Санкт-Петербургский государственный университет

Филологический факультет

Кафедра математической лингвистики

КУРСОВАЯ РАБОТА

на тему:

Опыт стилистической диагностики

текстовых источников Интернета

Выполнил:

студент II курса

Грановский Дмитрий

СПб 2006

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

1. Стилистическая диагностика текста и исследование языка интернета

2. Характеристика исследуемого материала: веблоги и форумы

2.1 Веблоги

2.2 Форумы

3. Методика исследования

3.1 Методика анализа веблогов

3.2 Методика анализа форумов

4. Исследуемые параметры текстов

5. Результаты исследования

6. Некоторые комментарии к результатам исследования

Заключение

Научная литература

СЛОВАРИ

#

# ВВЕДЕНИЕ

Классическое языкознание определяет язык как средство реализации познания окружающего мира, средство формирования представлений и идеалов на основе собранного познавательного материала и средство сохранения данной информации во всей ее совокупности в целях обмена опытом как в синхроническом, так и в диахроническом процессах.

Главнейшая функция языка – его коммуникативный аспект. Определенное сообщество, говорящее на том или ином языке, постоянно и неизбежно формирует его. При этом имеет место и обратное явление, т.е. язык и общество взаимодействуют, определяя динамику развития друг друга.

В последние 10 лет появилась и развилась особая среда языковой коммуникации – Интернет, ставший совершенно новым явлением в социолингвистике. Можно с достаточной уверенностью говорить о том, что в ближайшие годы Интернет станет объектом активных исследований в самых разных областях лингвистики в силу своей новизны, значительного количества материала и его качественного разнообразия.

В Интернете активизируются такие виды речевой деятельности, как письмо (в его компьютерно-печатной разновидности) и чтение, однако, несмотря на отсутствие говорения и аудирования в их привычном, традиционном понимании, «в сетевой сфере функционирования наблюдается тенденция к реализации письма – как говорения и чтения – как аудирования» [Трофимова 2005].

Интернет в России послужил, помимо прочего, индикатором языковой грамотности массовой аудитории. Наработанность грамматических навыков и орфоэпических умений проверяется именно в спонтанной речи. До Интернета спонтанная речь проявлялась, прежде всего, в своей устной форме, которая была одномоментной. В Интернете спонтанная разговорная устная речь неизбежно должна быть зафиксирована в письменной (а точнее, в печатной) форме, которая является, по сути, ее письменной расшифровкой. Следовательно, все речевые и языковые ошибки также фиксируются на письме. Формула «личность = текст или личность = язык» становится в Интернете аксиомой, так как все речевые индивидуальные особенности оказываются на виду в буквальном смысле слова. Так виртуальная личность реализует свой определенный стиль жизни, прежде всего и именно в стиле употребления языка.

# 1. Стилистическая диагностика текста и исследование языка интернета

В настоящее время учение о функционально-стилевой дифференциации литературного языка в достаточной степени разработано. В первую очередь его становление связано с именами Г.О. Винокура, В.В. Виноградова, Б.Н. Головина, М.Н. Кожиной и др. Основополагающим понятием здесь является понятие функциональных стилей как типов функционирования языка. Экстралингвистические факторы, условия, в которых используется язык, влияют на такие его параметры, как частотность употребления тех или иных лексем, речевых средств, доминирование грамматических конструкций определенного типа. Это и формирует систему функциональных стилей.

Функционально-стилистическая дифференциация языка происходит по формам речи (письменная и устная) и ее видам (монологический и полилогический). Принято выделять 5 базовых функциональных стилей: деловой, научный, публицистический, художественный и разговорный. Для каждого из них существует сложившаяся сфера применения (например, для делового стиля это разнообразная документация), основная функция (сообщение, воздействие, фиксация результатов познания мира), доминанта (точность, оценочность и т.д.) и масса статистических параметров, в соответствии с которыми и производится дифференциация. Понятно, что границы между стилями строго провести невозможно, и многие языковые явления находятся на стыке двух или более стилей. Так, научно-популярный стиль, выделяемый некоторыми исследователями из научного, обладает некоторыми признаками публицистического стиля, что соответствует его функции, а именно – донести сугубо научные факты до широкого круга читателей.

Появление Интернета, на наш взгляд, ставит под сомнение актуальность и полноту классической системы функциональных стилей и форм речи. В пользу этого можно привести следующие аргументы:

1. Монолог и полилог в рамках текстовых источников Интернета кажутся вполне равноправными, что характерно для разговорного стиля и в какой-то мере для публицистического. Монологический вид письменной речи представлен в данном исследовании веблогами, полилогический – форумами (веб-конференциями). Наряду с форумами можно выделить чаты (chatrooms).
2. Не сложилась какая-либо форма устной речи, что не характерно ни для разговорного стиля, ни для публицистического. Этот аргумент, однако, в ближайшие годы может быть опровергнут развитием подкастинга (podcasting) – своего рода устного аналога веблогов – и систем голосового общения через Интернет, массовое использование которых на сегодняшний день ограничено пропускной способностью и качеством каналов связи.
3. Ни один из существующих стилей не предусматривает в своей письменной форме наличия одновременно монолога и диалога, книжных и разговорных форм, целенаправленных орфографических ошибок и упрощений, псевдографики.

В ходе проведения исследования было выдвинуто 3 гипотезы:

1. Язык Интернета – самостоятельная форма существования языка, находящаяся на стыке письменной и устной.
2. Язык Интернета определяет собственный функциональный стиль, не являющийся разновидностью ни одного из существующих пяти стилей.
3. Функциональный стиль языка Интернета обладает некоторыми уникальными особенностями, но, в сущности, вписывается в существующую систему стилей, вероятнее всего, как вторая (после эпистолярной) письменная разновидность разговорного стиля.

Для проверки истинности какой-либо из гипотез, на наш взгляд, требуется обширный анализ текстовых Интернет-источников, выделение релевантных статистических параметров и сравнение их значений со значениями тех же параметров для текстов существующих функциональных стилей. Такие глобальные задачи данная работа перед собой не ставит. Мы ограничимся попыткой выработки методики создания репрезентативных выборок из исследуемых источников и критериев для их анализа.

# 2. Характеристика исследуемого материала: веблоги и форумы

## 2.1 Веблоги

Онлайн-дневники, или веблоги (weblogs), – сравнительно новое явление в мировом и русскоязычном Интернете.

Блог – сетевой дневник одного или нескольких авторов, состоящий из записей в обратном хронологическом порядке. Слово произошло от английского weblog – «веб-журнал». Совокупность блогов называется блогосферой.

Общая концепция веблогов такова: при регистрации пользователя (user) для него создается персональный раздел на сайте, в котором он может оставлять сообщения (entries). Также у пользователя существует так называемая «лента друзей», или «френдлента» (friends), представляющая собой список последних сообщений нескольких выбранных им других участников.

Для пользования такими сервисами не требуется специальных знаний в области веб-программирования и языков разметки гипертекста, что обусловливает высокие темпы развития онлайн-дневников. В настоящее время ежесуточное количество сообщений в русскоязычных веблогах можно оценить как приближающееся к 100 000 (включая комментарии). Крупнейшим блог-хостером на данный момент является открытый в 1999 г. «Живой журнал» (LiveJournal, <http://www.livejournal.com>), число пользователей которого в апреле 2006 г. превысило 10 миллионов. Около трети сообщений на русском языке публикуется именно там. В то же время существуют более или менее крупные системы онлайн-дневников исключительно на русском языке, среди которых можно выделить LiveInternet (<http://www.liveinternet.ru>). Данная работа опирается на материал «Живого журнала» в силу, во-первых, удобства составления выборок (см. методику анализа), и, во-вторых, большего числа пользователей, что по умолчанию обеспечивает большую репрезентативность.

## 2.2  Форумы

Форум, или веб-конференция, определяется как место сетевого общения нескольких авторов, где записи объединены общей темой. Набор записей, соответствующий какой-либо теме, получил название «ветки» (thread) или «топика» (topic). Форумы существуют в Рунете (русскоязычной части Интернета) гораздо дольше, чем веблоги. Оценить общее количество конференций в нем не представляется возможным, поскольку форум, как правило, имеется на каждом, даже не очень крупном сайте.

Для анализа мы выбрали два форума по следующим принципам:

* форумы должны быть крупными (как по числу зарегистрированных пользователей, так и по количеству ежесуточных сообщений);
* конференции должны быть разной тематики или не иметь ее вовсе;
* контингент пользователей форумов должен по возможности различаться (этот критерий обусловлен тем, что исследование противопоставляет не только форумы веблогам, но и форумы друг другу).

Форум №1 – форум студенческого сайта СПбГУ (<http://www.spbgu.ru/forums/>). Открыт весной 2003 г., число зарегистрированных пользователей – около 6900, число активных тем ежесуточно – 200-300. Форум позиционируется как самый крупный молодежный форум России.

Форум №2 – форум на <http://www.ixbt.com> – сайте, посвященном компьютерам (для исследования был выбран раздел форума «Общий», где обсуждаются темы, с компьютерами не связанные). Открыт в 2000 г., число зарегистрированных пользователей – около 48500, число активных тем ежесуточно – 600–700. Форум считается одним из крупнейших в Рунете.

# 3. Методика исследования

## 3.1 Методика анализа веблогов

Система веблогов позволяет пользователям оставлять свои комментарии к исходному сообщению. Однако мы считаем, что в таком случае совокупность этих комментариев практически не отличается от форума и, таким образом, не должна быть включена в наш анализ. Поэтому выборки делались исключительно из исходных сообщений авторов веблогов.

Очевидно, что целям работы удовлетворяет случайная выборка из веблогов русскоязычных пользователей. Можно считать, что с увеличением объема этой выборки ее репрезентативность также повышается. Таким образом, необходимо обеспечить:

* случайность выбора автора веблога;
* достаточный объем выборки.

В силу того, что возможность выбора случайного автора и случайной записи стандартными средствами портала не предусмотрена, отбор материала происходил следующим образом.

В русскоязычной части livejournal.com существует особое сообщество all\_ru («Френдлента ЖЖ-юзеров России», <http://all-ru.livejournal.com>), созданное исключительно для ведения френдленты всех пользователей, у которых в личных настройках страны установлена Российская Федерация и которые писали сообщения за последние несколько десятков минут. Соответственно, в каждый момент времени лента друзей данного сообщества представляет собой срез с совокупности недавних сообщений на русском языке. Если мы сделаем несколько таких срезов, каждый из которых содержит 50 сообщений, в разное время, то полученная выборка будет, несомненно, удовлетворять обоим вышеприведенным критериям и становиться все более релевантной по мере увеличения числа срезов.

## 3.2 Методика анализа форумов

Поскольку механизм форумов не содержит функции поиска случайного сообщения или случайной темы, мы ограничились тем, что из списка активных топиков (т.е. тех, в которых были сообщения за текущие сутки) случайным образом выбирались темы, из которых далее бралась первая страница, содержащая от 20 до 30 сообщений. Выбор только первой страницы обусловлен тем, что некоторые топики могут достигать значительной длины (на форуме iXBT.com нередки темы длиной более 100 страниц).

Для удобства общения на форумах принято цитировать то сообщение, на которое пишется ответ. Такие цитаты нами не учитывались, анализировался только авторский текст.

# 4  Исследуемые параметры текстов

1. Объем выборки – суммарное число сообщений, проанализированное в рамках исследования.
2. Репрезентативность – количество сообщений, несущих смысловую нагрузку, содержащих текст на русском языке. В рамках этого критерия из выборки исключались сообщения, содержащие:
	* только изображения (в т.ч. «смайлики»);
	* только ссылки;
	* только стихотворный текст;
	* текст на языке, отличном от русского;
	* результаты разнообразных онлайн-тестов.
3. Число словоупотреблений в выборке.
4. Среднее количество предложений в одном сообщении.
5. Среднее количество слов в одном сообщении.
6. Среднее количество слов в одном предложении.
7. Среднее число орфографических ошибок (в одном сообщении и на 1000 словоупотреблений). Явные опечатки (пропуски букв, замена одной буквы на другую, находящуюся рядом на стандартной клавиатуре) не учитывались. То, что мы ограничились лишь орфографическими ошибками, можно считать вполне рациональным, поскольку лексические ошибки достаточно полно характеризуют общую грамотность пишущего, а анализ других типов ошибок (синтаксических, словообразовательных, морфологических и др.) значительно усложнил бы исследование.
8. Среднее число орфографических упрощений (в одном сообщении и на 1000 словоупотреблений). Упрощением будем считать такую форму, которая формально является ошибочной, но при этом используется пользователем сознательно в силу, как правило, более краткой и приближенной к орфоэпической норме.
9. Среднее число словоупотреблений сленга (в одном сообщении и на 1000 словоупотреблений). Принимался во внимание как специализированный Интернет-сленг (например, широко известная аббревиатура ИМХО, происходящая от английской конструкции In My Humble Opinion — «по моему скромному мнению»), так и общераспространенный («крутой», «свалить», «тусоваться» и т.д.).
10. Среднее число словоупотреблений табуированной лексики (в одном сообщении и на 1000 словоупотреблений).

Несмотря на то, что подсчет распространенности каждой из четырех категорий лексики был проведен по двум параметрам (на 1 сообщение и на 1000 словоупотреблений), этого недостаточно. Предположим, что в нашей выборке 100 сообщений: 99 – коротких, но без ошибок, а последнее – длинное и со 100 ошибками. Опираясь только на первые два параметра, мы получим среднее число ошибок – 1 на сообщение, из чего, вероятно, составим неверное представление о грамотности в среде пользователей того или иного сервиса. Между тем, в такой выборке грамотность будет весьма высока. В связи с возможностью такой неверной интерпретации введем еще один числовой критерий – отношение числа сообщений, содержащих ошибки, к общему числу сообщений (репрезентативных). Назовем этот критерий встречаемостью.

Приведем некоторые комментарии к перечню параметров.

Параметры 1–2 показывают количество проанализированных структурных единиц текста (являющихся в целом нестандартными). Это, соответственно, отдельные сообщения в веблогах и веб-конференциях.

Параметр 3 характеризует величину выборки и является стандартным в большинстве подобных исследований.

Параметры 4–6 иллюстрируют объем сообщения и длину предложений внутри него, которая в общем случае коррелирует с их распространенностью.

Параметры 7–10 имеют непосредственное отношение к стилистике текста, причем логично предположить, что чем выше значение каждого параметра, тем более рассматриваемый текст является «неформальным», т.е. в рамках существующей системы приближенным к разговорному стилю.

Заметим, что 12 тыс. словоупотреблений – достаточно небольшая выборка, особенно по сравнению с национальными корпусами, однако «повседневные линейные языковые явления распределяются в текстах сравнительно стабильно, и их можно достоверно установить в относительно коротких текстовых сегментах (уже в объеме 1000 слов)» [Шимкова 2004].

В отношении орфографических ошибок, упрощений, сленговых и табуированных выражений автор в основном руководствовался собственным знанием русского языка и четырехлетним опытом веб-коммуникации, в сложных случаях прибегая к помощи русского орфографического словаря РАН под ред. В.В. Лопатина.

# 5. Результаты исследования

Практические результаты анализа представлены в следующей таблице:

|  |  |  |
| --- | --- | --- |
|  | Веблоги | Форумы |
|  | в целом | spbgu.ru | ixbt.com |
| Объем выборки, сообщений | 100 | 231 | 120 | 111 |
| из них репрезентативно | 73 | 230 | 119 | 111 |
| Словоупотреблений | 6040 | 5973 | 3014 | 2959 |
| Среднее число предложений | 9,973 | 3,070 | 3,017 | 3,126 |
| Среднее число слов | 82,740 | 25,970 | 25,328 | 26,658 |
| Средняя длина предложения | 8,296 | 8,459 | 8,395 | 8,528 |
| На 1 сообщение: |  |  |  |  |
| ошибок | 0,507 | 0,243 | 0,277 | 0,207 |
| упрощений | 0,384 | 0,083 | 0,084 | 0,081 |
| словоупотреблений сленга | 1,438 | 0,287 | 0,303 | 0,270 |
| табуированной лексики | 0,521 | 0,009 | 0,008 | 0,009 |
| На 1000 словоупотреблений: |  |  |  |  |
| ошибок | 6,126 | 9,376 | 10,949 | 7,773 |
| упрощений | 4,636 | 3,181 | 3,318 | 3,042 |
| словоупотреблений сленга | 17,384 | 11,050 | 11,944 | 10,139 |
| табуированной лексики | 6,291 | 0,335 | 0,332 | 0,338 |
| Встречаемость: |  |  |  |  |
| ошибок | 0,219 | 0,143 | 0,193 | 0,090 |
| упрощений | 0,192 | 0,057 | 0,059 | 0,054 |
| словоупотреблений сленга | 0,521 | 0,222 | 0,218 | 0,225 |
| табуированной лексики | 0,164 | 0,009 | 0,008 | 0,009 |

# 6. Некоторые комментарии к результатам исследования

Все значительные расхождения в значениях параметров между веблогами и форумами, на наш взгляд, продиктованы следующими причинами.

1. Сообщение в веблоге единично и представляет собой письменный монолог одного автора, в то время как совокупность сообщений в одном топике веб-конференции – диалог (полилог). Этим объясняется большее (в 3 раза) число предложений на 1 сообщение блога и, соответственно, среднее количество слов в нем. При этом значительного различия в длине предложений нет.

Помимо этого, языковая личность в диалоге ведет себя иначе, нежели в монологе. Осознанно или неосознанно, но коммуникант стремится создать определенное впечатление о себе (скорее всего, положительное) у партнера по коммуникации. Иными словами, мы считаем, что в письменном диалоге человек более склонен следить за речью, чем в письменном же монологе, и именно поэтому онлайн-дневники в 1,5 раза и более превосходят форумы по количеству упрощений, словоупотреблений сленга и табуированной лексики.

1. Веб-конференции модерируются, т.е. сообщение, содержащее в том числе запрещенную лексику, удаляется (хотя необязательно), а к его автору могут применяться санкции в виде запрета на размещение сообщений в конференцию или отдельные ее разделы на некоторое время. В связи с этим табуированная лексика в нашей выборке из форумов (оба из которых – модерируемые) практически отсутствует. Веблоги, как известно, в этом смысле предоставляют полную свободу слова.
2. Влияние технологий, типичных для блогов. В частности, результаты онлайн-тестов в сообщениях снижают репрезентативность выборки.
3. Влияние сложившихся традиций. Так, в веб-конференции не принято размещать сообщения, содержащие только изображения. При необходимости чаще дается внешняя ссылка.

Небезынтересна ситуация с грамотностью. Видно, что у веблогов выше встречаемость ошибок, а у конференций больше их частотность на 1000 словоупотреблений. Интерпретировать это можно следующим образом: грамотных людей на форумах больше, но если уж человек делает ошибки, то делает их много.

Также любопытно, что форум Санкт-Петербургского государственного университета заметно превосходит по числу ошибок форум iXBT.com. При этом значения остальных параметров, отнесенных нами к релевантным для стилистики, для этих двух конференций примерно одинаковы.

# Заключение

Стилистическая диагностика Интернет-источников представляется работой безусловно важной как с фундаментальной, так и с прикладной точки зрения, и довольно сложной. Как уже говорилось, новизна и разнообразие материала открывают массу возможностей для исследователя. В рамках данного исследования сделаны некоторые шаги к разработке эффективной методики создания текстовых выборок для анализа, что само по себе являлось одной из его задач. Далее, мы попытались выбрать некие критерии, по которым можно проводить собственно анализ. Можно говорить о том, что эти критерии были частично субъективны, но важнее то, что оценивалась (пусть и не абсолютно объективно) вся выборка одинаково. В дальнейшем, очевидно, имеет смысл прибегнуть к несколько иному набору параметров, например: соотношение частей речи, распределение падежных форм, частотность тех или иных синтаксических конструкций и т.п. Могут обнаружиться какие-либо уникальные для стилистической диагностики критерии[[1]](#footnote-1), противопоставляющие тексты Интернета всем остальным. Таким образом, планируется продолжать работу в этой области с учетом накопленного опыта.

# Научная Литература

1. Сичинава Д.В. Об одном лингвистическом параметре типологии текстов: коэффициент под/над. – Научно-техническая информация, серия 2, № 10, 2003, 27-35. – URL: <http://mitrius.narod.ru/Ling/Corpling/podnad.zip>
2. Трофимова Г.Н. Языковой вкус Интернет-эпохи в России (функционирование русского языка в Интернете: концептуально-сущностные доминанты): Монография. – 2005. – URL: <http://planeta.gramota.ru/gnt.html>
3. Функциональные стили и формы речи. / Под ред. проф. О.Б. Сиротининой. – Издательство Саратовского университета, 1993.
4. Шимкова М. Репрезентативность корпуса как лингвистическая проблема. // Сборник: Труды Международной конференции MegaLing-2005: Прикладная лингвистика в поиске новых путей. – СПб., 2005.

электронные источники

1. «Живой журнал». – URL: <http://www.livejournal.com>
2. Конференция iXBT.com. – URL: <http://forum.ixbt.com>
3. Форум СПбГУ. – URL: <http://www.spbgu.ru/forums/index.php>
4. Yandex. Поиск по блогам. – URL: <http://blogs.yandex.ru>

## СЛОВАРИ

1. Русский орфографический словарь Российской академии наук. / Отв. ред. В.В. Лопатин. Электронная версия. – 2002. – URL: <http://slovari.gramota.ru/portal_sl.html#lop>
1. Такие, например, как частотность конкретных слов: данная проблема затрагивается в статье [Сичинава 2003]. [↑](#footnote-ref-1)