КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 004.912

ПРИМЕНЕНИЕ КОМПЛЕКСА ИНСТРУМЕНТОВ УПРАВЛЕНИЯ КОРПУСАМИ ТЕКСТОВ ПРИ РЕШЕНИИ ЗАДАЧ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

С. А. Полицын, Е. В. Полицына

Московский авиационный институт (национальный исследовательский университет)

Поступила в редакцию 24.04.2019 г.

Аннотация. Одной из актуальных задач компьютерной лингвистики, необходимых для решения других задач, в т. ч. для использования методов машинного обучения, разработки и апробации новых алгоритмов, является задача составления, разметки и оперативного пополнения корпусов текстов. В статье освещается разработка и применение комплекса инструментов управления корпусами текстов, который позволит создавать субкорпуса по настраиваемому набору признаков.

Ключевые слова: корпус текстов, инструменты автоматизированного анализа текстов, разметка корпуса, краулер, управление корпусами текстов.

ВВЕДЕНИЕ

Количество обрабатываемой информации последние десятилетия и потребность в развитии методов и инструментов ее автоматизированной обработки только увеличиваются. Поэтому в настоящее время большое внимание уделяется развитию компьютерной лингвистики – направления в прикладной лингвистике, связанного с автоматизацией процессов сбора, обработки и разметки информации. Развитие средств автоматизированного анализа текстов позволило исследователям существенно сократить в том числе время на формирование текстовых корпусов для своих задач [8, 10]. В последнее время при решении задач компьютерной лингвистики успешно используется и все более набирающие популярность методы машинного обучения [11], использование которых во всех областях требует большого количества информации для обучения системы, компьютерная лингвистика – не исключение.

Текстовый корпус – это свод текстов и информации по ним, собранных в соответствии с определенными принципами, размеченных по определенному стандарту и обеспеченных поисковой системой [1]. Разработкой и созданием корпусов занимается специальный раздел языкознания «Корпусная лингвистика» [2].

Корпуса текстов применяются в лингвистических исследованиях различной направленности. Во-первых, на корпусах текстов проводятся разного рода лексические исследования, например [12]. Анализируются частоты употребления и различные формы слов, выявляются статистические закономерности и новые слова. Вторым направлением применения корпусов текстов в лингвистике является изучение исторического развития, диалектов языка, и различных форм словоупотреблений [13]. Применение корпусов текстов в задачах прикладной компьютерной лингвистики и в особенности, машинном обучении, где от количества и состава корпусов исходных текстовых данных, с помощью которых происходит обучение программной си-

стемы, еще больше повышают потребность в наличии корпусов текстов различной направленности.

ОБЗОР ЛИНГВИСТИЧЕСКИХ КОРПУСОВ ТЕКСТОВ

В течении последних десятилетий во многих странах ведутся работы по формированию лингвистических корпусов в целях изучения национальных языков. Первые корпуса появились в Великобритании в 60-е годы: Brown University Corpus и Lancaster/ Oslo-Bergen Corpus (LOB). Самым крупным британским корпусом является Британский Национальный Корпус (BNC) [3]. Эти корпуса содержат морфологическую разметку и имеют примерно один миллион словоупотреблений. В дальнейшем в эти корпуса была добавлена также и синтаксическая разметка. Самым крупным корпусом является Британский Национальный Корпус (BNC), насчитывающий около 100 миллионов слов [3].

В Германии самым крупным корпусом является корпус Института немецкого языка в Маннгейме [14]. Этот корпус содержит около двух миллионов словоупотреблений, имеет морфологическую и синтаксическую разметку, а также и автоматизированную систему поиска содержимого корпуса по морфологическим признакам словоформ. В Чехии – Чешский национальный корпус, отличительной особенностью которого является возможность получать все примеры употреблений вместе с контекстами, в которых словоформа встречается, частоту вхождения словоформы в корпус. Также есть морфологический анализатор, который позволяет проводить морфологический и контекстный анализ [4].

Одним из главных корпусов текстов на русском языке является Национальный корпус русского языка (НКРЯ) [6]. Корпус содержит около пятисот миллионов словоформ. В коллекции корпуса содержится множество типов текстов: исторические, литературные, диалектные, письменные, устные, современные, переводные. Корпус оснащен большим количеством разметок: лексической, морфологической, синтаксической, лексико-семантической и рядом других специализированных разметок. Особенностью корпуса является стихотворная разметка, которая позволяет искать стихотворные тексты с заданием различных параметров.

ПРОБЛЕМЫ КОРПУСНОЙ ЛИНГВИСТИКИ

На первых этапах развития корпусной лингвистики у корпусов текстов выявилась важная особенность – их узкая направленность. Каждый корпус текстов должен иметь разметку, а в некоторых случаях и набор информации, исходя из поставленных задач. Например, для исследований в области морфологии корпус должен содержать данные о характеристиках слова (часть речи, род, число, падеж – для имен существительных, вид, время, переходность – для глаголов и т. д.), для изучения выражений требуется поле для связи слов между собой, для создания классификаторов необходимо указание автора и предметной области и т. д.

Корпус текстов, особенно при его использовании для машинного обучения или проверки качества работы других алгоритмов и методов автоматического анализа текстов, должен содержать большой объем данных. Работы по реализации текстового корпуса требуют много времени и состоят из нескольких подзадач: сбор информации из самых разных источников и на разные тематики для следования принципу репрезентативности, обработка собранной информации, анализ обработанной и структурированной информации, формирование разметки для корпуса текстов.

Таким образом, из-за специфичности лингвистических корпусов и трудности их создания существует проблема ненужности корпусов после выполнения поставленных задач, так как практически всегда корпус создается под конкретную задачу. Такая проблема присуща корпусам меньшим, чем, например, НКРЯ, у которого большой арсенал разметок, что делает его почти универсальным, но у него отсутствует программный интерфейс или возможность получения большо-

го объема текстов для использования в программных инструментах. Кроме того, новые области применения инструментов компьютерной лингвистики появляются гораздо быстрее, чем новые тексты или виды разметки в корпусах. При решении прикладных задач тексты анализируются с самых разных точек зрения: являются ли тексты сообщений в социальных сетях мошенническими, какую эмоциональную окраску имеют и др. Решение таких задач, особенно средствами машинного обучения, требует составления специальным образом размеченного корпуса. После выполнения поставленных задач корпусы становятся не такими востребованными, как раньше, несмотря на то что проделана большая работа по его реализации. Решением этой проблемы может стать разработка комплекса инструментов для создания и разметки корпусов текстов, ориентированных на решение различных задач.

В качестве источников данных для корпусов могут служить различные пополняемые электронные библиотеки, собрания текстов по какой-либо тематике или энциклопедии, новостные ресурсы, открытые данные социальных сетей и др. Важным требованием для автоматического создания корпуса и его разметки является возможность получения текстов с заранее определенными людьми свойствами, например, тексты по тематике или по авторам. Таких свойств может быть много, поэтому требуется создание правил для расширяемой разметки корпуса [9], чтобы обеспечить возможность повторного использования имеющихся текстов, кроме того, необходим инструмент, который за относительно небольшое время позволит автоматически получать нужные тексты.

КОМПЛЕКС ИНСТРУМЕНТОВ УПРАВЛЕНИЯ КОРПУСАМИ ТЕКСТОВ

При разработке комплекса инструментов управления корпусами текстов учитывались как обозначенные выше потребности в различных сферах, так и выявленные недостатки существующих корпусов текстов. Комплекс инструментов должен обеспечить возможность создания и дополнения корпуса с универсальной расширяемой разметкой для накопления большого количества текстов разного объема с разными заданными характеристиками, должен предоставлять возможность ручного и автоматического дополнения его новыми текстами, дополнения существующей разметки в том числе с помощью сторонних программных инструментов, создания субкорпусов в зависимости от требований конкретного пользователя.

Ранее разработанные средства создания субкорпусов позволяют производить отбор текстов на основе фильтрации выбранных признаков, имеющихся на данный момент в разметке корпуса [9]. В создаваемые таким образом субкорпуса не попадают тексты, в которых отсутствует хотя бы один из фильтруемых признаков. Кроме того, для решения ряда задач существует необходимость формирования субкорпусов с учетом специальных характеристик, таких как объем текста (в знаках, словах, предложениях, абзацах), статистических характеристик (средняя длина слова, средняя длина предложения и т. д.) и др., которые могут быть получены автоматически.

На рис. 1 представлена общая схема работы комплекса инструментов управления корпусами текстов, основными частями которого являются:

1. Общий корпус текстов.

2. Набор инструментов поиска и извлечения текстов (краулеров) для пополнения корпуса.

3. Инструменты анализа текстов для автоматической разметки.

4. Инструмент выбора и разметки субкорпусов.

Инструмент выбора и разметки субкорпусов позволяют динамически добавлять разные виды разметки путем изменения конфигурационных файлов и добавления инструментов анализа текстов – размечающих утилит, на вход которым в заданном формате передаются данные о субкорпусе для разметки и необходимых характеристиках.

В качестве способа хранения данных было выбрано использование облачного решения

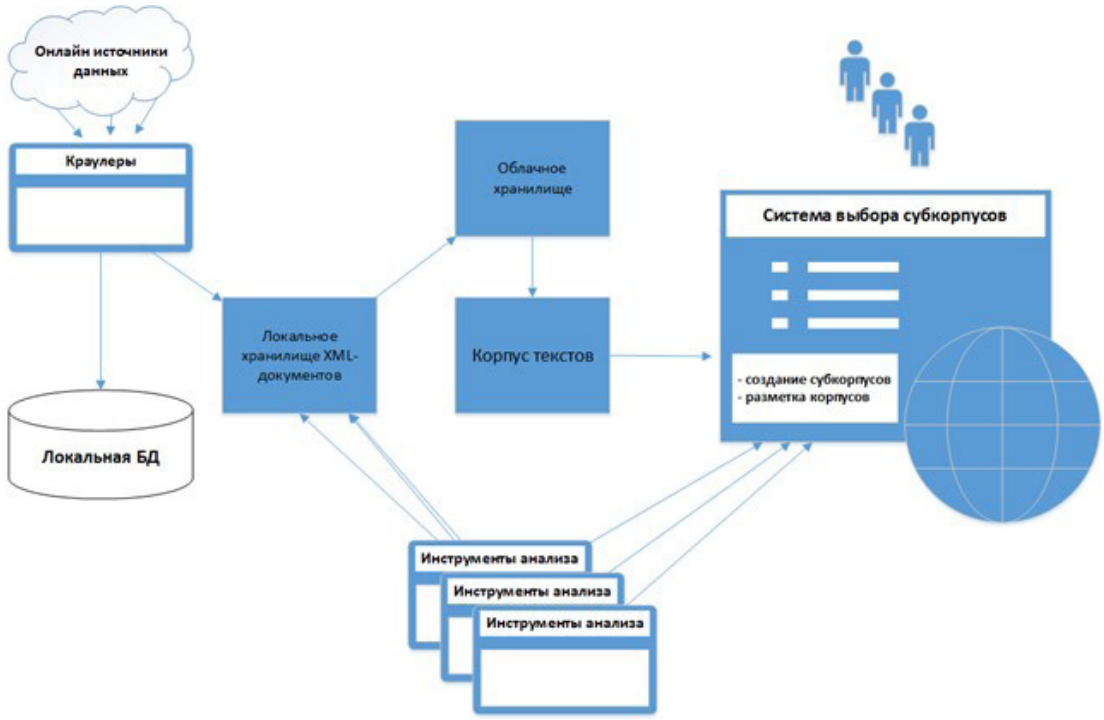

Рис. 1. Структура комплекса инструментов управления корпусами текстов

Яндекс-диск, что позволяет, с одной стороны иметь организованный совместный доступ к полному имеющемуся корпусу, с другой – для пользователя – получать необходимые субкорпуса без дополнительной нагрузки на комплекс инструментов целиком.

РАЗРАБОТКА ИНСТРУМЕНТОВ НАПОЛНЕНИЯ КОРПУСА ТЕКСТОВ

Инструменты поиска и извлечения текстов (краулеры) служат для сбора информации с веб-ресурса, на который они настроены, по ряду условий, заданных пользователем в конфигурационном файле. Важным критерием выбора ресурсов – источников текстов, является их достоверность. Для наполнения и последующего расширения корпуса подготовлены инструменты, позволяющие в автоматическом режиме загружать тексты научных статей [16]. В первую очередь для апробации этих инструментов был использован научный Интернет-ресурс «Научная электронная библиотека» [7] и ряд других.

После получения информации тексты конвертируются в XML-документ с расширяемой разметкой [9], содержащей информацию о загруженном материале и его исходный текст.

Для сбора данных и конфигурирования краулера необходимо провести анализ ресурса и определить, имеются ли фильтры по тематикам или другим критериям, необходимы ли разного рода проверки, например, возможно ли скачать текст, полная ли версия работы представлена, определение языка текста и др. Все эти возможности включаются в конфигурационный файл инструмента наполнения корпуса.

Часто тексты представлены в формате PDF (Portable Document Format), поэтому для сохранения текста и реализации разметки корпуса в комплексе инструментов реализована возможность получения из них текстов для дальнейшей обработки и преобразования в XML-файл. В настоящий момент только на одном из ряда проверенных ресурсов данные представляли собой изображения без возможности получения текста напрямую, если же таких ресурсов будет больше, потребуется внедрение средств распознавания изображений и выделения их них текста. Инструмент поиска и извлечения текстов разработан на языке Python с применением библиотеки Scrappy [15].

В зависимости от результатов проверок инструментом принимается решение о сбо-

{"subjects": "Общественные науки в целом", "name_book": "2003.01.022-025. ПОЛИТИКА КОНСТРУКЦИОНИЗМА. (СВОДНЫЙ РЕФЕРАТ)", "authors": ["ЯКИМОВА Е.В"], "key_word": null, "description": null, "book_text": " \n\n98 \n\nСОЦИАЛЬНАЯ ПСИХОЛОГИЯ \n\n \n \n \n \n \n \n \n\n(Сводный \n\n2003.01.022-025. ПОЛИТИКА КОНСТРУКЦИОНИЗМА. \nреферат). \n2003.01.022. VELODY I., WILLIAMS R. Introduction. // The politics of \nconstructionism. – L., etc., 1998. – P. 1–12. \n2003.01.023. LYNCH M. Towards a constructivist genealogy of social \nconstructivism // Ibid., P. 13–32. \n2003.01.024. GERGEN K.

Рис. 2. Пример записи о собранных данных в базе данных

ре данных по конкретной статье. Собранные данные и сам текст помещаются в базу данных sqlite3 (рис. 2) для реализации алгоритма обновления корпуса.

Затем на основе этих данных создаются XML-документы с использованием специально разработанной ранее разметки [9].

Для дополнения характеристик текста путем их автоматического получения из достоверных источников или путем возможного использования различных инструментов анализа текстов для автоматической разметки была расширена структура документа без потери обратной совместимости: для любого тэга можно указать атрибуты «auto» со значениями true/false (по умолчанию false) и «verify» со значениями true/false (по умолчанию true):

- «auto» = «true» – характеристика текста получена автоматически с помощью программных средств;
- «auto» = «false» – характеристика текста указана вручную экспертом или автором;
- «verify» = «true» – характеристика текста имеет достоверное значение, проверенное экспертом или однозначно определяемое с помощью программных средств;
- «verify» = «false» – характеристика текста, не проверенная экспертами.

Разработанный инструмент способен создавать корпусы текстов с динамической разметкой. На основе разработанного краулера путем настройки конфигурации могут быть созданы инструменты для других Интернет-ресурсов, например, Научной электронной библиотеки [7], Интернет-версии Большой российской энциклопедии и др.

ИНСТРУМЕНТЫ АНАЛИЗА ТЕКСТОВ ДЛЯ АВТОМАТИЧЕСКОЙ РАЗМЕТКИ КОРПУСОВ

Для возможности расширения разметки документов корпуса в дополнение к существующей была разработана возможность применения инструментов автоматического анализа текстов и их добавление в комплекс без участия разработчиков.

Дополнительная разметка может потребоваться в ходе подбора необходимых корпусов, например, для целей подготовки обучающей выборки при решении задач машинного обучения, где для некоторых алгоритмов требуются тексты определенного объема или богатого лексического наполнения (с большим размером словника).

Инструменты автоматического анализа текстов могут дополнять разметку корпуса статистической или лингвистической информацией.

Для добавления размечающей утилиты в комплекс инструментов нужно указать ее в файле конфигурации, представленном на рис. 3. В настоящее время поддерживаются утилиты на языках Java, C++, Python. В рамках курсовых и дипломных работ на кафедре 319 МАИ были разработаны инструменты, способные выделять из текста и добавлять разметку для числительных и др. частей речи, дат различных форматов, единиц измерения, получение словников по тексту по начальным формам слов, а также статистические данные о количестве букв, слов, и др. в текстах корпуса.

Требованиями к размечающим утилитам для их подключения в комплекс инструментов являются:

1. Работа с документами в XML-формате с описанным способом разметки.

2. Реализация программного интерфейса: при запуске на вход инструмент должен принимать следующие параметры: путь к директории с документами для разметки (может содержать вложенные директории), имя тега для добавления в документ результата разметки, дополнительные необязательные параметры.

Например, в результате запуска утилиты разметки во все документы в директории *texts*, будет добавлен тег *words_count* (рис. 4), содержащий количество слов в словнике:

java -jar simple_statistics_marker.jar .\\texts words_count

```
        <label>Количество символов</label>
        <techName>symbols_count</techName>
        <textLanguage>ru</textLanguage>
        <tagName>symbols_count</tagName>
        <path>/utils/symbols_count.jar</path>
        <parameters></parameters>
        <programmingLanguage>java</programmingLanguage>
        <runMode>interpretable</runMode>
        <operatingSystem>*</operatingSystem>
    </util>
    <util active="false">
        <label>Полный словник</label>
        <techName>words_list_all</techName>
        <textLanguage>en</textLanguage>
        <tagName>words_list_all</tagName>
        <path>/utils/wordsListCreator.exe</path>
        <programmingLanguage>C++</programmingLanguage>
        <runMode>executable</runMode>
        <operatingSystem>windows</operatingSystem>
    </util>
    <util active="true">
        <label>Количество слов</label>
        <techName>words_count</techName>
        <textLanguage>tk</textLanguage>
        <tagName>words_count</tagName>
        <path>/utils/words_count.py</path>
        <programmingLanguage>python</programmingLanguage>
        <runMode>interpretable</runMode>
        <operatingSystem>*</operatingSystem>
    </util>
</markup-utils>
```

Рис. 3. Конфигурационный файл для подключения инструментов разметки

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<doc>
    <category/>
    <author>Kris Kaspersky</author>
    <title>BIOS Setup - удаленный контроль</title>
    <keywords>ключевое слова, слова</keywords>
    <all_words_count auto="true" verify="false">3756</all_words_count>
    <text><![CDATA[аппаратные и программные комплексы на арене безопасности
```

Рис. 4. Пример файла корпуса с дополненной разметкой

ВОЗМОЖНОСТИ ИНСТРУМЕНТА ВЫБОРА И РАЗМЕТКИ СУБКОРПУСОВ

В результате начальной подготовки корпуса текстов и его наполнения на было добавлено в корпус и частично размечено около 40Гб текстов по различным тематикам и некоторыми дополнительными признаками. Разработанный инструмент управления корпусами обладает рядом возможностей:

1. Получение субкорпуса, составленного путем выбора текстов по сочетанию наличия или отсутствия определенных признаков (поддерживаются основные логические операции: «И», «ИЛИ», «НЕ»).

2. Добавление разметки в сформированный субкорпус в соответствии с выбранными утилитами разметки документов, доступными в комплексе инструментов.

3. Получение корпуса текстов целиком или сформированного субкорпуса с сервера на локальный компьютер.

4. Загрузив субкорпус на локальный компьютер, можно провести его разметку с помощью собственных утилит, либо утилит, поставляемых в составе комплекса.

5. Добавление новых текстов в корпус и разметки к новым или ранее загруженным текстам (функция доступна только администраторам комплекса инструментов).

В настоящее время на основе существующей системы управления корпусами текстов для настольных систем ведется ее расширения и разработка веб-версии, основным отличием которой будет сокращение нагрузки на компьютер пользователя за счет перенесения формирования субкорпусов и дополнительной разметки в облачную инфраструктуру, пользователь сможет получать архив с готовым размеченным субкорпусом. Еще одной особенностью веб-версии инструмента выбора и разметки корпусов является возможность добавления текстов в корпус для пользователей и реализация программного интерфейса для сторонних приложений.

В настоящее время наполнение корпуса производится инструментами наполнения корпуса, а также некоторые тексты, используемые при решении отдельных задач (выделения ключевых слов и словосочетаний, реферирования текстов, определения мошеннических текстов в социальных сетях, исследования словарей синонимов и др.) добавляются вручную администраторами после обработки инструментами разметки.

ЗАКЛЮЧЕНИЕ

Комплекс инструментов управления корпусами текстов, включающий в себя инструменты автоматического наполнения, управления, дополнительной разметки и получения субкорпусов, способен существенно облегчить решение задач корпусной лингвистики и подготовку данных для разработки новых алгоритмов и инструментов компьютерной лингвистики. Основной отличительной особенностью комплекса инструментов является возможность формирования собственных субкорпусов и добавления дополнительной разметки для выбранного субкорпуса. Широкие возможности по автоматизированному наполнению корпуса текстами за счет конфигурации инструмента поиска текстов позволяют применять корпус в задачах машинного обучения, например классификации текстов по различным критериям, где для повышения точности работы алгоритмов, требуется подготовка большой обучающей выборки по каким-либо определенным критериям.

СПИСОК ЛИТЕРАТУРЫ

1. Что такое корпус? [Электронный ресурс]. Режим доступа: http://velib.com/read_book/bez_avtora/vvedenie_v_korpusnuju_lingvistiku/glava_1_chto_takoe_korpus/, свободный – Заглавие с экрана. – (15.01.2019)

2. Корпусная лингвистика как раздел языкознания [Электронный ресурс]. Режим доступа: https://www.myfilology.ru/177/ korpusnaya-lingvistika-kak-razdel-yazykoznaniya/, свободный – Заглавие с экрана. – (15.01.2019)

3. British National Corpus [Электронный ресурс]. Режим доступа: http://www.natcorp.ox.ac.uk/, свободный – Заглавие с экрана. – (20.01.2019)

4. Český národní korpus [Электронный ресурс]. Режим доступа: https://korpus.cz/, свободный – Заглавие с экрана. – (15.01.2019)

5. ГИКРЯ – Генеральный Интернет-Корпус Русского Языка [Электронный ресурс]. Режим доступа: http://www.webcorpora.ru/, свободный – Заглавие с экрана. – (15.01.2019)

6. Национальный корпус русского языка [Электронный ресурс]. Режим доступа: http://www.ruscorpora.ru/, свободный – Заглавие с экрана. – (15.01.2019)

7. Научная электронная библиотека [Электронный ресурс]. Режим доступа: https://elibrary.ru/, свободный – Заглавие с экрана. – (15.01.2019)

8. Корпусная лингвистика [Электронный ресурс]. Режим доступа: http://lomonosov-fund.ru/enc/ru/encyclopedia:01210:article, свободный – Заглавие с экрана. – (15.01.2019)

9. *Полицын, С. А.* Применение корпуса текстов для автоматической классификации в комплексе инструментов автоматизированного анализа текстов / С. А. Полицын, Е. В. Полицына // Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. – 2018. – № 2. – С. 162–167.

10. Корпусная лингвистика [Электронный ресурс]. Режим доступа: http://corpora.iling.spb.ru/theory.htm, свободный – Заглавие с экрана. – (15.01.2019)

11. Машинное обучение для понимания естественного языка [Электронный ресурс]. Режим доступа: https://www.osp.ru/os/2016/01/13048649/, свободный – Заглавие с экрана. – (15.01.2019)

12. Лингвистические исследования на базе корпусов [Электронный ресурс]. – Режим доступа: https://www.myfilology.ru/177/ lingvisticheskie-issledovaniya-na-baze-korpusov, свободный – Заглавие с экрана. – (Дата обращения: 02.11.2018)

13. *Савчук, С. О.* Национальный корпус русского языка: перспективы использования в лингвистических исследованиях и в преподавании / С. О. Савчук // Вестник Азиатско-Тихоокеанской ассоциации преподавателей русского языка и литературы. – 2011. – № 2-3. – С. 62–67.

14. *Kupietz, M.* The German Reference Corpus DeReKo: A primordial sample for linguistic research / M. Kupietz [et al.] // In: Calzolari, N. et al. (eds.): Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010) (P. 1848–1854). Valletta, Malta: European Language Resources Association (ELRA).

15. Официальный сайт библиотеки Scrapy [Электронный ресурс]. – Режим доступа: https://scrapy.org/, свободный – Заглавие с экрана. – (Дата обращения: 23.04.2019)

16. *Полицына, Е. В.* Разработка комплекса инструментов для управления корпусами текстов / Е. В. Полицына, С. С. Попов // В сборнике: Информатика: проблемы, методология, технологии. Сборник материалов XIX международной научно-методической конференции. Под редакцией Д. Н. Борисова. – 2019. – С. 1621–1626.

Полицын Сергей Александрович – канд. техн. наук, доцент, институт № 3, кафедра 319, Московский авиационный институт (Национальный исследовательский университет).
E-mail: pul_forever@mail.ru

Полицына Екатерина Валерьевна – канд. техн. наук, доцент, институт № 3, кафедра 319, Московский авиационный институт (Национальный исследовательский университет).
E-mail: kathrin.beaver@mail.ru

THE COMPLEX OF TEXT CORPUS MANAGEMENT TOOLS USAGE IN SOLVING COMPUTER LINGUISTICS TASKS

S. A. Politsyn, E. V. Politsyna

Moscow Aviation Institute (National Research University)

Annotation. The task of creation, markup and keeping up-to-date of linguistic corpuses is very urgent today including machine learning needs, and approbation of new algorithms. The paper shows development of the set of programs for creating and managing text corpuses, and some applications of these programs which allows creating sub-corpuses basing on flexible set of parameters.
Keywords: automated text analysis tools, corpus of texts, linguistic markup, crawler, managing text corpuses.

Politsyn Sergey – candidate of technical sciences, associate professor, department 319, Moscow Aviation Institute (National Research University).
E-mail: pul_forever@mail.ru

Politsyna Ekaterina – candidate of technical sciences, associate professor, department 319, Moscow Aviation Institute (National Research University).
E-mail: kathrin.beaver@mail.ru