

# ЭКСПРЕСС-МОНИТОРИНГ СТИЛЕМЕТРИЧЕСКИХ ПОКАЗАТЕЛЕЙ ПЕРЕВОДНЫХ ТЕКСТОВ

Русяева Е.Ю., Ахобадзе Г.Н.

*Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия*

rusyaeva@ipu.ru, akhobadze@ipu.ru

*Аннотация. Представлен метод экспресс-мониторинга лингвистического анализа переводных письменных текстов как часть разработки по управлению крупномасштабной информационно-аналитической системой. Этот метод основан на сопоставлении стилистических показателей, при этом в качестве маркеров идентификации выступают служебные слова. Выявленные маркеры характеризуют идиостилевые особенности переводных текстов, а расчет их весовых коэффициентов и обобщенный расчет объема слов использованы для сравнительного анализа перевода текстов «ручным» и машинным способами.*

*Ключевые слова: крупномасштабная информационно-аналитическая система, переводной текст, стилеметрические показатели, лингвистический анализ, весовые коэффициенты, идиостилевые черты текста.*

## Введение

В современном мире невозможно обсуждать управление развитием без рассмотрения вопроса о создании крупномасштабных информационно-аналитических систем. Эти системы включают в себя инструменты для лингвистического анализа текстов и их перевода. Сейчас традиция перевода текстов (в том числе и аудио-) лишь человеком-переводчиком без помощи технических средств постепенно сходит на нет. С началом XXI века начали появляться и крупномасштабные сервисы автоматического перевода текстов, в том числе и бесплатные. К 2016 году прорыв в разработке нейросетей привел к улучшению автоматизации переводов, появились сервисы нейронного машинного перевода (НМП). Поскольку системы НМП имитируют работу нейронных сетей мозга человека, это позволяет им вычислять статистически наиболее вероятный вариант перевода на основе «обучающего набора данных» (ОНД). Так машинное обучение ведется на основе большого набора предложений на исходном языке и их перевода, выполненного человеком [1]. Но, как подчеркивают лингвисты [2], полностью автоматический перевод текстов не сможет заменить собой профессиональных переводчиков. В целом, как и во всей очень модной современной тенденции применения систем искусственного интеллекта (ИИ), в частности, GhatGPT, пока именно созданные людьми тексты составляют основу обучающей выборки, как и в случае с переводными текстами, так как без помощи людей переводчиков, редактирующих автоматические переводы, для их адекватной интерпретации, получается, что именно люди по-прежнему дают необходимые данные для машинного обучения (МО) и придают легитимность переводам.

Мы полностью разделяем эту мысль и отмечаем, что пока задачи исследований качества машинного и «человеческого» (далее будем называть такой перевод «ручным») текстов формальными методами специально не ставились [3]. Целью данного исследования, в развитие и продолжение темы [4, 5], является формальное доказательство различий текстовых переводов, выполненных в «ручном» (человеком) и автоматическом режиме. Комбинированный способ, когда человек переводит текст с помощью автоматического перевода или редактирует машинный текст, мы будем также условно относить к «ручному» переводу, кстати, это способ сейчас наиболее распространен, поскольку оперативен и эффективен. Применяемый нами метод заключается в сопоставлении и выяснении стилеметрических показателей двух письменных переводов с английского языка на русский язык текстов художественного и научного (гуманитарного направления) стилей.

Письменный перевод текста заключается фактически в создании нового текста на основе оригинала в текстовом варианте с сохранением первоначального смысла, логики построения, стилистики и терминологии [3]. А, как известно, наиболее удобным стилеметрическим инструментом и показателем идеостилевого авторского профиля [6, 7, 8] могут быть частоты служебных слов в текстах литературно-художественного стиля, на эту тему есть много интересных исследований [9, 10, 11]. Считаем, что особенно это касается текстов художественного стиля, где задачей переводчика является наиболее точно донести и авторский замысел [4], и идеостилевые авторские особенности.

То, что любому авторскому тексту (в том числе и переведенному) присуща некая авторская индивидуальность, не вызывает сомнения. Более того, даже переведенные переводчиком тексты тоже становятся в определенном смысле «уникальными», а автоматические переводы, выполненные без помощи редактирования переводчиков, так и не становятся полностью готовыми для восприятия читателя [2]. В нашу задачу как раз и входит мониторинг стилеметрических показателей,

определенных нами ранее как маркеры идентификации авторской текстовой индивидуальности [4], чтобы не только выявлять формальные показатели различий «ручного» и машинного перевода, но и делать это более оперативно, то есть проводить экспресс-мониторинг. Этот метод входит в состав и является частью разрабатываемой нами крупномасштабной вычислительной информационно-аналитической системы. В наши дальнейшие планы входит разработка полного комплекса методов лингвистического анализа различных текстов.

## 1. Теоретическая часть

В качестве маркеров идентификаторов нами были взяты такие наиболее часто встречающиеся в тексте простые, односоставные, непрямые служебные части речи в целях избегания омонимии. Этот выбор был подтвержден многочисленными исследованиями [6, 8, 9, 10, 11], согласно которым, наиболее часто встречающимися в текстах являются служебные (предлоги, союзы, частицы), а не знаменательные (существительные, глаголы, прилагательные) части речи.

Задача проверки работы метода на множестве переводных текстов пока не ставилась, мы только ищем подходы автоматического лингвистического анализа для повышения точности и упрощения самого процесса вычисления, пока говорим лишь о грамматическом качестве переводов на основании некоторых расчетов выявленных нами лингвистических особенностей. Глубокие лингвистические исследования русского языка [11, 12] подтверждают, что некоторые характерные особенности склада речи автора можно различать и по служебным частям речи – предлогам, союзам, частицам. Они отвечают не только за синтаксическую связность слов в предложении, и, хотя вроде бы по определению не имеют собственного лексического значения, тем не менее, в литературных текстах отчасти служат лексико-семантическими и стилистическими маркерами, некими показателями авторского стиля речи, построения текста наряду с синтаксическим разбиением и интонированием. Вот почему мы в разрабатываемом нами стилеметрическом методе анализа переведенных текстов ориентировались на рассмотрение частоты повторения служебных частей речи, предлогов, союзов и частиц. Причем, согласно предварительным анализам наиболее частотны и, следовательно, более показательными для анализа стали предлоги «в», «с» и «на», союзы «и», «а» и «но» и частицы «не», «бы» и «же».

Нами были детально проанализированы переводы произведений известных англоязычных авторов: английского писателя Джорджа Оруэлла, его знаменитый роман «1984» (написан в 1949 году), и научный текст американского писателя и бизнес-консультанта (современного гуру менеджмента) Адизеса Ицхака Калдерона «Управление в условиях кризиса».

Анализ переведенных текстов предполагает вычисление отношений  $n_{1в}/N_1 = K_{1в}$ ,  $n_{1на}/N_1 = K_{1на}$  и  $n_{1с}/N_1 = K_{1с}$ , где  $n_{1в}$  – количество предлогов «в» в тексте «ручного» перевода,  $n_{1на}$  – количество предлогов «на» в тексте «ручного» перевода,  $n_{1с}$  – количество предлогов «с» в тексте «ручного» перевода.  $N_1$  – это количество слов в тексте «ручного» перевода;  $n_{1и}/N_1 = K_{1и}$ ,  $n_{1а}/N_1 = K_{1а}$  и  $n_{1но}/N_1 = K_{1но}$ , где  $n_{1и}$  – количество союзов «и» в тексте «ручного» перевода,  $n_{1а}$  – количество союзов «а» в тексте «ручного» перевода,  $n_{1но}$  – количество союзов «но» в тексте «ручного» перевода. Далее рассчитаем  $n_{1не}/N_1 = K_{1не}$ ,  $n_{1бы}/N_1 = K_{1бы}$  и  $n_{1же}/N_1 = K_{1же}$ , где  $n_{1не}$  – количество частиц «не» в тексте «ручного» перевода,  $n_{1бы}$  – количество частиц «бы» в тексте «ручного» перевода,  $n_{1же}$  – количество частиц «же» в тексте «ручного» перевода. В соответствии с этими данными, вычисляем, соответственно, весовые коэффициенты  $K_{1в}$ ;  $K_{1на}$ ;  $K_{1с}$ ;  $K_{1и}$ ;  $K_{1а}$ ;  $K_{1но}$  и  $K_{1не}$ ;  $K_{1бы}$ ;  $K_{1же}$  выше указанных предлогов, союзов и частиц «ручного» перевода текста. Поясним, что с целью корректности анализа двух переведенных текстов, словесный объем переводимого текста при переводах остается неизменным.

Далее аналогичным образом подсчитывается количество предлогов  $n_{2в}$ ,  $n_{2на}$ ,  $n_{2с}$  в тексте машинного перевода; союзов  $n_{2и}$ ,  $n_{2а}$ ,  $n_{2но}$  в тексте машинного перевода; частиц  $n_{2не}$ ,  $n_{2бы}$ ,  $n_{2же}$  в тексте машинного перевода. Затем вычисляем весовые коэффициенты предлогов в тексте машинного перевода, соответственно,  $K_{2в}$ ;  $K_{2на}$ ;  $K_{2с}$ ; союзов в тексте машинного перевода –  $K_{2и}$ ;  $K_{2а}$ ;  $K_{2но}$  и частиц в тексте машинного перевода –  $K_{2не}$ ;  $K_{2бы}$ ;  $K_{2же}$ .

После вычисления весовых коэффициентов служебных частей речи (предлогов, союзов и частиц) в «ручном» и машинном переводах, можно произвести сравнение вычисленных однотипных весовых коэффициентов. Другими словами, в данном случае сравнению подвергаются коэффициенты:  $K_{1в}$ ,  $K_{2в}$ ;  $K_{1на}$ ,  $K_{2на}$ ;  $K_{1с}$ ,  $K_{2с}$ ;  $K_{1и}$ ,  $K_{2и}$ ;  $K_{1а}$ ,  $K_{2а}$ ;  $K_{1но}$ ,  $K_{2но}$ ;  $K_{1не}$ ,  $K_{2не}$ ;  $K_{1бы}$ ,  $K_{2бы}$ ;  $K_{1же}$ ,  $K_{2же}$ . Кроме того, при сопоставлении переведенных текстов оцениваются количество слов и средние значения весовых коэффициентов по предлогам, союзам и частицам  $C_{ср.п}$ ;  $C_{ср.с}$ ;  $C_{ср.ч}$ .

В ряде случаев для удобства обозначения размерности весовых коэффициентов и их средних значений (как относительных величин), они могут быть выражены в процентах. Для этого числовые значения весовых коэффициентов следуют умножить на 100 %.

## 2. Практическая часть

### 2.1. Текст художественного стиля

Как указано выше, к анализу были привлечены переводные тексты художественного и научного стиля, переводные «ручным» (в том числе и комбинированным) и машинным (только с помощью автоматических серверов перевода текстов) способами.

Художественное произведение: роман «1984» Джорджа Оруэлла, переведенный на русский язык («ручной» перевод, переводчик Юлия Шматько). Словарный объем текста (число слов в документе, статистика 6550. Это текст был сокращен для удобства анализа, полный текст «ручного» перевода) – N – в оригинале на английском языке 104160 слов / полный текст перевода N 82886 слов/.

Таблица 1. Текст романа «1984» Дж. Оруэлла, переведенный на русский язык («ручной» перевод)

№	Словарный объем текста (число слов в документе, статистика) – N	Частота появления предлогов			Весовые коэффициенты предлогов – K		
		в	на	с	K <sub>в</sub>	K <sub>на</sub>	K <sub>с</sub>
1	65550	1803	1020	670	0,0275	0,0155	0,0101
		Частота появления союзов			Весовые коэффициенты союзов – K		
		и	а	но	K <sub>и</sub>	K <sub>а</sub>	K <sub>но</sub>
		2006	191	479	0,0306	0,0029	0,0073
		Частота появления частиц			Весовые коэффициенты частиц – K		
		не	бы	же	K <sub>не</sub>	K <sub>бы</sub>	K <sub>же</sub>
		1248	212	144	0,0190	0,0032	0,0021

Роман «1984» Джорджа Оруэлла, переведенный на русский язык с оригинала (текст на английском языке, оригинал, взят также с тем же количеством слов: **65550**, как и отрывок «ручного» перевода. При машинном переводе количество слов уменьшилось, стало **53375**.)

Отмечаем, что при переводе количество слов уменьшается. Машинный перевод романа «1984».

Таблица 2. Текст романа «1984» Дж. Оруэлла, переведенный на русский язык (машинный перевод)

№	Словарный объем текста (число слов в документе, статистика) – N	Частота появления предлогов			Весовые коэффициенты предлогов – K		
		в	на	с	K <sub>в</sub>	K <sub>на</sub>	K <sub>с</sub>
1	53375	1463	820	530	0,0274	0,0153	0,0099
		Частота появления союзов			Весовые коэффициенты союзов – K		
		и	а	но	K <sub>и</sub>	K <sub>а</sub>	K <sub>но</sub>
		1509	194	347	0,0282	0,0036	0,0065
		Частота появления частиц			Весовые коэффициенты частиц – K		
		не	бы	же	K <sub>не</sub>	K <sub>бы</sub>	K <sub>же</sub>
		993	277	114	0,0186	0,0051	0,0021

По результатам табличных данных можно заключить, что, во-первых, количество слов в машинном переводе значительно меньше, чем в «ручном» переводе, и эти цифры при одном и том же взятом исходном тексте для переводимого текста составляют: 65550 («ручной» перевод) и 53375 – машинный перевод. Во-вторых, по весовым коэффициентам всех служебных частей речи в переведенных тестах наблюдается различие, за исключением случая весовых коэффициентов  $K_{1же} = 0.0021$  («ручной» перевод) и  $K_{2же} = 0,0021$  (машинный перевод). Из сравнения средних значений всех служебных частей речи видно, что при ручном переводе средние значения весовых коэффициентов по предлогам и союзам ( $C_{ср.п} = 0.0177$ ;  $C_{ср.с} = 0.0136$ ) больше чем в машинном переводе ( $C_{ср.п} = 0.0174$ ;  $C_{ср.с} = 0.127$ ). С другой стороны, среднее значение частиц в машинном переводе ( $C_{ср.ч} = 0,0086$ ) больше, чем в ручном переводе ( $C_{ср.ч} = 0.0081$ ).

Также выявлены максимальное и минимальное значения весовых коэффициентов: «ручной» перевод – весовой коэффициент по союзам  $K_{1и} = 0.0306$ ; минимальное значение при ручном и машинном переводах по частицам  $K_{1,2же} = 0.0021$ .

Большее суммарное употребление предлогов «в», «на» и «с», а также союзов «и», «а» и «но» в «ручном» переводе (1,77%, (1,36%) по отношению к машинному переводу (1,74%, 1,27%) и меньшее суммарное употребление частиц «не», «бы» и «же» в этих же переводах, соответственно (0,81%, 0,86%), указывает на детерминированность процедур переводов, связанную с характером процедуры переводов по разным, определенным задачи, алгоритмам. Этим же можно объяснить почти наличие разных значений вычисленных однотипных весовых коэффициентов служебных частей речи, характеризующих ручной и машинный переводы.

С позиции лингвистического анализа можно сделать вывод, что «ручной» перевод более широко использует стилеметрические показатели и, следовательно, стремится к большей точности в передаче идеостилевого авторского профиля, показателем которого служат выделенные нами служебные части речи как идентификационные маркеры.

## 2.2. Текст научного стиля (гуманитарного направления)

Текст Адизеса Ицхака Калдерона «Управление в условиях кризиса. Как выжить и стать сильнее». Серия «Методология Адизеса» (отрывок из «ручного» перевода английского текста на русском языке (перевод выполнен переводчиками издательства «Альпина Паблишер», число слов –15590). Полный тест «ручного» перевода: N – 76554 слов. Полный текст оригинала на английском языке – 89790 слов). Также отмечается уменьшение числа слов при переводе с оригинала.

Таблица 3. Текст И.Адизеса «Управление в условиях кризиса. Как выжить и стать сильнее», переведенный на русский язык («ручной» перевод)

№	Словарный объем текста (число слов в документе, статистика) – N	Частота появления предлогов			Весовые коэффициенты предлогов – K		
		в	на	с	в	на	с
1	15590	367	199	114	0,0235	0,0127	0,0073
		Частота появления союзов			Весовые коэффициенты союзов – K		
		и	а	но	и	а	но
		422	107	75	0,0270	0,0068	0,0048
		Частота появления частиц			Весовые коэффициенты частиц – K		
		не	бы	же	не	бы	же
310	26	32	0,0198	0,0016	0,0020		

Машинный перевод текста Адизеса Ицхака Калдерона «Управление в условиях кризиса» (машинный перевод отрывка оригинального текста в 15590 слов составил 14700 слов).

Таблица 4. Текст И.Адизеса «Управление в условиях кризиса. Как выжить и стать сильнее», переведенный на русский язык (машинный перевод)

№	Словарный объем текста (число слов в документе, статистика) – N	Частота появления предлогов			Весовые коэффициенты предлогов – К		
		в	на	с	$K_v$	$K_{на}$	$K_c$
1	14700	347	156	117	0,0236	0,0106	0,0079
		Частота появления союзов			Весовые коэффициенты союзов – К		
		и	а	но	$K_{и}$	$K_a$	$K_{но}$
		416	91	82	0,0282	0,0061	0,0055
		Частота появления частиц			Весовые коэффициенты частиц – К		
		не	бы	же	$K_{не}$	$K_{бы}$	$K_{же}$
314	29	36	0,0213	0,0019	0,0024		

Результаты данных таблиц 3 и 4 тоже показывают тенденцию уменьшения количества слов при «ручном» и машинном переводах («ручной» 15590, машинный – 14700). В этом случае по весовым коэффициентам всех служебных частей речи в переведенных тестах наблюдается различие. Согласно этим табличным данным для средних значений всех служебных слов оказывается, что при машинном переводе средние значения весовых коэффициентов союзов и частиц больше ( $C_{ср.с} = 0.0132$ ;  $C_{ср.ч} = 0.0085$ ) чем при «ручном» переводе ( $C_{ср.с} = 0.0128$ ;  $C_{ср.ч} = 0.0078$ ). А при «ручном» переводе, наоборот, среднее значение предлогов ( $C_{ср.в} = 0.0145$ ) больше, чем в машинном переводе ( $C_{ср.в} = 0.0140$ ). Что касается максимальных и минимальных значений весовых коэффициентов, то, по результатам таблиц, имеем: максимум  $K_{2c} = 0.0282$  (машинный перевод); минимум  $K_{1бы} = 0.0016$  («ручной» перевод).

Большее суммарное употребление союзов «и», «а» и «но», а также частиц «не», «бы» и «же» в машинном переводе (1,32%, (0.85%) по отношению к «ручному» переводу (1,28 %, 0.78%) и меньшее суммарное употребление предлогов «в», «на» и «с» в этих же переводах, соответственно (1,45%, 1.40%), указывает на детерминированность переводов, связанную с определенным характером процедуры переводов разными алгоритмами. Этим же можно объяснить наличие почти разных значений вычисленных однотипных весовых коэффициентов служебных частей речи, характеризующих «ручной» и машинный переводы.

Проведенный формальный анализ может служить также подтверждением, что «ручной» перевод более точен в передаче идиостилевого авторского профиля, показателем которого служат указанные служебные части речи как идентификационные маркеры.

### 3. Заключение

В исследовании проведен экспресс-мониторинг стилиметрических показателей переводных текстов литературно-художественного и научного (гуманитарного направления) стилей, переведенных «ручным» и машинным способами по наиболее частотным, определенных нами как маркеры, служебным частям речи. Анализ включает определение весовых коэффициентов этих маркеров, вычисляемых отношениями количества соответствующих служебных слов в тексте к общему объему слов рассматриваемого текста.

Разумеется, сравнительный анализ стилиметрических показателей, а именно, использование выделенных нами служебных слов как маркеров идентификации в переводных текстах Джорджа Оруэлла, роман «1984», и научного текста Адизеса Ицхака Калдерона «Управление в условиях кризиса», может быть продолжен. Однако приведенные примеры, по нашему мнению, достаточны для того, чтобы констатировать, что проведенный расчетный экспресс-мониторинг позволяет выявить некоторые лексические особенности в строении текстового перевода. Выделенные нами маркеры в целом характеризуют идиостилевые черты переводного текста и при переводах отчасти могут служить формальными показателями того, что «ручной» способ переводов достигает большей точности и, соответственно, лучшего грамматического качества.

## Литература

1. Переводчик – профессия на грани исчезновения. Электронный ресурс: <https://courier.unesco.org/ru/articles/perevodchik-professiya-na-grani-ischeznoveniya> (Дата обращения 25.05.2024).
2. Translation: from one world to another The UNESCO Courier April-June 2022 Электронный ресурс: (Дата обращения 20.05.2024).
3. Как проверить качество переводного текста Электронный ресурс: <https://livion.ru/news/kak-proverit-kachestvo-perevedennogo-teksta>. <https://livion.ru/news/vidy-pismennogo-perevoda> (Дата обращения 24.05.2024).
4. *Rusyaeva E.* Conceptual Approach in Structuring the Semantic Core Author's Text // Proceedings of the 16th International Conference Management of Large-Scale System Development (MLSD). Moscow: IEEE, 2023. С. <https://ieeexplore.ieee.org/document/10303783>.
5. *Rusyaeva E., Akhobadze G., Poltavsky A.* Fragment Analysis and Weighting Coefficients for Intelligent Text Monitoring // Proceedings of the 16th International Conference Management of Large-Scale System Development (MLSD). Moscow: IEEE, 2023. С. <https://ieeexplore.ieee.org/document/10304005>.
6. *Михеев М.Ю., Эрлих Л.И.* Идиостилевой профиль и определение авторства текста по частотам служебных слов // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2018. №2. С. 25-34
7. *Михеев М.Ю., Эрлих Л.И.* Текстовые скрепы и их частоты как различительный признак авторских идиостилей Электронный ресурс. <https://ruslang.ru/sites/default/files/doc/grigoriev2022/Mish.pdf> (Дата обращения 15.03.2024).
8. *Орлова М.В., Михеев М.Ю., Эрлих Л.И.* Об отличиях русского научного идиостиля от художественного по частотам употребления текстовых скреп/ Вопросы литературы. 2022. № 1. С.118-140.
9. *Кукушкина О.В., Поликарпов А.А.* Частотные и распределительные характеристики русских предлогов и синтаксем, с ними связанных (по «Ядерному компьютерному корпусу текстов русских газет конца XX-ого века») // Язык, сознание, коммуникация. Выпуск 47. М.: МАКС Пресс. том 47. 2013. с. 341-362
10. *Всеволодова М.В., Кукушкина О.В., Поликарпов А.А.* Русские предлоги и средства предложного типа. Материалы к функционально-грамматическому описанию реального употребления. Книга 1. Введение в объективную грамматику и лексикографию русских предложных единиц. М.: URSS, 2013. 304 с.
11. *Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С.* Автоматическая обработка текстов на естественном языке и анализ данных. М.: НИУ ВШЭ. 2017. 269 с.
12. *Потебня А.А.* Мысль и язык. Избранные работы / А. А. Потебня. – М.: Юрайт, 2023. – 238 с.