А. И. Чапля

ИСПОЛЬЗОВАНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ В СТИЛИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

Стилеметрия – это прикладная филологическая дисциплина, занимающаяся измерением стилевых характеристик текстов с целью систематизации и упорядочения (типологии, атрибуции, датировки, диагностики, реконструкции и т.п.) текстов и их частей [1, с. 420].

Для решения любой из перечисленных выше задач необходимо отобрать определенное число числовых параметров, вычисляемых по исследуемым текстам. Число таких параметров, предлагаемых разными исследователями, составляет не одну сотню. Их простейшая классификация приведена в работе [2, с. 209]. Ниже детально рассмотрен статистический метод принадлежности текстов к некоторому подъязыку или же языку в целом.

Для проведения исследования были использованы частотные словари, полученные по французским текстам нефтехимии, нефти и газу, электроники и публицистики. Каждый такой частотный словарь был получен из текстов

длиной в 100 000 словоупотреблений. В табл. 1 приведены данные о величинах абсолютной частоты (F) 520 словоформ, взятых из упомянутых частотных словарей.

Таблица 1

№ п/п	Словоформа	Абсолютные частоты			
		нефтехимия	нефть и газ	электроника	публицистика
1	de	5592	6388	6445	6462
5	les	1907	1936	1867	2052
15	dans	907	1000	977	1115
37	production	163	257	0	54
46	forage	16	184	0	0
52	produits	213	168	0	0
68	pression	153	133	0	0
70	raffinerie	30	132	0	0
91	partie	76	95	70	55
108	hydrocarbures	376	79	0	0
139	pays	17	69	0	152
147	temps	46	65	141	108
171	prix	31	57	0	113
187	fonotion	46	54	110	0
220	profondeur	8	49	0	0
227	appareil	4	7	49	0
520	vapeur	41	22	0	0

Отвлекаясь на время от лингвистического содержания задачи, представим себе, что нам заданы некоторые объекты C_1, C_2, ..., C_k (слова) и их признаки F_1, F_2, ..., F_k, т.е. частоты этих объектов (слов и словоформ) в разных подъязыках (отраслях, функциональных стилях).

Для объекта (слова) C_1 частоты F_1, F_2, ..., F_k в разных K частотных словарях являются его признаками, т.е.

$$C_1 (F_1^1, F_2^1, \ldots, F_k^1)$$
$$C_2 (F_1^2, F_2^2, \ldots, F_k^2)$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$C_I (F_1^{(i)}, F_2^{(i)}, \ldots, F_k^{(i)})$$

Для каждого объекта C_I (с соответствующим ранговым номером), т.е. для каждого слова частотных словарей соответствующего ранга, можно вычислить среднее арифметическое значение признака F_i [3, с. 102].

Если обозначить среднее арифметическое значение признака F_i через $\overline{X}_i$, то

$$\overline{X}_i = \frac{F_1^{(i)} + F_2^{(i)} + \cdots + F_k^{(i)}}{K}, \qquad (1)$$

т.е. среднее арифметическое значение F_i равно сумме всех признаков данного объекта, деленной на количество сравниваемых частотных словарей.

Каждую словоформу с частотами, употребленными в подъязыках, можно рассматривать как дискретный вариационный ряд [4, с. 131].

Тогда для него можно определить среднее линейное отклонение от среднего арифметического значения по формуле [5, с. 221]

$$\bar{e} = \frac{\sum_{i=1}^{n} (F_i - \overline{X})}{n} \qquad (2)$$

и дисперсию

$$\sigma^2 = \frac{\sum_{i=1}^{n}(F_i - \overline{X})^2}{n} \qquad (3)$$

Извлекая квадратный корень из величины σ^2, найдем среднее квадратическое отклонение от среднего арифметического значения

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(F_i - \overline{X})^2}{n}} \qquad (4)$$

Величина σ показывает, на сколько, в среднем отклоняются конкретные частоты словоформ (слов) от среднего их значения.

Для практических исследований представляется более удобным вычислять величину коэффициента вариации V, которая обычно выражается в процентах [3, с. 122]

$$V = \frac{\sigma}{\overline{X}} \cdot 100\% \qquad (5)$$

Величины σ и V являются мерами степени равномерности частот словоформ (слов). Чем меньше σ и V, тем однороднее изучаемая совокупность признаков $F_1, F_2, \ldots, F_k$ [5, с. 223].

Принимая во внимание, что в работах по лингвостатистике за максимальную ошибку принимают величину ± 30 % [6, с. 89], находим по формулам (1)–(5) для каждой словоформы (для каждого слова) все необходимые величины.

Если величина V получается менее чем 30 %, то частоты данной словоформы распределены в языке равномерно, т.е. можно считать, что такая словоформа (слово) относится в языку в целом.

Если же величина V оказывается больше 30 %, то можно предположить, что распределение частот словоформы (слова) неравномерно и такая словоформа специфична для какого-то конкретного подъязыка.

Для выяснения вопроса, к какому именно подъязыку относится данная словоформа (или данное слово), применяется *мода* [3, с. 109] D – значение наибольшего признака F_i. Тот подъязык, который содержит эту наибольшую частоту данной словоформы (слова), может считаться искомым.

Однако может случиться, что в двух подъязыках значение частот очень близки. Для учета таких случаев вводится дополнительная проверка степени близости двух наибольших частот F_1 и F_2. нами было принято априорно, что, если эти два значения отличаются друг от друга не более чем на 5 %, то они считаются близкими. Такая проверка осуществляется по следующей процедуре:

1. Находится величина ΔF, равная 5% от F_1:

$$\Delta F = \frac{F_1}{100} \cdot 5 \qquad (6)$$

2. Находится величина F_2^{I}:

$$F_2^{\text{I}} = F_1 - \Delta F \qquad (7)$$

3. Определяется величина разности $F_2^{\text{I}} - F_2$

Если $F_2^{\text{I}} - F_2 \le 0$, то F_1 и F_2 – близки.

Если же $F_2^{\text{I}} - F_2 > 0$, то F_1 и F_2 – не близки.

Фрагменты и рекомендации по итогам вычисления всех этих величин представлены в табл. 2.

Таблица 2

Фрагменты результатов машинного анализа ключевых слов (словоформ)

№ п/п	Словоформа	Статистические характеристики			Результат анализа и рекомендация
		$\overline{X}$	σ	V	
1	2	3	4	5	6
1	de	6221,7501	364,6179	5,86037	Слово принадлежит языку в целом
5	les	1940,5000	68,8785	3,54952	Слово принадлежит языку в целом
37	production	119,5000	99,6656	83,40225	Слово может быть ключевым в подъязыке нефти и газа
46	forage	50,0000	77,6401	155,28039	Слово может быть ключевым в подъязыке нефти и газа
52	produits	95,0000	96,3820	101,45479	Слово может быть ключевым в подъязыке нефтехимии
68	pression	72,0000	72,4189	100,58183	Слово может быть ключевым в подъязыке нефтехимии
70	raffinerie	40,4999	54,2286	133,89798	Слово может быть ключевым в подъязыке нефти и газа
91	partie	74,0000	14,3352	19,37198	Слово принадлежит языку в целом
108	hydrocarbures	113,7500	154,8069	136,09401	Слово может быть ключевым в подъязыке нефтехимии
139	pays	59,5000	59,1460	99,40505	Слово может быть ключевым в подъязыке публицистики
147	temps	97,0000	37,0337	41,14863	Слово может быть ключевым в подъязыке электроники
171	prix	50,2499	41,4691	82,52560	Слово может быть ключевым в подъязыке публицистики
187	fonotion	52,5000	39,0736	74,42599	Слово может быть ключевым в подъязыке электроники
220	profondeur	14,2499	20,3770	142,64571	Слово может быть ключевым в подъязыке нефти и газа
227	appareil	25,0000	23,0542	92,21713	Слово может быть ключевым в подъязыке электроники
236	pipe-line	14,0000	18,9208	135,14920	Слово может быть ключевым в подъязыке нефти и газа

299	monde	35,5000	39,8528	112,26156	Слово может быть ключевым в подъязыке публицистики
341	vitesse	49,7500	35,4920	71,34085	Слово может быть ключевым в подъязыке электроники
388	reforming	19,7499	21,0994	106,83274	Слово может быть ключевым в подъязыке нефтехимии
433	etat	30,2500	20,0795	66,37861	Слово может быть ключевым в подъязыке публицистики
512	nombreux	21,0000	2,9999	14,28571	Слово принадлежит языку в целом
520	vapeur	15,7500	17,1227	108,71568	Слово может быть ключевым в подъязыке нефтехимии

Вычисление величин V и D дает возможность ответить более или менее точно на ряд таких вопросов, как:

1. Принадлежит ли слово языку в целом, или определенному функциональному стилю?

2. Характерно ли слово для всего периода творчества определенного автора или оно специфично для отдельного периода его творчества?

3. Принадлежит ли слово языку в целом или оно специфично для текстов определенной специальности? и т.д.

Таким образом, использованный в настоящей работе метод может быть также применен при изучении стилей разных авторов, при изучении особенностей лексики одного автора в разные периоды его жизни, при изучении различных функциональных стилей (прозы, поэзии, драматургии), наконец, в изучении словарей различных специальностей и тем. Этот же метод может быть использован и при дешифровке анонимных текстов.

ЛИТЕРАТУРА

1. М*артыненко, Г. Я.* Стилеметрия / Г. Я. Мартыненко, С. В. Чебанов // Прикладное языкознание. – Санкт–Петербург : СПбГУ, 1996. – С. 420–435.

2. *Мальцева, Г. Ф.* Некоторые количественные приемы описания индивидуального авторского стиля / Г. Ф. Мальцева // Статистика текста: сб. статей. – Минск : Изд-во БГУ, 1969. – Т. 1. – С. 206–248.

3. *Шторм, Р.* Теория вероятностей. Математическая статистика. Статистический контроль качества / Р. Шторм. – М. : Мир, 1970.

4. *Маринеску, И.* Основы математической статистики и ее применение / И. Маринеску, Ч. Мойнягу, Р. Никулеску, Н. Ранку, В. Урсяну. – М. : Статистика, 1970.

5. *Суслов, И. П.* Общая теория статистики / И. П. Суслов. – М. : Статистика, 1970.

6. *Фрумкина, Р. М.* Статистические методы изучения лексики / Р. М. Фрумкина. – М. : Наука, 1964.