Санкт-Петербургский государственный университет

Филологический факультет

Кафедра математической лингвистики

Плетнева Анастасия Дмитриевна

Исследование значимых лингвистических характеристик в задаче автоматического определения типа автора

Выпускная квалификационная работа по направлению
45.03.02 «Лингвистика»,
образовательная программа
«Прикладная, экспериментальная и математическая лингвистика»

Научный руководитель:
доцент, к.ф.н.,
Хохлова Мария Владимировна

Санкт-Петербург
2017

Оглавление

Generating Table of Contents for Word Import ...

Введение

В настоящее время все больше людей — а число таких стремится к абсолютному большинству — используют социальные сети для выражения своего мнения и своих эмоций с помощью написания текстов, называемых блогами. При этом возникает такое удивительное явление, как язык Интернета, который сочетает в себе характеристики разговорного языка и чего-то нового, что не используется при письменном и устном общении: хэштеги, смайлики, ненормативная пунктуация.

Отмечается воздействие компьютеров и глобальной сети на русский язык с двух сторон: во-первых, происходит одновременное усложнение одних и упрощение других средств сравнительно с аналогичными в русском языке, не подвергшимися воздействию глобальной сети, а во-вторых, видна конкуренция норм письменного устного языков. В целом же, можно констатировать тот факт, что язык Интернета пока остается недостаточно изученным в современной лингвистике и находится под пристальным вниманием лингвистов [Селютин 2009].

В качестве материала для исследования были рассмотрены тексты блогов стажеров, которые участвовали в программах международных обменов от организации AIESEC. Стажеры выбирают волонтерскую программу по одному из семнадцати направлений, соответствующих целям устойчивого развития ООН, либо стажировку в профессиональной сфере (маркетинг или преподавание). В данной работе были использованы как раз такие блоги.

Выбор данного материала был продиктован следующими особенностями:

1. Стажировки по данным направлениям являются одинаковыми по времени (6-8 недель или 6 месяцев), а также абсолютное большинство стажеров ведут блоги для описания своего опыта, что позволяет собрать значительный объем текстов, в которых описаны самые разные ситуации.
2. Стажировки проходили в разных странах, а люди, участвовавшие в них, приезжали из разных городов России, что позволило исключить возможность преобладания определенного диалекта.

Цель научно-исследовательской работы состоит в изучении и сравнении лингвистических характеристик языка Интернета (тональности, типичных n-грамм и синтаксических структур) на основе анализа текстов интернет-блогов, которые помогут при автоматической обработке блогов и определении типа автора, написавшего блог.

В работе было произведено сравнение корпусов двух типов: 1) корпус, содержащий тексты стажировок волонтеров (370000 словоупотреблений); 2) корпус, содержащий тексты профессиональных стажировок (350000 словоупотреблений). С помощью лингвистических характеристик будет проверена гипотеза, что между блогами определенного типа и группой авторов наблюдается корреляция, поэтому тексты различных типов отличатся между собой. Авторы, которые вели записи о волонтерских стажировках, в большинстве своем являются студентами 2-3 курсов бакалавриата, их возраст находится в пределах 19-23 лет, обычно они пока не получают профессионального опыта. Стажировки посвящены волонтерству и длятся 6-8 недель.

В стажировках профессионального плана принимают участие недавние выпускники высших учебных заведений, им около 23-27 лет, у них уже либо был опыт работы, либо они уже задумывались о нем и именно за этим выбрали стажировку для получения такого опыта. Стажировки проходят в определенных сферах — маркетинг и преподавание — и продолжаются в среднем 6 месяцев.

Таким образом, мы явно видим, что типы авторов, написавших блоги различных корпусов, отличаются между собой, поэтому мы и попытались выявить конкретные различия в тональности текстов и в употреблении определенных n-грамм и синтаксических структур предложений.

Научная новизна и теоретическая значимость данной работы состоят в расширении лингвистических знаний о компьютерной коммуникации и блога как ее жанра. В работе рассмотрены языковые особенности интернет-языка и типичные характеристики блогов в соответствии с текущим уровнем развития Глобальной сети. Блоги изучены с точки зрения тональности и особенностей n-

грамм и синтаксических конструкций текстов, которые помогут при автоматическом определении типов авторов.

Задачи настоящей работы заключаются в следующем:

1. Создание корпусов блогов волонтерских и профессиональных стажировок с заменой всех эмотиконов на специально введенные хэштэги и составление частотного списка лексем.
2. Анализ тональности блогов с помощью словаря тональности проекта Linis Crowd, программы SentiStrength, а также оценки такого явления как использование эмотиконов и удлинения слов.
3. Извлечение n-грамм из текстов и исследование типичных словосочетаний.
4. Извлечение типичных для данных текстов синтаксических структур предложений с помощью языка регулярных выражений и программы SketchEngine.
5. Сравнение полученных результатов для корпусов двух типов.

Методы исследования: в ходе эксперимента были использованы корпусный, статистический и дистрибутивный анализ.

Глава I. Теоретические предпосылки исследования

1. Язык Интернета как объект лингвистического исследования

Язык реагирует на новые социальные и технологические тенденции (на изменение стиля жизни) «развитием литературно-художественных и публицистических стилей» и сопровождающим его «неуклонным пополнением словарного состава» [Ожегов 1974].

Интернет — это глобальное средство хранения информации, которое позволяет людям вступать в переписку, заказывать товары и услуги, узнавать новости, знакомиться с другими людьми со всего мира, изучать иностранные языки и т.д. Наиболее широко рассматриваются электронные средства коммуникации, использование которых стало возможным на базе глобальной сети. К электронным средствам коммуникации относятся: общение с помощью электронной почты, компьютерных видеоконференций, дискуссионных групп и чатов, электронных досок объявлений, интернет-блогов и т.п.

Теория дискурса играет большую роль в изучении особенностей Глобальной сети. Термин «дискурс» понимается разными исследователями по-разному. По Н.Д. Арутюновой, дискурс — это «связный текст в совокупности с экстралингвистическими — прагматическими, социокультурными, психологическими и др. факторами; текст, взятый в событийном аспекте… Дискурс — это речь, погруженная жизнь» [Арутюнова 1990]. Интернет представляет собой большой корпус текстовых данных, которые связаны различными средствами — ключевыми словами, гиперссылками, темами и т.д. Наблюдается рост количества исследования, которые посвящены анализу интернет-общения [Морослин 2009]. Интернет-общение (или компьютерное) — это общение между пользователями в социальных сетях [Горошко 2007; 2009]. Данное общение возможно с помощью такой характеристики как «компьютерный дискурс». Часто под этим понимаются все тексты, которые связаны с интернет-технологиями. Анализ текстов, которые используются в процессе компьютерной коммуникации, учитывает большой ряд различных явлений — в т. ч. прагматику и способы реализации компьютерных технологий.

Некоторые особенности интернет-общения представляют значительный интерес лингвистов: использование эмотиконов и смайликов, отступление от орфографических и пунктуационных норм, смешение лексики русского и иностранных языков и др. [Бергельсон 2002]. Именно поэтому язык Интернета рассматривают как особую разновидность литературного языка, обусловленный компьютерной коммуникацией.

Что касается общих тенденций глобальной сети, то среди них выделяют следующие:

1. одновременно протекающее усложнение одних и упрощение других средств по сравнению с аналогичными средствами в литературном языке. Различные дискуссии и полемики выражаются в более простой форме общения — интернет-чатах, а сложны описания эмоций заменяются специальными символами — смайликами;
2. конкурирующее воздействие норм письменной и устной речи [Иванов 2000].

2. Лингвистические особенности языка Интернета

Употребление русского литературного языка в социальных сетях Интернета привело к появлению особого явления — языка Глобальной сети. Его нельзя причислить к какой-либо разновидности социального диалекта или жаргона, которые отклонятся от существующей литературной нормы русского языка, поскольку сфера функционирования языка Интернета предельно широка и не ограничена использованием определенной группой лиц [Мичурин 2014]. При этом отмечается большое влияние иностранных языков на русский и создание особенных средств выражения эмоций и отношений к объектам — эмотиконов, смайликов, новых слов и различных графических средств.

Отмечается, что не все отклонения от норм литературного языка являются намеренными. В частности, при использовании языка Интернета часто встречаются ошибки, возникающие не столько из-за неграмотности пользователей, сколько из-за недостатка времени на проверку сообщений [Карнуп 2014]. Употребление большого количества аббревиатур, сокращений,

ненормативной пунктуации и т.п. объясняется законом экономии языковых средств для более быстрой скорости письма [Трофимова 2014]. В этом разделе перечислены основные особенности интернет-языка.

2.1. Фонологические особенности

Просодические средства языка Интернета являются одними из наименее изученных в этой области. Некоторые исследователи относят к фонологическому уровню различные типографические знаки, которые помогают интонационно выделять определенные части текста [Ахманова 1966]. Данное явление актуально для языка Глобальной сети с самого его создания. Например, слова, написанные буквами верхнего регистра, обозначает то, что собеседник говорит на повышенных тонах [Кувшинская 2014], а повторяющиеся многоточия — сомнение интернет-пользователей.

Графические средства, однако, не только помогают сделать акцент или интонационно выделить определенную часть текста, но и могут выступать в роли отсылочных знаков. Например, знак @ в сочетании с именем адресата обозначает в социальных сетях пользователя, к которому обращаются с определенным сообщением.

Для выражения эмоций в текстах также используются определенные сочетания типографических знаков, а именно эмотиконы. Они могут обозначать радость, грусть, сомнение и т.д., причем эти знаки являются общепринятыми и используются всеми пользователями Интернета во всех социальных сетях.

Однако такие фонологические особенности не являются характерной чертой только лишь электронного общения. В частности, отмечается, что в художественной литературе произносительные особенности используются для имитации живой, разговорной речи [Кузнецова 2009].

2.2. Морфемные и словообразовательные особенности

Язык Интернета заимствует множество английских слов или морфем, причем калькируются также аббревиатуры, которые также впоследствии могут заменять определенные корневые морфемы русских слов. В остальном процессы

и способы словообразования, типичные для русского языка, также применяются и в языке Глобальной сети.

Также для языка Интернета типичны заимствованные слова и выражения, построенные по типу просторечных [Плисецкая 2014]. Например, такими являются глаголы *апргрейдуться, флудить* и прилагательное *офлайновый*.

Иногда заимствованный элемент может писаться не кириллическими, а латинскими символами, например, *GIF-анимация*. Но в последнее время такие слова также начинают употребляться в кириллическом варианте (*гифка*).

Что касается моделей словообразования, то они такие же, как ив русском литературном языке. Например, частотны суффиксы *-ер* или *-щик\-чик*: *браузер, отладчик, загрузчик*.

Иногда заимствованные слова могут принимать облик уже существующих слов русского языка. Например, слово *емейл* часто употребляется в русском языке как *мыло*.

Продуктивным способом словообразования в языке Глобальной сети является также сложение русских слов с заимствованными из английского языка словами *Интернет* и *веб* и русским *сеть*: *веб-дизайн, интернет-технологии* и т.п.

2.3. Лексические и семантические особенности

В плане лексики и семантике происходит не только заимствование слов, описанное выше, но также и появление новых значений у русских слов. Например, слово *страница* практически синонимично слову *сайт*, но приобрело свое значение при развитии Глобальной сети.

В различных социальных сетях реальные имена заменяются на определенные никнеймы — то есть псевдонимы людей. Они служат средством самоидентификации и помогают интернет-пользователям рассказать об особенностях своего характера и привлечь к себе внимание.

На лексико-семантическом уровне языка Интернета также частотно использование аббревиатур [Куликова 2012]. Они распространены в первую очередь в английском языке (напр., *ASAP — as soon as possible; LOL — laughing out loud*). Интересно то, что эти аббревиатуры появляются и в русскоязычной

части Интернета, написанные кириллицей (*АСАП, ЛОЛ*) [Кувшинская]. Часто они используются теми людьми, которые хорошо знают английский язык и часто работают на нем, тем самым влияя на остальных пользователей рунета.

Наряду с аббревиатурами также употребляются и сокращения одного слова, которые происходят с помощью усечения основы. Такие слова не являются полноценными аббревиатурами [Русская грамматика §590] и принадлежат к просторечным, разговорным и употребляющимся в речи молодых людей (напр., *магаз, препод, инфа*).

2.4. Синтаксические и пунктуационные особенности

Существует определенная тенденция в языке Интернета — это стремление пользователей писать аграмматично, т.е. отклонению от синтаксических и пунктуационных норм из-за желания передать наибольшее количество информации в короткие сроки. Часто знаки препинания либо ставятся в неправильных местах, либо не ставятся вообще. Напр., в некоторых сокращениях знаки пропадают вообще (*тд* вместо *т.д.*).

Парцелляция также очень распространена в текстах языка Интернета. Это делается для акцента на каких-то объектах или темах в предложении (напр., *Ну. И. Что. Ты. Здесь. Делаешь.*). Таким способом могут одновременно выражаться определенные эмоции или чувства, напр., злость или ярость.

Знаки восклицания или вопроса также во многих случаях вообще не ставятся в предложениях. Они часто опускаются, т.к. пользователи считают излишними писать их, особенно когда, напр., вопросительные слова употребляются отдельно (*что*). Наблюдается и обратная ситуация: при большой эмоциональности сообщений знаки препинания употребляются во множественном числе, часто даже превышающем тройной вариант (напр., *что?????*).

3. Определение понятия «блог»

Блог определяется как веб-страница, содержащая личный онлайн-дневник с отзывами, комментариями и различными гипрессылками [Nowson 2006].

Термин «блог» был введен Йорном Баргером 17 декабря 1997 года. Короткую форму слова «блог» предложил Петр Мерхольз, который в шуточной форме упомянул его в своем блоге Peterme.com в апреле или мае 1999 года. Эван Уильямс из Pyra Labs использовали «блог» как существительное и глагол (англ. “to blog”, что означает «изменить свой блог или отправить на свой блог» [Merriam-Webster Online]), что способствовало созданию термина «блогер». В Pyra Labs был создан Blogger.com, что привело к популяризации блогерства.

В плане структуры блог состоит основной информации (т.е. поста) и комментариев, которые выражают мнение в связи с информацией в посте. Чаще всего блоги публичны, имеют читателей, вступающих в полемику с автором блога посредством комментариев к записи или в своем личном блоге. Записи в посте содержат текст, фотографии, графические элементы или мультимедиа. Записи в блоге обычно недлинные и сгруппированы в обратной хронологической последовательности.

На сегодняшний момент производится широкий анализ блогов, который и позволил выделить особенности этого жанра компьютерной коммуникации. При этом отмечаются наиболее важные функции блогов: обсуждение текущих новостей, высказывание собственного мнения, снятие накопленного эмоционального напряжения. На сегодняшний момент уже в 2017 г. Поисковая система Яндекс обнаруживает 5 млн блогов, причем это не все блоги, которые есть в русскоязычной части Интернета, а именно активные блоги интернет-пользователей, т.е. те, которые хоть раз обновлялись в текущем году. Для примера, отмечается, что за весь 2008 г. было написано лишь 3,8 млн блогов, так что рост использования этого средства интернет-коммуникации очевидно.

Главными целями ведения интернет-блогов являются информационная (желание сообщение чего-то кому-то) и контактно-устанавливающая. Тематика блогов также разнообразна: темы могут быть нейтральными (посты о погоде, культурных и спортивных событиях и др.), профессиональными, связанными с определенной областью деятельностью и выражающими социальные проблемы общества и личности.

Иногда для блогов устанавливаются определенные правила, например, запрет на употребление ненормативной лексики и пунктуации, ограничение на использование смайликов и эмотиконов и др. Конечно, нет никаких установленных правил для всех блогов, а их содержание различается в различных социальных сетях. Именно это представляет интерес для лингвистов: нарушение орфографических и пунктуационных правил, использование нецензурных выражений и т.д. дает возможность исследования блогов в качестве лингвистического объекта для определения интернет-языка как особого вида русского языка и для идентификации типа автора, написавшего блог [Иванов 2000].

4. Анализ тональности

Сейчас, в связи с постоянным ростом использования сети Интернет в нашей жизни, в ней появляется множество различных явлений, перечисленных в разделе 1.1. данного исследования. Также возрастает и эмоциональное воздействие Интернета на сознание человека. В сети Интернет пользователи находят все больше способов, чтобы передать свои чувства собеседнику, в том числе эмотиконы, различные картинки и ненормативную пунктуацию [Koltsova, Alexeeva, Kolcov 2016].

Язык Интернета все больше отдаляется от обычного письменного литературного языка, поэтому понятно, что для его исследования нужны другие методы. Сейчас широко разрабатываются различные инструменты для автоматической обработки текстов.

Блоги представляют повышенный интерес для всестороннего изучения лингвистами, в том числе в плане социологического анализа. Часто в блогах (и комментариях к ним) отражается настроение общества и его мнения на различные проблемы. Доказано, что Sentiment Analysis — то есть оценка тональности (эмоциональной окраски) текста — может показать поведение политического лидера в ходе предвыборной кампании или предсказать поведение той или иной группы лиц в реальном времени в зависимости от различных признаков [Кольцов, Павлова, Кольцова 2012].

Анализ тональности текста существует для того, чтобы определить эмоциональную окраску текста. Это необходимо для выявления отношения автора к определенной теме или предмету. Автоматическое определение тональности текста подразумевает под собой выделение определенных фрагментов текста, несущих на себе позитивную или негативную оценку по отношению к какому-либо объекту [Пазельская, Соловьев, 2011]. Таким образом, тональность текста определяется несколькими факторами: 1) субъект тональности (автор определенной цитаты); 2) тональная оценка (позитивная, негативная или нейтральная); 3) объект тональности.

Существует несколько методов для определения тональности текста: методы векторного анализа и поиск эмотивной лексики в тексте. Первый из них подразумевает сравнение определенных текстов с заранее размеченным корпусом по выбранной мере близости и дальнейшая классификация текста по позитивной или негативной шкале. Второй способ включает в себя поиск эмоционально окрашенной лексики (или других средств выражения тональности) в соответствии с заранее созданными словарями тональности.

Эмотиконы (напр., *:), ;)*), наряду с эмоционально окрашенными словами, несут на себе часть экспрессивной окраски предложений в тексте. Эмотиконы созданы для решения проблемы передачи эмоций, которые автор вкладывает в сообщение. Количество употреблений эмотиконов в Интернете постоянно растет, однако не всегда взаимосвязь между ними и сентиментами ясна. Смысл некоторых из них понятна без каких-либо дополнительных исследований (напр., эмотиконы, выражающие грусть или радость), они явно несут на себе часть экспрессивности текстов. Для того чтобы разобраться в их вариантах, начали создаваться специальные словари эмотиконов [Mander 2000].

В социальных сетях эмотиконы используются в разной степени. Отмечается, что в Твиттере (социальная сеть, посты в которой состоят из коротких сообщений не более чем 140 символов) пользователи употребляют наибольшее количество эмотиконов. Блоги русскоязычной сети ВКонтакте не могут похвастаться таким большим разнообразием этого средства выражения экспрессивности: в части

блогов — это неотъемлемая часть текстов, в других они почти вовсе не употребляются. Однако ни одно исследование полностью не было посвящено связи между эмотиконами и выражением эмоциональности текстов в социальной сети ВКонтакте.

При анализе тональности слова или другим средствам выражения экспрессивности обычно присваивается положительная, отрицательная или нейтральная оценка. Но существуют эмотиконы, которые нельзя причислить ни к одной из существующих оценок. Напр., *:/* обозначает чувство раздраженности, что может быть негативной эмоцией для одних людей и нейтральной для других. Для этой гипотезы было проведено исследование, в котором пользователи должны были сказать, относится ли определенный эмотикон с позитивной, негативной, нейтральной оценке или он не относится ни к одной из выше перечисленных. Некоторые эмотиконы получили очевидную оценку, позитивную и негативную, но таким из них, даже самые часто употребляемым (напр., *:p*), была присвоена разная оценка разными участниками исследования [Hao Wang, Jorge A 2015].

Открытым остается вопрос, действительно ли эмотиконы содержат в себе определенные сентименты. Для этого [Hao Wang, Jorge A 2015] было проведено исследование: из выборки блогов в социальной сети Твиттер были удалены все эмотиконы. Затем было использовано два наивных байесовских классификатора: один был обучен на выборке твитов с эмотиконами, а второй — без них. Затем был проведен анализ тональности обеих выборок. Результаты показали, что при анализе выборки без эмотиконов точность определения тональности была значительно ниже, чем при анализе выборки с эмотиконами.

Также в текстах блогах все чаще начинают появляться смайлики (графические изображения эмоций), которых существует огромное количество, в отличие от эмотиконов, которых существует ограниченное количество. Несмотря на это, было проведено исследование [Novak , Smailović, Sluban, Mozetič 2015], что пользователи употребляют смайлики всего в 4% текстов в социальной сети Твиттер и в 50% - в Инстаграм.

Ещё одним средством выражения эмоциональности является удлинение слов, т.е. повторение одной и той же гласной буквы. Согласно [Brody, Diakopoulos 2011], это частое явление в социальных сетях, напр., в Твиттер. Удлинение было обнаружено в одном из каждых шести твитов. Но это не случайное явление, оно появляется именно в субъективных (эмоционально окрашенных) словах, напр., *очень, здорово*. Стоит отметить, что исследований, направленных на изучение удлинения слов и других средств выражения эмоциональности, таких как ненормативную пунктуацию, чрезвычайно мало. Более того, в русскоязычной литературе они практически отсутствуют.

5. Автоматическое определение типа автора

В сегодняшнее время одной из главных задач при обработке и анализе естественного языка является определение и верификация авторства текстов. Авторство основывается на классификации текстов на определенные группы на основе лингвистических характеристик авторов [Романов 2010]. Кроме самого определения авторства, при котором учитывается стиль авторов, изучается социальный аспект, то есть то, как используется язык. Это помогает при идентификации аспектов профилирования, таких как возраст, пол или имеющееся образование.

Обычно задача автоматического профилирования используется в судебной экспертизе, маркетинге и при решении задач, связанных с безопасностью. С точки зрения судебной экспертизы, очень важно знать лингвистический профиль определенного круга лиц, напр., авторов оскорбительных писем, то есть выявить конкретные характеристики языка, проявляющиеся в сообщениях такого типа людей. Впоследствии эти языковые характеристики играют важную роль при доказательстве преступлений. Также авторское профилирование интересует маркетологов, для которых важно знать мнение клиентов о той или иной продукции. В этом им помогает выявление определенных характеристик потребителей, в частности, пол, возраст и материальное положение, которые определяются на основе отзывов на продукт.

Однако в данный момент времени фокус в задаче авторского профилирования все больше смещается на анализ блогов пользователей социальных сетей, использующих язык Интернета, и на определение того, как он отражает личностные характеристики человека. Однако все чаще информация, которую определенный человек указывает о себе, оказывается неверной, напр., пол или возраст. Именно поэтому очень важно описать демографический и психологический портрет пользователя на основе их текстов [Rangel, Rosso 2016].

Не существует единого мнения, какой же набор характеристик с наибольшей точностью указывает на авторство. Что касается английского языка, в работах использовались различные критерии. Исследователи [Koppel, Argamon, Shimoni 2003] связывали использование языка с такой демографической характеристикой как пол человека. Они анализировали официальные тексты, извлеченные из национального корпуса британского английского языка, рассматривая служебные слова в сочетании с частеречными особенностями слов. В результате удалось достигнуть точности 80% в определении гендерной принадлежности автора [Argamon, Koppel, Fine, Shimoni 2003]. Другие лингвисты также занимались обработкой официально-деловых текстов [Holmes, Meyerhoff 2003; Burger, Henderson 2011].

Исследователи [Rangel, Rosso 2016] также придерживаются похожих идей и классифицируют тексты в соответствии с половыми и возрастными различиями испаноговорящих авторов. В основном они занимаются автоматической идентификацией эмоций в сообщениях социальных сетей и их взаимосвязью с типом автора текстов. В свою очередь, эти лингвисты основываются на работе [Pennebaker 2011], который использовал набор психолингвистических характеристик, таких как частотное употребление определенных частей речи, типов глаголов и т.д., в задаче авторского профилирования. В частности, он пришел к выводу, что для письменной речи мужчин более характерными являются предлоги, поскольку они обычно более подробно описывают то, что их окружает. Для русского языка задача автоматического определения типа автора остается актуальной, так как было проведено немного исследований, направленных на

авторское профилирование русскоязычных текстов. Отмечается, что чаще всего используется понятие «авторский инвариант». Инвариантами могут быть: количество гласных и согласных в тексте, часто употребляемые автором слова и т.д. На материале русского языка впервые были использованы методы распознавания образов в задаче атрибуции анонимных текстов с учетом индивидуальных особенностей авторов [Марусенко 1990]. Данный способ авторского профилирования дал высокие результаты при обработке историко-литературных текстов [В поисках потерянного автора, 2001; Синелева, 2001]. В большинстве случаев метод позволяет четко классифицировать тексты в зависимости от стилистических характеристик произведений.

Изучалось также влияние возраста и гендерной принадлежности на стиль текстов блогов [Schler, Koppel, Argamon, Pennebaker 2006]. Исследователи обработали более 71 000 блогов и разработали определенный набор стилистических черт, таких как несловарные и служебные слова и гиперссылки, а также униграммы, которые несут в себе большое информационное содержание. Гендерная принадлежность автора определялась с точностью 80%, а идентификация возраста проводилась с 75% точностью. Ученые доказали, что языковые особенности авторов блогов коррелируют с возрастом, что отражается, например, в использовании предлогов и детерминантов. К этим характеристикам были впоследствии добавлены жаргонные слова и средняя длина предложений, что повысило точность до 80,3% в идентификации возраста автора и до 89,2% при определении гендерной принадлежности [Goswami, Sarkar, Rustagi 2009].

На конференции PAN по автоматическому авторскому профилированию в 2013 году тестировалось несколько подходов к решению этой задачи [Rangel, Rosso, Koppel, Stamatatos, Inches 2013]. Большинство из них принимало во внимание сочетания нескольких стилистических характеристик, таких как частота употребления знаков препинания, прописных букв, цитат и т.д. Некоторые из подходов были основаны на использовании эмоционально окрашенных слов. Наилучшие результаты были получены при обработке текстов при помощи анализа часто употребляемых коллокаций [Meina et al., 2013].

Анализ методов на представительных корпусах показал, что наилучшим классификатор является машина, основанная на методе опорных векторов. Однако сложно судить о точности этого классификатора для русского языка, так как испытания этого метода проводились на англоязычных текстах.

В свою очередь, искусственные нейронные сети дают хорошие результаты при авторском профилировании. Единственным недостатком являются временные затраты, связанные с обучением сети. Таким образом, полного исследования этого метода проведено не было.

Разработаны программы, которые комбинируют вышеуказанные методы с некоторыми другими, напр., марковскими цепями, информационной энтропией и т.д. [Дюрдева 2016]. Среди средств обработки текстов, дающих наиболее точные результаты (95-98%), выделяется программа «Авторовед». Она сочетает в себе метод опорных векторов и кластерный анализ, и для работы ей требуются тексты объемом 20000-25000 символов. Данное программное средство обладает дружественным интерфейсом и позволяет составить психологический портрет автора [Мощенкова, Кривицкая, Амосова 2014].

Существует несколько проблем в задаче автоматического профилирования. Известно, что объем данных, используемых в исследованиях, является важным фактором в алгоритмах машинного обучения такого типа. Было проведено несколько экспериментов по обработке коротких интернет-сообщений, содержащих не более 15 токенов в одном тексте [Zhang, Zhang 2010]. Результаты отличались по точности от данных других работ, занимающимися более длинными текстами. Для коротких текстов блогов точность идентификации гендерной принадлежности автора составила 72,1%, в то время как в других исследованиях та цифра превышала 80%.

Другой распространенной проблемой является необходимость получения маркированных данных для определения гендерной принадлежности и возраста авторов [Rangel, Rosso, Chugur et al 2013]. Проводились исследования произведений классической литературы малоизвестных писателей, в которых можно было легко использовать ручную разметку. Похожая задача является

гораздо более сложной для социальных сетей. Процесс разметки (т.е. определения, по крайней мере, возраста и пола авторов) должен быть автоматизирован, т.к. объем данных в современной Глобальной сети безграничен. На сегодняшний момент исследователи вручную размечают корпуса текстов [Nguyen, Gravel, Trieschnigg, Meder 2013] с большой вероятностью ошибки из-за неверно предоставленной личной информации авторами текстов. Например, исследовалась социальная сеть Netlog [Peersman, Daelemans, Vaerenbergh 2011], в которой авторы блогов сами указывают свой возраст и пол. Эта особенность привносит большую вероятность ошибок и неточностей при обработке таких блогов и авторском профилировании, однако в то же время отражает реальное состояние доступных для исследований данных.

Выводы к Главе I

Данная глава была посвящена теоретическим вопросам, касающимся языка Интернета, способам выражения экспрессивности в текстах и связанном с ними анализом тональности. Подробно описывались фонетические, лексические, семантические и синтаксические особенности интернет-текстов. Также рассматривался отдельный жанр этих текстов, а именно блоги и их характерные черты. Наконец, речь шла об анализе тональности, цель которого — оценка экспрессивности текстов, учитывающая эмоционально окрашенные слова, предложения и другие средства (в т.ч. эмотиконы, смайлики, удлинения слов и т.д.).

Такие лингвистические характеристики обычно используются при автоматическом определении типа автора. В главе рассматривались методы, лежащие в основе программных средств, применяемых при авторском профилировании, а также сами программы, направленные на решение данной задачи.

Таким образом, в этой главе были проанализированы такие понятия как «язык Интернета», «блог», «анализ тональности» и «авторское профилирвание». Они были рассмотрены в контексте текущих тенденций развития Глобальной

сети. Особое внимание было уделено характеристикам, которые помогают выражать эмоциональность в тексте.

В результате анализа были сделаны следующие выводы:

1. Язык Интернета — это особая разновидность русского литературного языка, на который сильно влияют интернет-технологии.

2. Особенности языка Глобальной сети проявляются на фонологическом, морфологическом, словообразовательном, семантическом и синтаксических уровнях. Наряду с обычными процессами, присущими русскому языку (напр., *заимствованию слов, усечению основы для образования новых слов*), в языке Интернета существуют явления, типичные для текстов Глобальной сети: использование большого количества эмотиконов, смайликов, удлинения слов и т.д.

3. Рассмотрен особый жанр блогов, в котором авторы выражают свое мнение и отношение к чему-либо. При этом у пользователей есть возможность прокомментировать определенные записи. Интересно, что все выше перечисленные особенности, присущие языку Интернета, в большой степени выражаются в блогах.

4. Практически все рассмотренные характеристики текстов интернет-блогов являются средствами выражения экспрессивности. Они используются при анализе тональности текстов, причем исследуются не только эмоционально окрашенные слова и предложения, но все особенности текстов Глобальной сети: *эмотиконы, смайлики, слова верхнего регистра* и т.д.

5. Определен термин «авторское профилирование» и проанализированы основные методы, использующиеся для решения данной задачи. Стоит отметить, что не существует конкретного набора лингвистических характеристик, необходимых для точного описания личностного портрета автора текста.

Таким образом, были рассмотрены не только характерные особенности языка Интернета и тенденции его развития, а также их значимость при анализе тональности текстов и авторском профилировании.

Глава II. Инструменты исследования

2.1. Sketch Engine

В данном исследовании нами был использован корпусный менеджер Sketch Engine [Kilgarriff, Rychly, Smrz, Tugwell 2004]. Он позволяет работать как с уже имеющимися корпусами текстов, так и с создаваемыми пользователями коллекциями. Среди инструментов, доступных в данной программе, можно назвать следующие:

1) модуль для создания пользовательских корпусов;

2) модуль для создания частотных списков;

3) модуль для создания профилей слов (*англ.* word sketches);

4) модуль для создания семантического тезауруса и др.

При помощи программы Sketch Engine были созданы два корпуса, в которые вошли тексты блогов волонтерских и профессиональных стажировок.

Была получена информация о частоте вхождений лемм, n-грамм и средств выражения эмоциональности в текст. Частотные списки лемм были получены с помощью модуля «список слов», который позволил создать списки лемм с указанием их частот вхождения (Рисунок 1).

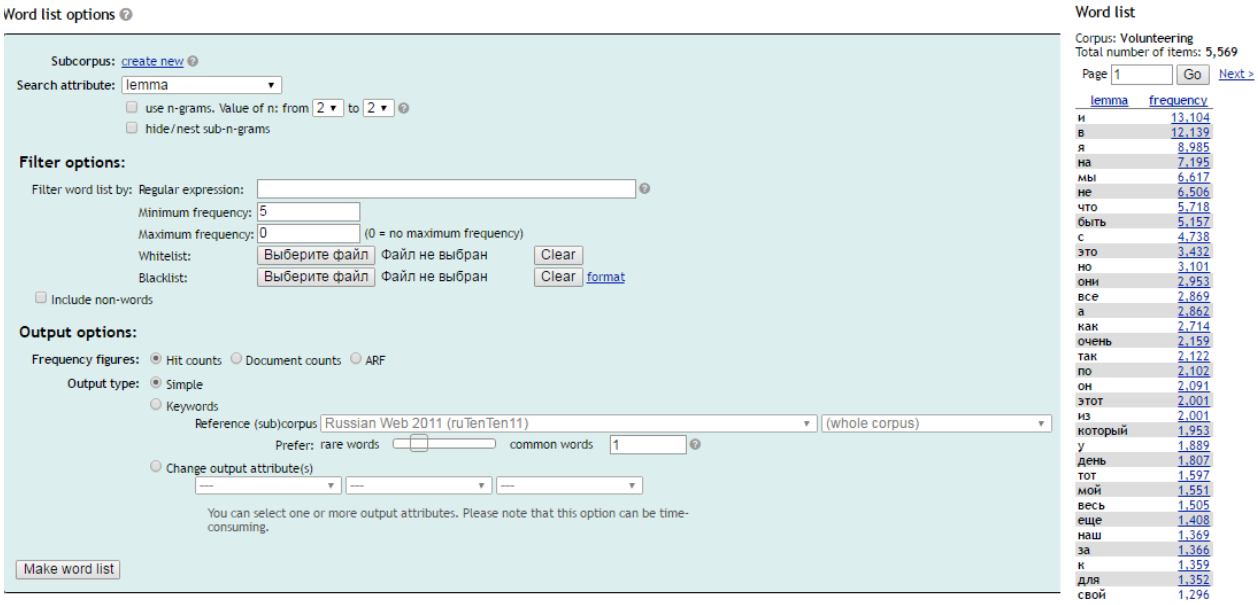


Рисунок 1 – Интерфейс создания списка слов в программе Sketch Engine и полученный результат

Из частотных списков были исключены «нежелательные» слова с помощью функции «список стоп-слов»: были отфильтрованы имена собственные, слова на иностранных языках, различные символы и их сочетания (см. приложение 2). Благодаря этому анализ тональности, характерных n-грамм и синтаксических структур был произведен с большей точностью и с меньшей тратой времени.

Для исследования синтаксиса словосочетаний и предложений были созданы частотные списки, позволяющие рассмотреть наиболее типичные из них. Для этого были сделан ряд запросов с использованием языка регулярных выражений. Следующие символы были использованы при написании запросов:

- символ * соответствует предыдущему элементу ноль или более раз;
- символ . обозначает один любой символ;
- символ | означает перечисление элементов и соответствует союзу *или*;
- символ *&* соответствует союзу *и*;
- квантификатор *{}* указывает, сколько раз может встречаться определенное выражение;
- последовательность .* обозначает любое количество любых символов между двумя частями регулярного выражения.

Для создания запроса в программе Sketch Engine, необходимо совершить поиск по корпусу при помощи функции “search”. Затем появляется страница для создания запроса (Рисунок 2).

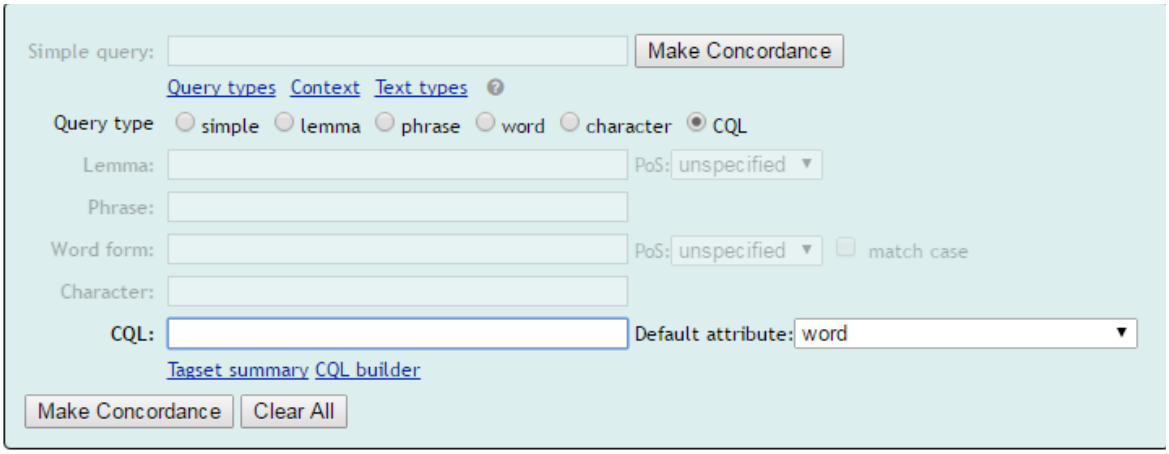


Рисунок 2 – Интерфейс запроса в программе Sketch Engine

В нашем исследовании мы использовали отдельный тип запросов CQL (Corpus Query Language). Это язык для расширенного поиска при помощи

регулярных выражений. В выпадающем окне “default attribute” доступен список атрибутов. В ходе нашей работы мы работали со следующими из них: word (словоформа), tag (морфологический тэг) и lemma (начальная форма слова).

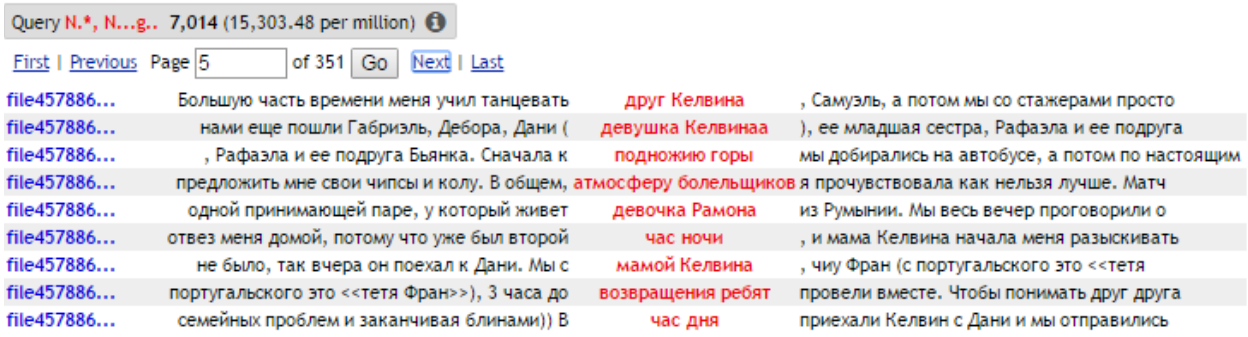
Также запрос можно усложнить, применив поиск, учитывающий морфологическую разметку корпуса. Ниже приведен пример расшифровки тэгов для имен прилагательных (Таблица 1).

Таблица 1 − Пример расшифровки тэгов для имен прилагательных

P	**Attribute (en)**	**Value (en)**	**Code (en)**
0	Category	Adjective	A
1	Type	qualificative	f
		possessive	s
2	Degree	positive	p
		comparative	c
		superlative	s
3	Gender	masculine	m
		feminine	f
		neuter	n
4	Number	singular	s
		plural	p
5	Case	nominative	n
		genitive	g
		dative	d
		accusative	a
		locative	l
		instrumental	i
6	Definiteness	short-art	s
		full-art	f

Для поиска при помощи морфологической разметки необходимо описать в терминах регулярных выражений цепочку, включающую информацию об исследуемом объекте. Запрос для одного слова выглядит следующим образом: оно

заключается в квадратные скобки (*[ ]*), в котором прописывается атрибут, а после знака равенства (=) указываются характеристики слова в двойных кавычках (*“ “*). При этом знаки препинания вводятся с помощью атрибута lemma после знака обратного слэша (напр., \. *(точка)*). Например, чтобы найти словосочетания типа существительное с другим существительным в Родительном падеже, необходимо ввести следующий запрос: [tag="N.*"][tag="N...g.."]. В результате применения подобного запроса будут обнаружены сочетания слов с правыми и левыми контекстами (Рисунок 3).


Query N.*, N...g.. 7,014 (15,303.48 per million)

First | Previous Page 5 of 351 Go Next | Last

file457886...	Большую часть времени меня учил танцевать	друг Келвина	, Самуэль, а потом мы со стажерами просто
file457886...	нами еще пошли Габриэль, Дебора, Дани (	девушка Келвинаа	), ее младшая сестра, Рафаэла и ее подруга
file457886...	, Рафаэла и ее подруга Бьянка. Сначала к	подножию горы	мы добирались на автобусе, а потом по настоящим
file457886...	предложить мне свои чипсы и колу. В общем,	атмосферу болельщиков	я прочувствовала как нельзя лучше. Матч
file457886...	одной принимающей паре, у который живет	девочка Рамона	из Румынии. Мы весь вечер проговорили о
file457886...	отвез меня домой, потому что уже был второй	час ночи	, и мама Келвина начала меня разыскивать
file457886...	не было, так вчера он поехал к Дани. Мы с	мамой Келвина	, чиу Фран (с португальского это <<тетя
file457886...	португальского это <<тетя Фран>>), 3 часа до	возвращения ребят	провели вместе. Чтобы понимать друг друга
file457886...	семейных проблем и заканчивая блинами)) В	час дня	приехали Келвин с Дани и мы отправились

Рисунок 3 – Результат поиска словосочетаний типа существительное с существительным в Родительном падеже

Для того чтобы создать частотный список элементов, необходимо выбрать ссылку “node forms” в разделе “frequency”. Таким образом, будет получен список слов с общим количеством найденных единиц по данному запросу, число вхождений определенного сочетания и их график распределения в соответствии с общим количеством элементов.

2.2. Веб-ресурс Linis Crowd

Это одна из первых работСловарь разрабатывался в рамках проекта, целью которой которого является являлось создание доступного для широких слоев пользователей ресурса на русском языке для анализа тональности. Тональные словари применяются в программном обеспечении, которое сопоставляет элементы словаря с найденными словами в тексте и, таким образом, определяет тональность всего текста.

Лексикон, использованный для создания словаря, собирался в два этапа: на первом был создан прототип словаря, куда были включены потенциально

окрашенные слова, а затем этот прототип был размечен волонтерами, которые присваивали словам оценки от -2 до 2 соответственно [Кольцов, Павлова, Кольцова 2012].

За основу прототипа тонального словаря был взят частотный словарь имен прилагательных из социальной сети Фейсбук, созданный Лабораторией цифрового общества [Лаборатория цифрового общества]. В списке содержалась 14933 прилагательных. Слова были предварительно оценены разработчиками, в результате чего получился словарь объемом 3293 прилагательных, которые создатели посчитали эмоционально окрашенными. Затем в данный список были внесены автоматически сгенерированные наречия (2310 слов) путем отсечения окончаний от прилагательных и добавления к основе окончаний наречий *(-о, -е, -и*). К получившему прототипу были добавлены эмоционально окрашенные слова из словаря И. Четверкина и Н. Лукашевич [Chetviorkin, Loukachevitch 2012]. 53 междометия были взяты из Объяснительного словаря русского языка [Морковкин 2012] и 1213 слова из ресурса, созданного Ю.В. Павловой [Павлова 2012] на основе словаря эмоциональных слов программы SentiStrength [Thelwall, Buckley, Paltoglou, Cai, Kappas 2010], переведенного с английского языка на русский. В итоге прототип словаря включал в себя 11869 слов из разных источников.

На втором этапе создания тонального словаря было собрано около 2000 постов самых популярных блоггеров интернет-журнала LiveJournal за период с марта 2013 по март 2014 года. Тексты блогов были написаны на девять тем, которые, по мнению авторов, можно было назвать темами, связанными с выражением мнения и эмоций (напр., социально-политическая тематика).

В результате был получен список из 7546 слов, среди которых оказались нейтральные (с оценкой 0), отрицательные (с оценками -1 и -2) и положительные (с оценками 1 и 2) слова. Распределение слов приведено в Таблице 2.

Таблица 2 − Распределение эмотивной лексики в словаре Linis Crowd

Оценка	**Количество слов**	**% слов**
-2	225	3
-1	1666	22
0	4753	63
1	853	11
2	49	1

Как можно заметить из данных Таблицы 2, наибольшее число слов получили оценку 0, т.е. были признаны волонтерами нейтральными.

В итоге, тональный словарь выглядит следующим образом: в нем представлен упорядоченный по алфавиту список слов с указанием среднего значения и дисперсии, вычисленных на основе оценок, приписанных волонтерами, а также общей оценки эмоциональности слова по шкале от -2 до 2. Отрывок данного словаря приведен в Таблице 3.

Таблица 3 – Отрывок из тонального словаря Linis Crowd

Слова	**Среднее значение**	**Дисперсия**	**Оценка**
безумие	-1	0	-1
безумство	-1	0	-1
безупречно	1,666667	0,471405	2
безупречный	1,111111	0,87489	1
безусловный	0	0	0
безуспешный	-1	0	-1
безучастный	-1,33333	0,471405	-1
безыдейный	-1,16667	0,372678	-1
безынициативный	-1	0	-1
безысходность	-0,75	0,829156	-1
безысходный	-1,33333	0,471405	-1

2.3. SentiStrength

SentiStrength — это программа для анализа тональности текста, разработанная Марком Тельволлем. Она была создана специально для оценки эмоций в коротких интернет-текстах. Программа использует разные признаки для оценки текста: отдельный словарь эмоционально окрашенных слов, эмотиконы, слова-усилители, слова-отрицания, идиоматические выражения и тому подобное.

Для каждого предложения внутри текста программа предоставляет информацию об эмоциональной окраске этих предложений: оценка от 1 до 5 соответствует положительно окрашенным словам, а от -1 до -5 — отрицательно окрашенным. Таким образом, если предложению приписаны оценки [3;-5], это значит, что оно содержит умеренную позитивную и сильную негативную оценки. Если предложение нейтрально по своей сути, то оценка будет равна [1;-1]. Две оценки нужны для того, чтобы наиболее адекватно оценить эмоциональную составляющую текста, так как, по мнению разработчиков, в любом предложении одновременно содержится как отрицательная, так и положительная стороны.

Программа была разработана для английского языка, но также адаптирована для ряда других языков, в том числе и русского [Павлова 2012]. В работе программы используется несколько основных файлов для более точной оценки эмоциональности текста. Самый основной из них — это словарь тональности, в котором указаны основные эмоционально окрашенные слова для данного языка, которым присвоена оценка от -5 до -1 и от 1 до 5. Существуют также дополнительные словари, содержащие эмотиконы и слова-усилители (например, *очень, чрезмерно, абсолютно*), закодированные от -5 до 5. Также при оценке используются списки со словами, выражающими отрицание, вопросительными местоимениями, идиоматическими, сленговыми фразами и выражениями, сокращениями, а также орфографическим словарем. По умолчанию, слова во всех файлах даются на английском языке. Для того чтобы использовать программу для текстов на другом языке, нужно заменить данные в файлах эквивалентами английскому языку. Согласно этим словарям программа определяет тональность текста (Рисунок 4).

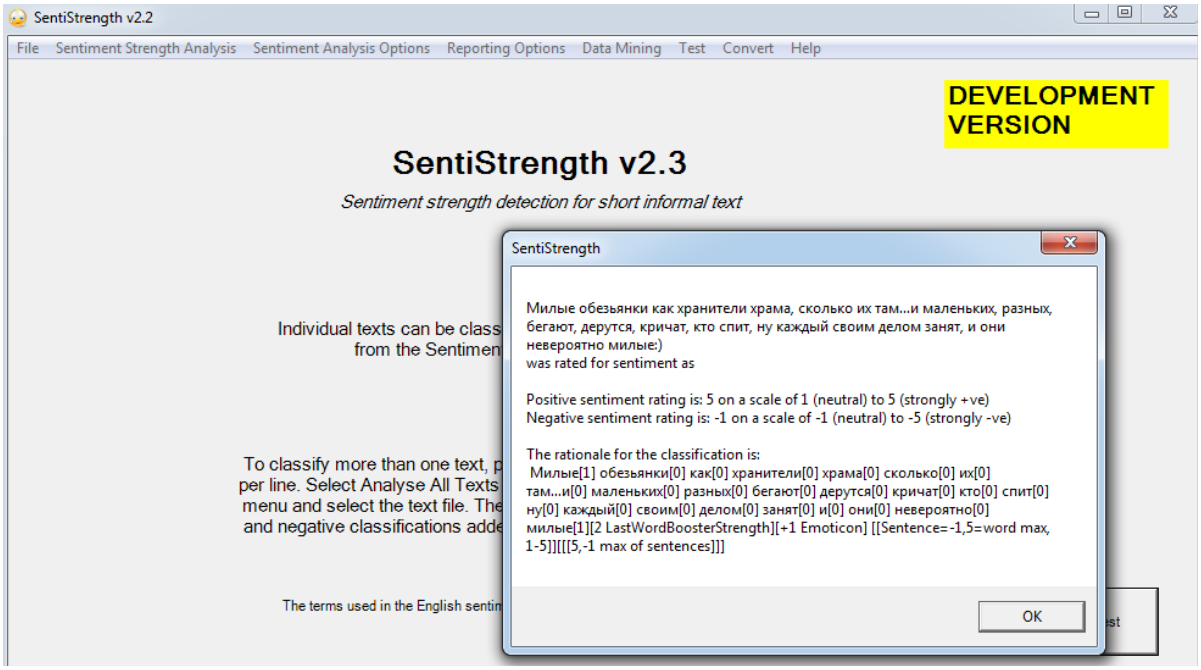


Рисунок 4 – Пример работы программы SentiStrength на материале отрывка текста блога

Программа SentiStrength анализирует и оценивает каждый элемент отрывка, присваивая им определенную оценку. В данном отрывке программой было выделено два положительно окрашенных слова, одно слово-усилитель (*невероятно*) и один эмотикон, что позволило определить общую положительную тональность предложения (оценка [4;-1]).

Выводы к Главе II

В данной главе рассмотрены основные инструменты, которые были использованы в нашем исследовании. В частности, рассказано о принципах работы программ Sketch Engine и SentiStrength, а также о свойствах краудсорсингового словаря тональности Linis Crowd. Были перечислены функции, которые использовались для этой работы. Основными положениями анализа стали следующие пункты:

1. Программа Sketch Engine позволяет пользователям создать нужные корпусы текстов, получить частотные списки лексем, n-грамм, предложений и различных средств выражения экспрессивности. При этом из списков можно исключить «стоп-слова», т.е. ненужные для определенного списка. Для создания частотных списков n-грамм и синтаксических структур

используется специальный язык запросов программы CQL и регулярные выражения.

2. Для оценки тональности слов был использован словарь тональности LinisCrowd, который состоит из списка эмоционально окрашенных слов и оценок, им присвоенных. Словам были приписаны оценки от -2 (негативные) до 2 (позитивные).

3. Одной из самых популярных программ для оценки тональности предложений используется программа SentiStrength. В ее работе используется несколько словарей: эмоционально окрашенных слов, эмотиконов, негативных и вопросительных слов и др. Именно они являются ключевыми признаками, на основе которых производится оценка. В работе программы применяются две шкалы: позитивная (с оценками от 1 до 5) и негативная (с оценками от -1 до -5), так как в одном предложении могут быть выражены как положительные, так и отрицательные сентименты.

Таким образом, нами были описаны и исследованы необходимые для нашего исследования инструменты.

Глава III. Исследование лингвистических характеристик

3.1. Общая характеристика практической части

Данная работа посвящена анализу корпусов текстов блогов волонтерских и профессиональных стажировок. Были созданы два корпуса, первый из которых состоит из 61 текста различных авторов общим объемом 370000 словоупотреблений. Второй корпус — корпус текстов о профессиональных стажировках — содержит 23 текста общим объемом 350000 словоупотреблений. Тексты были написаны авторами, которые в 2014-2017 годах побывали на волонтерских и профессиональных стажировках в 12 разных страна. Блоги оказались разными по объему, но у всех из них была одна цель: донести информацию о впечатлениях от пережитого опыта и поделиться результатами либо шестинедельной стажировки, направленной на реализацию одной из целей устойчивого развития ООН, либо профессиональной стажировки, длившейся от 3 до 18 месяцев.

Наша работа посвящена анализу интернет-блогов как части языка Глобальной сети согласно трем лингвистическим характеристикам: тональности, типичным n-граммам и синтаксическим структурам. Была рассмотрена и проверена гипотеза о том, что тексты авторов профессиональных и волонтерских блогов обладают различной тональностью и построены при помощи неодинаковых синтаксических структур в соответствии с различиями в возрасте, образовании и опыте работы авторов.

Таким образом, задачами текущего исследования являются:

1. Создание корпуса блогов волонтерских и профессиональных стажировок за 2014-2017 года на русском языке. При первичном исследовании этих блогов были замечены большое количество нетекстовых символов, а именно смайликов. Решено было заменить их определенными текстовыми тэгами (см. приложение 1).
2. Построение частотного списка слов волонтерских и профессиональных стажировок.

3. Сравнение частотных списков слов по волонтерским и профессиональным стажировкам по эмоциональной окраске с тональным словарем краудсорсингового веб-ресурса Linis Crowd.

4. Определение тональности отдельных предложений с помощью программы SentiStrength.

5. Исследование частоты появления эмотиконов и других средств выражения экспрессивности.

6. Формирование частотных списков n-грамм корпусов блогов профессиональных и волонтерских стажировок и выделение в них типичных словосочетаний.

7. Выявление типичных для корпусов блогов профессиональных и волонтерских стажировок структур предложений с помощью языка регулярных выражений.

3.2. Сравнение частотных списков со словарем тональности

Для проведения исследования был создан список лексем волонтерских и профессиональных стажировок. Для представления результатов сравнения этих лексем со списком слов из тонального словаря Linis Crowd были приведены несколько сопоставительных таблиц. Были рассмотрены только слова, которые также отражены в тональном словаре.

В словаре Linis Crowd словам присваивается негативная, нейтральная или положительная оценка (от -2 до 2) соответственно силами интернет-пользователей. По нашему мнению, слова с нейтральной оценкой не имеют большой значимости при определении тональности слов, поэтому такие слова исключались из анализа. При анализе блогов было выделено множество слов с позитивной оценкой. Всего эмоционально окрашенных слов, совпадающих со словарем тональности, оказалось 384 (Таблица 4). Список слов приведен в данной работе (см. приложение 3).

Таблица 4 − Распределение эмотивной лексики в корпусе волонтерских стажировок

Оценка	**Количество слов**	**Процент слов, %**
-2	16	4
-1	143	37
1	208	55
2	17	4

Результаты исследования эмотивной лексики текстов блогов волонтерских стажировок приведены в таблице 4. В ней указано количество слов с оценками от -2 до 2 (исключая 0), а также процент распределения этих слов в текстах. Результаты показывают, что слов с позитивной окраской наибольшее количество, что согласуется с первоначальным впечатлением от прочтения. Слов с маргинальными оценками (-2 и 2) обнаружено примерно одинаковое количество (4%), но слов с оценкой 1 больше, чем слов с оценкой -1 (55% vs.37%). Отметим, что некоторым (по большей части, нейтральным) словам была присвоена не соответствующая им оценка (например, *париж* с оценкой 1 или *тайвань* с оценкой -1), однако таких слов выявлено незначительное количество (примерно 3%) от всех слов, поэтому их, на наш взгляд, можно опустить.

То же исследование было проведено на материале корпуса профессиональных стажировок. Частотный список лексем был сравнен со словарем Linis Crowd. Нейтральные слова также исключались из анализа. В результате было выделено 269 слов (приложение 4), которые совпадают с данным словарем тональности. Из них лишь 46 слов оказались эмоционально окрашенными (Таблица 5).

Таблица 5 − Распределение эмотивной лексики в корпусе профессиональных стажировок

Оценка	**Количество слов**	**Процент слов, %**
-2	0	0
-1	10	21
1	32	71

2	4	8

Данные, приведенные в Таблице 5, позволяют сделать два вывода. Во-первых, уникальных эмоционально окрашенных слов в корпусе профессиональных стажировок в 8 раз меньше, чем в корпусе волонтерских стажировок. Во-вторых, слов с позитивной оценкой (1 или 2) гораздо больше, чем слов с негативной, причем слов с оценкой -2 не было обнаружено.

3.3. Анализ тональности текста с помощью программы SentiStrength

Как описано в предыдущем разделе, было проведено исследование эмоциональной окрашенности слов при помощи сравнения частотного списка слов корпуса со словарем тональности Linis Crowd. Чтобы убедиться в полученных результатах, которые показывают, что позитивно окрашенных слов больше, чем негативных, и, следовательно, общая тональность текста скорее позитивная, чем негативная, с помощью программы SentiStrength была проанализирована тональность предложений текста.

В задачи исследования не входила детальная адаптация программы к русскому языку, поэтому для настройки программы использовались уже готовые словари эмоционально окрашенных, негативных, вопросительных и других слов. Конечно, пришлось внести некоторые свои коррективы: например, в одном из словарей мы указали список тэгов, заменяющих смайлики в текущем исследовании.

Для того чтобы разметить предложения, весь корпус был помещен в текстовый файл формата .txt. Затем он был обработан программой, в результате чего всем предложения присвоены оценки по шкале от -5 до 5.

Для примера рассмотрим разбор одного из предложений программой. Она обрабатывает каждый элемент предложения (слово, эмотикон и т.д.) и присваивает ему оценку от -5 до 5. При этом если какой-либо элемент не найден ни в одном из словарей, то ему присваивается оценка 0 (таких элементов большинство). Затем оценка всех элементов сводится к единой — получается две оценки предложения:

положительная и отрицательная, так как в одном предложении могут содержаться разные сентименты (Рисунок 4).

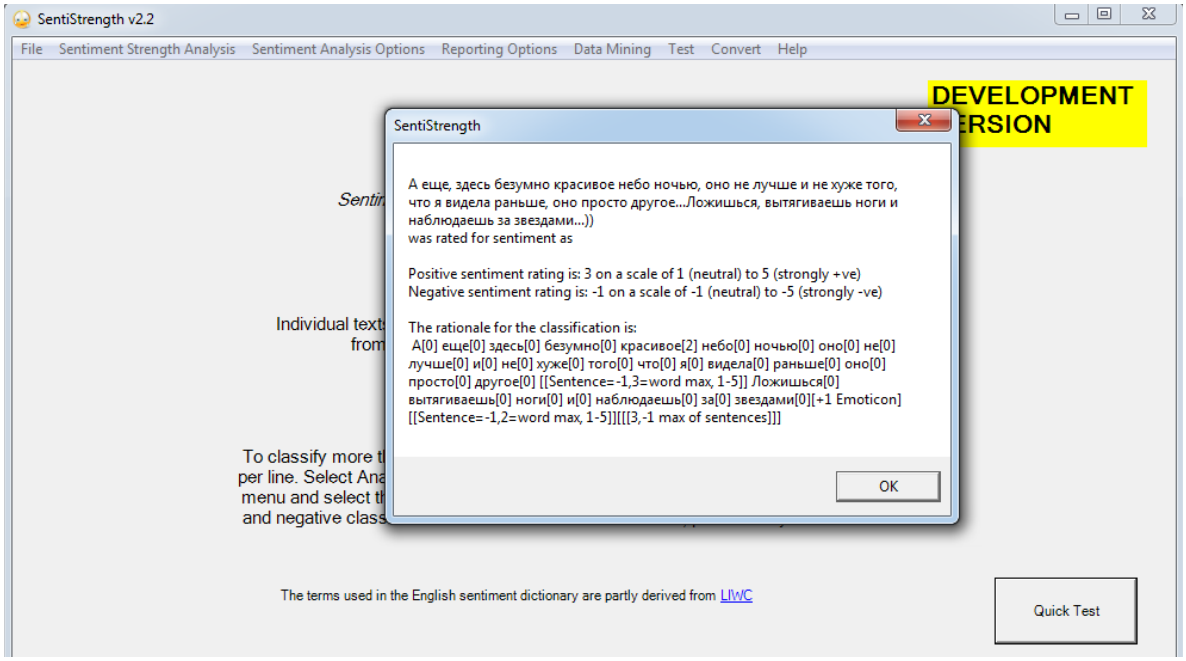


Рисунок 5 − Пример работы программы SentiStrength на материале небольшого отрывка текста

На Рисунке 5 показано, что предложению была присвоена оценка [3;-1], что означает, что оно явно положительное. В данном отрывке программы выделила одно эмоционально оценочное слово и эмотикон, которые повлияли на оценку всего предложения. Такие тесты можно проводить для одного предложения или небольшого куска текста. В нашем исследовании был обработан весь корпус волонтерских стажировок, результаты работы программы можно увидеть в следующей таблице. Всего было проанализировано 27307 предложений.

Всего возможных оценок предложений 25 ([-1;1], [-1;2], [-1;3], [-1;4], [-1;5] и т.д.). Среди них были выделены те, что характеризуют позитивные, негативные и нейтральные предложения (Таблица 6).

Таблица 6 − Распределение оценок тональности предложений корпуса волонтерских стажировок

Предложение	**Количество**	**Процент**	**Примеры оценок**
Положительное	8465	31	*[3;-1], [4;-2]*
Отрицательное	3550	13	*[1;-4], [2;-5]*

Нейтральное	15292	56	*[1;-1], [2;-2]*

Из данной таблицы видно, что нейтральных предложений наибольшее количество (56%). Например, в предложении ниже программа не обнаружила никаких соответствий между элементами предложения и специальными словарями, используемые при обработке предложению. Соответственно, данному предложению была присвоена оценка [1;-1]:

После пляжа мы с Келвином и Джозе пошли искать обменник, чтобы у меня наконец-то появились реалы, и Келвину не приходилось постоянно платить за нас двоих.

Однако это можно было предположить еще в предыдущем разделе — ситуация оказалась аналогичной той, что была при анализе эмоционально окрашенных слов. Больший интерес представляет то, что количество положительных предложений более чем в 2 раза превышает количество отрицательных предложений, что подтверждает теорию о положительной окрашенности текстов блогов вообще. Ниже в качестве примера приведены предложения с 1) положительной ([4;-1]) и 2) отрицательной ([1;-3]) оценками:

1) *В итоге, я очень рада, что она у меня наконец то есть!*
2) *Пока мы ехали в школу пошел сильный дождь и это, я скажу вам, не шутки.*

Та же работа была проделана для корпуса профессиональных стажировок. Было выделено 13068 предложений, что в 2 раза меньше, чем в корпусе волонтерских стажировок. При этом был применен тот же метод, что и при анализе первого корпуса: все предложения распределены на три группы: имеющие положительную, отрицательную и нейтральную тональность (Таблица 7).

Таблица 7 − Распределение оценок тональности предложений корпуса профессиональных стажировок

Предложение	**Количество**	**Процент**	**Примеры оценок**
Положительное	1852	14	*[3;-1], [4;-2]*

Отрицательное	1655	12	*[1;-4], [2;-5]*
Нейтральное	9561	74	*[1;-1], [2;-2]*

При анализе тональности предложений профессиональных блогов были выявлены интересные результаты. В текстах блогов корпуса профессиональных стажировок преобладают предложения с нейтральной эмоциональной окраской (74%) – это связано с использованием меньшего количества средств выражения экспрессивности по сравнению с текстами о волонтерских стажировках. Предложений с 1) позитивной и 2) отрицательной оценкой оказалось практически одинаковое количество (14% vs.12%), хотя общее количество предложений в 2 раза меньше, чем в корпусе волонтерских стажировок, как уже было отмечено выше:

1) А пока, наслаждаюсь моими последними деньками в Индонезии (завтракаю на нашей красивой веранде, ем больше фруктов, плаваю больше, гуляю больше, пою больше, загораю) [3;-1].

2) В воскресенье с утра был совсем какой-то кошмар, я вышла из дома и вдохнула просто гарь [1;-3].

3.4. Анализ средств выражения экспрессивности текстов о волонтерских стажировках

В данном разделе рассматривались различные способы выражения экспрессивности в тексте, исключая лексическую составляющую (результаты ее исследования описаны в предыдущих разделах этой главы). Для исследования были использованы возможности программы Sketch Engine, а также функции другой программы — Microsoft Word. С помощью первой создавался частотный список нетекстовых выражений эмоциональности текста (смотри приложение 5), а с помощью второй этот список дополнялся теми средствами, которые были отмечены при первоначальной, ручной обработке данных.

Результатами исследования будут являться таблицы, в которых отображено количество тех или иных способов выражения экспрессивности, а также их процентное соотношение в тексте. Данные таблицы сопровождаются примерами.

Кроме того, все средства выражения экспрессивности были разделены на две части: выражающие положительные и отрицательные эмоции. Рассмотрены следующие средства выражения: эмотиконы, смайлики, удлинения слов, ненормативная пунктуация, использование слов и выражений на любом другом языке, кроме русского, слова, написанные буквами верхнего регистра. Стоит заметить, что всего вхождений различных средств выражений эмоциональности, включая эмоционально окрашенные слова, 114219 в корпусе волонтерских стажировок. Но при исследовании было выяснено, что наибольшее количество из выше перечисленных средств занимают эмоционально окрашенные слова — 95983 вхождения. Поэтому на данном этапе они были исключены из анализа, а общее число средств, выражающих эмоциональность текста, таким образом, стало равняться 18236. Именно относительно данного числа будет посчитан процент появления того или иного явления в тексте.

3.4.1. Эмотиконы

Эмотикон (от англ. emotion icon — «эмоциональная иконка») — это пиктограмма, которая служит для выражения эмоции в языке Интернета. Она выражается в сочетании типографических знаков (например, *:), ;)* и другие). Эмотиконы могут не отображать фонетических, синтаксических, грамматических и других особенностей языка, при этом они помогают в уточнении самого высказывания и отношения к нему пишущего.

Частотный список использованных в данном корпусе эмотиконов приводится в приложении 3. Ниже мы приводим фрагмент таблицы, в которой в качестве примеров указаны эмотиконы, выражающие положительные и отрицательные эмоции. Отметим, что разметка (то есть определение, к какому типу относится определенный эмотикон) была сделана автором данного исследования (Таблица 8).

Таблица 8 − Распределение эмотиконов в текстах волонтерских стажировок

Эмотикон	**Количество, вхождения**	**Процент, %**	**Примеры**

Положительная эмоция	8098	44	*:), ;), ^_^, *_**
Отрицательная эмоция	2480	13	*:(, ((, oO*

Как видно из данной таблицы, эмотиконов, выражающих положительные эмоции в 4 раза больше тех, что выражают отрицательные:

Я накупила столько украшений, еды и прочего:))))))))))))))))))))))))))))))

При этом стоит отметить, что количество первых в тексте велико — 44%, а вместе они покрывают 57% из общего числа средств, выражающих эмоциональность текста.

3.4.2. Смайлики

В отличие от эмотиконов, смайлик — это графическое изображение человеческого лица, отображающего определенную эмоцию. Также, как и в случае с эмотиконами, они делятся на три группы: выражающие положительные, отрицательные эмоции и нейтральные смайлики (то есть те, которые не выражают определенной эмоции). Отметим, что в начале исследования в корпусе все смайлики (графические изображения) были заменены специальными тэгами для удобства обработки (Таблица 9).

Таблица 9 − Распределение смайликов в текстах волонтерских стажировок

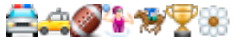
Смайлик	**Количество, вхождения**	**Процент, %**	**Примеры**
Положительная эмоция	446	2	
Отрицательная эмоция	100	0.5	
Нейтральные смайлики	579	3	

Таблица 9 показывает, что количество смайликов в текстах невелико (всего 5.5%), тогда как при первичной обработке данных казалось, что именно они

играют большую роль при передаче различных эмоций в тексте. В примере ниже смайлики заменены специальными тэгами:

[EMOJI] [EMOJI]Откуда такая возможность[EMOJI] [EMOJI]

3.4.3. Другие средства выражения эмоций

Оставшиеся средства выражения эмоций — удлинения слов, пунктуация, использование иностранных слов и слов, написанных верхним регистром — было решено рассматривать вместе. Несомненно то, что все они выражают эмоции в тексте, помогают уточнению определенных выражений (Таблица 10).

Таблица 10 − Распределение средств выражения эмоциональности в текстах волонтерских стажировок

Средство	**Количество, вхождения**	**Процент, %**	**Примеры**
Пунктуация	2822	15	*!!!, ..., ?!*
Удлинения слов	1001	5	*ооочень, эээ, привееет*
Иностранные слова	2237	12	*Buddy, hello*
Верхний регистр	463	2	*НЕТ, НО, БРАЗИЛИЯ*

Наибольший процент здесь занимает использование ненормативной пунктуации (15%). Стоит отметить, что при этом внутренняя организация предложения и, соответственно, пунктуация внутри него не нарушается, несмотря на ненормативность в конце предложений:

И тут такая новость в группе #aiesec_msc (скриншот прикрепляю), от которой все мысли сбежались в кучу……ГОА!!!!……МЕРОПРИЯТИЕ !!!!! …..МЕЖДУНАРОДНЫЙ ОПЫТ РАБОТЫ!!!!!.....ТАНЦЫ!!!!....это же правда #стажировкамечты !!!!

Далее идет использование иностранных слов (12%), которые могут использоваться как частично в структуре предложения (заменять определенные русские слова и выражения), так и все предложение может быть написано на иностранном языке. Данное явление объясняется либо написанием целых кусков

текста на не русском (чаще всего, английском) языке, либо языковыми лакунами в русском языке, вследствие чего авторы используют иностранные слова для их покрытия:

Вчера мы были на Средиземном море! Finally!

Верхний регистр и удлинения слов также нельзя причислить к наиболее используемым средствам выражения эмоций в тексте — их всего 2 и 5% соответственно:

В начале нам предстояло подняться на гору, затем спуститься с нее к пляжу, это было ооооооооооочень ОЧЕНЬ тяжело!

В предыдущем разделе был отмечен тот факт, что позитивно окрашенных предложений оказалось больше, чем негативных, что соответствует проценту ненормативной пунктуации в текстах, так как все эти знаки препинания были взяты из конца предложений.

3.5. Анализ средств выражения экспрессивности в текстах о профессиональных стажировках

В данном разделе представлены результаты, полученные в ходе анализа блогов о профессиональных стажировках. Их мы описываем отдельно от волонтерских блогов и в одном разделе, так как средств выражения эмоциональности таких текстов получилось гораздо меньше, чем в волонтерских блогах. Всего таких вхождений было выделено 9575 (Таблица 11). В отличие от волонтерских блогов, эмоционально окрашенные слова покрывают в тексте относительно маленький процент, поэтому они не были исключены из рассмотрения в этом разделе.

Таблица 11 − Распределение средств выражения эмоциональности в текстах профессиональных стажировок

Средство	**Количество, вхождения**	**Процент, %**	**Примеры**
Эмотивная лексика	1643	17	*Хороший, интересный, красивый*
Эмотиконы	3421	35	*:), ;), ^_^, *_*

Смайлики	463	5	😂😄
Пунктуация	2237	23	*!!!, ..., ?!*
Удлинения слов	77	1	*ооочень, эээ, приввееет*
Иностранные слова	1437	15	*Buddy, hello*
Верхний регистр	297	4	*НЕТ, НО, БРАЗИЛИЯ*

Таблица 11 позволяет сделать вывод, что эмотиконов среди средств выражения эмоциональности в текстах профессиональных блогов больше всего (35%):

Хочу Вам показать фотографии европейской еды))) у нас тут в маленьком Фуянге открывается европейский ресторан)) и там даже есть русское блюдо под названием чебуреки))))

Затем идет ненормативная пунктуация (23%) и иностранные слова (15%):

Спроси любого китайца – стоит ли ехать в Чунцин? – и он в первую очередь скажет, что стоит хотя бы ради того, чтобы попробовать местный hot pot.

Отметим, что данная статистика отличается от той, что была выявлена при исследовании волонтерских блогов, где эмоционально окрашенных слов было наибольшее количество (около 85%), здесь же их всего 17%.

3.6. Сравнение типичных n-грамм текстов блогов волонтерских и профессиональных стажировок

В исследовании с помощью программы Sketch Engine типичные для блогов n-граммы были выделены из текстов блогов волонтерских и профессиональных стажировок с помощью запросов в программе Sketch Engine и регулярных выражений, а также было проведено их сопоставление (запросы можно увидеть в Приложении 6). В работе рассмотрены следующие типа лексико-грамматических типов словосочетаний [Казаков 2013]:

- Именные словосочетания:

- Субстантивные словосочетания (главное слово — существительное): сочетания существительного с другим существительным в Родительном падеже, а также согласующихся в роде, числе и падеже прилагательного и существительного (напр., *центре города; последний день*);
- Адъективные словосочетания (главное слово — прилагательное): сочетания прилагательного с существительным с предлогом, а также прилагательного с наречием (напр., *похожее на парк; очень приятно*);
- Словосочетания с числительным в роли главного слова: сочетания числительного и существительного, числительного с существительным с предлогом (напр., *первый в жизни; несколько дней*);
- Словосочетания с местоимением в роли главного слова: сочетания согласующихся местоимения и прилагательного, а также местоимения с другим местоимением с предлогом (напр., *самое интересное; каждому из нас*);

- Глагольные словосочетания:
  - Словосочетания, распространенные существительным: сочетания глагола с существительным и глагола с существительным с предлогом (напр., *провести время; сижу в аэропорту*);
  - Словосочетания, распространенные наречием: сочетания глагола с наречием (напр., *уже писала*);
  - Словосочетания, распространенные глагольными формами:
    - Инфинитивом (напр., *ложиться спать*);
    - Деепричастием (напр., *ехали сидя; выйдя из машины*);
    - Причастием (напр., *созданным для продвижения*);
- Наречные словосочетания (напр., *впервые в жизни; очень быстро*).

Также составлены частотные списки междометий для текстов блогов волонтерских и профессиональных стажировок (пример частотного списка можно увидеть в Приложении 7).

В текстах блогов волонтерских стажировок было обнаружено 75645 типичных для русского языка словосочетаний. Их количество и процент распределения указан в приведенной ниже таблице (Таблица 12).

Таблица 12 − Распределение словосочетаний в текстах волонтерских стажировок

Тип словосочетания	**Подтип**	**Количество вхождений**	**Процент вхождений**
Именные	Субстантивные	7005	9
		28545	38
	Адъективные	414	0.6
		2883	4
	Числительное	108	0.2
		5028	7
	Местоимение	4153	5
		1050	1
Глагольные	Существительное	9005	12
		6501	9
	Наречие	6719	9
	Инфинитив	335	0.5
	Деепричастие	8	~0
		300	0.5
	Причастие	392	0.6
Наречные		941	1
		1856	2
Междометия		402	0.6

Данная таблица показывает, что в текстах блогов волонтерских стажировок больше всего субстантивных словосочетаний типа существительное с прилагательным, согласующихся в роде, числе и падеже (38%): *последний день, местных жителей, рабочий день*. Затем идут словосочетания типа существительное с другим существительным в Родительном падеже (9%): *восхода солнца, кусочке земли, чудес света* – и глагольные словосочетания с распределением в виде существительного (12 и 9% соответственно): *приехали в аэропорт, еду на стажировку, купили билеты* – и наречия (9%): *сильно отличается, очень любит, честно скажу*. Стоит отметить, что словосочетания с деепричастиями в текстах практически не употребляются.

Что касается текстов блогов профессиональных стажировок, то здесь было найдено 87601 словосочетание, что больше на 12000 элементов, чем в текстах блогов волонтерских стажировок (Таблица 13). При этом объем корпуса волонтерских стажировок больше на 20000 словоупотреблений.

Таблица 13 − Распределение словосочетаний в текстах профессиональных стажировок

Тип словосочетания	**Подтип**	**Количество вхождений**	**Процент вхождений**
Именные	Субстантивные	8603	10
		32588	39
	Адъективные	470	0.6
		3247	4
	Числительное	143	0.3
		5668	6
	Местоимение	4867	5
		1162	1
Глагольные	Существительное	10808	12
		7209	8
	Наречие	7745	9
	Инфинитив	426	0.6
	Деепричастие	7	~0
		335	0.4
	Причастие	460	0.6
Наречные		1235	1
		2225	2
Междометия		403	0.5

В текстах блогов профессиональных стажировок проценты распределения различных словосочетаний практически такие же, как и в текстах блогов волонтерских стажировок. Таким образом, несмотря на то, что блоги различаются по степени эмоциональности, они не отличаются синтаксисом словосочетаний.

3.7. Сравнение типичных синтаксических структур предложений текстов блогов волонтерских и профессиональных стажировок

В исследовании с помощью запросов с использованием регулярных выражений в программе Sketch Engine также были извлечены типичные для блогов синтаксические структуры предложений, и было проведено их сопоставление (запросы можно увидеть в Приложении 8). В работе

использовались следующие структурные схемы простых предложений [Русская грамматика §1913]:

- N1 V3s
- Vf3s Inf
- N2 (neg)Vf3s
- N1 - N1
- N1 - Adj1
- N1 - N2
- N1 — Inf
- Inf - N1
- Inf - Adv-o
- Praed Inf
- Praed (neg) N4/N2
- Praed part N2
- Adv quant - N2
- Нет N2
- Ни N2
- Никого (ничего) N2
- Никакого N2
- N1

Фразеологизированные схемы:

- N как N
- Не до N
- Inf так Inf
- Ох, ах, эх N
- Что за N

Именно эти структурные схемы рассматривались потому, что простые предложения с такими типами были найдены в текстах блогов. В ниже представленных таблицах приведено количество вхождений структур таких предложений и процент их распространения в текстах блогов волонтерских и профессиональных стажировок (Таблицы 14 и 15).

Таблица 14 − Распределение типов простых предложений в текстах волонтерских стажировок

Схема	**Количество вхождений**	**Процент вхождений**
N1 V3s	7338	73
Vf3s Inf	651	7

N2 (neg)Vf3s	337	3
N1 - N1	155	2
N1 - Adj1	45	0.4
N1 - N2	12	0.2
N1 — Inf	15	0.2
Inf - N1	4	0.1
Inf - Adv-o	650	7
Praed Inf	231	3
Praed (neg) N4/N2	7	0.1
Praed part N2	0	0
Adv quant - N2	0	0
Нет N2	112	2
Ни N2	51	0.4
Никого (ничего) N2	3	0.1
Никакого… N2	10	0.2
N1	87	0.7
N как N	2	0.1
не до N	7	0.1
Inf так Inf	1	0.1
ох, ах, эх N	1	0.1
что за N	16	0.2

Всего было найдено 9735 вхождений типичных структур в текстах блогов волонтерских стажировок. Пример частотного списка предложений, построенных по схеме N1 – N1, дан в Приложении 9. Для поиска таких предложений был сделан запрос [tag="N...n.*"][lemma="\-"][tag="N...n.*"]. Таким же образом были получены остальные частотные списки:

Для бразильцев сейчас зима, поэтому они не плавают, но для нас, стажеров, Бразилия - страна вечного лета, так что мы без зазрения совести искупались.

Большая часть предложений строится по структуре подлежащее со сказуемым (73%):

Так интересно бразильцы поют и хлопают во время задувания свеч именинником!

Сочетания глагола в 3 лице единственном числе с инфинитивом и инфинитива с наречием, оканчивающемся на *-о*, идут следом за предыдущей структурой (7%). Остальные структуры немногочисленны, и их распространение варьируется в диапазоне от 0.1 до 3%. Стоит отметить, что авторы блогов используют различные фразеологизированные структуры в своих текстах (5 разных типов):

Ну ничего, подумали мы, гулять так гулять.

Таблица 15 − Распределение типов простых предложений в текстах профессиональных стажировок

Схема	**Количество вхождений**	**Процент вхождений**
N1 V3s	3800	62
Vf3s Inf	816	13
N2 (neg)Vf3s	457	7
N1 - N1	214	3
N1 - Adj1	92	1
N1 - N2	0	0
N1 — Inf	6	0.4
Inf - N1	2	0.2
Inf - Adv-o	1	0.2
Praed Inf	560	9
Praed (neg) N4/N2	90	1
Praed part N2	1	0.2
Adv quant - N2	2	0.2
Нет N2	33	0.6
Ни N2	22	0.5

Никого (ничего) N2	0	0
Никакого… N2	4	0.3
N1	51	0.8
N как N	1	0.2
не до N	3	0.2
Inf так Inf	0	0
ох, ах, эх N	0	0
что за N	9	0.2

Таблица 15 показывает нам, что в текстах блогов профессиональных стажировок найдено 6164 вхождений структур простых предложений, причем количество типов меньше, чем в текстах блогов волонтерских стажировок (21 vs. 19). Это различие проявляется, по большей части, во фразеологизированных структурах: в текстах профессиональных блогов используется 3 типа, а их количество меньше, чем в волонтерских блогах:

А на вкус говядина как говядина.

В этих блогах также больше всего используется схема подлежащего со сказуемым, а также глагола в 3 лице единственном числе с инфинитивом. Однако очевидно, что несмотря на то, что вхождений различных структур в текстах профессиональных блогов меньше, чем в волонтерских, некоторые структуры употребляются в больше количестве. Это, например, 1) глагол в 3 лице единственном числе с инфинитивом (651 vs.816), 2) сочетания предикативов и фразеологизмов с модальными значениями возможности, целесообразности, волеизъявления, своевременности (напр., *можно, нельзя, велено, решено* и т.п.) с инфинитивом (231 vs.560), а также 3) предикатива с существительным в родительном или винительном падежах (7 vs.90). Мы видим, что схемы без подлежащего более распространены в текстах блогов профессиональных стажировок, чем волонтерских:

1) В заключении, <u>хочется сказать</u>, что еще в Новосибирске, один человек сказал мне: когда ты в Индонезии, ты просто счастлив и тебе не надо даже объяснять почему.

2) У каждого в жизни наступает момент, когда надо принять РЕШЕНИЕ, решение, которое должно направить вашу жизнь в новое неизведанное вами русло, к новым просторам и океанам.

3) Получив заказ, они высчитывают, сколько и какого нужно материала (в основном, конечно, кожзам, но есть и отличная кожа).

Напротив, авторы текстов блогов волонтерских стажировок больше используют схемы именных предложений, напр., предложения, состоящие из только 1) подлежащего (87 vs.51) или 2) структур с отрицательными частицами и существительными в родительном и винительном падежах (112 vs.33):

1) Девушка. Итог. Ночь.

2) Так много зелени и цветов здесь, совсем нет пыли , меньше народу и выглядит все как-то....прилично)))

Таким образом, тексты блогов волонтерских стажировок обладают большей фразеологизированностью и используют больше именных предложений, в то время как тексты блогов профессиональных стажировок обладают большей формальностью синтаксических структур.

Выводы к Главе III

В данной главе были описаны несколько проведенных нами экспериментов по сравнению текстов блогов волонтерских и профессиональных стажировок и выявили некоторые отличительные признаки этих текстов, которые подтверждаются тем, что авторы сами различаются возрастом, образованием и опытом работы. Для этого использовались корпусные и статистические методы. С помощью программ Sketch Engine и SentiStrength и тонального словаря LinisCrowd был проведен анализ тональности наших текстов и сравнили синтаксические структуры словосочетания и предложений. Ранее было отмечено, что авторы блогов отличаются по возрасту, образованию и опыту работы. В нашей работе рассмотрены три различных характеристики, которые помогли подтвердить, что и тексты, написанные ими, различаются своими структурами и средствами выражения эмоциональности. Ниже приводятся основные выводы:

1. Были проанализированы эмоционально окрашенные слова, предложения, а также иные средства выражения экспрессивности текстов (напр., *эмотиконы, смайлики, удлинения слов* и т.д.). В результате сравнения было обнаружено, что разных эмоционально окрашенных слов в текстах блогов волонтерских стажировок в 8 раз больше, чем в профессиональных. Но в обоих корпусах позитивно окрашенные слова преобладают над негативными.

2. В корпусе текстов блогов волонтерских стажировок было обнаружено в 2 раза больше предложений, чем в корпусе текстов блогов профессиональных стажировок. При этом в текстах блогов волонтерских стажировок позитивно окрашенных предложений также в 2 раза больше, чем негативных. Обратная ситуация наблюдается в корпусе текстов блогов профессиональных стажировок: там позитивных и негативных предложений примерно одинаковое количество. Это связано с личными особенностями авторов текстов: возрастом, образованием и опытом работы.

3. Распределение средств выражений экспрессивности в текстах разных блогов также неодинаково. Оно заключается в использовании эмоционально окрашенных слов: в текстах блогов волонтерских стажировок их наибольшее количество (85%), тогда как в блогах профессиональных стажировок их всего 17%. В обоих корпусах широко распространены ненормативная пунктуация, использование иностранных слов и эмотиконы.

4. В плане синтаксиса словосочетаний между разными текстами нет различий: именные и глагольные словосочетания широко употребляются в текстах блогов волонтерских и профессиональных стажировок, но в одинаковом процентном соотношении.

5. Что касается синтаксиса простых предложений, тексты блогов волонтерских стажировок обладают большей образностью за счет использования фразеологизированных структур и использования именных схем предложений. В то же время блоги профессиональных стажировок, то

они более формальны, т.к. там употребляется большое количество предложений с глаголами без подлежащего.

Таким образом, были выявлены определенные признаки, которые можно использовать при автоматических методах определения типа авторов текстов. Исследовались различные средства выражения экспрессивности текста и различные синтаксические структуры, и проводился анализ их распределения в текстах.

Заключение

Наше исследование было посвящено анализу текстов блогов волонтерских и профессиональных стажировок, которые были написаны участниками данных программ. Авторы этих текстов отличаются между собой по возрасту, образованию и опыту работы. Для того чтобы выявить определенные признаки, которые помогут в различении текстов, мы изучили теоретические вопросы, связанные с языком Интернета, определили жанр блогов Глобальной сети и рассмотрели способы применения анализа тональности к текстам блогов.

Мы подробно описывали фонетические, лексические, семантические и синтаксические особенности интернет-текстов, а также характерные черты блогов. Мы также рассмотрели способы выражения экспрессивности в таких текстах и то, как они применяются при анализе тональности.

Мы также перечислили основные инструменты нашего исследования и их функции, которые мы использовали. В частности, мы рассказали о принципах работы программ Sketch Engine и SentiStrength, а также о свойствах краудсорсингового словаря тональности LinisCrowd.

Основные выводы были сделаны в результате ряда экспериментов по изучению тональности и синтаксических структур текстов блогов. Как мы уже упоминали, авторы текстов различаются по возрасту, образованию и опыту работы. С помощью программ Sketch Engine и SentiStrength и тонального словаря LinisCrowd мы провели анализ тональности наших текстов и сравнили синтаксические структуры словосочетания и предложений. Основные выводы нашего исследования следующие:

1. Анализ тональности текстов показал большой процент использования различных средств выражения экспрессивности в текстах блогов волонтерских стажировок (114219 вхождений при общем объеме корпуса 370000 вхождений), в частности, эмоционально окрашенных слов, предложения, эмотиконов, удлинений слов и т.д. В блогах профессиональных стажировок ситуация иная: было извлечено всего лишь 9575 вхождений средств выражения экспрессивности при объеме корпуса

350000 словоупотреблений. Таким образом, тексты блогов волонтерских стажировок наиболее насыщенны в эмоциональном плане: авторы блогов о волонтерских стажировках используют в 12 раз больше различных средств выражения экспрессивной окраски, чем авторы блогов о профессиональных стажировках.

2. Были описаны основные синтаксические структуры словосочетаний в русском языке, которые были найдены в текстах блогов с помощью регулярных выражений. Процент их распределения в текстах блогов одинаковый, таким образом можно сделать вывод о том, что в плане синтаксиса словосочетаний различий практически нет.

3. Синтаксис простых предложений представляет собой существенное различие в текстах блогов. Тексты волонтерских стажировок являются более образными за счет использования различных фразеологизированных структур и именных предложений, в то время как блоги профессиональных стажировок напротив более формальны и содержат множество глагольных структур.

Таким образом, нами были выявлены основные характеристики, которые в дальнейшем возможно использовать при определении авторства интернет-блогов. В ходе работы была подтверждена гипотеза о том, что способ построения предложений коррелирует с типом автора. Авторы текстов волонтерских стажировок пишут более эмоционально, в их блогах присутствует большее количество не только тонально окрашенной лексики, но и иных средств выражения экспрессивности, в том числе для передачи устной речи. Пишущие о профессиональных стажировках используют конструкции с глаголами и строят более формальные высказывания.

Стоит отметить, что существует не так много исследований, посвященных анализу тональности средств выражения эмоциональности, кроме эмоционально окрашенных слов и предложений, а также типичным синтаксическим структурам словосочетаний и предложений.

В будущем можно продолжить исследование в данном направлении, используя дополненный ряд лингвистических характеристик, напр., морфологические и семантические признаки, а также в плане создания отдельных программ для определения авторства с применением данных признаков.

Перечень принятых терминов

Авторский инвариант — количественная или иная характеристика текстов, который определяет произведения одного автора.

Авторское профилирование — определение на основе определенных лингвистических характеристик принадлежность текста творчеству определенного автора.

Анализ тональности текста (англ. Sentiment analysis, Opinion mining) — это метод анализа содержания в компьютерной лингвистике, при котором автоматически в текстах выделяются эмоционально окрашенные слова и другие средства выражения экспрессивности.

Блог — это сетевой дневник. Место в Интернете (обычно в социальной сети), куда можно публиковать события из жизни и другую личную информацию.

Дискурс — связный текст в совокупности с различными социальными, культурными и др. факторами.

Искусственная нейронная сеть — математическая модель и её воплощение, которая построена на основании функционирования биологических нейронных сетей.

Кластерный анализ — статистический метод, при котором выполняется сбор данных, характеризующих выборку объектов, и классифицируются объекты по определенным группам.

Корпус — обработанная совокупность текстов, которая используется в качестве материала для исследования языка.

Метод опорных векторов — совокупность алгоритмов обучения с учителем, которые используются для задач классификации.

Норма языковая — совокупность используемых языковых средств, признаваемых носителями языка в качестве наиболее правильных в данный период времени.

Пост — отдельно взятое сообщение в блоге.

Регулярные выражения — формальный язык поиска и осуществления манипуляций с подстроками в тексте, основанный на использовании метасимволов.

Рунет — часть сайтов Интернета на русском языке.

Смайлик — это графическое изображение человеческого лица, отображающего определенную эмоцию.

Социальная сеть — бесплатная площадка в Интернете, где можно самостоятельно публиковать какую-то информацию и обмениваться ею с другими людьми.

Тэг — неструктурированное ключевое слово, относящееся к определенным частям информации.

Удлинение — множественное повторение в слове той или иной гласной буквы для выражения большей эмоциональности.

Хэштег — пометка или тег, который используется в блогах или социальных сетях для облегчения поиска сообщений по определенной теме.

Эмотикон (от английского emotion icon — «эмоциональная иконка») — это пиктограмма, которая служит для выражения эмоции в языке Интернета. Она выражается в сочетании типографических знаков (например, *:), ;)* и другие).

Список литературы

1. Арутюнова Н.Д. Дискурс [Текст] // Лингвистический энциклопедический словарь. М., 1990. С.136.
2. Ахманова, О.С. Словарь лингвистических терминов [Текст] // М.: Советская энциклопедия, 1966.
3. Бергельсон М.Б. Языковые аспекты виртуальной коммуникации (языковое поведение в сети Интернет) [Текст] // Вестн. МГУ. Сер. 19. Лингвистика и межкультурная коммуникация. 2002. №1. С.55-67.
4. В поисках потерянного автора: этюды атрибуции [Текст] / М. А. Марусенко, Б. Л. Бессонов, Л. М. Богданова и др. – СПб. : Филол. фак. С.-Петерб. гос. ун-та, 2001. – с. 209.
5. Горошко Е.И. Интернет-жанр и функционирование языка в Интернете: попытка рефлексии [Текст] / Е.И. Горошко // Жанры речи. - Саратов: Издательский центр «Наука», 2009. - Выпуск 6 «Жанр и язык». - С.11-127.
6. Горошко Е. И. Теоретический анализ Интернет-жанров [Текст] / Е.И. Горошко // Жанры речи. Выпуск 5 «Жанр и культура». - Саратов: Издательский центр «Наука», 2007.
7. Дюрдева П.С. Автоматическое определение автора текста на основе распределения частот буквосочетаний [Текст]: диплом. работа / Дюрдева Полина Сергеевна. — Санкт-Петербург, 2016. — с. 4-7.
8. Иванов Л. Ю. Язык Интернета: заметки лингвиста [Электронный ресурс] / Л. Ю. Иванов // Словарь и культура устной речи. - М.: Азбуковник, 2000. - С. 131-147. URL: www.ivanoff.ru/rus/ozhweb.htm. Дата обращения: 28.10.2016
9. Казаков В.П. Словосочетание. Аспекты характеристики словосочетания [Текст] / В.П. Казаков // Синтаксис современного русского языка. — СПбГУ, 2013. — с. 47-48.
10. Карнуп Е.В. Многоязычная коммуникация в сети Интернет как сфера реализации механизмов компрессии сообщений (на материале микроблогов

системы Твиттер) [Текст]: дис. … канд. филол. наук: 10.02.21 / Карнуп Екатерина Владимировна. — Санкт-Петербург, 2014. — с. 36-67.

11. Кольцов С. Н., Павлова Ю., Кольцова О. Ю. Метод автоматического анализа тональности текста в применении к социологическим задачам [Электронный ресурс] // Методическое пособие. М., 2012. URL: http://openbooks.ifmo.ru/ru/file/2203/2203.pdf. Дата обращения: 18.04.2017

12. Кувшинская Ю.М. Аббревиация в речи интернет-форумов [Текст] / Ю.М. Кувшинская // Современный русский язык в интернете. — М., 2014. — с. 23-38.

13. Кузнецова, Н. В. Фонетическое письмо в интернет-коммуникации в сопоставлении с другими типами текстов (по материалам национального корпуса русского языка, www.ruscorpora.ru) [Текст] / Н. В. Кузнецова // Духовные основы славянской культуры в народном сознании поколений. – Тюмень : Вектор бук, 2009. – С. 121–124.

14. Куликова А.В. Особенности Интернет-коммуникаций [Текст] / Куликова А.В. // Вестник Нижегородского университета им. Н.И. Лобачевского. Серия «Социальные науки», 2012, №4(28), с.19-24.

15. Лаборатория цифрового общества [Электронный ресурс]. URL: http://digsolab.ru/. Дата обращения: 20.05.2017.

16. Марусенко М. А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов [Текст] / М. А. Марусенко. – Л.: Изд-во Ленингр. ун-та, 1990. – с. 164.

17. Мичурин, Д. С. Прецедентный поликодовый текст в вербально-изобразительной коммуникации интернет-сообществ (на материале русскоязычных имидж-форумов) [Текст]: дис. … канд. филол. наук : 10.02.19 / Мичурин Дмитрий Сергеевич. – Тверь, 2014. – 162 с.

18. Морослин П.В. Структурно-семантические параметры веб-блогов как особого речевого жанра [Текст] // Вестник Тамбовского университета. Серия: гуманитарные науки. — 2009. - №12, с.332-337.

19. Мощенкова Д.С., Кривицкая Д.А., Амосова Н.С. Обзор программных продуктов разработанных для атрибуции художественных текстов [Электронный ресурс] // Молодежь и наука: сборник материалов X Юбилейной Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых с международным участием, посвященной 80-летию образования Красноярского края. — Красноярск: Сибирский федеральный ун-т, 2014. URL: http://elib.sfu-kras.ru/bitstream/handle/2311/17293/s43_010.pdf?sequence=1&isAllowed=y. Дата обращения: 15.05.2017.

20. Объяснительный словарь русского языка: Структурные слова: предлоги, союз, частицы, междометия, вводные слова, местоимения, числительные, связанные слова [Текст] // Гос. ин-т рус. яз. им. А. С. Пушкина; В. В. Морковкин, Н. М. Луцкая, Г. Ф. Богачёва и др.; Под ред. В. В. Морковкина. -2-е изд., испр. – М.: ООО «Издательство Астрель», 2003.

21. Ожегов С. И. Словарь русского языка [Текст]. — М., 1974.

22. Павлова Ю.В. Выявление социально значимых тем в блогах (на примере Живого Журнала) [Текст] // Магистерская диссертация, Высшая Школа Экономики, Санкт-Петербург, 2012.

23. Пазельская А.Г., Соловьев А.Н. Метод определения эмоций в текстах на русском языке [Электронный ресурс]. — М., 2011. URL: http://www.dialog-21.ru/media/1451/50.pdf. Дата обращения: 25.04.2017.

24. Плисецкая А.Д. О языковых и риторических стратегиях выражения оценки у пользователей социальной сети Фейсбук [Текст] / А.Д. Плисецкая// Современный русский язык в интернете. — М., 2014. — с. 83-92.

25. Романов А.С. Методика и программный комплекс для идентификации автора неизвестного текста [Текст]: дис. … канд. техн. наук : 05.13.18 / Романов Александр Сергеевич. – Томск, 2010. – с 5.

26. Русская грамматика [Текст] / Под ред. Н.Ю. Шведовой. Т.I. М., 1980. — с.262.

27. Русская грамматика [Текст] / Под ред. Н.Ю. Шведовой. Т.II. М., 1980. - с. 98-100.

28. Селютин А.А. Жанры как форма коммуникативного выражения онлайновой личности [Текст]/ А.А. Селютин// Вестник Челябинского государственного университета. - 2009. - №35 (173). - Филология. Искусствоведение. - Вып.37. - с. 138-141.

29. Синелева А. В. Атрибуция «Романа с кокаином»: лингвостатистическое исследование [Текст] : автореф. дис. … канд. филол. наук : 10.02.21 / Синелева Анастасия Васильевна ; Санкт-Петербургский гос. ун-т. – СПб: изд-во СПбГУ, 2001. – с. 22.

30. Тональный словарь веб-ресурса Linis Crowd [Электронный ресурс]. URL: http://linis-crowd.org/. Дата обращения: 21.10.2016

31. Трофимова, Г. Н. К вопросу о специфике функционирования русского языка в Интернете (норма и узус) [Электронный ресурс] / Г. Н. Трофимова. – 198 2001. – URL: http://www.dialog-21.ru/archive/2001/volume1/1_39.htm. Дата обращения: 25.09.2016.

32. Argamon S., Koppel M., Fine J., Shimoni A. Gender, genre, and writing style in formal written texts [Текст]. — TEXT, 23, 2003. — pp.321-346.

33. Brody S., Diakopoulos N. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs [Электронный ресурс]. — 2011. URL: http://www.nickdiakopoulos.com/wp-content/uploads/2007/05/BrodyDiakopoulosEMNLP11pp.pdf. Дата обращения: 24.04.2017

34. Burger J., Henderson J., Kim G., Zarrella G.. Discriminating gender on twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing [Текст]. — EMNLP ’11, Stroudsburg, PA, USA, 2011. — pp. 1301–1309.

35. Chetviorkin I., Loukachevitch N. Extraction of Russian Sentiment Lexicon for Product Meta-Domain [Текст] // In Proceedings of COLING 2012: Technical Papers, 2012. pp. 593-610.

36. Goswami S., Sarkar S., Rustagi M. Stylometric analysis of bloggers' age and gender [Текст]. — The AAAI Press, 2009. — pp. 214-217.

37. Hao Wang, Jorge A. Sentiment Expression via Emoticons on Social Media [Электронный ресурс]. — San Jose, USA, 2015. URL: https://arxiv.org/ftp/arxiv/papers/1511/1511.02556.pdf Дата обращения: 20.04.2017

38. Holmes J., Meyerhoff M. The Handbook of Language and Gender [Текст]. — Blackwell Handbooks in Linguistics, Wiley, 2003. — pp. 43-47.

39. Kilgarriff Adam, Rychly Pavel, Smrz Pavel, Tugwell David. The Sketch Engine. In Proc EURALEX 2004, Lorient, France; Pp. 105–116.

40. Koltsova O.Y., Alexeeva S.V., Kolcov S.N. An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media [Текст] // Компьютерная лингвистика и интеллектуальные технологии. 2016. — с. 277-287.

41. Koppel M., Argamon S., Shimoni A. Automatically categorizing written texts by author gender [Текст] / Literary and Linguistic Computing, 17 (4). — Amsterdam, 2003. — pp. 401–412.

42. Meina M., Brodzinska K., Celmer B., Czokow M., Patera, M., Pezacki J. et al. Ensemble-based classification for author profiling using various features notebook for PAN at CLEF 2013 [Электронный ресурс]. — 2013. URL: http://www-users.mat.umk.pl/~mich/pub/clef_2013.pdf. Дата обращения: 17.05.2017.

43. Merriam-Webster Online [Электронный ресурс]. URL: http://www.merriam-webster.com/. Дата обращения: 03.11.2016.

44. Nguyen D., Gravel R., Trieschnigg D., Meder T. "How Old Do You Think I Am?": A Study of Language and Age in Twitter [Текст]. — Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 2013. — pp. 439-447.

45. Novak P.K., Smailović J., Sluban B., Mozetič I. Sentiment of Emojis [Электронный ресурс]. — 2015. URL: https://doi.org/10.1371/journal.pone.0144296. Дата обращения: 17.04.2017.

46. Nowson S. The language of weblogs: a study of genre and individual differences. PhD Thesus (Unpublished manuscript) [Текст] // S. Nowson. — University of Edinburgh, 2006. — p. 279.

47. Peersman C., Daelemans W., Vaerenbergh L. Predicting age and gender in online social networks [Текст]. — In Proceedings of the 3rd international workshop on Search and mining user-generated contents, SMUC '11, New York, NY, USA, 2011. — pp. 37–44.

48. Pennebaker J.W. The secret life of pronouns: What our words say about us [Текст] / Bloomsbury Press. — 2011. — pp. 154-157.

49. Rangel F., Rosso P., Koppel, M., Stamatatos, E., Inches, G. Overview of the author profiling task at PAN 2013 [Электронный ресурс]. — 2013. URL: http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-RangelEt2013.pdf. Дата обращения: 17.05.2017.

50. Rangel F., Rosso P. On the impact of emotions on author profiling [Текст] // Information Processing & Management, Volume 52, Issue 1. — 2016. — pp. 73-92.

51. Schler J., Koppel M., Argamon S., Pennebaker J. Effects of age and gender on blogging [Текст]. — In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, AAAI, 2006. — pp. 199–205.

52. SentiStrength [Электронный ресурс]. URL: http://sentistrength.wlv.ac.uk/index.html. Дата обращения: 11.04.2017.

53. Sketch Engine [Электронный ресурс]. URL: https://the.sketchengine.co.uk/auth/corpora/. Дата обращения: 05.05.2016.

54. Thelwall M., Buckley K., Paltoglou G., Cai D., Kappas A. Sentiment strength detection in short informal text [Текст] // Journal of the American Society for Information Science and Technology, 61(12), 2010. — pp. 2544-2558.

55. WAN2TLK?: ltle bk of txt msgs [Текст] / editor Gabrielle Mander. — London: Michael O'Mara Books, 2000. — p. 96.

56. Zhang C., Zhang P. Predicting gender from blog posts [Электронный ресурс]. — Technical report, Technical Report. University of Massachusetts Amherst, USA, 2010. URL: http://web.stanford.edu/~pyzhang/papers/gender_prediction.pdf. Дата обращения: 18.05.2017.

Приложение 1. Список тэгов, заменяющих смайлики

[EMOJI] — любые смайлики, кроме ниже перечисленных;

[EMOJI_SAD] — смайлики, выражающие грусть;

[EMOJI_JOY] — смайлики, выражающие радость;

[EMOJI_HEART] — смайлики в виде сердца;

[EMOJI_SMILE] — смайлики в виде улыбки;

[EMOJI_CRY] — плачущие смайлики.

Приложение 2. Список стоп-слов

Индия
Китай
Колумбия
Богота
Индонезия
th
Россия
Вэньчжоу
i
Виви
Сиан
Пекин
Индро
Бали
Чаной
Тиан
P.S.
de
s
Боготе
colombia
Митча
Германия
Адит
dec
IMG_
Ботеро
Боготу
Шанхай
Чане
Apr
Москва
Картахена
Lombok
Чунцине
петербург
Санти
m
Rp
Медельин
колумбийского
jan
Темангтунга
Мелани
feb
Лилис
nob
Даниелы
Чунцина
денис
лера
Фанг
Ирфан
Темангтунге
я.
Боробудур
Картахены
Диего
aiesec
Синине
Сьёу
Ван
тао
la
rd
Wonosobo
Ява
Медельина
Даниела
Боливара
del
Айсек
Чунцин
Стивеном
xx
Ломбока
Boyolali
Dieng
Барранкильи
Барранкилье
Картахене
Санта-Марта
Тайрона
Картахену
Museo
карлос
колумбийских
Колумбийцы
Кумбум
Лхасу
Сычуань
Wumajie
Тыковки

x
аня
Таобао
Семаранг
АДита
Рендра
цу
Джокджу
сша
t
don
ява
d
China
AIESEC
gtalant
Трансмиленио
Juan
самара
Go
doctor
катя
Чака
Синин
Цинь
Сиане
Тыковкой
Yen
тыковку
Чану
европа
Бромо
Митчи
FB
Take
Атине
тайвань
nd
маша
st
Тревора
Яве
сате
Лисининг
Убуд
Баунти
Прамбанан

Приложение 3. Список эмоционально окрашенных слов текстов о волонтерских стажировках

Слова с эмоциональной окраской	**Оценка**
хороший	1
огромный	1
красивый	1
интересный	1
понравиться	1
проблема	-1
спасибо	1
счастливый	1
вкусный	1
приятный	1
любимый	1
нравиться	1
сладкий	1
приятно	1
добрый	1
достаточно	1
эффект	1
дружеский	1
блеск	2
отвратительный	-2
наслаждение	1
подделка	-1
освобождать	1
юмор	1
усталость	-1
мудрость	1
ненароком	-1
предавать	-1
фигня	-2
осуществить	1

Слова с эмоциональной окраской	**Оценка**
неудачный	-2
неправильно	-1
спрятать	-1
значительный	1
почетный	2
искренний	1
игровой	2
бесценный	2
расстраиваться	-1
стрелять	-1
падение	-1
пасть	-1
наркотик	-1
врать	-1
обижать	-1
продуктивный	1
тюрьма	-1
пьяный	-2
внушительный	1
пасмурный	-1
стойкий	1
улучшение	1
доверять	1
выгодный	1
популярный	1
праздник	1
замечательный	1
остановка	-1
нужный	1
крутой	1

Слова с эмоциональной окраской	**Оценка**
никакой	-1
легкий	1
особый	1
живой	1
отличный	2
помощь	1
толпа	-1
сожаление	-1
яркий	1
шок	-1
уверенный	1
мечта	1
удовольствие	2
удивительный	1
плохой	-1
вечеринка	1
низкий	-1
классный	1
наслаждаться	1
уставать	-1
ужасно	-1
плохо	-1
мусор	-1
непонятный	-1
чудесный	1
супер	1
опасный	-1
прекрасно	1
удача	1
незабываемый	1

Слова с эмоциональной окраской	**Оценка**
звезда	1
дикий	-2
радовать	1
потерять	-1
постараться	1
счастие	2
запрещать	-1
мечтать	1
заставлять	-1
тяжелый	-1
радость	2
восторг	2
развлечение	1
осознавать	1
грязный	-2
улыбка	1
уютный	1
покупка	-1
неплохой	1
обожать	1
идеальный	1
искусство	1
удобный	1
официальный	1
тайвань	-1
радоваться	1
грустно	-2
шумный	-1
попытаться	1
дружелюбный	1

Слова с эмоциональной окраской	**Оценка**
разнообразный	1
поражать	-1
поддерживать	1
холод	-1
музыкальный	1
светлый	1
грязь	-1
живописный	1
лошадь	1
смешной	1
шокировать	-1
детство	1
привлекать	1
захватывать	-1
набирать	-1
смерть	-1
комфортный	1
праздничный	1
порадовать	1
солнечный	1
отказываться	-1
непонятно	-1
невыносимый	-2
священный	1
ошибка	-1
рекомендовать	1
обучение	1
грустный	-1
заблудиться	-1
разводить	-1

Слова с эмоциональной окраской	**Оценка**
ненавидеть	-2
требовать	-1
война	-1
умирать	-1
трудный	-1
натуральный	1
спокойствие	1
позитивный	1
неплохо	1
убивать	-2
надежда	1
герой	1
красочный	1
гореть	-1
умный	1
государство	1
жуткий	-1
уважать	1
любоваться	1
злой	-1
правильный	1
удобно	1
дарить	1
скучный	-1
заботиться	1
гордиться	1
падать	-1
ленивый	-1
веселиться	1
непередаваемый	1

Слова с эмоциональной окраской	**Оценка**
положительный	1
целовать	1
уважение	1
добро	1
изучение	1
слабый	-1
предвкушение	1
понимание	1
гордость	1
увлекательный	1
скромный	1
черт	-1
золотой	1
отказывать	-1
волнение	-1
уверенность	1
приветливый	1
куртка	1
посуда	-1
прилетать	-1
наследие	1
раздражать	-1
успокаивать	1
чистота	1
удачно	1
подводить	-1
успешно	2
династия	1
улыбчивый	1
посоветовать	1

Слова с эмоциональной окраской	**Оценка**
посчастливиться	1
вдохновлять	1
благополучно	1
неописуемый	1
трястись	-1
купание	1
празднование	1
неудобно	-1
жертва	-2
комфортно	1
паника	-1
ошибаться	-1
магический	1
равнодушный	-1
тормозить	-1
справляться	1
пугать	1
разочарование	-1
удачный	1
вечный	1
крепкий	1
полюбоваться	1
сочный	1
могила	-1
экологический	1
вредный	-1
нервничать	-1
роскошный	1
печальный	-1
невольно	-1

Слова с эмоциональной окраской	**Оценка**
поздравлять	1
разрушать	-2
доброта	1
дырка	-1
развивать	1
испугаться	-1
подозревать	-1
утомительный	-1
комплимент	1
огонек	1
общительный	1
впечатлять	1
теряться	-1
хрустеть	-1
прилично	1
смущать	-1
долг	-1
легкость	1
преимущество	1
недовольный	-1
влюбляться	1
неприятный	-1
подружиться	1
записывать	-1
отсутствовать	-1
расплакаться	-1
местечко	1
грех	-1
назойливый	-1
упускать	-1

Слова с эмоциональной окраской	**Оценка**
необъятный	1
симпатичный	1
конструкция	1
твердый	-1
крохотный	-1
успех	1
улучшать	1
унывать	-1
сомнительный	-1
лекарство	1
шедевр	1
защита	1
гармония	1
кладбище	-1
восхитительный	2
терпеть	-1
впечатляющий	1
хитрый	1
испортить	-1
ценный	1
отражать	1
париж	1
зло	-1
нетерпение	-1
напугать	-1
познавательный	1
приправа	1
приемлемый	1
визжать	-1
столб	1

Слова с эмоциональной окраской	**Оценка**
безграничный	1
медленный	-1
драться	-1
мудрый	1
истинный	1
обманывать	-1
привлекательный	1
ненормальный	-1
ценность	1
невозможный	-1
радушный	1
мучить	-1
внимательный	1
загрязнение	-1
оригинальный	1
знаменательный	1
радостный	1
повар	1
призывать	1
образовательный	1
полноценный	1
немыслимый	-1
трудолюбивый	2
игнорировать	-1
рыдать	-1
заядлый	-1
нелепый	-1
восхищать	2
исследование	1
преследовать	-1

Слова с эмоциональной окраской	**Оценка**
страдание	-1
удобство	1
необыкновенный	1
стыдно	-2
очаровательный	2
проникать	-1
казахстан	-1
наказание	-2
познавать	1
достойный	1
величественный	2
безопасно	2
почет	1
доверие	1
комфортабельный	1
тощий	-1
наивный	-1
драгоценный	1
оружие	-1
скучноватый	-1
дружить	1
могучий	1
победа	1
элитный	1
негативный	-1
доброжелательный	1
опасаться	-1
романтичный	1
кризис	-1
хорошенько	1

Слова с эмоциональной окраской	**Оценка**
расплачиваться	-1
кошмар	-1
структура	2
неприятность	-1
несказанный	1
фешенебельный	1
жаловаться	-1
терпеливый	1
поцелуй	1
ненужный	-1
погибнуть	-2
познание	1
испортиться	-1
огорчать	-1
ругаться	-1
коза	-1
пролегать	-1
взятка	-1
уныние	-1
логотип	-1
достижение	1
хрупкий	-1
охрана	1
мертвый	-2

Приложение 4. Список эмоционально окрашенных слов текстов о профессиональных стажировках

Слова с эмоциональной окраской	**Оценка**
хороший	1
интересный	1
красивый	1
огромный	1
проблема	-1
вкусный	1
вечеринка	1
нравиться	1
приятный	1
толпа	-1
популярный	1
отличный	2
легкий	1
праздник	1
приятно	1
любимый	1
добрый	1
дружелюбный	1
замечательный	1
понравиться	1
счастливый	1
нужный	1
никакой	-1
остановка	-1
радость	2
шок	-1
наслаждаться	1
звезда	1
уставать	-1
плохо	-1

Слова с эмоциональной окраской	**Оценка**
мечта	1
неплохой	1
особый	1
обожать	1
мечтать	1
спасибо	1
плохой	-1
живописный	1
гордиться	1
уверенный	1
восторг	2
искусство	1
грязь	-1
удобный	1
счастие	2

Приложение 5. Частотный список средств выражения эмоциональности в текстах волонтерских блогов

Средство	**Количество, вхождения**
;), :),),)) и т.д.	7878
!, !!, ??, ?! и т.д.	4819
:(, (, ((и т.д.	2465
Иностранные слова	2237
… и т.д.	2089
EMOJI	579
Удлинения букв *о*	569
Слова верхнего регистра	463
^_^, ^^, *_*, :D, oO	235
Удлинения букв *а*	230
EMOJI_JOY	202
Удлинения других гласных букв	202
EMOJI_SMILE	158
EMOJI_HEART	86
EMOJI_SAD	65
EMOJI_CRY	35

Приложение 6. Поисковые запросы для нахождения словосочетаний с помощью языка регулярных выражений

Тип словосочетания	**Подтип**	**Запрос**	**Пример словосочетания**
Именные	Субстантивные	[tag="N.*"][tag="N...g.."] [tag = "A.*|P.*"][tag = "N.*"]	*Центре города; последний день*
	Адъективные	[tag = "A.*"][tag="Sp.*"][tag="N.*"] [tag="R.*"][tag="A.*"]	*Похожее на парк; очень приятно*
	Числительное	[tag = "M.*"][tag="Sp.*"] [tag="N.*"] [tag="M.*"][tag="N.*"]	*Первый в жизни; несколько дней*
	Местоимение	[tag="P.*"][tag="A.*"] [tag="P.*"][tag="Sp.*"][tag="P.*"]	*Самое интересное; каждому из нас*
Глагольные	Существительное	[tag="V.*"][tag="N.*"] [tag="V.*"][tag="Sp.*"][tag="N.*"]	*Провести время; сижу в аэропорту*
	Наречие	[tag="R.*"][tag="V.*"]	*Уже писала*
	Инфинитив	[tag="V.n.*"][tag="V.*"]	*Ложиться спать*
	Деепричастие	[tag="V.*"][tag="V.g.*"] [tag="V.g.*"][tag="Sp.*"] [tag="N.*"]	*Ехали сидя;* *Выйдя из машины*
	Причастие	[tag="V.p.*"][tag="Sp.*"][tag="N.*"]	*Созданным для продвижения*
Наречные		[tag="R.*"][tag="Sp.*"][tag="N.*"] [tag="R.*"][tag="R.*"]	*Впервые в жизни; очень быстро*
Междометия		[tag="I.*"]	*Хаха*

Приложение 7. Пример частотного списка именных (субстантивных) словосочетаний текстов блогов профессиональных стажировок (первые 100 словосочетаний)

Словосочетания	**Количество, вхождения**
Записки путешественника	54
мистера Индро	14
День Рождения	14
центре города	11
конце концов	9
уровнем моря	8
пару дней	7
центр города	6
день рождения	6
большинстве случаев	6
часах езды	5
количество людей	5
часть дня	4
часе езды	4
уголков мира	4
станции метро	4
пару раз	4
пару минут	4
отца Стивена	4
острова Ява	4
время года	4
вершине горы	4
Деда Мороза	4
что-то типа	3
частях города	3
часть пути	3
часов вечера	3
центру города	3
цвет лица	3
точки зрения	3

Словосочетания	**Количество, вхождения**
территории фабрики	3
слова благодарности	3
половину дня	3
половина пути	3
пол года	3
площадь Боливара	3
отец Стивена	3
образ жизни	3
новость дня	3
названия динозавров	3
миллионов человек	3
конце дня	3
интернс хауса	3
из-под крана	3
зону комфорта	3
директор школы	3
горы мусора	3
городка Чака	3
город Колумбии	3
время пребывания	3
время праздников	3
большинство мусульман	3
бабуле Стивена	3
Пару дней	3
языка жестов	2
юге Китая	2
шквал аплодисментов	2
членов семьи	2
черте города	2
чемпионата мира	2

Словосочетания	**Количество, вхождения**
частях страны	2
часть города	2
часть волос	2
часть Китая	2
части города	2
части Китая	2
часе ходьбы	2
центра города	2
цветами радуги	2
функционирования организации	2
фраза коллеги	2
форме лотоса	2
учитель математики	2
учитель информатики	2
упрощению работы	2
улицах Боготы	2
уголков планеты	2
тысячи Будд	2
тонкости функционирования	2
толпы туристов	2
толпа индийцев	2
течение года	2
территории музея	2
сфере преподавания	2
сумму денег	2
стороне улицы	2
стаканом воды	2
сотни туристов	2
сортов чая	2
следам путешествия	2

Словосочетания	**Количество, вхождения**
скоростью света	2
севере города	2
север Китая	2
свекрови Чаны	2
свадьбы ни-ни	2
родина камасутры	2
рода болезней	2
ресторане отеля	2
ремень безопасности	2
разрешения конфликтов	2

Приложение 8. Поисковые запросы для нахождения структурных схем простого предложения с помощью языка регулярных выражений

Структурная схема	**Запрос**	**Примеры**
N1 V3s	[tag="N.*"][tag="Vmi.*"]	*Холод стоял; день начался*
Vf3s Inf	[tag="V.i.3s.*"][tag="V.n.*"]	*Оставляет желать; будет означать*
N2 (neg)Vf3s	[tag="N...g.*"][tag="V.i.3s.*"]	*Времени остается*
N1 - N1	[tag="N...n.*"][lemma="\-"][tag="N...n.*"]	*Индонезия — страна; человек — возбудитель*
N1 - Adj1	[tag="N…n.*"][lemma="\-"][tag="A….n."][tag!="N...n.*"]{3,5}	*Водоемы — зеленые . Еще в самом начале*
N1 - N2	[tag="N...n.*"][lemma="\-"][tag="N...g.*"]	*Женщины — мужчин*
N1 — Inf	[tag="N...n.*"][lemma="\-"][tag="V.n.*|"]	*Турист — посетить*
Inf - N1	[tag="V.n.*"][lemma="\-"][tag="N...n.*"]	*Носить — часы*
Inf - Adv-o	[tag="V.n.*"][lemma="\-"][tag="R."]	*Ехать - обязательно*
Praed Inf	[word="можно|нельзя|надо…"][tag="V.n.*"]	*Надо сказать*
Praed (neg) N4/N2	[word="можно|нельзя|надо…"][tag="N...a.*|N...g.*"	*Нужно время*
Praed part N2	[tag="V.p.....s.*"][tag="N...g.*"]	*Лишено правды*
Adv quant - N2	[tag="R."][lemma="\-"][tag="N...g.*"]	*Совсем — одежды*
Нет N2	[word="нет"][tag="N...g.."]	*Нет денег; нет солнца*
Ни N2	[word="ни"][tag="N...g.."]	*Ни слова; ни телефона*
Никого (ничего) N2	[word="никого|ничего"][tag="N...g.."]	*Никого знакомого*
Никакого… N2	[word="никакого…"][tag="N...g.."]	*Никакого стыда*
N1	[lemma="\.|\!|\?\…"][tag="N…n.*"][lemma="\."]	*Продолжение; ночь*
N как N	1:[tag="N…n.*"][word="как"]2:[tag="N...n.*"]&1.tag=2.tag	*Работа как работа*
не до N	[word="не"][word="до"][tag="N.*"]	*Не до конца*
Inf так Inf	1:[tag="V.n.*"][word="так"]2:[tag="V.n.*"]&1.tag=2.tag	*Гулять так гулять*
ох, ах, эх N	[word="ох|ах|эх"][tag="N.*"]	*Ох рис*
что за N	[word="что"][word="за"][tag="N.*"]	*Что за стереотипы*

Приложение 9. Пример частотного списка структурной схемы простого предложений N1 - N1 текстов блогов профессиональных стажировок

Предложения	**Количество, вхождения**
усы - блеск	2
Чандигарх - город	2
Индонезия - страна	2
Индия - родина	2
Индия - место	2
человек - человек	1
феварль - апрель	1
украшения - заколка	1
туристы - сумки	1
сыр - панир	1
сын - отец	1
страна - ИНДОНЕЗИЯ	1
собор - церковь	1
слово - закон	1
семья - мусульмане	1
раз - Чане	1
праздник - песни	1
поэт - Хосе	1
питчер - перебежка	1
переводчики - парни	1
парень - Исхан	1
папа - директор	1
остановка - аэропорт	1
одежда - платье	1
овечки - жизнь	1
обстановка - минимализм	1
обряд - Дугжууба	1
образования - результат	1
название - улица	1
мусульмане - верующие	1

Предложения	Количество, вхождения
мужчины - родственники	1
месяцы - декабрь	1
место - мостики	1
место - вид	1
магазинчики - Йен	1
люди - плюющиеся	1
крем - всё	1
корабли - выбор	1
кондукторы - всё	1
комплекс - место	1
кожа - мешок	1
идея - девочки	1
занятие - простор	1
жизнь - ИГРА	1
животные - члены	1
диалог - встреча	1
дети - язык	1
городок - Санта	1
главное - Дух	1
воскресенье - выходной	1
виду - трава	1
вид - Чарминар	1
будущее - туман	1
брат - владелец	1
алкоголь - коньяк	1
аккорд - чистка	1
Ящерица - друг	1
Человек - существо	1
Цинь - Цинь	1
Холи - праздник	1

Предложения	**Количество, вхождения**
Холи - Буддизм	1
Фанг - няня	1
Тыковка - любимица	1
Танзания - рай	1
Тайвань - часть	1
Сиан - ведьма	1
Свадьба - Холи	1
Санкт-Петербург - Иркутск	1
Рендра - музыкант	1
Нескафе - преступление	1
Любовь - суть	1
Лисининг - девочка	1
Колумбия - рай	1
Коллеги - золото	1
Иян - сын	1
Итог - каноеинг	1
Индонезийцы - народ	1
Индия - центр	1
Инан - ученица	1
Жених - коллега	1
Дидик - учитель	1
Джо - образец	1
Джайпур - город	1
Гэри - ребята	1
Вино - тысячи	1
Весна - пора	1
Буддизм - Кинематограф	1
Автор - архитектор	1