А.Я. Шайкевич
A.Ya. Shaikevich

МЕРЫ ЛЕКСИЧЕСКОГО СХОДСТВА ЧАСТОТНЫХ СЛОВАРЕЙ

MEASURES OF LEXICAL SIMILARITY BETWEEN FREQUENCY DICTIONARIES

Аннотация. Предлагаются две меры лексического сходства частотных словарей (ЧС). Первая мера учитывает общее лексическое покрытие популяции двумя частотными словарями. Наибольший вклад в результат вносят частые слова. Вторая мера применима только к ЧС подкорпусов, она опирается на число общих лексических маркеров и в меньшей мере зависит от самых частых слов. Два подхода иллюстрируются данными ЧС русского, английского и китайского языков.

Ключевые слова. Лингвостатистика, подкорпусы.

Abstract. Two measures of lexical similarity between frequency dictionaries (FDs) are proposed. The first statistic measures common coverage of the population by two FDs. The main contribution to the result is made by most frequent words. The second measure is applicable to subcorpora of some general corpus. This statistic is based on the number of common keywords of FDs and to a lesser extent depends on most frequent words. The examples are drawn from FDs of Russian, English and Chinese.

Keywords. Statistical linguistics, subcorpora.

В лингвостатистике иногда возникает необходимость сравнить лексику двух частотных словарей и количественно определить их лексическое сходство.

В качестве простейшей меры такого сходства будем использовать формулу

$$C_{xy} = \Sigma \, min \, \{px_i, py_i\} \qquad (1)$$

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

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

Можно заранее ожидать, что наибольший вклад в общую сумму внесут самые частые слова. Чтобы получить более дифференцированные результаты модифицируем формулу 1 следующим образом:

$$C_{xy} = \Sigma\, min\, \{px_i, py_i\} / 0{,}5\, (\Sigma\, px_i + \Sigma\, py_i), \qquad (2)$$

где $\Sigma\, px_i$ и $\Sigma\, py_i$ подсчитываются для определенной зоны рангового словаря.

Покажем расчет на примере пяти самых частых слов из Частотного словаря современного русского языка, построенного на Национальном корпусе русского языка [Ляшевская, Шаров 2009: 419]. Сравним относительные частоты слов в Художественной литературе (ХЛ) и в Публицистике (Пуб.):

	ХЛ	Пуб.	min		ХЛ	Пуб.	min
и	.0365	.0355	.0355	на	.0167	.0167	.0167
в	.0262	.0364	.0262	я	.0183	.0104	.0104
не	.0223	.0169	.0169	Σ	.1200	.1159	.1057

C = .1057/.1180 = .896

Подсчитаем коэффициент лексического сходства (*C*) для жанров этого же словаря, добавив к ХЛ и Пуб. еще два типа текстов: Устная речь (УР) и Прочее. Результаты приведены в табл. 1. В правой верхней части матрицы даются значения *C* для ста самых частых слов рангового словаря, в левой нижней части – для следующих четырехсот слов.

Таблица 1. Матрица лексического сходства жанров НКРЯ

	УР	ХЛ	Пуб.	Проч.
УР		.751	.667	.670
ХЛ	.661		.892	.852
Пуб	.561	.716		.933
Проч	.527	.700	.859	

Публицистика и прочая литература близки друг к другу. От них сильно отличается Художественная литература, совсем далека от них Устная речь.

Еще в 1960-е годы был создан Brown University Corpus of American English со следующими жанровыми подкорпусами (указана доля жанра в общем объеме – 1 миллион словоупотреблений):

```
A. Press: Reportage   8.8%   K. Fiction: General  5.8%
B. Press: Editorial   5.4%   L. Fiction: Mystery
C. Press: Reviews     3.4%      and Detective      4.8%
D. Religion           3.4%    M. Fiction: Science  1.2%
E. Skills and Hobbies 7.2%    N. Fiction: Adventure
F. Popular Lore       9.6%      and Western        5.8%
G. Belles Lettres,            P. Fiction: Romance and
   Biography, etc.   15.0%      Love Story         5.8%
H. Miscellaneous      6.0%    R. Humor             1.8%
J. Learned and
  scientific Writing 16.0%
  [Kučera, Francis 1967: 277–293, 100 самых частых слов]
```

Точно такой же объем и расклад жанров находим в Ланкастерском корпусе китайского языка (Lancaster Corpus of Mandarin Chinese – Oxford Text Archive), только жанр N сменил ярлык, здесь он называется Martial Art Fiction.

Таблица 2. Матрица лексического сходства жанров корпуса университета Брауна

	A	B	C	D	E	F	G	H	J	K	L	M	N	P	R
A		89	88	85	87	90	87	84	87	78	76	78	77	74	82
B			90	92	90	92	92	87	90	78	75	81	75	74	85
C				88	87	91	90	84	88	78	75	78	76	74	86
D					87	90	92	86	90	78	74	80	75	76	82
E						90	87	88	88	77	73	77	75	73	81
F							94	86	90	81	78	81	80	78	85
G								85	89	82	79	87	80	78	86
H									90	70	66	71	68	65	75
J										78	68	75	75	69	77
K											92	89	94	91	87
L												88	92	92	87
M													87	87	88
N														90	88
P															87
R															

Сто самых частых слов (со средним значением *C* = .827), ясно указывают на существование двух жанровых кластеров – ДЕЛОВОГО (A, B, C, D, E, F, G, J) со значениями *C* от .87 до .92 и ЛИТЕРАТУРНОГО (K, L, M, N, P) со значениями *C* от .87 до .94.

Результаты для китайского корпуса приведены в табл. 3. В правой верхней части матрицы даются значения *C* для ста самых частых слов рангового словаря, в левой нижней части – для следующих 1085 слов.

Среднее значение *C* в первой сотне китайских слов равно 74,1 (значительно ниже, чем в английском языке), во второй зоне оно равно 49,2. ЛИТЕРАТУРНЫЙ кластер помимо жанров K, L, M, N, P включает также R. ДЕЛОВОЙ кластер охватывает жанры B, C, D, E, F, J. Жанры A (репортаж) и G (биографии и эссе) располагаются между этими двумя кластерами.

Таблица 3. Матрица лексического сходства жанров Ланкастерского корпуса

	A	B	C	D	E	F	G	H	J	K	L	M	N	P	R
A		82	78	77	77	85	87	67	76	76	76	78	75	76	70
B	62		88	82	86	86	79	70	86	69	67	76	62	73	62
C	49	66		83	75	78	73	72	86	62	68	70	61	64	57
D	50	51	42		80	85	77	65	86	68	73	74	68	69	61
E	55	50	38	55		83	77	65	79	68	73	71	68	68	61
F	68	60	45	58	66		88	64	80	76	82	87	75	78	70
G	69	53	41	54	53	69		59	72	86	91	86	83	86	79

	A	B	C	D	E	F	G	H	J	K	L	M	N	P	R
H	41	49	51	33	34	36	32		70	50	55	53	48	50	47
J	53	61	53	58	54	60	47	48		64	68	71	62	72	57
K	59	41	28	44	50	58	68	22	35		87	84	82	91	83
L	62	46	33	48	54	63	70	27	42	64		86	85	87	81
M	50	43	33	45	46	54	56	26	41	49	58		78	88	77
N	52	38	25	44	49	51	65	20	32	68	66	51		81	82
P	58	42	30	44	51	61	69	22	37	74	70	57	65		82
R	51	36	25	37	45	51	59	19	30	64	60	48	58	62	

По-другому выглядит конфигурация основных жанров в корпусе русской прозы 1850–1870-х гг. Здесь представлено восемь жанров – роман, повесть, рассказ, очерк, историческая проза, воспоминания, драма и прочее. Межжанровые значения *C* представлены в табл. 4 (справа зона 1–100 рангового словаря, слева – зона 601–3600).

В обеих зонах драма противостоит остальным жанрам, которые расположились цепочкой именно в том порядке, что дается в табл. 4. В зоне 601–3600 начинает проглядывать не столько жанровое, сколько собственно лексическое своеобразие.

Таблица 4. Матрица лексического сходства основных жанров прозы 1850–1870-х гг.

	Дра-ма	Ро-ман	Пов.	Расск.	Очерк	Ис-тор.	Восп.	Проч.
Драма		77	80	79	76	74	72	71
Роман	68		97	93	92	90	88	86
По-весть	70	88		95	92	90	89	85
Рассказ	68	81	84		92	92	88	86
Очерк	63	80	79	80		91	90	92
Истор.	58	71	71	72	68		87	88
Вос-пом.	61	77	76	72	76	64		90
Проч.	54	70	69	69	74	64	74	

Предложенная мера лексического сходства кажется очень естественной, ею можно воспользоваться при ЧС 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 меры, опирающиеся именно на различия в ЧС. Рассмотрим подход, при котором сравниваемые ЧС относятся к какому-то уже существующему (или специально конструируемому на этом случай) корпусу с числом подкорпусов более двух. Для каждого из подкорпусов можно статистическими средствами выделить лексические маркеры. Используем для этого формулу

$$S = (x – m — 1) / \sqrt{m}, \quad (3)$$

где x –частота слова в подкорпусе, m – математическое ожидание этой частоты в рамках данной нулевой гипотезы.

Выбрав некоторый порог (скажем, $S > 3$), будем считать лексическими маркерами все слова, чьи S превысили данный порог.

Получив набор лексических маркеров для каждого из подкорпусов, мы для пары сравниваемых подкорпусов (X и Y) строим четырехклеточную таблицу вида

	X	не-X
Y	a	b
не-Y	c	d,

где a – число общих маркеров X и Y, b – число маркеров Y, отсутствующих в X, c – число маркеров X, отсутствующих в Y, d – число маркеров, отсутствующих как в X, так и в Y.

Воспользуемся простейшим коэффициентом связи, введенным Дж. Юлом еще в 1900 г. [Upton, Cook: 435]:

$$Q = (ad – bc)/(ad+bc), \quad (4)$$

принимающим значения от –1 до +1.

Коэффициенты связи, полученные по формулам (3) и (4), не прошли еще проверки на широком материале ЧС. Исходя из общих соображений, можно ожидать интересных результатов при разнообразии подкорпусов в рамках общего ЧС. Складывается впечатление, что в зоне тысячи самых частых слов этот подход себя оправдывает.

Новые лингвостатистические инструменты открывают перед исследователем и новые (хотя и неясные) перспективы.

Литература

1. *Ляшевская О.Н., Шаров С.А.* (2009), Частотный словарь современного русского языка на материале Национального корпуса русского языка. М.

2. *Kučera F.* (1967), Computational analysis of presentday American English, Providence.

3. *Upton G., Cook I.* (2014), A Dictionary of Statistics, OUP.

References

1. *Lyashevskaya O.N., Sharov S.A.* (2009), Chastotnyj slovar' sovremennogo russkogo yazyka na materiale Natsional'nogo korpusa russkogo yazyka. [Frequency Dictionary of Contemporary Russian on the Material of the National Corpus of Russian Language]. Moscow.

2. *Kučera F.* (1967), Computational analysis of presentday American English, Providence.

3. *Upton G., Cook I.* (2014), A Dictionary of Statistics, OUP.

Шайкевич Анатолий Янович

Институт русского языка им. В.В. Виноградова РАН (Россия)

Shaikevich Anatole

The Vinogradov Institute of Russian Language of the Russian Academy of Sciences (Russia)

E-mail: lingstat@yandex.ru