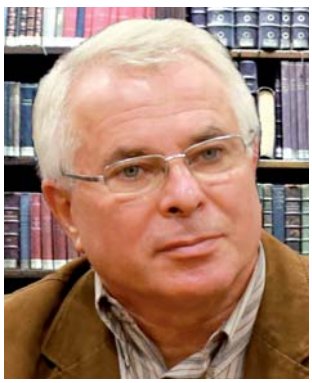
КОМПЬЮТЕРНЫЙ АНАЛИЗ ЛЕКСИКИ УЧЕБНЫХ ТЕКСТОВ РКИ



М.И. ШАПОВАЛОВ
mi.shapovalov@mpgu.su
канд. техн. наук,
д-р инж. наук (ФРГ), доцент
Московского педагогического
государственного университета
Москва, Россия

Ключевые слова:
корпусный подход, языковой корпус, учебное пособие, читаемость, лексический охват

DOI: 10.37632/PI.2020.282.5.010

В статье рассматривается процедура осуществления лингвистической оценки учебных пособий РКИ с позиций корпусного подхода. Получение и изучение количественных языковых данных позволяет составить представление о качестве материала.

Рост интереса к изучению РКИ требует обратить особое внимание на качество учебных материалов.

Чаще всего в процессе преподавания РКИ учителя комбинируют несколько учебников и пособий, а также свои наработки и тексты из Интернета. Естественно, выбор материала в значительной степени определяет результат. В настоящей работе мы рассмотрим качество материала только с одной стороны, а именно – проанализируем лексику и «читабельность» учебных текстов.

В основу Государственного стандарта по русскому языку положена серия лексических минимумов, соответствующая общеевропейской системе тестирования CERF. При включении слова в минимум составители руководствуются такими критериями, как частотность, стилистическая принадлежность, словообразовательный потенциал и т.п. К сожалению, решения часто основаны на опыте и интуиции, поэтому достаточно субъективны.

Уже само количество единиц в лексических минимумах различных уровней может вызвать вопросы. В таблице 1 представлены значения для лексического минимума ТРКИ и кембриджского CEF. Там же указано реальное количество уникальных слов, использованных в сертификационных тестах. Для шестого уровня (C2) лексический минимум не предусмотрен.

Таблица 1

Количество единиц для лексического минимума ТРКИ и кембриджского CEF

Уровень	Уровень РКИ	Шкала CERF Лексический минимум	Тест РКИ Количество уникальных слов	Кембриджский экзамен CEF Лексический минимум
1	A1	780	832	1500
2	A2	1300	1352	2000
3	B1	2300	2806	3000
4	B2	5000	2724	3500
5	C1	11000	3841	4100

Существуют методические соображения, предполагающие, что переход на каждый следующий уровень должен сопровождаться удвоением тезауруса. Это и дает число 11000 на пятом уровне.

В то же время созданные на материалах Национального корпуса русского языка [2] «Частотный словарь художественной литературы» содержит 5000 единиц, «Словарь значимой лексики художественной литературы» – 1000, «Частотный словарь живой устной речи» – менее 5000, «Словарь значимой лексики живой устной речи» – 780 единиц [1].

Обратим внимание на тот факт, что проверка реального количества уникальных слов, использованных в сертификационных тестах начиная

с четвертого уровня, дает значения, близкие к экзамену CEF, что существенно отличается от заявленных лексических минимумов шкалы CERF.

Мы ссылаемся именно на «Частотный словарь современного русского языка» под ред. О.Н. Ляшевской и С.А. Шарова, так как словари, составленные на основе лингвистических корпусов, наиболее востребованы благодаря своей репрезентативности и полноте. Кроме общего частотного и алфавитно-частотного списка лемм, в словаре содержится ряд показателей, весьма полезных для анализа текстов. Далее мы покажем, как их использование поможет получить достаточно интересные результаты.

Возвращаясь к лексическим перечням, отметим, что компьютерные технологии позволяют существенно облегчить их анализ для текстов любого размера. Что касается рекомендаций преподавателю относительно оптимального объема словника, представляется разумным использовать показатель покрытия (лексического охвата), распространенный в прикладной лингвистике. Упрощая, можно сказать, что это количество слов, которые совпали в тексте и словнике.

Если говорить об обучении русскому языку в специальных предметных областях, следует отметить, что использование корпусных инструментов позволяет проще решать задачи учебной лексикографии. Лексические списки можно автоматически извлекать из реальных текстов, относящихся к данной предметной области.

Именно здесь может помочь дополнительная информация из частотного словаря Национального корпуса русского языка. Дело в том, что попытка использовать значение частоты при создании словников для специальной предметной области приводила к тому, что на первом месте стояли предлоги. На самом деле, следует обратить внимание на показатель D – коэффициент Жуйана. Частотный словарь разбит на 100 сегментов, условно соответствующих функциональным стилям. Если слово используется в разных областях (например, «вода» – используется в бытовой речи, химии, физике, географии и т.д.), то значение коэффициента Жуйана близко к 100 (для слова «вода» – 96). У терминов значение коэффициента существенно меньше. Таким образом, алгоритм выделения терминов для формирования словника может выглядеть следующим образом:

1. Создать корпус текстов предметной области (это может быть набор учебников по дисциплине для студентов или специальные тексты для подготовки специалистов).

2. Провести лемматизацию корпуса (привести все слова к нормальной форме). Это делается с использованием интернет-сервисов типа Advego [5].

3. Отсортировать леммы по возрастанию коэффициента Жуйана, используя частотный словарь Национального корпуса русского языка.

На первом месте окажутся именно термины. Если слова нет в словаре, то, скорее всего, это термин или имя собственное. Если из текста нужно выбрать все слова общей лексики, то сортировать надо по убыванию значения коэффициента Жуйана. Вопрос о граничном значении коэффициента, определяющем, термин ли это, будет рассмотрен ниже на примере конкретного анализа учебного текста.

Теперь от размера словника перейдем к анализу частот. В различных субтестах количество незнакомых слов предполагается от 1,5 до 4%. Компьютерный анализ сертификационных тестов с использованием частотного словаря Национального корпуса русского языка помогает понять, насколько лексика теста отражает реальную частоту использования слов в русском языке. На данном этапе будем принимать во внимание только частоту, хотя в дальнейшем будем учитывать и другие показатели.

Анализ теста А2 (базовый уровень) показывает, что частота по Корпусу превышена более чем в 10 раз у 28% слов. У 5% более чем в 100 раз. Если для таких слов, как *знать, мочь, прослушать, говорение, синонимичный* (табл. 2), это объясняется значительным количеством комментариев методического характера, то для слов *машина, быль, пленник, строгать* (табл. 3) объяснятся, на наш взгляд, только недостаточным вниманием к лексическому составу учебных текстов.

Таблица 2

Превышение частоты по сравнению с НКРЯ

	Превышение частоты по сравнению с НКРЯ
знать	4069
мочь	1653
прослушать	378
говорение	210
синонимичный	270

Таблица 3

Превышение частоты по сравнению с НКРЯ

	Превышение частоты по сравнению с НКРЯ
машина	1050
быль	540
пленник	494

	Превышение частоты по сравнению с НКРЯ
строгать	4
минуть	427
наслать	420
экскурсовод	378
тесто	285
геолог	219
несовременный	189
кавказский	173

В то же время у 13% слов в тесте частота ниже, чем по корпусу. Совсем низкая частота у слов, представленных в таблице 3.

Таблица 4

глаз
то
иметь
надо
под
за
такой
дело
тот
для
вот
от

Это тем более странно, что обычно предлоги имеют максимальную частоту в тексте, из-за чего, как правило, исключаются из частотного анализа как стоп-слова.

В методических указаниях к Субтесту 2 «Чтение» в качестве характеристики презентируемого материала указывается: «Предъявляются тексты смешанного типа (повествовательного характера с элементами описания). Тематика текстов актуальна для социально-бытовой и социально-культурной сфер общения. Это тексты, максимально приближенные к аутентичным текстам страноведческого, информационно-публицистического и социально-бытового характера».

Нами была проведена проверка с использованием случайным образом сформированной выборки из Интернета текстов информационно-публицистического и социально-бытового характера с целью установить, насколько лексика теста соответствует лексике реальных текстов. Результат: 25% совпадений для текстов информационно-публицистического и 21% социально-бытового характера. Возможно, конечно, что на выборке большего размера цифры изменятся, но, тем не менее, такой результат заставляет задуматься о лексическом составе учебных текстов.

К сожалению, недостаточное внимание к лексике свойственно и некоторым учебным пособиям. Мы попытались оценить, достаточен ли лексический запас, получаемый студентом-иностранцем из учебного пособия по РКИ, для чтения хотя бы обычного школьного учебника.

Сразу следует сказать, что речь идет только об общей лексике. С помощью описанной ранее методики термины, встречающиеся в учебниках, выведены за рамки рассмотрения.

Для анализа использовались: учебное пособие «Русский как иностранный», учебники «Алгебра и начала математического анализа 10 класс», «Всеобщая история. XX век 11 класс», «География: Экономическая и социальная география мира 10–11 класс». Учитывая негативные результаты анализа, точное название учебника РКИ не указываем.

На первом этапе была поставлена задача – определить граничное значение коэффициента Жуйана, позволяющее отнести слово к термину. Работа была проведена на примере математических терминов. Мы использовали «Краткий словарь терминов в математике» Е. Половинкиной и С. Шакировой [4]. По частотному словарю Национального корпуса русского языка для каждого слова из словаря терминов в математике были определены значения частоты и коэффициента Жуйана.

Анализ показал, что слова из словаря терминов разбились на три группы: слова общей лексики, используемые в том числе в математике (фигура, норма, метр, сумма, линия, угол, пример). Слова из языка математики, используемые достаточно широко в общей лексике (радиус, вектор, квадрат, график, градус, сектор). И, наконец, математические термины, практически не используемые в разговорной речи (аксонометрия, асимптота, дискриминант, коллениарность, мантисса, радиан).

Мы проанализировали значение коэффициента Жуйана для всех слов из словаря терминов в математике. Слова с коэффициентом Жуйана выше 80 соотносятся с группой 1, от 21 до 80 соотносятся с группой 2, с коэффициентом меньше 21 – с группой 3.

Описанный алгоритм позволил нам в дальнейшем при анализе текстов всех учебников исключить из рассмотрения слова с коэффициентом Жуйана меньше 21 – они относятся к терминам, и работать только со словами общей лексики.

На втором этапе осуществлялся семантический анализ учебного пособия «Русский как

иностранный». В качестве инструмента использовался семантический онлайн-анализатор текста Advego.

Первое, на что следует обратить внимание: частота слов в пособии существенно отличается от соответствующих частот в Корпусе русского языка.

– Частота слов существенно выше, чем в частотном словаре, – 32,25%.

– Частота занижена в 2 раза – 17,87%.

– Частота занижена в 3 раза – 10,79%.

– Частота занижена более чем в 3 раза (до 15) – 39,09%.

Если считать, что частота появления слова в учебном пособии отражает «важность» слова, можно сделать вывод, что учебное пособие не ориентировано на частотный словарь современного русского языка.

Тексты школьных учебников также были обработаны семантическим анализатором Advego, и далее проверялось совпадение слов в учебном пособии и текстах учебников (лексический охват).

– Математика: совпало 364 слова, или 28,5%.

– География: совпало 1270 слов, или 14,9%.

– История: совпало 915 слов, или 18%.

Отсюда следует простой вывод, что студенты, учившиеся русскому языку по указанному пособию, способны понять максимум 30% слов общей лексики из школьного учебника.

В заключительной части исследования анализировались интегральные характеристики учебных текстов. Исследователи достаточно давно обратили внимание на тот факт, что разные тексты воспринимаются по-разному в зависимости от таких факторов, как средняя длина слова в тексте, количество слов длиной более 5 символов, средняя длина предложения, количество знаков препинания и т.д. Для оценки «читабельности», т.е. параметра, оценивающего, насколько текст сложен для восприятия, служат сервисы статистического анализа текстов.

Обычно применяются 5 формул:

– Формула Flesch-Kincaid,

– Индекс Колман-Лиау,

– Automatic Readability Index,

– SMOG,

– Формула Дэйла-Чейла.

Эти формулы, адаптированные для русского языка, используются сервисом Readability [3]. Индекс Dale-Chall, кроме того, анализирует, входят ли слова, встречающиеся в тексте, в 3000 наиболее частотных слов.

Учитывая, что помимо индекса читабельности сервис выдает такие дополнительные параметры, как возраст аудитории, есть смысл использовать подобные инструменты для предварительного оценивания качества учебных текстов (в данном случае мерой служит сложность текста – соответствует ли она возрасту ученика). Естественно, следует делать поправку на то, что ученики не являются носителями языка. Скорее всего, «поправочный коэффициент» может быть получен только опытным путем.

Нами были проверены тексты некоторых сертификационных тестов трех уровней. Результаты:

ТРКИ 1: Уровень читабельности: 4.47. Аудитория: 4–6-й класс (возраст – примерно 9–11 лет).

ТРКИ 3 (С1): Уровень читабельности: 14.42. Аудитория: I–III курсы вуза (возраст – примерно 17–19 лет).

С2: Уровень читабельности: 4.28. Аудитория: 4–6-й класс (возраст – примерно 9–11 лет).

Как мы видим, разброс весьма велик. А ведь уровень ТРКИ 1 (В1) считается достаточным для начала обучения в российских учебных заведениях, в том числе в университетах. Вызывает удивление читабельность текста для уровня С2 – уровня носителя языка. Все это еще раз заставляет задуматься о качестве подготовки учебных текстов и текстов, используемых при тестировании.

Заметим, что использованные нами методы анализа могут применяться и для организации компьютерной проверки результата Субтеста 5 «Говорение». Традиционно этот вид тестирования сложнее всего автоматизировать.

В последнее время появилось значительное количество систем на базе сервиса распознавания голоса Google. Обладая простым интерфейсом, они отлично подходят для неподготовленных пользователей. Например, такая система распознавания реализована в сервисе Google Диск.

Нами был проведен эксперимент по распознаванию короткого текста из сертификационного теста уровня А2. Для носителей языка распознавание составило 100%. При проверке в группе студентов (15 магистрантов первого курса из КНР) уровень распознавания не поднимался выше 70%. Для автоматизации проверки использовалась та же простая программа с использованием функций EXCEL, что и в случае анализа учебника РКИ. Следует отметить, что пороговый уровень распознавания (вернее, его градации, соответствующие разным оценкам) должен определяться исходя из опыта.

Мы вынесли за рамки настоящей работы исследование текстов РКИ с использованием таких характеристик, как плотность новых слов и индекс повторяемости. Эти показатели обычно используются преподавателями иностранных языков.

Показатель плотности новых слов демонстрирует, насколько равномерно распределены

новые слова по курсу (или по уровням). Исходя из методических соображений, нагрузка по уровням должна распределяться по возможности равномерно. Если преподаватель пользуется только одним учебником, можно не обращать на это внимания, но как только в курсе встречаются учебные материалы разных авторов, проблема становится серьезной. Проверка может быть легко проведена с использованием тех же самых инструментов, которые мы описывали выше. Следует обратить внимание, что на самом деле это все та же проблема размера словника на разных уровнях.

Что касается индекса повторяемости лексики – его значение тесно связано с предыдущей характеристикой текста и вычисляется так же легко. Обычно считают, что оптимальный индекс повторяемости лексики порядка 95%.

Завершая нашу работу, следует сказать, что проблема исследования качества текста (в том числе учебного) достаточно давно решается в рамках так называемой поисковой оптимизации (SEO-оптимизации). Это набор методов, алгоритмов и сервисов, позволяющих улучшить качество текстов на сайтах Интернета.

Существующие сервисы (в значительной части бесплатные) позволяют получить ответы на такие вопросы:

– насколько точно заголовок раскрывает суть текста страницы (содержит ли заголовок нужные ключевые слова);

– можно ли по первому абзацу догадаться о содержании (присутствуют ли ключевые слова в первом абзаце или первых 80 словах);

– насколько текст соответствует заданным ключевым словам;

– нет ли в тексте бессмысленных участков, которые можно убрать без потери качества материала (присутствуют ли в тексте ключевые слова, взятые из семантического ядра);

– можно ли в тексте быстро отыскать основные тезисы, выводы, решения (для этого следует проверить плотность ключевых слов в первом абзаце, в основной части и в заключении);

– нет ли в тексте слов и словосочетаний, плотность которых превышает 4% (такая ситуация носит название «переспам» и свидетельствует о низком качестве стилистики текста).

Существуют интернет-сервисы, позволяющие анализировать морфологические факторы текста, оценивающие долю разных частеречных классов в тексте, а также присутствие слов с определенной словообразовательной структурой (например, существительных с тем или иным суффиксом).

Использование этих сервисов позволяет существенно повысить качество учебных текстов и обучения в целом.

Для обеспечения высокого качества учебных материалов в преподавании РКИ необходима проверка учебных текстов (особенно самостоятельной разработки и взятых из Интернета) с использованием методов квантитативной и корпусной лингвистики.

В процессе анализа следует обращать особое внимание на лексический состав текстов и тестов.

Критерием, определяющим размер лексического минимума (особенно в специальных областях), является степень покрытия (лексический охват).

Использование в анализе таких показателей слов в частотном словаре Национального корпуса русского языка, как коэффициент Жуйана, позволяет организовать автоматическое выделение терминов в различных предметных областях, автоматизировать создание словника предметной области и существенно повысить качество анализа.

Предложенные автором алгоритмы дают возможность автоматизированной проверки результатов Субтеста 5 «Говорение» (даже в дистанционном режиме) и позволяют выявлять основные ошибки в произношении.

Дальнейшими направлениями анализа могут служить измерение плотности новых слов и определение индекса повторяемости лексики, что позволит организовать объективную оценку языкового содержания учебного курса. ■

ЛИТЕРАТУРА

1. Ляшевская О.Н., Шаров С.А. Новый частотный словарь русской лексики dict.ruslang.ru/freq.php.
2. Национальный корпус русского языка. ruscorpora.ru/.
3. Оценка читабельности текста. readability.io/.
4. Половинкина Е., Шакирова С. Словарь математических терминов mat-analiz.ru/index/0-79 .
5. Семантический анализ текста онлайн. advego.com/text/seo/.

References

1. Lyashevskaya O.N., Sharov S.A. Novyy chastotnyy slovar russkoy leksiki. dict.ruslang.ru/freq.php.
2. Natsionalnyy korpus russkogo yazyka.ruscorpora.ru/.
3. Otsenka chitabelnosti teksta. readability.io/.
4. Polovinkina E., Shakirova S. Slovar matematicheskikh terminov. mat-analiz.ru/index/0-79.

5. Semanticheskiy analiz teksta onlayn. advego.com/text/seo/.

M.I. Shapovalov
Moscow State pedagogical University
Moscow, Russia

COMPUTER ANALYSIS OF VOCABULARY OF TRAINING TEXTBOOK

Corpus approach, language corpus, training textbook, readability, lexical coverage.

The article looks at the framework of the linguistic evaluation of training textbook from the corpus approach standpoint. Acquiring and analyzing the quantitative language data allows to create a better picture of quality of the training textbook.

ПОЗДРАВЛЯЕМ ПОЗДРАВЛЯЕМ ПОЗДРАВЛЯЕМ

Когда думаешь о преподавателе, невольно вспоминаешь дисциплину, лекции по которой он читал. Когда я думаю о Борисе Ивановиче Фоминых, вспоминаю синтаксис. Почему этот преподаватель решил посвятить свою жизнь изучению данного раздела языкознания? Можно ли понять характер и душу профессора через преподаваемую им дисциплину? Чем живет человек, когда заканчиваются лекции и семинары? Как хочется знать больше о том, кто тебя учит! Желание знать о преподавателе – это и есть любовь.

Мне всегда хотелось узнать о Борисе Ивановиче больше, несмотря на некоторый страх перед экзаменом по его дисциплине. Синтаксис изучает процесс упорядочивания письменной речи. Наверное, поэтому Борис Иванович стремился упорядочить мышление своих студентов, подвести их к системе. А я не сразу это осознала и впитала. Но работа в школе помогла мне понять, что в Институте русского языка им. А.С. Пушкина меня учили правильно.

Благодаря Борису Ивановичу я научилась смотреть на языкознание как на «точную науку». До сих пор храню в электронном виде контрольные работы по синтаксису. Борис Иванович всегда подбирал удачные примеры предложений на разные виды синтаксических конструкций.

Скрупулезное отношение к работе и четкость – вот главные черты преподавательского стиля Бориса Ивановича. Эти черты он воспитывает в своих студентах. А его ум и знание дисциплины вызывают восхищение. Наверное, все, кто учился у него, в результате поняли, что ни одного слова филолог не должен произносить зря, исключительно ради красоты речи: в каждой фразе должны быть ясность и определенность, должна чувствоваться структура.

От всей души я поздравляю Бориса Ивановича с юбилеем и желаю ему крепкого здоровья, любви и поддержки со стороны близких и коллег, нескончаемого желания приносить пользу своим умом, знаниями и опытом.

Бутяйкина Татьяна, выпускница 2016 г.