Пермский филиал федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики» Факультет экономики, менеджмента и бизнес-информатики

Горбунова Мария Андреевна

Доклад на тему

«РАЗРАБОТКА СИСТЕМЫ СБОРА СТАТИСТИКИ ИСПОЛЬЗОВАНИЯ МАРКЕРОВ СТИЛЯ В НАУЧНЫХ ПУБЛИКАЦИЯХ НА АНГЛИЙСКОМ ЯЗЫКЕ»

Пермь, 2017 год

Корпусная лингвистика является одним из наиболее перспективных и востребованных направлений в области исследований естественного языка. Данная наука занимается разработкой, созданием и использованием, так называемых текстовых корпусов, состоящих из документов, объединённых по какому-либо признаку. С помощью корпусов лингвисты решают широкий спектр задач, например, подсчитывают статистику использования различных языковых единиц, представление лингвистических данных, реализация математического анализа текста и т.д.

Проведение анализа текста без использования компьютерных инструментальных средств – это достаточно затратный по времени и силам процесс, поэтому необходимо использовать программные инструментальные средства. На данный момент существует большое количество инструментов для обработки корпусов текстов. Наиболее распространенными являются AntConc[1], WordSmith Tools[2], Gate Developer[3], Sketch Engine[4] и CQPweb[5], каждая из которых предоставляет пользователю возможность хранить и редактировать корпуса текстов. Как и любые системы, вышеупомянутые средства имеют ряд недостатков:

- Во-первых, большинство из них являются платными, что является препятствием для использования в академических целях.

- Во-вторых, для пользования данными системами необходимо знание определенного языка, чаще всего английского.

- В-третьих, обработка текста влечет за собой хранение дополнительных файлов (например, файл с оригинальным текстом, файл с обработанным текстом, файл содержащий комплекс правил обработки данного текста и т.д.), которые чаще всего занимают много памяти компьютера.

- В-четвертых, большинство представленных инструментов не имеют средств визуализации обработанных корпусов.

- В-пятых, подобные системы имеют недостаточные средства сбора статистики, так как они не адаптируемы к потребностям пользователей. Но необходимо отметить, что подобный функционал должен быть в системах такого рода, так как с появлением корпусов, объёмы исследуемых данных в разы увеличились, и ручная обработка стала весьма трудозатратной.

Таким образом, статистическая обработка языка ускоряет исследование, а также позволяет с помощью методов математической статистики подтвердить или опровергнуть гипотезу о каком-либо языковом явлении. Из чего следует, что данная работа является актуальной и несет в себе новое решение для рассматриваемой проблемы.

Одним из инструментов для обработки корпусов текстов является программа Gate Developer. Данная программа позволяет создавать и изменять правила по обработке языка. Стоит отметить, что данный инструмент широко применяется для обработки корпусов текстов. Программа является бесплатной и предоставляется с открытым исходным кодом, что очень удобно для дальнейшего расширения функционала. Так же, как и в системах упомянутых выше, в программе Gate Developer нет инструментария, позволяющего генерировать отчеты о собранной статистике.

Объектом исследования в данной работе выступают инструментальные средства обработки текстовых корпусов. Тогда как предметом – средства сбора статистической информации при обработке корпусов текстов, встроенные в инструментарий.

Таким образом, целью данной работы является разработка средства сбора статистики на базе программы Gate Developer для отображения информации о корпусе текстов, обладающего такими свойствами как генерация отчетов о собранных данных, на основе параметров, введенных пользователем, и формирование структуры отчета.

Для достижения данной цели необходимо выполнить ряд задач:

1. Провести анализ существующих средств обработки лингвистических корпусов. Выявить основные достоинства и недостатки рассмотренных программ. Рассмотреть примеры реализации плагинов в программе Gate Developer. Рассмотреть средства генерации отчетов.

2. Сформулировать функциональные требования к разрабатываемому продукту. Выбрать инструментальные средства реализации.

3. Разработать архитектуру модуля для сбора статистики.

4. Разработать плагин для сбора статистики и базовые отчетные формы.

Для исследования объекта необходимо применить такие методы как:

- моделирование информационных систем (для моделирования плагина и его архитектуры);

- формализация (при представлении требований к продукту);

- статистика (для сбора статистических данных в корпусах текстов);

- вероятностно-теоретические методы (при расчёте дополнительных показателей, выводимых в отчете по сбору статистики).

На данный момент реализовано огромное количество систем по обработке естественного языка. Для выяснения необходимого функционала для разрабатываемого программного модуля, необходимо провести обзор систем-аналогов, выявить существующие средства для сбора статистики и генерации отчетов, и каким образом они реализованы.

Для сравнительного анализа систем были сформулированы следующие критерии:

1. Возможность генерации отчетов.

2. Поддержка различных форматов отчетов.

3. Список параметров, заносимых в отчет.

4. Формирование пользовательских отчетов.

5. Возможность расширения функционала, а именно встраивания дополнительных программных модулей в систему.

6. Язык разработки.

Проанализируем наиболее популярные системы обработки естественного языка (Табл. 1).

**Таблица 1. Сравнение средств обработки языка**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Название системы** | **Генерация отчета** | **Форматы экспорта** | **Параметры отчета** | **Пользовательские отчеты** | **Расширение функционала** | **Язык разработки** |
| AntConc | + | + (.txt, .html, .xml, .ant) | Hit, KWIC,File | - | + | Perl |
| WordSmith Tools | + | + (.txt, .xml, .xls, .rtf, WordList files) | Word, Frequency, Texts, Lemmas, Set | - | - | Python |
| Sketch Engine | - | + (.txt, .xml) | Word, Co-occurrence count, Candidate count | - | + | C++, Python, JavaScript, Jquery |
| Gate Developer | - | + (.xml) | Type, Set, Start, End, ID, Features | - | + | Java |
| CQPWeb | + | + (.txt) | No, Search Results, No of occ., percent | - | + | Perl, R, PHP |

По итогам анализа систем обработки естественного языка, можно сделать ряд выводов. Ни одна из рассмотренных систем не позволяет пользователю вносить собственные параметры в отчет. Некоторые из систем не предоставляют возможности генерировать отчеты, также, как и не во всех реализована функция сбора статистики. В большинстве из рассмотренных систем отчет представляет собой файл в формате «\*.txt», с ограниченным набором параметров. Данный формат является универсальным для дальнейшей обработки, но не обеспечивает пользователю четкого графического отчета о произведенной обработке текста. Одна из систем, WordSmith Tools, позволяет отобразить отчет посредством MS Excel, что является преимуществом перед другими инструментами.

Таким образом, стоит отметить, что возможность генерировать пользовательские отчеты с графическим отображением полученных данных не реализована ни в одной из рассмотренных систем, тем не менее, подобная функция является полезной при представлении результатов обработки текста.

Проанализировав средства для обработки языка было выявлено, что ни одно из них, включая Gate Developer, не предоставляет пользователю возможность сгенерировать отчет. Данный факт является показателем того что системы недостаточно доработаны, так как отсутствие генерации отчетов — это существенный недостаток. Для того чтобы построить программный модуль, который выполнял бы такую функцию, необходимо просмотреть решения, которые существуют на рынке в данный момент и выбрать подходящую систему. Определим, что является генератором отчетов.

Генератор отчетов – это библиотека или программа, позволяющая представить данные в структурированном виде, удобном для чтения и печати [8]. Чаще всего для пользователя достаточно затруднительно, после завершения работы с какими-либо данными, формировать отчет вручную. Инструменты для генерации отчетов значительно облегчают задачу.

Стоит отметить, что генераторы отчетов нужны не только пользователям, но и разработчикам, так как система, оснащенная функцией выведения отчетности, имеет некоторое преимущество перед аналогами.

Для дальнейшей разработки программного модуля необходимо проанализировать имеющиеся на рынке средства генерации отчетов и выбрать тот, который максимально удовлетворяет требованиям. Были выбраны 5 генераторов отчетов: Crystal Reports, StimulSoft Reports, Jasper Report, Next Report, BIRT. Для оценивания каждого из инструментов были разработаны следующие критерии:

1. Язык внедрения.

2. Возможность создания диаграмм.

3. Форматы документов отчетов, выгружаемых пользователю.

4. Возможность выбора визуального оформления отчета (шрифты, цветовое оформление, отступы и т.д.).

5. Стоимость программы.

6. Открытый код.

7. Возможность работы с базами данных.

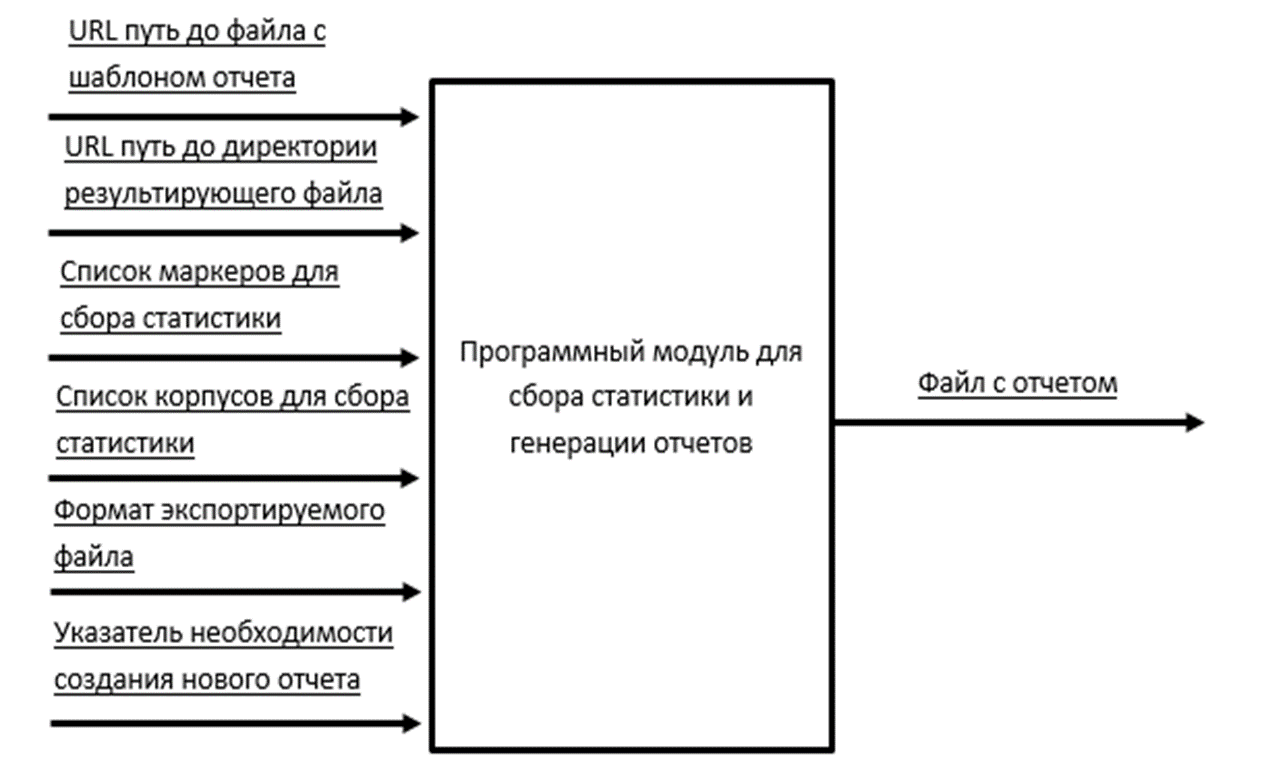
Результаты сравнения генераторов отчетов представлены ниже (Табл. 2).

**Таблица 2. Сравнение генераторов отчетов**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Язык внедрения** | **Возможность создания диаграмм** | **Форматы отчетов, выгружаемых пользователю** | **Визуальное оформление** | **Стоимость** | **Открытый код** | **Возможность работы с базами данных** |
| **Crystal Reports** | Visual Basic, C++,C#, Java и т.д. | + | RPT, RPTR, HTML, MS Word, MS Excel, ODBC, PDF, RTF, CSV, TTX, TXT, XML | + | ~ 75 000 руб. | - | + |
| **StimlSoft Reports** | JavaScript, HTML5, PHP, Java, и Flex | + | PDF, XPS, MS Power Point, HTML, MHT, TXT, RTF, MS Word, MS Excel, CSV, DBF, XML | + | ~ 29 000 руб. | + (при определенных условиях покупки) | + |
| **Jasper Report** | Java | + | PDF, RTF, HTML, XLS, CSV и XML | + | Бесплатно | + | + |
| **Next Report** | Java | + | HTML, EXCEL, EXCELX, PDF, DOCX, RTF,  CSV, TSV, TEXT и XML. | + | Бесплатно | - | + |
| **BIRT** | Java | + | HTML, PDF, Excel, Power Point,  Word | + | Бесплатно | + | + |

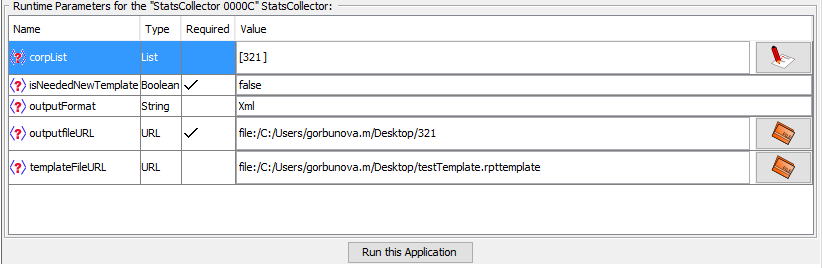
Таким образом, среди всех рассмотренных средств генерации отчетов, наиболее подходящим для решения поставленной задачи является BIRT. Данный инструмент поддерживает внедрение с помощью языка программирования Java, позволяет визуализировать данные в виде диаграмм, а также выбрать параметры отображения отчета. Также выбранное средство предоставляется бесплатно, что является важным критерием при академических исследованиях.

Перед тем как приступить к непосредственной разработке программного модуля, необходимо описать какие параметры должны быть использованы. Так как плагин должен генерировать отчет на основе шаблона и экспортировать готовый отчет, в качестве входных параметров необходимо получать путь до файла с шаблоном отчета и путь до директории, в которой необходимо сохранить отчет. Схема передачи параметров изображена на рис. 1.

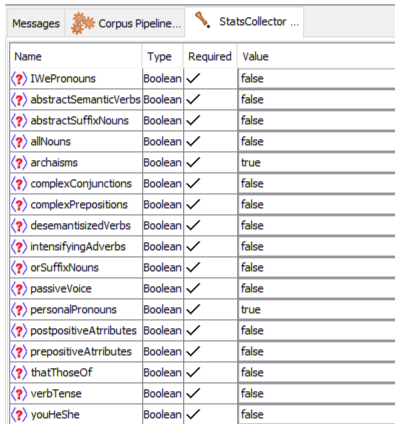


**Рисунок 1. Параметры входа/выхода**

Полный листинг добавления параметров можно увидеть в приложении Б. В результате выполнения плагина, в окне редактора параметров обрабатывающего ресурса отображаются добавленные параметры (рис. 2) и список необходимых маркеров (рис. 3).



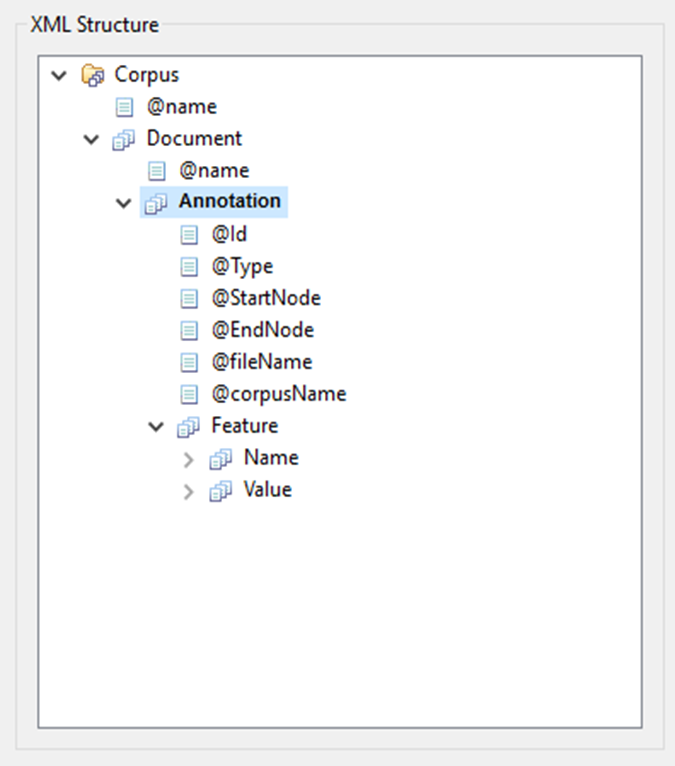
**Рисунок 2. Параметры ввода**



**Рисунок 3. Список маркеров**

Ранее было сказано, что в системе Gate Developer уже существует плагин для сбора статистических данных. На данный момент StatsCollector, программный компонент по сбору статистики, экспортирует результат работы в виде файла в формате \*.txt. Данный формат не удобен для дальнейшей обработки, к тому же результирующий файл не обладает необходимой для этого структурой. Таким образом, в плагине необходимо реализовать алгоритм, который бы выгружал аннотированные тексты в структурированную форму (XML) и передавал их в генератор отчетов, в качестве источника данных. В зависимости от значения параметра corpList, плагин формирует .xml документ с данными корпуса или же документ с данными из нескольких корпусов, с помощью функции getcorp(). В качестве входного параметра getcorp() принимает название корпуса, и формирует строку с данными. Для работы с XML и его обработкой, была использована библиотека DOM4J. В случае, если плагин применяется к нескольким корпусам, метод применяется к каждому из них и формирует общую строку для набора корпусов.

Размеченные корпуса в формате XML после обработки обладают структурой, изображенной на рис. 4.



**Рисунок 4. Структура документа**

Для того чтобы сгенерировать отчет, пользователю необходимо передать в плагин URL адрес расположения шаблона отчета. В текущем контексте шаблон рассматривается как спроектированная структура отчета. Для создания и редактирования шаблонов отчетов существует независимый от среды разработки Eclipse интерфейс, к которому можно получить доступ посредством введения значения true во входной параметр isNeededDrawTemplate. После введения данного параметра необходимо нажать кнопку “Run Application”, после чего пользователь видит окно.

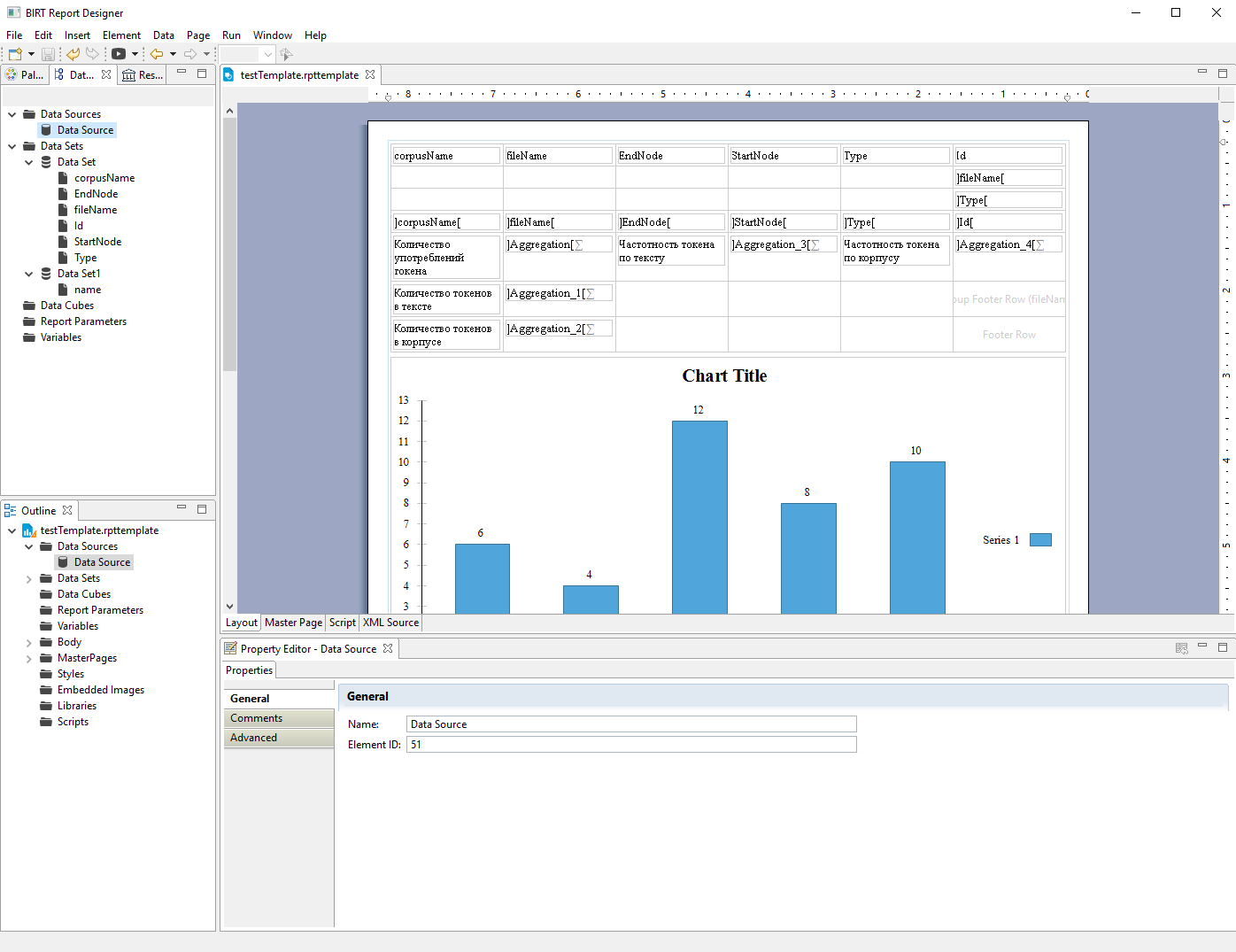
С помощью интерфейса пользователь может добавлять элементы из палитры и таким образом заполнять отчет. Так как на основе этого шаблона в дальнейшем будет генерироваться плагин необходимо, чтобы в нем была указана структура данных. Рассмотрим добавление данных на примере вышеуказанного xml-документа.

Сначала необходимо создать шаблон (.rpttemplate) или дизайн(.rptdesign),в данном случае это будет шаблон. После этого перейти на вкладку Data Explorer и с помощью контекстного меню вызвать окно создания источника данных.

Далее необходимо последовательно выбрать «XML Data Source», а затем указать путь до самого источника данных. Схему данных программа определяет автоматически.

Далее необходимо создать DataSet, так же как для предыдущего действия, для этого необходимо воспользоваться контекстным меню. После того как заданы название и другие настройки набора данных, необходимо создать структуру колонок, выбрав их из структуры xml-файла.

После добавления данных пользователь может выполнять действия с ними, а также использовать различные элементы для создания шаблона. В конце работы необходимо сохранить шаблон в формате \*.rpttemplate (или \*.rptdesign). Указанные процедуры являются необходимыми, так как структура данных, включенная в шаблон, позволяет загрузить результаты вычисления статистики в отчет и отобразить их, так как было задано пользователем.



**Рисунок 5. Пример шаблона отчета**

На рис.5 представлен пример шаблона. На нем мы можем увидеть функцию Aggregation, которая позволяет выполнять стандартные математические функции, а также функции математической статистики, что является применимым к данной работе.

**Библиографический список**

1. Трофимов Сергей Варианты использования (Use Case) [Электронный ресурс]. URL: http://www.caseclub.ru/articles/use\_case.html (дата обращения: 06.03.2017).

2. Cunningham H. [и др.]. Developing language processing components with GATE (a user guide) // University of Sheffield. 2001. № Gate 2 (2006). C. 1–457.

3. Laurence Anthony’s Homepage [Электронный ресурс]. URL: http://www.laurenceanthony.net/ (дата обращения: 03.02.2017).

4. GATE.ac.uk - index.html [Электронный ресурс]. URL: https://gate.ac.uk/ (дата обращения: 14.03.2017).

5. The Stanford Natural Language Processing Group [Электронный ресурс]. URL: https://nlp.stanford.edu/software/tagger.shtml (дата обращения: 16.03.2017).

6. Лингвистический Энциклопедический Словарь. Кореферентность [Электронный ресурс]. URL: http://tapemark.narod.ru/les/243a.html (дата обращения: 16.03.2017).

7. Чанкер для русского языка [Электронный ресурс]. URL: http://web-corpora.net/wsgi/chunker.wsgi/npchunker/ (дата обращения: 16.03.2017).

8. генератор отчетов - это... Что такое генератор отчетов? [Электронный ресурс]. URL: http://dic.academic.ru/dic.nsf/eng\_rus/56647/генератор (дата обращения: 14.03.2017).

9. Crystal Reports [Электронный ресурс]. URL: https://msdn.microsoft.com/ru-ru/library/bb126227(v=vs.90).aspx (дата обращения: 15.03.2017).

10. Генератор отчетов для Java. Экспорт отчетов в PDF, MS Office, Open Office и т.д. :: Стимулсофт [Электронный ресурс]. URL: https://www.stimulsoft.com/ru/products/reports-java (дата обращения: 15.03.2017).

11. Бизнес аналитика. Генератор отчетов для JavaScript, .NET, ASP.NET, WPF, Flex, PHP, Silverlight, Java, ASP.NET MVC, WinRT, HTML5. Дизайнер отчетов. Создание отчетов. Сервер отчетов. Бизнес-аналитика. :: Стимулсофт [Электронный ресурс]. URL: https://www.stimulsoft.com/ru (дата обращения: 15.03.2017).

12. JasperReports® Library | Jaspersoft Community [Электронный ресурс]. URL: http://community.jaspersoft.com/project/jasperreports-library (дата обращения: 15.03.2017).

13. Business reporting software - NextReports Designer, Engine and Server [Электронный ресурс]. URL: http://www.next-reports.com/ (дата обращения: 15.03.2017).

14. BIRT Home [Электронный ресурс]. URL: http://www.eclipse.org/birt/ (дата обращения: 15.03.2017).

15. About [Электронный ресурс]. URL: http://www.eclipse.org/birt/about/project-organization/ (дата обращения: 15.03.2017).

16. Hardie A. The IMS Open Corpus Workbench ( CWB ) CQPweb System Administrator ’ s Manual 2014. C. 1–45.

17. Scott M. WordSmith Tools step by step 2010.

18. Talk T., Diachronic T., Corpus E. A Guide to using AntConc C. 1–9.

19. Сравнение утилит анализа корпусов — NLPub [Электронный ресурс]. URL: https://nlpub.ru/Сравнение\_утилит\_анализа\_корпусов (дата обращения: 21.05.2017).

20. Каталог лингвистических программ и ресурсов в Cети, ч.1 / Linguistics Software Catalogue, ch.1 [Электронный ресурс]. URL: http://rvb.ru/soft/catalogue/c01.html (дата обращения: 21.05.2017).

21. Sketch Engine | language corpus management and query system [Электронный ресурс]. URL: https://www.sketchengine.co.uk/ (дата обращения: 04.02.2017).

22. Обзор платформы Eclipse - как её использовать | Статьи о программном обеспечении [Электронный ресурс]. URL: https://hightech.in.ua/content/art-eclipse-platform (дата обращения: 23.03.2017).

23. Welcome to NetBeans [Электронный ресурс]. URL: https://netbeans.org/ (дата обращения: 23.03.2017).

24. Обзор NetBeans IDE | Обзор IDE [Электронный ресурс]. URL: https://aboutide.wordpress.com/2011/03/23/обзор-netbeans-ide/ (дата обращения: 23.03.2017).

25. IntelliJ IDEA [Электронный ресурс]. URL: http://jetbrains.ru/products/idea/ (дата обращения: 23.03.2017).

26. SAP Crystal Reports 2011 User вЂTM s Guide ■ SAP Crystal Reports 2011 2012.