УДК 81-114.4

СОЗДАНИЕ ВСПОМОГАТЕЛЬНОГО ИНФОРМАЦИОННОГО РЕСУРСА ДЛЯ АНАЛИЗА УЧЕБНЫХ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Ю.Н.Баранова,
Национальный исследовательский университет – Высшая школа экономики,
e-mail: ligros7@gmail.com

Т.С. Елипашева,
Национальный исследовательский университет – Высшая школа экономики,
e-mail: t.elipasheva@gmail.com

Аннотация. В данной статье освящается тема создания вспомогательного информационного ресурса «Лексикатор» для педагогов и студентов, занимающихся обучением и изучением русского языка как иностранного. С этой целью реализован функционал для автоматического выделения сложных синтаксических и лексических структур, а так же система оценки сложности текста для чтения по ряду распространённых индексов и собственному индексу «Лексикатора». Ожидается, что разрабатываемый ресурс будет использован для подготовки, определения сложности и анализа учебных текстов в ходе подбора материалов для занятий.

Ключевые слова: русский язык как иностранный, компьютерная лингвистика, лексическая сложность, синтаксическая сложность, сложность для чтения

CREATING OF AN INFORMATION RESOURCE FOR EDUCATIONAL TEXT ANALYSIS IN RUSSIAN

Yu.N.Baranova,
National Research University – Higher School of Economics,
e-mail: ligros7@gmail.com

T.S. Elipasheva,
National Research University – Higher School of Economics,
e-mail: t.elipasheva@gmail.com

Abstract. This article tells about creation of an information resource «Leksikator» for teachers and students engaged in teaching and learning Russian as a foreign language. To do this functional for automatic selection of the complex syntactic and lexical structures, as well as the system for evaluation of text complexity for reading on a number of common indices and an authentic

«Leksikator» index are implemented. It is expected that the resource will be used for the analysis and preparation material for lessons and evaluation of educational text complexity.

Keywords: Russian as a foreign language, computer linguistics, lexical complexity, syntactic complexity, reading complexity.

Введение

С ростом информатизации общества всё большее внимание уделяется внедрению компьютерных технологий во все сферы деятельности человека, неотъемлемой частью которых является сфера образования. Это служит причиной постоянно увеличивающегося числа информационных ресурсов, отвечающих потребностям хранения, обработки и распространения учебных и вспомогательных материалов.

Большое количество таких ресурсов создаётся в рамках обучения иностранному языку, что обусловлено процессами глобализации как тенденции развития современного общества, приводящей к повышению потребности в качественных учебных материалах. Важными критериями при подборе материалов для обучения являются их актуальность и адаптированность для обучающихся. Одной из неотъемлемых частей процесса обучения языку является чтение, что часто ставит перед преподавателями и студентами ряд технических трудностей, связанных с проблемами поиска или понимания текстов определённого уровня сложности. Если в рамках решения задачи подбора актуального материала могут использоваться корпусы современных текстов, то проблема адаптации к уровню знания языка носит более нетривиальный характер. Причиной этого служит высокая сложность анализа конструкций языка, которая обычно производится педагогом вручную, занимая большое количество времени и являясь достаточно рутинной процедурой.

С целью снижения сложности описанной выше задачи по адаптации текстовых учебных материалов в рамках учебных курсов по русскому языку как иностранному разработан ресурс, позволяющий автоматизировать ряд наиболее трудоёмких операций по работе с такими текстами, включая определение уровней лексических и синтаксических конструкций, основных параметров текста и объективные меры сложности текста для чтения.

Общее описание возможностей

Так как ресурс создан с целью автоматизации работы по подготовке учебного материала, его основной функционал заключается в нахождении сложных конструкций в обрабатываемом учебном тексте.

Задача адаптации текста подразумевает обработку на двух уровнях – синтаксическом и лексическом. На уровне синтаксической обработки необходимо обнаруживать синтаксические конструкции, которые могут быть слишком сложными для изучающих язык. Аналогичная, по сути, задача стоит при обработке на уровне лексики – выделять следует ту лексику, которая вызывает затруднение у учеников. Выбор незнакомых слов на уровне проверки

лексики основан на словарях лексических минимумов различных уровней, выбор между которыми можно осуществлять в интерфейсе «Лексикатора».

Графически представление учебного текста с выделением сложных синтаксических и лексических конструкций выглядит как подсветка указанных конструкций во введённом в текстовое поле интерфейса материале. Поля ввода графического пользовательского интерфейса рассчитаны на длину учебных текстов в пределах четырёх тысяч символов.

Для формальной оценки сложности текста для чтения используется нескольких распространённых индексов читаемости (индекс Дейла-Холла, индекс Флеша-Кинкейда и т.д.), а так же собственного индекса, вычисляемого по расширенным статистическим метрикам, извлечённым из текста. Извлечённые метрики доступны в графическом интерфейсе ресурса и могут использоваться для проведения собственного анализа.

Описание пользовательского интерфейса

Сайт, где располагается ресурс «Лексикатор», на данный момент находится на этапе активной разработки, но основная функциональность, описанная в статье, уже присутствует в пользовательском интерфейсе. Как можно увидеть на Рисунке 1, интерфейс имеет достаточно простую и понятную пользователю структуру. На вход подается некоторый исходный текст, который необходимо обработать. Реализовано два уровня обработки, которые располагаются отдельными вкладками в нижней части сайте под полем для введения исходного текста:

- Лексический уровень обработки - выделение слов, не входящих в лексический минимум для конкретного выбранного уровня владения языком: A1, A2 или B1. При выполнении лексической обработки необходимо выбрать уровень языка. По нажатию одной из трёх кнопок содержимое поля «Исходный текст» обрабатывается и выводится в нижнем текстовом поле в размеченном виде. Слова, входящие в лексический минимум, не маркированы, слова, не входящие в лексический минимум подсвечены жёлтым цветом.
- Структурный уровень обработки – выделение структурных осложнителей текста. При структурном уровне обработки выбор уровня владения языком блокирован. Сложные конструкции определены сводом разработанных правил, процесс составления которых описывается ниже. В конце процесса обработки в результирующем тексте жёлтым цветом подсвечиваются конструкции, которых не должно быть в простом предложении.

На панели «Параметры» отображаются статистические показатели наиболее значимых параметров, которые используются для оценки сложности текста и подсчета различных индексов определения сложности текста для чтения. При разработке было принято решение, что оптимальным является постоянное отображение лишь четырёх параметров:

- Среднее количество слов в одном предложении текста;

- Процент слов в предложении, не входящих в словарь лексического минимума;
- Средняя длина слова в буквах;
- Средняя длина слова в слогах.

Полный список параметров, используемых при вычислении индексов сложности для чтения, описан далее и может быть отображён путем выбора опции «Расширенный список параметров». Сами значения индексов расположены слева от поля ввода исходного текста:

- Индекс Дейла-Холла
- Индекс Флеша-Кинкейда
- Индекс Лексикатора

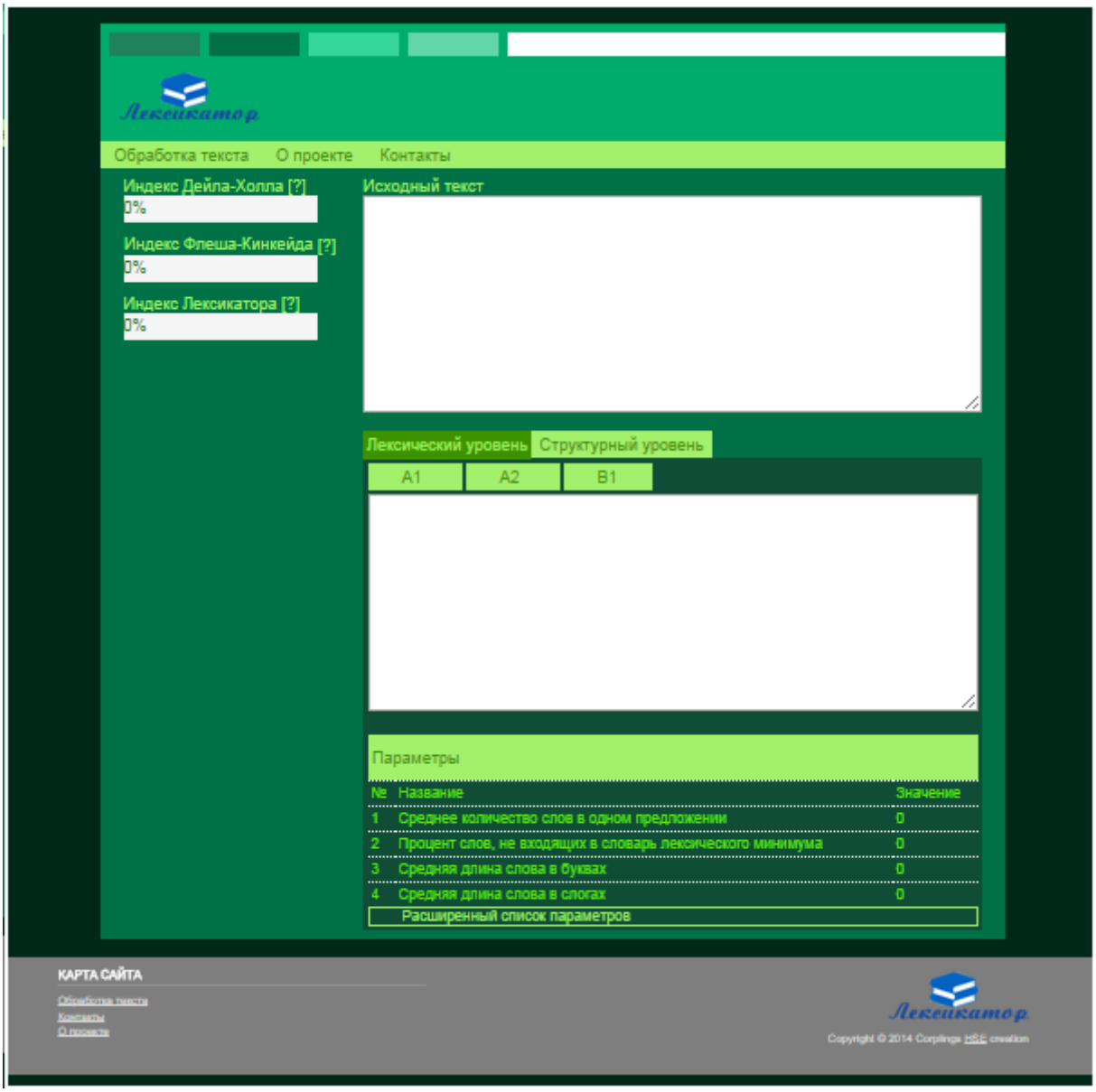

Рисунок 1. Пользовательский интерфейс ресурса «Лексикатор»

Для удобства пользователя при наведении на знак вопрос рядом с названием индекса на панели выводится справка с описанием индекса и алгоритмом его подсчета, что отражено на Рисунке 2.

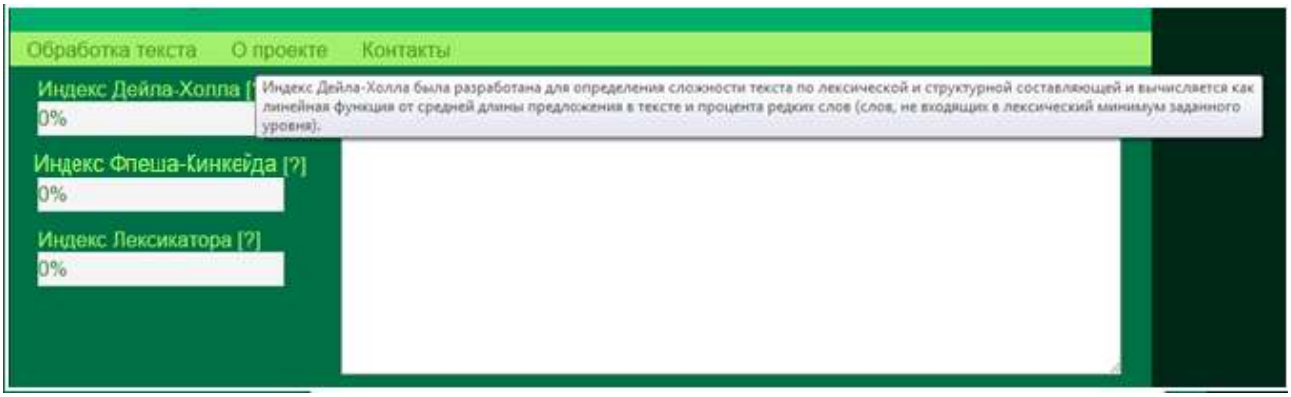

Рисунок 2. Пример всплывающей подсказки

Ресурс снабжен информацией о целях проекта, актуальности задачи и руководством пользователя. Данная информация располагается на вкладке «О проекте». Раздел «Контакты» предоставляет информацию о контактных данных команды разработчиков информационного ресурса.

Лексический уровень обработки

Как уже было указано выше, задача обработки на лексическом уровне состоит в поиске и отображении в тексте (графически - подсветке) сложных для выбранного уровня владением языка слов.

В ходе решения задачи реализации лексической подсветки, командой, работающей над ресурсом «Лексикатора», рассматривались три лексических уровня языка согласно государственному стандарту изучения русского языка как иностранного[1]: элементарный (A1), базовый (A2) и первый сертификационный (B1). Основная работа по обнаружению сложных конструкций в дальнейшем велась именно с учётом этого набора уровней и именно они представлены в графическом интерфейсе ресурса.

Каждый уровень охватывает умения обучающегося в областях аудирования, чтения, письма и говорения. На данный момент разработанный ресурс позволяет развивать речевые умения лишь в области работы с текстовой информацией, в чём изначально и состояла исследовательская цель проекта. Для каждого уровня владения русского языка в области чтения существуют определенные нормы (Таблица №1).

Уровень	Требования	Объем текста	Кол-во незнакомых слов	Лексический минимум
Элементарный (A1)	• читать текст с установкой на общий охват его содержания; • определить тему текста;	250-300 слов.	1-2%.	780 единиц

	• понять достаточно полно и точно основную информацию текста, а также некоторые детали, несущие важную смысловую нагрузку.			
Базовый (А2)	• читать текст с установкой на общий охват его содержания; • определить тему текста: понять его основную идею; • понять как основную, так и дополнительную информацию, содержащуюся в тексте, с достаточной полнотой, точностью и глубиной.	600-700 слов.	3-4 %.	1300 единиц
Первый сертификационный (В1)	• использовать различные стратегии чтения в зависимости от коммуникативной установки; • определить тему текста, понять его основную идею; • понять как основную, так и дополнительную информацию, содержащуюся в тексте, с достаточной полнотой, точностью и глубиной; • интерпретировать информацию, изложенную в тексте, выводы и оценки автора.	800-1000 слов.	до 5-7 %.	2300 единиц

Таблица №1. Таблица уровней владения русским языком как иностранным.

Остановиться на таком наборе уровней было решено по причине того, что уровни владения языком выше, чем B1, подразумевают способность обучающегося читать и понимать практически неадаптированные тексты.

Для каждого из описанных в Таблице 1 уровней был выделен некоторый набор слов в начальной форме, соответствующий определенному лексическому уровню. Наборы слов соответствуют словарям лексических минимумов уровней A1, A2, B1 из учебных пособий по изучению русского языка как иностранного [2, 3, 4]. Следует учесть, что каждый словарь более высокого уровня включает в себя словари всех уровней до него.

Для реализации запланированного функционала ресурса стояла задача распознавания структур и слов, не входящих в лексический минимум выбранного уровня, и выделение их для пользователя с целью последующего анализа лексической сложности текста, используя зрительное восприятие. Решение данной задачи было получено путём выполнения следующих шагов:

- Создание словарей на основе лексических минимумов трёх выбранных уровней изучения языка (A1, A2, B1);
- Получение обрабатываемого текста, введённого пользователем ресурса;
- Выполнение процессов токенизации и лемматизации (разбиение на отдельные слова и приведение этих слов к начальной форме);
- Проверка полученных токенов на наличие их в словаре выбранного лексического уровня;
- Подсветка токенов, не прошедших проверку (не обнаруженных в лексическом минимуме заданного уровня).

Для процесса токенизации и лемматизации использовалась свободно распространяемая библиотека PyMorphy для языка Python 2.7.

Таким образом, подсветка лексически сложных элементов текста реализована с помощью сопоставления и нахождения исключений между набором лексем текста и словаря рассматриваемого уровня изучения языка.

Причиной, по которой данный метод не может быть единственной составляющей оценки сложности текстового материала для обучающегося, является то, что помимо незнакомой лексики в нём могут встречаться сложные конструкции - вводные слова, обороты, предложения могут быть неполными и т.д. С целью учёта такого рода осложнителей было принято решение обрабатывать текст на втором уровне, так же доступном отдельной вкладкой на графическом интерфейсе ресурса – уровне выделения сложных конструкций.

Результат лексической обработки текста представлен на Рисунке 3.

Рисунок 3. Пример выделения лексических структур

Структурный уровень обработки

Данный уровень анализа позволяет определить наличие сложных синтаксических структур в тексте. Для решения этой задачи был разработан ряд правил, определяющих, какого рода конструкции в предложении считаются слишком сложными для восприятия обучающихся языку. В связи с тем, что не существует чётких границ соответствия сложности различных структурных осложнителей уровням изучения языка, все правила созданы для одного общего уровня.

Для составления правил использовались справочники по синтаксису русского языка. Выделенные на основе их правила затем были формализованы и записаны с помощью синтаксиса языка регулярных выражений, используя

теорию контекстно-свободных грамматик, и затем интегрированы с общим программным кодом на языке Python. В данном случае формализация является необходимым действием для трансляции естественного языка с помощью машинного кода. Результатом стала возможность графического отображения определенных структур, в том числе синтаксических и семантических, которые составляют трудность для восприятия в процессе обучения на базовом и первом уровне изучения языка.

Правила структурированы в соответствии с синтаксической системой русского языка:

- Коммуникативные – модальность, эмоциональная окрашенность
- (междометия, вводные компоненты, обращения);
- структурные: любые обороты, присоединительные конструкции, вставные конструкции;
- структурно-семантические.

Примеры правил:

Правило 1. Экспликаторы субъективной модальности (вводные слова):

1) , *Prnt* , (уточнение: , _ Prnt , ПОСЛЕ ЗАПЯТОЙ ЕСТЬ ПРОБЕЛ)
2) <начало предложения> *Prnt,*

Где Prnt - обозначение для вводных конструкций. Список вводный слов оформлен в отдельный словарь и составляет 366 единиц.

Правило 9. Наличие излишнего количества однородных членов:

1) структуры, где больше трёх членов (два и три члена возможны) однородного ряда (для существительных глаголов и прилагательных);
2) структуры, где больше двух членов однородного ряда (для причастий, наречий, деепричастий, предикативов).

Так же в правилах описаны следующие основные пункты, которые должны графически выделяться как сложные конструкции:

- модальные частицы (“что за”, “только”, “лишь” и т.д.);
- модальный постфикс “-то” (“какой-то”, “чей-то” и т.д.);
- междометия (“ах”, “ой” и т.д.);
- сложные составные слова (“диван-кровать” и т.д.);
- глагольные односоставные предложения;
- сравнительные обороты;
- синтаксические сращения;
- сравнительные придаточные обороты;
- сопоставительные придаточные обороты;
- прочие сложные конструкции, описанные формальными правилами.

Всего список правил содержит 20 общих пунктов с подпунктами, поясняющими схемы предложения в виде регулярных выражений и подробным списком слов и оборотов, которых не должно быть в предложениях текста для того, чтобы они считались простыми.

Первоначальной идеей разработки такого свода синтаксических правил была открываемая ими возможность автоматического упрощения предложения путём удаления всех сложных структур. Данная идея впоследствии претерпела изменения ввиду потери смысла предложения после удаления некоторых сложных структур и свелась к визуальному их выделению в тексте. В целом, задача автоматического упрощения текста может быть решена при разработке более детальных правил, на разработку которых необходимо большое количество человеческих ресурсов и времени.

Индексы определения сложности текста для чтения

Выделение сложных структур в обрабатываемом тексте позволяет получить только его приблизительную оценку, которую должен выполнять пользователь ресурса исключительно на основе своего опыта. Для того, чтобы реализовать возможность получения объективной оценки сложности учебного материала, используется вычисление ряда различных индексов сложности текста для чтения.

Сложность для чтения может быть представлена как функция, которая сопоставляет множеству признаков, извлечённых из текста, определённый уровень сложности из заранее определённых по какой-либо системе классификации. Традиционно, признаки, которые выделяются для характеристики рассматриваемых текстов, делятся на две группы - лексические параметры и синтаксические параметры. Конкретный список параметров зависит от индекса, который используется для вычисления показателя сложности текста для чтения.

На данный момент реализовано три различных индекса, каждый из которых затем становится самостоятельной характеристикой текста и не зависит от других индексов: индекс Дейла-Холла, индекс Флеша-Кинкейда и собственный индекс «Лексикатора».

Одной из самых распространённых метрик сложности текста для чтения является индекс Флеша-Кинкейда (Flesch, 1948; Kincaid et al., 1975). Он представляет собой линейную функцию среднего количества слогов в слове (что является лексическим параметром) и средней длины предложения в текста (что является синтаксическим параметром). Не смотря на небольшое количество используемых для оценки материала характеристик, данный индекс показывает высокую степень релевантности полученной по нему оценки и реальной сложности восприятия текста обучающимся.

Не менее эффективным и одновременно простым является второй по распространённости индекс - индекс Дейла-Холла (Chall and Dale, 1995). Он определяет сложность как линейную функцию как среднюю длину предложения в тексте (синтаксическая составляющая оценки, аналогична одной из составляющих индекса Флеша-Кинкейда) и процента редких слов (лексическая составляющая оценки). В случае работы с описанными выше уровнями изучения русского языка, за процент редких слов при расчёте принимаются все слова и их формы, не входящие в словари лексических минимумов соответствующих уровней. Таким образом, любое слово за

пределами словаря лексического минимум считается редким словом при подсчёте индекса Дейла-Холла.

Собственный индекс «Лексикатора» разработан на основе более широкого списка параметров и за счёт этого его уровень точности предсказания сложности для чтения выше, чем у перечисленных ранее простых индексов. Список параметров для извлечения из текста был составлен на основе работы [5] и затем адаптирован к задаче предсказания сложности текста на русском языке путём удаления и добавлении ряда параметров [1]. В результате итоговый список параметров выглядит следующим образом:

- Среднее количество слов в одном предложении текста;
- Средняя длина одного слова в предложении;
- Длина текста в буквах;
- Длина текста в словах;
- Средняя длина предложения в слогах;
- Средняя длина слова в слогах;
- Процент слов в 3 слога и больше;
- Процент слов в 4 слога и больше;
- Процент слов в 5 слога и больше;
- Процент слов в 6 слога и больше;
- Средняя длина предложения в буквах;
- Средняя длина слов в буквах;
- Процент слов длиной в 5 букв и больше;
- Процент слов длиной в 6 букв и больше;
- Процент слов длиной в 7 букв и больше;
- Процент слов длиной в 8 букв и больше;
- Процент слов длиной в 9 букв и больше;
- Процент слов длиной в 10 букв и больше;
- Процент слов длиной в 11 букв и больше;
- Процент слов длиной в 12 букв и больше;
- Процент слов длиной в 13 букв и больше;
- Процент слов в предложении, не входящих в словарь лексического минимума;
- Средняя длина предложения в словах.

Большинство из перечисленных выше параметров относительные (рассчитываются относительно длины предложения/текста). В дальнейшем эти параметры и параметры для индексов Дейла-Холла и Флеша-Кинкейда используются заранее обученной моделью на основе алгоритма машинного обучения.

В итоге, благодаря разработанной модели машинного обучения, на графическом интерфейсе ресурса сложность для чтения представляется не числом, которое может вызвать затруднение для его корректного толкования, а процентной шкалой соответствия обрабатываемого текста одному из допустимых уровней изучения языка. В данном случае это уровни A1, A2 и B1, так как дальнейшие уровни изучения подразумевают возможность понимания практически не адаптированных текстов и речи.

Пример отображения расширенного списка параметров в графическом пользовательском интерфейсе представлен на Рисунке 4.

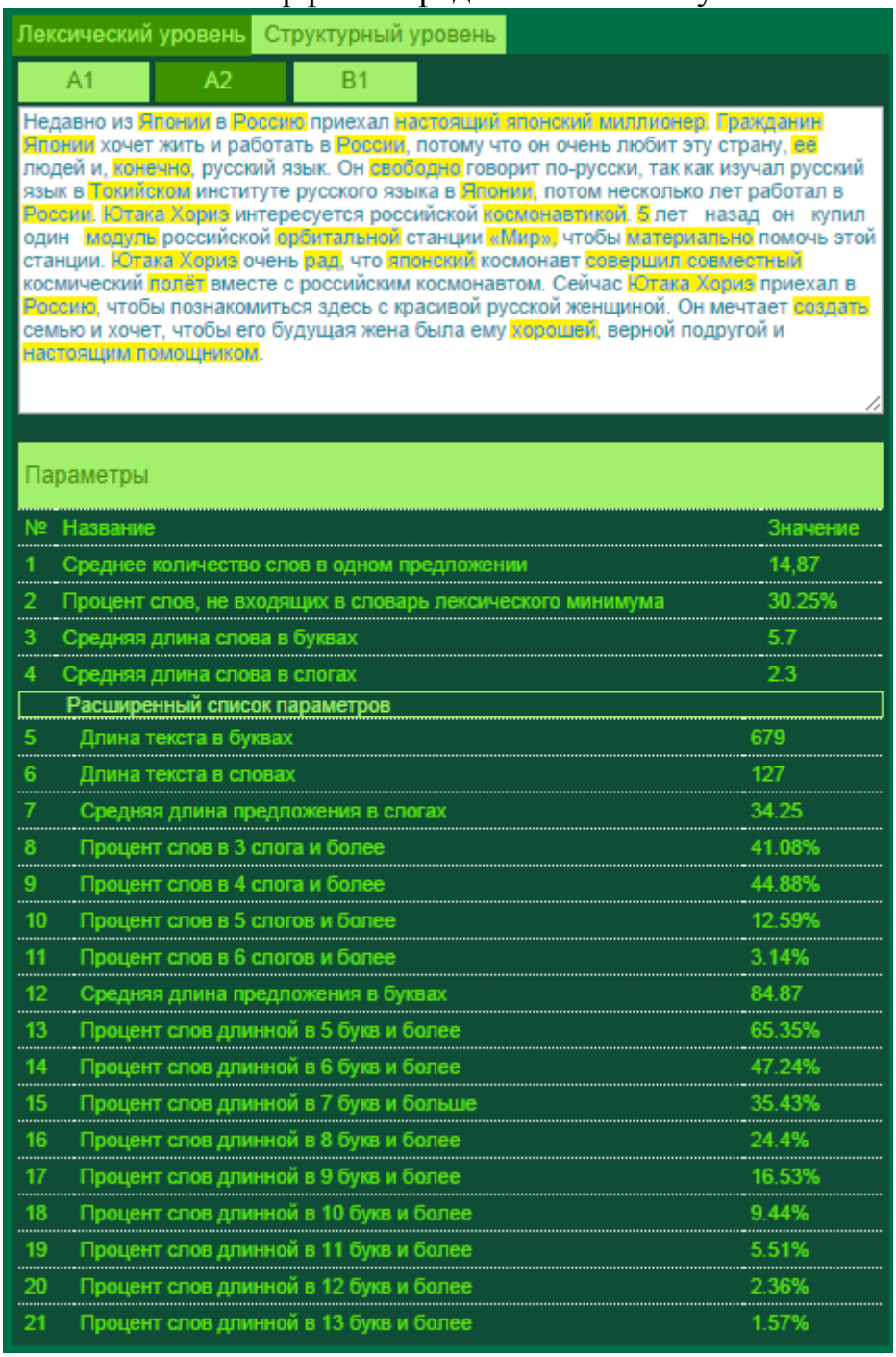

Рисунок 4. Пример отображения расширенного списка параметров

Таким образом, пользователь ресурса может однозначно определить, что тот или иной текст с большой степенью вероятности содержит синтаксические и лексические параметры, характерные для указанного «Лексикатором» уровня изучения языка. Стоит так же учесть, что результаты определения всех перечисленных индексов сложности для чтения доступны в интерфейсе и благодаря разным параметрам, используемым для вычисления, результаты сопоставления текстов ими различным уровням изучения языка могут быть различными. Тем не менее, сильный разброс для этих индексов так же не

характерен. Например, при получении наибольшего процента соответствия текста уровню А1 при использовании индекса «Лексикатора» и такого же результата с помощью индекса Дейла-Холла, вне зависимости от результатов индекса Флеша-Кинкейда можно утверждать, что текст наиболее характерен для уровня А1. Согласно описанному примеру, вычисление нескольких индексов по различным параметрам одновременно для одного и того же текста повышает вероятность корректного определения их уровня соответствия перечисленным уровням изучения языка.

В случае необходимости извлечения промежуточных результат вычисления – параметров исследуемого текста было обеспечено выведение вычисленных параметров и отображение их в пользовательском интерфейсе ресурса. Параметры текста могут быть использованы при проведении собственной оценки текста пользователем ресурса. Доступны краткий и расширенный список извлечённых параметров. В расширенный список входят все параметры, указанные выше при описании собственного индекса «Лексикатора». Краткий список параметров выглядит следующим образом:

- Среднее количество слов в одном предложении текста;
- Процент слов в предложении, не входящих в словарь лексического минимума;
- Средняя длина слова в буквах;
- Средняя длина слова в слогах.

В целом, описанный функционал позволяет получить достаточно объективную оценку исследуемого текста с учётом сложности синтаксических и лексических структур и возможностью просмотра в рамках дополнительной оценки извлечённых из текста основных характеристик пользователем ресурса.

Перспективы разработки

На данный момент ресурс «Лексикатор» обеспечивает следующие возможности:

- возможность обработки введённого пользователем текста с выделением сложных лексических и синтаксических структур на одном из трёх поддерживаемых уровнях изучения русского языка как иностранного (А1, А2, В1);
- получение объективной оценки сложности введённого текста для чтения на определённом уровне сложности по трём индексам;
- получение статистических характеристик обрабатываемого текста.

Одним из направлений дальнейшей разработки является увеличение набора индексов, по которым может оцениваться сложность текста для чтения с целью выявления того, какой из них даёт наилучшие результаты применительно к русскому языку. В случае приемлемых результатов работы таких индексов, полученные оценки сложности должны коррелировать с субъективными оценками учителей, использующих ресурс.

В рамках этой задачи, а так же в рамках создания базы учебных текстов, предполагается добавить возможность сохранения введённого текста по

соглашению с пользователем вместе с его соотнесением данного текста с одним из трёх уровней изучения языка. Ещё одним полем, запрашиваемым в контексте описанной выше задачи, может быть поле, описывающее тип текста (например, проза, публицистика, научные тексты и так далее), что так же может послужить двум целям одновременно: улучшению точности сопоставления сложности текста для чтения с определённым уровнем изучения языка, так как в рамках различных типов текста вес извлечённых из него параметров и сами эффективные параметры оценки могут быть различными, и возможность выбора текстовых фрагментов из общей базы текстов по заданному типу. Данную базу предполагается сделать открытой для общего использования, что облегчит поиски материалов для подготовки к занятиям с обучающимися.

Заключение

В рамках описанного выше исследования был разработан и предоставлен для общего использования в сети Интернет ресурс для студентов и преподавателей, позволяющий частично автоматизировать процессы подбора и адаптации актуальных для учащихся учебных материалов по русскому языку как иностранному. Это приводит к возможности существенного увеличения скорости работы над подготовкой учебных материалов для занятий и повышению качества этих материалов.

Разработанный ресурс получил название «Лексикатор» и имеет довольно простой и дружественный для пользователя с любым уровнем работы с компьютером интерфейс, описание которого приведено выше. В перспективе данный информационный ресурс может развить свой функционал до полноценной базы текстовых примеров для занятий с возможностью классификации по уровням обучения языку и выбором жанра текста.

В данной научной работе использованы результаты, полученные в ходе выполнения проекта «Адаптация языкового материала НКРЯ для электронного учебника "Русский язык как иностранный"», выполненного в рамках Программы «Научный фонд НИУ ВШЭ» в 2013-14 гг., грант № 13-05-0031.

Литература

1. Nikolay Karpov, Julia Baranova, Fedor Vitugin. Single-sentence Readability Prediction in Russian // Analysis of images, social networks, and texts, Yekaterinburg, 2014.
2. Андрюшина Н.П. Лексический минимум по русскому языку как иностранному. Элементарный уровень. Общее владение / Андрюшина Н.П., Козлова Т.В., ред. 4-е изд., испр. и доп. Санкт-Петербург, Златоуст, 2012, 80 с.
3. Андрюшина, Н.П. Лексический минимум по русскому языку как иностранному: Базовый уровень. Общее владение / Н.П. Андрюшина, Т.В. Козлова. – М.–СПб. : ЦМО МГУ – «Златоуст», 2000. – 116 с.
4. Андрюшина Н.П. Лексический минимум по русскому языку как иностранному. Первый сертификационный уровень. Общее владение /

Андрюшина Н.П. и др., ред. 5-е изд., испр. и доп. Санкт-Петербург, Златоуст, 2011, 200 с.
5. Владимирова Т.Е. Государственный стандарт по русскому языку как иностранному. Элементарный уровень / Владимирова Т.Е. и др. – 2-е изд., испр. и доп. – М. – СПб.: “Златоуст”, 2001. – 28 с.
6. Невдах М.М. 2008. Разработка метода автоматизированной оценки сложности учебных текстов для высшей школы, Международная научная конференция: "Теория вероятностей, случайные процессы, математическая статистика и приложения".