

ОБЪЕКТИВНЫЕ ПАРАМЕТРЫ СУБЪЕКТИВНОЙ ОЦЕНКИ КАЧЕСТВА ПЕРЕВОДА

Л. В. Елизарова, К. К. Дмитриева

Российский государственный педагогический университет имени А. И. Герцена

OBJECTIVE PARAMETERS FOR SUBJECTIVE EVALUATION OF TRANSLATION QUALITY

L. V. Elizarova, K. K. Dmitrieva

Herzen State Pedagogical University of Russia

Аннотация: статья посвящена изучению современных технологий оценки качества перевода. В разное время в зависимости от вектора теоретических исследований, потребностей индустрии перевода и задач образовательной деятельности в центре внимания находились различные аспекты качества перевода. Неизменным оставался интерес к изучению критериев оценки, обусловленный стремлением к выявлению универсальных, максимально объективных критериев оценивания. Экспертная оценка, выполняемая человеком вручную с опорой на типологию ошибок, долгое время оставалась единственным способом оценивания результатов перевода. Под влиянием жестких требований современного общества к оперативной передаче больших объемов информации возникает потребность в оценке качества материалов, выполненных с помощью машинного перевода. В связи с этим происходит модификация экспертной оценки и начинается активная разработка и внедрение автоматических метрик. Эволюция экспертных технологий в профессиональной сфере – от формальных средств оценивания до систем, нацеленных на оценку текстов различной жанровой принадлежности – предопределяет необходимость комплексного изучения практических решений в тесной взаимосвязи с теоретическими положениями. С целью преодоления недостатков, связанных, прежде всего, со степенью объективности оценки, на практике все чаще предпринимаются попытки усовершенствовать набор оценочных параметров, апробировать возможные варианты их комбинаций. Поиск оптимальных решений приводит к возникновению различных версий систем оценивания. В статье рассматривается типология «Многомерных метрик качества», в которой отражается актуальный на данный момент подход к систематизации ошибок по категориям и критериям критичности. Свою версию указанной типологии недавно представила компания Google. Цель исследования заключается в выявлении отличительных особенностей традиционной и обновленной версий типологии. Итоги сравнительно-сопоставительного анализа ключевых параметров типологии, а также результаты ее апробации при оценке переведенных материалов компании PROMT позволили описать влияние произведенных модификаций на степень объективности оценки качества перевода. Вопрос о целесообразности использования нескольких версий одной системы оценивания остается открытым для дальнейшего изучения.

Ключевые слова: качество перевода, экспертная оценка, типология «Многомерных метрик качества», типология ошибок, параметры оценивания, критичность ошибок.

Abstract: the article explores modern tools of translation quality evaluation. Various aspects of translation quality became the focus of attention at different periods of time depending on the approach of research, the needs of the translation industry, and the educational tasks. The interest in the study of evaluation criteria has remained unchanged as the researchers strive to identify universal and the most objective evaluation criteria possible. For a long time, manual evaluation by experts based on the error typology remained the only way to evaluate the translation quality. Modern society demands rapid transmission of large volumes of information, thus creating a



need to assess the quality of materials produced by machine translation. These changes lead to modification of the expert evaluation and active development and implementation of automatic metrics. The evolution of expert tools – from formal evaluation tools to systems aimed at evaluating texts of different genres – calls for a comprehensive study of practical solutions in close interrelation with theoretical provisions. In order to overcome the disadvantages, which are primarily associated with the degree of evaluation objectivity, researchers make an increasing number of attempts to improve the set of evaluation parameters and to test their possible combinations. The search for the optimal solutions brings about various versions of evaluation systems. The article considers the Multidimensional Quality Metrics typology, which represents the cutting edge approach to error systematisation by categories and severity criteria. Google has recently presented its version of this typology. The study aim is to identify the distinctive features of the traditional and the updated versions of the typology. We compared their key parameters and tested the typology in evaluation of the materials translated by PROMT. The results revealed the impact of the modifications on the degree of objectivity in translation quality evaluation. Further study is necessary to determine whether it is reasonable to use several versions of one evaluation system.

Key words: translation quality, expert evaluation, Multidimensional Quality Metrics typology, error typology, evaluation parameters, error severity.

Введение

В настоящее время *качество перевода* рассматривается как одно из ключевых понятий теории и практики перевода. Вопросы оценки качества переведенных материалов находятся в центре внимания исследователей, преподавателей и представителей переводческой отрасли. В рамках каждого из указанных направлений – научном, дидактическом и прикладном соответственно – решаются разные задачи. Но все они объединены общей целью, определяемой как достижение или выполнение качественного перевода, соответствующего объективным параметрам оценивания. Единство цели предполагает многообразие путей решения, каждый из которых включает, прежде всего, определение критериев оценки, варианты их комбинаций в зависимости от вида перевода, условий его реализации и многих других факторов. В стремлении улучшить качество перевода исследователи предлагали различные классификации оценочных параметров, что привело к значительному увеличению числа критериев и к затруднениям в их использовании. При этом общее свойство всех имеющихся подходов заключалось в том, что в них «дается ряд норм, критериев и требований, которые переводчик должен соблюдать» [1, с. 126].

Наиболее изученными и оптимальными в количественном отношении представляются принципы оценки качества перевода, выработанные в рамках лингвистического и коммуникативно-функционального подходов. Но независимо от выбранной концепции исследователи всегда стремились выявить соотношение между адекватностью и эквивалентностью как экспертными категориями оценки. Дифференциация указанных категорий рассматривается как условие для дальнейшего выделения критериев, связанных с понятием «переводческая ошибка» [2, с. 102]. На современном этапе развития переводоведения сформировалось достаточно четкое понимание важности параметров, определяемых переводческой

эквивалентностью, для текстоцентрической концепции, с одной стороны, и приоритета категории адекватности при оценке перевода с учетом условий коммуникативной ситуации – с другой [2; 3]. Исследователи А. В. Соснин, Ю. В. Балакина, А. Н. Кащин также дифференцируют данные категории, определяя «процессуальную сущность» эквивалентности, другими словами, ее ориентированность на результат перевода. Адекватность, по мнению авторов, является дискурсивной по своей сути и «указывает на соответствие процесса перевода данным коммуникативным условиям» [1, с. 125]. Необходимость включения параметров коммуникативной ситуации для объективной оценки качества перевода убедительно аргументировал В. В. Сдобников: «универсальное требование к качеству перевода: быть способным удовлетворять потребности коммуникантов и инициатора перевода в рамках их предметной деятельности в определенной коммуникативной ситуации» [3, с. 41]. Результаты перевода, оцененного только на основе текстовых категорий и изолированного от реальных условий, носят субъективный характер и не могут рассматриваться как достоверные.

Комплексное изучение лингвистических и коммуникативных критериев представляется перспективным для дальнейших исследований, в том числе в сфере дидактики. Разделяя мнение исследователей о трудностях реализации такого подхода при обучении переводу, представляется целесообразным разделять критерии оценки академической успеваемости обучающихся и профессиональной деятельности [3; 4]. В дидактическом аспекте, на наш взгляд, представляется перспективным анализ критериев оценки, используемых для решения учебных и переводческих задач.

В определенной оппозиции по вопросам оценки качества к исследовательскому и дидактическому аспектам находится переводческая отрасль, в реальных условиях которой отсутствует возможность де-

тального обсуждения результатов и выставления оценки вручную. С развитием систем машинного перевода и последующей интеграцией нейронных сетей в профессиональный инструментарий на практике усилилась тенденция к формализации процесса оценки качества перевода. В определенном смысле базой для формализации является типология переводческих ошибок, включенная в алгоритм присвоения «веса» различного рода неточностям и математического способа вычисления итоговой оценки. Не останавливаясь подробно на оценке объективности результатов с помощью такой формальной или универсальной схемы, следует отметить преобладание на рынке переводческих услуг тенденции к оптимизации выполняемых видов работ и острую потребность в средствах оценивания.

Разработанные к настоящему времени метрики автоматической оценки качества как максимально формальные средства лишь частично удовлетворяют требованиям объективного оценивания индустрии перевода. Изначально метрики представляли собой только формулы, основанные на статистическом сравнении предложенного перевода с эталонным образцом перевода, выполненным профессиональным переводчиком (BLEU, WER, TER). Со временем на основе формул стали создаваться отдельные программы, а не так давно программы стали обучать с помощью нейронных сетей. В 2022 г. на Седьмой конференции по машинному переводу (WMT22), которая является одним из самых авторитетных мероприятий в переводческой отрасли, были представлены новые метрики автоматической оценки, подавляющее большинство из которых созданы на основе нейронных сетей [5]. Активное применение метрик для автоматической оценки машинного перевода продолжится при условии понимания их сильных и слабых сторон для решения переводческих задач. Вместе с тем, по мнению А. В. Соснина, Ю. В. Балакиной и А. Н. Кашихина, установивших взаимосвязь экспертных категорий адекватности и эквивалентности и автоматических метрик, последние «практически невозможно использовать для оценки человеческого перевода по следующей причине: большинство из современных метрик – N-граммные» [1, с. 140]. Прежде всего, исключается оценка перевода художественных текстов, но более формальные и «сухие», к которым авторы относят научные статьи, техническую документацию, могут быть переведены с «N-граммной точностью»; следовательно, этот перевод может быть оценен с помощью метрик [1].

Несмотря на стремительные улучшения в работе метрик, экспертная оценка, традиционно выполняемая человеком, по-прежнему остается основным видом оценивания. С экспертной оценкой неразрывно связано понятие «типология оценивания», описа-

нием компонентов которой активно занимаются как теоретики, так и практики перевода. В этом смысле данное явление можно считать точкой соприкосновения теоретических и практических подходов.

Самые первые версии типологии были достаточно простыми. Например, консультативный комитет по автоматической обработке языков (ALPAC – Automatic Language Processing Advisory Committee) оперировал всего двумя характеристиками текста: условная понятность (intelligibility), которая оценивалась по балльной шкале от 1 до 9, и точность перевода (fidelity или accuracy), которая оценивалась по шкале от 0 до 9 [6, с. 50]. Современные типологии включают два критерия оценки качества перевода: критичность и категории ошибок. Критерий категории ошибок представляет собой классификацию ошибок, состоящую из нескольких уровней. Критерий критичности ошибки определяет количественный штраф или вес ошибки. Далее для вычисления индекса качества перевода, как правило, используется математическая формула.

К преимуществам применения типологий следует отнести систематизацию ошибок по категориям и по критерию критичности, которая помогает экспертам более объективно оценивать качество текстов. Вместе с тем все типологии характеризуются различной степенью субъективности. Например, эксперты сами устанавливают уровни критичности для ошибок какой-либо категории. Отмечая эволюцию типологий в части содержательных и формальных характеристик, следует отметить, что она была и остается неотъемлемой частью экспертной оценки.

Цель данной работы заключается в выявлении отличительных особенностей двух версий типологии «Многомерных метрик качества» (MQM). К задачам исследования также относятся определение преимуществ и недостатков каждой из версий, сравнение эффективности версий в части выявления и объективного оценивания различных типов ошибок, а также оценка целесообразности их использования экспертами.

Материалы и методы исследования

В настоящее время типология «Многомерных метрик качества» (Multidimensional Quality Metrics, MQM) является одной из самых распространенных типологий. В 2021 г. она была признана «золотым стандартом» для экспертной оценки Metrics Shared Task на конференции по машинному переводу [7]. Типология была разработана компанией TAUS в сотрудничестве с исследовательским институтом DFKI в 2014 г. в рамках проекта ЕС «QTLanch Pad». Изначально она называлась DQF-MQM, так как в ней использовалась классификация ошибок TAUS «Динамическая модель оценки качества» (DQF, Dynamic

Quality Framework), которая составляла ядро типологии. В дальнейшем MQM разработала собственную усовершенствованную классификацию ошибок, которая сейчас и составляет ядро типологии, и отбросила «приставку» DQF [8]. Выбор данной типологии для нашего исследования обусловлен также появлением ее обновленной версии, разработанной компанией Google.

Как и в случае с другими типологиями, MQM включают два критерия оценки качества перевода: критичность и категории ошибок. Критерий категории ошибок представляет собой классификацию ошибок, состоящую из трех уровней, каждый из которых более специфичен, чем предыдущий. Критерий критичности ошибки определяет количественный штраф или вес ошибок. Сам процесс оценивания проходит в оценочной таблице, созданной для программы Excel [8]. Эксперты последовательно оценивают каждый сегмент текста, отмечая в таблице отдельные типы ошибок и присуждая каждой из них определенный уровень критичности. На заключительном этапе оценивания перевода автоматически высчитывается суммарный штраф и итоговая оценка текста с помощью встроенной в Excel формулы. Суммарный штраф за все ошибки непосредственно влияет на оценку качества текста. Итоговая оценка текста (Overall Quality Score) рассчитывается по тематической формуле, представленной ниже:

$$OQS = 1 - \frac{APT}{EWC},$$

где APT (Absolute Penalty Total) – сумма всех штрафных баллов в тексте, EWC (Evaluation Word Count) – количество слов в тексте. Чем больше суммарный штраф за все ошибки, тем ниже итоговая оценка текста и, следовательно, качество перевода. Отличительной особенностью типологии является отсутствие связи между двумя критериями оценки, благодаря чему сохраняется гибкость оценки.

Таким образом, типология MQM представляет собой качественное решение для целей оценки качества перевода. Она не только повышает объективность оценивания с помощью введения четкой системы распределения баллов и определенной типологии ошибок, но и сохраняет гибкость экспертной оценки благодаря тому, что два критерия оценки (критичность и категории ошибок) никак не связаны между собой.

В последние годы исследователи и компании пытаются усовершенствовать процесс оценки качества перевода с помощью внесения изменений в уже существующие типологии. В 2020 г. компания Google представила свою версию MQM, в которой предпринимается попытка повысить объективность экспертной оценки. В нашем исследовании используется

метод сравнительно-сопоставительного анализа, направленный на выявление отличительных особенностей двух версий типологии MQM. С помощью количественного анализа изучается влияние количественных изменений в параметрах типологии и степени объективности оценки. Контекстуальный анализ используется для определения содержательных характеристик переводческих решений с учетом изменений в количестве и составе уровней оценивания.

В качестве материалов исследования были использованы тексты с Пятой конференции по машинному переводу (WMT20), так как версия MQM Google тестировалась на материалах с данной конференции. В качестве оригинальных текстов были взяты материалы, предоставленные разработчиками конференции для задания «newstest2020», в котором проверялась работа систем машинного перевода. В качестве переводов использовались переводы оригиналов на русский язык, поданные на данную конференцию компанией PROMPT, разрабатывающей одноименную систему машинного перевода [9]. Предоставленные на перевод тексты различались по объему и тематике: от обычных новостных заметок до статей медицинской тематики с использованием профессиональной лексики. С учетом источников оригинальные материалы можно охарактеризовать как тексты массмедийного дискурса или в другой терминологии как газетно-публицистические тексты. Наличие в исходных текстах элементов разных стилей в совокупности с тематическим разнообразием способствуют появлению в тексте перевода ошибок различных типов и уровней критичности, что позволит более полно оценить эффективность типологии на практике.

Результаты исследования

В версии MQM Google предпринимается попытка снизить субъективность оценки, связав два критерия через систему оценивания, поэтому наибольшему количеству изменений подвергся критерий критичности ошибок.

В оригинальной типологии MQM выделяются четыре уровня критичности.

1. Критическая ошибка – штраф 25. Всего одна критическая ошибка делает перевод автоматически непригодным для использования.

2. Значительная ошибка – штраф 5. Может затруднить понимание текста.

3. Незначительная – штраф 1. Ошибка присутствует, но не затрудняет понимание.

4. Нулевая – штраф 0. Текст не содержит ошибок, но на данный сегмент текста необходимо обратить внимание [10, с. 110].

Они не привязаны к определенным типам ошибок, поэтому эксперты могут присуждать любые уровни критичности любым ошибкам.

Компания Google решила «пожертвовать» данной гибкостью ради повышения объективности оценки. Она связала два критерия оценки, присвоив определенным категориям ошибок определенные штрафные баллы. В табл. 1 представлена система оценивания MQM Google [11, с. 1465].

Т а б л и ц а 1

Система оценивания MQM Google

Критичность	Категория	Штрафной вес
Грубые ошибки	Комплексные ошибки	25
	Все остальные	5
Незначительные ошибки	Естественность/ пунктуация	0,1
	Все остальные	1
Нейтральные ошибки	Все	0

Категории «Естественности/пунктуации» был присвоен вес 0,1 балла, так как по данным исследователей Google большинство ошибок в данных категориях не относятся непосредственно к переводу, например, расстановка пробелов или использование кавычек. Данные ошибки легко устраняются с помощью дополнительных алгоритмов и чаще всего не влияют на понимание смысла предложения [11, с. 1464]. Более того, Google ограничила максимальный штраф за сегмент текста 25 баллами, что отразила в категории «Комплексные ошибки», так как данные ошибки, затрагивающие сразу несколько категорий, влияют на сегмент в целом. Для повышения объективности оценки компания также убрала критический уровень, аргументируя свое решение тем, что разли-

ца между критической и грубой ошибкой часто зависит от контекста и, следовательно, является субъективной.

В версии Google изменения коснулись типологии ошибок MQM. Оригинальная типология MQM состоит из 3 уровней спецификации. В табл. 2 представлены первые два уровня типологии, так как последний, третий уровень, редко используется экспертами на практике в силу своей излишней спецификации [10, с. 109].

Сравнительный анализ оригинальной версии MQM и версии, предложенной компанией Google, позволил выявить количественные и качественные различия в системе оценивания.

К количественным изменениям относится сокращение компанией Google 3-го уровня в типологии ошибок. Данное решение представляется логичным из-за излишней детализации оригинальной типологии ошибок MQM. Ошибки на 3-м уровне отражают дополнительную информацию, которая практически не влияет на оценку перевода. Например, 3-й уровень категории «Точность», подкатегории «Неверный перевод (Mistranslation)» включал в себя: *двусмысленный перевод, перевод даты/времени, перевод имени собственного, технический неверный перевод, перевод числа, буквальный перевод, перевод единицы измерения и перевод единицы, которую не надо переводить*. Результаты сравнения демонстрируют схожесть некоторых категорий 3-го с категориями 2-го уровня «Соблюдение традиций написания текстов на языке перевода»: *перевод единицы измерения и формат единицы измерения*. Из-за большого количества категорий, сходства в их названиях и т. д. повышается степень субъективности экспертной оценки.

Т а б л и ц а 2

Типология MQM

Первый уровень	Второй уровень
Точность (Accuracy)	Добавление, Опускание, Неточный перевод, Излишняя спецификация, Недостаточная спецификация, Непереведенный сегмент
Терминология (Terminology)	Ошибка в выборе термина, Термин из другой предметной области, Непостоянное употребление термина
Естественность (Fluency)	Грамматика, Регистр, Орфография, Кодировка символов, Непоследовательность
Стиль (Style)	Неловкий стиль, Корпоративный стиль, Непоследовательность при передаче стиля, Несоответствующий стиль
Дизайн (Design)	Графики и таблицы, Длина текста, Формат текста, Переносы слов, Локальное форматирование, Разметка и др.
Соблюдение традиций написания текстов на языке перевода (Locale convention)	Формат написания адресов, дат, имен, валют, единиц измерения, телефонных номеров, времени, Специфическая пунктуация, Специфические сочетания клавиш
Истинность (Verity)	Полнота описания в переводе, Соответствие целевой аудитории, Культуро-специфичные единицы, Соответствие правовым нормам, Пригодность для использования
Другое (Other)	

Во-первых, эксперты по-разному интерпретируют названия специфичных категорий. Во-вторых, из-за громоздкости типологии увеличивается время на поиск нужного соответствия, вследствие чего не исключены ошибки при выборе предложенных вариантов.

Кроме сокращения уровней изменился количественный состав внутри оставшихся уровней. Например, были полностью удалены категории 1-го уровня «Истинность» и «Дизайн». В категории «Стиль» была сохранена только одна подкатегория – «Неловкий стиль». Однако наряду с сокращениями компания Google добавила несколько категорий 1-го уровня, например, категорию «Ошибка в оригинале (Source error)», в которой отмечаются случаи, когда ошибка в переводе вызвана ошибкой в оригинале текста. Также была добавлена категория «Комплексные ошибки (Non-translation)», в которой отмечаются ошибки, попадающие сразу в несколько категорий.

Выявленные количественные несоответствия в категориях ошибок влияют на степень объективности оценки качества. Следующим шагом в нашем исследовании было изучение качественных изменений указанного критерия на практике.

Компания Google добавила в классификацию категорию «Комплексных ошибок». Данное решение освобождает экспертов от решения вопросов оценки, возникающих в случае, когда одна единица в переводе попадает сразу под две или более категории ошибок.

Пример 1

Комплексная ошибка

Оригинал	Перевод
...a couple of quick concerned phone calls from palace to Downing Street should have been enough to persuade the PM to shorten his pre-Queen's Speech prorogation to the usual length of a few days or a week (https://drive.google.com/drive/folders/)	...пары быстрых обеспокоенных телефонных звонков из дворца на Даунинг-стрит должно было быть достаточно, чтобы убедить премьера сократить его пророгацию речи перед королевой до обычной продолжительности в несколько дней или неделю (https://drive.google.com/drive/folders/)

В примере 1 при переводе сохранено некорректное написание термина *пророгация*. Кроме формальной ошибки представляется проблематичным понимание данного понятия в тексте, предназначенном для массовой русскоязычной аудитории. Пророгация – это юридический термин, тесно связанный с Вестминстерской системой государственного управления, и как любая специальная единица, является в большинстве случаев не известным широкой публи-

ке. Оптимальным переводческим решением с учетом характера реципиента будет вариант *отсрочка*. В случае оценивания перевода данной единицы с помощью типологии MQM может возникнуть вопрос, к какой категории следует отнести ошибку. А в случае ее принадлежности к двум категориями, актуальным становится другой вопрос, как оценивать в дальнейшем: выставлять штрафной балл за каждую категорию отдельно или за передачу единицы в целом. Подобные ситуации повышают субъективность оценки, так как эксперты могут прийти к разным решениям этой проблемы. В связи с этим решение Google добавить дополнительную категорию для «Комплексных ошибок» представляется обоснованным.

Добавление другой категории «Ошибка в оригинале» позволяет сразу же выявлять сегменты текста с ошибкой в оригинале, так как все они будут сгруппированы в одну категорию. Одновременно встает вопрос о необходимости выделения отдельной категории 1-го уровня для решения данной проблемы. Во-первых, если ошибка была изначально допущена в оригинале, система машинного перевода не сможет самостоятельно ее выявить и, следовательно, переведет текст неверно. В данном случае нецелесообразно присваивать машинному переводу штраф, так как программа перевода не является причиной ошибки. Следовательно, все ошибки в данной категории будут нейтрального уровня критичности и будут весить 0 баллов. Во-вторых, в оригинальной типологии данную функцию маркера выполняет нейтральный уровень критичности.

Таким образом, изменения в версии MQM Google в части классификации ошибок направлены на повышение объективности экспертной оценки. Компания уменьшила количественные показатели типологии, прежде всего, ограничив вариативность выбора типа ошибок. Такая прозрачная и удобная в работе типология позволит более продуктивно оценивать текст перевода. Однако для повышения объективности компания исключила две ключевые категории ошибок 1-го уровня, что нивелировало ее преимущества в части гибкости и полноты анализа текста.

Вторым критерием оценки качества MQM, в котором произошли наибольшие изменения, является критерий критичности ошибок.

Компания Google исключила критический уровень ошибки (25 штрафных баллов), так как по данным исследователей компании разница между критической и грубой ошибкой зависит от контекста и поэтому не объективна [11]. В оригинальной типологии MQM под критическими ошибками понимаются ошибки, которые кардинально меняют смысл текста перевода, в то время как грубые – значительно затруд-

няют его понимание. Мы считаем, что это важное деление, которое чаще всего возможно использовать объективно.

Пример 2

Неверный перевод

Оригинал	Перевод
<i>She conceded that people were perfectly entitled to behave foolishly in her shot while she was reporting from a public place</i> (https://drive.google.com/drive/folders/)	<i>Она признала, что люди имеют полное право вести себя глупо, когда она стреляла из общественного места</i> (https://drive.google.com/drive/folders/)

В тексте оригинала под единицей *in her shot* подразумевается кадр репортажа. Из-за неверного перевода весь сегмент, а в итоге весь текст, стал непригодным для использования, так как журналистка из жертвы превратилась в нападающую. Более того, в сегменте, где в оригинале говорилось только про съемку репортажа и попадание людей в кадр, появилась стрельба в общественном месте. Из примера 2 становится очевидным тот факт, что указанную ошибку необходимо отнести к критическим, а не к грубым. Штрафные баллы, присуждаемые за данный сегмент, будут в пять раз меньше в типологии MQM Google, чем в типологии MQM: 5 баллов к 25 соответственно.

Максимальные штрафные баллы (25 баллов) в версии типологии MQM Google присуждаются только за грубые ошибки комплексного характера. Однако мы считаем, что данное решение требует дальнейшего осмысления, потому что критичность ошибки никак не связана с количеством категорий, под которые она попадает. Так, например, комплексная ошибка может относиться к грубым ошибкам, но при этом не заслуживать 25 штрафных баллов.

Пример 3

Грубая комплексная ошибка

Оригинал	Перевод
<i>And Scotland, which has not been the primary target of this destabilisation, must now soon be given a second chance to consider its position, in relation to a UK so radically changed by Brexit</i> (https://drive.google.com/drive/folders/)	<i>И Шотландия, которая не была главной целью этой дестабилизации, должна вскоре получить второй шанс рассмотреть свою позицию по отношению к Великобритании, столь радикально изменившейся в результате Brexit</i> (https://drive.google.com/drive/folders/)

В примере 3 описывается случай, когда единица «Brexit» в переводе попадает под несколько подкатегорий типологии MQM Google: *непереведенный*

текст, грамматика, несогласованность терминологии, а также неверный формат передачи имени собственного. Это делает перевод единицы комплексной ошибкой, а по уровню критичности он соответствует уровню грубой ошибки: значительно затрудняет понимание текста, однако все же не искажает смысл. Несмотря на это, в рамках типологии MQM Google за данную ошибку присуждают максимальные 25 баллов, в то время как за по-настоящему критическую ошибку из предыдущего примера – всего 5. В связи с этим привязка максимального штрафного количества баллов к комплексным ошибкам на практике может повышать недостоверность оценки качества перевода и неверно отражать качество как сегмента, так и всего текста в целом.

Изменения компании Google в критерии критичности и системе оценивания MQM в целом представляются недостаточно обоснованными. Практика показала, что объединение критерия критичности с критерием категорий ошибок не приводит к повышению объективности, скорее наоборот, может повысить недостоверность перевода, так как ограничивает возможность экспертов оценивать качество перевода с учетом контекста.

По итогам проведенного исследования следует также отметить, что основные изменения в версии MQM Google коснулись системы оценивания и распределения баллов. К наиболее удачным решениям компании Google следует отнести сокращение уровней типологии, добавление категории «Комплексные ошибки», позволившей единообразно оценить ошибки, которые можно отнести сразу к нескольким категориям. Данные изменения позволили повысить уровень объективности данной типологии. В то же время решения MQM Google в отношении критерия критичности ошибок, прежде всего исключение критического уровня из типологии, нуждаются в дальнейшем осмыслении и доработке.

Открытым остается вопрос о перспективах использования каждой из указанных версий. С одной стороны, необходимость обновления систем оценивания является очевидной. С другой стороны, увеличение количества версий одной и той же типологии может привести к затруднениям в реализации, что подтверждает опыт проведения Седьмой конференции по машинному переводу (WMT22). В 2022 г. проверка текстов осуществлялась, помимо экспертов конференции, двумя независимыми компаниями: Google и Unbabel. Каждая из компаний использовала типологию «Многомерных метрик качества» (MQM). Однако из-за того, что Google использовала собственную версию с измененной системой оценки, а Unbabel – традиционную, организаторам конференции пришлось применять дополнительные меры, чтобы привести данные к одной шкале. Ожидаемо,

что в процессе приведения двух наборов данных к одной шкале, часть данных была потеряна [5, с. 50].

Таким образом, в результате анализа изменений MQM Google мы пришли к выводу, что на данном этапе целесообразным представляется рассматривать типологию MQM Google как экспериментальное решение, а не как успешно модифицированную версию типологии MQM, прежде всего из-за решений, повлиявших на усиление субъективности оценки.

Заключение

Оценка качества перевода включает широкий спектр вопросов, изученных в разной степени в зависимости от задач исследовательской, образовательной и профессиональной деятельности. Для каждого из указанных направлений актуальным остается поиск и реализация оптимальных средств оценивания. И если в академических кругах ведется дискуссия о необходимости формализации параметров оценки, возможности разработки моделей для объективной оценки текстов, то в индустрии перевода уже используются различные технологии для экспертной и автоматической оценки качества. С учетом особенностей их реализации в современных условиях профессиональной деятельности происходит модификация систем оценивания. Современный подход к параметрам типологии позволяет оценить качество перевода текстов разной жанровой принадлежности. Анализ критериев, применяемых на практике, представляется необходимым для теоретического осмысления в части соотношения используемых категорий с основными параметрами оценки, выявления их преимуществ и недостатков. Перспективным представляется установление взаимосвязи оценки качества машинного перевода и способов улучшения его технологий, а также возможности применения современных средств для оценки перевода, выполненного человеком. Практическая значимость исследования заключается в применении его результатов для разработки и совершенствования технологий постредактирования машинного перевода. При обучении переводу итоги работы могут быть использованы для формирования у обучающихся более полного представления о наборе параметров оценки качества, используемых в профессиональном переводе, а также для развития умения выбирать оптимальные средства оценивания, основанные на понимании преимуществ и недостатков различных версий.

ЛИТЕРАТУРА

1. Соснин А. В., Балакина Ю. В., Кацкихин А. Н. Взаимосвязь экспертных категорий и автоматических метрик, используемых для оценки качества перевода // Вестник Санкт-Петербургского университета. Язык и литература. 2022. № 19 (1). С. 125–148.

2. Ивлева А. Ю. Проблема оценки качества перевода : подходы к решению // Вестник Волгоградского государственного университета. Сер. 2, Языкознание. 2018. Т. 17, № 1. С. 98–106.

3. Сдобников В. В. Оценка качества перевода в аудитории и в реальной жизни // Перевод и сопоставительная лингвистика. Екатеринбург : Изд-во Урал. гуманитарного университета, 2015. Вып. 11. С. 41–45.

4. Ивлева А. Ю. Оценка качества профессионального и учебного перевода : проблемы и перспективы // Вестник Пермского государственного политехнического университета. Проблемы языкознания и педагогики. 2018. № 1. С. 8–18.

5. Results of WMT22 Metrics Shared Task : Stop Using BLEU – Neural Metrics Are Better and More Robust / M. Freitag et al. // Association for Computational Linguistics : Proceedings of the Seventh Conference on Machine Translation (Abu Dhabi, United Arab Emirates, 7–8 December 2022). Abu Dhabi, 2022, Pp. 46–68.

6. Улиткин И. А. Автоматическая оценка качества машинного перевода научного текста : 5 лет спустя светлой памяти моего учителя Нелюбина Л. Л. // Вестник Московского государственного университета. Сер.: Лингвистика. 2022. № 1. С. 47–59.

7. Results of the WMT21 Metrics Shared Task : Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain / M. Freitag et al. // Association for Computational Linguistics : Proceedings of the Sixth Conference on Machine Translation (Online, 10–11 November 2021). Online, 2021. Pp. 733–774.

8. MQM: официальный сайт. URL: <https://themqm.org/>

9. Results of the WMT20 Metrics Shared Task / Nitika Mathur et al. // Association for Computational Linguistics : Proceedings of the Fifth Conference on Machine Translation (online, 19–20 November 2020). Online, 2020, Pp. 688–725.

10. Рычихин А. К. О методах оценки качества машинного перевода // Системы и средства информатики. 2019. № 29 (4). С. 106–118.

11. Experts, Errors, and Context : A Large-Scale Study of Human Evaluation for Machine Translation / M. Freitag et al. // Transactions of the Association for Computational Linguistics. 2021. № 9. Pp. 1460–1474.

ИСТОЧНИК

Материалы конференции WMT20_data. URL: https://drive.google.com/drive/folders/1n_alr6WFQZfw4dcAmyx-ow4V8FC67XD8p

REFERENCES

1. Sosnin A. V., Balakina Yu. V., Kashhikhin A. N. Vzaïmosvyaz' ehkspertnykh kategorij i avtomaticheskikh metrik, ispol'zuemykh dlya otsenki kachestva perevoda [Interdependence of expert categories and automated metrics applied to evaluate translation quality]. In: *Vestnik Sankt-Peterburgskogo universiteta. YAzyk i literatura*. 2022. No. 19 (1). Pp. 125–148.

2. Ivleva A. YU. Problema otsenki kachestva perevoda: podkhody k resheniyu [The Problem of Translation Quality

Assessment: Approaches to Solution]. In: *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2, Yazykoznanie*. 2018. Vol. 17, No. 1. Pp. 98–106.

3. Sdobnikov V. V. Otsenka kachestva perevoda v auditorii i v real'noj zhizni [Translation quality assessment in a translation class and in real life]. In: *Perevod i sopostavitel'naya lingvistika*. 2015. No. 11. Pp. 41–45.

4. Ivleva A. YU. Otsenka kachestva professional'nogo i uchebnogo perevoda: problemy i perspektivy [Assessment of professional and academic translation quality: problems and perspectives]. In: *Vestnik Permskogo natsional'nogo issledovatel'skogo politekhnicheskogo universiteta. Problemy yazykoznanija i pedagogiki*. 2018. No. 1. Pp. 8–18.

5. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust / M. Freitag et al. In: *Association for Computational Linguistics: In Proceedings of the Seventh Conference on Machine Translation (Abu Dhabi, United Arab Emirates, 7–8 December 2022)*. Abu Dhabi, 2022. Pp. 46–68.

6. Ulitkin I. A. Avtomaticheskaya otsenka kachestva mashinnogo perevoda nauchnogo teksta: 5 let spustya svetloj pamyati moego uchitelya Nelyubina L. L. In: *Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta. Seriya: Lingvistika*. 2022. No. 1. Pp. 47–59.

Российский государственный педагогический университет имени А. И. Герцена

Елизарова Л. В., кандидат филологических наук, доцент кафедры перевода

E-mail: lyudmilaelizarova@yandex.ru

Дмитриева К. К., студент 4-го курса института иностранных языков, профиль «Перевод и переводоведение»

E-mail: kdmtrieva2002@gmail.com

Поступила в редакцию 25 ноября 2023 г.

Принята к публикации 26 марта 2024 г.

Для цитирования:

Елизарова Л. В., Дмитриева К. К. Объективные параметры субъективной оценки качества перевода // Вестник Воронежского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2024. № 2. С. 30–38. DOI: <https://doi.org/10.17308/lic/1680-5755/2024/2/30-38>

7. Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain / M. Freitag et al. In: *Association for Computational Linguistics: Proceedings of the Sixth Conference on Machine Translation (Online, 10–11 November 2021)*. Online, 2021. Pp. 733–774.

8. MQM: Official Website. Available at: <https://themqm.org>

9. Results of the WMT20 Metrics Shared Task / Nitika Mathur et al. In: *Association for Computational Linguistics: Proceedings of the Fifth Conference on Machine Translation (online, 19–20 November 2020)*. Online, 2020. Pp. 688–725.

10. Rychikhin A. K. O metodakh otsenki kachestva mashinnogo perevoda [On methods of machine translation quality assessment]. In: *Sistemy i sredstva informatiki*. 2019. No. 29 (4). Pp. 106–118.

11. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation / M. Freitag et al. In: *Transactions of the Association for Computational Linguistics*. 2021. No. 9. Pp. 1460–1474.

SOURCE

Materialy konferentsii WMT20_data. Available at: https://drive.google.com/drive/folders/1n_alr6WFQZfw4dcAmyxow4V8FC67XD8p

Herzen State Pedagogical University of Russia
Elizarova L. V., Candidate of Philology, Associate Professor of the Translation and Interpreting Department
E-mail: lyudmilaelizarova@yandex.ru

Dmitrieva K. K., Student, 4th year of Study, Institute of Foreign Languages, Main Field of Study: Translation and translation studies
E-mail: kdmtrieva2002@gmail.com

Received: 25 November 2023

Accepted: 26 March 2024

For citation:

Elizarova L. V., Dmitrieva K. K. Objective parameters for subjective evaluation of translation quality. Proceedings of Voronezh State University. Series: Linguistics and Intercultural Communication. 2024. No. 2. Pp. 30–38. DOI: <https://doi.org/10.17308/lic/1680-5755/2024/2/30-38>