Comparative Quality Assessment of Human and Machine Translation with Best-Worst Scaling

Bettina Hiebl and **Dagmar Gromann**
University of Vienna, Austria
{bettina.hiebl, dagmar.gromann}@univie.ac.at

Abstract

Translation quality and its assessment are of great importance in the context of human as well as machine translation. Methods range from human annotation and assessment to quality metrics and estimation, where the former are rather time-consuming. Furthermore, assessing translation quality is a subjective process. Best-Worst Scaling (BWS) represents a time-efficient annotation method to obtain subjective preferences, the best and the worst in a given set and their ratings. In this paper, we propose to use BWS for a comparative translation quality assessment of one human and three machine translations to German of the same source text in English. As a result, ten participants with a translation background selected the human translation most frequently and rated it overall as best closely followed by DeepL. Participants showed an overall positive attitude towards this assessment method.

1 Introduction

Human and machine translation quality and their assessment have been of importance in research and industry alike (Harris et al., 2016). Different concepts in the field of translation studies include those focusing on preserving the purpose of the source text in the translation, such as the Skopos theory (Reiss and Vermeer, 1984), on the target text as central point in the analysis of quality as Ammann (1990), or on pragmatic aspects of translation as House (2015).

Quality Assessment (QA) approaches in the field of Machine Translation (MT) include QA frameworks for assessment by humans and by machines. Very well-known automated metrics that compare candidate translations to reference translations are, for example, BLEU (Papineni et al., 2002), NIST (Doddington, 2002), and METEOR (Banerjee and Lavie, 2005). MT Quality Estimation (Specia and Shah, 2018) represents a fairly new approach that instead of using reference translations trains machine learning models to predict the output quality of a specific MT system. Human assessment of machine translation consists of human ranking (Macháček and Bojar, 2013), overall assessment (Bojar et al., 2017) or error classification (Popović, 2018) and is generally considered subjective, time-consuming and therefore expensive.

Best-Worst Scaling (BWS) (Louviere and Woodworth, 1990) is an annotation method that addresses these limitations, since it allows for subjective and time-efficient annotations. Annotators are provided with n items in a set and are asked to select the best and the worst items from the set. With a set of four items, this simultaneously leads to a ranking with one clear best item, two that are better than the fourth, and a fourth worst item. BWS has successfully been applied to annotating emotion intensities (Mohammad and Bravo-Marquez, 2017), evaluating stakeholder priorities in health matters (Hollin et al., 2022), assessing consumer preferences in wine attributes (Stanco et al., 2020), among many other application scenarios. BWS has also been applied to assess gender-fair language strategy preferences in translation (Paolucci et al., 2023). However, to the best of our knowledge, comparative translation quality assessment with BWS has not been proposed before.

In this paper, we propose a comparative analysis of five sets of four German translations of the same English source text, one human and three machine translations from Google Translate, DeepL and Microsoft Bing Translator. Ten master students of translation studies or multilingual technologies selected the best and the worst option from the set and rated the best from +4 to 0 and the worst from 0 to -4 in an online survey. This rating provides an overall score for each translation method, but also allows for a more detailed analysis on how high or low each method is assessed. In contrast to ranking, e.g. Bojar et al. (2013), not each translation needs to be annotated with a rank label for each set, but only the best and the worst. Furthermore, the agreement between choices and ratings can be directly assessed without having to calculate an inter-annotator agreement. The translations were selected across domains and consisted of one paragraph from non-fiction books, which required a comparatively low level of domain expertise from participants. In addition, participants were invited to leave comments on each set in free text fields and evaluate the overall method at the end of the survey. The results showed an overall positive attitude towards BWS. Since translation quality assessment by humans in itself is rather subjective, we believe that BWS provides a viable, time-efficient and easy to implement alternative for comparing translations, which can be a comparison of MT systems, of human translations, or, as in this case study, to compare both.

2 Preliminaries

As a basis for the study presented below, we provide an exemplary overview of selected work on MT quality assessment as well as combined assessment of human and machine translation, with no claims regarding completeness. In addition, we will briefly introduce the concept of BWS and typical use cases.

2.1 Translation Quality Evaluation

The evaluation of translation quality has received much attention in translation studies and is a topic that is open for debate. Proposed methods to quality analysis range from source-oriented functionalist approaches (Reiss and Vermeer, 1984; Nord, 1997) to target-text quality analysis, e.g. Ammann (1990), a focus on pragmatic aspects, e.g. House (2015), and analysis based on comprehensibility dimensions (Göpferich, 2008). A common denominator for translation quality in translation studies and machine translation are the concepts of the source text-focused adequacy or accuracy and the target text-focused fluency (Castilho et al., 2018). The Multidimensional Quality Metric (MQM) (Lommel et al., 2014) proposes a framework for translation quality evaluation to be applicable to human and machine translation alike. To this end, a catalogue of known quality issues that can be used as an error typology is presented. Another similar error typology that also considers automation was proposed by Popović (2018). As a mid-way between error classification and overall rating or ranking, Popović (2020) propose to mark all words, phrases, and sentences of a target text that are problematic in terms of comprehensibility and adequacy.

Well-known automated metrics to evaluate machine translation are BLEU (Papineni et al., 2002), NIST (Doddington, 2002), and METEOR (Banerjee and Lavie, 2005), which rely on existing reference translations. In order to avoid having to use reference translations, the idea of Machine Translation Quality Estimation (MTQE) (e.g. Specia and Shah (2018)) was proposed, in which machine learning models are trained to predict the translation quality of MT models. For instance, COMET (Rei et al., 2020) takes the source text into account in training a multilingual MT evaluation model and seeks to assimilate human rankings. Toral and Way (2018) used both BLEU as well as human assessment. In a similar fashion, Webster et al. (2020) compared English to Dutch literary translations by humans and the NMT-based systems Google Translate and DeepL, assessing them using manual annotation as well as different metrics in order to get insights into lexical richness, cohesion, syntactic and stylistic parameters. They found NMT to follow the sentence structure of the source text more closely and that human translation tends to have more lexical richness and local cohesion. Several others (Ortiz-Boix and Matamala, 2017; Jia et al., 2019) focus on comparing human translations to post-edited MT output or on the influence of machine translation errors or quality on post-editing effort and performance (Carl and Báez, 2019; Munkova et al., 2021).

Approaches to quality evaluation that are closest to BWS relate to ranking and rating of translations. In the Workshop on Machine Translation (WMT) starting from 2013 (Macháček and Bojar, 2013; Bojar et al., 2013) five different machine translations with 30 or less words for one source text were humanly ranked relative to each other, allowing for ties in rank. These collected rank labels were then used to assign an overall score to each MT system. In 2017 (Bojar et al., 2017), moved from pair-wise ranking to direct assessment of one machine translation with a reference translation on a 0-100 rating scale by means of crowd-sourcing. The yearly collocated WMT Metrics Shared Task (Freitag et al., 2023) asked professional annotators to label problematic sequences with an MQM error category and severity to be compared with automated evaluation metrics and approaches. With BWS annotators only select the best and the worst translation from a set instead of assigning error categories, rank labels or scores to all translations. Furthermore, there is no need to additionally calculate inter-annotator agreements, since both the annotator's choices and numbers assigned to the best and the worst options allow for a direct comparison of translations/systems, especially given the negative scores for worst translations. In other words, the method as such is designed to provide a comparative score between translations and annotators as a result across all sets.

2.2 Best-Worst Scaling

BWS(Louviere and Woodworth, 1990; Louviere et al., 2015) was developed by Louviere in the 1980s for measuring a list of objects by dividing them into subsets, which are measured on one or more underlying, latent, subjective scales by selecting the best and worst option of each set (Louviere et al., 2015), allowing for comparative rating. Its underlying concept is random utility theory (RUT) (Thurstone, 1927), which assumes that humans are rational decision-makers trying to maximize utility when making choices (Cascetta, 2009), but acknowledges that the utilities have a random component (Louviere et al., 2015). Originally applied mostly in the field of psychology, BWS has been used in different fields, such as health, agriculture, environment, business, linguistics, transportation, and other fields, within the last two decades (Schuster et al., 2024). In the context of translation, Balducci Paolucci et al. (2023) conducted a case study focusing on gender-fair language in translation from English to German, using BWS and a Likert scale in order to evaluate preferences of specific gender-fair translation strategies. To the best of our knowledge, the proposed study is the first to use BWS for assessing human and machine translation quality in comparison.

3 Method

In order to assess the translation quality of different machine translation systems as well as human translation, a combination of two methods for measuring subjective assessment was used: Best-Worst Scaling (BWS) and a Likert scale. BWS (Louviere et al., 2015) was used to select the subjectively best and worst translation, whereas the Likert scale (Likert, 1932) was used to rate the quality of the selected best and worst translations. These methods were used jointly in order to not only rate whether a specific translation was perceived to be the best or worst of a set, but also how high or low the selected translations are rated.

3.1 Text Selection

Five original English text passages of non-fiction books as well as their officially published human translations were selected for the case study, in order to guarantee having a good quality human translation as well as texts on different, slightly specific, non-fiction topics, which are history, politics, finance, biology, and physics. The selected texts were taken from the following books:

- Set 1, History: Queen of our Time: The Life of Elizabeth II by Robert Hardman (Pan Macmillan, 2022)
- Set 2, Politics: A Promised Land by Barack Obama (Penguin Books, 2020)
- Set 3, Finance: Bitcoin for Dummies by Peter Kent and Tyler Bain (Wiley, 2023)
- Set 4, Biology: Seven and a Half Lessons about the Brain by Lisa Feldman Barrett (HarperCollins, 2020)
- Set 5, Physics: Quantum Physics for Dummies by Steve Holzner (Wiley, 2013)

Each of the English text passages consists of one to three sentences. The length of the original texts ranges from 31 to 41 words and from 197 to 270 characters. The statistics on the words and characters per text and translation are shown in Table 1.

Set	W EN	C EN	W HT	C HT	W GT	C GT	W DL	C DL	W BT	C BT
1	36	197	38	262	37	251	36	252	34	235
2	46	270	47	345	45	326	45	328	44	317
3	31	206	31	241	35	247	35	249	35	247
4	38	217	52	333	37	261	41	265	38	262
5	46	268	47	329	42	282	42	292	45	299

Table 1: Counts of Words (W) and Characters (C) for the source texts (EN) and the translations by a human (HT), Google Translate (GT), DeepL (DL), and Microsoft Bing Translator (BT)

3.2 Translation Selection

As human translation for each of the texts, the published translation of the books was used. With the exception of Quantum Physics for Dummies, for each of the books, only one German translation has been published so far, namely *Queen of our Times: Das Leben von Elizabeth II (Bastei Entertainment, 2022); Ein verheißenes Land (Penguin Verlag, 2020); Bitcoin für Dummies (Wiley, 2023); and Siebeneinhalb Lektionen über das Gehirn (Rowohlt, 2023)*. For the book with multiple translations the 3rd edition of the book published in 2020, i.e. *Quantenmechanik für Dummies (Wiley, 2020)*, was selected.

For the MT examples, the NMT systems Google Translate[1], DeepL Translate[2], and Microsoft Bing Translator[3] were selected due to their wide usage, popularity, free availability and ease of access.

3.3 Participant Selection

Major criteria for participant selection were a background in translation studies and a very good command of the English and German language. These language skills are required because the English source texts were displayed alongside the German translations in this survey. The selected participants are considered expert annotators in comparison to annotators of other ranking or rating methods that were based on crowd-sourcing (Bojar et al., 2013) or language proficiency (Freitag et al., 2023) without necessarily a professional background in translation, however, the participants selected were no domain experts. The target group consisted of experienced master's students in their last year of studies. Participants who are currently enrolled in a more technical translation master's program, were expected to have a bachelor's degree in translation studies.

3.4 Survey Design

After an introductory description of the survey and some general demographic questions and questions on the background/education of the participants, the survey also comprised some questions on MT background and use of the participants. The entire survey including the source text and translations can be found in Appendix A. The tool Questionstar[4] was used for conducting the survey.

Before starting the main part of the survey, participants were shown a short explanation of BWS and an example of how to rate the selected options. The major part of the survey consisted of five sets of each a source text in English and its four translations to German. Participants were asked to rate the best selected option on a scale from 4 (highest score) to 0 (lowest score) and the worst from from 0 (highest score) to -4 (lowest score). Additionally, participants were invited to provide comments on their choices or the text/translations in a free text field.

The four different translations of the texts were arranged in different order in each set. Reordering translation options between sets is necessary for three main reasons: (1) ensure that participants are not inadvertently biased towards selecting specific options due to translation patterns of individual MT systems, (2) ensure that participants make a reflected choice and not randomly select options, e.g. always first as best and second as worst, and (3) make it harder to be biased by trying to single out a specific choice, which in this case study is the one human translation. The second reason is one very commonly applied for these types of surveys to allow researchers to single out participants that simply click through the sets, without taking the survey seriously. As regards the third reason, the reordering makes sure that participants are not biased towards always selecting the one option where the human translation supposedly is as

[1] https://translate.google.com
[2] https://www.deepl.com/translator
[3] https://www.bing.com/translator
[4] https://www.questionstar.de

best. This reordering was done using all 24 possible different variations of combining four systems (permutations), reordering them using the *RAND* function in Excel and selecting the first 5 instances for Set 1 to 5. The order per set is shown in Table 2. The order was the same for all participants.

In the last section of the questionnaire, participants were asked about the difficulty of selecting the best respectively the worst translation, for their overall opinion on this method for translation quality assessment, for their experience in assessing translation quality, and for any further comments they would like to share.

In order to evaluate the survey design and measure the approximate completion duration, a PhD student of translation studies was invited to pilot the survey. Especially the length of the chosen texts and their translations are an important factor in the design, since cognitive and temporal overload of participants are to be avoided. The pilot study resulted in an estimated duration of 35 minutes and no negative feedback regarding text length, survey length, or clarity of instructions.

3.5 Analysis

The numeric BWS ratings are summed up by translation option across all sets and all participants and divided by the number of times the item was selected and rated. This provides one overall score for all translation options. Furthermore, the number of times an option was selected at all, as best, and as worst are analyzed and presented. While theoretically it could happen that one option is never selected in the entire survey, this is practically unlikely. However, should this be the case, then the option is considered neither the best nor the worst and results in a score of zero. Additionally, all free text comments, demographic data and other answers were analyzed.

4 Results

In this section, the participants' profiles, their BWS ratings for the five sets of translations as well as the corresponding Likert scale ratings will be presented, followed by an analysis of the free text answers and experience with the BWS method. In total, the overall completion time for the entire survey ranged between 20 and 35 minutes.

4.1 Participants

Out of the ten participants, nine identified as female and one as male; 30% are between 18 and 24 years old, 60% are between 24 and 34 years old, and one person is between 35 and 44 years old. All of them had a bachelor's degree as the highest completed degree, 90% in translation studies and 10% in romance studies. Asked to rate their proficiency in English according to the Common European Framework of Reference for Languages (CEFR), seven candidates selected C2, two candidates C1, and one candidate B2. In addition, they had to rate their proficiency in German according to the CEFR, for which eight candidates selected C2, one C1, and one B2. The expected level of German proficiency for this degree program is C1 (CEFR). Therefore, all candidates have a sufficiently high command of English and German and an education related to languages. This is important, since the survey showed the English source texts alongside the German translations. In addition, nine out of ten participants indicated to have some translation experience and the remaining person to have more than 8 years translation experience.

To complement the profiles, the candidates were asked regarding their use of MT tools. Two candidates indicated to use it once a month, seven several times per week, and one person daily. Regarding the purpose of the use of MT, the selected options were privately (5), work other than professional translation (7), work for professional translation (3). For this question, more than one option could be selected. When asked to indicate whether they have a preferred MT system and if so, which one(s), eight participants mentioned DeepL, one person DeepL and Google Translate, and one person Google Translate. The overall satisfaction with MT quality was indicated as very satisfied (2), somewhat satisfied (6), and neither satisfied nor not satisfied (2). The options not very or not at all satisfied were not selected.

4.2 BWS Ratings

From the four translations across five texts, each translation method was selected more than once as best or worst. Table 3 shows the overall averaged and summed ratings and total number of times each translation method was selected. The best summed rating was attributed to the human translation with 33 points from the Likert scale

Set	Option 1	Option 2	Option 3	Option 4
1	BT	HT	DL	GT
2	GT	HT	BT	DL
3	HT	BT	DL	GT
4	DL	HT	BT	GT
5	HT	DL	BT	GT

Table 2: Order of translations per set

across all sets and participants, closely followed by DeepL with 25. The Microsoft Bing Translator and Google Translation were mostly rated negatively, resulting in a score of -20 for the former and -25 for the latter ranked in the overall last place. As can be seen from the detailed BWS rating results in Table 4, this last position and worst result can be attributed to a collective choice as worst translation by all 10 participants and a very negative score on the Likert scale in Set 2 on politics and an overall low selection rate in other sets (see Table 3). Even the overall best option of human translations was assigned a number of negative ratings across sets, but still achieved enough positive selections and ratings in total to result as best option. It is the overall number of times the option was selected as best/worst and scored highly/poorly that finally counts.

The average rating in Table 3 is calculated as the sum of the positive and negative ratings divided by the total number of times the translation method was selected. The average rating for DeepL amounted to 1.04, being slightly higher than the human translation (0.89), while the average ratings for Google Translate (-1.25) and Microsoft Bing Translator (-1.05) were negative. Overall, human translations were selected as the best version 24 times, i.e., in 48% of all cases, whereas the translations by DeepL were selected as best in 30% of all cases. Google Translate and Microsoft Bing Translator clearly lagged behind. While the former was selected more frequently as best and less frequently as worst than the latter, the scores associated with both options still made the former the overall worst option across sets and domains.

The detailed results of the combined BWS & Likert Scale ratings for each participant and each translation output are shown in Table 4. Each of the ten participants rated one of the four presented translations per set as best and one as worst, resulting in a total of 50 selections for each best (rated from 4 highest score to 0 lowest score) and worst (rated from 0 highest score to -4 lowest score). Positive ratings are highlighted in green, negative ones in red, and "neutral" ones in gray. In addition, the overall results are color-coded according to the source of the translation, i.e. whether it is human translation or produced by Google Translate, DeepL Translate, or Microsoft Bing Translator.

As shown in Table 4, the human translations (HT) are selected as best option in three sets on history, politics, and biology, as among the worst in Set 3 on finance, and as clearly the worst in Set 5 on physics. The translation output of DeepL receives an overall positive evaluation in four out of five sets, were only Set 4 in biology results in a finally negative rating. In Set 4 on biology 70% of the participants selected it as the worst option. As regards Microsoft Bing Translator, it is evident that it was the least selected best or worst option in total with 19 selections, where Google Translate obtained only one more selection with a total of 20. Both obtained very negative ratings in one set, Set 1 for the former and Set 2 for the latter. Interestingly BT is the only option not to be selected at all in one set. It can be seen from these results that it is not only the number of times a translation mode is being selected, but also the exact scores associated with a translation. A translation selected considerably less frequently than the human translation (37 times as opposed to 24 times) can still obtain rather positive results if the individually, per-set attributed scores are overall more positive.

4.3 Ratings & Participant Comments per Set

The full source texts and the translations are provided as part of the survey shown in Appendix A. For each individual set participants had the option to comment on their choices of best and worst as well as their evaluations of the translations in a free-text field. For Set 1 on history, a paragraph describing the role of the Lord Chamberlain in the

	HT	**GT**	**DL**	**BT**
Sum Rating	33	-25	25	-20
Avg. Rating	0.89	-1.25	1.04	-1.05
Sum Rating without Set 3	36	-22	10	-17
Times Selected Best	24 (48%)	7 (14%)	15 (30%)	4 (8%)
Times Selected Worst	13 (26%)	13 (26%)	9 (18%)	15 (30%)
Total Selected	37	20	24	19

Table 3: Average, summed and total BWS rating results

	Set 1 (History)				**Set 2 (Politics)**				**Set 3 (Finance)**				**Set 4 (Biology)**				**Set 5 (Physics)**			
Target	**1.1**	**1.2**	**1.3**	**1.4**	**2.1**	**2.2**	**2.3**	**2.4**	**3.1**	**3.2**	**3.3**	**3.4**	**4.1**	**4.2**	**4.3**	**4.4**	**5.1**	**5.2**	**5.3**	**5.4**
Mode	**BT**	**HT**	**DL**	**GT**	**GT**	**HT**	**BT**	**DL**	**HT**	**BT**	**DL**	**GT**	**DL**	**HT**	**BT**	**GT**	**HT**	**DL**	**BT**	**GT**
P1	-3		2		-3	3					2	-3		-2		2		1	-2	
P2		2		-1	-3	1			3	-1			-4	3			-1		3	
P3	-2	1			-2	2			-3		1			4	-4		3	-1		
P4	-2	3			-4	4			-3		2		-4	3				-2	2	
P5	-3	3			-4	3			-2	2			-2	3			-3	4		
P6	0			2	-4			3	2	0			0	0			-4			3
P7	-1		3		-4	3				-1	3		-2	3				3		-1
P8	-3	3			-4			4	-1		4		-1		4		-4	4		
P9	-2			3	-4	4			-1		3		-2			3	-2	4		
P10	-4	3			-4	2			2	-3				-3		2	-1			1
Sum	**-20**	**15**	**5**	**4**	**-36**	**22**	**0**	**7**	**-3**	**-3**	**15**	**-3**	**-15**	**11**	**0**	**7**	**-12**	**13**	**3**	**3**
Best	0	6	2	2	0	8	0	2	3	1	6	0	0	6	1	3	1	5	2	2
Worst	9	0	0	1	10	0	0	0	5	4	0	1	7	2	1	0	6	2	1	1

Table 4: Detailed BWS rating results per participant, strategy, and text (HT = Human Translation, GT = Google Translate, DL = DeepL Translate, BT = Microsoft Bing Translator)

Royal Household, was selected. As shown in Table 4, the translation rated as best most often and receiving the best ratings, is the human translation, whereas the translation by Microsoft Bing Translator is chosen as worst most often and receives the worst rating. The trickiest part of the paragraph for translation was the half-sentence in brackets after describing the Lord Chamberlain as a chairman, saying “it has yet to be a woman”. The human translator opted for translating this as “eine Frau konnte sich für dieses Amt noch nicht durchsetzen” (so far no woman has not yet been able to win this office), whereas the translation by Microsoft Bing Translator reads “es ist noch keine Frau” (it is not yet a woman), the one by DeepL “eine Frau hat es noch nicht gegeben” (there has not yet been a woman), and the one by Google Translate “eine Frau ist es bisher noch nicht” (it is not yet a woman). In evaluating the comments, it turned out that this half-sentence was the crucial reason for participants to select HT as the best and BT as the worst option. Other comments reflected on the different translations for “non-executive”, arguing that “nicht geschäftsführend” sounds more natural than “nicht-exekutiv” and the translations for “Royal Household”, with participants expressing differing opinions on translating it as “königlicher Haushalt” or “Königshaus”.

The paragraph selected for Set 2 on politics is written by Barack Obama describing how his interest in books provided him with knowledge helping him during high school and college. For this set too, the human translation gets the highest overall ratings (22) and is selected as best option and GT as the worst. The most challenging part of this paragraph according to participants’ comments should be at the end of the sentence, when he refers to “bull sessions”, i.e., informal talks/discussions. While all MT systems translate this literally as “Bullensitzungen”, the human translator uses “Diskussionsrunden” (discussion group meetings). Other than that, the most apparent difference between the translation by Google Translate and all other options commented on by most participants is a problem in authenticity and fluency, with changes in the word order contributing grammatical issues, which finally result in its overall selection as worst translation.

For Set 3 on finance, the selected paragraph is a fairly general one about the influence of the launch of Bitcoin on blockchain and cryptocurrency. For this set, the translations produced by Google Translate and Microsoft Bing Translator were exactly the same, so strictly speaking in this set only three different translations were compared. If these two identical translations were counted as one, the summed rating would be -6, once selected as best and 5 times as worst. These should have clearly been eliminated/changed by the researchers before distributing the survey. Interestingly, the translation by DeepL, which differs only slightly from the ones by the two mentioned above, was rated as the best 6 times. The overall rating for the human translation in this set is negative (-3) and it was chosen as worst option by 5 participants. Most of the participants mentioned deciding for best and worst either due to the first sentence or due to the ending. Four of the six participants who selected DeepL's translation commented that their choice was because they liked how the first sentence was phrased, which is "im Bereich Blockchain und Kryptowährung" (in the field of blockchain and cryptocurrency), which no other version used. Four out of the five participants who chose the human translation as the worst option, commented that they did not like the last sentence of the text, i.e., the combination of "Achtung" (Attention) followed by a comma and ending the sentence with an exclamation mark. All of these make it sound more colloquial in German than the English source, and one person commented that they did not like the addition of the word "wahre" (actual) to the "revolution" in the first sentence of the human translation, as this changes the tone of the sentence. Some participants commented that they would prefer "Achtung" to the wording used by all three MT systems "seien Sie gewarnt" (be warned).

The paragraph in Set 4 on biology compares energy efficiency to a financial budget. The human translation is selected as best translation 6 times. The Google Translate version has been selected as best three times with a total rating of 7. The translation produced by DeepL with a total rating of -15 is selected as worst most often (6 times). One specificity of the results for this set is that P6 chose HT as best and DL as worst, but rated them both with "0", which indicates that for this participant neither of the translations was particularly good or bad. Interestingly, for this set several participants commented that deciding on the best and worst translations is difficult without the context or information on the use case, as the human translation is translated much more loosely than the others. Those who selected the translation by DeepL as worst, commented that it is either not coherent or incomprehensible, hard to read, or sounds artificial. The most challenging part of this paragraph was the second half of the last sentence, i.e. "tracks resources like water, salt, and glucose as you gain and lose them." As one participant phrased it, the wording DeepL used "wie Sie zu- und abnehmen" in German sounds as if referring to gaining and losing weight. The participants who chose the human translation as the best version commented that it sounded most natural, was easy to understand, and translated in a creative and not too literal way. One participant who selected the human translation as worst commented that, depending on the context, this translation could also be the best translation, but that without context they do not perceive it as faithful to the source text enough.

The paragraph selected for Set 5 on physics was concerned with black bodies and the spectrum of light emitted by them. The human translation was clearly rated worst for this set with a total rating of -12 and the translation by DeepL clearly rated best with an overall rating of 13. With this set, several participants who selected the human translation as worst mentioned that they did not like that instead of opting for the literal translation of the field "physics" it was translated as "die Physiker" (the physicists), which is not only less general than the field but also adds a masculine gender in German. It was also criticized that in the human translation the word "sogenannten" (so-called) was added. However, one participant who chose the human translation as worst argued that the line between best and worst translation was very thin in this case. The arguments for selecting the translation by DeepL as the best one were that it translates "physics" literally while being most fluent, adding no additional words, and being most appropriate and straightforward.

4.4 General Comments by Participants

At the end of the questionnaire the participants were asked about how easy/difficult they found selecting and rating the best and worst translations overall. The selected options for the degree of difficulty to select the best translation were somewhat easy (4), neutral (1), somewhat difficult (4) and very difficult (1). Rating the difficulty of the best translation per set was judged to be somewhat easy by two participants and somewhat difficult by eight, so the big majority found it to be more difficult to rate than to select the best option. Selecting the worst translation was found to be very easy (3), somewhat easy (3), neutral (1), somewhat difficult (2), and very difficult (1). Rating the translation selected as worst was considered to be somewhat easy by four participants, neutral by three, and somewhat difficult by three.

Participants' free-text comments on the comparative approach of BWS for choosing the best and worst translation were mostly positive. Only one participant mentioned that the approach does not allow for assessing the options regarding more than one dimension making the approach much more subjective. Overall, the approach was found to be a promising and interesting approach and useful for translation quality assessment. According to the participants, the availability of different solutions make you aware of several ideas you would not have considered on your own. Also, it was commented that finding the worst option was easiest and finding the best option much harder, but all in all assessment was easier with more options than having to grade one option would be, although sometimes the best option might have been a combination of several of the available options. It was also commented that in some sets context was missing and might have changed the outcome.

Regarding their experience with assessing (human or machine) translation quality, one participant indicated having used the MQM error typology and one participant having used the MQM-DQA metrics before. Three participants indicated that they do translation quality assessment to a certain degree in translation classes. Several indicated to use some sort of quality assessment on MT or comparing different alternatives for translations of specific sentence parts before deciding which one to use. Participants further indicated that translation quality assessment is fairly subjective except when there are indisputable errors, but also stressful and tiring in general.

General comments on the survey included that the difficulty of selecting and rating the best and worst translations differed for the sets, which is why it would have been better to have the questions on the difficulty for each set rather than in general at the end. Two participants commented that they assume that one of the translations was always human, which they attributed to the fact that it was less faithful/close to the original, arguing that more context would be needed for more reliable decision-making. Since the instructions explicitly stated that the comparison was between human and machine translations, it could be that the one human translation included per set stood out so clearly that this fact became evident to these participants or that this was the expectation by the participants.

5 Discussion

The major objective of the proposed study was to evaluate whether the method of BWS can effectively be applied to a comparative analysis of translation by humans and/or machines. The purpose of this case study was also to show that this method directly leads to a comparative result without the need of any further inter-annotator agreement calculations or scoring methods of the individual participating translation modes. Overall, it can be stated that the results show a very clear preference for human translations and DeepL from the set of selected MT systems. While a particularly unpopular result for a single set can influence the overall rating, the trends for the participating translation modes are still clearly visible from the final overall results. The one extremely negative rating for Google Translate and Microsoft Bing Translator might have contributed to the overall negative rating, however, the fact that both were not selected as often as the other modes contributed just as much. This statement can be particularly reinforced by the fact that also the human translations ranged among the worst for particular sets. Thus, even though individual sets might influence the final result, the overall tendency of being a viable or less preferable translation option can be deduced from the results. The decision to indicate to participants that there is a direct comparison between different translation modes, i.e., human and machine translation, is entirely open for the proposed method and could easily be adapted.

Human translations were overall selected and rated as the best option, however, it should be noted that each set contained a translation by a different professional translator. This can particularly be noticed by the strong differences in ratings across the sets, which, however, could have been the case with the same human translators for each set. Nevertheless, in future work, it would be interesting to repeat the experiment with human translations from a single professional translator or maybe even two human translations in the individual sets, restricting the number of domains to specific sets of expertise.

BWS is considered a perfectly equipped annotation or prioritization method of subjective nature, which means each person can take a subjective decision. Nevertheless, the overall results return tendencies, especially for translation quality, where at times the selection corresponds to a 100%. As indicated by the comments of the participants, a slight variation in wording or a divergence in the selection of just one word can already influence the decision on whether to select a translation option or not and as best or worst. The advantage of BWS is that strong variations in one set still allow for a tendency and trend in the overall results in the end. Without any further context on the topic, participants selected translations that are more faithful to the source text, which in many cases was one of the MT options. As a matter of fact, the lack of context is the most substantial limitation of this case study, limited source and target texts to less than 50 words. Thus, it should be considered for future surveys how BWS can be provided with more context without risking a cognitive and temporal overload of participants. As a method, it still provides a viable alternative to direct assessment with reference translations and ranking methods, especially considering the number of helpful comments left for this case study.

In terms of limitations, it has to be acknowledged that the number and especially scope of evaluated source texts and translations is strongly limited. Only five sets of individual paragraphs were evaluated in this study. In addition, only the language combination of translating from English to German was considered, which is in favor of training settings of major MT systems. Furthermore, the number of participants was limited to 10. While this is a small number, it, however, shows that BWS is adequate for different sizes of participation numbers. In this study, the objective was less to reach a wide audience but rather to make sure participants have a translation background and experience with quality assessment, in order to test this novel method and to obtain feedback on its efficacy and user-friendliness. In this regard, it is within human nature that it is easier to exclude an option we clearly like least, i.e., select the worst option, rather than identifying the best among four options, which is indicated by the ratings and comments of the participants at the end of the survey. While in this study students with translation backgrounds participated, it would be interesting to repeat this study with professional and experienced translators, in which case the domain should be limited to their respective expertise. Nevertheless, in this case study, the level of required expertise and technical vocabulary was intentionally kept at a low level to facilitate participation by language rather than domain experts.

6 Conclusion

The study showed that assessing translation quality with BWS is an easy to implement and understand method, which can be successfully administered without lengthy explanations and returns interesting results. The two major benefits to be expected from BWS for translation quality assessment are time efficiency and subjective decisions. Even though the selected number and sizes of translations was small, the survey also only took between 20 and 35 minutes to be completed. While each participant made individual choices for each set, the subjective decisions still provide an overall tendency on which translation method and origin might be preferable for these domains and text genres. It is interesting that in the comments participants remarked on the fact that this is a highly subjective exercise, which, however, when evaluating the overall results is not negative. Quite to the contrary, with BWS and the rating of each selected best and worst option the results show that an effective and consistent comparison of translation quality can be achieved with this method.

For future research we suggest using longer texts in order to provide more context for the MT systems, as well as to perform studies on a larger scale and with professional translators of more experience. In addition, it would be interesting to directly compare this quality assessment method with previously, state-of-the-art methods related to ranking and rating the overall quality of translation, be it machine and/or human, which is part of our future endeavours. In addition, it is interesting to see how this method can be applied to different types of methods related to machine and/or human translation, such as pre-editing, post-editing, and specific translation strategies.

References

Ammann, Margret. 1990. Anmerkungen zu einer theorie der Übersetzungskritik und ihrer praktischen Anwendung. *TEXTconTEXT*, 5:209–250.

Balducci Paolucci, Angela, Manuel Lardelli, and Dagmar Gromann. 2023. Gender-fair language in translation: A case study. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 13–23.

Banerjee, Satanjeev and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In Bojar, Ondrej, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Herve Saint-Amand, Radu Soricut, and Lucia Specia, editors, *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In Bojar, Ondřej, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark, September. Association for Computational Linguistics.

Carl, Michael and M Cristina Toledo Báez. 2019. Machine translation errors and the translation process: a study across different languages. *Journal of Specialised Translation*, 31:107–132.

Cascetta, Ennio, 2009. *Random Utility Theory*, pages 89–167. Springer US, Boston, MA.

Castilho, Sheila, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to human and machine translation quality assessment. In *Translation Quality Assessment: From Principles to Practice*, volume 1 of *Machine Translation: Technologies and Applications*, pages 9–38. Springer, Cham, Switzerland.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA.

Freitag, Markus, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore, December. Association for Computational Linguistics.

Göpferich, Susanne. 2008. *Textproduktion im Zeitalter der Globalisierung: Entwicklung einer Didaktik des Wissenstransfers*. Studien zur Translation ; 15. Stauffenburg, Tübingen, 3. aufl. edition.

Harris, Kim, Aljoscha Burchardt, Georg Rehm, and Lucia Specia. 2016. Technology landscape for quality evaluation: Combining the needs of research and industry. In *LREC Workshop on Translation Evaluation*, pages 50–54.

Hollin, Ilene L, Jonathan Paskett, Anne LR Schuster, Norah L Crossnohere, and John FP Bridges. 2022. Best–worst scaling and the prioritization of objects in health: a systematic review. *Pharmacoeconomics*, 40(9):883–899.

House, Juliane. 2015. *Translation quality assessment: Past and present*. Routledge.

Jia, Yanfang, Michael Carl, and Xiangling Wang. 2019. How does the post-editing of neural machine translation compare with from-scratch translation? a product and process study. *The Journal of Specialised Translation*, 31(1):60–86.

Likert, Rensis. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12:0455–463.

Louviere, Jordan J and George G Woodworth. 1990. Best worst scaling: A model for largest difference judgments [working paper]. *Faculty of Business*.

Louviere, Jordan J, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Macháček, Matouš and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In Bojar, Ondrej, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Herve Saint-Amand, Radu Soricut, and Lucia Specia, editors, *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria, August. Association for Computational Linguistics.

Mohammad, Saif and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In Ide, Nancy, Aurélie Herbelot, and Lluís Màrquez, editors, *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada, August. Association for Computational Linguistics.

Munkova, Dasa, Michal Munk, Katarina Welnitzova, and Johanna Jakabovicova. 2021. Product and process analysis of machine translation into the inflectional language. *SAGE Open*, 11(4):21582440211054501.

Nord, Christiane. 1997. *Translating as a purposeful activity: Functionalist approaches explained*. Routledge.

Ortiz-Boix, Carla and Anna Matamala. 2017. Assessing the quality of post-edited wildlife documentaries. *Perspectives*, 25(4):571–593.

Paolucci, Angela Balducci, Manuel Lardelli, and Dagmar Gromann. 2023. Gender-fair language in translation: A case study. In Vanmassenhove, Eva, Beatrice Savoldi, Luisa Bentivogli, Joke Daems, and Janiça Hackenbuchner, editors, *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 13–23, Tampere, Finland, June. European Association for Machine Translation.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Popović, Maja. 2018. Error classification and analysis for machine translation quality assessment. In Moorkens, Joss, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*, pages 129–158. Springer International Publishing, Cham.

Popović, Maja. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Reiss, Katharina and Hans J. Vermeer. 1984. *Grundlegung einer allgemeinen Translationstheorie*, volume 147. Max Niemeyer Verlag, Tübingen.

Schuster, Anne L.R., Norah L. Crossnohere, Nicola B. Campoamor, Ilene L. Hollin, and John F.P. Bridges. 2024. The rise of best-worst scaling for prioritization: A transdisciplinary literature review. *Journal of Choice Modelling*, 50:100466.

Specia, Lucia and Kashif Shah. 2018. machine translation quality estimation: applications and future perspectives. In *Translation Quality Assessment: From Principles to Practice*, volume 1 of *Machine Translation: Technologies and Applications*, pages 201–235. Springer, Cham, Switzerland.

Stanco, Marcello, Marco Lerro, and Giuseppe Marotta. 2020. Consumers' preferences for wine attributes: A best-worst scaling analysis. *Sustainability*, 12(7):2819.

Thurstone, Louis L. 1927. A law of comparative judgment. *Psychological review*, 34(4):273–286.

Toral, Antonio and Andy Way, 2018. *What level of quality can neural machine translation attain on literary text?*, pages 263–287. Springer.

Webster, Rebecca, Margot Fonteyne, Arda Tezcan, Lieve Macken, and Joke Daems. 2020. Gutenberg goes neural: Comparing features of dutch human translations with raw neural machine translation outputs in a corpus of english literary classics. *Informatics*, 7(3):32.

Appendix A. Questionnaire

Best-Worst-Scaling NMT vs HT

Q1

Dear participant,

Thank you for participating in this study on the comparison of machine and human translations. The aim of this survey is to assess translation quality of translations produced by machine translation systems and human translators.

For each example, you will be shown the English source text as well as four different translations. In a first step, please choose which of the four options, according to your opinion, is the best translation and which one you consider the worst option. You will then be asked to rate how good the best option is from 4 (highest score) to 0 (lowest score) and how bad the worst option is in your mind from 0 (highest score) to -4 (lowest score). Please also feel free to comment on your choice, e.g. why one translation is better or worse than the others in your opinion.

Completing the survey should take approximately 20 to 25 minutes and is entirely anonymous. Please provide free text comments in English. Your data and the answers you provide will be recorded for scientific purposes only and analyzed and published anonymously.

Thank you for participating and we hope you enjoy it!

Q2

How old are you?

□ 18-24 years

□ 25-34 years

□ 35-44 years

□ 45-54 years

□ 55-64 years

□ Older than 65 years

Q3

What gender do you identify as?

□ Female

□ Male

□ Diverse/Non-binary

□ Prefer not to say

Q4

What is the highest degree you have completed?

□ Bachelor's degree (BA, BSc,...)

□ Master's degree (MA, MSc,...)

□ Diploma (Mag.)

□ Doctorate (Dr., PhD)

□ Other, please specify as a comment:________________

Q5

Please indicate the field in which you obtained that degree.

____

Q6

If you have a degree in translation studies/transcultural communication, please indicate your language combination.

A language ____

B language ____

C language(s) ____

Q7

Please indicate your proficiency in **English** according to the Common European Framework of Reference for Languages (CEFR).

□ C2 - Proficient user (Mastery)

□ C1 - Proficient user (Effective operational proficiency)

□ B2 - Independent user (Vantage)

□ B1 - Independent user (Threshold)

□ A2 - Basic user (Waystage)

Q8

Please indicate your proficiency in **German** according to the Common European Framework of Reference for Languages (CEFR).

□ C2 - Proficient user (Mastery)

□ C1 - Proficient user (Effective operational proficiency)

□ B2 - Independent user (Vantage)

□ B1 - Independent user (Threshold)

□ A2 - Basic user (Waystage)

Q9

How would you rate the amount of experience you have translating?

□ I have a lot of translation experience

□ I have some translation experience

□ I have no translation experience

Q10

What is your current profession?

Q11

How many years of professional experience in the translation sector do you have?

□ None

□ Up to 1 year

□ 1 to 3 years

□ 4 to 8 years

□ More than 8 years

Q12

How often do you use machine translation tools (e.g. DeepL, Google Translate, Microsoft Bing Translator, etc.)?

□ Never

□ Up to 5 times per year

□ Once a month

□ Once per week

□ Several times per week

□ Daily

Q13

What do you use machine translation for? (please select all that apply)

□ Professional translation

□ Other work purposes

□ Privately

Q14

If you use MT, is there an MT system you prefer? If so, please indicate below which ones you prefer.

____

Q15

How satisfied are you with the machine translation results in general?

□ Very satisfied

□ Somewhat satisfied

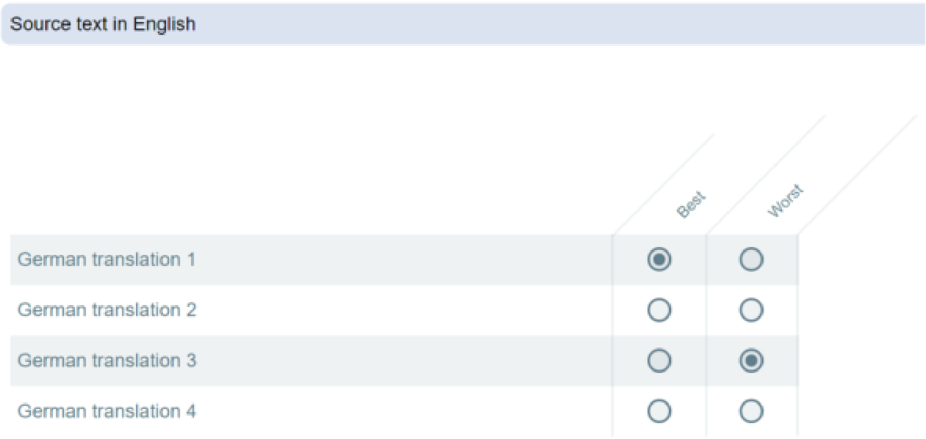
□ Neither satisfied nor not satisfied

□ Not very satisfied

□ Not at all satisfied

Q16

In each of the following five sections, you will be shown a set of four German translations for one English paragraph. Please indicate which of the four translations - according to your opinion - is the **best** translation and which one the **worst** . You will then be asked to rate the chosen options as well as to motivate your choice. **Please do not select the same translation for best and worst option and make sure to only select one option as best and exactly one as worst, which means two options will remain unrated.** Please find below an example of how it should be done (if there is one box selected in each of the columns, this is correct).



Source text in English

	Best	Worst
German translation 1	●	○
German translation 2	○	○
German translation 3	○	●
German translation 4	○	○

Q17

Set 1:

English original: At the top of the Royal Household is the Lord Chamberlain, often likened to a non-executive chairman (it has yet to be a woman). He is appointed on a part-time basis to oversee the whole operation.

	Best	Worst
An der Spitze des königlichen Haushalts steht der Lord Chamberlain, der oft mit einem nicht-exekutiven Vorsitzenden verglichen wird (es ist noch keine Frau). Er wird auf Teilzeitbasis ernannt, um den gesamten Betrieb zu beaufsichtigen.	□	□
An der Spitze des Britischen Hofes steht der Lord Chamberlain, oft verglichen mit einem	□	□

	Best	Worst
nicht geschäftsführenden Vorsitzenden (eine Frau konnte sich für dieses Amt noch nicht durchsetzen). Er wird auf Teilzeitbasis eingestellt, um den gesamten Betrieb zu leiten.		
An der Spitze des Königshauses steht der Lord Chamberlain, der oft mit einem nicht geschäftsführenden Vorstandsvorsitzenden verglichen wird (eine Frau hat es noch nicht gegeben). Er wird auf Teilzeitbasis ernannt, um den gesamten Betrieb zu überwachen.	□	□
An der Spitze des königlichen Haushalts steht der Lord Chamberlain, der oft mit einem nicht geschäftsführenden Vorsitzenden verglichen wird (eine Frau ist es bisher noch nicht). Er wird auf Teilzeitbasis ernannt, um den gesamten Betrieb zu überwachen.	□	□

Q18

On a scale from **0 (lowest score)** to **4 (highest score)** , how good would you rate the translation you selected to be the **best** one?

□ 4 (highest score)

□ 3

□ 2

□ 1

□ 0 (lowest score)

Q19

On a scale from **0 (highest score)** to **-4 (lowest score)** , how bad would you rate the translation you selected to be the **worst** one?

□ 0 (highest score)

□ -1

□ -2

□ -3

□ -4 (lowest score)

Q20

Pleasecomment on what makes the one translation the best and the other the worst in your opinion. Feel free to add any other commentsyou would like to share on this text or the translations.

Q21

Set 2:

English original: My interest in books probably explains why I not only survived high school but arrived at Occidental College in 1979 with a thin but passable knowledge of political issues and a series of half-baked opinions that I'd toss out during late-night bull sessions in the dorm.

	Best	Worst
Mein Interesse an Büchern erklärt wahrscheinlich, warum ich nicht nur die Highschool überlebte, sondern 1979 auch mit einem dürftigen, aber passablen Wissen über	□	□

Best-Worst-Scaling NMT vs HT

	Best	Worst
politische Themen und einer Reihe unausgegorener Meinungen, die ich während der nächtlichen Bullensitzungen von mir gab, am Occidental College ankam der Schlafraum.		
Mein Interesse an Büchern erklärt vermutlich, warum ich nicht nur die Highschool überstand, sondern 1979 beim Eintritt ins Occidental College über ein zwar dünnes, aber einigermaßen passables Politikwissen verfügte und ein paar halb gare Ansichten entwickelt hatte, die ich bei nächtlichen Diskussionsrunden im Studentenwohnheim zum Besten gab.	□	□
Mein Interesse an Büchern erklärt wahrscheinlich, warum ich nicht nur die High School überlebte, sondern 1979 auch mit einem dünnen, aber passablen Wissen über politische Themen und einer Reihe von unausgegorenen Meinungen am Occidental College ankam, die ich während nächtlicher Bullensitzungen im Wohnheim ausstieß.	□	□
Mein Interesse an Büchern erklärt wahrscheinlich, warum ich nicht nur die Highschool überlebte,	□	□

	Best	Worst
sondern 1979 am Occidental College ankam, mit einem dünnen, aber passablen Wissen über politische Themen und einer Reihe halbfertiger Meinungen, die ich während der nächtlichen Bullensitzungen im Studentenwohnheim in die Runde warf.		

Q22

On a scale from **0 (lowest score)** to **4 (highest score)** , how good would you rate the translation you selected to be the **best** one?

□ 4 (highest score)

□ 3

□ 2

□ 1

□ 0 (lowest score)

Q23

On a scale from **0 (highest score)** to **-4 (lowest score)** , how bad would you rate the translation you selected to be the **worst** one?

□ 0 (highest score)

□ -1

□ -2

□ -3

□ -4 (lowest score)

Q24

Best-Worst-Scaling NMT vs HT

Pleasecomment on what makes the one translation the best and the other the worst in your opinion. Feel free to add any other commentsyou would like to share on this text or the translations.

Q25

Set 3:

English original: The launch of Bitcoin set off a revolution in blockchain and cryptocurrency. There are now more than 13,000 different cryptocurrencies. (Most, be warned, are essentially valueless and will remain that way.)

	Best	Worst
Der Start des Bitcoin-Netzwerks löste eine wahre Blockchain- und Kryptowährungsrevolution aus. Inzwischen gibt es über 13.000 verschiedene Kryptowährungen. (Achtung, die meisten davon sind im Wesentlichen wertlos und werden es auch bleiben!)	□	□
Die Einführung von Bitcoin löste eine Revolution in der Blockchain und Kryptowährung aus. Mittlerweile gibt es mehr als 13.000 verschiedene Kryptowährungen. (Die meisten, seien Sie gewarnt, sind im Wesentlichen wertlos und werden es auch bleiben.)	□	□
Die Einführung von Bitcoin löste eine Revolution im Bereich Blockchain und Kryptowährung aus. Inzwischen gibt es mehr als 13.000 verschiedene	□	□

	Best	Worst
Kryptowährungen. (Die meisten, seien Sie gewarnt, sind im Wesentlichen wertlos und werden es auch bleiben.)		
Die Einführung von Bitcoin löste eine Revolution in der Blockchain und Kryptowährung aus. Mittlerweile gibt es mehr als 13.000 verschiedene Kryptowährungen. (Die meisten, seien Sie gewarnt, sind im Wesentlichen wertlos und werden es auch bleiben.)	□	□

Q26

On a scale from **0 (lowest score)** to **4 (highest score)** , how good would you rate the translation you selected to be the **best** one?

□ 4 (highest score)

□ 3

□ 2

□ 1

□ 0 (lowest score)

Q27

On a scale from **0 (highest score)** to **-4 (lowest score)** , how bad would you rate the translation you selected to be the **worst** one?

□ 0 (highest score)

□ -1

□ -2

□ -3

□ -4 (lowest score)

Q28

Pleasecomment on what makes the one translation the best and the other the worst in your opinion. Feel free to add any other commentsyou would like to share on this text or the translations.

Q29

Set 4:

English original: You can think about energy efficiency like a budget. A financial budget tracks money as it’s earned and spent. A budget for your body similarly tracks resources like water, salt, and glucose as you gain and lose them.

	Best	Worst
Sie können sich die Energieeffizienz wie ein Budget vorstellen. In einem Finanzbudget werden die Einnahmen und Ausgaben des Geldes erfasst. Ein Budget für Ihren Körper erfasst in ähnlicher Weise Ressourcen wie Wasser, Salz und Glukose, wie Sie sie zu- und abnehmen.	□	□
Am besten stellen Sie sich die Frage der Energieeffizienz so vor, als würden Sie ein Haushaltsbuch führen: Dabei notieren Sie, wie viel Geld hereinkommt und wie viel ausgegeben wird. Für Ihren Körper heißt das, dass Sie Ressourcen wie Wasser, Salz und Glukose eintragen und festhalten,	□	□

	Best	Worst
wie viel Sie davon aufnehmen oder verbrauchen.		
Sie können sich Energieeffizienz wie ein Budget vorstellen. Ein Finanzbudget verfolgt, wie Geld verdient und ausgegeben wird. Ein Budget für Ihren Körper verfolgt in ähnlicher Weise Ressourcen wie Wasser, Salz und Glukose, während Sie sie gewinnen und verlieren.	□	□
Sie können sich Energieeffizienz wie ein Budget vorstellen. Ein Finanzhaushalt erfasst das verdiente und ausgegebene Geld. Ein Budget für Ihren Körper erfasst in ähnlicher Weise Ressourcen wie Wasser, Salz und Glukose, während Sie diese aufnehmen und verlieren.	□	□

Q30

On a scale from **0 (lowest score)** to **4 (highest score)** , how good would you rate the translation you selected to be the **best** one?

□ 4 (highest score)

□ 3

□ 2

□ 1

□ 0 (lowest score)

Q31

On a scale from **0 (highest score)** to **-4 (lowest score)** , how bad would you rate the translation you selected to be the **worst** one?

- □ 0 (highest score)
- □ -1
- □ -2
- □ -3
- □ -4 (lowest score)

Q32

Pleasecomment on what makes the one translation the best and the other the worst in your opinion. Feel free to add any other commentsyou would like to share on this text or the translations.

Q33

Set 5:

English original: Physics in the late 19th and early 20th centuries was concerned with the spectrum of light being emitted by black bodies. A black body is a piece of material that radiates corresponding to its temperature — but it also absorbs and reflects light from its surroundings.

	Best	Worst
Die Physiker beschäftigten sich im späten 19. und frühen 20. Jahrhundert vor allem mit dem Lichtspektrum, das von sogenannten schwarzen Körpern ausgesendet wird. Ein schwarzer Körper ist ein Stoff, der wie alle anderen Körper seiner Temperatur entsprechend strahlt, aber auch Licht	□	□

Best-Worst-Scaling NMT vs HT

	Best	Worst
aus seiner Umgebung absorbiert und reflektiert.		
Die Physik des späten 19. und frühen 20. Jahrhunderts beschäftigte sich mit dem Lichtspektrum, das von schwarzen Körpern abgestrahlt wird. Ein schwarzer Körper ist ein Stück Material, das entsprechend seiner Temperatur strahlt - aber auch Licht aus seiner Umgebung absorbiert und reflektiert.	□	□
Die Physik des späten 19. und frühen 20. Jahrhunderts beschäftigte sich mit dem Spektrum des Lichts, das von Schwarzen Körpern emittiert wird. Ein schwarzer Körper ist ein Stück Material, das entsprechend seiner Temperatur strahlt – aber es absorbiert und reflektiert auch Licht aus seiner Umgebung.	□	□
Die Physik im späten 19. und frühen 20. Jahrhundert befasste sich mit dem Spektrum des von schwarzen Körpern emittierten Lichts. Ein schwarzer Körper ist ein Stück Material, das entsprechend seiner Temperatur strahlt – aber auch Licht aus seiner Umgebung absorbiert und reflektiert.	□	□

Q34

On a scale from **0 (lowest score)** to **4 (highest score)** , how good would you rate the translation you selected to be the **best** one?

□ 4 (highest score)

□ 3

□ 2

□ 1

□ 0 (lowest score)

Q35

On a scale from **0 (highest score)** to **-4 (lowest score)** , how bad would you rate the translation you selected to be the **worst** one?

□ 0 (highest score)

□ -1

□ -2

□ -3

□ -4 (lowest score)

Q36

Pleasecomment on what makes the one translation the best and the other the worst in your opinion. Feel free to add any other commentsyou would like to share on this text or the translations.

Q37

How easy/difficult was selecting the best translation in general for you?

very easy	somewhat easy	neutral	somewhat difficult	very difficult
□	□	□	□	□

Q38

How easy/difficult was rating the best translation for you?

very easy	somewhat easy	neutral	somewhat difficult	very difficult
□	□	□	□	□

Q39

How easy/difficult was selecting the worst translation in general for you?

very easy	somewhat easy	neutral	somewhat difficult	very difficult
□	□	□	□	□

Q40

How easy/difficult was rating the worst translation for you?

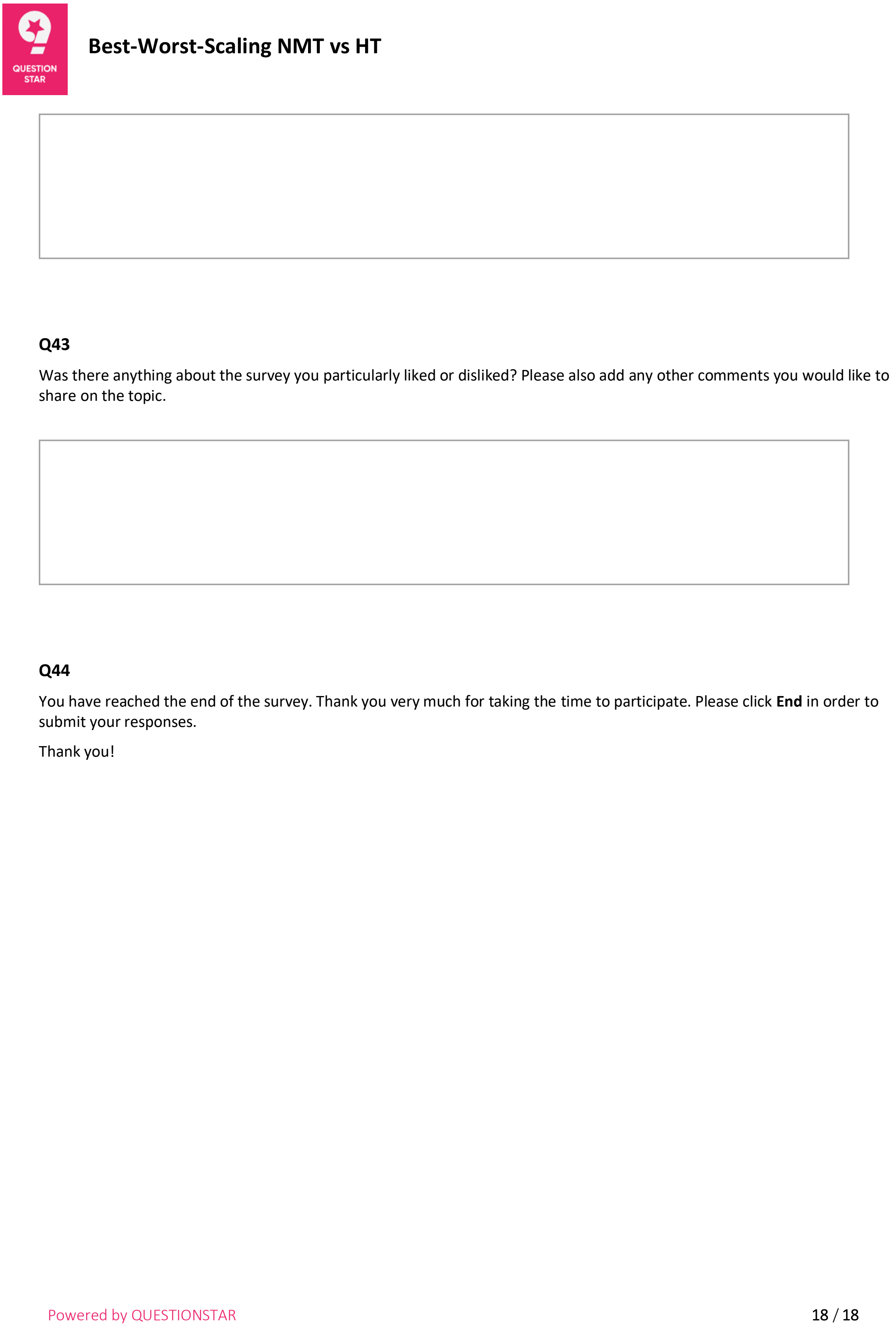
very easy	somewhat easy	neutral	somewhat difficult	very difficult
□	□	□	□	□

Q41

What do you think about this comparative approach of picking out the best and the worst translation?

Q42

What is your experience with assessing (human or machine) translation quality?



Best-Worst-Scaling NMT vs HT

Q43

Was there anything about the survey you particularly liked or disliked? Please also add any other comments you would like to share on the topic.

Q44

You have reached the end of the survey. Thank you very much for taking the time to participate. Please click **End** in order to submit your responses.

Thank you!