ЕВРАЗИЙСКОЕ НАУЧНОЕ ОБЪЕДИНЕНИЕ
EURASIAN SCIENTIFIC ASSOCIATION
www.esa-conference.ru

УДК 748,1751

Данные статистического анализа текстов, порождаемых в состоянии стресса и в «обычном» состоянии

Атаманова Ольга Викторовна, канд. филол. наук, преподаватель
Петрова Екатерина Дантесовна, преподаватель
Салищева Ольга Гурьевна, преподаватель
Военная академия связи им. С.М. Буденного (г. Санкт-Петербург)

В работе представлены некоторые данные статистического анализа текстов, порожденных в стрессовом состоянии и в период стабилизации стресса. Стресс сопровождается обеднением лексики информанта, что отражается на изменении статистических параметров. Построены графики зависимости «ранг-частота» на основе частотных словарей.

Ключевые слова: *психологический стресс, статистический анализ текстов, закон Ципфа, распределение Парето-Ципфа-Мандельброта-Лотки, частотный словарь, обеднение лексики, зависимость «ранг-частота».*

Можно ли использовать статистические параметры, полученные в ходе лексико-статистического анализа текстов, для определения психологического состояния респондента? Если мы станем изучать тексты устной монологической речи одного и того же человека, порожденной в различных условиях с точки зрения стресса, то есть в состоянии стресса и в «обычном» состоянии, когда отсутствуют факторы, провоцирующие стресс, то в большинстве случаев мы будем наблюдать следующую картину. При всем разнообразии наших индивидуальных особенностей и привычек, то есть при разнообразии речевого поведения каждого из нас, найдется ряд особенностей речи, которые, вместе взятые, однозначным образом указывают на изменение сознания респондента. К частным случаям измененного состояния сознания человека относят состояние психологического стресса [1, с. 14-16].

Речь человека, пребывающего в состоянии беспокойства, по мере нарастания этого беспокойства, становится все более отрывистой и все менее связной. По-видимому, та энергия, которую стресс так или иначе забирает у каждого из нас, оставляет все меньше возможностей для абстрактного мышления. Итак, мы все меньше абстрагируемся от реальности, от повседневности, что, конечно, отражается на лексике говорящего, и слова, обозначающие отвлеченные понятия, слова с абстрактным значением все меньше присутствуют в ней, и вместе с тем, слова, относящиеся к повседневной жизни адресанта, используются все чаще. С точки зрения морфологии можно говорить о все меньшем количестве прилагательных, наречий, о преобладании коротких слов, служебных слов (по данным статистического анализа, доля служебных слов в речи всегда существенна, однако резкое увеличение их количества, безусловно, указывает на изменения состояния сознания человека). Обычно в речи такого адресанта присутствует довольно много слов-заместителей, заполнителей молчания, слов-паразитов и поисковых слов [2, с. 80-85]. Такие особенности изменения речи человека под влиянием стресса, безусловно, отразятся, в частности, на свойствах частотного словаря, построенного на основе оцифрованного устного текста адресанта. В целом можно говорить об обеднении лексики адресанта под влиянием психологического стресса [4, с. 48].

Можно ли каким-то образом измерить степень такого обеднения? И, в связи с этим, рассуждать о возможно более сильно или слабо выраженном стрессе? Можно подсчитать число рангов частотного словаря, соответствующее таким «обедненным» словоформам. И само это число считать показателем обеднения словаря для данного адресанта и при данных обстоятельствах. Итак, M – значение ранга, соответствующего последней «обедненной» словоформе.

Однако, анализируемые тексты имеют, вообще говоря, разную длину. Более удачной характеристикой обеднения лексики адресанта может служить показатель, равный отношению числа M к длине текста, на основе которого составлен словарь.

Существуют и другие статистические параметры, весьма «чувствительные» к свойствам анализируемого текста, связанным с изменением состояния сознания человека – каждое такое изменение немедленно сказывается на поведении параметров текста.

В целом при анализе текстов больших объемов говорят о некоем «предельном» состоянии или о состоянии насыщения выборки (в данном случае текст мы считаем выборкой, характеризующей случайный процесс порождения речи), когда вероятность появления в тексте того или иного слова обратно пропорциональна рангу этого слова (номеру строки в частотном списке). В этом состоит «классический» закон Ципфа, характерный для аналитических языков [3, с. 27-28; 5]. О «предельности» закона говорят в том смысле, что вероятность есть предельное значение частоты появления слова в тексте при неограниченном увеличении объема текста. Такой закон может быть выражен формулой

$$p = \frac{k}{i} \quad (1)$$

где p – вероятность появления слова в тексте, i – ранг, k – эмпирически подобранный коэффициент.

Если это выражение прологарифмировать, а затем построить график билогарифмической зависимости «ранг-частота» (по горизонтали – логарифм ранга i, по вертикали – логарифм вероятности p), то мы получим убывающую линейную функцию.

Указанный закон Ципфа справедлив, как уже было сказано, для аналитических языков, в которых изменения частей речи (склонение, спряжение) осуществляется не с помощью изменения окончаний, а

ЕВРАЗИЙСКОЕ НАУЧНОЕ ОБЪЕДИНЕНИЕ
EURASIAN SCIENTIFIC ASSOCIATION
www.esa-conference.ru

только благодаря служебным словам (так происходит, например, в английском языке). В русском языке, как мы знаем, изменение окончаний приводит к появлению новых форм слова. На уровне компьютерного анализа две разные словоформы соответствуют двум различным рангам частотного списка. Поэтому в случае синтетического языка, к каким относится, в частности, русский язык, зависимость несколько иная, график принимает вид параболы, а прямая является линией аппроксимации, к которой наш график приближается при неограниченном росте объема текста. Для таких случаев справедливо распределение Парето-Ципфа-Мандельброта-Лотки, существующее на основе расширенного закона Ципфа для синтетических языков:

$$p = \frac{k}{(i+\rho)^{\gamma}} \quad (2)$$

Здесь p – вероятность появления слова в тексте, i – ранг, k, ρ, γ – эмпирически подобранные коэффициенты [3, с. 35-37].

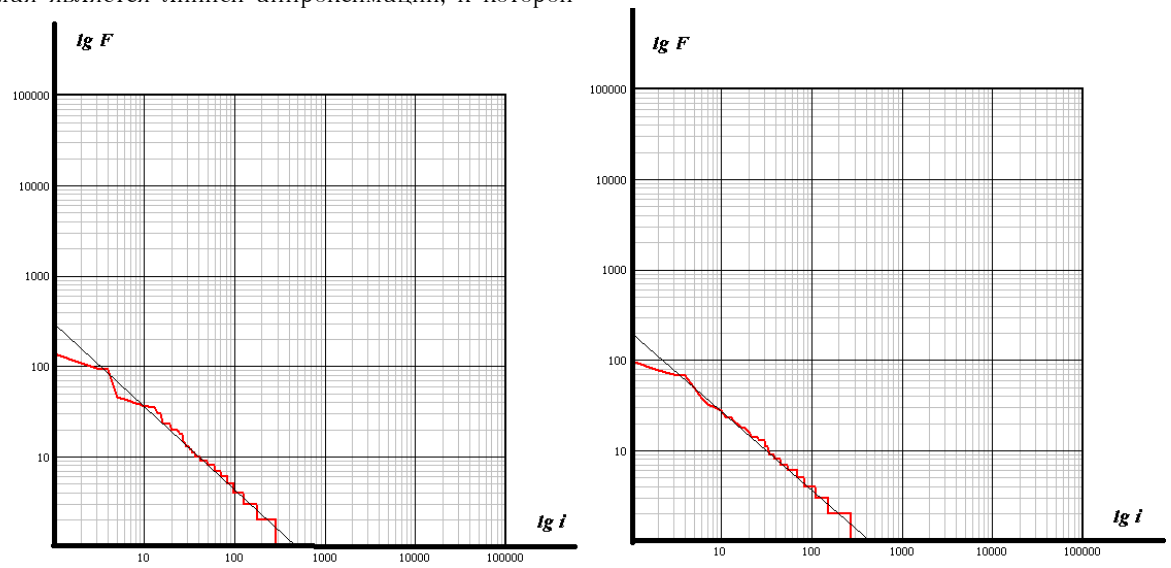

Рис. 1. **График зависимости «ранг-частота» для испытуемой в период выраженного стресса (левая диаграмма) и в период стабилизации стресса (правая диаграмма). По горизонтали – логарифмические значения рангов, по вертикали – логарифмические значения частот**

В качестве примера рассмотрим графики билогарифмической зависимости «ранг-частота» для одного и того же респондента Ж. (рис.1). Это пациентка Института акушерства и гинекологи им. Д.О. Отта, переживающая предродовой стресс (график зависимости «ранг-частота», построенный на основе текста, порожденного в этот период, соответствует левой диаграмме на рисунке), а затем проходящая период стабилизации после благополучного завершения родов (это правая диаграмма). Основные данные статистического анализа представлены в таблице.

Таблица 1. **Сравнительная таблица параметров распределения для информанта Ж. в период выраженного стресса и в период стабилизации**

Параметр	Период выраженного стресса	Период стабилизации стресса
Длина текста (N)	2812	1443
Объем словаря (V)	985	532
Отношение объема словаря к длине текста	0,350	0,368
Тангенс угла наклона аппроксимирующей прямой (γ)	0,918	0,867
Показатель синтетичности языка (ρ)	1,330	3,014
Среднее значение частоты ($F_{ср}$)	2,855	2,712

Из рисунка и из данных таблицы можно видеть, что в период выраженного стресса отношение объема словаря к длине текста меньше, чем в период стабилизации (соответственно, 0,350 и 0,368), то есть на текст данного объема приходится меньшее количество словоформ в период стресса, чем в период стабилизации, что указывает на обеднение словаря респондента, о котором говорилось выше. «Частотность» слов в целом выше в случае стресса (среднее значение частоты слова в тексте равно 2855 по сравнению с 2712 при стабилизации), что опять же указывает на большую повторяемость слов и обеднение словаря, связанное со стрессом. Аппроксимирующая прямая при уменьшенном стрессе проходит более полого, нежели при усиленном (тангенс угла наклона прямой равен, соответственно, 0,918 и 0,867), а значит, в период выраженного стресса мы приближаемся к предельному, «насыщенному» состоянию быстрее, чем в период стабилизации в силу все той же повторяемости слов, которая увеличивается по мере увеличения нестабильности психологического состояния человека.

В целом можно сказать, что уровень нестабильности данного информанта не слишком высок с учетом того, что разница статистических показателей не так велика, как бывает в других случаях.

www.esa-conference.ru

Итак, мы рассмотрели пример статистического анализа двух текстовых фрагментов, порожденных одним и тем же человеком, для разных периодов – периодов сильного беспокойства и периода относительной стабилизации состояния сознания. Как нам кажется, такое сравнение весьма полезно, так как с учетом навыков речевого поведения мы можем четко видеть разницу в статистических показателях в разные периоды, обнаружить «чувствительность» этих показателей, весьма высокую. И, если так, то психологическое состояние человека можно определять по статистическим свойствам порожденного текста, что, безусловно, имеет практическую пользу.

Литература:

1.Людвиг А. Измененные состояния сознания // Чарльз Тарт. Измененные состояния сознания. М. : Эксмо, 2003. – С. 14–37.

2.Носенко Э. Л. Эмоциональное состояние и речь. Киев, Вища школа, 1981. – 195с.

3.Пиотровский Р.Г. Лингвистическая синергетика: исходные положения, первые результаты, перспективы. СПб : Филол. фак-т СПб гос. ун-та, 2006. – 158 с.

4.Спивак Д.Л. Измененные состояния сознания: психология и лингвистика. СПб : Ювента : СПб : Филол. фак-т СПб гос. ун-та, 2000. – 293 с.

5.Mandelbrot B. On the Theory of Word Frequencies and on Related Markovian Models of Discourse// Structure of Language and Its Mathematical Aspects. Providence, Rhode Island: American Mathematical Society, 1961. – pp. 190-219.