Modeling Quality of Experience in German Automatic Text Summarization and Machine Translation

Dinh Nam Pham[1], Vivien Macketanz[1], Shushen Manakhimova[1], Sebastian Möller[1,2]
[1] Speech and Language Technology Lab, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany
[2]Quality and Usability Lab, Technische Universität Berlin, Germany
Correspondence: shushen.manakhimova@dfki.de

Abstract

We present a model for predicting the Quality of Experience (QoE) of German machine-generated text from Automatic Text Summarization (ATS) and Machine Translation (MT). Based on previously established quality dimensions, we fine-tuned BERT for ATS and ELECTRA for MT, which performed best per task. Adding linguistic features further improved accuracy. For ATS, BERT excelled as a multi-target regressor; for MT, separate ELECTRA models performed best. Our results show that combining linguistic features with language models enables robust QoE prediction.

1 Introduction

The Quality of Experience (QoE) framework is well established in domains like video streaming, gaming, or cloud computing, but it remains under-explored for evaluating machine-generated text. As tools for automatically generated text become increasingly widespread, understanding how users perceive output quality gains importance. QoE can be described as the user's subjective degree of delight or annoyance when consuming a service, determined by how well the delivered content meets their expectations for utility or enjoyment in a given context (Le Callet et al., 2012). Modeling QoE for machine-generated text lets developers optimize and monitor the perceived quality of translation and summarization systems without repeatedly relying on expensive human tests.

This study adapts the QoE framework to assess perceived quality in German machine-generated text for two key tasks: Machine Translation (MT) and Automatic Text Summarization (ATS). Building on our previous work that identified relevant perceptual dimensions and produced high-quality human ratings (Manakhimova et al., 2025), we train machine learning models to predict these ratings. Since human evaluations are reliable but costly and hard to scale, we propose an automated prediction approach that combines neural language models with linguistic features.

We fine-tune pre-trained German language models (BERT and ELECTRA) for regression and enhance their [CLS] token embeddings with a 17-dimensional vector of linguistic features capturing readability, lexical diversity, and structural aspects. The [CLS] token is a special classification token automatically prepended to each input sequence. After transformer encoding, its corresponding final hidden state serves as a fixed-size summary representation of the entire sequence. This representation, concatenated with the linguistic feature vector, forms the basis for our regression predictions.

Our main contributions are threefold. First, we formalize QoE prediction for machine-generated text as a multivariate regression problem. Second, we compare multiple multi-output regression strategies (including single-target, multi-target, and multi-task learning) to identify the most effective approach for each task. Third, we demonstrate that integrating linguistic features significantly enhances performance, even in low-resource settings.

Our results offer insights into modeling user-perceived quality in natural language generation, highlighting the benefits of combining neural representations with linguistically motivated features for scalable QoE prediction.

2 Related Work

2.1 Automatic Quality Assessment

Traditional text quality assessment typically measures adherence to formal standards, whereas Quality of Experience (QoE) emphasizes user perception, as established in multimedia research. Subjective evaluation methods like Semantic Differential (SD) scaling (Osgood, 1957) commonly used to capture these perceptions. SD scaling presents respondents with pairs of opposing adjectives and

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 169–175, 2025

asks them to rate a concept or stimulus along a Likert-like scale anchored by these terms.

Readability and text complexity have been researched extensively. Some studies have leveraged language models alone for these tasks, such as (Martinc et al., 2021), while others have focused on linguistic feature-based approaches to readability and complexity across languages, e.g., (Santucci et al., 2020; Štajner and Hulpus, 2020; Seiffe et al., 2022). These studies demonstrate that linguistic features remain highly relevant for evaluating textual properties, even in the era of neural language models.

A study particularly relevant to our approach is Anschütz and Groh (2022), which also integrates a language model with linguistic features for complexity prediction. Their method aligns with our goal of combining automatically extractable linguistic properties with deep learning models to enhance prediction performance.

For German, Naderi et al. (2019) used subjective judgments to train a readability model, underscoring the value of perception-based assessment. More recently, studies explore large language models (LLMs) for reference-free quality evaluation (Chen et al., 2023; Huang et al., 2023) and text simplification quality estimation (Kriz et al., 2020).

Building on these efforts, our work integrates perceptual evaluation with linguistic and neural predictors. Unlike prior research focused on single dimensions (e.g., readability), we model multiple quality dimensions per task, providing a more comprehensive view of user-perceived quality in machine-generated text.

2.2 Linguistic Text Features

Our approach incorporates a targeted set of linguistic features that influence readability, complexity, and overall textual perception. Classic indicators include average sentence length (Pitler and Nenkova, 2008) and word length (McNamara et al., 2014), with longer sentences and words generally signaling increased difficulty. We use established readability metrics: Flesch Reading Ease (Flesch, 1948), Wiener Sachtextformel (Bamberger and Rabin, 1984) (German-specific), Coleman-Liau index (Coleman and Liau, 1975), Gunning Fog (Gunning, 1952), and SMOG (Mc Laughlin, 1969). These assess text difficulty via sentence and word characteristics. To measure lexical variation, we include Type-Token Ratio (TTR) (Richards, 1987) and lexical density (McNamara et al., 2014), defined as the ratio of content words (nouns, verbs, adjectives, adverbs) to the total number of words. Higher values suggest richer, denser language impacting perceived informativeness and complexity.

Together, these features provide insights into multiple dimensions of text quality and complement neural methods in modeling user-perceived QoE.

3 Methodology

3.1 Data

Our experiments build on our prior work (Macketanz et al., 2022; Manakhimova et al., 2025) that examined human perception of machine-generated German text quality using crowdsourced SD surveys to identify subjective quality dimensions in ATS and MT.[1]

Automatic Text Summarization (ATS). The ATS data is based on the GeWiki corpus, which contains preprocessed German Wikipedia articles across domains such as people, science, and politics (Frefel, 2020). Summaries were obtained both from the SwissText & KONVENS 2020 shared task and generated internally using a range of extractive and abstractive methods, such as Lead-3 (Dohare and Karnick, 2017), TextRank (Mihalcea and Tarau, 2004), Pointer-Generator (See et al., 2017), Transformer (Vaswani et al., 2017), Convolutional Self-Attention Transformer (Li et al., 2019), and BERT-Transformer (Devlin et al., 2019). The resulting summaries typically average 37 words (about 2–3 sentences). To ensure a diverse quality range, all summaries underwent an error-type annotation to include various error types and error severities. We further made sure to include different summary lengths.

Machine Translation (MT). For the MT corpus, we selected English–German translations from top-, middle-, and bottom-ranked systems in the WMT19 News Translation task (Barrault et al., 2019), aiming to create a dataset that spans a range of translation quality.[2] Analogously to the ATS corpus, we then performed an error-type annotation to include data points with various translation error types and severities.

[1] Note that this paper highlights only key aspects of our foundational research regarding the dataset creation and quality dimensions identification. For a detailed description of dimension identification and human ratings, see the cited work.

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

3.2 Human Ratings and Quality Dimensions

For each text type, we created tailored adjective pairs (e.g., *grammatical–ungrammatical*) and had German native speakers rate texts on 7-point Likert scales (0–6). The survey ran on the Crowdee platform[3]. Raters were native German speakers from the DACH region (Germany, Austria, Switzerland) and were compensated at a rate of 2 Euro per 10-minute task, in line with minimum wage requirements. Annotators were provided with the texts (each text type separately) and instructed to judge each one using the adjective pair scales, focusing on the language quality while disregarding the content of the sentences to the best of their ability. To ensure data quality, we applied pre-task qualifications and filtered results using the Inconsistency Score (Naderi, 2018) among other criteria. The cleaned datasets comprise 91 ATS and 106 MT texts[4], each rated by 10–20 annotators, prioritizing rating reliability over quantity.

We performed an exploratory factor analysis (EFA) of human ratings, which revealed four latent quality dimensions per task, and validated the structure in a follow-up study using four adjective pairs per text type; the reduced instruments reproduced the original factors with strong correlations to the full scales. For **ATS**, the dominant factor is *linguistic logic* (≈ coherence/consistency), capturing internal cohesion and semantic plausibility; the remaining factors are *complexity* (syntactic/lexical richness and associated reading effort), *clarity* (readability and focus), and *predictability* (discourse flow and logical progression). For **MT**, the factors are *precision* (≈ adequacy/completeness), *complexity* (as above), *grammaticality* (≈ fluency/surface well-formedness, including spelling and punctuation), and *transparency* (≈ coherence/naturalness). We retain the inductive labels but provide these parenthetical alignments to established constructs to facilitate interpretation and comparison with prior work. The resulting human-derived dimensions serve as ground-truth targets for our predictive modeling.

For the model training, we use the mean rating per dimension per text. Each dataset includes the text plus four columns for the dimension ratings (see Tables 1 and 2). Ratings range from 0 to 6 (0 being the lowest perceived quality, and 6 the highest), framing this as a multi-output regression task with four targets. Each test item has a corresponding Mean Opinion Score (MOS) per dimension, calculated as the arithmetic mean of individual human ratings. MOS is a standard practice in QoE research (ITU-T, 2016).

Table 1: Excerpt from the ATS dataset illustrating its structure, with Mean Opinion Scores per test item and quality dimension. Text samples were shortened.

Text Sample	Logic	Compl-exity	Cla-rity	Predict-ability
"35,2 Vol. ist ein deutscher Kriminalfilm aus dem Jahr 2012 [...]."	2.2692	4.1923	3.0962	2.4615
"Bansin ist ein Ortsteil der Gemeinde Heringsdorf im Landkreis [...]."	4.8125	4.0000	5.1875	3.3750
"Veikko Lahti war ein finnischer Ringer. Er war Olympiasieger [...]"	2.0909	1.8182	3.0909	2.6364

Table 2: Excerpt from the MT dataset illustrating its structure, with Mean Opinion Scores per test item and quality dimension. Text samples were shortened.

Text Sample	Preci-sion	Compl-exity	Trans-parency	Gramm-aticality
"Ihre Zurückhaltung ist alles auf das Timing zurückzuführen."	3.0345	1.8966	3.0862	3.6207
"Seine Nachrichten führten jedoch nur zu einem grauen [...]."	2.6842	2.1579	2.7368	1.7368
"Sicher ist, dass der Konflikt der Mitte des 17. Jahrhunderts [...]."	3.4118	4.1765	3.8235	2.8235

3.3 Model Selection

Given the limited size of our datasets and the resource-intensive demands of further data collection, we leveraged pre-trained German language models for training and evaluation to capitalize on their existing knowledge.

We first evaluated five pre-trained models to select the best performer for further experiments: the BERT variants "bert-base-german-uncased" (Bayerische Staatsbibliothek, 2025b), "bert-base-german-cased" (Bayerische Staatsbibliothek, 2025a), "gbert-base," "gbert-large," and the ELECTRA model "gelectra-large" (Chan et al., 2020), all sourced from Huggingface[5]. For each, we replaced the output layer with a linear layer of size 4 to predict all four quality dimensions simultaneously in a multi-target regression framework.

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

[4] You can find the ATS and MT datasets without the ratings here: https://github.com/DFKI-NLP/TextQ

[5] https://huggingface.co/

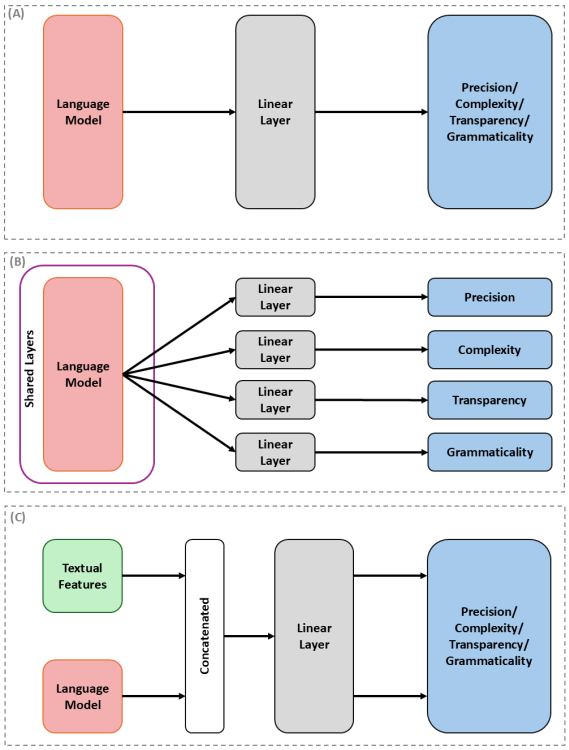

Figure 1: Model architectures with the MT labels for (A) fine-tuned language model, (B) multi-task learning model and (C) textual feature integration.

3.4 Multi-Label Regression Approach

Continuing with the best-performing pre-trained model for each dataset, we explored different approaches for the four-label regression task. The baseline models (from Section 3.3) output all four labels simultaneously in a multi-target regression setup. In addition, we trained separate single-target models (one per quality dimension) and evaluated their performance individually.

Given evidence that multi-task learning (MTL) can outperform single-task models in similar multi-target regression settings (Mohtaj et al., 2023), we also implemented an MTL model for each dataset. Our MTL architecture shares a single pre-trained backbone (BERT or ELECTRA, as chosen in Section 3.3) across all tasks but uses separate task-specific linear regression heads. Each dimension is predicted by an independent 1D regression head that processes the pooled [CLS] token embedding from the shared backbone. The overall loss is the sum of all task losses with equal weighting, treating all dimensions as equally important.

3.5 Linguistic Feature Integration

Lastly, we evaluated the benefit of combining statistical and linguistic text features with language model embeddings, leveraging both transformer architectures and features in a hybrid approach. We extracted 17 features, grouped into four main categories. The first category, structural metrics, includes average sentence length, the percentage of words with six or more letters, and the average number of characters per word. The second category covers syllabic properties, such as average syllables per word, the percentage of monosyllabic words, and the percentage of words with three or more syllables. Third, we incorporated readability formulas, including the Flesch Reading Ease, all four variants of the Wiener Sachtextformel, as well as the SMOG, Coleman-Liau, and Gunning Fog indices. Finally, the fourth category captures lexical diversity through measures like TTR, lexical density, and the number of unique tokens.

These features were min-max normalized to [0,1] per dataset split, with scaling parameters applied consistently to validation and test sets to avoid data leakage (see Section 3.6). The normalized features were concatenated with the [CLS] embedding from the selected model and fed into the regression head according to the multi-target regression method chosen. Figure 1 illustrates the model architectures for MT.

3.6 Experimental Setup

We used consistent hyperparameters and data splits across all experiments for fair comparison: learning rate 2×10^{-5}, MSE loss, 30 epochs, batch size 16, and AdamW optimizer with weight decay 0.01.

Due to the small dataset size, we employed a 7-fold cross-validation. Each dataset was randomly shuffled and split into seven equal folds $F = f_1, ..., f_7$. For each round, fold f_i served as test set, fold f_{i-1} (or f_7 if $i = 1$) as validation, and the remaining five folds for training. Thus, each fold was used once as validation and once as test. During training, we monitored validation RMSE per epoch and saved the best checkpoint for final evaluation on the test fold. Reported metrics are the average over the seven test folds.

4 Results and Discussion

We primarily use RMSE to evaluate model performance, as it is a standard regression metric that penalizes outliers quadratically and measures error on the original rating scale.

Table 3: The performance of the fine-tuned language models on ATS.

Model	MAE	RMSE	R^2
bert-base-german-uncased	0.7965	0.9670	0.1026
bert-base-german-cased	0.7696	0.9390	0.1510
gbert-base	0.8448	1.0248	-0.0076
gbert-large	**0.6958**	**0.8588**	**0.2939**
gelectra-large	0.7780	0.9356	0.1683

Table 4: The performance of the fine-tuned language models on MT.

Model	MAE	RMSE	R^2
bert-base-german-uncased	0.9911	1.1877	0.1085
bert-base-german-cased	1.0066	1.2295	0.0476
gbert-base	0.9966	1.2292	0.0613
gbert-large	0.8961	1.0669	0.2719
gelectra-large	**0.8500**	**1.0133**	**0.3727**

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (1)$$

where y_i is the true value, $\hat{y}_i$ is the predicted value, and n is the number of samples. Additionally, we report mean absolute error (MAE) and the coefficient of determination (R^2) as supplementary metrics. For each model, we average RMSE, MAE, and R^2 scores across all four quality dimensions.

The fine-tuning results of pre-trained language models on our datasets are shown in Tables 3 and 4. For ATS, gbert-large performed best, while gelectra-large was top for MT.

MAE values are lower than RMSE, as expected (Willmott and Matsuura, 2005). The ratio between RMSE and MAE can provide additional insight into the shape of the error distribution (Karunasingha, 2022). Specifically, a lower ratio indicates a platykurtic distribution, which has lighter tails and fewer extreme errors than a normal distribution, while a higher ratio would point to a leptokurtic distribution with heavier tails and more outliers. The RMSE/MAE ratios of our best models (ATS: 1.23, MT: 1.19) are less than 1.25, which is shown to be consistent with a platykurtic error distribution (Karunasingha, 2022). In such cases, RMSE was demonstrated to be a more reliable performance measure than MAE (Karunasingha, 2022), supporting RMSE as a suitable metric for our experiments. Moreover, according to (Inagata et al., 2025), ratios between approx. 1.14 and 1.25 indicate good forecasting performance. Our models' ratios fall within this range, confirming their robustness. Based on these findings, we selected gbert-large for ATS and gelectra-large for MT for further experiments.

Table 5: The performance of the multi-label regression approaches and textual feature integration for ATS.

Model	MAE	RMSE	R^2
Multi-target regression	**0.6958**	**0.8588**	**0.2939**
Single-target regression	0.7408	0.8990	0.1895
Multi-task learning	0.7525	0.9092	0.2001
Multi-target regression + textual features	**0.6764**	**0.8308**	**0.3225**

Table 6: The performance of the multi-label regression approaches and textual feature integration for MT.

Model	MAE	RMSE	R^2
Multi-target regression	0.8500	1.0133	0.3727
Single-target regression	**0.8207**	**0.9785**	**0.3867**
Multi-task learning	0.8661	1.0208	0.2898
Single-target regression + textual features	**0.7885**	**0.9259**	**0.4510**

The results of the multi-label regression approaches and integration of textual features are shown in Tables 5 and 6. For both text types, the multi-task model performed worst, though with only slightly higher errors. This may be due to limited training data restricting model complexity, which is currently limited to a linear layer over language model embeddings. Larger datasets could enable more complex architectures (e.g., LSTM/GRU) and better leverage MTL. The multi-target regression model achieved the lowest RMSE for ATS, while single-target regression was best for MT. Adding textual features to these models significantly improved performance across all metrics.

These findings highlight the advantage of combining neural embeddings with linguistic features. While BERT and ELECTRA embeddings capture semantic information, textual statistics add complementary insights not fully captured by the language models. Our hybrid models achieved RMSEs of 0.8308 (ATS) and 0.9259 (MT), with MAEs of 0.6764 and 0.7885, corresponding to average deviations of 11.27% and 13.14% on the 0–6 scale. This confirms the effectiveness of our approach in predicting human-annotated quality dimensions and validates modeling QoE through these identified dimensions.

5 Conclusion and Future Work

We present an approach to modeling QoE for machine-generated German text, focusing on ATS and MT. By fine-tuning pre-trained language models and integrating linguistic features, we significantly improved the prediction of perceptual quality dimensions for both text types. Our experiments showed gbert-large performed best for ATS, and gelectra-large for MT. Conducted in a low-resource setting, these results deepen understanding of QoE in NLP and highlight the value of combining neural and linguistic features for quality estimation.

To advance this work, we plan to expand our datasets with LLM-generated items rated via crowdsourcing. As initial variance analyses show no significant differences, we will integrate these with the original data. We are also developing a set of ~100 automatically extractable linguistic features to enhance model inputs and to train linear regression models for feature selection to identify the most predictive features per dimension, improving interpretability and robustness.

Limitations

We acknowledge several limitations of our study. First, the dataset of human ratings is relatively small. However, to ensure rating reliability and mitigate common issues in crowdsourced data, we prioritized collecting more ratings per item over a larger number of texts. We are currently addressing the dataset size by augmenting it with LLM-generated items, which will also help align the data with recent advances in text generation.

Second, our research is limited to the German language, so generalizability to other languages remains uncertain. Nonetheless, we help fill a gap, as most prior work in this area has focused on English.

Finally, our models have only been evaluated on MT and ATS outputs. It remains an open question how well they generalize to other types of machine-generated text; a direction we plan to explore in future studies.

Acknowledgments

We would like to thank our colleagues Aleksandra Gabryszak and Polina Danilovskaia for their valuable assistance with dataset quality control and training an early model.

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

Miriam Anschütz and Georg Groh. 2022. TUM social computing at GermEval 2022: Towards the significance of text statistics and neural embeddings in text complexity prediction. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 21–26, Potsdam, Germany. Association for Computational Linguistics.

Richard Bamberger and Annette T Rabin. 1984. New approaches to readability: Austrian research. *The Reading Teacher*, 37(6):512–519.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. 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)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Bayerische Staatsbibliothek. 2025a. bert-base-german-cased (revision 43cce13).

Bayerische Staatsbibliothek. 2025b. bert-base-german-uncased (revision b705f0e).

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of Large Language Models for reference-free text quality evaluation: An empirical study. *arXiv preprint arXiv:2304.00723*.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Shibhansh Dohare and Harish Karnick. 2017. Text summarization using abstract meaning representation.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Dominik Frefel. 2020. Summarization corpora of Wikipedia articles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6651–6655, Marseille, France. European Language Resources Association.

Robert Gunning. 1952. The technique of clear writing. *(No Title)*.

Hui Huang, Shuangzhi Wu, Xinnian Liang, Bing Wang, Yanrui Shi, Peihao Wu, Muyun Yang, and Tiejun Zhao. 2023. Towards making the most of LLM for translation quality estimation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 375–386. Springer.

Tomoya Inagata, Yuji Mizuno, Keita Matsunaga, Fujio Kurokawa, Masaharu Tanaka, and Nobumasa Matsui. 2025. Parameter evaluation method for power demand forecasting methodology of a clinic. *IEEE Transactions on Industry Applications*, 61:930–939.

ITU-T. 2016. Mean opinion score (MOS) terminology. Recommendation ITU-T P.800.1 (07/2016).

Dulakshi Santhusitha Kumari Karunasingha. 2022. Root Mean Square Error or Mean Absolute Error? Use their ratio as well. *Information Sciences*, 585:609–629.

Reno Kriz, Marianna Apidianaki, and Chris Callison-Burch. 2020. Simple-QE: Better automatic quality estimation for text simplification. *arXiv preprint arXiv:2012.12382*.

Patrick Le Callet, Sebastian Möller, Andrew Perkis, and 1 others. 2012. Qualinet white paper on definitions of quality of experience. *European network on quality of experience in multimedia systems and services (COST Action IC 1003)*, 3(2012).

Ke Li, Yuntian Deng, and Yoon Kim. 2019. 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)*, pages 404–413.

Vivien Macketanz, Babak Naderi, Steven Schmidt, and Sebastian Möller. 2022. Perceptual quality dimensions of machine-generated text with a focus on machine translation. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 24–31.

Shushen Manakhimova, Vivien Macketanz, and Sebastian Möller. 2025. Quality of experience of german machine translation and automatic text summarization. Halle/Saale. In press.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411.

Salar Mohtaj, Vera Schmitt, Razieh Khamsehashari, and Sebastian Möller. 2023. Multi-task learning for German text readability assessment. In *CLiC-it*.

Babak Naderi. 2018. *Motivation of Workers on Microtask Crowdsourcing Platforms*, 1st edition. Springer Publishing Company, Incorporated.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective assessment of text complexity: A dataset for German language. *Preprint*, arXiv:1904.07733.

Charles Egerton Osgood. 1957. *The Measurement of Meaning*. University of Illinois Press, Urbana,.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195.

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

Valentino Santucci, Filippo Santarelli, Luciana Forti, and Stefania Spina. 2020. Automatic classification of text complexity. *Applied Sciences*, 10(20):7285.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Laura Seiffe, Fares Kallel, Sebastian Möller, Babak Naderi, and Roland Roller. 2022. Subjective text complexity assessment for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 707–714.

Sanja Štajner and Ioana Hulpus. 2020. *When shallow is good enough: Automatic assessment of conceptual text complexity using shallow semantic features*. European Language Resources Association, ELRA-ELDA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Cort J. Willmott and Kenji Matsuura. 2005. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in assessing average model performance. *Climate Research*, 30:79–82.