ГИПЕРТЕКСТ, КОНТЕКСТ И ПОДТЕКСТ В ПОИСКОВО-АНАЛИТИЧЕСКОЙ СИСТЕМЕ “ГАЛАКТИКА-ZOOM”

HYPERTEXT, CONTEXT AND SUBTEXT IN SEARCH ENGINE “GALAKTIKA-ZOOM”

И.И.Богатырева *(bogatyreva@galaktika.ru),*

А.В.Антонов *(alexa@galaktika.ru),*

Е.С.Курзинер *(koorz@galaktika.ru)*

Корпорация “Галактика”, Москва

ПАС “Галактика-Zoom” осуществляет автоматическое выделение ключевых слов текстовой выборки, создает т.н. информационный портрет. Алгоритмически инфопортрет представляет собой самые характерные для запросного текста слова и словосочетания. Фактически инфопортрет – это **парадигматический контекст** запроса, или **гипертекст** выборки. Внутри этой информационной парадигмы образуются смысловые синтагмы с новым, не существующим в синтагматическом контексте, смыслом, или **подтекст.**

Поисково-аналитическая система “Галактика-Zoom” позволяет эффективно решать задачи поиска и анализа больших объемов информации в текстовых базах. Главная отличительная особенность данной системы заключается в том, что в случае каждого конкретного запроса происходит построение т.н. “информационного портрета” – т.е. выявление упорядоченных по значимости ключевых слов и словосочетаний, характерных как для данной выборки в целом, так и для каждого документа этой выборки. Подобная упорядоченность отражает ранг частотности ключевой темы (слова или словосочетания) выборки на фоне этой же темы в целой базе. При этом главные темы выборки (или отдельно взятого документа) не просто выстраиваются по частоте их встречаемости, а их отбор фактически характеризует отличие данной выборки (или документа) от всех остальных документов текстовой базы, т.е. мы получаем портрет интересующего нас объекта *на фоне* всего остального, и этот портрет как информационно, так и лексически представляет собой *парадигматический контекст запроса.* Внутри полученной информационной парадигмы образуются некие смысловые синтагмы с новым (или несколько иным) смыслом, т.е. мы получаем своего рода *подтекст* запрашиваемого текста.

В настоящее время вышеназванная система используется, в основном, как удобный инструмент поиска и анализа текстов СМИ или же какой-то специфической документации определенной организации или компании. В рамках данной статьи мы хотим показать и другие возможности этого инструмента, попробовав применить его для исследования художественных текстов.[1]

Как известно, всякий художественный текст является чем-то цельным, но одновременно и многоструктурным, причем цельность художественного текста предполагает множественность его интерпретаций и, следовательно, является источником множества потенциальных структур. Анализируя и интерпретируя текст, мы фактически осуществляем переход от его линейного пространства к нелинейному - семантическому. Художественный текст содержит не только сообщения в их явной форме, но и нечто, находящееся вне этих рамок, - некий смысловой довесок, именуемый в лингвистической литературе *подтекстом*. Среди лингвистов и литературоведов до сих пор нет единства мнений ни о статусе, ни о типологии подтекста, да и само его определение не имеет однозначной интерпретации. Но в любом случае нет сомнений в том, что подтекст – это вполне конкретная реальность, в большинстве случаев сознательно запрограммированная автором текста. В подтексте с наибольшей очевидностью проявляется взаимозависимость и системность всех составляющих элементов текста. Каждый элемент текста как бы нанизывается на две оси координат: *контекст* и *подтекст.*

[1]Подобного рода исследование уже проводилось: с помощью представляемой поисково-аналитической системы был проделан сравнительный анализ лексического состава и стилистических особенностей произведений Н.В.Гоголя, Л.Н.Толстого, Ф.М.Достоевского, А.П.Чехова и М.А.Булгакова (см. материалы “Диалога-2002”, том 2 “Прикладные проблемы”, статья А.В.Антонова и Е.С.Курзинер “Автоматическое определение тематики большого необработанного текстового массива”).

В литературе можно встретить точки зрения, согласно которым подтекст можно рассматривать как явление *семантическое, прагматическое* и как факт *формальной* структуры текста.

В нашем исследовании мы будем понимать подтекст как явление семантическое. В работах, где представлено подобное понимание подтекста, данный термин часто используется как дублет термина *смысл*. Причем, некоторые авторы им обозначают не всякий смысл, а лишь тот, который рассчитан на понимание посвященных, избранных, т.е. *эзотерический* смысл. Многие исследователи понимают подтекст более широко – как не выраженное словами (глубинное, или дополнительное) значение. Приведем несколько определений подтекста в рамках данного подхода.

Подтекст - это сознательно или бессознательно создаваемая говорящим часть семантической структуры текста, доступная восприятию в результате особой аналитической процедуры, предполагающей переработку эксплицитной информации и вывод на ее основе дополнительной информации.[2]

Подтекст - скрытый, отличный от прямого значения высказывания смысл, который восстанавливается на основе контекста с учетом ситуации. В театре подтекст раскрывается актером с помощью интонации, паузы, мимики, жеста.[3]

Подтекст - не выраженный явным образом, отличный от непосредственно воспринимаемого при чтении фрагмента текста смысл, восстанавливаемый читателем (слушателем, адресатом) на основании соотнесения данного фрагмента текста с предшествующими ему текстовыми фрагментами как в рамках данного текста, так и за его пределами – в созданных ранее текстах (“своих” или “чужих”).[4]

Таким образом, мы подходим к еще одному пониманию подтекста – *интертекстуальному*. Соотнося два текстовых фрагмента из разных текстов, К.Ф.Тарановский подтекстом называет *ранее существовавший текст, отраженный в данном*. Тарановский вывел такое понимание подтекста при изучении литературы эпохи модернизма в книге 1976 г. “Очерки о Мандельштаме”. Понимаемый таким образом подтекст, по мнению Б.Гаспарова, выполняет интегрирующую функцию в тексте, в который он инкорпорирован: он позволяет увидеть подразумеваемые смысловые мотивировки, объясняющие связь между отдельными элементами текста, до того казавшимися соположенными случайно; в конечном счете текст после осознания присутствующих в нем подтекстов предстает для нас более связным и осмысленным.[5]

В настоящем исследовании мы понимаем подтекст в русле семантических и интертекстуальных теорий, причем не сужая его до некоего эзотерического смысла, доступного лишь посвященным, но допуская, что в ряде случаев и такое понимание подтекста оказывается допустимым и имеющим право на существование.

Следует сразу оговорить и наше понимание ещё одного термина, который будет использоваться в нашей работе и который также понимается и определяется неоднозначно. Речь идет о термине *гипертекст***.** Согласно классическому определению, данному Теодором Нельсоном, гипертекст – это форма письма, которое ветвится или осуществляется по запросу, это как бы нелинейное письмо. Гипертекст представляет собой крайне расплывчатое и тем не менее широко используемое в современной литературе понятие. Этим термином могут обозначить Интернет, энциклопедию, справочник, т.е. любой текст, в котором обнаруживаются какие-либо ссылки на фрагменты из других текстов. В принципе под термином *гипертекст* может пониматься не только *текст*, организованный по-особому, но и *метод* объединения нескольких документов, *механизм*, позволяющий эти документы определенным образом организовать, *форму* организации материала и т.п. *Гипертекст* – это одновременно и процесс, и результат этого процесса, в то время как *текст* в его традиционном понимании – это всё-таки именно результат. Мы понимаем *гипертекст* как особый тип текста, который устроен таким образом, что он представляет собой некоторую *систему*, и даже *иерархию текстов*, представляющих собой одновременно и единство, и множество текстов.

Перейдем теперь непосредственно к самому исследованию и его результатам. Настоящее исследование проводилось по двум базам – базе литературных текстов, где представлены как художественные произведения разных авторов и жанров (классические тексты русской и переводной зарубежной поэзии и прозы, приключенческая литература, фантастика, фэнтези и т.п.), так и философские сочинения различных школ и направлений, публицистика и др., и базе, где представлены тексты из СМИ (причем, эта база ежедневно пополняется).

Мы поставили перед собой следующую задачу: попытаться проанализировать тексты из разных баз при помощи поисково-аналитической системы “Галактика-Zoom”, задавая в качестве текста запроса достаточно известное выражение, про которое мы точно знаем, кто, когда и в каком конкретном контексте его произнес и какой изначальный смысл в него вложил. Поскольку ПАС “Галактика-Zoom” не просто находит тексты, где встречается фраза из нашего запроса, но и выдает в качестве результата их анализа ключевые, наиболее значимые слова и словосочетания (типа “определение” + “определяемое слово”) – т.н. информационный портрет – найденной выборки (1), каждого найденного документа (2) и выстраивает эти документы в порядке максимального соответствия их индивидуального информационного портрета портрету всей выборки (3), мы

[2] http://www.ruthenia.ru/annalystxt/Podtxt.htm
[3] Большой энциклопедический словарь.
[4] http://www.krugosvet.ru/
[5] Б.Гаспаров. В поисках другого. - НЛО, №14, 1996.

получаем возможность увидеть следующую (как выяснилось, в ряде случаев довольно любопытную, хотя и вполне ожидаемую и объяснимую) картину:

1) наш информационный портрет дает совершенно другое понимание известной фразы из нашего запроса;

2) один и тот же запрос, проведенный по разным базам, дает нам совершенно разные информационные портреты;

3) в полученном информационном портрете соединяются как исконные, так и абсолютно новые смыслы.

Полученный информационный портрет, внутри которого объединяются выделенные смысловые синтагмы, в ряде случаев несет определенный, иногда достаточно неожиданный новый смысл, *подтекст* наших найденных текстов. И таким образом сама запросная фраза обрастает новыми смыслами и несет теперь в себе то, что даже не предполагалось в исходном *контексте* первоисточника: новые *контексты* её употребления заставляют читателя воспринимать её иначе, чем в тексте-источнике. Любопытной иногда оказывается и сама выборка текстов, фактически представляющая собой *гипертекст*, объединяющий порой неожиданные и разнородные тексты, и дающая читателю возможность увидеть или выстроить такие связи, которые до этого ему даже не могли прийти в голову.

Продемонстрируем всё вышесказанное конкретными примерами. Введем в качестве фразы для поиска цитату из “Фауста” Гете *“вечная женственность”* (поиск по литературной базе). Полученная картина оказалась достаточно целостной: первые по значимости 15 текстов, в которых обнаруживается заданная нами фраза, являются в большинстве своем философскими трудами (Н. Бердяева, В.Соловьёва, Д. Андреева и др.), есть среди них и “Фауст” Гете, а также лирика Блока, обращенная к идеалу Прекрасной Дамы и вечной женственности. В середине и ближе к концу выборки встречаются и труды по психологии (З.Фрейда, К.Г.Юнга). Остальные тексты в нашей выборке – художественные.

Напомним, в каком контексте употребляется эта фраза у Гете. В переводе Б. Пастернака вторая книга “Фауста” заканчивается следующими строками:

> Всё быстротечное – символ, сравнение,
> Цель бесконечная – здесь, в достижении.
> Здесь заповеданность истины всей.
> Вечная женственность тянет нас к ней.

Анализ информационного портрета выборки текстов, где встречается данная цитата, дает интересные результаты. С большим отрывом в рейтинге[6] на первом месте находится словосочетание *сексуальный акт* (рейтинг 113, 87). За ним следуют:

мистический – рейтинг 32,18,
сексуальный – рейтинг 29,25,
культ – рейтинг 23,33,
акт– рейтинг 22,41,
божество, мужской, женственный, творческий, религиозный, божественный, духовный и т.п. Таким образом, “вечная женственность” предстает перед нами в сексуально-мистическом контексте. Женщина при этом видится как *дева, прекрасная дама, красота, божество*; слова *Богородица, Мария* и *Христос* отсылают нас к сакральной трактовке образа женщины. При этом с ними как минимум одинаковы по значимости, а в ряде случаев и гораздо более значимы в инфопортрете слова *сексуальный, акт, влечение, страстный, поцелуй, грех* и т.п., что говорит о чрезвычайно важной роли сексуального подтекста в образе Вечной Женственности. Итак, изначально метафизический, предельно абстрактный смысл данного понятия конкретизируется в литературных и философских сочинениях, приобретает новые очертания и наполняется новым мистико-эротическим смыслом.

Достаточно интересную картину в ряде случаев даёт сопоставление результатов одного и того же запроса по разным базам. В качестве примера возьмем фразу *“без гнева и пристрастия”* и проведем такой сопоставительный анализ. Полученные информационные портреты оказались довольно любопытными. При анализе литературной базы был получен следующий информационный портрет 26 документов, где встретилась данная фраза:

весь ход – рейтинг 154,20
цитата – рейтинг 118,78
возмездие – рейтинг 102,65
Сталин – рейтинг 86,04
германский – рейтинг 85,51
Гитлер – рейтинг 68,70
виновный – рейтинг 66,87

[6] Рейтинг (вес, коэффициент значимости) – это величина, показывающая соответствие данного слова или словосочетания тематике данной выборки относительно всех других выборок. Более подробное описание как самого понятия рейтинга, так и строгая математическая формула, по которой определяется коэффициент значимости, или вес слов и словосочетаний соответствующего инфопортрета, представлены в работах: А.В.Антонов, В.С.Мешков. Современные проблемы поисковых систем и некоторые пути их преодоления. – Сер. Аналитика – Капитал. М., 2000; А.В.Антонов. Информационно-поисковая система Galaktika-ZOOM с элементами анализа на гипермассивах информации. - Сб. ВИНИТИ №8, 2001.

чудовищный – рейтинг 52,04
Европа – рейтинг 51,77
немецкий – рейтинг 42,93
исследование – рейтинг 39,61 и т.д.

Налицо желание и попытка рефлексировать на тему тоталитаризма и империализма 30-х-40-х годов XX века, осмыслить происходящее в нашей стране и в Европе, проследить противостояние нацистской и коммунистической идеологии и результаты этого противостояния. Слова *виновный, чудовищный, возмездие* свидетельствуют о стремлении к оценке этого периода, одного из самых коротких и при этом достаточно значимых в мировой истории. При этом вслед за Тацитом литераторы пытаются дать максимально объективную оценку (см.: *весь ход, исследование, материал*) наиболее спорному и неоднозначному периоду новейшей истории человечества, который и сейчас оказывает влияние на общественные отношения и мало кого оставляет равнодушным. Представляются любопытными не только *контекст*, в котором встретилась данная фраза в литературной базе, или же вполне конкретный *подтекст,* за ней стоящий, но и полученный *гипертекст* – сама выборка литературных текстов. Подавляющее большинство в ней составляют не философские, не публицистические и даже не художественные тексты, где речь идет о данном периоде в нашей истории (как, например, романы Ю.Семёнова), а боевики (напр., А.Быстрова), фантастические романы (С.Лема и др.), рассказы и романы, относящиеся к жанру фэнтези (напр., книги Н.Перумова и т.п.).

Посмотрим теперь, какой информационный портрет и какую выборку документов дает нам этот же запрос по базе СМИ. Основные блоки главных тем этой выборки такие.

1). Слова и словосочетания, характеризующие современную политическую или социальную ситуацию: *социальная революция, политическая система, чудовищная конституция, реакционные губернаторы, административная вертикаль, унитарное государство, силовые структуры, демократические ценности.*

2). Слова и словосочетания, описывающие проблемы, связанные с религиозной и межнациональной враждой: *радикальный ислам, национальные конфликты, Приднестровье, Карабах, священная корова*.

3). Слова и словосочетания, связанные с понятием собственности: *тотальная приватизация, одна комната, государственная собственность*.

В этом информационном портрете мы даже при большом желании не увидим стремления современных журналистов осмыслить прошлое страны и влияние этого прошлого на современность. Все внимание СМИ приковано к различным текущим проблемам. Большую значимость в этом портрете имеют слова *демократический, реакционный, революционный, радикальный*, что, по всей видимости, указывает на стремление журналистов дать оценку ситуации в стране в *настоящее* время. Пресса, в отличие от литературы, словно скользит по поверхности происходящих в социуме событий и, не видя (и даже не желая и не пытаясь видеть) их глубинных корней, стремится проследить лишь внешние, очевидные причинно-следственные связи.[7] И даже фраза, обращенная изначально к историческим изысканиям, в СМИ приобретает будничную окраску и адресуется лишь к повседневности; употребленная в этом контексте, она утрачивает свои корни.

Впрочем, подобные результаты дают и другие запросы, проведенные по базе СМИ. Так, информационный портрет библейской цитаты *"блажен, кто верует"* выглядит весьма показательно. Проиллюстрируем это на примере части полученного списка главных тем документов, где употреблялась данная фраза:

божия мать – рейтинг 146,76
Христос – рейтинг 29,23
молитва – рейтинг 17,46
демократия – рейтинг 17,15
религия – рейтинг 16,87
оппозиция – рейтинг 16,72
потрясение – рейтинг 16,02
церковь – рейтинг 15,82
парламентские выборы – рейтинг 13,96
вера – рейтинг 12,92
социалистический – рейтинг 12,29
коррупция – рейтинг 10,34
демократический – рейтинг 9,02
молиться – рейтинг 8,76
социализм – рейтинг 8,55 и т.д.

Как мы видим, взятая из сакрального текста и определяющая отношения человека с Богом, в прессе эта фраза употребляется в двух основных контекстах, равных (!) по значимости, - религиозном и политическом. В одном ряду, практически с одним и тем же рейтингом, стоят *молитва* и *демократия, религия* и *оппозиция, церковь* и *парламентские выборы* и т.п. Таким образом, изначально сакральная и однозначно понимаемая фраза

[7] Следует особо отметить, что речь идет *исключительно* о тех документах, которые оказались в нашей выборке по конкретному запросу. Всё вышесказанное ни в коей мере не претендует на далеко идущие обобщения относительно характера и содержания текстов современных журналистов.

в нашей выборке приобретает двойную контекстуальную окраску и словно рисует нам две грани веры современного общества: веру в Бога и веру в силу государства и правительства, веру в политические начинания.

Поскольку задачей данного исследования было продемонстрировать некоторые *возможности* использования поисково-аналитической системы "Галактика-Zoom" при анализе художественных текстов, исходя из того, что вышеназванная система, как и любой другой инструмент, основанный на строгих математических законах, сможет проверить (и соответственно подтвердить или опровергнуть) наши субъективные впечатления, связанные с восприятием и пониманием литературных произведений, мы попробовали проанализировать нашим инструментом ряд известных романов и сопоставить полученный автоматическим образом беспристрастный инфопортрет с нашим восприятием и трактовкой этих же источников. Посмотрим на "верхушку" выданного нам информационного портрета романа "Мастер и Маргарита":

Маргарита – рейтинг 79,69
Воланд – рейтинг 48, 55
прокуратор – рейтинг 44,30
Коровьев – рейтинг 42,12
Азазелло – рейтинг 34,07
Пилат – рейтинг 28,54
Берлиоз – рейтинг 27,79 и далее (выборочно в порядке их значимости)
Иван, кот, Левий, мастер, мессир, Иешуа, Фагот, Патриарший пруд, Иуда, пятый прокуратор, арестант, Лысая гора, трамвай, луна, бал, сеанс, кровавый подбой, клетчатый и т.д.

Была проведена кластеризация элементов инфопортрета[8], и мы получили следующие результаты (инфопортреты выделенных кластеров приводятся выборочно):

Азазелло, Степа, Воланд, кот, Коровьев, Маргарита;
Прокуратор, Пилат, Иуда;
Иван, Стравинский, Берлиоз, кот, Воланд, Варенуха;
Варенуха, Воланд, администратор, Римский;
Пилат, первосвященник, Каифа, Маргарита, заговорить, игемон, арестант;
прокуратор, казнь, Пилат, заговорить, зарезать;
Берлиоз, Азазелло, Максимилиан, сумасшедший, казнь;
палач, Левий, передний, столб;
Берлиоз, Аннушка, профессор, Иисус, неизвестный, Иванович, поэт;
Берлиоз, Воланд, покойный, телеграмма, жилец, Римский, профессор.

Представляется, что вышеприведенная картина выглядит достаточно убедительной, т.к. реально отражает те связи и отношения, которые внимательный читатель обнаружит в тексте романа М.А.Булгакова. Мы же можем констатировать тот факт, что наша система справилась с поставленной перед ней задачей и грамотно выделила как *ключевые слова* данного текста, так и его *основные сюжетные линии* (фактически именно они даны в инфопортретах разных кластеров), и это позволяет нам надеяться, что и другие исследования (см. примеры, приведенные выше) являются достаточно достоверными и объективными.

Подведем итоги нашего эксперимента. Как нам кажется, информационный портрет текстов, выдаваемый поисково-аналитической системой "Галактика-Zoom" по определенному запросу, в ряде случаев обнаруживает то, что не лежит на поверхности, но представляет собой скорее скрытую, извлекаемую путем истолкования информацию. Благодаря нахождению рядом в информационном портрете главных тем выборки (т.е. как бы в новом контексте), иногда формируется совершенно неожиданный подтекст либо какого-то текста, либо запросной фразы, приобретающей новые (или скрытые?) смыслы (см., например, вышепроведенный анализ результатов запроса "*блажен, кто верует*"). Как мы видели, при этом происходит своего рода наращение смыслов слов или фраз, или же актуализация их скрытых смыслов, что создаёт новое видение текста и его оценку, углубляет наше о нем представление, а иногда и *пере-* или *по*ворачивает его, выявляя неожиданные грани и оттенки, создавая (или проявляя?) смысловую многоплановость и объёмность. Безусловно, мы отдаем себе отчет в том, что ПАС "Галактика-Zoom" - это *инструмент*, а не разум: *окончательные* выводы о результатах объективного исследования, проведенного с помощью точного инструмента, должен сделать человек.

[8] Необходимо пояснить, что в нашей базе литературные тексты загружены не как отдельные документы, а постранично. Так, роман представляет собой некоторое *множество* документов, и кластеризация проводилась как в отношении этих *документов*, так и в отношении *ключевых слов* полученного инфопортрета. Множество документов, составляющих роман, было разбито на кластеры по общности тем, и особенность нашего алгоритма кластеризации состоит в том, что базовыми единицами для проведения кластеризации являются слова и словосочетания, составляющие инфопортрет документов. Формальное описание данного алгоритма представлено в тексте доклада, принятого для представления на конференции "Диалог 2006": А.В.Антонов, С.Г.Баглей, В.С.Мешков, А.В.Суханов "Кластеризация документов с использованием метаинформации".