкорпуса, составляющего часть более широкого корпуса текстов. Удобным инструментом для решения этой задачи в случае конкретного слова служит формула

$$S = (f–m–1)/\sqrt{m},$$

где f — частота слова в подкорпусе, m — математическое ожидание частоты слова на основе данных широкого корпуса.

Как к первому примеру обратимся к слову *пауза*. В общем корпусе текстов Чехова оно встретилось 422 раза; зная долю драмы (0,06) и предполагая равномерное распределение нашего слова, мы ожидаем, что в драме *пауза* встретится 422 × 0,06 = 23 раза (m). В действительности, его частота (f) в драме составляет 380. Подставляя значения f и m в формулу, получаем S = (380 – 23 – 1)/4,7 = 356/4,7 = 75 — исключительно высокую величину, ставящую слово во главу рангового списка специфических слов драмы: *пауза* (75), *уходить* (42), *входить* (37), *ты* (36), *да* (частица) (31), *я* (30), *смеяться* (28), *садиться* (26), *вот* (24), *вы* (24). Эти десять слов можно считать функциональными маркерами чеховской драмы; к ним присоединяются *в сторону, вбегать, вскакивать, вставать, дверь, занавес, идти, кричать, напевать, нет, плакать, прощай*(*те*), *рыдать, сцена, хохотать*. Этим конструктивным элементам драмы противостоят маркеры, мотивированные содержанием отдельных текстов: *ау, брас, вишневый, генеральша, имение, картограмма, лесничество, любить, медальон, нянечка, пистолет, подагра, сад, торги, траур, фрак, целовать, чайка, шлем*.

Наша формула восходит к гипотезе о распределении Пуассона, в максимальной степени ее эффективность проявляется при низких m. Обратимся к переписке Чехова, где подтекстами выступают письма к определенному адресату. Примером могут служить письма Чехова к брату Александру. Лексическими маркерами (S>3) оказались 122 слова. Конструктивными элементами этого подкорпуса могут считаться обращения к адресату: *ты* (44), *Гусев* (15), *штаны* (15), *брат* (11), *Сашечка* (10), *двуличновольнодумствующий* (4); а также шутливые подписи: *благодетель* (17), *Antonius* (6). Сфера «контента» распадается на несколько областей:

1. *Николка, родственник, родитель, отец, тетка*;
2. *Суворин, «Новое время», Лейкин*;
3. *гонорар, деньги, копейка, контора, рубль*;
4. *цуцык, дети, чадо*;

5. *будь здрав, здравствовать, ибо, оный, сей*;
6. *жопа, нужник, ххх*.

Что касается подкорпуса писем в целом, который составляет 37% общего корпуса Чехова, то здесь надо учитывать тенденцию нашей формулы завышать S при больших значениях m. Если в драме лишь слово *пауза* имело S>50, в письмах находим пять таких слов: *я, ваш, А. Чехов, писать, письмо*. У 63 маркеров подкорпуса писем S>18 (в драме таких слов было 12), например, *А. Чехов, будет, будь здоров, вчера, вы, выслать, жму руку, здоров, здоровье, ибо, кланяться, книга, май, Мелихово, многоуважаемый, написать, насчет, Петербург, погода, поклон, получить, привет, пьеса, рассказ, сей, Суворин, Таганрог, твой, телеграмма, я, Ялта*. Общее число маркеров составляет здесь 875.

Лексические маркеры диалогов персонажей у Чехова можно сравнить с маркерами диалогов в русской прозе 1850–1870 гг. Вот как выглядят ранги S верхушки ранговых словарей:

	50/80	Чехов		50/80	Чехов		50/80	Чехов
я	2	1	ваш	8	10	будет	16	19
вы	1	2	знать	20	11	же	19	21
ты	3	3	вот	13	12	что	10	22
не	9	4	ведь	15	13	мой	21	23
а	6	5	кто	30	14	твой	23	26
да(союз)	7	6	мы	13	15	какой	22	35
ну	4	7	нет	11	17	-то	5	36
это	12	8	так	12	17	уж	24	48
да(частица)	14	9	-с	18	17	ли	17	>100

За исключением *-то* и *ли* конструктивный компонент двух подкорпусов практически совпадает. Среди маркеров Чехова с низкими S ясно обнаруживаются две струи:

1. *ага, али, ан, вовсе, вон*(=там), *грош, дьявол, дяденька, жулик, заместо, зря, ирод, кажись, леший, наплевать, начальство, папочка, побей Бог, покеда, поп, пороть, почем, почитай, пропить, свинья, сволочь, скотина, тапереча, хам, харя, шабаш, шляться*.

2. *во имя, действительно, доказать, наука, необходимо, предрассудок, Россия, собственно говоря, супруга, сэр, трудиться, философствовать, цивилизация*.

В докладе также рассматриваются интригующие маркеры беллетристики Чехова.