О. Н. Камшилова
O. N. Kamshilova

УЧЕБНЫЙ КОРПУС ТЕКСТОВ: РАБОТА НАД ОШИБКАМИ

LEARNER CORPORA: ERROR STUDY AND CORRECTION

Аннотация. В докладе обобщается опыт использования учебных корпусов текстов (LC) как «корпусов ошибок» и оцениваются новые перспективы применения LC-технологий. Описывается опыт корпусного анализа «перепроизводства» грамматических структур (на примере структуры SVC) в речи русских школьников, изучающих английский язык.

Abstract. The paper reviews the use of Learner Corpora as “error collections” and new prospects for LC technologies. It also focuses on a corpus analysis of SVC structure overuse by Russian EFL schoolchildren.

1. Учебные корпусы текстов: «учиться на ошибках»

Учебные корпусы текстов (Learner Corpora, далее LC) изначально создавались с целью мониторинга и анализа ошибок, допускаемых при овладении инофонами чужим языком. Корпусные технологии позволили обнаружить наиболее распространенные ошибки в словоупотреблении и словообразовании, характер которых заставил пересмотреть содержание многих обучающих материалов, поскольку выявленные отклонения от нормы свидетельствовали о влиянии интерференции родных языков инофонов и так называемой «промежуточной грамматики» или интеръязыка, от чего не предупреждает ни один традиционный учебник или учебный словарь. Таким образом, известная поговорка «на ошибках учатся» достаточно точно характеризует появление новых методик, подкрепленных материалом представительных эмпирических баз данных. На основе исследований в LC

создаются словари и учебники нового типа, включающие предупреждающие такие ошибки комментарии[1].

Почти 30-летний опыт работы с LC показал, что потенциал их далеко не исчерпан. Внедрение новых технологий и расширение сфер применения LC тоже, в некотором смысле, можно считать «работой над ошибками». Отметим основные направления в совершенствовании LC-технологий[2]:

- При сохранении основных принципов проектирования и строительства LC развитие технологий предлагает новые форматы и процедуры. Многие корпусы включают аудио и видео материалы, сканы и pdf-файлы оригинальных рукописей информантов, что требует разработки процедур их встраивания в корпус и инструментов дальнейшей обработки.
- Кроме обязательной морфологической и синтаксической разметки, возникает необходимость в просодической разметке аудио и видео материалов. Новые запросы создателей LC сегодня – семантическая и дискурсивная разметки.
- Специфическая для LC задача – разметка ошибок (error tagging). Комплекс проблем, которыми занимаются сегодня, связан с классификацией ошибок и стандартизацией их кодирования. Достаточно трудоемкой задачей считается и сам процесс обнаружения ошибок в тексте, который обычно проводится вручную.
- Создание LC начиналось с письменных текстов на английском языке. Сегодня они активно создаются на материале разных языков как иностранных – венгерского, румынского, финского, эстонского, чешского и других. Технологии LC используются также в исследовании языка детей-билингвов, родных языков «потенциальных» билингвов – школьников,

---

[1]Например, словарь *Cambridge Advanced Learner's Dictionary (2008),* УМК издательства Macmillan *Global (2012)* и др.

[2] См. материалы международной конференции LLLC2012, посвященной созданию и функционированию учебных корпусов в Оулу (Финляндия), проходившей в октябре 2012 года – URL: **http://www.oulu.fi/hutk/sutvi/oppijankieli/LLLC/LLLC2012_ abstracts.pdf**

живущих в зонах контактирующих языков, например, немецкого и итальянского на севере Италии (провинция Больцано)[3].

- Накопление корпусных данных и совершенствование разметки обусловливает обращение к статистическим процедурам обработки извлекаемой информации. Применение статистических методов позволяет строить гипотезы и доказательно обосновывать тенденции, характеризующие процесс освоения чужого / родного языка[4].
- Большинство известных LC фиксируют определенный этап языковой компетенции. Новое направление в LC – создание лонгитюдных корпусов, накопление текстов одного и того же автора (авторов) в течение некоторого времени, что позволяет представить процесс овладения языком в динамике.
- LC перестают быть только базой для извлечения ошибок, но становятся полезным и эффективным средством обучения. Примером последнего может быть обучающая система Т. Кобба, созданная на основе оригинального учебного корпуса[5].

2. «Невидимые» ошибки: перепроизводство грамматических структур

В «работе над ошибками» в LC можно выделить два аспекта, определяющие эффективность использования LC вообще – это извлечение и классификация ошибок. Работа с грамматическими ошибками, например, должна находить и систематизировать не только те, которые обнаруживаются по формальным (морфологическим) признакам, но и не имеющие таковых, что делает ошибки второго рода «невидимыми», нераспознаваемыми зачастую даже при анализе вручную. Они обнаруживаются

---

[3] См. *Dickinson, Ledbetter; Durst, Szabó; Gerstenberger; Ivaska; Kikerpill, Klaas-Lang, Praakli ; Štindlová, Rosen; Alderson; Glasnieks, Abel, Lyding* на **http://wwwoulu.fi/hutk/sutvi/oppijankieli/LLLC/LLLC2012_abstracts.pdf**

[4] См. *Min et al., Lee et al.* в этом сборнике

[5] см. разделы Tutorial и Teachers на **http://www.lextutor.ca/**

только при накоплении достаточно большого эмпирического материала. Речь идет о так называемом «перепроизводстве» (overuse) грамматических структур. Рассмотрим пример выявления таких ошибок на материале учебного корпуса SPbEFL LC[6].

Исследования показали, что иностранцы, говорящие / пишущие на английском языке, чаще других используют структуру SVC (с субъектным предикативом – *John is* ***a student/clever***) со связкой *be*[7]. Перепроизводство этой базовой структуры отмечается и в петербургском корпусе. В речи носителей языка она также высокочастотна, особенно в разговорной речи и академической прозе[8]. Показательно заполнение компонентов структуры в речи носителей: позиция субъектного предикатива С в разговорной речи преимущественно замещается существительным (более 50%), в отличие от академических текстов, где фиксируется абсолютное преимущество прилагательного[9].

Тексты петербургского корпуса относятся к разговорному регистру, но демонстрируют прямо противоположную тенденцию: в позиции С преобладает прилагательное (около 59%), доля существительного заметно ниже (около 36 %):

Таблица 1. Заполнение позиции предикатива (С)
в текстах корпуса SPbEFL LC

	Прил. (AP)	**Сущ. (NP)**	**Сл.соч. с предл.**	**Прид. предл.**	**Инфинитив (InfP)**

[6] SPbEFL LC (Корпус текстов петербургских школьников, изучающих английский язык) – URL: **www.spbeflcorp.ru**

[7] *Hinkel E.* Simplicity without elegance: Features of sentences in L1 and L2 Academic texts // TESOL Quarterly, Vol.37, No.2, 2003. P. 275–300*; Mauranen A.* The Corpus of English as Lingua franca in Academic Settings // TESOL Quarterly, Vol.37, No.3, 2003. P. 515–527.

[8] *Biber D., Johansson S., Leech G., Conrad S., Finegan E.* Longman Grammar of Spoken and Written English. Harlow: Longman, 1999. P. 437.

[9] там же, P.446.

			(PP)	(CC)	
S <BE> C (1325 hits)	779	475	18	40	13

Обращение к данным корпуса LSWE (Longman Spoken and Written English Corpus)[10] дает возможность сравнить наиболее частотные предикативные прилагательные и существительные в речи носителей и в текстах SPbEFL LC (табл. 2 и 3).

Только три прилагательных из петербургского корпуса попадают в первую дюжину высокочастотных предикативных прилагательных в текстах носителей языка (*good, different, difficult* – табл. 2). Причем высокая частота *different* и *difficult* в речи носителей фиксируется в академическом (формальном) регистре, но не в разговорном[11]. Поэтому употребление прилагательных формального стиля в речи школьников в этом контексте – явное нарушение нормы, которое в практике обучения языку не осознается как ошибка и не корректируется.

Таблица 2. Сравнение высокочастотных предикативных прилагательных с учетом регистра[12]

SPbEFL	**LWSEC**	**Conversation (per million)**	**Academic (per million)**
(1) good	(2) good	Over 200	Over 20
(4) different	(5) different	Over 40	Over 100
(12) difficult	(7) difficult		Over 100

[10] там же, P. 446

[11] Сравнение употреблений прилагательного *difficult* в текстах корпуса SPbEFL с примерами из BNC делает разницу в регистрах очевидной.

[12] В скобках указан ранг в первой дюжине предикативных прилагательных.

Таблица 3. Сравнение высокочастотных предикативных существительных

	SPbEFL	**LWSE (conversation)**
1	*proper name*	crap
2	idea	*proper name*
3	thing	home
4	friend	no way
5	problem	people
6	webster	matter
7	film	thing
8	place	time
9	child /children	trouble
10	dog	way

Возможно, что информанты корпуса испытывают влияние родного языка – прилагательное *трудный* имеет широкую сочетаемость, как в атрибутивной, так и в предикативной функции: *трудная задача / тема, трудный язык / спектакль, трудные жильцы; выборы / экзамены были трудными, учеба была трудной, расставаться было трудно*[13]. Перепроизводство (overuse) структур с *difficult* в этом случае можно рассматривать как явное свидетельство «своей грамматики», интерференции родного языка.

Сравнение частотных списков существительных в роли субъектного предикатива (табл. 3) обнаруживает больше совпадений: высокий ранг имен собственных, прономинального *thing*. Сопоставимы общереферентные употребления *child /children* и *people*, оценочные *problem* и *trouble*. Однако список *top*-существительных в LWSE носит ярко выраженный модальный («attitudinal, marking the stance of the speaker»)[14]

[13] Примеры из НКРЯ.

[14] *Biber D., Johansson S., Leech G., Conrad S., Finegan E.* Longman Grammar of Spoken and Written English. Harlow: Longman, 1999. P. 450.

характер, в то время как в списке SPbEFL находим имена с нейтральной оценкой (*friend, film, Webster, dog*). Оценочный компонент в текстах информантов корпуса SPbEFL чаще выражен прилагательным-определением к предикативному существительному (*a good / favourite / expensive / new film*). Таким образом, можно предположить, что авторы петербургского корпуса, в отличие от носителей языка, для выражения оценки предпочитают использовать предикативные и атрибутивные прилагательные.

Обнаруженное с помощью корпусного анализа «перепроизводство» структуры SVC в корпусе SPbEFL в количественном отношении не слишком противоречит нормам английского языка. Гораздо важнее выявленные «невидимые» ошибки, значительные отклонения от нормы, которые приводят к серьезным нарушениям в речевом регистре, что не корректируется в процессе обучения и не учитывается учебными материалами, которые должны быть ориентированы на предупреждение ложных решений, вызванных влияниями интерференции и интеръязыка. Определить эти ложные решения возможно с помощью обращения к данным учебных корпусов.

Совершенно очевидно, что в практике создания и использования LC последовательно меняются и корректируются задачи – от описательных и объяснительных к прикладным, а именно: от накопления и систематизации данных об ошибках, исследования системных характеристик интеръязыка и языковых процессов освоения чужого / родного языка к пониманию того, что данные LC могут стать основой обучающих систем и материалов. Однако первичная цель создания таких корпусов – накопление «отрицательного материала», выявление и анализ ошибок, позволяющие вести разносторонний мониторинг процессов овладения языком, остается по-прежнему актуальной.