ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ
(Вычислительные системы)

2002 год Выпуск 171

УДК 519.769 : 801.314.4

ИСПОЛЬЗОВАНИЕ L-ГРАММНЫХ ХАРАКТЕРИСТИК ДЛЯ АНАЛИЗА ВАРИАТИВНОСТИ ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ [1]

В.Д. Гусев, Н.В. Саломатина

Введение

Данная работа является продолжением наших исследований по количественной оценке вариативности языковых единиц, принадлежащих разным иерархическим уровням. Ранее в качестве объектов исследования выступали корни [1] и канонические формы слов [2, 3], а также словосочетания, построенные на "игре слов" [4]. В этой работе мы анализируем тексты, являющиеся переводами одного и того же произведения (будем в дальнейшем называть их параллельными). Большой интерес к такого рода объектам наблюдается не только у лингвистов, но и у биологов [5], музыковедов [6], специалистов в области сжатия данных [7] и информационного поиска [8].

Как правильно отмечено в [9], детальное сравнение разных переводов одного и того же произведения — "дело трудоемкое и неблагодарное". Обычно такой анализ проводится на качественном уровне. *Целью данной работы* является *количественное исследование сходства и различия параллельных текстов*, приемов варьирования, используемых переводчиками при описании одной и той же ситуации, и индивидуальных особенностей их

[1] Работа выполнена при финансовой поддержке РФФИ, проект 00-06-80420.

стиля. Исходные тексты мы представляем в виде совокупности повторяющихся цепочек из L подряд следующих слов ("L-граммы на словах"), где значение L меняется от 1 до максимально возможного (L_{max}), определяемого наиболее длинным повтором в тексте. Ранее система представления подобного типа использовалась нами при анализе слитных, т.е. не содержащих разделителей, текстов, таких как ДНК-последовательности, знаменные песнопения и т.п., но цепочки формировались из символов, т.е. элементов более низкого иерархического уровня [10].

Для устранения вариативности на уровне словоизменительной парадигмы каждая словоформа заменялась своей основой, получаемой путем усечения окончаний и возвратных частиц. Реализация этой процедуры в автоматическом режиме не гарантирует 100%-ной точности, но она и не требуется для достижения нашей цели. Заметим также, что последовательное наращивание длин повторяющихся цепочек ($L = 1, 2, 3, \ldots$) эквивалентно расширению контекста, что позволяет устранять неоднозначности, возникающие при усечении отдельных словоформ, уже на уровне биграмм и триграмм.

Использование L-граммных характеристик в сочетании с позиционной информацией позволяет: 1) проводить сравнение параллельных текстов без предварительного их выравнивания (процедуры выравнивания достаточно трудоемки и хорошо работают лишь на близких текстах); 2) получать количественные оценки близости текстов; 3) фиксировать наиболее характерные отличия между текстами и классифицировать их как *случайные* (неизбежные при *независимом* переводе одного и того же текста разными людьми) или *систематические* (подразумевающие целенаправленную стратегию дистанцирования от имеющегося известного перевода).

§ 1. L-граммные характеристики текста

Рассматривая каждую словоформу (или ее основу в случае усечения) как отдельный символ из большого, но ограниченного алфавита, можно перенести на естественноязыковые (структурированные) тексты понятия частотной характеристики и частотного спектра, введенные нами в [10] применительно к слитным (неструктурированным) символьным последовательностям.

Пусть T — текст, N — длина текста (число содержащихся в нем словоформ). Назовем L-граммой цепочку из L подряд следующих словоформ. Полное число L-грамм, содержащихся в тексте, равно $N - L + 1$. Число различных L-грамм обозначим через M_L (очевидно, что $M_L \le N - L + 1$). *Частотная характеристика порядка* L текста T есть совокупность элементов $\Phi_L(T) = \{\varphi_{L_1}, \varphi_{L_2}, \ldots, \varphi_{L_i}, \ldots, \varphi_{L_{M_L}}\}$, где элемент φ_{L_i}, $1 \le i \le M_L$, есть пара (i-я L-грамма, F_{L_i} — частота ее встречаемости в тексте). В ряде случаев полезно указывать не только частоту встречаемости конкретной L-граммы, но и места ее вхождения в текст. Соответствующую характеристику будем называть *частотно-позиционной*. Часто L-граммы, однократно встречающиеся в тексте, не включаются в $\Phi_L(T)$. Такую характеристику будем называть *усеченной*.

Полным частотным спектром текста назовем совокупность частотных характеристик $\Phi(T) = \{\Phi_1(T), \Phi_2(T), \ldots, \Phi_{L_{max}}(T)\}$, где $L_{max} = L_{max}(T)$ — длина максимальной повторяющейся цепочки в тексте. Выбор L_{max} в качестве порогового значения в $\Phi(T)$ выглядит достаточно естественно, поскольку при $L > L_{max}$ частотные характеристики становятся неинформативными (все L-граммы имеют единичную частоту, а $M_L = N - L + 1$). На практике, начиная с относительно небольших значений L, в $\Phi(T)$ уже используют усеченные характеристики.

При наличии двух текстов T_1 и T_2 (как в нашем случае) удобно ввести понятие *совместной частотной характеристики* порядка L текстов T_1 и T_2:

$$\Phi_L(T_1, T_2) = \{\varphi_{L_1}(T_1, T_2), \varphi_{L_2}(T_1, T_2), \ldots, \varphi_{L_{M_L}}(T_1, T_2)\},$$

где $M_L = M_L(T_1, T_2)$ — количество различных L-грамм, общих для обоих текстов ($0 \le M_L(T_1, T_2) \le \min(M_L(T_1), M_L(T_2))$), а элемент $\varphi_{L_i}(T_1, T_2)$, $1 \le i \le M_L(T_1, T_2)$, есть тройка: (i-я общая L грамма, частота ее встречаемости в T_1 — $(F_{L_i}(T_1))$, частота ее встречаемости в T_2 — $(F_{L_i}(T_2))$). Соответственно *совместный частотный спектр двух текстов* можно определить как совокупность совместных частотных характеристик

$$\Phi(T_1, T_2) = \{\Phi_1(T_1, T_2), \Phi_2(T_1, T_2), \ldots, \Phi_{L_{max}(T_1,T_2)}(T_1, T_2)\},$$

где $L_{\max}(T_1, T_2)$ — длина максимального общего повтора у текстов T_1 и T_2.

Наряду с L-граммами из $\Phi(T_1, T_2)$, общими для обоих текстов, интерес представляют L-граммы, встретившиеся только в одном из текстов. Обозначим через $D_L(T_i)$ множество L-грамм (вместе с их частотами), присутствующих только в тексте T_i, $i = 1, 2$. В теоретико-множественных терминах $\Phi_L(T_1, T_2)$ можно трактовать как пересечение множеств $\Phi_L(T_1)$ и $\Phi_L(T_2)$, а $D_L(T_i)$ — как соответствующие *дополнения*. На практике целесообразно рассматривать лишь *усеченные дополнения*, содержащие L-граммы с частотой $F \geq 2$, поскольку только высокочастотные L-граммы из дополнений могут трактоваться как неслучайные, отражающие специфику конкретного текста.

На основе характеристик $\Phi_L(T_1)$, $\Phi_L(T_2)$ и $\Phi_L(T_1, T_2)$ можно вычислять различные теоретико-множественные меры близости (см. [10, 11]) между текстами T_1 и T_2 (например, отношение числа элементов в пересечении двух множеств к числу элементов в объединении и др.). В первом приближении о сходстве текстов можно судить уже по значениям параметров $M_L(T_1, T_2)$, $L = 1, 2, \ldots$ и $L_{\max}(T_1, T_2)$, хотя они не являются нормированными.

§2. Исходный материал

Рассматриваются два перевода на русский язык широко известной книги Алана А. Милна "Винни-Пух". Первый, ставший "каноническим", принадлежит Б. Заходеру (1960 г., изд-во "Детский мир"), второй — В. Веберу и Н. Рейн (1999 г., изд-во "ЭКСМО-пресс"). В последнем случае имело место "разделение обязанностей": В. Вебер переводил "прозу", а Н. Рейн — "стихи". Текст А. Милна содержит 49269 словоупотреблений, переводы Б. Заходера и В. Вебера с Н. Рейн, соответственно, 39806 и 43554 словоупотреблений.

Несколько меньший объем текста у Заходера (по сравнению с Вебером) объясняется тем, что его перевод фактически является пересказом, о чем он сам уведомляет читателя ("пересказал Борис Заходер"). Формально в нем на две главы меньше, чем в оригинале у Милна и в переводе Вебера, есть изменения даже в

порядке следования глав. Наиболее существенные отступления от оригинала наблюдаются в стихотворной части, где Заходер–переводчик сознательно уступает место Заходеру–поэту с его самобытным стилем, чувством ритма и остроумной "игрой слов" (см., например, его песенку "про со́рок пя́ток", которые трансформируются в "сороко́в пятко́в"). Заходер не всегда в точности следует оригиналу и в повествовательной части текста, но его "добавления", как правило, не затрагивают сюжетную линию и носят характер добродушно–иронических пояснений и комментариев. Вот, к примеру, как переведен Заходером и Вебером один и тот же фрагмент из оригинала:

Милн: "What about a mouthful of something?" Pooh always like a little something at eleven o'clock in the morning, and he was very glad to see Rabbit getting out the plates and mugs.

Вебер: "Как насчет того, чтобы перекусить?" Винни–Пух всегда любил перекусить в одиннадцать часов утра и жутко обрадовался, увидев как Кролик достает тарелки и миски.

Заходер: "А не пора ли чем–нибудь подкрепиться?" Винни–Пух был всегда не прочь немного подкрепиться, в особенности часов в одиннадцать утра, *потому что в это время завтрак уже кончился, а обед еще и не думал начинаться*. И, конечно, он страшно обрадовался, увидев, что Кролик достает чашки и тарелки.

Здесь выделенная курсивом вставка Заходера очень удачно поясняет, что выбор момента времени (одиннадцать часов) отнюдь не случаен.

Перевод В. Вебера и Н. Рейн формально более полон (содержит две пропущенные Заходером главы) и ближе к оригиналу, но и они, провозгласив принцип "невмешательства" (В. Вебер: "...ничего не выдумывать, ничего не привносить своего" — см. [9]), далеко не последовательны в его реализации, о чем свидетельствует следующий небольшой пример, описывающий ситуацию, когда объевшийся медом Винни–Пух застревает на выходе из норы угощавшего его Кролика:

Милн: "Hallo, are you stuck?" — he asked.

Заходер: "Ты что — застрял?" — спросил он.

Вебер: "Привет, ты что ли застрял?" — *с любопытством* спросил он.

Здесь привнесенная Вебером вставка ("с любопытством") предполагает как бы "отстраненно-созерцательное" отношение Кролика к произошедшему событию, тогда как из последующего текста видно, что он сильно огорчен как обжорством медведя-жонка, так и тем, что ему самому в дальнейшем, возможно, не удастся воспользоваться своей "парадной дверью".

Интересные соображения по поводу того, с какой целью делаются разные переводы одного и того же произведения вообще и Винни-Пуха в частности, приведены в уже цитировавшейся выше статье А. Борисенко [9]. Одно из соображений сформулировано следующим образом: "Вебер изо всех сил старается сделать "не так, как у Заходера", что связывает ему руки во многих отношениях". Это подразумевает, что Вебер целенаправленно варьирует текст Заходера, даже там, где он близок к оригиналу, но не все поддается варьированию без ущерба качеству. Забегая вперед, отметим, что полученные нами количественные оценки различия двух текстов хорошо согласуются с этим выводом.

§3. Процедура выделения основы слова

Учет словоизменительной парадигмы, как уже отмечалось выше, осуществлялся путем замены словоформ, представленных в тексте, их основами. Под основой слова понималась та его часть, которая остается после удаления окончания.

Схема выделения основы слова включает следующие шаги.

1. По имеющемуся списку из 75 окончаний (12 однобуквенных, 46 двухбуквенных, 17 трехбуквенных) в анализируемом слове определяется сначала наиболее длинное окончание. Выделенная основа проверяется (верифицируется) по словарю канонических форм, где каждая каноническая форма разбита на морфемы (использовался словарь Д. Уорта объемом свыше 100 тыс. канонических форм). Если основа представлена в словаре, процедура считается успешно завершенной.

2. Если основа не найдена (отказ), что может произойти по разным причинам, делается ряд дополнительных проверок с учетом наиболее вероятных причин отказа:

а) проверка на наличие неучтенного *чередования гласных и согласных* осуществляется путем варьирования выделенной основы в соответствии с заданными правилами ("кресло — кресел", "брать — беру", "печь — пеку",...). Преобразованная основа ищется в словаре канонических форм и, если поиск оказывается успешным, процедура выделения основы считается завершенной;

б) проверка на наличие *супплетивных форм*, примерами которых могут служить местоимения в косвенных падежах ("я", "меня", "мне"), некоторые глаголы ("ехать", "принять") и т.п., осуществляется путем поиска в словаре исключений, который формируется заранее. Если основа в словаре не найдена, переходим к следующему шагу.

3. Предполагая, что основа могла быть не найдена из-за неправильно выделенного окончания, что возможно из-за омонимии самих окончаний ("ами — и", "ому — у"), а также совпадения окончания с частью основы ("ет" в слове "привет"), уменьшаем на единицу длину окончания, выделяем новый вариант основы и повторяем процедуру поиска основы в словаре (шаги 1, 2).

4. Если и после выполнения шагов 1, 2, 3 основа не найдена, считаем, что она совпадает со всей словоформой. По этой схеме, в частности, завершается обработка имен собственных и других слов, которых нет в словаре. Например, в именах героев ("Кристофер Робин", "Винни-Пух"), когда они представлены в косвенных падежах, основа отождествляется со всей формой ("Кристофера Робина", "Кристоферу Робину" и т.п.).

Описанный способ выделения основы является приближенным, так как канонической форме в словаре не поставлен в соответствие вектор окончаний всей ее парадигмы. Поэтому разрешить омонимию окончаний и окончания с частью основы невозможно. Выборочная проверка процедуры на относительно небольших фрагментах текста дает оценку погрешности на уровне 15 % от общего числа рассмотренных словоформ. При этом примерно 46 % ошибок обусловлено омонимией окончаний (или основы и окончания), 31 % — недоучетом супплетивных форм, 21 % — отсутствием слов в словаре (имена собственные и т.п.), 2 % — омонимией словоформ. Для наших целей такая ошибка

является вполне допустимой, особенно если учесть, что в характеристиках 2-го и более высокого порядка наличие контекста уже помогает разрешить омонимию.

§4. L-граммные характеристики оригинала и двух его переводов

Выделение L-грамм из текста проводилось с помощью скользящего окна, охватывающего L подряд следующих слов и сдвигающегося вдоль текста с шагом в одно слово. Знаки препинания игнорировались. Основы слов, образующих L-связную цепочку, отделялись друг от друга косой чертой. В табл.1 приведены интегральные характеристики полных частотных спектров L-грамм исходного текста (Мяли, английский вариант) и двух его переводов на русский (Заходер и Вебер). Отметим, что в английском варианте процедура выделения основы не использовалась. Интегральные характеристики содержат информацию о разнообразии цепочек длины L ($L = 1, 2, \ldots, L_{\max}$), наличии среди них повторяющихся L-грамм и о максимальных частотах повторений при каждом L.

В табл.2 для разных L приведены примеры наиболее частых L-грамм, встретившихся в каждом из текстов. Номер L-граммы при каждом значении L совпадает с ее рангом — местом в упорядочении (чем больше ранг, тем меньше частота встречаемости F соответствующей L-граммы). Существенные различия в значении рангов одной и той же L-граммы в двух упорядочениях обычно не являются случайными и требуют интерпретации.

Обратим внимание на следующие особенности данных, представленных в табл.1 и 2.

1. Различие в длинах текстов Заходера ($N = 39806$ словоупотреблений) и Вебера ($N = 43554$) объясняется, в основном, тем, что перевод Вебера формально более полный и в нем содержатся фрагменты оригинала, выпущенные Заходером. Различия в длинах оригинала ($N = 49269$) и перевода Вебера, по-видимому, объясняются лексико-грамматическими особенностями двух языков. Число словоформ в английском тексте формально увеличивается за счет использования артиклей ("a",

Таблица 1

Количественные данные о разнообразии связанных цепочек из L слов, наличии среди них повторений и максимальной кратности повторений в текстах Милна, Заходера и Вебера

N	Вебер 43560			Заходер 49806			Милн 49269		
L	Кол-во разных цепочек	Кол-во кратных цепочек	Макс-я частота цепочки	Кол-во разных цепочек	Кол-во кратных цепочек	Макс-я частота цепочки	Кол-во разных цепочек	Кол-во кратных цепочек	Макс-я частота цепочки
1	4797	2500	1779	4230	2211	1711	2720	1672	2204
2	28141	5429	260	24661	4921	234	21470	6471	303
3	40445	2028	26	36375	2090	58	39583	4809	78
4	42689	560	6	39017	644	7	46365	2001	17
5	43346	190	4	39568	222	6	48203	820	10
6	43460	89	4	39693	119	3	48734	437	9
7	43501	52	3	39743	78	3	48953	268	4
8	43519	34	2	39765	58	3	49048	188	4
9	43527	25	2	39777	46	3	49100	150	4
10	43531	20	2	39786	37	2	49127	126	3
11	43534	16	2	39791	31	2	49147	107	3
12	43536	13	2	39795	26	2	49163	92	3
13	43537	11	2	39798	22	2	49176	79	3
14	43538	9	2	39800	19	2	49188	67	3
15	43539	7	2	39802	16	2	49198	57	2
16	43540	5	2	39803	14	2	49205	49	2
17	43541	3	2	39804	12	2	49211	42	2
18	43541	2	2	39805	10	2	49215	37	2
19	43541	1	2	39806	8	2	49219	32	2
	$L_{max} = 19$			$L_{max} = 25$			$L_{max} = 39$		

"the"), прининфинитивной частицы "to", а также вспомогательных глаголов в глагольных конструкциях. Как видно из табл.2 ($L = 1$), все они относятся к числу высокочастотных словоформ, хотя иногда частота оказывается завышенной из-за омонимии (как в случае с "to").

2. Несколько удивляет существенное различие в объемах словарей 1-грамм в оригинале ($M_1 = 2720$) и обоих переводах (Заходер: $M_1 = 4227$; Вебер: $M_1 = 4803$). Возможное объяснение состоит в том, что Милн, ориентируясь на детскую аудиторию, сознательно ограничивал свой словарь и увеличивал уровень повторности в тексте (уже начиная с $L = 2$, число повторяющихся L-грамм в английском тексте значимо ("в разы") превышает аналогичный показатель для русских текстов).

3. Максимальное разнообразие повторяющихся L-грамм для всех текстов наблюдается при $L = 2$, но многие повторяющиеся комбинации из двух слов не являются синтаксически и семантически завершенными конструкциями (например, "in the", "Пух и", "*и* не").

4. Состав наиболее высокочастотных 1-грамм определяется, в основном, тремя факторами. *Первый* отражает специфические *лексико-грамматические особенности конкретного языка* (наличие артиклей, прининфинитивных частиц, способы образования глагольных форм, способы реализации отрицания (частица "не" в русских текстах входит в первую пятерку по частоте встречаемости) и т.п.). *Второй фактор* отражает наличие *общеязыковых инвариантов*, представленных "служебной лексикой" (различные союзы ("and", "but", "so", "и", "а", "но"), местоимения ("he", "I", "you", "they", "it", "он", "я", "ты", "они"), предлоги ("in", "for", "on", "at", "в", "на", "с") и т.п.). *Третий фактор* непосредственно отражает *тематику произведения*, в данном случае состав действующих лиц ("Pooh", "Пух"; "Piglet", "Пятачок", "Хрюка"; "Christopher", "Robin", "Кристофер", "Робин"; "Eeyore", "Иа"; "Rabbit", "Кролик"). По близости мест, занимаемых некоторыми из перечисленных объектов в соответствующих упорядочениях (см. "Piglet", "Пятачок", "Хрюка"),

Таблица 2

Фрагменты полных частотных спектров L-грамм
для текста Милна и его переводов

L	№	Милн		Заходер		Вебер	
		L-грамма	F	L-грамма	F	L-грамма	F
1	1	and/	2204	и/	1711	и/	1779
1	2	the/	1526	сказа/	934	что/	941
1	3	he/	1343	он/	898	не/	938
1	4	to/	1287	что/	882	в/	835
1	5	said/	1259	не/	801	Пух/	760
1	6	a/	1113	в/	742	я/	638
1	7	it/	1105	Пух/	692	а/	638
1	8	Pooh/	894	я/	646	он/	632
1	9	I/	877	а/	619	на/	605
1	10	of/	802	бы/	525	с/	463
1	11	was/	739	это/	466	Хрюка/	442
1	12	you/	718	на/	459	то/	399
1	13	Piglet/	553	Пятачок/	425	так/	379
1	14	in/	543	все/	394	все/	365
1	15	that/	505	как/	358	это/	352
1	16	his/	399	так/	352	как/	340
1	17	what/	399	ты/	303	ты/	334
1	18	had/	372	то/	302	Иа/	328
1	19	so/	368	с/	300	Кролик/	325
1	20	as/	355	Кролик/	293	но/	298
1	21	for/	353	но/	241	к/	275
1	22	on/	348	Робин/	237	его/	270
1	23	is/	334	Кристофер/	235	бы/	269
1	24	Christopher/	318	Иа/	233	Робин/	263
1	25	they/	315	они/	230	Кристофер/	262
1	26	Eeyore/	310	ес/(ли,ть)	225	если/	252
1	27	all/	310	потому/	224	потому/	249
1	28	at/	307	Тигр/	222	у/	219
1	29	Robin/	305	очень/	211	же/	218
1	30	but/	297	его/	206	они/	208
2	1	Christopher/ Robin/	303	Кристофер/ Робин/	234	Кристофер/ Робин/	260
2	2	said/Pooh/	287	сказа/Пух/	204	потому/что/	93

Продолжение таблицы 2

		Милн		Заходер		Вебер	
L	№	*L*-грамма	*F*	*L*-грамма	*F*	*L*-грамма	*F*
2	3	it/was/	170	потому/что/	104	а/потом/	92
2	4	and/then/	168	сказа/Кролик/	84	что/то/	76
2	5	of/the/	167	сказа/Пятачок/	81	Пух/и/	75
2	6	and/he/	166	и/он/	80	Крошк/Ру/	67
2	7	he/said/	158	сказа/он/	71	ничего/не/	51
2	8	in/the/	150	Пух/и/	69	и/не/	49
2	9	he/had/	135	что/он/	68	о/чем/	49
2	10	said/Piglet/	127	Крошк/Ру/	63	и/Хрюка/	47
3	1	said/Christopher/ Robin/	78	сказа/Кристофер/ Робин/	58	Пух/и/Хрюка/	26
3	2	Winnie/the/ Pooh/	49	Пух/и/Пятачок/	26	Кристофер/ Робин/и/	24
3	3	and/then/he/	43	сказа/Пух/я/	18	о/том/что/	18
3	4	Pooh/and/ Piglet/	35	Кристофер/ Робин/и/	15	себе/под/нос/	17
3	5	Christopher/ Robin/and/	32	сказа/Пух/а/	15	горшк/с/мед/	13
3	6	there/was/a/	32	потому/ что/он/	13	с/Кристофером/ Робином/	13
3	7	said/Pooh/I/	30	и/сказа/что/	13	в/том/что/	12
3	8	it/was/a/	28	а/я/дума/	13	ответи/Кристофер/Робин/	12
3	9	of/the/ forest/	25	нет/сказа/ Пух/	12	спроси/Кристофер/Робин/	11
3	10	a/long/time/	24	Кристофер/ Робин/сказа/	12	с/дн/ рождени/	10
4	1	for/a/long/ time/	17	поздравля/с/ дн/рождени/	7	поздравля/с/ дн/рождени/	6
4	2	the/Hundred/ Acre/Wood/	13	медвед/с/ опилк/в/	6	понима/о/ чем/я/	5
4	3	in/the/middle/ of/	13	с/опилк/в/ голов/	6	к/тому/мест/ где/	5
4	4	the/top/ of/the/	13	сказа/Кристофер/Робин/и/	6	во/втор/полови/дн/	5
4	5	said/Winnie/ the/Pooh/	12	на/тот/ случа/если/	6	друз/и/родич/ Кролик/	5

можно предполагать, что речь идет о *различных переименованиях* одного и того же лица.

Особо следует отметить глагол "said" ("сказал"), занимающий очень высокие места в упорядочениях у Милна и Заходера (соответственно, 5–е и 2–е), но не попавший в первые пять десятков у Вебера ($F = 135$, 55–е место). Этот факт нельзя отнести к разряду случайных. Высокая по любым меркам частота встречаемости глагола "said" у Милна — это сознательно используемый прием, учитывающий роль фактора повторности в восприятии детьми описаний различных ситуаций. Заходер уловил эту стилистическую особенность и сохранил ее в своем переводе. Вебер также уловил ее, но в своем стремлении сделать не так, как у Заходера, он "поправил" не только Заходера, но и самого Милна, предложив широкий спектр возможных переводов слова "said", далеко не всегда тождественных по смыслу и соответствующих ситуации. О том, что эта деятельность велась целенаправленно и с размахом, свидетельствует следующий список слов, использованных Вебером в качестве эквивалента "said" (указаны их частоты у Заходера и Вебера): "воскликнул" ($F_{\text{Зах.}} = 1$, $F_{\text{Веб.}} = 57$); "ответи" ($F_{\text{Зах.}} = 30$, $F_{\text{Веб.}} = 157$); "кивну" ($F_{\text{Зах.}} = 17$, $F_{\text{Веб.}} = 61$); "продолжи" ($F_{\text{Зах.}} = 1$, $F_{\text{Веб.}} = 21$); "прошепта" ($F_{\text{Зах.}} = 1$, $F_{\text{Веб.}} = 13$) и т.д.

5. Состав высокочастотных биграмм и триграмм в значительной мере предопределен составом высокочастотных одиограмм, которые объединяются в связки с помощью соединительных союзов ("Пух/и/Пятачок/", "и/он/"), предлогов ("in/the/", "о/том/что/"), глагольного управления ("said/Pooh/", "сказа/он/") и согласования (например, в наименованиях "Кристофер/Робин/"). При $L = 4$ на первый план начинают выходить устойчивые словосочетания, связанные с тематикой произведения ("the/Hundred/Acre/Wood/", "поздравля/с/дн/рождени/", "во/втор/половин/дн/" и т.п.). Максимальные повторы во всех трех текстах связаны с дублированием одной из песенок Винни-Пуха в одной и той же главе ($L_{\max}$(Милн)= 39, $L_{\max}$(Заходер)= = 25, $L_{\max}$(Вебер)= 19). Следует отметить, что эта характеристика не является "устойчивой": достаточно незначительной вариации, чтобы разрушить длинный повтор. Так, перестанов-

ка двух слов при повторении песенки в переводе Вебера снизила длину максимального повтора до 19.

6. В связи с тем, что в состав высокочастотной лексики входят не только служебные слова, но и слова, определяющие тематику произведения (в частности, имена действующих лиц), возникает вопрос, можно ли формально разделить эти две категории слов, не прибегая к истолкованию их значений. Подобного рода вопросы типичны для задач информационного поиска, где важно, как минимум, уметь выделять из текста набор ключевых слов и словосочетаний. Нам представляется, что очень полезной в этом плане является позиционная информация, т.е. информация о местах вхождения в текст конкретного слова или словосочетания. Общая закономерность состоит в том, что *слова, определяющие содержание текста*, как правило, *распределены* по нему *неравномерно* даже при относительно высокой частоте встречаемости. Если рассмотреть для иллюстрации позиционные распределения в английском тексте двух слов: "Eeyore" (кличка ослика) и "all" (все, всё), занимающих 26-е и 27-е места в частотном упорядочении ($F = 310$ для каждого объекта), то окажется, что первое слово не встречается в позициях $1 \div 5552$ (при длине текста $N = 49269$), а также $12504 \div 16181$, $27557 \div 30535$, $38942 \div 42248$, т.е. максимальные "гэпы" (фрагменты, свободные от данного слова) имеют размеры порядка 4–5 тыс.слов. Аналогичный показатель для второго объекта ("all") составляет около тысячи слов, что свидетельствует о гораздо более равномерном (практически случайном) распределении его по тексту. Длина максимального "гэпа" — далеко не единственный показатель, по которому можно судить о неравномерности позиционного распределения конкретного слова. Существуют различные статистические критерии для выявления позиционных аномалий (см. для обзора [12]). В общем случае вопрос о взаимосвязи информационной значимости слова с его позиционным распределением требует специального изучения.

§5. Совместные частотные характеристики переводов Заходера и Вебера и их дополнения

Обозначим текст, соответствующий переводу Заходера, через $T_{\text{З}}$, а переводу Вебера и Рейн — через $T_{\text{В}}$. Тогда в приводимых ниже таблицах $M_L(T_{\text{З}}, T_{\text{В}})$ означает число разных L-грамм, общих для $T_{\text{З}}$ и $T_{\text{В}}$, $\rho_L^{\text{З}} = M_L(T_{\text{З}}, T_{\text{В}})/M_L(T_{\text{З}})$ — доля, которую составляют общие L–граммы от полного словаря L–грамм у Заходера (в процентах), соответственно, $\rho_L^{\text{В}} = M_L(T_{\text{З}}, T_{\text{В}})/M_L(T_{\text{В}})$ — доля общих L-грамм у Вебера. Условимся далее под $D_L(T_{\text{З}}|F > 1)$ понимать совокупность L-грамм, представленных только в $T_{\text{З}}$ и имеющих частоту $F > 1$. Аналогичная характеристика для текста Вебера — $D_L(T_{\text{В}}|F > 1)$. В терминах §1 $D_L(T_{\text{З}}|F > 1)$ и $D_L(T_{\text{В}}|F > 1)$ — это усеченные дополнения. Число различных L–грамм, представленных в усеченных дополнениях, будем обозначать, соответственно, как $|D_L(T_{\text{З}}|F > 1)|$ и $|D_L(T_{\text{В}}|F > 1)|$, а максимальные частоты L-грамм из дополнений — как $F_{\max}(D_L(T_{\text{З}}))$ и $F_{\max}(D_L(T_{\text{В}}))$.

В табл.3 для различных L представлены перечисленные выше параметры совместных частотных характеристик и дополнений. В табл.4 и 5 показаны фрагменты этих характеристик. Поскольку главная наша цель состоит в выявлении наиболее существенных различий в параллельных переводах, мы включили в табл.4 минимальное число L–грамм, достаточно полно представленных в обоих текстах, дополнив их значительным количеством "контрастных" L-грамм, т.е. таких, у которых частоты встречаемости в $T_{\text{З}}$ и $T_{\text{В}}$ сильно отличаются. При каждом значении L равномерно представленные в обоих текстах L–граммы отделены от контрастных строкой звездочек."Контрастные" L-грамммы в табл.4 расположены в соответствии с убыванием модуля разности рангов каждой L-граммы в упорядочениях $\Phi_L(T_{\text{З}})$ и $\Phi_L(T_{\text{В}})$. Группы равночастотных L–грамм в этих упорядочениях характеризовались средним значением ранга. Предельным проявлением "контрастности" являются L-граммы, вошедшие в дополнения (табл.5): частота встречаемости их в одном из текстов равна нулю.

Отметим наиболее существенные особенности данных, представленных в табл. 3–5.

Таблица 3

Динамика изменения с ростом L параметров совместных частотных характеристик и усеченных дополнений для текстов $T_З$ и $T_В$

L	Совместные частотные характеристики			Дополнения ($F > 1$)			
				Вебер		Заходер	
	$M_L(T_В, T_В)$	$\rho_L^В(\%)$	$\rho_L^З(\%)$	$F_{max}(D_L)$	$\lvert D_L \rvert$	$F_{max}(D_L)$	$\lvert D_L \rvert$
1	2618	54,5	61,9	442	602	69	402
2	7229	25,7	29,3	47	2254	81	1866
3	4066	10,6	11,2	26	1231	26	1291
4	1806	4,2	4,6	5	424	6	503
5	856	2,0	2,2	4	173	6	190
6	411	0,9	1,0	4	87	3	105
7	203	0,5	0,5	3	52	3	71
8	102	0,2	0,3	2	34	3	56
9	51	0,1	0,1	2	25	3	46
10	26	0,06	0,07	2	20	2	37
11	15	0,03	0,04	2	16	2	31
12	8	0,02	0,02	0	0	2	26
13	4	0,01	0,01	0	0	2	22
14	1	0,002	0,002	0	0	2	19

Таблица 4

Фрагменты совместных частотных характеристик текстов Заходера и Вебера
($F_З$ и $F_В$ – частоты встречаемости L–граммы в соответствующих текстах)

L	Общая L-грамма	$F_З$	$F_В$	L	Общая L-грамма	$F_З$	$F_В$	L	Общая L-грамма	$F_З$	$F_В$
1	и/	1711	1779	1	палк/	11	26	2	воздушн/шарик/	5	20
1	что/	882	941	1	ужасн/	32	13	2	сказа/Пух/	204	6
1	не/	801	938	1	киваю/	17	61	3	Кристофер/	15	24
1	в/	742	835	1	ответи/	30	157		Робин/и/		
1	Пух/	692	760	1	тигр/	222	41	3	о/том/что/	7	18
1	я/	646	638	1	сказа/	934	136	3	спроси/Кристофер/	11	11
1	а/	619	638	1	уж/	92	200		Робин/		
1	он/	898	632	1	же/	129	218	3	с/дн/рождени/	10	10
1	на/	459	605	2	Кристофер/Робин/	234	260	3	потому/что/он/	13	8
1	с/	300	463	2	потому/что/	104	93	3	и/кролик/Ру/	10	9
*	**********	***	***	2	а/потом/	54	92	3	все/в/порядк/	11	8
1	немножко/	24	1	2	что/то/	35	76	3	Кристофер/	10	8
1	минутк/	14	1	2	Пух/и/	69	75		Робин/не/		
1	сперва/	13	1	2	Кролик/Ру/	63	57	3	я/хочу/сказа/	12	7
1	помча/	13	1	2	ничего/не/	54	51	3	во/всяк/случа/	5	9
1	увы/	12	1	2	и/не/	33	49	**	**********	***	***
1	поближе/	10	1	2	о/чем/	33	49	3	сказа/Пух/и/	11	1
1	очевидн/	10	1	2	я/не/	52	47	3	прин/в/голов/	10	1
1	воскликну/	1	57	**	**********	***	**	3	и/знаком/Кролик	9	1
1	продолжи/	1	21	2	сказа/Сов/	36	1	3	что/он/не/	8	1
1	репейник/	1	18	2	Пух/сказа/	31	1	3	не/мож/бы/	7	1
1	огляде/	1	17	2	по/дорог/	20	1	3	и/тут/же/	1	9
1	свидани/	1	16	2	бы/так/	20	1	3	со/всех/сторон/	1	7
1	слабеньк/	1	14	2	как/будто/	17	1	3	себе/под/нос/	4	17
1	прошепта/	1	13	2	что/там/	10	1	3	сказа/Кристофер/	58	5
1	дважды/	1	12	2	дорог/мой/	10	1		Робин/		
1	половин/	1	12	2	я/сам/	9	1	4	поздравля/с/	7	6
1	поскольку/	1	11	2	ой/Иа/	9	1		дн/рождени/		
1	шар/	26	2	2	знаком/Кролик/	9	1	4	ты/понима/что/я/	5	3
1	чем-нибудь/	17	2	2	киваю/Пух/	1	16	4	с/этими/слов/он/	3	4
1	подкрепи/	14	2	2	нет/ответи/	1	16	4	добр/утр/Кристофер/	3	3
1	пауз/	2	16	2	до/свидани/	1	14		Робин/		
1	ах/	28	3	2	и/теперь/	1	13	4	Кенга/и/Кролик/Ру/	3	3
1	домик/	3	52	2	ответи/Кролик/	1	11	**	**********	***	***
1	Пятачок/	425	4	2	сказа/Кролик/	34	2	4	сказа/Кристофер/	6	1
1	ох/	16	3	2	да/да/	16	2		Робин/и/		
1	поздорова/	3	26	2	ну/вот/	15	2	4	ничего/не/говори/потому/	5	1
1	закрича/	28	4	2	два/раз/	10	2	4	ему/прин/в/голов/	5	1
1	вновь/	4	47	2	на/что/	2	17	4	как-нибудь/и/	4	1
1	заяви/	4	19	2	спасибо/тебе/	2	12		друг/раз/		
1	переспроси/	4	16	2	зади/лап/	2	12	4	понима/о/чем/я/	1	5
1	туда/	30	6	2	мож/бы/	45	4	4	а/потом/Кристофер/	1	4
1	согласи/	6	22	2	сказа/он/	71	5		Робин/		
1	палоч/	28	12	2	ответи/Пух/	5	30	4	ни/о/чем/не/	2	5

Таблица 5

Статистика L-грамм, встретившихся только в одном из текстов (фрагменты дополнений)

Перевод Б.Заходера			Перевод В.Вебера и Н.Рейн		
L	L-грамма	$F_З$	L	L-грамма	$F_В$
1	Иа-Иа/	69	1	Хрюка/	442
1	здравствуй/	26	1	Тигер/	183
1	опилк/	17	1	Хоботун/	35
1	тирлим-бом-бом/	14	1	разуме/	26
1	Слонопотам/	14	1	малявк/	25
1	чертополох/	13	1	хо-хо/	19
1	совсем-совсем/	12	1	родич/	15
1	например/	12	1	озабоченн/	14
1	пап/	12	1	умишк/	14
1	поспешн/	11	1	Скорабуду/	13
1	мам/	11	1	бубни/	13
1	дремуч/	10	1	праздничн/	12
1	все-все-все/	9	1	вырва/	12
1	влез/	9	1	трам-пам/	12
1	тсс/	9	1	подели/	11
1	огляну/	8	1	выда/	10
1	хвалебн/	8	1	кольц/	9
1	Щасвирнус/	8	1	ворот/	9
1	локт/	8	1	полюбопытствова/	9
1	шумелку/	7	1	предупреди/	9
1	ай-ай-ай/	6	1	коротк/	9
1	никогда-никогда/	6	1	напрыгива/	8
2	сказа/Пятачок/	81	1	тра-ля-ля/	8
2	сказа/Иа/	58	1	фут/	8
2	и/Пятачок/	42	1	ш-ш-ш/	7
2	нет/сказа/	23	1	сочинени/	7
2	воздушн/шар/	13	2	и/Хрюка/	47
2	дремуч/лес/	10	2	вроде/бы/	22
2	родственник/и/	9	2	реши/что/	19
2	грусти/сказа/	9	2	спроси/Хрюка/	19
2	маленьк/существ/	8	2	и/уж/	14
2	по/правд/	8	2	очен/даже/	14
2	хвалебн/песн/	7	2	воскликну/Пух/	14
2	здравствуй/Пух/	6	2	конечн/же/	13
2	не/спеш/	6	2	ответи/Кристофер/	12

Перевод Б.Заходера			Перевод В.Вебера и Н.Рейн		
L	L-грамма	$F_З$	L	L-грамма	$F_В$
2	дом/Иа/	6	2	слабеньк/умишк/	11
2	еще/немножк/	6	2	и/родич/	11
2	рыб/жир/	6	2	озабоченн/спроси/	11
2	но/увы/	6	2	радостн/воскликну/	10
2	я/личн/	6	2	друз/и/	10
3	Пух/и/Пятачок/	26	2	к/пример/	10
3	сказа/Пух/я/	18	2	к/домик/	9
3	сказа/Пух/а/	15	2	так/ли/	8
3	а/я/дума/	13	2	раз/уж/	8
3	нет/сказа/Пух/	12	2	праздничн/обед/	8
3	потому/что/ведь/	11	2	долг/науз/	6
3	ну/что/ж/	9	3	Пух/и/Хрюка/	26
3	родственник/и/знаком/	9	3	ответи/Кристофер/Робин/	12
3	и/вдруг/он/	7	3	друз/и/родич/	9
3	опилк/в/голов/	7	3	не/так/ли/	8
3	в/дремуч/лес/	6	3	на/случа/если/	8
3	очень/маленьк/существ/	6	3	и/родич/кролик/	7
3	по/правд/говор/	6	3	да/кивну/Пух/	6
3	глупеньк/мой/мишк/	5	3	втор/половин/дн/	6
3	и/тому/подобн/	5	3	со/слабеньк/умишк/	6
4	медвед/с/опилк/в/	6	3	нет/ответи/Пух/	5
4	с/опилк/в/голов/	6	3	так/уж/получи/	5
4	Пух/сказа/Кристофер/Робин/	5	4	во/втор/половин/дн/	5
4	родн/и/знаком/Кролик/	5	4	друз/и/родич/Кролик/	5
4	ну/так/что/же/	4	4	ты/понима/о/чем/	4
4	на/ветк/сиде/Пух/	4	4	мишк/со/слабеньк/умишк/	4
4	он/толсте/не/стан/	4	4	и/тут/же/добави/	4
4	да/сказа/Кристофер/Робин/	4	4	не/так/уж/и/	4
4	родственник/и/знаком/Кролик/	4	4	бубни/себе/под/нос/	4

1. Цепочки, общие для $T_Э$ и $T_В$, составляют заметную долю от словарей L–грамм в каждом из текстов лишь для значений $L = 1, 2, 3$ (см. столбцы $\rho_L^Э$ и $\rho_L^В$ в табл.3). Это означает, что при $L > 3$ дополнения $D_L(T_Э)$ и $D_L(T_В)$ будут мало отличаться соответственно от $\Phi_L(T_Э)$ и $\Phi_L(T_В)$, т.е. дополнения информативны лишь для небольших значений L. Для текстов, не являющихся параллельными, доля общих цепочек с ростом L убывает еще быстрее. Динамика изменения этой зависимости характеризует степень близости текстов.

2. Максимальное разнообразие цепочек, представленных в совместных частотных характеристиках, наблюдается при $L = 2$. Число кратных ($F > 1$) цепочек, представленных в дополнениях, также достигает максимума при $L = 2$. Это согласуется с выводом о том, что максимальное разнообразие повторяющихся L–грамм для одного текста имеет место при $L = 2$ (см. § 4).

3. Максимальные частоты L–грамм из дополнений не являются "устойчивыми" характеристиками параллельных текстов. Так при $L = 1$ $F_{max}(D_1(T_В)) = 442$ (достигается на слове "Хрюка"), а $F_{max}(D_1(T_Э)) = 69$ (достигается на слове "Иа-Иа"). Столь большое различие между максимальными частотами 1–грамм из дополнений объясняется тем, что Вебер переименовал заходеровского "Пятачка" (одного из главных действующих лиц) в "Хрюку", но вместе с тем несколько раз он все же использовал слово "пятачок", подразумевая "кончик носа" у поросенка. Поскольку заглавные и строчные буквы при формировании L–грамм не различаются, слово "Пятачок", встречающееся в переводе Заходера столь же часто, как "Хрюка" в переводе Вебера, не вошло в состав $D_1(T_Э)$, а представлено в совместной частотной характеристике 1–го порядка текстов $T_Э$ и $T_В$. В то же время максимальная частота L–грамм из $D_1(T_Э)$, достигаемая на слове "Иа–Иа" (кличка ослика), явно занижена, поскольку Заходер кроме "Иа–Иа" часто использует сокращенный вариант (просто "Иа"), который трактуется как самостоятельное слово и к тому же фигурирует у Вебера (т.е. "Иа" не входит в состав $D_1(T_Э)$).

Аналогичные коллизии возникают и при $L = 2$. Здесь уже $F_{max}(D_2(T_Э))$ существенно превышает не только $F_{max}(D_2(T_В))$, но и $F_{max}(D_1(T_Э))$, что нетипично, поскольку с увеличением L

максимальные частоты обычно падают. Объяснение состоит в том, что максимум для $D_2(T_Э)$ достигается на биграмме "сказа/Пятачок/" ($F = 81$), которая состоит из двух высококонтрастных однограмм ("сказа" и "Пятачок", см. табл.4, $L = 1$). Вебер, как мы уже отмечали в § 4, сознательно избегает частого употребления слова "сказал", что снижает частоты соответствующих биграмм. В итоге самой частой биграммой в $D_2(T_В)$ оказывается "и/Хрюка/" ($F = 47$), но она значительно уступает по частоте рекордсмену из $D_2(T_Э)$ ("сказа/Пятачок/").

4. Свойство контрастности, понимаемое как резкое нарушение баланса в распределении конкретной L–граммы по двум текстам, теряет свою силу с увеличением L. Это объясняется быстрым уменьшением максимальных частот L-грамм с ростом L. Из табл.4 видно, что уже при $L = 4$ баланс частот (максимальной и минимальной) составляет 6:1 (чаще 5:1), так что говорить о "резком" нарушении баланса можно лишь с некоторой натяжкой.

5. Контрастные L–граммы служат для формирования более длинных цепочек ($(L + 1)$-, $(L + 2)$-грамм и т.д.), входящих уже в состав дополнений. Например, контрастные 1–граммы "сказа/", "Пятачок/", "шар/", "пауз/", "домик/" и др. (см. табл.4) в некотором смысле предопределяют появление в составе дополнений $D_2(T_Э)$ и $D_2(T_В)$ таких биграмм как "сказа/Пятачок/", "и/Пятачок/", "воздушн/шар/", "долг/пауз/", "к/домик/" и т.п. (см. табл.5).

6. Из определения $D_L(T)$ следует, что любая цепочка текста, содержащая в себе L–грамму из дополнения, также входит в состав дополнения, но более высокого порядка. Это означает, что высокочастотные 1–граммы из $D_1(T_Э)$ или $D_1(T_В)$ с большой вероятностью входят в состав наиболее частых биграмм из $D_2(T_Э)$ (соответственно $D_2(T_В)$), последние могут входить в состав триграмм из $D_3(T_Э)$ (или $D_3(T_В)$) и т.д. Например, слово "Хрюка" (рекордсмен по частоте в $D_1(T_В)$) входит в состав самой частой биграммы "и/Хрюка/" из $D_2(T_В)$, последняя, в свою очередь, входит в состав высокочастотной триграммы "Пух/и/Хрюка/" из $D_3(T_В)$. Суммируя выводы пп.5 и 6, можно сказать, что L–граммы из дополнений с большой вероятностью содержат в

себе цепочки из дополнений и пересечений более низкого порядка, причем цепочки из пересечений, как правило, являются контрастными.

Следует, однако, отметить, что не всегда наличие коротких контрастных или "суперконтрастных" (из дополнений) L–грамм предопределяет существование длинных L–грамм с теми же свойствами. Иногда первичными являются именно длинные цепочки, повторяющиеся в одном из сравниваемых текстов, но отсутствующие в другом. При этом отдельные элементы таких цепочек тоже могут обладать свойством контрастности, но оно будет иметь вторичный характер, т.е. вытекать из контрастности исходных цепочек. В качестве примера можно указать на два оборота, детализирующие портрет Винни–Пуха: "медведь с опилками в голове" (Заходер) и "мишка со слабеньким умишком" (Вебер). Фрагменты этих оборотов встречаются в дополнениях $D_1(T_З) \div D_4(T_З)$ (1–й вариант) и, соответственно, $D_1(T_В) \div D_4(T_В)$ (2–й вариант). Из–за варьирования часто́ты отдельных фрагментов (к примеру, таких как "опилк/" в первой фразе и "умишк/" — во второй) превышают частоты самих фраз, однако все вхождения этих фрагментов в $T_З$ и $T_В$ имеют место лишь в контексте указанных фраз.

7. Анализ переводов Заходера и Вебера указывает на наличие нескольких механизмов, обеспечивающих появление "первичных" контрастных L–грамм, т.е. таких, контрастность которых не обусловлена контрастностью содержащихся в них цепочек или содержащих их надцепочек.

Среди возможных механизмов отметим:

— *блочные вставки, делеции и замены.* Так, слово "малявка", фигурирующее в дополнении первого порядка у Вебера ($F = 25$), — это персонаж главы, выпущенной Заходером при переводе, но включенной Вебером (больше этот персонаж нигде не встречается). Слова "котлетный" ($F = 6$) и "конфетный" ($F = 6$) — это вольный перевод Н. Рейн слова "cottleston" в песенке из главы 6 ($F = 19$). Заходер же в этом месте вообще не придерживается оригинала и вставляет свою песенку–загадку (пример блочной замены), не содержащую указанных слов;

— *использование* авторами параллельных переводов *разного числа синонимов* для обозначения одного и того же лица, объекта, явления. Например, у Заходера ослик имеет два имени ("Иа" и "Иа-Иа"), а у Вебера — одно ("Иа"). Поэтому "Иа-Иа" является высокочастотной 1-граммой (входит в состав $D_1(T_З)$), а "Иа" — нет. Другой пример связан с оборотом "one day" ($F = 14$), который Заходер переводит как "однажды" ($F = 17$), а Вебер как "однажды" ($F = 11$) и "как-то раз" ($F = 6$). В итоге последняя L-грамма оказывается контрастной;

— *индивидуальные стилистические приемы.* Чрезвычайно высокая частота использования слова "said" у Милна — пример такого приема. Заходер также очень часто использует слово "сказал", но он лишь следует Милну, стараясь точнее передать особенности его стиля. Вебер же этого не делает и предлагает целый спектр далеко не синонимичных переводов слова "said" (см. табл. 6), многие из которых обладают свойством контрастности. Индивидуальный стилистический прием, используемый только Заходером, — двукратное (а порой и трехкратное) усиливающее повторение отдельных слов: "очень-очень", "совсем-совсем", "никогда-никогда", "все-все-все", "ай-ай-ай". Все они относятся к числу контрастных. Визитная карточка Вебера — слово "разумеется" ($F = 26$ в $D_1(T_В)$). В оригинале ему часто (но далеко не всегда) соответствует оборот "of course", у Заходера — "конечно";

— *целенаправленное варьирование оригинала или уже имеющегося перевода* (последний вариант встречается чаще). Поскольку отступление от оригинала не должны быть значительными, речь идет преимущественно о синонимическом варьировании. Строго обосновать целенаправленный характер варьирования невозможно за исключением очевидных случаев, когда переводчик привносит что-то свое, чего не было в оригинале. Однако существуют косвенные признаки, позволяющие судить о влиянии имеющегося перевода на последующие. Некоторые из них обсуждаются в следующем параграфе.

§6. Детализация различий двух переводов

В табл. 6 приведены примеры соответствий между отдельными контрастными L-граммами, характерными для одного из тек-

Таблица 6

Примеры соответствий между контрастными L-граммами (помечены звездочкой) и их аналогами в других текстах ($F_З$, $F_В$ и $F_М$ — частоты L-грамм в $T_З$, $T_В$ и $T_М$)

	Заходер			Вебер			Милн	
№	L-грамма	$F_З$	$F_В$	L-грамма	$F_З$	$F_В$	L-грамма	$F_М$
1	Пятачок/	425	4	Хрюка/*	0	442	Piglet/	553
2	Тигр/(а)	222	0	Тигер/*	0	224	Tigger/	183
3	Иа-Иа/*	69	0	Иа/	233	328	Eeyore/	310
4	Слонопотам/	14	0	Хоботун/*	0	35	Heffalump/	46
5	Бук/(а, и)*	11	0	Вузл/(а)	0	6	Woozle/(s)	11
6	метр/	5	0	фут/*	0	8	foot/	6
7	тсс/*	9	0	шшш/	0	7	hush/	10
8	тирлим-бом-бом/*	14	0	трам-пам/	0	12	tiddely-pom/	20
9	(ворра)[5]*	2	0	(ворра)[4]	0	2	(*worra*)[5]	2
10	аа/*	19	1	— (ах, да, понятно)			oh/	134
11	чертополох/*	13	0	репейник/	1	18	thistle/(s)	10
12	тревожн/	7	0	озабоченн/*	0	14	anxiously/	11
	испуганн/	4	0					
13	ошлк/в/голов/	7	0	слабеньк/умишк/*	0	11	very/little/brain/	8
	глупеньк/	7	1				silly/old/Bear/	6
14	Щасвирнус/	8	0	Скорабуду/*	0	13	Backson/	12
15	родственник/	9	0	друз/и/родич/	0	9	friends/and/	11
	и/знаком/*						relations/	
16	пятнист/или/	4	0	пятнаст/или/	0	3	the/Spotted/or/	3
	травоядн/*			цветоядн/			Herbaceous/	
17	очень/маленьк/	9	0	очень/	10	7	a/very/	8
	существ/*			маленьк/			small/animal/	
18	здравствуй/*	26	0	привет/	15	59	hallo/	57
							good/morning/	14
19	наскакива/	7	2	напрыги-	0	16	bounced/	13
	наскочи/	7	1	ва/(ть, л)*			bouncing/	6
20	дом/	176	154	домик/*	3	52	house/	123
21	шар/*	26	2	шарик/	17	41	balloon/	39
22	подкре-	14	2	перекуси/	0	5	— (см. при-	
	пи/(ться)*			перехвати/	0	1	мер в разд.2)	

Продолжение таблицы 6

	Заходер			Вебер			Милн	
№	L-грамма	$F_З$	$F_В$	L-грамма	$F_З$	$F_В$	L-грамма	$F_М$
23	хитр/ западн/(я)	3	0	хитроумн/ западн/(я)*	0	4	cunning/trap/	3
24	дремуч/ лес/	10	0	столетн/лес/*	0	12	the/Hundred/ Acre/Wood/	13
25	рыб/жир/*	6	0	пивн/дрожж/	0	5	medicine/	11
26	сказа/	934	136	сказа/	934	136	said/	1259
				воскликну/	1	57		
				ответи/	30	157		
				кивну/	17	61		
				поздорова/	3	26		
				продолжи/	1	21		
				прошепта/	1	13		
				заяви/	4	19		
				согласи/	6	22		
				вырва/(лось)	0	12		
27	закрича/*	28	4	— (проверещал, заверещал, завопил)			shouted/ cried/	9 31
28	немножко/*	24	1	— (немного, немно- жечко, несколько,...)			a/little/ ... (something, futher, longer, ...)	 8 4 4
29	минутк/*	14	1	— (чуть, немного, ...)			for/a/moment/ wate/a/moment/	11 5
30	— (размышляя, раздумывая, недоумевая)			гада/(я)*	0	8	wondering/	22
31	шумелк/ кричалк/ ворчалк/	13 9 3	1 0 0	бубнилк/*	0	18	a/noise/ a/ .../hum/ .../murmured/	29 14 4
32	Ягуляр/*	6	0	Ягулар/	0	5	Jagular/	6
33	искпедицию/	4	0	икспедицию/*	0	6	Expotition/	10
34	огляну/	8	0	огляде/*	1	17	looked/round/	8
35	очень-очень/*	14	2	—			—	
36	—			очень/даже/*	0	14	—	
37	например/*	12	0	—			—	
38	—			к/пример/(у)*	0	10	—	

стов (Заходера или Вебера) и их аналогами в другом тексте и оригинале. Нумерация в таблице фиксирует исходные контрастные L–граммы, но не их аналоги, поскольку аналогов может быть несколько или не быть вовсе. Аналоги могут отличаться по длине от исходной L–граммы и не всегда удовлетворяют свойству контрастности. Когда аналогов слишком много или вообще нет, в соответствующей графе ставится прочерк, означающий отсутствие устойчивой закономерности. Иногда в скобках при этом приводятся примеры аналогов (если они есть). Наличие двух прочерков в строке (к примеру, у Милна и Вебера или у Милна и Заходера) означает, что контрастная L–грамма, закрепленная за данной строкой, может рассматриваться как проявление авторского стиля (в данном случае, соответственно, Заходера или Вебера).

Анализ табл.6 и материалов, не вошедших в нее, позволяет выделить следующие схемы варьирования, имевшие место при переводе оригинала.

1. *Переименования действующих лиц* (см. №№1 ÷ 5 в табл.6). К ним прибегают и Заходер и Вебер. В некотором смысле они неизбежны, поскольку многие имена у Милна придуманные (см. №№3 ÷ 5), часто искаженные в соответствии с тем, как их могли бы произнести дети (например, "Tigger" вместо "Tiger"). Заходер нашел чрезвычайно удачные эквиваленты для действующих лиц. Находки Вебера не отличаются фантазией: "Тигер" и "Вузла" — это буквальные заимствования англоязычных терминов их текста Милна; "Хоботун" же совсем не передает то ощущение мощи и силы, которое исходит от "Слонопотама" — комбинации из "слона" и "гиппопотама" у Заходера ("Heffalump" у Милна — также комбинация из искаженного "elephant" ("слон") и глагола "lump" ("to lump along" — тяжело ступать)). К переименованиям можно отнести и соответствия между отдельными лицами и местоимениями, обусловленные различиями в схеме изложения (от первого лица или от третьего). Так, слово "папа" (имеется в виду отец Кристофера Робина) встречается только у Заходера ($F = 12$), ведущего изложение от третьего лица. У Вебера ему соответствует местоимение "я", поскольку диалоги папы с Кристофером Робином ведутся от первого лица.

2. *Переход из одной системы мер и весов в другую* (см. №6). Заходер при переводе Милна заменял элементы английской метрической системы ("футы", "ярды", "фунты" и т.п.) русскими аналогами ("метры", "килограммы" и т.п.). Вебер все "вернул на место", вновь прибегнув к англоязычной лексике.

3. *Свобода в выборе звукоподражаний, восклицаний (междометий)* (см. №№7 ÷ 10). И Заходер и Вебер в этом вопросе не слишком придерживаются оригинала, хотя вариант Заходера в строке 8 выглядит предпочтительнее. Достаточно редкий случай, когда оба переводчика не отклонились от оригинала, представлен в строке 9, но и здесь Вебер вместо 5–ти кратного повторения слова "ворра" (как у Милна и Заходера) ограничился 4–х кратным. Многие междометия, число которых в английском тексте значительно выше, чем в русских, при переводе игнорируются, особенно Вебером. Например, междометие "oh" в оборотах типа "Oh! I see" или "Oh! — said Pooh" переводится Заходером как "А–а" ("А–а, понятно" или "А–а, — сказал Пух"), тогда как Вебер в этих случаях часть междометий пропускает, а для других не пытается найти устойчивого эквивалента ("ах", "да", "жаль", "ну, конечно" и т.п. — см. №10).

4. *Синонимические замены* (см. №№11 ÷ 25, 32 ÷ 34). Этот класс преобразований очень широк. Он включает в себя:

— *замены* (см. №№11, 12, 18, 22, 23), *вставки* и *делеции* (№17), а также *перестановки слов* (№15);

— *локальные варьирования* внутри слов (см. №16 ("пятнистый" — "пятнастый"), №34 ("оглянулся" — "огляделся")); к разновидностям локального варьирования можно отнести использование *уменьшительных форм* (характерно для Вебера (см. №№20, 21)) и *различных схем намеренного искажения слов* (№№32, 33);

— *замены* одних *словосочетаний* другими (см. №№13, 14, 24, 25).

5. *Разбиения* (замена часто встречающихся слов группами близких по смыслу (в конкретной ситуации) слов) (см. №№26 ÷ 30). Разбиение можно трактовать как варьирование, направленное на устранение "избыточных" повторов с помощью синонимичных подстановок. Оно вполне оправдано, когда

стремятся получить лексически сбалансированный и не перенасыщенный стереотипными конструкциями текст. Однако, если избыточность вводится автором в текст сознательно (например, в расчете на детское восприятие), использование данного приема нежелательно.

6. *Обвединения* (замена близких по смыслу слов одним единственным). Примером такого рода является использование Вебером обобщающего названия "бубнилки" для многочисленных песенок Винни–Пуха, среди которых Милн, а следом за ним и Заходер, выделяли различные "жанры" ("шумелки", "кричалки", "ворчалки" и т.п. — см. №31). Игнорирование Вебером "жанровой специфики" явно обеднило текст, не говоря уж о появлении "фирменных" оборотов типа "гордо *побубнивая* себе под нос" (у Заходера — "напевая свою песенку").

7. *Варьирование без сохранения позиционных соответствий* (см. №№35 ÷ 38). Речь идет об индивидуальных стилистических приемах (соответственно, Вебера и Заходера) синонимичных по смыслу, но позиционно не согласованных. Так, Заходер использует слово "например" (и только его) в ситуациях, когда ни Милн, ни Вебер ничего подобного не употребляют. Аналогично, Вебер использует синонимичный оборот "к примеру" (и только его), никак не привязываясь позиционно к местам вхождения слова "например" в текст Заходера. Возможно , слово "варьирование", рассматриваемое в контексте установления взаимосвязи между переводами Заходера и Вебера, лишь условно применимо для описания подобной ситуации. Можно, однако, говорить, о взаимосвязи каждого из переводов с текстом оригинала и тогда этот термин оказывается вполне уместным (см. п.8 и табл.6).

8. *Индивидуальные стилистические приемы* (даже не связанные отношением синонимии) могут рассматриваться как *элементы варьирования оригинала*. Из разъяснения, данного в начале этого раздела относительно наличия двух прочерков в одной строке табл.6, следует, что контрастная L–грамма, характеризующая индивидуальный стилистический прием переводчика, не является отражением какой–либо конструкции, представленной в оригинале или уже имеющемся переводе. Иными словами, речь идет о привнесении в перевод чего–то своего, т.е. о варьи-

рования оригинала. Кроме отмеченных в конце раздела 5 индивидуальных стилистических приемов укажем еще ряд типично "заходеровских" и "веберовских" оборотов, не вошедших в табл.6 (в скобках указаны частоты встречаемости оборота в $T_З$ и $T_В$ соответственно). *Заходер*: "по-моему" (34, 5); "может быть" (45, 4); "как будто" (17, 1); "вот как" (15, 2); "ну вот" (15, 2); "ну что ж" (9,0); "все время" (17, 3); "по правде говоря" (6, 0) и др.; *Вебер*: "вновь" (4, 47); "точно" (6, 34); "вроде бы" (0, 22); "конечно же" (0, 13); "и тут же" (1, 9); "поскольку" (1, 11); "со всех сторон" (1, 7) и др.

Какие же соображения, подтверждающие, что Вебер делал свой перевод "с оглядкой" на текст Заходера, можно привести по итогам рассмотрения табл.6? Укажем некоторые из них.

1. Удивляет "вето", наложенное Вебером на использование слова "здравствуй". Возможная причина состоит в том, что это слово является основным у Заходера для обозначения приветствий (см. №18).

2. Слово "thistle(s)" в словарях и Заходером переводится как "чертополох" (см. №11). Вебер переводит его как "репейник", что менее точно, поскольку репейник — это обобщающее название для целой группы колючих растений (терн, лопух, волчец, чертополох). Возможная причина опять же — сделать не так, как у Заходера.

3. Словосочетание "friends and relations " (см. №15) Заходер переводит двояко: как "родственники и знакомые" (основной вариант) и "родные и знакомые". Тем самым он сильно ограничивает выбор еще одного синонима для "relations", опять же, если исходить из необходимости "сделать иначе". Видимо поэтому Вебер при переводе этого слова использует устаревшую форму "родичи", частота встречаемости которой в современных текстах на порядок уступает варианту Заходера ("родственники"). Аналогичные соображения можно привести и по поводу использования Вебером слова "пятнистый" (у Заходера — "пятнистый" — №16).

4. Выражение "the Spotted or Herbaceous" (соединение разнородных терминов, дословно переводящееся как "пятнистый или травянистый") Заходер, отступая от оригинала, перевел как "пятнистый или травоядный" (см. №16). Вебер мог бы вернуться к

оригиналу, т.е. использовать термин "травяной" (или "травянистый"), но он, продолжая полемизировать с Заходером, придумывает новый термин "цветоядный". К подобному же "словотворчеству" Вебер прибегает, заменяя глагол "наскакивать" (у Заходера) своим собственным "напрыгивать" (см. №19). А глагольные формы от слова "Пух", использованные Н. Рейн в одной из песенок гл. 5 (ч. 2), могут смело претендовать на включение в словарь ненормативной лексики.

5. Слово "medicine" (см. №25 в табл.6) Милн всюду использует в смысле "лекарство", особо не уточняя о каком лекарстве идет речь. Заходер наряду с этим термином использует уточняющий ("рыбий жир"), подчеркивая, что речь идет об очень неприятном для многих детей лекарстве. Вебер, следуя Милну, также мог бы ограничиться словом "лекарство", но он, неявно признавая, что находка Заходера была удачной, вводит свой уточняющий термин "пивные дрожжи".

6. Особенно трудной оказалась для Вебера проблема подбора синонимичных выражений для фразеологизмов. Приведем несколько примеров на эту тему:

Заходер		*Вебер*
"остановился как вкопанный"	⟺	"застыл *как памятник*"
"побежал домой, что было духу"	⟺	"*со всех лап* бросился к своему домику"
"помчался прочь, продолжая вопить"	⟺	"пустился бежать, *вопя благим матом*"
"издал отчаянный, жалобный вопль"	⟺	"испустил *вопль грусти и отчаяния*"

Рассмотренные в данном разделе отличия двух переводов позволяют на наш взгляд сделать следующие

В ы в о д ы

1. И Заходер и Вебер в точности не следуют оригиналу. Однако первый об этом предупреждает сам ("пересказал Борис Заходер"), тогда как второй настаивает на близости своего перевода оригиналу ("ничего не привнося своего"[9]). Реально дело обстоит таким образом: в тех местах, где Заходер почти дословно

следует Милну, Вебер варьирует Заходера. Там же, где Заходер отходит от оригинала, Вебер следует Милну или, в свою очередь, отходит от оригинала.

2. Если отвлечься от композиционных изменений, внесенных Заходером, можно сказать, что он сделал максимально *"русифицированный"* и максимально *ориентированный на детскую аудиторию* того времени (60-е годы) перевод "Винни- Пуха", *сохранив стилистические приемы и находки Милна. "Русификация"* проявляется *в выборе имен действующих лиц* ("Тигра" вместо "Тигер" у Вебера, "Бука" вместо "Вузла"), *замене терминов* из английской системы мер и весов русскими эквивалентами, а строк из популярных английских песенок — аналогичными русскими ("в лесу родилась палочка"), и даже в *игнорировании отдельных атрибутов английского быта*, нетипичных для России (не упоминается слово "камин", устранены "свечи" из традиционного набора по случаю дня рождения и т.п.). Сам Заходер говорит об этом в предисловии к первому изданию следующим образом: "Я решил сперва выучить Винни и его друзей объясняться по–русски, что, уверяю вас, было тоже нелегко".

Ориентация на детскую аудиторию проявляется в *выделении заглавными буквами* и многократном *повторении наиболее важных* в сюжетном отношении *терминов* ("Очень Хитрая Западня", "Полезный Горшок" и т.п.), в *повышенной экспрессивности изложения*, достигаемой использованием многочисленных восклицаний (междометий), усиливающих тандемных повторов ("совсем–совсем", "очень–очень"), в *намеренном упрощении лексики* (Заходер избегает сложных и иноязычных слов, типа "меланхолично", "разумеется", "умиротворенный", "галион", присущих лексикону Вебера) и в других стилистических приемах и находках типа классификации песенок на "шумелки", "кричалки", "ворчалки" и т.п. Заметим попутно, что Заходер очень бережно относится к такого же рода находкам Милна: фразу о том, что пристыженный Пятачок решил "убежать из дому и стать моряком", он перенес из выпущенной им при переводе главы в конец главы 5.

3. *Вебер* в навязчивом стремлении "сделать все не так, как у Заходера" *вернул* в свой текст всю *англоязычную лексику*, пол-

ностью проигнорировав желание Заходера "выучить Винни и его друзей объясняться по-русски". Он не задумался, почему диалоги Кристофера Робина с отцом построены Заходером *от третьего лица*, и, изложив все *от первого лица*, получил фразы-ребусы типа: "И *я* того же мнения, — согласился с ним *я*", или "*Я* промахнулся? — спросил *ты*". *Вебер впрямую не воспользовался стилистическими приемами Заходера*, но, несомненно, при переводе многих фраз *учитывал расстановку акцентов, сделанную Заходером*. К примеру, фразу "And it must be a Cunning Trap" Заходер переводит как "И это должна быть Очень Хитрая Западня", добавляя слово "очень" и акцентируя тем самым внимание на том, что реализация плана поимки Слонопотама будет непростой. Вебер отслеживает этот акцент, не предписываемый Милном, и даже усиливает его "в своей манере": "Но это должна быть ну о-о-очень хитроумная западня".

Основным "достижением" Вебера следует считать то, что он, дистанцируясь от Заходера, *исказил стиль самого Милна*, которому Заходер старался следовать. Наиболее ярко это проявилось на схеме варьирования типа "разбиение" (см. в табл.6 далеко не полный список вариантов перевода Вебером глагола "said"). Суть схемы можно пояснить следующим примером. Фразу "So Pooh pushed and pushed and pushed his way through the hole, and at last he got in" Заходер переводит, сохраняя "характер повторности" (точные повторы кратности 3): "И Винни полез в нору. Он *протискивался, протискивался, протискивался* и, наконец, очутился там". Вариант Вебера выглядит так: "И Винни-Пух стал *вползать, втискиваться, вталкиваться, ввинчиваться* в узкую нору, пока не пролез в жилище Кролика". Вебер не только усложнил "характер повторности" (синонимичные повторы кратности 4), и сделал его неуловимым для детей, но и заставил усомниться в синонимичности ("ввинчивание" предполагает "вращение" — вариант немыслимый при заползании в нору).

Давая общую оценку переводу В. Вебера и Н. Рейн, можно сказать, что они, не сделав шага вперед, отодвинулись, как минимум, на два шага назад, исказив стиль оригинала и затруднив восприятие текста детьми.

Заключение

В данной работе техника L–граммного анализа текстов, разработанная нами ранее для изучения длинных неструктурированных последовательностей (типа ДНК–молекул), переносится на тексты естественного языка. Под L–граммой в этом случае понимается цепочка из L подряд следующих слов текста (или их основ). L–граммный анализ применим как к одному, так и к группе текстов. L–граммные частотно–позиционные характеристики содержат информацию о словарях L–грамм в каждом тексте ($L = 1, 2, 3$ и т.д.), их пересекаемости, частоте встречаемости и распределении по текстам каждой L–граммы.

С помощью L–граммного анализа проведено количественное исследование *сходства и различия параллельных текстов* (в данном случае переводов на русский язык известной книги Алана А. Милна "Винни–Пух") без предварительного их выравнивания. Прослежена динамика изменения частотно–позиционных характеристик с ростом L. Показана особая роль *"контрастных" L–грамм* (т.е. L–грамм, представленных преимущественно (или только) в одном из сравниваемых текстов) в выявлении *композиционных изменений* в параллельных текстах, *индивидуальных стилистических приемов* авторов переводов, а также проявлений *целенаправленного варьирования оригинала или уже имеющегося перевода*. Проведена *классификация* основных *схем варьирования*. Указаны возможности использования *позиционной информации* для разделения высокочастотной лексики на "служебную" и "тематическую".

По итогам исследования сделан вывод о том, что перевод В. Вебера и Н. Рейн, сделанный в 1999 г., реализует стратегию сознательного дистанцирования от уже имеющегося ("канонического") перевода Б. Заходера, что приводит в итоге к существенному искажению стиля оригинала и может затруднить восприятие текста детской аудиторией.

Литература

1. ГУСЕВ В.Д., САЛОМАТИНА Н.В. Определение и анализ ближайших окрестностей корней слов русского языка // Обнаружение эмпирических закономерностей. — Новосибирск, 1999. — Вып. 166: Вычислительные системы. — С. 80–103.

2. ГУСЕВ В.Д., САЛОМАТИНА Н.В. Электронный словарь паронимов: версия 1 // НТИ, сер 2. Информационные процессы и системы. — 2000. — №6. — С. 34–41.

3. ГУСЕВ В.Д., САЛОМАТИНА Н.В.Электронный словарь паронимов: версия 2 // НТИ, сер 2. Информационные процессы и системы. — 2001. — №7. — С. 26–33.

4. ГУСЕВ В.Д., САЛОМАТИНА Н.В. Количественные исследования вариативности языковых единиц // Тр. международной научно–практической конференции KDS – 2001. Т. 1. — С.–Петербург, 2001. — С. 186–193.

5. УОТЕРМЕН М.С. Выравнивание последовательностей // Математические методы для анализа последовательностей ДНК /Под ред. М.С. Уотермена. — М.: Мир, 1999. — С. 85–120.

6. ЗАРИПОВ Р.Х. Машинный поиск вариантов при моделировании творческого процесса. — М.: Наука, 1983. — 232 с.

7. BELL Timothy C., CLEARY John G., WITTEN Ian H. Text compression. — Prentice Hall, Inc. — 1990. — P. 317.

8. LYON Caroline, MALCOLM James, DICKERSON Bob. Detecting short passages of similar text in large document collection // Proc. Conf. On Empirical Methods in Natural Language Processing (EMNLP 2001). — Carnegie Mellon University, Pittsburg, USA. — 2001, June 3,4.

9. БОРИСЕНКО А. Песни невинности и песни опыта. О новых переводах "Винни–Пуха"// Иностранная литература. — 2002. — №4. — "Трибуна переводчика".

10. ГУСЕВ В.Д. Характеристики символьных последовательностей. // Машинные методы обнаружения закономерностей. — Новосибирск, 1981. — Вып. 88: Вычислительные системы. — С. 112–123.

11. FINDLER N. V., Van LEEUWEN J. A family of similarity measures between two strings // IEEE Trans. On Pattern Analysis and Machine Intelligence. — 1979. — Vol.PAMI–1, №1. — P. 116–118.

12. ГУСЕВ В.Д., НЕМЫТИКОВА Л.А., САЛОМАТИНА Н.В. Выявление аномалий в распределении слов или связанных цепочек символов по длине текста //Настоящий сборник. — С. 51–74.

Поступила в редакцию
14 января 2003 года