***А.Н. Лапошина***

***Т.С. Веселовская***

***Л.Ю. Жильцова***

***О.Ф. Купрещенко***

***М.Ю. Лебедева***

**КОРПУСНОЕ УЧЕБНИКОВЕДЕНИЕ: В ПОИСКАХ ОБЪЕКТИВНЫХ КРИТЕРИЕВ ОЦЕНКИ УРОВНЯ УЧЕБНИКОВ ДЛЯ БИЛИНГВОВ[[1]](#footnote-2)**

**Аннотация.** В статье предлагается способ объективизации информации об уровне сложности учебников русского языка для детей 7–10 лет со вторым родным русским (РК2Р) с помощью количественного анализа данных корпуса учебников русского языка TIRTEC. Работа содержит сравнительный анализ лексических и синтаксических показателей учебных текстов для детей РК2Р с текстами для детей, изучающих русский как родной и как иностранный.

**Ключевые слова.** Корпус учебников, корпусное учебниковедение, русский язык как иностранный, русский как второй родной.

## **1. Введение**

Сбор репрезентативной коллекции языковых учебников и её последующий количественный анализ способен дать ответы на целый ряд исследовательских вопросов. В частности, такие коллекции могут служить для изучения лексического состава учебника или полноты описания определенных методических категорий [Gabrielatos 2006], поиска формальных показателей, указывающих на сложность учебного текста [Laposhina et al. 2018; Reynolds R. 2016], исследования культурной информации и стереотипов, транслирующихся посредством учебных материалов [Al Jumiah 2016, Sun & Kwon 2020].

С развитием технологий сбора и обработки текстовых данных всё чаще подобные исследования выполняются в русле корпусного учебниковедения (англ. corpus-based textbook analysis): создаются специальные корпусы учебников, разрабатывается методика педагогической разметки таких корпусов [Volodina 2014].

Задача классификации учебных пособий по уровню языковой сложности крайне важна при обучении языку, поскольку позволяет подобрать оптимальное пособие для учащегося исходя из его языковой подготовки, возраста, целей и пр. Одной из самых проблемных зон этой области методики преподавания РКИ можно назвать оценку уровня пособий для детей, для которых русский язык является не единственным языком повседневной коммуникации, из-за разнородности этой группы и отсутствия регламентирующих документов и стандартов для этой категории учащихся.

В настоящей работе в качестве одного из решений этой проблемы предлагается система оценивания языкового учебного материала по уровню сложности на примере анализа учебников по русскому языку для детей младшего школьного возраста с разным уровнем владения русским языком.

**2. Материалы и методы**

Для данного исследования была отобрана коллекция из 18 пособий по русскому языку из корпуса TIRTEC (Text-Image Russian Textbook Corpus). Корпус содержит тексты учебников русского языка для детей младшего школьного возраста с разным языковым опытом, от учебников русского как иностранного до учебников, использующихся в российской школе, снабженные педагогической разметкой, которая включает информацию об авторах и предполагаемой целевой аудитории учебника, типе текстового блока (например, *текст, упражнение, справочная информация)* и мн.др.  [Лапошина et al. 2019].

Поскольку в фокусе внимания этой статьи находится группа пособий для детей со вторым родным русским (РК2Р), мы отобрали 5 пособий для данной группы учащихся (далее в таблицах обозначаются индексом 2), 3 пособия по русскому как иностранному (индекс 1), линейку из учебников русского языка в русских школах за рубежом 1-3 классов (индекс 3), и линейку из учебников русского языка для российских школ 1-3 классов (индекс 4).

Все приведенные ниже расчеты производились с помощью программного кода на языке Python и морфологического анализатора pymystem3[[2]](#footnote-3).

**3. Анализ лексических показателей уровня сложности учебных материалов**

Для сравнительного анализа лексического состава выбранных пособий мы произвели расчет общего объема учебника в словах, количества уникальных лексем, представленных в учебнике, а также долю слов, входящих в лексические списки, которые могут быть релевантны в данном случае.

Одним из наиболее разработанных показателей доступности лексики текстов, предназначенных для изучающих язык как иностранный, является процент лексики из лексических минимумов [Karpov et al. 2014, Reynolds 2016]. Поскольку к настоящему моменту не существует специальных лексических списков для детей, изучающих русский язык, мы использовали в качестве ориентира Лексический минимум для взрослых, начинающих изучать  русский язык как иностранный (далее ЛМ А1)[[3]](#footnote-4), который содержит лексические единицы элементарного уровня – например, *город, рука, бутерброд*.

Ещё одним показателем доступности лексики учебных текстов традиционно считается частотность [Chen, Meurers 2016]. Для этих подсчетов мы выбрали список 5 тысяч самых частотных слов русского языка Нового частотного словаря русской лексики (далее – ЧС 5 000)[[4]](#footnote-5), однако стоит отметить, что данный список может быть не вполне релевантен для задачи оценки детской литературы - например, в первую тысячу частотных слов входят такие лексемы как *суд, советский* и пр.  Для получения информации о частотности слова по коллекции текстов, более близкой детской аудитории, мы использовали список 5 тысяч самых частотных слов на материале корпуса детской литературы Деткорпус[[5]](#footnote-6) (далее – Деткорпус 5 000). В расчетах использовались все фрагменты учебников (тексты, упражнения, справочная информация и пр.), за исключением текстов, обращенных к  учителю или родителю.

*Таблица 1.* **Лексический состав учебников русского языка для разных групп учащихся**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Индекс** | **Пособие** | **ЛМ А1** | **ЧС 5 000** | **Деткорпус** **5 000** | **Всего слов** | **Уникальных лексем** |
| 1 | Сорока[[6]](#footnote-7) | 83 | 92 | 95 | 5931 | 743 |
| 1 | Шаг за шагом | 80 | 90 | 89 | 5852 | 741 |
| 2 | Лена и Миша | 70 | 89 | 93 | 16645 | 1603 |
| 2 | Истоки | 67 | 87 | 93 | 22271 | 2631 |
| 2 | У костра | 62 |  84 | 88 | 12599 | 2600 |
| 3 | Дронов 3 | 60 | 83 | 88 | 14567 | 1930 |
| 3 | Дронов 2 | 58 | 82 | 88 | 13886 | 1975 |
| 3 | Дронов 1  | 57 | 81 | 88 | 13938 | 2093 |
| 2 | Совёнок | 57 | 78 | 87 | 6427 | 1042 |
| 1 | Говорю-пишу | 57 | 74 | 82 | 7299 | 1480 |
| 4 | Канакина 1 | 53 | 78 | 84 | 11861 | 1855 |
| 4 | Канакина 2  | 51 | 77 | 82 | 28304 | 3390 |
| 4 | Канакина 3 | 48 | 75 | 81 | 38334 | 4542 |
| 2 | Крокодил | 39 | 63 | 70 | 5771 | 1940 |

Таблица 1 содержит информацию о лексическом составе четырех групп анализируемых учебников.  Отсортированная по убыванию процента лексики из ЛМ А1, гипотетически она должна показывать плавный рост индекса группы учебников (от 1 – русский как иностранный к 4 – русский как родной) по мере уменьшения процента лексики из списка элементарного уровня. Однако в реальности мы наблюдаем более сложную картину. Некоторые пособия оказываются ожидаемо простыми по этому параметру (Сорока, Шаг за шагом), тогда как другие заметно сближаются с учебниками для российской школы (Говорю-пишу) или даже оказываются сложнее их («Крокодил»). Если линейка учебников для российских школ (Канакина) выстраивается в логичную цепочку уменьшения процента элементарной лексики от 1 к 3 классу, то линейка для зарубежных школ (Дронов) не столь последовательна: минимальные показатели демонстрируют учебники 1 и второго класса.

Несмотря на разный объем и состав лексических списков (объем ЛМ А1 составляет 900 лексем; лексика ЧС 5000 и Деткорпуса 5000 пересекается в 60% случаях), результаты подсчетов по спискам достаточно сильно коррелируют друг с другом: сортировка таблицы по убыванию процента лексики из двух других списков практически не меняет порядок учебников на получившейся шкале. Для большей точности подсчетов в будущих исследованиях будет выбран  какой-то один список или предложен скомпилированный из них единый список.

Информативным показателем целевой группы учебника может также стать общее количество слов и количество уникальной лексики учебника. Здесь на фоне стремительного роста объема лексики от 1 к 3 классу учебника Канакиной выделяется стабильность этого параметра по всей аналогичной линейке Дронова.

Отношение общего количества слов к количеству уникальной лексики иллюстрирует повторяемость, отработанность лексических единиц. По этому параметру нижние строчки  занимают пособия для РК2Р, Крокодил и У костра.

**4. Анализ синтаксических показателей**

Для того, чтобы проиллюстрировать возможные отличия пособий на синтаксическом уровне, приведем несколько характеристик,  традиционно используемых для задачи измерения сложности и связности текста [Graesser et al. 2014, Reynolds 2016]. Формула читабельности Флеша представлена в адаптированном варианте для русского языка [Оборнева 2006]. В расчетах использовались только законченные фрагменты текстов, исключались грамматические упражнения, списки лексики и пр.

*Таблица 2.* **Показатели синтаксической сложности учебников русского языка для разных групп учащихся**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Индекс** | **Пособие** | **Медианное значение формулы Оборневой** | **Средняя длина предложения** | **Среднее кол-во подчинит. союзов на предложение** |
| 1 | Сорока | 94 | 4 | 0.2 |
| 2 | Совёнок | 93 | 3.3 | 0.3 |
| 1 | Шаг за шагом | 91 | 3.8 | 0.3 |
| 1 | Говорю-пишу | 88 | 5 | 0.2 |
| 3 | Дронов 1 | 81 | 5.7 | 0.4 |
| 3 | Дронов 2 | 79 | 6.8 | 0.5 |
| 2 | У костра | 77 | 6 | 0.6 |
| 4 | Канакина 1 | 75 | 6.6 | 0.5 |
| 3 | Дронов 3 | 74 | 7.2 | 0.6 |
| 2 | Истоки | 74 | 9.7 | 1 |
| 2 | Лена и Миша | 74 | 9.1 | 1 |
| 4 | Канакина 3 | 72 | 7 | 0.6 |
| 4 | Канакина 2 | 71 | 7.3 | 0.7 |
| 2 | Крокодил | 59 | 11.2 | 0.8 |

Таблица 2 демонстрирует шкалу учебников, выстроенную по убыванию индекса читабельности Флеша, то есть по возрастанию уровня сложности. Несмотря на то, что все учебники, кроме пособия Крокодил попадают в категорию трактовки *65*–*100 баллов* – *очень легко читается* [Оборнева 2006],что достаточно закономерно для учебников для детей 7–10 лет, наглядным становится их распределение внутри этой группы. Верхние строчки занимают все 3 учебника для детей-иностранцев. Подтверждается замеченная ранее в Таблице 1 существенная разница в уровне сложности учебников 1 и 2 класса линейки для российских школ Канакиной. Выявлены учебники, по нескольким параметрам структурной сложности  превосходящие линейку для российских школ: Истоки, Лена и Миша. Пособие Крокодил, оказавшись самым сложным по данным обеих таблиц, требует более детального анализа и, вероятнее всего, корректировки методического описания предполагаемой аудитории учебника.

**5. Выводы**

Таким образом, одним из подходов к решению методической проблемы отсутствия единой системы маркировки и классификации учебников русского языка как второго родного (РК2Р) может стать сравнительный анализ коллекции текстов таких учебников и их сопоставления с двумя методическими «полюсами», между которыми они находятся: учебников русского как родного и русского как иностранного. Подобное сравнение позволяет отобрать лингвистические показатели, способные лечь в основу составления протокола оценивания и маркировки учебников для билингвов.

### **Литература:**

1. *Лапошина А.Н., Веселовская Т.С., Купрещенко О.Ф.* (2019),Иллюстративно-текстовый корпус учебников русского языка для детей младшего школьного возраста: концепция и методика создания. Труды международной конференции «Корпусная лингвистика – 2019», с. 63–71.
 2. *Оборнева И.В.* (2006). Автоматизированная оценка сложности учебных текстов на основе статистических параметров: дис... канд. пед. наук. М., 120 с.

3. *Al Jumiah A.* (2016), Language, Power, and Ideology in High School EFL Textbooks in Saudi Arabia. (Doctoral of Philosophy), University of New Mexico, Albuquerque, New Mexico.

4. *Chen X., Meurers D.* (2016). Characterizing text difficulty with word frequencies. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 84–94.

5. *Gabrielatos C.* (2006). Corpus-based evaluation of pedagogical materials: If-conditionals in ELT coursebooks and the BNC. Paper presented at 7th Teaching and Language Corpora Conference, Paris, France.

6. *Graesser A. C., McNamara D. S., Cai Z., Conley M., Li H., Pennebaker J.* (2014), Coh-Metrix measures text characteristics at multiple levels of language and discourse. The Elementary School Journal, 115(2), pp. 210-229.

7. *Karpov N., Baranova J., Vitugin F.* (2014), Single-sentence readability prediction in Russian. In Proceedings of Analysis of Images, Social Networks, and Texts conference (AIST), pp. 91–100.

8. *Reynolds R.* (2016), Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications, pp. 289–300.

9. *Sun W., Kwon J.* (2020), Representation of monoculturalism in Chinese and Korean heritage language textbooks for immigrant children. Language Culture and Curriculum, 33, pp. 402–416.

10. *Volodina E., Pilán I., Eide S. R., Heidarsson, H.* (2014), You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In Proceedings of the third workshop on NLP for computer-assisted language learning, pp. 128–144.

**References**

1. *Laposhina А.N., Veselovskaya Т.S., Kupreshchenko O.F.* (2019),Illyustrativno-tekstovyj korpus uchebnikov russkogo yazyka dlya detej mladshego shkol'nogo vozrasta: koncepciya i metodika sozdaniya [Text-image corpus of Russian language textbooks for primary school: concept and method of creation]. Trudy mezhdunarodnoj konferencii «Korpusnaya lingvistika – 2019» [Proceedings of the International Conference «Corpus Linguistics – 2019»]. St. Petersburg, 2019.

2. *Oborneva I.V.* (2006). Avtomatizirovannaya ocenka slozhnosti uchebnyh tekstov na osnove statisticheskih parametrov [Automatic assessment of the complexity of educational texts on the basis of statistical parameters]. Мoscow.

3. *Al Jumiah A.* (2016), Language, Power, and Ideology in High School EFL Textbooks in Saudi Arabia. (Doctoral of Philosophy), University of New Mexico, Albuquerque, New Mexico.

4. *Chen X., Meurers D.* (2016). Characterizing text difficulty with word frequencies. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 84–94.

5. *Gabrielatos C.* (2006). Corpus-based evaluation of pedagogical materials: If-conditionals in ELT coursebooks and the BNC. Paper presented at 7th Teaching and Language Corpora Conference, Paris, France.

6. *Graesser A. C., McNamara D. S., Cai Z., Conley M., Li H., Pennebaker J.* (2014), Coh-Metrix measures text characteristics at multiple levels of language and discourse. The Elementary School Journal, 115(2), pp. 210-229.

7. *Karpov N., Baranova J., Vitugin F.* (2014), Single-sentence readability prediction in Russian. In Proceedings of Analysis of Images, Social Networks, and Texts conference (AIST), pp. 91–100.

8. *Reynolds R.* (2016), Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications, pp. 289–300.

9. *Sun W., Kwon J.* (2020), Representation of monoculturalism in Chinese and Korean heritage language textbooks for immigrant children. Language Culture and Curriculum, 33, pp. 402–416.

10. *Volodina E., Pilán I., Eide S. R., Heidarsson, H.* (2014), You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In Proceedings of the third workshop on NLP for computer-assisted language learning, pp. 128–144.

**CORPUS-BASED ANALYSIS OF RUSSIAN TEXTBOOK FOR BILINGUAL CHILDREN: TOWARDS DATA-DRIVEN TEXTBOOK EVALUATION**

A.N. Laposhina, T.S. Veselovskaya, L.Yu. Zhiltsova, O.F. Kupreshshenko, M.Yu. Lebedeva

**Abstract.** The article is devoted to the discussion of the way to obtain objective information about the level of difficulty of Russian language textbooks for children 7–10 years old with second native Russian (RSNL) by means of quantitative analysis of TIRTEC Russian language textbook corpus data. This paper contains a comparative analysis of the lexical and syntactic indicators of the educational texts for RSNL children with the texts for children learning Russian as a native and as a foreign language. The results of this analysis allow us to determine the place of each textbook on the scale of progressive complexity of the language material and can be used to clarify information about the target audience of the textbook.

**Keywords.** Corpus of textbooks, textbook analysis, Russian as a foreign language, Russian as a second native language.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Лапошина Антонина Николаевна**

Государственный институт русского языка им. А.С. Пушкина (Россия).

**Laposhina Antonina**

Pushkin State Russian Language Institute (Russia).

***E-mail: antonina.laposhina@gmail.com***

**Веселовская Татьяна Сергеевна**

Государственный институт русского языка им. А.С. Пушкина (Россия).

**Veselovskaya Tatyana**

Pushkin State Russian Language Institute (Russia).

***E-mail: veselovskayats@gmail.com***

**Жильцова Людмила Юрьевна**

Государственный институт русского языка им. А.С. Пушкина (Россия).

**Zhiltsova Lyudmila**

Pushkin State Russian Language Institute (Russia).

***E-mail: jiltsova.ludmila@gmail.com***

**Купрещенко Ольга Федоровна**

Государственный институт русского языка им. А.С. Пушкина (Россия).

**Kupreshchenko Olga**

Pushkin State Russian Language Institute (Russia).

***E-mail: ofkupr@gmail.com***

**Лебедева Мария Юрьевна**

Государственный институт русского языка им. А.С. Пушкина (Россия).

**Lebedeva Maria**

Pushkin State Russian Language Institute (Russia).

***E-mail: m.u.lebedeva@gmail.com***

1. Работа выполнена при финансовой поддержке РФФИ, проект 17-29-09156. [↑](#footnote-ref-2)
2. https://github.com/nlpub/pymystem3 [↑](#footnote-ref-3)
3. Андрюшина Н.П., Козлова Т.В. Лексический минимум по русскому языку как иностранному. Элементарный уровень. СПб., 2012. [↑](#footnote-ref-4)
4. Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. [↑](#footnote-ref-5)
5. detcorpus.ru [↑](#footnote-ref-6)
6. В тексте работы приводятся сокращенные авторские обозначение учебников и пособий, цифра соответствует классу. Полный список и расшифровка указаний доступны по ссылке: https://digitalpushkin.tilda.ws/istochniki [↑](#footnote-ref-7)