zhurnal jeksperimental'nogo obrazovanija, 2012, no. 6, pp. 107—109 (in Russian).
14. **Buravcev A. V.** Seryj upravlencheskij analiz, *Perspektivy nauki i obrazovanija*, 2017, no. 5 (29), pp. 74—79 (in Russian).
15. **Tsvetkov V. Ya.** Kognitivnye aspekty postroenija virtual'nyh obrazovatel'nyh modelej, *Perspektivy nauki i obrazovanija*, 2013, no. 3, pp. 38—46 (in Russian).
16. **Tsvetkov V. Ya.** Incremental Solution of the Second Kind Problem on the Example of Living System, *Biosciences biotechnology research Asia*, November 2014, vol. 11 (Spl. Edn.), pp. 177—180.
17. **Doha E. H., Abd-Elhameed W. M., Youssri Y. H.** Second kind Chebyshev operational matrix algorithm for solving differential equations of Lane—Emden type, *New Astronomy*, 2013, vol. 23, pp. 113—117.
18. **Robert C. P.** Monte Carlo methods, John Wiley & Sons, Ltd, 2004.
19. **Bautista J., Pereira J.** Ant algorithms for assembly line balancing, *Ant algorithms*, 2002, pp. 49—61.
20. **Dorigo M., Di Caro G., Gambardella L. M.** Ant algorithms for discrete optimization, *Artificial life*, 1999, vol. 5, no. 2, pp. 137—172.
21. **DeVore R. A., Temlyakov V. N.** Some remarks on greedy algorithms, *Advances in computational Mathematics*, 1996, vol. 5, no. 1, pp. 173—187.
22. **Wang Gai-Ge, Guo Lihong, Gandomi Amir H.** et al. Chaotic krill herd algorithm, *Information Sciences*, 2014, vol. 274, pp. 17—34.

УДК 004.912 DOI: 10.17587/it.24.515-523

Ю. С. Шучалова, приглашенный преподаватель, iusshuchalova@hse.ru,
В. В. Ланин, ст. преподаватель, vlanin@hse.ru,
Национальный исследовательский университет "Высшая школа экономики", г. Пермь

Исследовательский портал для анализа и оценки стиля научных публикаций

Описан этап проектирования портала для проведения корпусных исследований английского языка. Сформулированы требования к решению, показаны лингвистические подходы к решению поставленных задач. Приведен процесс моделирования системы и рассмотрены особенности реализации с учетом специфики предметной области. Для интеграции гетерогенных компонентов предложена сервисная архитектура.

Ключевые слова: *корпусные исследования, исследовательский портал, академический английский язык, моделирование информационных систем*

Введение

Широкое распространение задач и инструментов, связанных с информационными технологиями, в науках о естественных языках позволило лингвистам наряду с традиционными методами развивать такое направление, как корпусная лингвистика. Данное направление основано на использовании так называемых корпусов, под которыми понимается представительная коллекция текстов по определенной тематике, доступная в электронной форме [1], т. е. набор документов или записей, собранных по определенному принципу: общий стиль, общий автор, общий жанр и т. д. Таким образом, в отличие от классической лингвистики, корпусная лингвистика ориентирована на изучение практического применения языка, а не его теоретических аспектов. Многие традиционные задачи языкознания решаются с помощью текстовых корпусов. Например, изучение отдельных словоупотреблений, общей динамики языка, машинный перевод и обучение языку (*computer-assisted language learning*), некоторые виды анализа текста. В целом поиск новых возможностей применения корпусов в различных областях — одна из основных задач корпусной лингвистики.

Корпусная лингвистика [2] тесно связана с информационными технологиями. Сбор и обработка обширных корпусов требуют использования программного обеспечения и технологий, допускающих совместную работу множества удаленных пользователей. Программное обеспечение, предназначенное для работы лингвистов, как правило, можно разделить на общее — для выполнения базовых функций (например, автоматической морфологической разметки) — и специализированное, созданное для решения конкретных задач. Одной из таких задач является проверка соответствия текста академическому стилю. При написании научных статей на английском языке студенты сталкиваются с необходимостью использования стандартных языковых средств для соблюдения академической стилистики текста.

В то же время должен быть соблюден некий баланс в количестве этих языковых средств, чтобы не нарушалась читаемость текста. Сложность состоит в том, что не существует общепринятых стандартов, которых можно придерживаться при написании статьи на английском языке, так как стилистика и степень читаемости текста оцениваются экспертами.

Существующие инструменты для рецензирования и методы анализа стиля [3] при оценке качества и читаемости опираются на ограниченное число признаков (терминология, формальный стиль речи, читаемость текста), в то время как в методической литературе и научных статьях представлено гораздо больше характеристик академического стиля. Кроме того, изначально сформулированные правила определения качества текста необязательно могут отражать реальную ситуацию использования различных признаков научного стиля на практике. Именно такую возможность предоставляют методы корпусной лингвистики.

Таким образом, существует потребность в создании приложения, которое имело бы возможность оценивать научные статьи, написанные носителями на английском языке и одобренные к публикации влиятельными рецензируемыми изданиями (например, Springer), по некоторым лингвистическим характеристикам, которые в основном определяют академический стиль письменной речи. Усредненные значения встречаемости значимых характеристик, определенные в ходе исследования корпуса статей по определенной тематике, смогут служить ориентиром при написании собственного материала.

Для проверки значимости различных лингвистических характеристик при оценке качества стиля и читаемости текста будет проводиться исследование, основанное на сравнении встречаемости маркеров стиля в научных статьях, взятых в качестве эталонных, и учебных работ высокого и низкого качества (по мнению эксперта). Результаты исследования могут быть использованы для частичной автоматизации работы эксперта при проверке соответствия статьи стилистическим требованиям, а также для формирования рекомендаций. Наиболее подходящей для реализации метода анализа и оценки стиля научных текстов на английском языке с помощью эталонных корпусов, по мнению авторов, является форма исследовательского портала.

Целью работы является проектирование архитектуры портала для анализа и оценки стиля научных публикаций с указанием необходимых свойств системы и возможных средств ее реализации. Идентификация функционального стиля текста, в том числе научного (академического), — одна из базовых задач лингвистики в целом [4]. Использование методов корпусной лингвистики, основанных на анализе коллекций текстов, позволяет рассмотреть не только теоретические признаки научных текстов, но и особенности их применения на практике.

Формирование требований к решению

Для реализации портала необходимо формализовать требования к функциональности портала, которые основываются на теоретических аспектах метода. В рассматриваемом методе оценка качества стиля статьи проводится на основе сравнения с корпусом научных статей, написанных носителями на английском языке и признанных качественными (прошедших рецензирование экспертами), по некоторым лингвистическим характеристикам, которые в основном определяют академический стиль письменной речи. Эти характеристики далее в тексте будем называть качественными критериями или маркерами академического стиля.

В качестве теоретической базы для создания приложения экспертом был предоставлен список качественных критериев ("маркеров") академического стиля речи. Список составлен на основе справочных и учебных материалов, а также интернет-ресурсов по обучению академическому письму.

Все вошедшие в список критерии можно разделить на три группы: лексические, грамматические и синтаксические. Внутри групп также возможно разделение критериев, например, по особенностям их проверки. Рассмотрим критерии, входящие в каждую группу, подробнее.

Лексические критерии можно условно разделить на три подгруппы:

- критерии частотности появления в тексте конкретных слов, терминологии;
- критерии частотности появления в тексте слов, соответствующих определенным словообразовательным схемам;
- критерии частотности появления в тексте слов определенных частей речи.

К первой подгруппе относятся следующие критерии:

- активное использование терминологии, соответствующей предметной области;

- близкая к нулевой встречаемость личных местоимений;
- десемантизированные глаголы;
- глаголы широкой абстрактной семантики;
- усилительные наречия.

Вторая группа включает в себя такие маркеры, как:
- наличие абстрактных существительных, образованных с помощью суффиксов;
- наличие суффиксов, обозначающих термины и технические понятия.

Наконец, к третьей группе мы отнесем следующие критерии:
- номинативность текста — преобладание существительных;
- предположительно низкая встречаемость личных местоимений.

В группу грамматических маркеров входят два критерия:
- широкое использование глаголов в пассивном залоге;
- преобладание глаголов настоящего времени.

Синтаксические критерии, как и лексические, можно разделить на подгруппы:
- критерии, описываемые структурами;
- критерии, учитывающие встречаемость определенных союзов, предлогов, средств связи и др.

В первую подгруппу входят следующие критерии:
- преобладание предложений с простой, сложноподчиненной или сложносочиненной структурой;
- наличие постпозитивных и препозитивных определений;
- преобладание препозитивных определительных групп.

Во второй группе выделяются следующие маркеры:
- использование двойных и составных союзов;
- использование слов, являющихся в литературном языке архаизмами;
- составные предлоги;
- средства логической связи.

Предполагается, что числовые оценки значений описанных выше характеристик для определенной предметной области возможно получить, проведя анализ корпуса текстов, состоящего из статей по соответствующей тематике. Разрабатываемый сервис для анализа и оценки стиля должен получать на вход корпус текстов, состоящий из научных статей на английском языке, посвященных сходной тематике. Последнее уточнение нужно для более удобного выделения терминов в тексте, например, с использованием терминологического словаря на заданную тематику.

После автоматической токенизации текста и нанесения разметки на слова и конструкции, которые описаны в качественных критериях академической речи, приложение предоставляет пользователю возможность проверить и редактировать автоматически нанесенную разметку.

Проанализировав аннотированный корпус, приложение выдает по каждому из критериев, представленных выше, статистическую информацию, например:

— среднее число маркеров данного типа, встречающееся в документах корпуса;

— наибольшее и наименьшее количественные значения характеристики;

— опционально: распределение числа маркеров в каждом документе;

— для критериев, оценивающих встречаемость частей речи, типов предложений и др. процентное соотношение относительно общего числа слов/предложений.

Должен учитываться тот факт, что размер статей может быть различным, поэтому статистическую информацию необходимо отражать как в абсолютных числах, так и относительно общего числа слов в документе. Полученная в результате работы сервиса информация может быть использована экспертом для интерпретации и проверки ее валидности.

Для формирования рекомендаций сервис должен получить отдельную статью, разметить ее по тем же типам аннотации (маркерам стиля), которые оцениваются в корпусе, и сравнить расстояния между значениями статистических характеристик корпуса и статьи по некоторой метрике.

Кроме функциональных требований, выделяются также нефункциональные, касающиеся особенностей целевой аудитории портала и особенностей использования.

Академический стиль речи и методы его идентификации и анализа

Научный (академический) стиль речи представляет научную сферу общения и речевой деятельности, связанную с реализацией науки как формы общественного сознания [6]. Иными словами, это особый функциональный стиль речи, который в широком смысле используют для сообщения нового знания

о действительности и доказательства ее истинности. Этот стиль используют в научных статьях, учебной литературе, монографиях и т. д.

В настоящее время у людей, занимающихся научными исследованиями, нередко возникает необходимость освоения академического стиля английского языка, для того чтобы их научные работы и статьи могли быть признаны международным академическим сообществом. Для помощи в изучении особенностей данного стиля создаются справочные и учебные материалы, а также обучающие интернет-ресурсы. Существует множество учебных пособий на английском языке, предназначенных для обучения академическому письму. Методические пособия могут быть адресованы исследователям и студентам [21], преподавателям [12, 14, 17, 23, 26], редакторам изданий [15, 25]. Имеются также учебные пособия от русскоязычных авторов [8—10].

Кроме того, доступны интернет-ресурсы, содержащие рекомендации к написанию научных текстов. В частности ресурсы, называемые *academic phrasebanks*, содержат списки конструкций, которые могут быть использованы в тех или иных ситуациях: когда нужно выразить критический взгляд, обозначить дистанцию между приводимым мнением и мнением автора, описать классификацию, привести результаты сравнения и примеры, ввести понятие, и т. д. Пример подобного ресурса — *Academic Phrasebank (University of Manchester)* [11].

Особенностям академического стиля английского языка были посвящены многие научные работы, в том числе русскоязычные диссертации и исследования как 80-х годов прошлого века, так и относительно недавнего времени. В исследованиях охвачены как общие структурные и функциональные особенности научных текстов [4], так и узкие темы, например, выражение экспрессивности [7], причинно-следственных отношений и др.

Классификация текстов, в том числе по функциональным стилям (разговорный, научный, художественный и др.) — одна из задач обработки естественного языка. Определение функционального стиля текста можно использовать в информационном поиске, машинном переводе, генерации текстов [5] для получения более точных и удовлетворяющих пользователя результатов.

В работе [5] представлены методы машинного обучения для классификации текстов, в которых использованы в том числе описанные выше признаки, разделенные авторами на лексические (число слов, N-грамм, глубина дерева синтаксического разбора и т. д.) и квантитативные (число символов, число слогов и т. д.). Классифицируемые документы сопоставляют с векторами признаков.

Подход к анализу академического стиля был представлен в работе [20]. Выделяются признаки текста, отвечающие за формальный стиль речи (пассивный залог, субъективные выражения, вопросы), читаемость (союзы и другие соединяющие фразы, использование существительных вместо глаголов) и научный язык (по списку из 200 научных слов и др.), и на основе выделенных признаков создается самоорганизующаяся карта (*Self-Organizing Map* — особая разновидность нейронной сети).

Еще один метод был использован для анализа характеристик в заголовках статей по теме "Computer Science". Значения параметров, таких как длина заголовка, использование пунктуации и предлогов, частота слов, были исследованы на примере корпуса статей из научных журналов [12].

Метод, рассматриваемый в данной работе, предполагает сравнение статей, стиль которых признан экспертами качественным, со статьями, которые, несмотря на соблюдение рекомендаций из методической литературы, являются плохо читаемыми и не соответствуют стандартам. Сравнение проводят по признакам, связанным с лексическими и синтаксическими характеристиками текста, которые были выделены при анализе учебных пособий и научных работ об особенностях академического стиля речи. В качестве опоры для исследования берется коллекция "эталонных" статей из рецензируемых источников, каждой из которых ставится в соответствие вектор, описывающий встречаемость тех или иных лингвистических характеристик. В дальнейшем, при сравнении эталонных статей с работами студентов будет исследоваться влияние каждой из этих характеристик на оценку качества и читаемости текста.

Особенностью этого метода является ориентация на практическое применение теоретических рекомендаций по соблюдению академического стиля. Проведение исследований на коллекции текстов, собранных по общим признакам — в данном случае по языку и стилю, — относит метод к области корпусной лингвистики.

Существуют различные инструменты для работы с корпусами текстов — от обычных сайтов для простого просмотра и поиска данных до систем, позволяющих создавать собственные приложения. Некоторые из этих

инструментов могут быть использованы при реализации метода анализа и получения оценки соответствия статей академическому стилю английского языка, описанного в работе [22].

Моделирование портала

На рис. 1 (см. вторую сторону обложки) представлен порядок действий пользователя на портале.

Под выбором сценария подразумевается выбор процесса анализа корпуса или формирования рекомендаций для отдельной статьи. В случае последнего варианта пользователь загружает статью перед переходом к выбору корпуса. Помимо пользователя в качестве акторов выделены различные компоненты, которые могут представлять собой отдельные сервисы (см. таблицу).

После открытия корпуса пользователь может отредактировать разметку вручную (данная активность опущена на диаграмме) или перейти на следующий шаг анализа. Рассмотрим вариант, когда пользователю необходимо автоматически разметить корпус, создав собственный тип аннотаций.

Представленная последовательность описывает цикл работы пользователя портала по основным сценариям без учета некоторых вариантов использования, которые будут восстановлены при дальнейшей реализации портала. При проектировании также были выделены компоненты, сервисы портала, которые выполняют отдельную смысловую функцию.

Архитектура портала

Как показано в предыдущем разделе, различные функции портала могут быть разделены на взаимодействующие друг с другом, но сравнительно независимые компоненты-сервисы. Подобное разделение хорошо подходит для распределенных систем и позволяет оптимально распределять нагрузку между серверами и повысить надежность системы.

Функции описанных сервисов могут быть условно разделены на функции интерфейса, отвечающие за взаимодействие с пользователем, бизнес-логику — основные инструменты портала, предназначенные для анализа корпусов и формирования рекомендаций для отдельных статей, и работу с данными. Портал должен иметь базу данных для хранения пользовательской информации, а также отдельное хранилище для корпусов, так как они могут занимать большие объемы памяти. Таким образом, архитектура портала является частным случаем трехслойной архитектуры. Упрощенное описание архитектуры представлено на рис. 2 (см. вторую сторону обложки). Трехслойная архитектура предполагает наличие следующих элементов: слой интерфейса (представления), слой приложения (домена, бизнес-логики), сервер БД (источник данных).

Портал является браузерным приложением и работает с тонким клиентом. Слой интерфейса отвечает за взаимодействие с пользователем, а также содержит некоторые связанные с этим модули, такие как визуальный редак-

Описание сервисов портала

Обозначение	Название компонента	Функции
System	Компонент взаимодействия с пользователем (далее Система)	Осуществляет взаимодействие с пользователем (функции интерфейса), управление остальными сервисами (функции диспетчера)
Search Engine	Сервис поиска корпусов	Выполняет поиск корпуса по запросу
Corpus Visualizer	Визуальный редактор разметки	Визуализирует разметку корпуса для пользователя, предоставляет возможность ручного редактирования разметки (добавления/редактирования/удаления аннотаций)
PR Editor	Визуальный редактор лексико-синтаксических шаблонов	Предоставляет интерфейс для создания лексико-синтаксических шаблонов и формирует на их основе обрабатывающие ресурсы для автоматической разметки корпуса
Annotator	Компонент разметки корпуса	Наносит разметку документа/корпуса автоматически, используя обрабатывающие ресурсы, соответствующие типам аннотаций
Statistics Processor	Компонент сбора статистики	Собирает статистические данные на основе разметки корпуса
Report Generator	Компонент формирования отчетов	Формирует отчеты-анализы и отчеты-сравнения, используя статистические данные; проводит сравнение характеристик документа и корпуса

тор разметки и визуальный редактор лексико-синтаксических шаблонов. Последние два модуля могли бы быть вынесены на слой бизнес-логики, однако их роль больше связана с интерфейсом и обеспечением более удобной и понятной для пользователя работы по анализу корпусов, нежели с основной задачей портала.

Слой приложения содержит модули для выполнения основных функций по работе с корпусами, а также социальную часть портала. Сервис поиска корпусов позволит оперативно подбирать корпус в зависимости от указанных параметров, ключевых слов или даже, возможно, от всего текста загруженной пользователем статьи. Разработка этого модуля чрезвычайно важна для портала, так как качество оценки стиля статьи по реализуемому методу будет напрямую зависеть от степени семантической близости пользовательской статьи и эталонного корпуса.

Социальная часть портала включает в себя форум, инструкции по работе с порталом и коллекцию материалов, в том числе элементы, созданные пользователями и переведенные ими в общий доступ: например, новые обрабатывающие ресурсы, списки слов и т. д. с описаниями, а также результаты анализов корпусов, проведенных с помощью портала. Теоретически социальная часть может считаться частью слоя взаимодействия с пользователем, однако вследствие более тесной связи с базой данных и необходимости обеспечивать интеллектуальный поиск (по аналогии с поиском корпусов) в данной архитектуре он является частью слоя приложения. Стоит отметить, что указанные в данном слое сервисы, как и модули визуальных редакторов в слое представления, являются достаточно независимыми и могут быть развернуты на различных серверах в случае большой нагрузки.

Наиболее важным для портала является сервис обработки корпусов, который и выполняет основную функцию системы — анализ корпусов и формирование рекомендаций по качеству стиля для отдельных статей. Согласно предложенному методу анализа корпусов его следует выполнять в три этапа, каждый из которых реализован в отдельном компоненте архитектуры. Предполагается, что данный сервис будет активно использовать ресурсы библиотеки GATE Embedded для работы с корпусами текстов и созданные с ее помощью специальные элементы.

Первый компонент — компонент разметки корпуса. Он отвечает за обработку выбранного корпуса с помощью имеющихся ресурсов, созданных на основе имеющихся плагинов среды GATE (например, базовые плагины для токенизации, разбиения на предложения, морфологического разбора), специализированных, изначально добавленных в систему обрабатывающих ресурсов (к таким будут относиться средства для нанесения маркеров стиля по критериям, обозначенным в предыдущих разделах) или обрабатывающих ресурсов, созданных пользователем с помощью визуального редактора лексико-синтаксических шаблонов. Последние будут также использовать специализированные плагины, написанные с помощью библиотек GATE Embedded, только настроенные на выделение пользовательских лингвистических конструкций, списков слов и др.

Разметка корпуса является необходимым для дальнейшего анализа этапом, однако при использовании корпусов, уже имеющихся в системе, дополнительная разметка может не понадобиться, особенно при использовании портала "не лингвистами", для оценки качества стиля собственных статей. Поэтому работа данного компонента может быть пропущена. Стоит уточнить, что при разметке корпуса с помощью некоторых лексико-синтаксических шаблонов может понадобиться ручное редактирование аннотаций, поэтому между этапом разметки и этапом сбора статистики нужна возможность возвращения результатов в интерфейс для работы с пользователем. Этим, в частности, объясняется разделение компонентов разметки и сбора статистики.

Таким образом, компонент разметки корпуса получает на вход корпус и список обрабатывающих ресурсов, аннотации от которых должны быть нанесены на корпус. Для новых корпусов, которые еще не имеют разметки и, соответственно, аннотаций никаких типов, перед запуском обрабатывающих ресурсов по умолчанию проводится базовая обработка от токенизации до морфологического разбора. На выходе сервис, реализуемый данным компонентом, выдаст корпус с обновленной разметкой. Корпус и разметка принимаются и выдаются в формате, пригодном для вывода в интерфейс, обработки с помощью ресурсов GATE и последующих компонентов. При необходимости можно встроить механизмы конвертации в удобные для пересылки данных форматы и обратно для модулей визуализации и компонентов сервиса обработки корпусов.

Вторым этапом как анализа, так и формирования рекомендаций по стилю статьи, явля-

ется сбор статистики по имеющейся разметке. Данный сервис принимает в качестве входных данных: размеченный корпус; список "тегов", т. е. типов аннотаций, которые будут учитываться при сборе статистики; типы статистических характеристик, которые должны быть вычислены для отчета.

Этот компонент подсчитывает базовые характеристики, например, число слов в документе, число аннотаций определенного типа в документе, число предложений, содержащих такие аннотации и т. д. Затем на их основе вычисляются указанные во входных данных статистические характеристики — как для корпуса в целом, так и для отдельных документов. В качестве выходных данных выводятся вычисленные значения характеристик в удобном формате, например, XML-файла.

Последний компонент сервиса отвечает за создание отчетов по анализу корпуса и формирование рекомендаций для отдельной статьи на основе сравнения с эталонным корпусом по теме. Для анализа корпуса необходимо только вывести его в определенном удобном для пользователя формате, формирование же рекомендаций требует отдельной обработки статистических характеристик документа и эталонного корпуса. Результат также выводится в удобном для пользователя формате в виде рекомендаций с указаниями на количественные различия между документом и средними результатами по корпусу.

Последний слой архитектуры — работа с данными: базой пользователей и их документов, корпусов, отчетов и других файлов, а также отдельное хранилище корпусов. Хранилище корпусов должно быть достаточно большим или по крайней мере масштабируемым, так как в ходе развития популярности портала предполагается размещение крупных коллекций документов.

Предложенная архитектура обеспечивает высокую степень масштабируемости и надежность системы за счет полной или относительной независимости отдельных ее компонентов. Важным при дальнейшей разработке продукта будет являться выб ор стандартов передачи данных между отдельными сервисами для повышения скорости, так как корпус текстов может быть достаточно объемным — вплоть до нескольких тысяч документов.

Заключение

В ходе данного исследования были выделены функциональные и нефункциональные требования к реализации портала для анализа и оценки стиля научных публикаций на английском языке. В целом есть два основных пути использования портала. Первый путь — это анализ статистических характеристик разметки корпуса для профессиональных лингвистов, проводящих исследования собственных корпусов. Второй путь — сравнение характеристик статьи пользователя с характеристиками эталонного корпуса для пользователей, которым не требуются дополнительные настройки с точки зрения лингвистики, только рекомендации системы. Первый сценарий требует большой гибкости сервиса и работы с объемными данными, второй — максимальной простоты интерфейса. Функциональные требования к порталу представлены в виде диаграмм вариантов использования, основные процессы жизненного цикла системы и взаимодействие компонентов внутри системы также описаны с помощью диаграмм.

Функции портала естественным образом подразделяются на функции взаимодействия с пользователем, основные сервисы и работу с данными. В связи с этим для реализации портала предложена трехслойная архитектура, каждый слой которой составляют взаимодействующие друг с другом, но фактически независимые компоненты. Подобная архитектура способствует улучшению масштабируемости и надежности системы, так как каждый компонент может располагаться на отдельных серверах. Весь портал и хранилища данных могут быть размещены в облачном сервисе.

Список литературы

1. **Шаров С. А.** Представительный корпус русского языка в контексте мирового опыта // НТИ. Сер. 2. 2003. № 6. С. 9—17.

2. **Корпусная** лингвистика / Теория. URL: http://corpora.iling.spb.ru/theory.htm (дата обращения: 04.02.2017).

3. **Scholz T., Conrad S.** Style Analysis of Academic Writing // Natural Language Processing and Information Systems: proceedings of 16th International Conference on Applications of Natural Language to Information Systems (Alicante, Spain, June 28—30, 2011). 2011. P. 246—249.

4. **Бартков Б. И., Минина Л. И., Миронец Ю. А.** Структурные и функциональные особенности научного текста. Академия наук СССР, Дальневосточный науч. центр, Кафедра иностранных языков, 1985. 147 с.

5. **Ермакова Л. М., Абашев М. А., Никитин Р. В., Ушаков Р. И.** Методы автоматической классификации текстов по функциональным стилям // Вестник Пермского университета. Математика. Механика. Информатика. 2014. Вып. 4 (27). С. 78—83.

6. **Научный** стиль — Стилистический энциклопедический словарь русского языка. URL: http//stylistics.academic.ru/89/Научный_стиль (дата обращения: 05.06.2015).

7. **Скрипак И. А.** Языковое выражение экспрессивности как способа речевого воздействия в современном научном дискурсе: на материале статей лингвистического профиля на русском и английском языках. Дис. канд. филол. наук. Ставрополь, 2008. 199 с.

8. **Смирнова Л. Н.** Курс английского языка для научных работников. Л.: Наука, 1971. 330 с.

9. **Смирнова Л. Н.** Scientific English. Английский язык для научных работников. Курс для начинающих. Л.: Наука, 1980. 245 с.

10. **Цвиллинг М. Я.** Научная литература: язык, стиль, жанры. М.: Наука, 1985. 336 с.

11. **Academic** Phrasebank. URL: http://www.phrasebank.manchester.ac.uk (дата обращения: 20.01.2015).

12. **Anthony L.** Characteristic features of research article titles in computer science // IEEE Transactions on Professional Communication. 2001. N. 44 (3). P. 187—194.

13. **Guse J.** Communicative Activities for EAP. Cambridge: Cambridge University Press, 2013. 322 p.

14. **Hamp-Lyons L., Heasley B.** Study writing. A course in writing skills for academic purpose. Cambridge: Cambridge University Press, 2013. 213 p.

15. **IEEE** — The world's largest professional association for the advancement of technology. URL: https://www.ieee.org/index.html (дата обращения: 05.06.2015).

16. **Jacobs P. S., Krupka G. R., Rau L. F.** Lexico-Semantic Pattern Matching as a Companion to Parsing in Text Understanding // Workshop on Speech and Natural Language Colocated with the 6th Human Language Technology Conference. 1991. P. 337—341.

17. **Jordan R. R.** English for academic purposes. A guide and resource book for teachers. Cambridge: Cambridge University Press, 2012. 404 p.

18. **Khosmood F., Levinson R. A.** Automatic natural language style classification and transformation / BCS Corpus Profiling Workshop, 2008. URL: https://style.soe.ucsc.edu/sites/default/files/CP08-KL-camera.pdf (дата обращения: 08.02.2017).

19. **Laurence** Anthony's AntConc. URL: http://www.laurenceanthony.net/software/antconc/ (дата обращения: 08.02.2017).

20. **Luyckx K., Daelemans W.** Shallow text analysis and machine learning for authorship attribution // Computational Linguistics in the Netherlands 2004: selected papers from the Fifteenth CLIN Meeting. Utrecht, LOT. 2005. P. 149—160.

21. **Scolz T., Conrad S.** Style Analysis of Academic Writing// Natural Language Pocessing and Information Systems: 16th International Conference on Appications of Natural Language to Information Systems, Proceedings. NLDB 2011, Alicante, Spain, June 28—30, 2011. P. 246—249.

22. **Siepmann D., Gallagher J. D., Hannay M., Mackenzie J. L.** Writing in English: a guide for advanced learners. Francke Verlag, 2011. 469 p.

23. **Strinyuk S. A., Shuchalova Y., Lanin V.** Academic Papers Evaluation Software // Application of Information and Communication Technologies (AICT), 2015 9th International Conference (14—16 Oct. 2015. Rostov-on-Don). IEEE, 2015. P. 506—510.

24. **Swales J. M., Feak C. B.** Academic writing for graduate students. Essential tasks and skills. 3d. Ed. The University of Michigan, 2014. 418 p.

25. **TÜPH** Научный портал. URL: https://kaemus.psych.ut.ee/&lang=Rus (дата обращения: 15.05.2017).

26. **Turabian K.** A Manual For Writers of Term Papers, Theses and Dissertations. 7th ed. The University of Chicago Press, 2007. 470 p.

27. **Wallwork A.** English for academic research: vocabulary exercises. London: Springer, 2013. 193 p.

Yu. S. Shuchalova, Invited Lecturer, e-mail: iusshuchalova@hse.ru,
V. V. Lanin, Senior Lecturer, e-mail: vlanin@hse.ru,
National Research University Higher School of Economics, Perm

Research Portal for Scientific Publication Style Analysis

As any other genre, academic paper can be characterized by its own specific rules and features. Academic writing has investigated in terms of grammar, structure, genre and other crucial features, however, not enough attention has been paid to building a systematic approach. Recommendations given in guides and handbooks for both competent and novice academic writers in English are not systematized and sometimes even have obvious internal contradictions. The project named "Paper Cat" is aimed to provide corpus analysis tools for scientific publications in English. Authors assume that it possible to develop special software tools which can be able to perform automatic analysis based on natural language processing and corpus linguistic methods. The article describes a development of the Internet portal as a part of a project. Functional and non-functional requirements to the system are given and appropriate portal architecture is suggested. Heterogeneous components integrate due to service architecture. Implementation features are discussed taking into account the specifics of the subject area. Described portal can be used both for research and study aims.

Keywords: *corpus studies, research portal, academic English, information system modeling*

DOI: 10.17587/it.24.515-523

References

1. **Sharov S. A.** Predstavitel'nyj korpus russkogo jazyka v kontekste mirovogo opyta, *NTI*, Ser. 2. 2003. N. 6. P. 9—17 (in Russian).

2. **Korpusnaja** lingvistika / Teorija, available at: http://corpora.iling.spb.ru/theory.htm (available date: 04.02.2017) (in Russian).

3. **Scholz T., Conrad S.** Style Analysis of Academic Writing. *Natural Language Processing and Information Systems: 16th International Conference on Applications of Natural Language to Information Systems* (Alicante, Spain, June 28—30, 2011), 2011, pp. 246—249.

4. **Bartkov B. I., Minina L. I., Mironec Ju. A.** *Strukturnye i funkcional'nye osobennosti nauchnogo teksta*. Akademija nauk SSSR, Dal'nevostochnyj nauch. centr, Kafedra inostrannyh jazykov, 1985, 147 p. (in Russian).

5. **Ermakova L. M., Abashev M. A., Nikitin R. V., Ushakov R. I.** Metody avtomaticheskoj klassifikacii tekstov po funkcional'nym stiljam, *Vestnik Permskogo universiteta. Matematika. Mehanika. Informatika*, 2014, no. 4 (27), pp. 78—83 (in Russian).

6. **Nauchnyj stil'** — Stilisticheskij jenciklopedicheskij slovar' russkogo jazyka / Eds by M. N. Kozhina; 2nd edition, Moscow, Flinta: Science, 2006, 696 p. (in Russian).

7. **Skripak I. A.** Jazykovoe vyrazhenie jekspressivnosti kak sposoba rechevogo vozdejstvija v sovremennom nauchnom diskurse: na materiale statej lingvisticheskogo profilja na russkom i anglijskom jazykah. Phd thesis. Stavropol', 2008. 199 p. (in Russian).
8. **Smirnova L. N.** *Kurs anglijskogo jazyka dlja nauchnyh rabotnikov.* Leningrad: Nauka, 1971, 330 p. (in Russian).
9. **Smirnova L. N.** *Scientific English. Anglijskij jazyk dlja nauchnyh rabotnikov. Kurs dlja nachinajushhih.* Leningrad: Nauka, 1980, 245 p. (in Russian).
10. **Cvilling M. Ja.** *Nauchnaja literatura: jazyk, stil', zhanry*, Moscow: Nauka, 1985, 336 p. (in Russian).
11. **Academic** Phrasebank, available at: http://www.phrasebank.manchester.ac.uk/ (date of access: 20.01.2015).
12. **Anthony L.** Characteristic features of research article titles in computer science, *IEEE Transactions on Professional Communication*, 2001, no. 44 (3), pp. 187—194.
13. **Guse J.** *Communicative Activities for EAP.* Cambridge, Cambridge University Press, 2013, 322 p.
14. **Hamp-Lyons L., Heasley B.** *Study writing.* Cambridge, Cambridge University Press, 2013, 213 p.
15. **IEEE** — The world's largest professional association for the advancement of technology, URL: https://www.ieee.org/index.html (date of access: 05.06.2015).
16. **Jacobs P. S., Krupka G. R., Rau L. F.** Lexico-Semantic Pattern Matching as a Companion to Parsing in Text Understanding, *Workshop on Speech and Natural Language colocated with the 6th Human Language Technology Conference*, 1991, pp. 337—341.
17. **Jordan R. R.** English for academic purposes, Cambridge, Cambridge University Press, 2012, 404 p.
18. **Khosmood F., Levinson R. A.** Automatic natural language style classification and transformation, *BCS Corpus Profiling Workshop*, 2008, available at: https://style.soe.ucsc.edu/sites/default/files/CP08-KL-camera.pdf (date of access: 08.02.2017).
19. **Laurence** Anthony's AntConc, available at: http://www.laurenceanthony.net/software/antconc/ (date of access: 08.02.2017).
20. **Luyckx K., Daelemans W.** Shallow text analysis and machine learning for authorship attribution, *Computational Linguistics in the Netherlands 2004: selected papers from the Fifteenth CLIN Meeting*, van der Wouden T. [Ed.], e. a., Utrecht, LOT, 2005, pp. 149—160.
21. **Scolz T., Conrad S.** Style Analysis of Academic Writing, *Natural Language Pocessing and Information Systems: 16th International Conference on Appications of Natural Language to Information Systems*, Proceedings. NLDB 2011, Alicante, Spain, June 28—30, 2011, pp. 246—249.
22. **Siepmann D.** *Writing in English*: A Guide for Advanced Learners / J. D. Gallagher, M. Hannay; J. L. Mackenzie. — UTB, 2011, 469 p.
23. **Strinyuk S. A., Shuchalova Y., Lanin V.** Academic Papers Evaluation Software, *Application of Information and Communication Technologies (AICT), 2015 9th International Conference (14—16 Oct. 2015. Rostov-on-Don)*: IEEE, 2015, pp. 506—510.
24. **Swales J. M., Feak C. B.** *Academic writing for graduate students. Essential tasks and skills.* The University of Michigan, 2014, 418 p.
25. **TÜPH.** Nauchnyy portal, available at: https://kaemus.psych.ut.ee/&lang=Rus (date of access: 15.05.2017) (in Russian).
26. **Turabian K.** A Manual For Writers of Term Papers, Theses and Dissertations. — 7th ed. — The University of Chicago Press, 2007, 470 p.
27. **Wallwork A.** *English for academic research: vocabulary exercises*, London, Springer, 2013, 193 p.

УДК 004.421.6 DOI: 10.17587/it.24.523-528

Н. И. Лиманова, д-р техн. наук, проф., e-mail: Nataliya.I.Limanova@gmail.com,
М. Н. Седов, аспирант, e-mail: SedovMN@inbox.ru,
Поволжский государственный университет телекоммуникаций и информатики, г. Самара

Алгоритм нечеткого поиска реквизитов физических лиц в базах данных на основе метрики Левенштейна

При передаче данных от одного учреждения к другому возникает проблема персональной идентификации физических лиц, у которых частично или полностью не совпадают реквизиты. В работе представлен алгоритм нечеткого поиска, использующий модифицированную метрику Левенштейна, позволяющий выполнять поиск физических лиц в базе данных на основе нечеткого сравнения. Алгоритм реализован на языке PL-SQL в СУБД Oracle 11g.

Ключевые слова: *межведомственный информационный обмен, нечеткое сравнение, поиск персональных данных, функция интеллектуального сравнения, персональный идентификационный номер (ПИН)*

Введение

В процессе обработки информации о физических лицах в базах данных для удобства обработки каждому набору реквизитов физических лиц (таких как ФИО, адрес, номера паспорта, СНИЛС и т. п.) присваивается так называемый персональный идентификационный номер (ПИН). В случае обработки или передачи данных о физическом лице вся привязка осуществляется именно к этому ПИНу. При осуществлении обмена информацией о физических лицах между различными учреждениями возникает проблема сопоставления реквизитов из одной базы данных реквизитам в другой. Если проводить данное сопоставление методом простого сравнения реквизитов (метод прямого сравнения), то в случае ошибочных данных,