Perceptual Quality Dimensions of Machine-Generated Text with a Focus on Machine Translation

Vivien Macketanz
German Research Center for AI
vivien.macketanz@dfki.de

Babak Naderi
Quality and Usability Lab, TU Berlin
babak.naderi@tu-berlin.de

Steven Schmidt
Quality and Usability Lab, TU Berlin
steven.schmidt@tu-berlin.de

Sebastian Möller
Quality and Usability Lab, TU Berlin
sebastian.moeller@tu-berlin.de

Abstract

The quality of machine-generated text is a complex construct consisting of various aspects and dimensions. We present a study that aims to uncover relevant perceptual quality dimensions for one type of machine-generated text, that is, Machine Translation. We conducted a crowdsourcing survey in the style of a Semantic Differential to collect attribute ratings for German MT outputs. An Exploratory Factor Analysis revealed the underlying perceptual dimensions. As a result, we extracted four factors that operate as relevant dimensions for the Quality of Experience of MT outputs: precision, complexity, grammaticality, and transparency.

1 Introduction

In recent years, automatically generated text has increasingly gained importance, e.g., chatbots, automatic summarizations, or machine translations. Although the quality of such texts has greatly improved over time, it has not yet reached human parity (Toral et al., 2018). Therefore, the quality of machine-generated text is of ongoing interest to the research community and is further important for gaining acceptance in different applications.

The Quality of Experience (QoE) is defined as "the degree of delight or annoyance of the user of an application or service" (Le Callet et al., 2012). This means that the QoE is a subjective perception that needs to be quantified in empirical studies (Möller and Raake, 2014). While there are standardized methods for auditory and visual media, such as ITU P.800, P.910, or BT.500, the QoE of text has been mostly disregarded until now.

The perceptual quality of machine-generated text is a highly complex construct. Many aspects and dimensions play a crucial role; hence, it is the object of investigation of various research areas. We suggest that a multi-dimensional prediction model covering a wide variety of aspects is the best approach to assess the quality of machine-generated text. To the best of our knowledge, no such model exists. Therefore, we are developing a prediction model for the quality of German machine-generated text, specifically, Machine Translation (MT). We aim to create our model based on a combination of linguistic data and automatically extractable factors that can predict the QoE of MT outputs. Our first milestone is identifying relevant perceptual quality dimensions, the foundation of our model. We achieved this milestone by conducting a crowdsourcing study in the style of a Semantic Differential and subsequently extracting the quality dimensions through an Exploratory Factor Analysis.

2 Related Work

This section provides an overview of the existing metrics for capturing the performance or quality of MT systems. The first category of metrics is automatic methods, which have the advantage of being fast, low-cost, and reproducible. The most commonly used metrics are BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), COMET (Rei et al., 2020), and PRISM (Thompson and Post, 2020). Metrics like TER (Snover et al., 2009) measure the translation edit rate, and quality estimation methods (Blatz et al., 2004; Specia et al., 2009) can predict the quality without access to the reference translation(s). However, one shared shortcoming of all these automatic metrics is that, as opposed to our approach, they are not based on relevant quality dimensions and thus lack diagnostic power.

The second category of metrics is subjective methods for directly measuring quality that are more costly yet more reliable. There are large-scale human rankings that are often conducted in international conferences in order to compare the performance and/or quality of several MT systems (Callison-Burch et al., 2007; Bojar et al., 2015). The Multidimensional Quality Metrics (MQM) is a framework for the manual assessment of translation

Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval), pages 24 - 31
May 27, 2022 ©2022 Association for Computational Linguistics

quality (Lommel et al., 2014b). Additionally, test suites have recently regained more importance. A test suite is a challenge set created to systematically analyze the behavior of MT systems in different aspects, e.g., (Guillou and Hardmeier, 2016), (Isabelle et al., 2017), or (Burchardt et al., 2017).

While the mentioned techniques focus on capturing the performance or quality of MT systems, they cannot sufficiently capture the QoE by users of MT output as QoE is the only technique that is not measured by pre-defined criteria. Instead, QoE is based on identifying relevant criteria (i.e., quality dimensions) in a real-world scenario.

3 Experimental Setup

We conducted a study to identify relevant dimensions for the quality of machine-generated text, specifically German MT outputs. We did so by utilizing a crowdsourcing survey in which participants had to rate MT outputs. Our corpus contained English to German translations from the submissions to the News translation task of the Fourth Conference on Machine Translation (WMT19)[1]. We chose this data for our corpus as we needed test sentences from several MT systems with varying translation quality. Furthermore, the data is freely available for research purposes[2]. We extracted a set of translations from six submitted systems that appeared at the top, the middle, and the bottom of the ranking of WMT19 systems (Barrault et al., 2019), resulting in a corpus of 11,922 sentences. A linguistic expert created a sub-corpus for the survey, dedicating around 15 hours to carefully extract translations varying in length, quality, and error types. The sub-corpus consists of 45 sentences.[3]

The survey was conducted as a Semantic Differential (SD) (Osgood et al., 1957). An SD is a rating scale that measures a person's attitude towards an entity, here: our test sentences. The participants were asked to rate their perception of the test items on a scale between two polar adjectives, e.g., "grammatical – ungrammatical". All adjective pairs used in the study can be found in Table 2 in the Appendix. The adjective pairs were carefully selected by a linguist who is experienced in MT evaluation and thereafter discussed with another linguist to cover all potentially relevant aspects for the perceptual quality of the test sentences.

We would like to emphasize that while we are using MT as an example text type, the focus of our study lies on the quality of *machine-generated text*. Therefore, we solely work with the MT outputs and do not take the source sentences and concomitant quality aspects into account (as opposed to approaches that focus on the quality of MT).

3.1 Antonym pair identification study

We first ran a small-scale preliminary study with 14 participants to confirm our antonym pairs. The participants were colleagues and mostly linguistic experts. Our test set comprised 15 sentences from the sub-corpus. The first part of the study consisted of the SD; the participants were instructed to rate the quality of each sentence based on 38 adjective pairs serving as endpoints of a 7-point Likert scale ranging from -3 to +3. As we are solely focusing on the intrinsic quality, they were instructed to rate only the quality of the language but not of the translation itself. The adjective pairs were hand-selected by a linguistic expert, experienced in the evaluation of MT, to cover as many aspects of machine-translated text as possible. In the second part, the participants had to rate each adjective pair on its suitability to evaluate language on a 5-point scale. In addition, they were also encouraged to provide feedback regarding the suitability and to suggest other potential adjective pairs. Based on the rating of the adjective pairs, we removed all adjective pairs with a mean value of less than 3.2 and a standard deviation of more than 1.2. As a result, we reduced the number of adjective pairs to 20.

3.2 Crowdsourcing study

The main study was conducted as a crowdsourcing survey with Crowdee[4]. 141 crowdworkers participated in the study. The survey followed the IRB guidelines of our institution, and participants were paid according to the minimum wage law. Crowdworkers stayed anonymous, no personal information was collected in the survey[5]. The study was accessible to native speakers only as a good knowledge of German was required. As we wanted the participants to evaluate the language itself (and not the content of the test sentences), they were instructed to base their ratings exclusively on the

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

[2] cf. Licensing of Data https://www.statmt.org/wmt19/translation-task.html

[3] https://github.com/DFKI-NLP/TextQ

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

[5] Crowdee's privacy Statement can be found here: https://www.crowdee.com/privacy-statement

language of the sentences and ignore the meaning of the sentences as best as they could. They were only informed that the sentences might contain errors, but not that the sentences were outputs of English to German MT. The full instructions can be found in Table 3 in the Appendix.

The adjective pairs were randomized per participant, and so was the order of the polarity. All 45 sentences from the sub-corpus were used. While this is a comparably small number of test items, we argue that we can still draw significant conclusions as the items were hand-picked by an expert to cover as many different linguistic aspects as possible. Based on the feedback we received from the preliminary study, we decided to present only three test sentences to each participant, as the rating is very time-consuming. Each sentence had to be rated based on all 20 antonym pairs. Completing the full survey was expected to take around 10 minutes.

Following (Naderi et al., 2015), we incorporated a test condition for the majority of the sentences[6]. The test condition is based on calculating an Inconsistency Score (IS) (Naderi, 2018) on repeated adjective pairs. Altogether, we collected up to 30 ratings of all adjective pairs per sentence. The average working time amounted to 392.1 seconds.

4 Multidimensional Analysis

QoE can be formalized as a multidimensional perceptual space where the defining parameters function as dimensions. It is the aim of the multidimensional analysis to identify those dimensions for the QoE of MT output.

4.1 Data cleansing

While crowdsourcing studies have many benefits, one shortcoming is that there might be crowdworkers who do not work thoroughly, eventually leading to noisy data (Naderi et al., 2015). Thus, we had to cleanse the data to filter out invalid ratings. [7] We did so in three steps: First, we eliminated ratings of participants that completed the survey in 40% or less of the expected 10 minutes. Thus, participants who finished the questionnaire in 240 seconds or less were excluded from the analysis. Second, we excluded all ratings of participants who provided the same value for every adjective pair for every sentence, assuming they were not reading the test material. Lastly, we calculated the IS (Naderi, 2018). While it is known that the degree of variance in human evaluation of translation is high (Lommel et al., 2014a), the IS allows filtering out outliers that show a higher degree of variance than expected under normal conditions. The IS calculation is based on the test conditions of the repeated adjective pairs. For details of the calculation, the interested reader is referred to Naderi (2018).

The data cleansing removed 6,800 ratings, resulting in 14,200 ratings. The average working time after the data cleansing amounted to 473.31 sec.

4.2 Exploratory Factor Analysis

We conducted an Exploratory Factor Analysis (EFA) in SPSS (IBM Corp.). Factor analysis is a technique for identifying common factors (i.e., latent variables) that explain the correlation among a set of observed variables. The extraction method used was Maximum Likelihood; The rotation method was PROMAX with Kaiser Normalization, leading to non-orthogonal dimensions.

It is important to balance the statistical goodness-of-fit and the interpretability of the resulting dimensions (Wältermann et al., 2010). Our data contained several adjective pairs with low communalities and/or cross-loadings differing by less than 0.2. Our interpretation is that these pairs are not specific enough or are related to other, irrelevant aspects. Thus, we removed those attributes for the sake of interpretability. The dimension reduction revealed four factors for eight polar adjective pairs. Pearson's chi-squared test for the goodness of fit was p = 0.36 ($\chi^2 = 2.06$, df = 2). The Kaiser-Meyer-Olkin value was quite high at 0.901, indicating that the data is adequate for a factor analysis.

The distribution of the adjective pairs on the four factors and the explained percentage of variance can be seen in Table 1. Note that the adjectives are translated into English for better understanding. The four adjective pairs *unambiguous – ambiguous* (German: eindeutig – mehrdeutig), *precise – vague* (präzise – ungenau), *complete – incomplete* (vollständig – lückenhaft), and *clear – chaotic* (klar – wirr) are loading on factor 1 (F1). F1 explains 53.2% of the variance. Factor 2 (F2) is loaded by the two adjective pairs *direct – ponderous* (direkt – umständlich) and *simple – complicated* (einfach – kompliziert) and explains an additional 8.4% of the variance. Only one adjective pair is loading on Fac-

[6] 30 of the 45 test sentences were rated with the test condition, as we ran the survey in two batches and included the test condition only in the second batch.

[7] Crowdworkers were paid regardless of their ratings.

	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 the adjective pairs (English translations) on the factors and % of explained variance.

tor 3 (F3): *grammatical – ungrammatical* (grammatisch – ungrammatisch) and another 10.5% of the variance is explained by F3. The fourth factor (F4) is also loaded by one adjective pair only, namely *neat – confusing* (übersichtlich – verwirrend), and it explains an additional 8.0% of the variance.

The adjective pairs loading on F1 are all describing characteristics related to precision; hence, this factor is labeled *precision*. The adjective pairs loading on F2 are related to complexity; thus, F2 is labeled *complexity*. F3 is labeled *grammaticality*, and F4 is labeled *transparency*. The *precision* and *transparency* factors seem to overlap while the remaining factors are more easily separable in their meaning.

4.3 Quality dimensions

Former commonly used quality aspects for MT were *fluency* and *adequacy* (cf., e.g., the MQM metrics mentioned in Section 2). While our study has not tested for extrinsic *adequacy*, as we only presented the MT outputs and not the source sentences, other authors have already stated that *fluency* is not the central problem in MT nowadays (Bentivogli et al., 2016). Neural MT has become more fluent, with MT errors being more subtle and thus harder to spot. Our study confirms this claim as the analysis has brought out four other relevant quality dimensions: *precision*, *complexity*, *grammaticality*, and *transparency*. Interestingly, our 20 antonym pairs did include the adjective pair *fluent – non-fluent*, as we covered a wide variety of translation issues. However, we had to eliminate this pair during the EFA due to discriminant validity issues.

Looking at our four dimensions, the factor *precision* seems to refer to the clarity and completeness of the text. The factor *complexity* presumably refers to the textual complexity, and sentences with a high rating for the adjectives *complicated* and *ponderous* in our study generally tend to be longer. More interesting findings arise when looking further into our data: Sentences with a high rating for the factor *grammaticality* tend to miss words, contain spelling or punctuation errors, or hold mistranslations. Interestingly though, these sentences tend to be shorter rather than longer. Our theory is that the longer and therefore more convoluted a sentence is, the more difficult it is to spot grammar errors, and, consequently, other factors like *complexity* become more relevant. Our last dimension, *transparency*, seems less tangible than the other dimensions. We theorize that it refers to the lucidity of the text. It seems similar to *precision*, and there is indeed a higher correlation (0.748).

As a final remark, we would like to point out that the identification of the dimensions in the multidimensional analysis is strongly dependent on the data (Wältermann et al., 2010), i.e., the choice of test sentences and antonym pairs. While we collected a large number of data points, validating these is the subject of future work. Hence, we cannot guarantee that the identified quality dimensions cover all potential perceptions completely. Furthermore, as the survey was conducted with German native speakers, the majority of the participants can be assumed to be WEIRD participants[8] (Henrich et al., 2010) which leads to a demographic bias. Our findings cannot be assumed to be valid for other languages and/or participant groups.

5 Conclusion and Outlook

We present a study exploring the relevant quality dimensions for MT outputs. We identified antonym pairs of a Semantic Differential in a preliminary study and used these attributes to rate 45 German test sentences. We then carried out an Exploratory Factor Analysis that resulted in the extraction of four relevant quality dimensions: *precision*, *complexity*, *grammaticality*, and *transparency*. According to our study, these are the quality dimensions that are relevant for the QoE, i.e., the subjective perception of a user of a text.

Our ultimate goal is to develop a prediction model to assess the quality of machine-generated text. We focus on two text types: Machine Translation and Automatic Text Summarization (ATS). Our next step is to identify the relevant quality dimensions for ATS. To do so, we are currently

[8] WEIRD stands for **w**estern, **e**ducated, **i**ndustrialized, **r**ich, and **d**emocratic participants

conducting another crowdsourcing study with an adapted set of adjective pairs. The focus on two different types of machine-generated texts allows us to compare the (potential) differences in the perceptive quality dimensions and enables us to draw generalizations for other text types.

Simultaneously, we are working on the quantification of the quality dimensions for MT. As the factor analysis conducted in the study at hand is highly complex, we are developing a simplified survey in which we present only one representative antonym pair per dimension. If the result of the follow-up study verifies our current study, we can assume our dimensions to be accurate.

Further steps will involve correlating automatically extractable text parameters and quality dimensions, and building and testing various prediction models. These efforts should ultimately result in a quality prediction model for MT, ATS, and potentially other types of machine-generated text.

Other potential future work includes analyzing the possible overlap between the four dimensions at hand and other existing quality metrics, e.g., MQM. Furthermore, it would be of interest to expand the analysis to other languages, as it might also counteract the WEIRD bias.

Acknowledgements

The present study was funded by the Deutsche Forschungsgemeinschaft (DFG) through the project “Analyse und automatische Abschätzung der Qualität maschinell generierter Texte”, project number 436813723. We thank all colleagues who participated in our preliminary study.

References

Satanjeev Banerjee 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.

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.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING ’04, page 315–es, USA. Association for Computational Linguistics.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, et al. 2015. Findings of the 2015 Workshop on Statistical Machine Translation.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. A linguistic evaluation of rule-based, phrase-based, and neural MT engines. *The Prague Bulletin of Mathematical Linguistics*, 108(1):159–170.

Chris Callison-Burch, Cameron Shaw Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

IBM Corp. IBM SPSS Statistics for Macintosh. Version 28.0.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A Challenge Set Approach to Evaluating Machine Translation.

Patrick Le Callet, Sebastian Möller, Andrew Perkis, et al. 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).

Arle Lommel, Maja Popovic, and Aljoscha Burchardt. 2014a. Assessing inter-annotator agreement for translation error annotation. In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*.

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

Sebastian Möller and Alexander Raake. 2014. *Quality of experience: advanced concepts, applications and methods*. Springer.

Babak Naderi. 2018. *Motivation of workers on micro-task crowdsourcing platforms*. Springer.

Babak Naderi, Ina Wechsung, and Sebastian Möller. 2015. 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)*, pages 1–2. IEEE.

Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. 1957. *The measurement of meaning*. 47. University of Illinois press.

Kishore Papineni, 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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece. Association for Computational Linguistics.

Lucia Specia, Craig Saunders, Marco Turchi, Zhuoran Wang, and John Shawe-Taylor. 2009. Improving the confidence of machine translation quality estimates.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation.

Marcel Wältermann, Alexander Raake, and Sebastian Möller. 2010. Quality dimensions of narrowband and wideband speech transmission. *Acta Acustica united with Acustica*, 96(6):1090–1103.

A Appendix

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

Table 2: Complete list of polar adjective pairs used in the study in the German original and translated into English for better understanding.

German original
Willkommen zur Umfrage In dieser Umfrage sollst du die Sprache von verschiedenen Sätzen anhand einer Adjektivliste bewerten. Hierzu werden dir insgesamt 3 Sätze auf je 4 Seiten gezeigt. Die Sätze können fehlerhaft sein, müssen aber nicht. Bitte bewerte jeden dieser 3 Sätze in Hinblick auf die verwendete Sprache (inklusive Satzzeichen) mit Hilfe der Adjektivliste. Die Adjektivliste enthält 22 gegesätzliche Adjektivpaare, die an den beiden Enden einer Skala von -3 bis +3 stehen. Bitte schiebe für jedes Adjektivpaar den Slider auf der Skala dorthin, wo der Wert deiner Meinung nach die Sprache des jeweiligen Satzes am besten beschreibt. Versuche, den Inhalt der Sätze nicht in deine Bewertung miteinfließen zu lassen. Alle deine Antworten aus dem folgenden Fragebogen werden anonym behandelt und dienen ausschließlich dem Zweck dieser wissenschaftlichen Arbeit. Achtung: Das Ergebnis dieser Umfrage ist sehr wichtig für uns und andere Wissenschaftler, die in diesem Bereich arbeiten. Wir verfügen über Methoden um die Einheitlichkeit deiner Antworten zu überprüfen. Wir werden diese Methoden nutzen, um die Qualität der abgeschickten Aufgaben zu bewerten. Crowdworker, die qualitativ hochwertige Antworten geben, werden zu weiteren Untersuchungen eingeladen, zu denen sie exklusiven Zugang erhalten. Auf der nächsten Seite wirst du zunächst ein Beispiel sehen, bevor es losgeht.
English translation
Welcome to the survey In this survey, you are supposed to evaluate the language of different sentences with the help of an adjective list. You will be shown 3 sentences altogether, distributed over 4 pages each. Die sentences might, but don't have to, contain errors. Please evaluate each of the 3 sentences with regard to the language used (including punctuation) with the help of the adjective list. The adjective list contains 22 polar adjective pairs which are located on both ends of a scale from -3 to +3. Please move the slider for each adjective pair to the point on the scale where the value describes the language of the respective sentence best in your opinion. Try to not let the content of the sentences influence your evaluation. All your answers in the following survey will be handled anonymously and exclusively serve the aim of this scientific work. Note: The result of this survey is very important for us and other scientists working in this area. We are equipped with methods to check your answers for consistency. We will use these methods to evaluate the quality of the completed task. Crowdworkers that provide high-quality answers will be invited to further surveys to which they will receive exclusive access. On the next page, you will first see an example before the survey starts.

Table 3: Instructions for the crowdsourcing survey in the German original and translated into English for better understanding.