А.Б. Борунов, В.Т. Малыгин

Международный гуманитарно-лингвистический институт, г. Москва
Финансовый университет при Правительстве РФ, Владимирский филиал, г. Владимир

ИССЛЕДОВАНИЕ АНГЛОЯЗЫЧНОГО АВТОРСКОГО КОРПУСА РЭГУ Н. МИТРЫ: ОПЫТ ОБРАБОТКИ ТЕКСТА КОМПЬЮТЕРНОЙ ПРОГРАММОЙ "WORDSMITH TOOLS"

THE ENGLISH-LANGUAGE WRITER RAGHU N. MITRA'S CORPUS STUDY: THE EXPERIENCE OF WORD-PROCESSING WITH COMPUTER PROGRAM "WORDSMITH TOOLS"

Ключевые слова: компьютерные исследования корпуса, Р. Н. Митра, статистическая лингвистика, методика обработки корпуса программой "WORDSMITH TOOLS"

Keywords: a corpus computer research, R. N. Mitra, statistical linguistics, a method of word-processing with program "WORDSMITH TOOLS"

В статье описывается проведение экспериментального статистического исследования авторского корпуса с применением компьютерной программы "WORDSMITH TOOLS", целью исследования является описание методики подготовки корпуса, функционала программы обработки текста, определение исходных данных авторского корпуса для дальнейшего автоматизированного составления алфавитно-частотного конкорданса языка писателя и авторских словарей.

This article describes a statistical research of the author's corpus, using computer program "WORDSMITH TOOLS", the aim of the study is to describe the methods of a corpus preparation, the functionality of the text processing software, the initial data of the writer's corpus for further preparation of the alphabetical-frequency concordance of the writer's language and writer's dictionaries.

Компьютерные исследования текста начали развиваться в СССР и за рубежом еще в середине XX в., с внедрением первых ЭВМ, которые позволили автоматизировано обрабатывать большие тексты, а во-второй половине XX в. появляются крупные исследования в данной области, отметим, например, следующие научные публикации «Вопросы статистической стилистики» (Вопросы…, 1974), «Ученые записки Тартуского университета. Квантитативная лингвистика и автоматический анализ текстов» (Ученые записки…, 1990) и другие.

Компьютерные исследования корпуса предполагают обработку больших объемов текстов для подтверждения гипотезы и формулирования выводов, однако, в середине XX в. ЭВМ не обладали теми ресурсами как сегодня и не могли полностью удовлетворить запросы исследователей.

Сейчас компьютеры стали более совершенными, автоматическая обработка больших корпусов текстов занимает несколько минут, кроме того мир стал компьютеризирован, человечество не представляет свою жизнь без IT-технологий и интернета, поэтому в XXI в. наблюдается обращение исследователей к вопросам лингвостатистики, компьютерной обработки текстов, созданию корпусов, применению компьютерных программ при составлении конкордансов и глоссариев. Современные компьютеры позволяют быстро обрабатывать введенную информацию, составлять отчеты по частотному словоупотреблению, например, алфавитно-частотные или частотные конкордансы, выделить коллокаты, кроме того, компьютер позволяет сохранять результаты в удобном для дальнейшей работы формате. О возросшем интересе к компьютерной обработке текста в XXI в. косвенно свидетельствует факт появления при университетах специализированных кафедр, так, например, в РГСУ в 2011 была создана кафедра компьютерной лингвистики, одноименная кафедра появилась в МФТИ, в РГПУ им. А.И. Герцена образовалась кафедра прикладной лингвистики, а в МГЛУ появился Институт прикладной и математической лингвистики, формирование подобных кафедр и научных школ, занимающихся проблемами квантитативной лингвистики, лингвостатистики, компьютерной лингвистики, корпусных исследований наблюдается и во многих других вузах нашей страны и зарубежья. Развитием профильных кафедр в ВУЗах, удобством обработки текстов для лингвиста, открытие новых перспектив исследования объясняет тот факт, что в русле данной проблематики был защищён ряд кандидатских диссертаций, назовём например следующие, Кириллов М.А. «Лингвостатистический анализ художественного текста на материале коротких рассказов Ф. С. Фицджеральда» (Кириллов, 2002), Клочко А.Д. «Оптимизация индивидуальных лингвистических исследований средствами специализированной базы данных» (Клочко, 2006), Краев С. В. «Ядерные служебные слова в русском подъязыке информатики: квантитативно-квалитативное исследование» (Краев, 2007), а также совсем недавно докторская диссертация Воевудской О.М. «Концепция идеографического словаря основного лексического фонда германских языков» (Воевудская, 2016) и ряд других.

На сегодняшний момент в научном обиходе сосуществуют несколько, кажущихся на первый взгляд, синонимичных понятий «прикладная лингвистика», «лингвостатистика» (статистическая лингвистика), «корпусная лингвистика», «квантитативная лингвистика», «компьютерная лингвистика». Прикладная лингвистика кажется наиболее широким понятием, включающим в себя остальные науки. Другие частные области прикладной лингвистики, а именно «лингвостатистика», «корпусная лингвистика», «квантитативная лингвистика», «компьютерная лингвистика» еще не получили явного разграничения в работах отечественных учёных, так «лингвостатистика» и «квантитативная лингвистика» часто выступают полными синонимами, как и пара «корпусная и компьютерная лингвистика», ввиду смежности области исследования – корпус исследуется компьютерными программами, либо с

квалитативной лингвистикой – опираясь на то, что проводится кватитативный анализ корпуса. Отметим, что предметную область компьютерной лингвистики успешно описал в своей статье Яцко В.А. (Яцко, 2014), остальные дисциплины всё ещё обладают некой «транспарентностью» и требуют уточнений.

В данной статье мы намеренно не выставляем в заглавие одну из перечисленных выше дисциплин, а оперируем термином *корпусные исследования*, которое представляется нам наиболее удачно отвечающим задачам исследования, так как, в первую очередь, ставится задача квантитативного характера, а именно обработка текста компьютерными программами и экспериментальный автоматизированный количественный подсчет ЛЕ корпуса для дальнейшего составления частотного конкорданса авторского корпуса.

В наши дни написано множество разнообразных программ компьютерного анализа текста, отвечающих разным задачам в области корпусных исследований, так например программы для частотного анализа корпуса, программы для автоматического моделирования словарей, программы определения авторства, программы автоматизированного перевода и электронные словари и другие.

Фактическим материалом для исследования послужили данные анализа англоязычных текстов Рэгу Н. Митры с помощью программного обеспечения WordSmith Tools.

Материалом для компьютерной обработки послужили 4 книги Р.Н. Митры на английском языке:

1. «Очень банальная страсть» “A Very Insipid Passion” (опубликован в США в 1999 г. и на английском языке в России в 2002 г. Перевод названия романа приводится по изданию: Mitra, R.N. A Very Insipid Passion. – M.: Manager, 2002) (Mitra, 2002). Отметим, что в России было опубликовано только 2 произведения Р. Н. Митры и только на английском языке. Данные о переводе его произведений на русский язык на данный момент отсутствуют;

2) «Грехопадение» “Impute Fall to Sin” (опубликован в США в 2003 г. и на английском языке в России в 2005 г. Перевод названия романа приводится по изданию: Mitra, R.N. Impute Fall to Sin. – M.: Manager, 2005) (Mitra, 2005);

3) «Дождь из теней» “A Rain Full of Ghosts” (опубликован в США в 2004 г.) (Mitra, 2004);

4) «Если бы не смерть» “If there wasn't death” (опубликован в США в 2007 г.) (Mitra, 2007), а также 2 фрагмента других произведений писателя, представленных в свободном доступе в сети Интернет (“As in the falling of an eyelid”, “At The Davies”).

Для проведения анализа текста использовалась компьютерная программа WordSmith Tools (WordSmith Tools). Данная программа текстового анализа обладает широким функционалом, а именно: позволяет проводить анализ коллокат (контекстное словоупотребление - *семантический* анализ), поиск слов по тексту(ам), лемматизация (выделение лемм – *морфологический* анализ); анализ длины слов в тексте(ах), построение

списка многочленных сочетаний, создание списка ключевых слов, диаграммы ключевых слов, т.е. работа с токенами (*графематический* анализ). Все эти процедуры составляют этапы комплексного статистического анализа корпуса, одним из основных результатов которого является построение частотного/алфавитно-частотного конкорданса.

Кроме того, программа включает в свой состав несколько полезных утилит, а именно генерация списка слов для заданной совокупности текстовых файлов (позволяет сохранять результаты в формате pdf, Excel и других форматах); разбиение больших текстов на совокупность фрагментов; пакетного редактирования множества текстов и другие.

Проведение лингвостатистического анализа корпуса подразумевает под собой ряд этапов:

1) первоначальный подготовительный этап;

2) машинная обработка;

3) анализ результатов.

Опишем методику проведения исследования. Задача первоначального *подготовительного этапа* состояла в том, чтобы отобрать тексты для создания корпуса, в нашем случае книги Р. Н. Митры. В дальнейшем необходимо было перевести данные книги в электронный вид путем сканирования, автоматически перевести полученное сканирующим устройством изображение в текст, далее провести вычитку результатов ручным способом, чтобы предотвратить погрешности распознавания текста программным обеспечением, в нашем случае ABBYY FineReader.

Данный подготовительный процесс представляется наиболее трудоёмким, так как требует продолжительной механической работы, а именно сканирования нескольких сотен страниц вручную, автоматического распознавания текста и перевода его в формат .doc, затем исследователь должен провести сверку отсканированного и распознанного текста с печатной версией книги, исправить неточности и погрешности автоматического распознавания текста, т.е. внимательно просмотреть все страницы полученного документа.

Итогом данного этапа стало то, что каждый отдельный файл с произведением был проверен на наличие погрешностей распознавания, найденные недочеты были исправлены, затем все тексты были объединены в единый файл (авторский корпус). Промежуточным итогом первого этапа стало создание электронного авторского корпуса текстов Рэгу Н. Митры. После вычитки текст переводится из формата .doc в формат .txt блокнот. Данный формат является наиболее подходящим для компьютерной обработки текста программой WordSmith Tools.

Переходим ко *второму этапу* машинной обработки полученного корпуса. В интерфейс программы загружаем полученный текстовый файл (полученный нами авторский корпус англоязычных текстов Р. Н. Митры) с расширением .txt.

Программа WordSmith Tools позволяет проводить ряд операций с корпусом, в первую очередь, рассмотрим базовую функцию, которой

является статистическая обработка текста, с построением частотного/алфавитно-частотного конкорданса загруженного файла.

Рассмотрим общие данные статистики, предлагаемой данной компьютерной программой, которые включают:

а) число файлов, подвергающихся анализу (number of files involved in the word-list);

б) размер файла(ов) (в байтах) (file size (in bytes, i.e. characters);

в) число слов «токены» (running words in the text (tokens);

г) «токены», используемые для построения списка (tokens used in the list);

д) число различных типов слов (no. of different words (types);

е) отношение разных слов к общему числу произнесенных слов (type/token ratios);

ж) количество предложений в тексте (no. of sentences in the text);

з) значение длины предложения (mean sentence length (in words);

и) количество абзацев в тексте (no. of paragraphs in the text);

к) значение длины абзаца (mean paragraph length (in words);

л) количество n-буквенных слов (the number of n-letter words).

Нами было проведено 2 эксперимента с корпусом, преследовавших цель получения основных статистических параметров авторского корпуса. В эксперимент № 1 был включен весь корпус (4 произведения и 2 фрагмента – единым корпусом), в эксперимент № 2 мы включили 4 корпуса (из каждого произведения сформировали отдельный корпус, 2 фрагмента не были включены из-за их незначительного объёма).

Рассмотрим результаты эксперимента № 1, которые занесены в таблицу 1. При обработке общего файла корпуса текстов Р. Н. Митры были выделены исходные данные взятых для анализа текстов, которые составляют – 414 311 слов (столбец “tokens used for word list”), либо 2 273 083 знаков с пробелами (столбец “file size”).

Табл. 1 Статистические данные обработки авторского корпуса

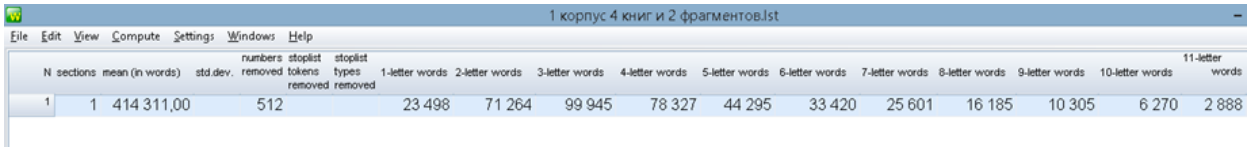
Отметим, что в англоязычной литературе используется термин “token”, что переносится в отечественную науку, где исследователи чаще прибегают к термину «токен», который понимается «как последовательность буквенных и/или цифровых символов, отделенную слева и справа знаками форматирования текста и/или препинания. Разбивка текста на токены называется токенизацией, а программы, выполняющие токенизацию, – токенайзерами» (Яцко, 2014: 26).

Программа позволяет подсчитать количество токенов с учетом их частотности в анализируемом отрывке (англоязычный термин *“type-token ratio”*). Токены подразделяются на уникальные и общие. Термин

«уникальный токен» используется для обозначения токена без учета количества его повторов в тексте, а термин «общий токен» – количество токенов с учетом их частотностей. Отметим, что количество общих токенов, как правило, больше количества уникальных токенов. Это позволяет при взвешивании терминов использовать вероятностные величины и устранить зависимость весовых коэффициентов от размера текста. Согласно машинному подсчету вероятностный коэффициент общих токенов в текстах Р. Н. Митры равно 4,68. Данная методика используется в психолингвистике при анализе спонтанной речи испытуемых в качестве оценок объема словаря используются две основных меры: коэффициент лексического разнообразия (отношение разных слов к общему числу слов, “type-token ratio”). В основе методики лежит идея о том, что чем больше словарный запас испытуемого, тем вероятнее встретить в его речи большее количество разных слов. Значения коэффициентов могут рассчитываться на фиксированное количество слов (токенов), на фиксированное количество высказываний.

В таблице 1 приведены данные STTR (“standardised or mean type/token ratio”), которые включают в себя каждую из форм слова, например “say” и “says” считаются разными единицами. Согласно информации с сайта разработчика программы, данные STTR рассчитываются следующим образом: “The number shown is a percentage of new types for every n tokens. That way you can compare type/token ratios across texts of differing lengths. This method contrasts with that of Tuldava who relies on a notion of 3 stages of accumulation. The WordSmith method of computing STTR was my own invention but parallels one of the methods devised by the mathematician David Malvern working with Brian Richards (University of Reading)” (WordSmith Tools). Согласно данным таблицы 1 STTR в авторском корпусе Р. Н. Митры составляет 55,8.

Кроме того, программа позволила нам подсчитать количество n-буквенных слов в корпусе, данные приведены в таблице, таблице 2, таблице 3. Данные указывают на то, что наиболее частотными в произведениях Р. Н. Митры являются 3-х буквенные слова, что может быть объяснено включением в подсчет определенного артикля “the” (24% от всего корпуса). Количество слов, состоящих из 1, 2 и 3 букв составляет около 47% корпуса, что объясняется обилием служебных слов: артиклей, предлогов, союзов состоящих в английском языке преимущественно от 1 до 3 букв. Отметим присутствие в корпусе 16, 17 и 20 буквенных слов, составляющих около 0,01% от всего корпуса. При проведении дальнейшего исследования было бы интересно сравнить количество n-буквенных слов в авторском корпусе Р. Н. Митры с процентном соотношением n-буквенных слов английского языка, а также с авторским корпусом другого писателя.

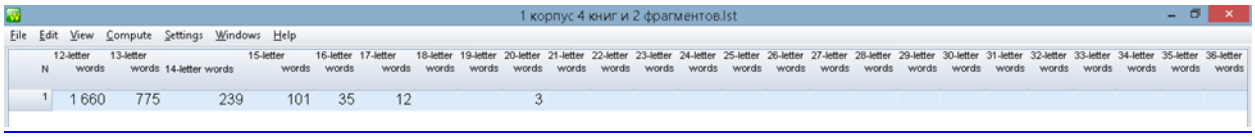


1 корпус 4 книг и 2 фрагментов.lst

File Edit View Compute Settings Windows Help

	N	sections	mean (in words)	std.dev.	numbers removed	stoplist tokens removed	stoplist types removed	1-letter words	2-letter words	3-letter words	4-letter words	5-letter words	6-letter words	7-letter words	8-letter words	9-letter words	10-letter words	11-letter words
1	1	1	414 311,00		512			23 498	71 264	99 945	78 327	44 295	33 420	25 601	16 185	10 305	6 270	2 888

Табл. 2 Статистические данные обработки авторского корпуса (продолжение)



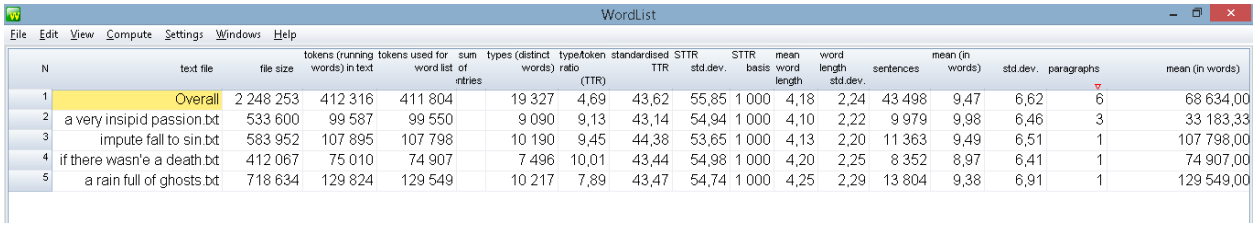
1 корпус 4 книг и 2 фрагментов.lst

File Edit View Compute Settings Windows Help

N	12-letter words	13-letter words	14-letter words	15-letter words	16-letter words	17-letter words	18-letter words	19-letter words	20-letter words	21-letter words	22-letter words	23-letter words	24-letter words	25-letter words	26-letter words	27-letter words	28-letter words	29-letter words	30-letter words	31-letter words	32-letter words	33-letter words	34-letter words	35-letter words	36-letter words
1	1 660	775	239	101	35	12			3																

Табл. 3 Статистические данные обработки авторского корпуса (продолжение)

Перейдем к обработке данных эксперимента № 2. В данном случае интерес представляет сравнение 4-х корпусов произведений.

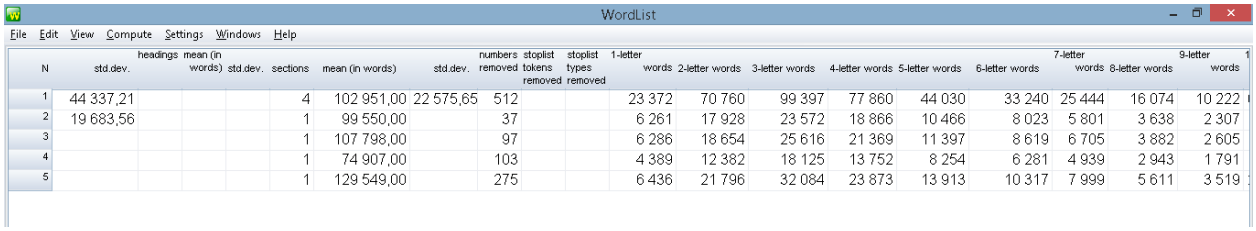


WordList

File Edit View Compute Settings Windows Help

N	text file	file size	tokens (running words) in text	tokens used for word list	sum of entries	types (distinct words)	type/token ratio (TTR)	standardised TTR	STTR std.dev.	STTR basis	mean word length	word length std.dev.	sentences	mean (in words)	std.dev.	paragraphs	mean (in words)
1	Overall	2 248 253	412 316	411 804		19 327	4,69	43,62	55,85	1 000	4,18	2,24	43 498	9,47	6,62	6	68 634,00
2	a very insipid passion.txt	533 600	99 587	99 550		9 090	9,13	43,14	54,94	1 000	4,10	2,22	9 979	9,98	6,46	3	33 183,33
3	impute fall to sin.txt	583 952	107 895	107 798		10 190	9,45	44,38	53,65	1 000	4,13	2,20	11 363	9,49	6,51	1	107 798,00
4	if there wasn'e a death.txt	412 067	75 010	74 907		7 496	10,01	43,44	54,98	1 000	4,20	2,25	8 352	8,97	6,41	1	74 907,00
5	a rain full of ghosts.txt	718 634	129 824	129 549		10 217	7,89	43,47	54,74	1 000	4,25	2,29	13 804	9,38	6,91	1	129 549,00

Табл. 4 Статистические данные обработки текстов Р.Н. Митры

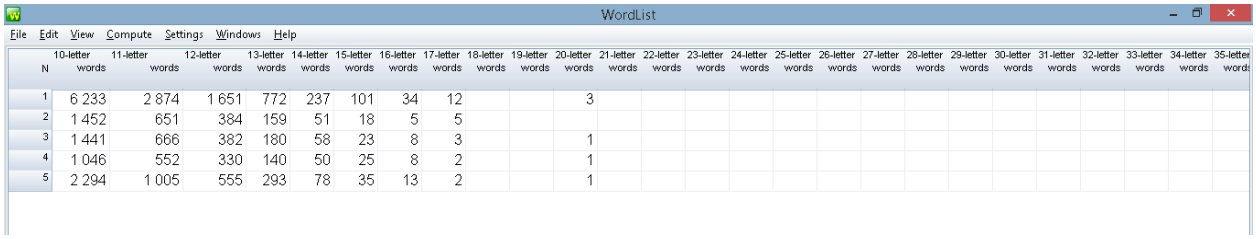


WordList

File Edit View Compute Settings Windows Help

N	std.dev.	headings mean (in words)	std.dev.	sections	mean (in words)	std.dev.	numbers removed	stoplist tokens removed	stoplist types removed	1-letter words	2-letter words	3-letter words	4-letter words	5-letter words	6-letter words	7-letter words	8-letter words	9-letter words
1	44 337,21			4	102 951,00	22 575,65	512			23 372	70 760	99 397	77 860	44 030	33 240	25 444	16 074	10 222
2	19 683,56			1	99 550,00		37			6 261	17 928	23 572	18 866	10 466	8 023	5 801	3 638	2 307
3				1	107 798,00		97			6 286	18 654	25 616	21 369	11 397	8 619	6 705	3 882	2 605
4				1	74 907,00		103			4 389	12 382	18 125	13 752	8 254	6 281	4 939	2 943	1 791
5				1	129 549,00		275			6 436	21 796	32 084	23 873	13 913	10 317	7 999	5 611	3 519

Табл. 5 Статистические данные обработки текстов Р.Н. Митры (продолжение)



WordList

File Edit View Compute Settings Windows Help

N	10-letter words	11-letter words	12-letter words	13-letter words	14-letter words	15-letter words	16-letter words	17-letter words	18-letter words	19-letter words	20-letter words	21-letter words	22-letter words	23-letter words	24-letter words	25-letter words	26-letter words	27-letter words	28-letter words	29-letter words	30-letter words	31-letter words	32-letter words	33-letter words	34-letter words	35-letter words
1	6 233	2 874	1 651	772	237	101	34	12			3															
2	1 452	651	384	159	51	18	5	5																		
3	1 441	666	382	180	58	23	8	3			1															
4	1 046	552	330	140	50	25	8	2			1															
5	2 294	1 005	555	293	78	35	13	2			1															

Табл. 6 Статистические данные обработки текстов Р. Н. Митры (продолжение)

Обрабатывая корпусы отдельных произведений Р. Н. Митры, отметим, что сравниваемые параметры сходны (табл. 4, табл. 5, табл. 6), а именно:

а) число токенов в каждом из 4 исследуемых произведений находится в диапазоне 75 000 – 130 000 в каждом произведении;

б) самые длинные слова в произведениях состоят из 20 букв;

в) вероятностный коэффициент общих токенов колеблется в диапазоне 43 – 44.

В данной статье представлены основные исходные данные авторского корпуса, которые могут быть использованы при продолжении исследования, а именно при проведении процедуры лемматизации, выделении коллокат,

построении алфавитно-частотного конкорданса и авторского словаря языка писателя.

Литература

1. Воевудская О.М. Концепция идеографического словаря основного лексического фонда германских языков. Дис. докт. филол. наук. – Воронеж, 2015. – 450 с.
2. Вопросы статистической стилистики. Сборник статей. – Киев: Из-во «Наукова думка», 1974. – 331 с.
3. Кириллов М.А. Лингвостатистический анализ художественного текста: на материале коротких рассказов Ф. С. Фицджеральда. Дис. канд. филол. н. – Иваново, 2002. – 292 с.
4. Клочко А.Д. Оптимизация индивидуальных лингвистических исследований средствами специализированной базы данных. Дис. канд. филол. н. – Армавир, 2006. – 263 с.
5. Краев С.В. Ядерные служебные слова в русском подъязыке информатики: квантитативно-квалитативное исследование. Дис. канд. филол. н. – Екатеринбург, 2007. – 306 с.
6. Ученые записки Тартуского университета. Квантитативная лингвистика и автоматический анализ текстов. – Тарту: Тартуский ун-т, 1990.
7. Яцко В.А. Предметная область компьютерной лингвистики // Вестник ИГЛУ. – Иркутск. – №2 (27). , 2014. – С. 24 - 35.
8. Mitra R.N. If there wasn't death. – Denver, Colorado: Outskirts Press Inc., 2007. – 230 p.
9. Mitra R.N. Impute Fall to Sin. – M.: Manager, 2005. – 336 p.
10. Mitra R.N. A Rain Full of Ghosts. – Baltimore: Publish America, 2004. – 366 p.
11. Mitra R.N. A Very Insipid Passion. – M.: Manager, 2002. – 336 c.
12. Mitra R.N. As in the falling of an eyelid (отрывок части книги). – [Электронный ресурс]. – Режим доступа: URL: http://www. members.tripod.com/~ShibaHill/eyelid.html, свободный. Дата обращения: 20. 02.2016.
13. Mitra R.N. At The Davies: A Novel of Medical Life (отрывок части книги). – [Электронный ресурс]. – Режим доступа: URL: http://www.members.tripod.com/~ShibaHill/atthedavies.html, свободный. Дата обращения: 20.02.2016.
14. WordSmith Tools [Электронный ресурс]. – Режим доступа: http://www.lexically.net/wordsmith/index.html свободный. – Загл. с экрана. – Яз. англ., дата обращения: 20.02.2016.

(0,5 п.л.)