УДК 81-114.2+81´32

ТЕМАТИЧЕСКАЯ ОРГАНИЗАЦИЯ ИНТЕРНЕТ-ЭГОИСТОРИИ В РАНГОВОМ РАСПРЕДЕЛЕНИИ ЦИПФА

А. С. Инфантьева, П. А. Катышев

THEMATIC ORGANIZATION OF INTERNET-EGOHISTORY IN THE ZIPF RANK DISTRIBUTION

A. S. Infantyeva, P. A. Katyshev

Статья посвящена тематическому анализу такого типа текстов, как Интернет-эгоистория, которая является разновидностью рассказа о себе, оформленного в текст личного Интернет-дневника. Тематический анализ проводится на основе метода рангового распределения Ципфа, согласно которому частота встречаемости слова в тексте обратно пропорциональна его рангу. Материалом для анализа являются 10 полных Интернет-эгоисторий.

Paper is devoted to a thematic analysis of Internet-egohistory, which is a kind of self-story and executed the text of a personal online diary. Thematic analysis is conducted on the basis of rank Zipf distribution, according to which the frequency of words in the text is inversely proportional to its rank. Material for analysis are 10 full Internet-egohistory.

Ключевые слова: Интернет-эгоистория, ранговое распределение Ципфа, тематическое поле, частотность слова.

Keywords: Internet-egohistory, Zipf rank distribution, subject field, the frequency of word/

В настоящее время для лингвистики характерен возрастающий интерес к изучению новых дискурсивных практик, одной из которых является Интернет-эгоистория. В отечественные филологические исследования термин «эгоистория» был введен Ю. Л. Троицким в значении сюжетного повествования о собственной жизни, имеющего свою интригу и не совпадающего ни с автобиографией, ни с автопсихологией [8, с. 58]. Одной из разновидностей рассказов о себе является текст личного Интернет-дневника – так называемого блога.

Целью данного исследования является выявление текстовых показателей, согласно которым в эгоисторическом тексте возможно выявить признаки состояния сознания и речемыслительной деятельности субъекта.

В качестве метода исследования выбран анализ рангового распределения Ципфа, необходимый для выявления закономерностей тематической организации Интернет-эгоисторий, созданных наркозависимыми.

Тематическая организация – это распределение тем в целостном тексте. Центральным понятием в исследовании тематической организации является тематическое поле. Лингвистика традиционно рассматривает его как лексико-семантическую группу слов, связанных парадигматическими отношениями [2].

По определению, данному А. Гурвичем и используемому в нарративном анализе в социологии, тематическое поле – это «совокупность данных, актуальных для определённой темы и являющихся фоном для её раскрытия» [11, с. 10].

Обобщая оба понимания, можно определять тематическое поле как сумму частей целого текста, репрезентирующих события или ситуации, представленные субъектом в отношении к определенной теме и образующие фон, на котором выделяется тема.

Цель анализа тематических полей (thematic field analysis) – исследование «механизмов временной и тематической организации нарратива», доказывающих, что «выбор рассказчиком эпизодов обусловлен контекстом его интерпретации всей своей жизни» [11, с. 10].

Из такого свойства нарративного текста, как вербализация актуальной информации непосредственно следует вывод, что степень выраженности той или иной темы прямо пропорциональна значимости идентификации субъекта с опытом, относящимся к ней.

Избирательность вербализации опыта в нарративе связана с тем, что актуализация объекта повествования «изоморфна самоактуализации субъекта» [8, с. 21]. Таким образом, организуя значимый опыт в тематические поля, субъект действует аналогично пациенту психотерапевта – переосмысляет значимые события через их вербализацию. И это особенно важно для субъектов, «испытывающих трудности в определении своей социальной идентичности» [9, с. 59].

Под индексами темы отдельного текста могут пониматься наиболее частотные существительные, интенсивность использования которых выражается в долях покрываемости всего текста. Для её вычисления используется метод рангового распределения Ципфа.

Согласно закону Дж. Ципфа, распределение слов естественного языка подчиняется закону, который можно сформулировать следующим образом: если к какому-либо достаточно большому тексту составить список всех встретившихся в нем слов, затем расположить эти слова в порядке убывания частоты их встречаемости в данном тексте и пронумеровать в порядке от 1 (порядковый номер наиболее часто встречающегося слова) до R, то для любого слова произведение его порядкового номера (ранга) в таком списке и частоты его встречаемости в тексте будет величиной постоянной, имеющей

примерно одинаковое значение для любого слова из этого списка. Аналитически закон Дж. Ципфа может быть выражен в виде $f_n = c$, где f – частота встречаемости слова в тексте; n – ранг (порядковый номер) слова в списке; c – эмпирическая постоянная величина [10].

Таким образом, относительная частота слова (F) в тексте обратно пропорциональна рангу слова (n), причём «выполнение данного рангового распределения – это признак "правильности" данного текста» [5, с. 9]. Принципиальные отклонения от нормы в ранговом распределении слов текста свидетельствуют об иных закономерностях организации текста.

Проблема связи частоты слова и его ранга соединяется с проблемой оценки лексического богатства текста или совокупности текстов. Таким образом, степень преобладания наиболее частотных слов в тексте обратно пропорциональна лексическому составу текста: чем чаще употребляется слово, тем меньше слов в тексте.

Число, показывающее сколько раз встречается слово в тексте, называется частотой вхождения слова. Если расположить частоты по мере убывания и пронумеровать, то порядковый номер частоты называется рангом частоты. Вероятность обнаружения слова в тексте равна отношению частоты вхождения слова к числу слов в тексте. Дж. Ципф определил, что если умножить вероятность обнаружения слова в тексте на ранг частоты, то получившаяся величина приблизительно постоянна для всех текстов на одном языке.

График зависимости ранга от частоты представляет из себя равностороннюю гиперболу. Ципф также установил, что зависимость количества слов с данной частотой от частоты постоянна для всех нормальных текстов в пределах одного языка и также является гиперболой. Исследование вышеуказанных зависимостей для различных текстов показали, что наиболее значимые слова текста лежат в средней части диаграммы, так как слова с максимальной частотой, как правило, являются предлогами, частицами и указательными местоимениями [5], а редко встречающиеся слова в большинстве случаев не имеют решающего значения.

Интенсивность использования наиболее частотных существительных равна покрываемости текста участком словаря (10 первых существительных в частотном списке) и обозначается Z_n. Таким образом, Z_n вычисляется как $F(x_{n1})+ F(x_{n2})\ldots+ F(x_n)$, где x – слово с рангом n. Частота слова в тексте, обозначаемая $F(x_n)$, является отношением количества словоупотреблений одного слова к объёму текста в словоупотреблениях.

Поскольку в настоящее время одной из ключевых проблем общества является наркомания [1, 4, 7], методы лингвистики могут быть использованы для повышения качества диагностики этого заболевания на ранних этапах. Кроме того, материал Интернет-эгоисторий является наиболее доступным и удобным для изучения состояния сознания наркозависимых.

Таким образом, материал Интернет-эгоисторий наркозависимых может выступать не только в качестве контрольной группы в данном исследовании, но и подтвердить возможность диагностирования наркозависимости на основании текста.

Материалом данного исследования являются пять полных нормальных (непатологичных) Интернет-эгоисторий и пять Интернет-эгоисторий наркозависимых личностей.

Анализ на основе закона рангового распределения Дж. Ципфа на материале Интернет-эгоисторий проводился в три этапа:

1) составление тезауруса каждого из исследуемых текстов;

2) распределение рангов для первых 10 по частотности лексем;

3) вычисление процентного соотношения словоупотреблений и словарного объёма текста и показателя Z_n.

На данном примере видно, каким образом проводился анализ:

04 марта 02:01

Знаешь тебя не хватает. Привыкла к тебе, нельзя так. Нехватает твоих прикосновений! Передать никак. Ты! зачем ты обещал, если знал что не сможешь? Или ты надеялся успеешь? Успеешь погладить меня, погладить как тогда. Чтобы был полет. Погладить. помнишь, ты поразил меня? Помнишь. Это всегда помнишь.

Помнишь, ломало? Доза выросла, ломки были не передать. Ты чувствуешь что надо, а сил нет. Ходишь по комнате без сил, а собраться выйти никак. Ты тогда сам удивился дозам, но поставил. сам лег. это и теперь остается полетом. Была вторая сильная волна прихода. Как передать... И полетела... Я чувствовала свое тело лежащее рядом с тобой и летела... ты лежал, я поняла ты ведешь меня. Времени нет. Я лечу и ты летишь. Мы вернулись. я лежала, боясь чувствовать не передаваемый полет, не помню, когда это кончилось зато помню полет.

Объём текста в словоупотреблениях составляет 116 единиц. Таким образом, Z_n для данного текстового фрагмента составляет 69,6 %. Индекс покрываемости Интернет-эгоистории, из которой взят данный фрагмент, составляет 27,4 %, и данный фрагмент текста является одним из самых плотных по количеству употреблений 10 самых частотных слов.

Контрольными данными для данного анализа являются показатели Z_n, полученные по результатам исследования беллетристических, эпистолярных и разговорных текстов, приведенного в коллективной монографии [6].

Z_n для беллетристических, эпистолярных и разговорных текстов удерживается в интервале от 0.93 до 1,45 % [6, с. 81]. По результатам данного исследования, в Интернет-эгоисториях Z_n варьируется от 4,5 до 27,4 %. Такой разброс обусловлен интенсивностью записей в Интернет-эгоистории: чем чаще добавляются записи, тем вероятней доминирование в тексте одной и той же темы.

В Интернет-эгоистории со средней интенсивностью записей 1\20,6 (одна запись за 20.6 дней) имеет Z_n = 4,5 %; при интенсивности 1\1,07 – Z_n = 27,4 %. Прямая зависимость Z_n от интенсивности следует из особенностей Интернет-эгоистории как типа дискурса: текст посвящен субъекту, тематически ограничен повествованием о его жизни и дискретен, т. е. состоит из добавляемых записей, малых по объёму.

При высокой интенсивности записей внимание субъекта сосредоточивается на определённой теме, волнующей его, поэтому Z_n в таком тексте превышает 15 % – максимально возможный Z_n в нормальных традиционных текстах [6, с. 81].

Тезаурус данного фрагмента, таким образом, составляет 71 лексему:

ты 11	прикосновение 1	удивиться 1
помнить 9	зачем 1	мой 1
полет 6	обещать 1	но 1
лететь 6	если 1	поставить 1
я 6	знать 1	оставаться 1
лечь 5	смочь 1	вторая 1
не 5	или 1	волна 1
передать 4	надеяться 1	приход 1
погладить 4	чтобы 1	свой 1
сила 4	поразить 1	тело 1
чувствовать 3	это 1	рядом 1
тогда 3	всегда 1	понять 1
и 3	ломать 1	вести 1
хватать 2	вырасти 1	время 1
это 2	ломка 1	нет1
быть 2	что 1	мы 1
а 2	надо1	вернуться1
сам 2	нет 1	боясь 1
был 2	ходить 1	когда 1
доза 2	по 1	кончиться 1
никак 2	комната 1	зато 1
успеть 2	собраться 1	к 1
как 2	так 1	нельзя 1
знать 1	привыкнуть 1	выйти 1

Тексты Интернет-эгоисторий наследуют отдельные свойства малых нарративных жанров. Как и в ранее исследованных малых текстах с объёмом слов порядка $N = 10^3$ и $N = 10^4$ [5, с. 15], у частотных слов Интернет-эгоисторий с объёмом $N = 10^1$ и $N = 10^2$ достаточно высокий индекс покрываемости, что обусловлено именно малым объёмом текста.

Ранговое распределение в текстах наркозависимых характеризуется ещё более высоким Z_n, что связано с особой ролью существительных при психопатологии речи: «По мере утяжеления психического заболевания доля именных единиц в патологическом тексте имеет тенденцию к увеличению» [6, с. 82]. Максимальный Z_n = 31,1 % в данном случае не показателен, так как близок к максимальному Z_n в нормальных Интернет-эгоисториях (27,4 %).

Минимальный Z_n = 14,8 % является очень высоким, что говорит об узкой тематической направленности Интернет-эгоисторий наркозависимых и подтверждает тенденцию к сведению «истории жизни к истории наркотизации» [9, с. 59] и нескольких ведущих мотивов к одному – влечению к психоактивным веществам [3, с. 327].

Ранговое распределение слов свидетельствует о том, что тематическая организация Интернет-эгоисторий характеризуется более высокой степенью покрываемости текста частотными существительными: их максимальный показатель покрываемости в два раза превышает аналогичный, описывающий ранговое распределение слов в традиционных нарративных малых жанрах.

Такое повышение индекса покрываемости сопровождается интенсивностью добавляемых записей и выдвижением на первый план какой-либо определенной темы. На этом фоне Интернет-эгоистории наркозависимых характеризуются максимально высоким Z_n, отличным от показателей нормальных текстов, что демонстрирует зацикленность субъекта на доминирующей теме – влечении к психоактивным веществам.

Литература

1. Глухарева, А. Н. Депрессивные идеи (клинико-семантический анализ речевого поведения при депрессивных расстройствах) / А. Н. Глухарева: автореф. дис... канд. мед. наук. – СПб., 2000.

2. Гореликова, М. И. Лингвистический анализ художественного текста / М. И. Гореликова, Ф. М. Магомедова. – М., 1983.

3. Елшанский, С. П. Семантика внутреннего восприятия при зависимостях от психоактивных

веществ (на модели опийной наркомании) / С. П. Елшанский. – М, 2004.

4. Микиртумов, Б. Е. Клиническая семантика психопатологии / Б. Е. Микиртумов. – М, 2007.

5. Орлов, Ю. К. Модель частотной структуры лексики / Ю. К. Орлов // Исследования в области вычислительной лингвистики и лингвостатистики. – М., 1976.

6. Пашковский, В. Э. Психиатрическая лингвистика / В. Э. Пашковский, В. Р. Пиотровская, Р. Г. Пиотровский. – М., 2009.

7. Польская, Н. А. Психопатология: от переживания к нарративу / Н. А. Польская: монография. – Саратов, 2004.

8. Троицкий, Ю. Л. Эгоистория / Ю. Л. Троцкий // Дискурс. – 1996. – № 1.

9. Фоломеева, Н. М. Наркомания как форма девиантного поведения / Н. М. Фоломеева, И. И. Шурышина, Н. А. Новикова, Т. В. Чешнева. – М., 1997.

10. Фрумкина, Р. М. Роль статистических методов в современных лингвистических исследованиях // Математическая лингвистика / Р. М. Фрумкина. – М., 1973.

11. Rosenthal, G. The Narrated Life Story: On the Interrelation Between Experience, Memory and Narration / G. Rosenthal // Narrative, Memory & Knowledge: Representations, Aesthetics, Contexts. – pub. University of Huddersfield, 2006. – URL: http://www2.hud.ac.uk/hhs/nme/books/2006/Chapter_1_-_Gabriele_Rosenthal.pdf (октябрь 2009).