QUALITY OF EXPERIENCE OF GERMAN MACHINE TRANSLATION AND AUTOMATIC TEXT SUMMARIZATION

Shushen Manakhimova[1], *Vivien Macketanz*[1], *Sebastian Möller*[1,2]

[1]*Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI),* [2]*Technische Universität Berlin*

shushen.manakhimova@dfki.de

Abstract: Quality of Experience measures the user's subjective perception of an interaction with a system or product—a framework widely applied in media technology, We extend the application of Quality of Experience to machine-generated text, focusing on identifying and validating quality dimensions. We investigate two text-generation tasks: Machine Translation and Automatic Text Summarization. This work centers on empirical studies, leveraging human evaluation to quantify perceived quality dimensions for both text types.

1 Introduction

Machine-generated text plays a crucial role in applications such as chatbots, translation services, and text-to-speech systems. However, its impact on user satisfaction is often overlooked. Although established metrics such as BLEU, ROUGE, and METEOR are invaluable for benchmarking, they often show weak correlations with human judgments [1, 2, 3] and neglect the subjective aspects of user satisfaction and real-world usability. Understanding the subjective perception of the quality of these outputs remains largely unexplored.

Recent work in Machine Translation (MT) and summarization evaluation have shifted focus toward human-centric approaches. Frameworks such as InteractEval [4] and HOPE [5] integrate human judgments into the assessment process, yet they do not explicitly focus on perception or satisfaction.

One promising alternative is to adopt a Quality of Experience (QoE) — a framework that evaluates user satisfaction beyond mere technical performance [6]. Traditionally employed in multimedia domains such as speech, audio, and video communications [7], QoE measures the user's subjective perception of their interaction with a system or product. This user-centric framework is equally relevant to text-based systems, particularly as LLM-generated text becomes omnipresent across various applications.

To successfully adapt the QoE framework to machine-generated text, we first need to understand the dimensions that shape the experience. In this work, we aim to determine the aspects of machine-generated text that affect its perceived quality. Specifically, we focus on two prominent text-generation tasks: MT and Automatic Text Summarization (ATS).

2 Related Work

QoE research integrates perspectives from multiple fields, including social psychology, cognitive science, economics, and engineering. In QoE quality is defined as a multidimensional construct that extends beyond objective performance, with dimensions emerging dynamically from real-world interactions [8]. In this context, QoE is seen as an event within a perceptual space, with psychophysical methods used to extract and quantify its dimensions [9].

Among the cornerstone methods in QoE research are subjective tests. Methods like Semantic Differential (SD) scaling—which uses bipolar scales with antonym labels [10]—and the Mean Opinion Score (MOS) are widely used to measure overall satisfaction in multimedia applications [11]. Factor analysis further helps reveal latent quality dimensions.

These subjective evaluation methods have already proven effective in evaluating textual data. For instance, an automated readability assessment for German sentences, leveraging a corpus annotated by second-language learners on a 7-point Likert scale was developed in [12]. Each sentence was rated by 10 or more participants, and the MOS was calculated to quantify readability. While the work in [12] focused on readability as a specific linguistic property, we extend these subjective evaluation techniques to broader dimensions of text quality.

Our previous work [8] investigated perceptual quality dimensions in MT, identifying precision, complexity, grammaticality, and transparency as key factors influencing user satisfaction. The current paper extends the research to the quantification of the previously identified quality dimensions, as well as another text type. By leveraging methods such as crowdsourcing, Semantic Differential (SD) scaling, and Exploratory Factor Analysis (EFA), we develop a framework for evaluating machine-generated text through a user-centric lens.

3 Empirical Studies

We have conducted experimental research in the form of several crowdsourcing studies with German-speaking participants to identify and verify the relevant quality dimensions for the two text types MT and ATS.

3.1 Datasets

3.1.1 Machine Translation Corpus

For the text type MT, we created a corpus that was based on translations from English to German that were generated for the News Translation task of the Conference on Machine Translation 2019 (WMT19)[1]. To create a corpus that contains a wide quality range of translations with a variety of translation error types, we selected translations from the systems that had been ranked best, worst, and mediocre in the shared task [13]. Following the system selection, we selected a subset of all the systems' translations as our corpus. Then, a simple error-type annotation was conducted by several linguists. The annotation did not follow any error type classification but rather focused on the most prominent and frequent errors observed in the translations. This was done to ensure that our dataset consisted of sentences with various error types and severities.

This approach ensured that our dataset included sentences with diverse errors and varying degrees of severity. We also made sure that the test sentences included translations of varying lengths based on word count. These steps helped to maintain a well-balanced corpus containing translations with varying levels of correctness and sentence lengths.

3.1.2 Automatic Text Summarization Corpus

The dataset for the text type ATS consists of German summaries of the GeWiki corpus of varying quality[2]. We obtained the summaries from two sources. The first source is the SwissText & KONVENS conference 2020[3]. The conference organizers generously provided these summaries for our research. The second source consists of internally generated summaries. The

[1] http://www.statmt.org/wmt19/index.html

[2] https://github.com/domfr/GeWiki

[3] https://swisstext-and-konvens-2020.org/

summaries from both sources were produced using a range of extractive and abstraction summarization approaches. Extractive methods, such as Lead-3 [14], select the first three sentences of a document as a summary, while TextRank [15] is an unsupervised graph-based ranking algorithm that identifies the most important sentences in a text. Abstractive methods include Pointer-Generator [16], which combines copying from the source text with novel word generation, as well as Transformer-based models like Transformer [17] and Transformer SE, which utilize self-attention mechanisms to capture long-range dependencies and improve summary coherence. Additionally, the dataset includes Convolutional Self-Attention Transformer [18] and BERT-Transformer [19]. Analogously to the corpus creation for MT, we performed a simple error type analysis on the summaries to select a subset of summaries with varying error types and error severities for the final ATS corpus. We also made sure to include summaries of different lengths, ranging from one to five sentences, resulting in a well-balanced corpus containing summaries with varying levels of correctness and lengths.

3.1.3 Polar Adjective Pairs

For our crowdsourcing experiments, we additionally needed two sets of polar adjective pairs relating to the two text types. These adjective pairs were used in the experiments to evaluate the summaries and translations. Therefore, we needed adjective pairs that could reflect the characteristics of the machine-generated texts. The adjective pairs always consisted of a positive adjective and its polar negative complement, e.g., *grammatical - ungrammatical*. As the characteristics of translations and summaries differ from one another, we also created distinct sets per text type. The sets can be found in Tables 4 and 5 in the Appendix. The creation of the adjective sets involved several steps: Referring to our error type analysis of the MT and ATS corpora, we identified a list of adjective pairs that best depicted the errors present in the texts per text type. The selection of relevant adjective pairs was also conducted with the help of linguists to ensure they accurately reflected linguistic nuances. The list contained about 40 adjective pairs per text type, with some but not all overlapping between the text types. The goal was to create a set of adjective pairs that should cover as many aspects of the language of the texts as possible. Therefore, the meaning of some of the pairs (partly) overlapped.

We then conducted a small-scale pre-study, instructing the participants to evaluate the language of the presented texts from our corpora with the help of all adjective pairs. In the second part of the pre-study, the participants had to evaluate how useful they found the adjective pairs for the task they had just completed. Based on the result of this evaluation, the adjective sets were both reduced to around 20 adjective pairs each which were then used for the crowdsourcing studies.

3.2 Quality Dimension Identification

3.2.1 Experiment Set-up

We conducted two quantitive crowdsourcing studies (one per text type MT and ATS) to identify the relevant quality dimensions for the QoE. The studies were based on an evaluation employing an SD: The participants were presented translations or summaries along with the instruction to evaluate the language of the texts with the help of the adjective sets. Each adjective in a pair represented the endpoints on a Likert scale from -3 to +3. The instructions specified that the content of the texts should be left out of the evaluation—to the extent that it is possible to separate content from language. Furthermore, the participants were not informed that the texts they were seeing had been generated by machines, but only that the texts might or might not contain errors.

The structure of the surveys was as follows: (1) There was an introduction to the experiment, explaining the set-up and providing an example of a polar adjective pair. (2) The introduction was followed by a test text that had to be evaluated, which served two purposes: Firstly, the participants could gain a good understanding of the set-up of the experiment. Second, the test texts were either distinctly high or low in quality. If participants selected a value for the test adjective pair that did not align with the text's quality, they were prompted to reconsider their evaluation. This feedback created the impression of "being observed" [20], resembling the Hawthorne Effect in psychology, where study participants alter their behavior in response to observation [21]. (3). In the main part of the experiment, each text was presented separately and had to be evaluated on all 20 adjective pairs before the next text was presented. For each adjective pair, a slider had to be set to an integer number between -3 and +3 to represent the participant's perception of the text. Each participant was presented with three texts in total. (4) After the evaluation part, the participants were given the possibility to provide feedback on the survey.

The survey completion was expected to take around 10 minutes. We created around 15 versions of each of the surveys with 45 translations and 40 summaries altogether. In total, 350 participants completed the survey for the MT texts and 425 for the ATS texts. The difference in participant numbers between the two text types is due to the need to rerun the survey when there were too many unusable ratings. The experiment was carried out on the Crowdee platform[4], a mobile crowdsourcing system that supports multiple languages, including German [22]. This ensured that the survey interface was fully localized for our target audience, improving user experience and understanding. Participants had to be on a native speaker level of German to conduct the survey. They could participate up to five times in the study and were presented with different versions of the survey if they did participate multiple times. According to the platform's filters, participants were (a) self-identified native German speakers and (b) living in Germany, Austria, or Switzerland (the DACH region). Since participants chose to join and needed internet access, the sample might include younger or tech-savvy people and might not fully reflect the broader German-speaking population in the DACH region.

3.2.2 Analysis and Results

Before the result data could be analyzed, it had to be cleaned. One shortcoming of crowdsourcing studies is the risk of noisy data caused by crowdworkers who do not work diligently [20]. Therefore, the cleaning process is an important initial step. Our process consisted of the following phases: We first eliminated all ratings of a participant if they completed the survey in 240 seconds or less (40% of the expected 10 minutes). We then excluded all ratings of a participant if they had selected the same value for every adjective pair of a sentence as we assumed these participants were not filling out the survey conscientiously. In the last step, we calculated the inconsistency score (IS) [23]. To calculate the IS, we incorporated test conditions in the surveys by repeating two adjective pairs per sentence. While the degree of variance in human evaluation of translation is high [24], the IS allows for filtering out outliers that show a higher degree of variance than expected under ordinary conditions. Cleaning the data resulted in around ten to 20 ratings per test item.

To reveal the common factors that explain the correlation among the adjective pairs, we conducted an EFA per text type in SPSS [25]. We used Maximum Likelihood as the extraction method and PROMAX with Kaiser Normalization as the rotation method, leading to non-orthogonal dimensions. The analysis revealed that for both text types, several adjective pairs showed low communalities and/or cross-loadings with a difference of less than 0.2. We as-

[4] https://www.crowdee.com/

sumed that these pairs are too close to one another in their meaning or generally not specific enough; Therefore, we removed those attributes to balance the statistical goodness-of-fit and the interpretability of the resulting dimensions [26].

For both text types, the dimension reduction resulted in four factors with eight adjective pairs. For MT, Pearson's chi-squared test for the goodness-of-fit was p = 0.36 (χ^2 = 2.06, df = 2), and for ATS, it was p = 0.63 (χ^2 = 0.92, df = 2).

The distribution of the adjective pairs on the four factors and the explained percentage of variance can be seen in Tables 1 and 2. Note that the adjective pairs are translated into English for better understanding.

	F1	F2	F3	F4
unambiguous – ambiguous	,757			
precise – vague	,947			
complete – incomplete	,822			
clear – chaotic	,580			
direct – ponderous		,806		
simple – complicated		,923		
grammatical – ungrammatical			,958	
neat – confusing				,915
% of variance	53,2	8,4	10,5	8,0

Table 1 – Loadings of adjective pairs (English translations) on factors and % of explained variance for Machine Translation.

	F1	F2	F3	F4
precise – vague	.884			
complete – incomplete	,942			
coherent – incoherent	,844			
logical – illogical	,796			
simple – complicated		1,002		
straightforward – complex		,783		
unambiguous – ambiguous			,729	
predictable – unpredictable				,704
% of variance	54,4	14,3	10,6	2,6

Table 2 – Loadings of adjective pairs (English translations) on factors and % of explained variance for Automatic Text Summarization.

Each of the four factors F1 to F4 for both text types refers to an underlying quality dimension with the respective adjective pair(s) loading on each dimension. For MT, all four attributes loading on F1 refer to aspects of *Precision*; The two attributes loading on F2 refer to aspects of *Complexity*; The single attributes loading on F3 and F4 refer to aspects of *Grammaticality* and *Transparency*. For ATS, the four attributes loading on F1 refer to aspects of *Linguistic Logic*; The two attributes loading on F2 refer to aspects of *Complexity*, which is the only quality dimension shared by both text types. The single attributes loading on F3 and F4 refer to aspects of *Clarity* and *Predictability*.

Table 3 gives an overview of typical traits that we connect with the quality dimensions.

3.3 Quality Dimension Quantification

In a follow-up experiment, we repeated the crowdsourcing studies for both text types with a smaller set of adjective pairs for the SD to confirm the identified quality dimensions. The adjective pairs with the highest loadings per factor were chosen as representatives for the respective quality dimensions revealed by the EFA, resulting in four adjective pairs per text type. We then correlated the results of the two experiments per text type to quantify the quality dimensions and test the validity and reliability of the framework. In this experiment, 425 participants rated the translations, and 120 participants rated the summaries. Again, the variation in participant numbers results from unusable ratings, which necessitated multiple experiment runs to gather sufficient total ratings.

The steps we took for the correlation entailed the same stages for both text types. We first conducted Grubbs's test [27] to identify outliers and excluded sentences from the correlation if they were significant outliers in two or more of the four factors. After this initial step of filtering, all Spearman values were around 0.8 for both text types, which was the first indicator of a high correlation between the two groups. In the next step, we wanted to test if there was

Factor	Quality dimension	Characteristics
MT – F1	Precision	clear and complete phrasing unambiguous meaning
MT & ATS – F2	Complexity	easily comprehensible not circumlocutory
MT – F3	Grammaticality	correct spelling and punctuation no missing words
MT – F4	Transparency	clear and coherent reasonably structured
ATS – F1	Linguistic Logic	accurate phrasing cohesive
ATS – F3	Clarity	direct and clear language content easily understandable
ATS – F4	Predictability	logical and expected structure methodical and coherent

Table 3 – Quality dimensions and their characteristics.

a significant difference between the groups. We therefore first conducted the Jarque–Bera test [28] which revealed that the data was normally distributed across all factors in all text types. Then, we performed Levene's test [29] to assess the equality of variances. For all factors but F2 of MT and F4 of ATS, Levene's test uncovered that the variances were equal, therefore, we performed a one-factor ANOVA (analysis of variance) [30] on those factors. For the factors that did not show homogeneity of variance, we calculated Welch's t-test [31] instead of an ANOVA.

For ATS, the analysis revealed that there was no statistically significant difference between the two groups for either factor. For MT, our statistical analysis revealed that for all factors but F2 (complexity), there was no statistically significant difference between the identification and quantification experiment. However, the difference between the two groups in F2 is 0.5 and thus relatively small on a 7-point scale. Furthermore, the correlation between the two groups on this factor remains high. Since we cannot conclusively determine whether the significant difference arose from the varying number of presented adjective pairs or simply from the different crowdworkers who participated in the studies, we consider these differences as tolerable.

Our findings suggest that the reduced set of adjective pairs captures the key aspects of the quality dimensions well while remaining consistent and reliable across both experiments. Validating these quality dimensions allows for a simpler and more efficient evaluation process in future studies, reducing the mental effort of participants while still providing detailed analysis. The strong correlations across text types show that these dimensions work well for both MT and ATS. The difference in F2 for MT might indicate that complexity is seen differently in translations than in summaries. For example, translations might need simpler language to stay fluent, while summaries might focus more on structural and conceptual complexity to convey concise information. This highlights the need to adjust quality evaluation frameworks to fit different NLP tasks.

4 Results

Using Exploratory Factor Analysis (EFA) and structured crowdsourcing, our analysis identified four quality dimensions for each text type,

For MT, the four quality dimensions identified are: **Precision**, which reflects clear, complete, and unambiguous phrasing; **Complexity**; **Grammaticality**, related to correct spelling,

punctuation, and sentence structure, and **Transparency**, relevant to coherence and clarity of organization.

For ATS, the identified quality dimensions are: **Linguistic Logic**, capturing accurate phrasing and cohesive structure, which explained the largest variance; **Complexity**, similarly identified in MT; **Clarity**, relating to easily understandable language; and **Predictability**, representing logical structure.

These results reveal both shared and task-specific dimensions between MT and ATS. **Complexity** is the only dimension shared across both text types, while the other dimensions are unique to their respective tasks.

A second round of crowdsourcing, employing a reduced set of four polar adjective pairs per dimension, confirmed the reliability of these findings. Spearman correlation values of approximately 0.8 and ANOVA results showing only a minor 0.5 difference for MT's Complexity on a 7-point scale demonstrate high consistency and robustness. These findings provide a foundation for applying the QoE framework to text evaluation.

We believe that text quality in MT or summarization pipelines may directly impact user satisfaction in NLP applications such as speech-to-speech translation and conversational agents. Applying QoE methods to both audio or video and text in these applications can help establish more comprehensive evaluation standards for NLP systems.

5 Future Work

Building on the results outlined in this paper, our next step will be to develop a prediction model for estimating user-perceived text quality in MT and ATS systems. We are currently working on a baseline regression model. In the meantime, we are working on identifying and selecting the relevant linguistic features that correspond to each quality dimension defined in our study.

Following this, we plan to fine-tune or train an LLM for the task. A key component of our research will involve comparing the performance of the regression-based model with the LLM-based approaches to determine which method yields better results in capturing and evaluating these subjective quality dimensions.

In the future, we aim to extend this approach to other NLP tasks and usage contexts. Such predictive tools could provide a cost-effective alternative to purely empirical assessments.

Additionally, we are expanding our datasets to incorporate LLM-generated text to reflect the growing prevalence of machine-generated content. To ensure that these models are grounded in a robust empirical foundation, we will conduct controlled laboratory and crowdsourcing experiments to evaluate the reliability and stability of the identified quality dimensions, particularly for LLM-generated data.

Our ultimate goal is for other researchers and practitioners to adopt or adapt our framework for additional language pairs, text types, and contexts, thereby contributing to standardization efforts in the evaluation of machine-generated texts.

6 Limitations

While this study successfully identified and validated quality dimensions for evaluating MT and ATS, further exploration is needed to broaden the scope and applicability of our findings. Future research should examine correlations and interactions among these dimensions to better understand their combined impact on perceived quality. Additionally, validating these dimensions in diverse contexts—across different languages, text types, and evaluation scenarios—will further establish the framework's robustness and aid in standardizing text evaluation methodologies.

Acknowledgments

Firstly, we would like to thank Dominik Frefel, Manfred Vogel, and Fabian Märki for generously providing their corpus of GeWiki summaries they created for the SwissText & KONVENS conference 2020. Furthermore, we are thankful to all our colleagues who allocated some time to participate in our pre-study. Lastly, we are grateful to our colleague Aleksandra Gabryszak who generously provided her time to share her extensive knowledge about automatic text summarization with us.

The present research was funded by the Deutsche Forschungsgemeinschaft (DFG) through the project "Analyse und automatische Abschätzung der Qualität maschinell generierter Texte", project number 436813723.

References

[1] REITER, E.: *A structured review of the validity of BLEU*. *Computational Linguistics*, 44(3), pp. 393–401, 2018. doi:10.1162/coli_a_00322. URL https://aclanthology.org/J18-3002.

[2] CALLISON-BURCH, C., M. OSBORNE, and P. KOEHN: *Re-evaluating the role of Bleu in machine translation research*. In D. MCCARTHY and S. WINTNER (eds.), *11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 249–256. Association for Computational Linguistics, Trento, Italy, 2006. URL https://aclanthology.org/E06-1032/.

[3] LOMMEL, A.: *Blues for bleu : Reconsidering the validity of reference-based mt evaluation*. 2016. URL https://api.semanticscholar.org/CorpusID:16858249.

[4] CHU, S., J. KIM, and M. YI: *Think together and work better: Combining humans' and llms' think-aloud outcomes for effective text evaluation*. 2024. URL https://arxiv.org/abs/2409.07355. 2409.07355.

[5] GLADKOFF, S. and L. HAN: *HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation*. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK, and S. PIPERIDIS (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 13–21. European Language Resources Association, Marseille, France, 2022. URL https://aclanthology.org/2022.lrec-1.2/.

[6] MÖLLER, S. and A. RAAKE: *Quality of experience: Advanced concepts, applications and methods*. Springer, 2014.

[7] BRUNNSTRÖM, K., K. DE MOOR, A. DOOMS, S. EGGER-LAMPL, M.-N. GARCIA, T. HOSSFELD, S. JUMISKO-PYYKKÖ, C. KEIMEL, C. LARABI, B. LAWLOR, P. LE CALLET, S. MÖLLER, F. PEREIRA, M. PEREIRA, A. PERKIS, A. PINHEIRO, U. REITER, P. REICHL, R. SCHATZ, and A. ZGANK: *Qualinet White Paper on Definitions of Quality of Experience*. 2013.

[8] MACKETANZ, V., B. NADERI, S. SCHMIDT, and S. MÖLLER: *Perceptual quality dimensions of machine-generated text with a focus on machine translation*. In A. BELZ, M. POPOVIĆ, E. REITER, and A. SHIMORINA (eds.), *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pp. 24–31. Association for Computational Linguistics, Dublin, Ireland, 2022. doi:10.18653/v1/2022.humeval-1.3. URL https://aclanthology.org/2022.humeval-1.3.

[9] MÖLLER, S., M. WÄLTERMANN, and M.-N. GARCIA: *Features of Quality of Experience*, pp. 73–84. 2014. doi:10.1007/978-3-319-02681-7_5.

[10] OSGOOD, C. E.: *The Measurement of Meaning*. University of Illinois Press, Urbana,, 1957.

[11] *Itu-t rec. p.10/g.100 (11/2017) vocabulary for performance, quality of service and quality of experience*. 2017. URL https://api.semanticscholar.org/CorpusID:210127568.

[12] NADERI, B., S. MOHTAJ, K. ENSIKAT, and S. MÖLLER: *Subjective assessment of text complexity: A dataset for german language*. 2019. URL https://arxiv.org/abs/1904.07733. 1904.07733.

[13] BARRAULT, L., O. BOJAR, M. R. COSTA-JUSSÀ, C. FEDERMANN, M. FISHEL, Y. GRAHAM, B. HADDOW, M. HUCK, P. KOEHN, S. MALMASI, C. MONZ, M. MÜLLER, S. PAL, M. POST, and M. ZAMPIERI: *Findings of the 2019 Conference on Machine Translation (WMT19)*. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 1–61. Association for Computational Linguistics, Florence, Italy, 2019. URL `http://www.aclweb.org/anthology/W19-5301`.

[14] DOHARE, S. and H. KARNICK: *Text summarization using abstract meaning representation*. 2017. URL `https://api.semanticscholar.org/CorpusID:260498172`.

[15] MIHALCEA, R. and P. TARAU: *Textrank: Bringing order into texts*. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411. 2004.

[16] SEE, A., P. J. LIU, and C. D. MANNING: *Get to the point: Summarization with pointer-generator networks*. *arXiv preprint arXiv:1704.04368*, 2017.

[17] VASWANI, A., N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, L. KAISER, and I. POLOSUKHIN: *Attention is all you need*. In *Advances in Neural Information Processing Systems*, pp. 5998–6008. 2017.

[18] LI, K., Y. DENG, and Y. KIM: *Convolutional self-attention networks*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 404–413. 2019.

[19] DEVLIN, J., M.-W. CHANG, K. LEE, and K. TOUTANOVA: *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. 2019.

[20] NADERI, B., I. WECHSUNG, and S. MÖLLER: *Effect of being observed on the reliability of responses in crowdsourcing micro-task platforms*. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 1–2. IEEE, 2015.

[21] MELAMED, A. W.: *The hawthorne studies: A review of the literature*. *The Journal of Psychology*, 123(6), pp. 669–675, 1989.

[22] NADERI, B., T. POLZEHL, A. BEYER, T. PILZ, and S. MÖLLER: *Crowdee: mobile crowdsourcing micro-task platform for celebrating the diversity of languages*. In *Interspeech*. 2014. URL `https://api.semanticscholar.org/CorpusID:22508487`.

[23] NADERI, B.: *Motivation of workers on microtask crowdsourcing platforms*. Springer, 2018.

[24] LOMMEL, A., M. POPOVIC, and A. BURCHARDT: *Assessing inter-annotator agreement for translation error annotation*. In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*. 2014.

[25] IBM CORP.: *IBM SPSS Statistics for Macintosh. Version 28.0*. 2021.

[26] WÄLTERMANN, M., A. RAAKE, and S. MÖLLER: *Quality dimensions of narrowband and wideband speech transmission*. *Acta Acustica united with Acustica*, 96(6), pp. 1090–1103, 2010.

[27] GRUBBS, F. E.: *Sample criteria for testing outlying observations*. *Annals of Mathematical Statistics*, 21(1), pp. 27–58, 1950. doi:10.1214/aoms/1177729885.

[28] JARQUE, C. M. and A. K. BERA: *Efficient tests for normality, heteroscedasticity, and serial independence of regression residuals*. *Econometrica*, 48(3), pp. 467–478, 1980. doi:10.2307/1911969.

[29] LEVENE, H.: *Robust tests for equality of variances*. In I. OLKIN, H. HOTELLING ET AL. (eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pp. 278–292. Stanford University Press, 1960.

[30] FISHER, R. A.: *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.

[31] WELCH, B. L.: *The generalization of "student's" problem when several different population variances are involved*. *Biometrika*, 34(1-2), pp. 28–35, 1947. doi:10.1093/biomet/34.1-2.28.

Appendix

	German original	English translation
Group 1: final list of adjective pairs that are loading on the underlying factors, with the pairs used for the quality dimension quantification highlighted in boldface	direkt – umständlich eindeutig – mehrdeutig **einfach – kompliziert** **grammatisch – ungrammatisch** klar – wirr **präzise – ungenau** **übersichtlich – verwirrend** vollständig – lückenhaft	direct – ponderous unambiguous – ambiguous **simple – complicated** **grammatical – ungrammatical** clear – chaotic **precise – vague** **neat – confusing** complete – incomplete
Group 2: list of adjective pairs that were removed during the factor analysis for the sake of interpretability	flüssig – holprig formell – informell geordnet – durcheinander geschrieben – gesprochen höflich – unhöflich kongruent – inkongruent konsistent – inkonsistent logisch – unlogisch menschlich – technisch muttersprachlich – fremdprachlich persönlich – unpersönlich professionell – laienhaft	fluent – non-fluent formal – informal orderly – messy written – spoken polite – impolite congruent – incongruent consistent – inconsistent logical – illogical human – technical native – foreign-language personal – impersonal professional – unprofessional
Group 3: list of adjective pairs that were removed after the preliminary study	aktiv – passiv angemessen – unangemessen angenehm – unangenehm bedeutungsvoll – bedeutungslos bekannt – unbekannt förmlich – lässig gebildet – ungebildet gut – schlecht hochwertig – minderwertig informativ – nichtssagend kreativ – simpel lustig – ernst optimal – suboptimal praktisch – unpraktisch stilvoll – stillos vertraut – fremd vorhersehbar – unberechenbar warm – kalt weich – hart zweckorientiert – zweckfrei	active – passive appropriate – inappropriate pleasant – unpleasant meaningful – meaningless known – unknown formal – casual educated – uneducated good - bad valuable – poor informative – bland creative – simple funny – serious optimal – suboptimal practical – impractical classy – unclassy familiar – foreign predictable – unpredictable warm – cold soft – hard purposeful – purposeless

Table 4 – Complete list of polar adjective pairs used in the experiments for the text type MT in the German original and translated into English for better understanding.

	German original	English translation
Group 1: final list of adjective pairs that are loading on the underlying factors, with the pairs used for the quality dimension quantification highlighted in boldface	**eindeutig – mehrdeutig**	**unambiguous – ambiguous**
	einfach – kompliziert	**simple – complicated**
	logisch – unlogisch	**logical – illogical**
	präzise – ungenau	precise – vague
	simpel – komplex	straightforward – complex
	vollständig – lückenhaft	complete – incomplete
	vorhersehbar – unberechenbar	**predictable – unpredictable**
	zusammenhängend – unzusammenhängend	coherent – incoherent
Group 2: list of adjective pairs that were removed during the factor analysis for the sake of interpretability	anspruchsvoll – anspruchslos	sophisticated – unsophisticated
	ausführlich – knapp	elaborate – terse
	direkt – umständlich	direct – ponderous
	fehlerfrei – fehlerhaft	error-free – faulty
	fließend – holprig	fluent – non-fluent
	geordnet – ungeordnet	ordered – disordered
	grammatisch – ungrammatisch	grammatical – ungrammatical
	klar – wirr	clear – chaotic
	sinnvoll – sinnlos	meaningful – meaningless
	übersichtlich – verwirrend	neat – confusing
	verständlich – unverständlich	comprehensible – incomprehensible
	widerspruchsfrei – widersprüchlich	consistent – contradictory
	wortreich – wortarm	verbose – concise
Group 3: list of adjective pairs that were removed after the preliminary study	konsistent – inkonsistent	consistent – inconsistent
	tiefgehend – oberflächlich	thorough – cursory
	informativ – uninformativ	informative – uninformative
	nicht wiederholend – wiederholend	non-repetitive – repetitive
	geeignet – unpassend	suitable – unsuitable
	kongruent – inkongruent	congruent – incongruent
	pragmatisch – unpragmatisch	pragmatic – impragmatic
	eigenständig – uneigenständig	independent – dependent
	konkret – abstrakt	concrete – abstract
	kompatibel – inkompatibel	compatible – incompatible
	normal – skurril	normal – eccentric
	beständig – wechselhaft	stable – volatile
	prägnant – ungenau	concise – imprecise
	sinnig – unsinnig	meaningful – nonsensical
	kohärent – inkohärent	coherent – incoherent
	bekannt – fremd	familiar – unfamiliar
	einheitlich – uneinheitlich	uniform – inconsistent

Table 5 – Complete list of polar adjective pairs used in the experiments for the text type ATS in the German original and translated into English for better understanding.