АНАЛИЗ L-ГРАММНЫХ СЛОВАРЕЙ ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ[1].

В.Д. Гусев, Н.В. Саломатина

Данная работа является продолжением наших исследований по количественной оценке вариативности языковых единиц разных иерархических уровней. Ранее объектами исследования выступали корни [1] и канонические формы слов [2], а также словосочетания, построенные на «игре слов» [3]. В этой работе мы анализируем тексты, являющиеся переводами одного и того же произведения (будем называть их параллельными). Интерес к объектам такого рода наблюдается не только у лингвистов, но и у биологов [4], музыковедов [5], специалистов в области информационного поиска [6]. Как отмечено в [7], 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 стиля.

1. L-граммные характеристики текста[2]

Пусть T — текст, N — длина текста в словоформах. Назовем L-граммой цепочку из L подряд следующих словоформ. Полное число L-грамм в тексте равно $N-L+1$. Обозначим число различных L-грамм через M_L ($M_L \le N-L+1$). *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, частота F_{L_i} ее встречаемости в тексте$>$. *Частотно-позиционной* назовем характеристику, где для каждой L-граммы указаны места ее вхождения в текст. *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.

При наличии двух текстов T_1 и T_2 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 $M_L(T_1,T_2)$ — 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 $\phi(T_1,T_2)$ = $\{\phi_1(T_1,T_2)$,

[1] Работа выполнена в рамках проекта № 03-06-80118, поддержанного грантом РФФИ

[2] Ранее система представления подобного типа использовалась нами при анализе слитных, не содержащих разделителей, текстов (ДНК-последовательности, знаменные песнопения), но цепочки формировались из символов, т.е. элементов более низкого иерархического уровня [8].

$\phi_2(T_1,T_2)$, …, $\phi_{L_{\max}(T_1,T_2)}(T_1,T_2)\}$, 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 основе характеристик $\phi_L(T_1)$, $\phi_L(T_2)$ и $\phi_L(T_1,T_2)$ можно вычислять различные теоретико-множественные меры близости между текстами T_1 и T_2 (см. [8]).

2. L-граммные характеристики оригинала и переводов

Рассматриваются два перевода на русский язык книги Алана А. Милна «Винни-Пух». Один принадлежит Б. Заходеру (1960 г., изд-во «Детский мир»), второй — В. Веберу и Н. Рейн (1999 г., изд-во «ЭКСМО-пресс»). Текст А. Милна (T_M) содержит 49269 словоупотреблений, переводы Б. Заходера (T_3) и В. Вебера(T_B), соответственно, 39806 и 43554 словоупотреблений.

Для устранения вариативности на уровне словоизменительной парадигмы каждая словоформа была заменена основой, получаемой путем усечения окончания. Реализация этой процедуры в автоматическом режиме не гарантирует 100 %-ной точности, но она и не требуется для достижения нашей цели. Последовательное наращивание длин повторяющихся цепочек эквивалентно расширению контекста, что позволяет устранять возникающие неоднозначности. Выделение L-грамм из текста проводилось скользящим окном, охватывающим L подряд следующих слов и сдвигающимся вдоль текста с шагом в одно слово. Знаки препинания игнорировались. Основы слов в L-связной цепочке отделялись друг от друга косой чертой.

Объемы словарей переводов оказались в среднем в 1,65 раза больше объема словаря оригинала. Высокочастотные 1-граммы включают служебную и общеупотребительную лексику, а также слова, отражающие тематику текста. Для разделения лексики на служебную и тематическую может быть использована позиционная информация. Общая закономерность состоит в том, что *слова*, *определяющие содержание текста*, как правило, *распределены* по нему *неравномерно* даже при относительно высокой частоте встречаемости. Максимальные расстояния между соседними вхождениями тематического слова обычно в несколько раз выше аналогичного показателя для служебного слова при сопоставимых частотах обоих слов. Максимальное разнообразие кратных L-грамм наблюдается при L=2. Доля общих цепочек двух переводов является существенной лишь для L=1,2,3.

Индивидуальные стилистические приемы или целенаправленное варьирование проявляют себя на уровне «контрастных» (т.е. представленных преимущественно в одном из текстов) L-грамм. Максимально контрастными по определению являются L-граммы из дополнений. Анализ контрастных L-грамм, примеры которых приведены в табл.1, позволил выделить следующие схемы варьирования, имевшие место при переводе оригинала:

1).*Переименования действующих лиц* (см. № 1÷4). К ним прибегают и Заходер и Вебер.

2). *Переход из одной системы мер и весов в другую* (см. № 5, Заходер).

3). *Смена лица*, от которого ведется изложение (см. № 6, Заходер).

4). *Свободный выбор звукоподражаний, восклицаний* (см. № 7÷9). И Заходер и Вебер не слишком придерживаются оригинала. Пример в строке 9 является образцом целенаправленного варьирования.

5). *Синонимические замены* (см. № 10÷21, 25, 26). Этот класс преобразований включает в себя:

a) *замены* (см. № 10, 11, 16, 19, 20), *вставки*, *делеции* (№ 15), а также *перестановки слов* (№ 13);

b) *локальное варьирование внутри слов* (см. № 14), в частности, использование *уменьшительных форм* (характерно для Вебера (см. № 18)) и *различных схем намеренного искажения слов*

(№ 25);

с) *замены* одних *словосочетаний* другими (см. № 12, 21).

6). *Разбиения* (замена часто встречающихся слов группами близких по смыслу — см. № 22, 23).

7). *Объединения* (замена близких по смыслу слов одним единственным). (см. № 24).

8). *Варьирование без сохранения позиционных соответствий* (см. № 26, 27). Речь идет об индивидуальных стилистических приемах синонимичных по смыслу, но позиционно не согласованных.

Какие же соображения, подтверждающие, что Вебер делал свой перевод «с оглядкой» на Заходера, можно привести по итогам рассмотрения табл.1?

(1) Удивляет «вето», наложенное Вебером на использование слова «здравствуй». Возможная причина состоит в том, что это слово часто используется Заходером для приветствий (см. № 16). (2) Слово «thistle(s)» в словарях и у Заходера переводится как «чертополох» (см. № 10), у Вебера — «репейник», а это обобщающее название для группы колючих растений (терн, лопух, чертополох). Возможная причина — сделать не так, как у Заходера.

(3) Словосочетание «friends and relations » (см. № 13) Заходер переводит как «родственники и знакомые» и «родные и знакомые», тем самым ограничивая выбор еще одного синонима для «relations». Видимо поэтому Вебер использует устаревшую форму «родичи», частота встречаемости которой в современных текстах на порядок уступает варианту Заходера.

(4) Выражение «the Spotted or Herbaceous» (дословно: «пятнистый или травянистый») Заходер, отступая от оригинала, перевел как «пятнистый или травоядный» (см. № 14). Вебер, мог бы вернуться к оригиналу, но он, продолжая полемизировать с Заходером, придумывает термин «цветоядный». Аналогичная ситуация возникает с глаголом «наскакивать» (у Заходера), вместо которого Вебер пользуется собственным — «напрыгивать» (см. № 17).

(5) Особенно трудным оказался для Вебера подбор синонимичных выражений для фразеологизмов (Заходер: «остановился как вкопанный» ⇔ Вебер: «застыл *как памятник*»; Заходер: «побежал домой, что было духу» ⇔ Вебер: «*со всех лап* бросился к своему домику»).

Рассмотренные в данном разделе отличия двух переводов позволяют сделать следующие

Выводы

1). 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 привносить своего» [7]). Однако 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 и хорошо выявляемых с помощью позиционного анализа, можно сказать, что он сделал максимально «*русифицированный*» и *ориентированный на детскую аудиторию* перевод «Винни-Пуха», *сохранив стилистические приемы и находки Милна*.

Табл.1 Соответствия контрастных L-грамм (помечены звездочкой) и их аналогов*

№	Заходер			Вебер			Милн	
	L-грамма	$F_З$	F_B	L-грамма	$F_З$	F_B	L-грамма	F_M
1	Пятачок/	425	4	Хрюка/*	0	442	Piglet/	553
2	Тигр/(а)	222	0	Тигер/*	0	224	Tigger/	183
3	Слонопотам/	14	0	Хоботун/*	0	35	Heffalump/	46
4	Бук/(а, и)*	11	0	Вузл/(а)	0	6	Woozle/(s)	11
5	Метр/	5	0	Фут/*	0	8	Foot/	6
6	Пап/(а)*	12	0	Я/	646	638	I/	877
7	Тирлим-бом-бом/*	14	0	Трам-пам/	0	12	Tiddely-poom/	20
8	(ворра) ^{5} *	2	0	(ворра) ^{4}	0	2	(worra) ^{5}	2
9	Аа/*	19	1	- (ах, да, понятно,…)			Oh/	134
10	Чертополох/*	13	0	Репейник/	1	18	Thistle/(s)	10
11	Тревожн/ Испуганн/	7 4	0 0	Озабоченн/*	0	14	Anxiously/	11
12	Опилк/в/голов/ Глупеньк/	7 7	0 1	Слабеньк/умишк/*	0	11	Very/little/brain/ Silly/old/Bear/	8 6

13	Родственник/ И/знаком/*	9	0	Друз/и/родич/	0	9	Friends/and/ Relations	11
14	Пятнист/или/ Травоядн/*	4	0	Пятнаст/или/ Цветоядн/	0	3	The/Spotted/or/ Herbaceous/	3
15	Очень/маленьк/ Существ/*	9	0	Очень/маленьк/	0	7	a/very/ small/animal/	8
16	Здравствуй/*	26	0	Привет/	15	59	Hallo/ Good/morning/	57 14
17	Наскакива/ Наскочи/	7 7	2 1	Напрыгива/(ть, л)*	0	16	Bounced/ Bounsing/	13 6
18	Шар/*	26	2	Шарик/	17	41	Balloon/	39
19	Подкрепи/(ться)*	14	2	Перекуси/ Перехвати/	0 0	5 1	-	
20	Хитр/западн/(я)	3	0	Хитрумн/западн/(я)*	0	4	Cunning/trap/	3
21	Дремуч/лес/	10	0	Столетн/лес/*	0	12	Hundred/acr/wood/	13
22	Сказа/*	934	136	Сказа/ Воскликну/ Ответи/ Поздорова/ Вырва/(лось)	934 1 30 3 0	136 57 157 26 12	Said/	1259
23	Немножко/*	24	1	- (немного,немно- жечко,несколько,…)			A/little/(something, Futher, Longer, …)	8 4 4
24	Шумелк/ Кричалк/ Ворчалк/	13 9 3	1 0 0	Бубнилк/*	0	18	A/noise A/…/hum/ …/murmured	29 14 4
25	Икспедицию/	4	0	Икшпедицию/*	0	6	Expotition/	10
26	Очень-очень/	14	2	Очень/даже/	0	14	-	
27	Например/*	12	0	К/пример/(у)*	0	10	-	

* Примечание: а) F_M , F_3 и F_B — частоты L-грамм, соответственно, в T_M , T_3 и T_B ;

б) прочерк в графе означает отсутствие устойчивой конструкции.

3). Вебер в стремлении «сделать не так, как у Заходера» вернул в свой текст всю англоязычную лексику, полностью проигнорировав желание Заходера «выучить Винни и его друзей объясняться по-русски». Дистанцируясь от Заходера, он невольно искажает стиль Милна, ориентированный на детское восприятие. Наиболее ярко это проявилось на схеме варьирования типа «разбиение» (см. в табл.1 далеко не полный список не всегда синонимичных вариантов перевода Вебером глагола «said»).

Заключение

С помощью L-граммного анализа проведено количественное исследование сходства и различия параллельных текстов (переводов на русский язык 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 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 реализует стратегию сознательного дистанцирования от ставшего «каноническим» 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 2 // НТИ, сер.2. Информационные процессы и системы. — 2001. — № 7. — С. 26-33.
3. Гусев В.Д., Саломатина Н.В. Количественные исследования вариативности языковых единиц // Труды международной научно-практической конференции KDS-2001. Т. 1. — С.-Петербург, 2001. — С. 186-193.
4. М.С. Уотермен. Выравнивание последовательностей // В кн. «Математические методы для анализа последовательностей ДНК» (под ред. М.С. Уотермена). — М.: Мир, 1999. — С. 85-120.
5. Р.Х. Зарипов. Машинный поиск вариантов при моделировании творческого процесса. — М.: Наука, 1983. — 232 с.
6. Caroline Lyon, James Malcolm, Bob Dickerson. 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. — June 3,4 2001.
7. А. Борисенко. Песни невинности и песни опыта. О новых переводах «Винни-Пуха». // Иностранная литература. — 2002. № 4. — «Трибуна переводчика».
8. Гусев В.Д. Характеристики символьных последовательностей. // Машинные методы обнаружения закономерностей. — Новосибирск, 1981. — Вып. 88: Вычислительные системы. — С. 112-123.