О. Н. Камшилова, Т. А. Сосунова

КОРПУСНЫЙ АНАЛИЗ ГРАММАТИЧЕСКОГО СОСТАВА УЧЕБНЫХ ДИАЛОГОВ

Проекту SPbEFL LC[1], который создавался как «учебный корпус» (Learner Corpus), точнее, корпус текстов, порождаемых обучающимися иностранному языку, уже десять лет. Цель подобных проектов заключается в выявлении наиболее значимых ошибок, которые допускают обучающиеся в чужом языке, в особенности таких, которые не учитываются в дидактических материалах, в том числе ошибок национально-специфических, то есть связанных с влиянием родного языка обучающихся. На материале этого корпуса проведен целый ряд исследований, наиболее значимые результаты которых отражены в [1]. Однако до сих пор оставался мало исследованным подкорпус скриптов диалогов на английском языке, порожденных информантами (русскоязычными школьниками) в ответ на задание-стимул.

Предварительный анализ этих диалогов показал, что кроме очевидных ошибок (грамматика глагольных форм, порядок слов, артикли, предлоги, лексические ошибки и пр.) тексты диалогов обучающихся производят впечатление искусственности, «простоты без элегантности» [2]. Очевидно, что часть этого впечатления создается за счет многочисленных повторов слов и реплик партнера, заполненных пауз хезитации (*So...,Can you tell me..., And..., Well, yes...,Maybe...),* а также повторов формулировок задания, которые делают старательные попытки информантов варьировать микротемы внутри диалога нарочитыми и поверхностными. И все-таки за этими видимыми причинами неудач, на наш взгляд, кроются более веские основания, которые отличают диалогические тексты учащихся от обучающих образцов, тем более, от естественной диалогической речи носителей языка.

Более глубокое проникновение в причины неестественности учебных диалогов дает возможность корпусного анализа текстов сопоставимых корпусов. С этой целью был создан корпус диалогов-образцов из учебных пособий по английскому языку уровня Pre-Intermediate и Intermediate издательств Oxford University Press, Cambridge University Press, Express Publishing, Macmillan и Longman – Textbook Dialogue Corpus (TDC). Этот корпус сопоставим с корпусом диалогов SPbEFL LC, далее – Learner Dialogue Corpus (LDC) по объему (10 821 с/у и 11 561 с/у соответственно), по уровню владения языком информантов корпуса SPbEFL LC, учебным программам, темам диалогов и времени сбора материалов для корпуса.

Процедура анализа была подсказана работой Х. Ли и А.К. Фанга [3], в которой анализировался грамматический состав (grammatical composition) текстов диалогов матерей и детей – фрагмент известного корпуса CHILDES (*Child Language Data Exchange System*)[2] – в терминах частотных распре-

[1] SPbEFL LC – «Учебный корпус текстов петербургских школьников, изучающих английский язык» – URL: www.spbeflcorp.ru

[2] CHILDES – URL: http://childes.psy.cmu.edu

делений основных грамматических классов слов. Количественный анализ в корпусе основывался на сравнении числа словоупотреблений (токенов) и типов словоформ (типов токенов) основных грамматических классов слов в материнских и детских репликах, которые стали основой двух сопоставляемых корпусов текстов. Исследование показало, что согласно коэффициенту Пирсона, который показывает степень связи, корреляции сопоставляемых массивов текстов, материнские и детские тексты чрезвычайно близки, как по числу токенов разных грамматических классов слов (r =0,936), так и по типам токенов в этих грамматических классах (r=0,991). Однако распределение грамматических классов слов по сопоставляемым корпусам оказалось различным. Например, в текстах детей значительно преобладают существительные и местоимения, как наиболее референтно-прозрачные номинации, а употребление прилагательных, наречий, глаголов значительно уступает показателям материнских текстов. Напротив, в речи матерей наиболее представленным классом являются глаголы. Эти и другие наблюдения по корреляции инпута (речи матерей) и аутпута (речи детей) дали авторам возможность предположить, что проведенное исследование может оказаться полезным не только при исследовании процесса овладения ребенком родным языком, но и при обращении к изучению проблем овладения вторым или иностранным языком [3, с. 95].

На основе данной методики был проведен сравнительный анализ корпусов текстов диалогов-образцов («инпут») и текстов ученических диалогов («аутпут»). Анализ опирается на сравнение распределений существительных, глаголов, прилагательных, наречий, местоимений, союзов, предлогов и междометий (см. табл. 1 и 2) в текстах корпусов.

Таблица 1

Распределение грамматических классов слов
в корпусе диалогов-образцов (TDC)

Часть речи	Кол-во типов токенов (типов словоформ)	Кол-во токенов (словоупот-реблений)	Количественное соотношение токенов (токен / всего токенов %)	Количественное соотношение типов (тип / всего типов%)
Существительные	538	1353	12,5%	40%
Глаголы	302	2625	24,26%	22,47%
Прилагательные	133	425	3,93%	9,9%
Наречия	88	809	7,48%	6,55%
Местоимения	44	2179	20,14%	3,27%
Предлоги	26	851	7,86%	1,86%
Союзы	15	353	3,26%	1,12%
Междометия	32	531	4,59%	2,38%
Другое	*166*	*1695*	*15,66%*	*12,35%*
Всего	**1 344**	**10 821**	**100.00%**	**100.00%**

Таблица 2

Распределение грамматических классов слов
в корпусе диалогов учащихся (LDC)

Часть речи	Кол-во типов токенов (типов словоформ)	Кол-во токенов (словоупот-реблений)	Количественное соотношение токенов (токен / всего токенов %)	Количественное соотношение типов (тип / всего типов%)
Существительные	474	1823	15,77%	44,76%
Глаголы	201	2630	22,75%	18,98%
Прилагательные	145	572	4,95%	13,7%
Наречия	73	740	6,41%	6,89%
Местоимения	45	2227	19,26%	4,25%
Предлоги	23	852	7,37%	2,17%
Союзы	15	786	6,8%	1,42%
Междометия	35	644	5,57%	3,31%
Другое	*48*	*1287*	*11,13%*	*4,53%*
Всего	**1 059**	**11 561**	**100.00%**	**100.00%**

Значительная степень корреляции текстов двух корпусов подтверждается коэффициентом корреляции Пирсона[1]. Коэффициент корреляции токенов (словоупотреблений) основных грамматических классов слов в сопоставляемых корпусах оказался очень высоким: r=0,9462. При вычислении коэффициента корреляции типов токенов (типов словоформ) этих грамматических классов в обоих корпусах был также получен высокий результат: r=0,9639.

Однако, как и в сравнении материнских и детских текстов, основные различия обнаружились в распределении грамматических классов слов и их реальном лексическом наполнении.

В исследовании [3] существенным различием материнских и детских текстов (инпута и аутпута) оказалась асимметрия в распределении основных грамматических классов – существительных и глаголов: в детских текстах количество существительных значительно превышало количество глаголов, в то время как в материнских текстах б*о*льшая доля принадлежала глаголам. Это объясняется тем, что дети быстрее осваивают существительные (и местоимения), чем глаголы, прилагательные, наречия, семантика которых более абстрактна, в отличие от имен существительных и местоимений, которые в речи ребенка конкретно-референтны [3, с. 100].

Рисунок 1 в более явной форме демонстрирует, какие грамматические классы слов преобладают в диалогической речи информантов корпуса и в диалогах-образцах. В обоих корпусах количественно преобладают глаголы (как в речи матерей, т.е. взрослой речи), местоимения и существительные. Обращает на себя внимание количество союзных употреблений в текстах школьников.

[1] Вычисление коэффициента было осуществлено автоматически с помощью программного обеспечения IBM SSPS Statistics 22IBM SSPS Statistics 22 – URL: www-01.ibm.com/support/

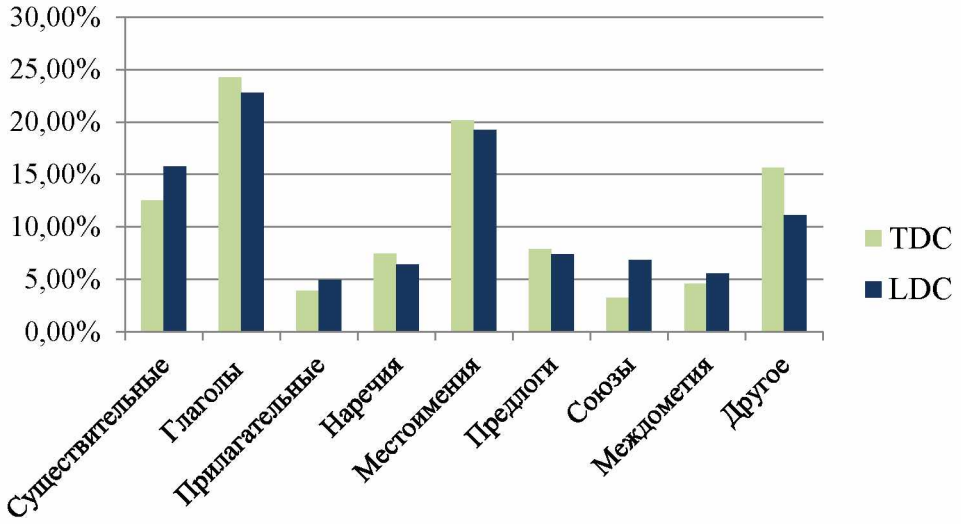

Рис. 1. Количественное соотношение грамматических классов слов (количество токенов) в обоих корпусах

Рисунок 2 дает представление о разнообразии репертуара глаголов, существительных и других грамматических классов:

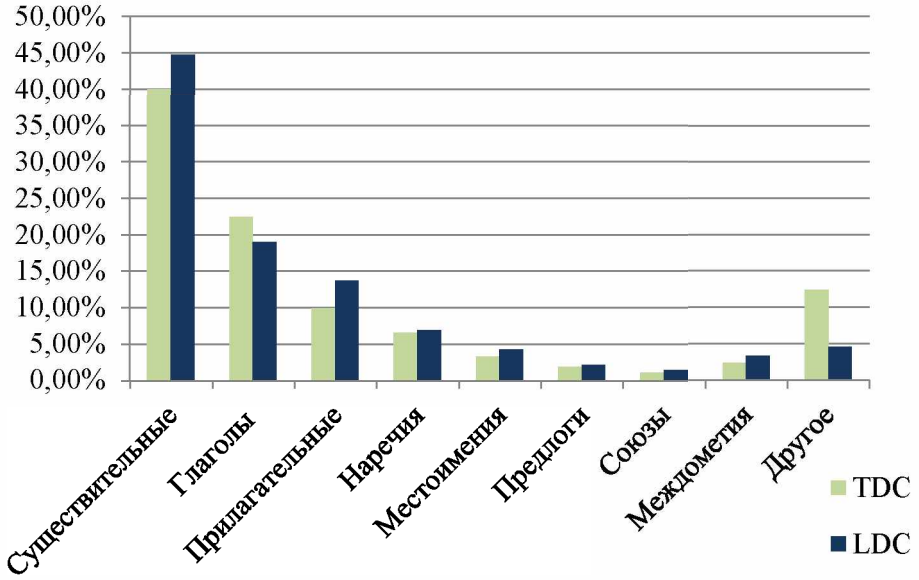

Рис. 2. Количественное соотношение грамматических классов слов (количество типов токенов) в обоих корпусах

Мы видим, что в обоих корпусах наиболее разнообразно представлены существительные, причем в ученических диалогах количество типов именных словоформ больше, чем в диалогах-образцах (как в речи детей), но их глагольный репертуар *уже* – свидетельство того, что обучающиеся избегают глагольных синонимов и сложных глагольных форм. Общее количество союзов в диалогах школьников почти в 2 раза (рис. 1) больше, чем в диалогах-образцах, но по репертуару они близки. Такой показатель – результат перепроизводства (overuse) в диалогической речи школьников союза *and*. Примечательно, что он

встречается в начале вопросительных предложений (*And what about/And do you/And what do you..*?), выполняя функцию слова-связки (linking word) и используется как некая «универсальная» форма вовлечения собеседника в разговор. Учащиеся заполняют им паузы хезита-ции, когда вспоминают или подбирают нужное слово (*So... and... I like very much that you are listening to rock music/ comedies and ...or horror.../Oh, it is a pity and... Oh, maybe we should go to my home. / But my friends go and...and I can't go with ...without them*).

Союз *and* эквивалентен русским союзам *и, а,* которые имеют высокую частоту в разговорной речи[1] и выполняют функцию связи предложений в высказывании. Поэтому частое употребление информантами союза *and* может быть вызвано интерференцией родного языка.

Эти и другие наблюдения, сделанные при сопоставлении распределений грамматических классов слов в диалогических корпусах текстов, позволяют заключить, что при очень высокой степени корреляции их грамматического состава тексты диалогов учащихся значительно отличаются от текстов диалогов-образцов из учебников репертуаром лексических средств, составляющих эти классы. Особо скудным представляется глагольный репертуар из-за повтора наиболее простых глаголов, которые усваиваются учащимися еще на самых ранних этапах обучения и часто выполняют в их речи роль неравнозначной замены, становясь основой конструкций внутренней, переходной грамматики и являясь тем самым интерференцией родного языка учащихся. Кроме того, глагольный репертуар школьников обеднен использованием простых глагольных форм и избеганием видо-временных форм. Наблюдается явный дефицит во владении так называемыми комментат*о*рными глаголами – модальными, фазовыми и грамматическими, которые составляют вершину глагольного списка текстов диалогов-образцов. Более разнообразен репертуар существительных, что, весьма вероятно, как и в речи детей, осваивающих язык, может быть следствием референтной прозрачности именных номинаций.

Данные проведенного корпусного исследования могут внести поправки в методики и дидактические материалы обучения диалогической речи.

ЛИТЕРАТУРА

1. *Камшилова, О. Н.* Учебный корпус: потенциал, состав, структура / О. Н. Камшилова – СПб. : ООО «Книжный дом», 2012. – 56 с.
2. *Hinkel, E.* Simplicity without elegance: Features of sentences in L1 and L2 Academic texts / E. Hinkel // TESOL Quarterly – Vol.37. – № 2. – 2003. – P. 275–301.
3. *Li.* Word frequency of the CHILDES corpus: Another perspective of child language features / Li, Hanhong, Fang, C. Alex // ICAME Journal. – № 35 – 2010. – P. 95–112.

[1] Устный подкорпус Национального корпуса русского языка состоит из 189 481 предложения, в которых союз *и* встречается в начале предложения 5 511 раз, союз *а* – 16 236 раз.