УДК 37.09:811.111 (Организация обучения. Английский язык)

К АНАЛИЗУ ЛИНГВИСТИЧЕСКИХ ОШИБОК УЧАЩИХСЯ НА БАЗЕ УЧЕБНОГО КОРПУСА ТЕКСТОВ

© 2017 Е.П. Соснина
Соснина Екатерина Петровна, кандидат технических наук, доцент кафедры «Прикладная лингвистика», декан гуманитарного факультета. E-mail: ling@ulstu.ru

Ульяновский государственный технический университет. Ульяновск, Россия

Статья поступила в редакцию 25.10.2017

В статье рассматривается проект по разработке и использованию специального учебного корпуса, состоящего из сочинений российских младших школьников (8–10 лет), изучающих английский язык в школах с углубленным изучением иностранного языка города Ульяновска. Проект нашего корпуса насчитывает около 500 англоязычных письменных работ учащихся по наиболее распространенным темам стандартных учебных пособий, по которым учителя работают со своими учениками 2–3 классов. *Целью нашего проекта является* выявление и анализ лексических, орфографических и грамматических ошибок школьников младших классов. Для этого исследуются способы классификации лексических и грамматических ошибок, выбираются методики их кодирования и разметки в электронном корпусе, подсчета ошибок по типам, что позволяет провести всесторонний лингвистический анализ ошибок учащихся для лингводидактических целей. Разработанный метаязык использовался для обозначения ошибок, содержащихся в корпусе, и для получения статистических данных корпуса при помощи программ-конкордансов. Общее количество проанализированных и размеченных ошибок в корпусе более 3 500, из них более полутора тысяч орфографических и лексических. Орфографические ошибки составляют большинство, а именно 56 % от общего количества всех допущенных лексических ошибок, по сравнению с количественными показателями других ошибок. На базе компьютерного анализа корпуса также получен вывод о том, что из более двух тысяч грамматических ошибок чаще всего в письменных работах учащихся младших классов встречаются ошибки в употреблении артиклей и пунктуации. При этом распространенность частотных типов ошибок обоих классов в текстах сочинений школьников имеет, на наш взгляд, психолингвистические основания. Результаты проекта и проведенных исследований показывают перспективность корпусного подхода в лингводидактике, становлении общей теории ошибок, а также в коррекционной педагогике.
Ключевые слова: корпусная лингвистика, учебный корпус, английский как второй иностранный язык, прикладная лингвистика.

1. *Введение.* С развитием прикладных лингвистических технологий стали возможны более эффективные исследования на базе различного рода корпусов текстов, как например, разработка учебных корпусов текстов, ориентированных на дидактический формат для анализа языка детей и взрослых, изучающих иностранный язык как второй иностранный.

Создание таких корпусов, в которых отражались бы все особенности языка учащихся, необходимо для изучения проблем овладения иностранным языком и выработки коррекционной методики при дальнейшем преподавании. Кроме того, при изучении электронного корпуса со специальной разметкой ошибок, исследователи имеют доступ не только к ошибкам учащихся, но также ко всему языку учащихся данной возрастной или гендерной группы. База данных языка учащихся, будучи достаточно репрезентативной и полной, может стать очень полезным средством для всех, кто хочет выяснить, как изучается язык и как сделать процесс обучения более эффективным. Впрочем, для повышения ценности корпуса и для извлечения наиболее значимых лингвистических моделей, кроме разметки ошибок, также часто добавляются различные виды аннотаций: разметка частей речи, семантическая, дискурсивная разметка или грамматический разбор.

Направление, связанное с созданием учебных корпусов письменной или устной речи, сейчас продолжает интенсивно развиваться, взять хотя бы активные исследования и публикации признанного эксперта в этой области С. Гранже [1, 2], доступные также в ее личном проекте https://uclouvain.be/fr/repertoires/sylviane.granger , ее коллег и учеников [3, 4]. Также отметим множество реальных больших проектов в этой области корпусной лингвистики, такие как Cambridge Learner Corpus (Кембриджский учебный корпус), ICLE (Международный корпус для изучающих английский язык), LLC (учебный корпус Лонгман) [5]. Между тем в России эта область научного лин-

гвистического исследования пока развивается не так активно и исследований на базе учебных корпусов в российской педагогической науке, на наш взгляд, недостаточно.

2. Прикладные аспекты разработки учебного корпуса. 2.1. Разработка учебного корпуса. В УлГТУ на базе кафедры «Прикладная лингвистика» реализовано несколько проектов учебных корпусов, в том числе проект по созданию и анализу электронного учебного корпуса *письменных текстов начального уровня обучения иностранному языку* [6]. Подобного рода проект, но ориентированный на иные педагогические задачи и иную возрастную группу (старшие классы), разрабатывался в Санкт-Петербурге [7]. Проект нашего корпуса насчитывает около 500 англоязычных сочинений по наиболее распространенным темам, изучаемым учениками 2–3 классов в языковых школах города Ульяновска, например, «Моя семья», «Праздники», «Спорт», «Мое любимое животное», «Мой день», «Моя школа» и др.

Первоначальной целью создаваемого учебного корпуса являлось выявление и анализ лексических, орфографических и грамматических ошибок школьников младших классов. Для этого исследуются способы классификации лексических, орфографических и грамматических ошибок, выбирается методика их кодирования в корпусе, а также методика количественного автоматизированного подсчета ошибок по типам, что позволяет провести их всесторонний лингвистический и лингводидактический анализ.

Кроме того, для адекватной оценки и анализа грамматических ошибок школьников, изучающих английский язык на начальном этапе, необходимо четко представлять себе те разделы лексики и грамматики, которые входят в требования образовательного стандарта и подлежат обязательному изучению в начальной школе. Также следует учитывать учебные пособия для данного уровня обучения иностранным языкам, например, учебники английского языка И.Н. Верещагиной и Т.А. Притыкиной [8, 9].

На основе анализа способов кодировки и разметки ошибок, используемых в современных учебных корпусах, была выбрана простая и достаточно эффективная модель разметки, предложенная исследователем ланкастерского университета Юкио Тоно [10]. Данная модель была модифицирована с учетом специфики корпуса и целями исследования. Ошибочная конструкция помечается тегами <LE-код ошибки> </LE-код ошибки>, например <LE-1> – это открывающий тег для орфографической ошибки, а </LE-1> – ее закрывающий тег. Таким образом, можно легко автоматически определить, где начинается и где заканчивается ошибка. Разработанный метаязык использовался для обозначения ошибок, содержащихся в корпусе, и для получения статистических данных корпуса при помощи программ-конкордансов.

2.2. *Классификация и анализ лексических ошибок.* При начальном лингвистическом анализе письменных работ школьников выявляются различные типы ошибок, которые возникают вследствие расхождений в лексико-грамматическом строе родного и иностранного языков. Их можно условно разбить на классы (орфографические, лексические, грамматические и др.), а внутри классов выявить наиболее типичные случаи. Следует отметить, что общее количество проанализированных и размеченных ошибок в корпусе более 3 500, из них более полутора тысяч лексических.

Лексические ошибки в основном связаны с нарушением лексических речевых норм. Традиционно сюда относят употребление слов в ненормативных с точки зрения языковой системы значениях, нарушения лексической сочетаемости, повторы и тавтологию. Такие типы ошибок представлены в любой из классификаций [11, 12, 13, 14]. Например, по М.С. Соловейчик [14] лексические ошибки делятся на ряд подтипов: 1) употребление одного слова вместо другого (low – small); 2) плеоназмы (very many); 3) нарушение законов семантической сочетаемости слов; 4) словотворчество и другие.

Одна из интересных ошибок «словотворчества» в нашем корпусе закодирована как <*LE-7*>. Например,

I want to be a <*LE-7*> *pianist* </*LE-7*>.
My mother is <*GE-3*> <*LE-1*> *meneger* </*LE-1*> *of* <*LE-7*> *reklama* </*LE-7*> </*GE-3*>.

Орфографические ошибки заключаются в нарушении существующих в английском языке орфограмм. В нашем корпусе они обозначены кодами <LE-1> и <LE-2>. В корпусе выделяется несколько подтипов орфографических ошибок, например:

- ✓ пропущенная буква в слове (*panthrs – panthers*);
- ✓ неправильный порядок букв в слове (*hathc – hatch*);
- ✓ неправильное написание слова (*animols – animals, wene – when, doog – good*);
- ✓ недописанное слово (*ca – can*);

- ✓ лишняя буква в слове (*hedgehofg – hedgehog*);
- ✓ написание слова через дефис (*teeth-tusk*);
- ✓ упущенный дефис (*short haired*);
- ✓ неправильное написание названий и имен собственных (*Nikulaevih*).

Кодом <*LE-2*> мы обозначали довольно распространенный случай для данной возрастной группы, т.е. слова, которые написаны в сочинениях на английском языке русскими буквами, например:

His name is <LE-2> *Сережа* </LE-2>.
I live in <LE-2> *Ульяновск* </LE-2>.

Одним из характерных для данного возраста типов ошибок в частности является *ошибка в графике,* т.е. выборе заглавных и прописных букв. В нашем корпусе она закодирована как <LE-3>. Ее частотность показывает на достаточную распространенность данного типа ошибки. Например:

<GE-1> In </GE-1> <LE-3> *tuesday* </LE-3> and <LE-3> *thursday* </LE-3> I go to the swimming pool.
We <GE-6> are </GE-6> celebrate <LE-3> *new* </LE-3> <LE-3> *year* </LE-3>.
His name is <LE-3> *kesha* </LE-3>.

Орфографические ошибки составляют большинство, а именно 56 % от общего количества всех допущенных лексических ошибок, по сравнению с количественными показателями других ошибок (см. таб. 1, представляющую выборку по статистическим данным по ряду орфографических и лексических ошибок).

Таб. 1 Выборка по базовым типам наиболее распространенных лексических ошибок во 2–3 классах
(Basic types of the most widespread lexical mistakes of the 2nd–3d grade pupils)

Кодировка ошибок	Количественное соотношение	Процентное соотношение (%)
<LE-1>	849	56%
<LE-3>	358	23 %
<LE-7>	10	1 %

2.3. *Классификация и анализ грамматических ошибок.* Согласно учебным пособиям, используемым в школе, и психолингвистическому уровню детей на данном этапе обучения грамматика языка изучается в минимальном объеме, в игровой форме, а основной акцент делается на развитие коммуникативных навыков и активное пополнение словарного запаса. Предполагается, что полученные грамматические знания позволяют учащимся составлять небольшие письменные тексты, содержание которых соответствует изучаемой тематике. Тем не менее, школьники допускают большое количество ошибок в образовании форм слова, построении предложений, т.е. грамматические ошибки разнообразны и распространены, а их классификация достаточно широкая. В нашем проекте мы выделили базовые разновидности и разметили их тегами.

Грамматические ошибки связаны со случаями несоблюдения грамматических, а именно морфологических и синтаксических норм языка. Приведем лишь некоторые примеры наиболее распространенных ошибок и их частотность (см. таб. 2):

✓ *<GE-1> – Ошибка в употреблении предлогов, в т.ч. пропуск*

Her birthday is <GE-1> </GE-1> the twenty-first of December.
I saw an elephant <GE-1> *in* </GE-1> the Zoo.
<GE-13> In summer I swim <GE-1> *on* </GE-1> the sea <LE-1> usally </LE-1> </GE-13>.

✓ *<GE-2> и <GE-3> – Ошибки в употреблении артикля*

I like <GE-2> *the* </GE-2> English and <GE-2> *the* </GE-2> Russian.
My father is <GE-3> very clever man </GE-3>.

✓ *<GE-4> – Ошибка в использовании частей речи*

My mother and my father can <GE-4> *driver* </GE-4> but I cannot <GE-4> *driver* </GE-4>.

✓ *<GE-6> – Отсутствие сказуемого в предложении*

We<GE-6> </GE-6> from Russia.

✓ *<GE-9> – Ошибка в согласовании подлежащего и сказуемого*

I <GE-9> *lives* </GE-9> in Russia.

✓ *<GE-12> – Пунктуационная ошибка*

I don't have lunch at <GE-2> the </GE-2> school <GE-12> . </GE-12> Because I'm not hungry.
They have a son <GE-12> , </GE-12> <LE-1> Serg </LE-1>.

✓ *<GE-15> – Ошибка в построении структуры предложения*

<GE-15> My family <GE-6> </GE-6> five </GE-15>.
<GE-15> Him second year </GE-15>.

Таб. 2 Выборка по базовым типам наиболее распространенных грамматических ошибок во вторых, третьих классах (Basic types of the most widespread grammatical mistakes of the 2nd–3d grade pupils)

Код ошибки	Количество ошибок	Процент ошибок
<GE-1>	121	6%
<GE-2>	100	5%
<GE-3>	502	26%
<GE-6>	125	6%
<GE-9>	179	9%
<GE-12>	435	23%
<GE-15>	24	1%

На базе компьютерного анализа корпуса был получен вывод о том, что из более двух тысяч грамматических ошибок чаще всего в письменных работах учащихся младших классов встречаются ошибки в употреблении артикля и пунктуации. Причем, высокая частотность пунктуационных ошибок была для нас неожиданной. Кроме того, довольно большое количество ошибок школьники допускают при согласовании подлежащего и сказуемого, в предлогах и некоторых других грамматических явлениях, в частности, в структуре предложений.

3. Заключение. Корпус создавался с целью лингвистического анализа типовых лексических и орфографических ошибок в иноязычной речи школьников младшего возраста. На первом этапе разработки проекта были проанализированы типовые лексические и грамматические ошибки. Наиболее частотными словарными ошибками являются орфографические. Их появление в текстах сочинений имеет, на наш взгляд, психолингвистические основания.

На основе анализа грамматических ошибок можно сделать вывод, что объем «заученных» грамматических явлений и структур для данного уровня недостаточен и приводит к когнитивным сбоям при использовании языка. Отметим, что даже поверхностный анализ обоих классов ошибок, допускаемых детьми в иноязычной речи, подтверждает то, что основной причиной ошибок является действие законов аналогии и интерференции, т.е. стремление уподобить новое известному, построение иноязычных структур по моделям родного языка. Особенно это характерно для начального уровня обучения иностранному языку.

На наш взгляд, результаты исследований показывают перспективность корпусного подхода в лингводидактике, становлении общей «теории ошибок» [15], а также в коррекционной педагогике, так как, по мнению многих педагогов-практиков, ошибки в словах обладают «возрастной устойчивостью», т.е. одни и те же разновидности встречаются в речи учащихся и младших, и старших классов с незначительными изменениями. При этом, ошибки также обладают «качественной устойчивостью» (неправильная сочетаемость, графика, порядок слов, проблемы с артиклями и др.). Кроме того, следует помнить, что основы иностранного языка закладываются именно на начальном этапе обучения, и поэтому своевременный акцент и отработка проблемных моментов поможет при освоении более сложных языковых явлений.

1. Granger, Sylviane (2003): The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research. TESOL Quarterly 37:3, 538–545.
2. Granger, Sylviane. (2003) Error-tagged learner corpora and CALL: a promising synergy. In: CALICO Journal, Vol. 20, no. 3, p. 465–480.
3. Bestgen, Yves; Granger, Sylviane. Categorizing spelling errors to assess L2 writing In: International Journal of Continuing Engineering Education and Life-Long Learning, Vol. 21, no. 2/3, p. 235–252.
4. Granger, S., Gilquin, G., Meunier, F. (2015). The Cambridge Handbook of Learner Corpus Research (2015). P. 1–748.
5. Мальцева, М.С. Учебный корпус (learner corpus) как база для лингвистического и лингводидактического анализа в рамках методики преподавания иностранных языков // Социально-экономические явления и процессы. 2011. № 9 (31). С. 209–212.
6. Соснина, Е.П. Корпусный подход к классификации и анализу лингвистических ошибок учащихся // Актуальные задачи лингвистики, лингводидактики и межкультурной коммуникации: сб. науч. ст. Ульяновск, УлГТУ, 2006. С.32–36.

7. Камшилова, О.Н. Исследовательский потенциал корпуса английских текстов петербургских школьников: анализ интерязыка // Известия Российского государственного педагогического университета им. А.И.Герцена. 2009. № 104. С. 114–123
8. Верещагина, И.Н., Притыкина, Т.А. Английский язык: учеб. для II кл. шк. с углубл. изуч. англ. яз. М., Просвещение, 1991. 240 с.: ил.
9. Верещагина, И.Н., Притыкина, Т.А. Английский язык: учеб. для III кл. шк. с углубл. изуч. англ. яз. М., Просвещение, 1994. 352 с.: ил.
10. Tony McEnery, Richard Xiao, Yukio Tono. (2006) Corpus-based Language Studies: An Advanced Resource Book. Taylor & Francis, p. 386.
11. Фоменко, Ю.В. Типы речевых ошибок. Новосибирск, НГПУ, 1994. 60 с.
12. Цейтлин, С.Н. Речевые ошибки и их предупреждение. М., УРСС, 2013. 187 с.
13. Цейтлин, С.Н. Ошибки в письменной речи учащихся и способы их классификации // Русский язык в школе. 1984. № 2. С. 40–46.
14. Соловейчик, М.С. Нарушение языковых норм в письменной речи младших школьников // Начальная школа. 1979. №4. С. 19–23.
15. Новицкая, И. В., Вакалова, А. Е. Теория ошибки в свете различных подходов // Молодой ученый. 2016. №26. С. 788–794.

RESEARCH OF PUPILS' LINGUISTIC ERRORS ON THE BASE OF LEARNER CORPUS

© 2017 E.P. Sosnina

Ekaterina P. Sosnina, PhD in Computer Science, Assosiate professor of the Applied Linguistics Department, Dean of the Faculty of Humanities. E-mail: ling@ulstu.ru

Ulyanovsk State Technical University. Ulyanovsk, Russia

The paper presents an on-going research project to develop the specialized learner corpus that consists of essays written by young (8–10 years old) language learners at schools with a profound teaching of the foreign language in Ulyanovsk, Russia. The learner corpus contains about 500 works, produced on the most widespread topics of the standard educational materials used by English teachers of the 2nd and 3d grades at Russian schools. The main research areas of the project are the identification and analysis of lexical, spelling and grammatical errors in the essays. We have studied the methods of classification of vocabulary and grammar-related mistakes as well as techniques of their coding and tagging in the electronic corpus. We also provide the errors statistics, which allows us to carry out the comprehensive linguistic analysis of each mistake type for the applied-linguistic purposes. The developed meta-language was used to tag the mistakes found in the corpus and to obtain the statistical data with the help of the open concordance programs. The total number of the analyzed and tagged mistakes in the learner corpus is more than 3,500, among which we identify 1,500 spelling and lexical errors. Spelling errors make the majority and are 56% of the total of all the lexical mistakes in comparison with the quantitative indices of other mistakes. By the computer analysis of more than 2,000 grammatical errors in the learner corpus we found out that the most frequent mistakes in written works by the elementary graders are the wrong usage of articles and punctuation marks. In our opinion, the prevalence and frequency of the tagged error types of both classes in the essays of the focus group of schoolchildren have the certain psycholinguistic reasons. The analysis results are useful in applied linguistics and didactic domains, and prove the potential of using corpus techniques with ESL learners and teachers.
Keywords: Corpus Linguistics, Learner Corpus, English as the Second Language (ESL), Applied Linguistics.

1. Granger, Sylviane (2003): The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research. TESOL Quarterly 37:3, 538–545.
2. Granger, Sylviane. (2003) Error-tagged learner corpora and CALL: a promising synergy. In: *CALICO Journal*, Vol. 20, no. 3, p. 465–480.
3. Bestgen, Yves; Granger, Sylviane. Categorizing spelling errors to assess L2 writing In: *International Journal of Continuing Engineering Education and Life-Long Learning*, Vol. 21, no. 2/3, p. 235–252.
4. Granger, S., Gilquin, G., Meunier, F. (2015). The Cambridge Handbook of Learner Corpus Research (2015). P. 1–748.
5. Mal'tseva, M.S. Uchebnyi korpus (learner corpus) kak baza dlya lingvisticheskogo i lingvodidakticheskogo analiza v ramkakh metodiki prepodavaniya inostrannykh yazykov (Learner Corpus as the Basis for Linguistic and Linguo-Didactic Analysis in Foreign Languages Teaching Methodology). *Sotsial'no-ekonomicheskie yavleniya i protsessy.* 2011. № 9 (31). S. 209–212.
6. Sosnina, E.P. Korpusnyi podkhod k klassifikatsii i analizu lingvisticheskikh oshibok uchashchikhsya (Corpus-Based Approach to Classification and Analysis of Linguistic Mistakes Made by Students). *Aktual'nye zadachi lingvistiki, lingvodidaktiki i mezhkul'turnoi kommunikatsii*: sb. nauch. st. Ul'yanovsk, UlGTU, 2006. S.32–36.

7. Kamshilova, O.N. Issledovatel'skii potentsial korpusa angliiskikh tekstov peterburgskikh shkol'nikov: analiz interyazyka (Research Potential of a Learner Corpus of English Texts of St. Petersburg School Students: Interlanguage Analysis). *Izvestiya Rossiiskogo gosudarstvennogo pedagogicheskogo universiteta im. A.I. Gertsena*. 2009. № 104. S. 114–123
8. Vereshchagina, I.N., Pritykina, T.A. Angliiskii yazyk (The English language): ucheb. dlya II kl. shk. s uglubl. izuch. angl. yaz. M., Pro-sveshchenie, 1991. 240 s.: il.
9. Vereshchagina, I.N., Pritykina, T.A. Angliiskii yazyk (The English language): ucheb. dlya III kl. shk. s uglubl. izuch. angl. yaz. M., Prosveshchenie, 1994. 352 s.: il.
10. Tony McEnery, Richard Xiao, Yukio Tono. (2006) Corpus-based Language Studies: An Advanced Resource Book. Taylor & Francis, p. 386.
11. Fomenko, Yu.V. Tipy rechevykh oshibok (Types of Speech Mistakes). Novosibirsk, NGPU, 1994. 60 s.
12. Tseitlin, S.N. Rechevye oshibki i ikh preduprezhdenie (Speech Mistakes and Their Prevention). M., URSS, 2013. 187 s.
13. Tseitlin, S.N. Oshibki v pis'mennoi rechi uchashchikhsya i sposoby ikh klassifikatsii (Mistakes in the Written Speech of Pupils and Methods of Their Classification). *Russkii yazyk v shkole*. 1984. № 2. S. 40–46.
14. Soloveichik, M.S. Narushenie yazykovykh norm v pis'mennoi rechi mladshikh shkol'nikov (Violation of Language Norms in the Written Speech of Schoolchildren at Elementary School). *Nachal'naya shkola.* 1979. №4. S. 19–23.
15. Novitskaya, I.V., Vakalova, A.E. Teoriya oshibki v svete razlichnykh podkhodov (The Theory of a Mistake in the Light of Various Approaches). *Molodoi uchenyi.* 2016. №26. S. 788–794.