When the Scale is Unclear – Analysis of the Interpretation of Rating Scales in Human Evaluation of Text Simplification

Regina Stodden[1]

[1]*Heinrich Heine University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany*

Abstract

In the evaluation of text simplification, human ratings are of the highest importance as automatic metrics are not yet sufficient. However, so far, no best practices for human evaluation of text simplification exist. Hence, several different rating scales and definitions of evaluation dimensions are used to evaluate text simplification system outputs. Also, the scales lack some analysis regarding their reliability and interpretation. Therefore, in this paper, we analyse the interpretation of the scales of the evaluation dimensions meaning preservation, and simplicity based on simplification pairs with no change. Our analysis shows that annotators differently interpreted the scale of the simplicity dimension: on the one hand, the lowest value was interpreted to describe that the simplified sentence is more complex than the original sentence, and on the other hand, that a simplified sentence is as complex as the original sentence. Overall, the paper emphasises that best practices for human evaluation of text simplification are demanded to reduce misinterpretation of the scales.

Keywords

text simplification, human evaluation, scale interpretation

1. Introduction

Text simplification is the manual or automatic process of generating a simpler version of a complex text or sentence by preserving its meaning. Simplified texts are easier to understand, for example, for non-native speaker or people with lower literacy. Besides simplicity, meaning preservation and grammaticality are important criteria for a good simplification of a text. Thus, these criteria are also used to evaluate automatic text simplification systems [1]. Therefore, the original text and its generated simplified version are aligned to a simplification pair. This pair can be evaluated manually or automatically [1].

So far, manual evaluation of text simplification is the most reliable evaluation method to judge text simplification [1], as for example the existing automatic evaluation metrics either focus only on lexical changes, e.g., SARI [2], or the meaning preservation, e.g., BLEU [3]. Nevertheless, human evaluation also has its weakness because no best practice for text simplification

Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021), co-located with SEPLN 2021. September 21st, 2021 (Online). Saggion, H., Štajner, S. and Ferrés, D. (Eds).

regina.stodden@hhu.de (R. Stodden)
https://user.phil.hhu.de/~stodden (R. Stodden)
0000-0002-7470-0961 (R. Stodden)

evaluation exists. Currently, three dimensions are most often used in research, i.e., meaning preservation, simplicity and fluency [1]. Even if there is agreement on these dimensions, there is no agreement on the question and scale used for evaluation (see [4, 5, 6, 7]). Even if Likert scales [8] are often used in text simplification evaluation and other evaluation tasks, many options exist for using and interpreting a Likert scale [9].

In this paper, we analyse the interpretation of different existing scales, including Likert scales, of human evaluation in 6 text simplification datasets. We investigate whether different scale interpretations exist by looking at human ratings of simplification pairs for which the original and the simplified sentences are identical. In detail, we answer the following research questions: I) Do human annotators agree on one label in the judgment of simplicity of identical sentence pairs, e.g., the middle or the lowest score value? II) Do human annotators agree on one label in the judgment of meaning preservation of identical sentence pairs, i.e., the highest score value? III) Do human annotators stick to their interpretation of a rating scale in all of their ratings?

In the following, we will first summarise the state of the art in current manual evaluation of text simplification. Then, we describe our methods and data and build our hypotheses. Afterwards, we present our results, conclude with some final interpretation and discussion of the results and mention possible future works.

2. Related Work

The human evaluation of natural language processing tasks is very costly and time-consuming, hence, automatic metrics are developed and optimised. For text simplification evaluation also some evaluation directions exist, e.g., evaluation on multiple references [2], evaluation without any reference [10] or evaluation of structural simplifications [11]. But all of these metrics still have some limitations, hence, they should be only used for quickly comparing and assessing different text simplification systems [1].

For a more detailed evaluation, human evaluation is required. In human text simplification evaluation, common evaluation dimensions exist, .i.e., meaning preservation, simplicity, and grammatically [1], but there is no agreement on the questions asked per evaluation dimension or the scale used for evaluation.

For the same dimensions, e.g., fluency (also called grammaticality), several definitions and questions exist: in [11], they ask the raters if the output sentence is grammatical, [12] ask if the simplified sentence is grammatical and fluent, and [13] state that ”fluency indicates if the output is syntactically correct.”. Even if the statements sound similar, they emphasise different points and, hence, the raters may focus on different points during the rating. Especially if a rater is not an expert in text simplification, minor differences may lead to incomparable results. There is also a discussion of whether sentence pairs should be rated by experts or crowd workers of the target group [1].

Furthermore, there is no agreement on a rating scale, most approaches prefer Likert scales (see [6, 5]) but others prefer continuous scales (see [7]). However, Likert scales are also differently used, e.g., a scale ranging from 1 to 5 (see [5]) or -2 to +2 (see [6]). Following [9], Likert scales can also be differently used regarding other aspects, e.g., single-item vs. multi-item, same distance between consecutive points (ordinal vs. interval), odd or even number of points, each point labeled vs. only end points labeled, descending vs. ascending order, negatively or positively stated items.

In text simplification evaluation, the most common rating scales are 5 point Likert-scales, e.g., [5], a scale from -2 to +2, e.g., [6], and a continuous scale from 0 to 100, e.g., [14, 7]. On the one hand, [7] argue that a continuous scale leads to more consistency in inter-annotator agreement in text simplification evaluation as already proofed for machine translation. On the other hand, [6] prefer a Likert scale with negative to positive scale points including a neutral middle point because they are helpful to rate sentence pairs in which the simplified sentence is more or equally complex as the original sentence. However, both scales include a middle point element. Following [9, 15], annotators interpret the middle point as, e.g., ”undecided”, ”neutral”, or ”no opinion”, which might be not always the interpretation the scale developers have intended. Overall, the different scales and their interpretations make it difficult to compare the ratings of different system outputs and, therefore, distort text simplification evaluation.

3. Method

3.1. Data

As we want to analyse the interpretation of rating scales by different annotators, a dataset with ratings of at least 2 annotators is required. Therefore, in our analysis, we focus on QATS [16][1], HSplit [6][2], PWKP test [11][3], ASSET [7][4], human-likert and system-likert [14][5], and Fusion [18, 19][6]. An overview of their relevant evaluation dimensions, scales and number of raters per dataset is given in Table 1.

Additionally, in all datasets, grammaticality is also rated. However, it is rated only absolutely on the simplified sentence and not in relation to the original sentence, so we do not consider it in the analysis. The simplicity rating of QATS is also not considered for the same reason.

3.2. Hypotheses

The dataset selection already showed the differences between the human evaluation in text simplification. Even if the name and the idea behind the evaluation dimensions are very similar, the judgements are collected I) on scales with different sizes, i.e., 3, 5 and 100, II) on scales with different point names, i.e., ”good”, to ”bad” or ”strongly disagree” to ”strongly agree”,

[1] The data is available online https://qats2016.github.io/shared.html.

[2] The human judgements of HSplit are available online https://github.com/eliorsulem/simplification-acl2018.

[3] Due to a currently dead link to the system outputs of the sentence pairs, we instead copied the system outputs provided in EASSE [17] in the given order. However, the sentence pairs of 2 system outputs could not be found. Hence, our version of the dataset contains only 500 sentence pairs. The human judgements are available online https://github.com/eliorsulem/SAMSA/blob/master/Human_evaluation_benchmark.ods. The original sentences and system outputs are available in EASSE https://github.com/feralvam/easse/tree/master/easse/resources/data.

[4] The human judgements of ASSET are available online https://github.com/facebookresearch/asset/tree/master/human_ratings.

[5] The human judgements are available online http://dl.fbaipublicfiles.com/questeval/simplification_human_evaluations.tar.gz.

[6] The data will be available here https://cs.pomona.edu/~dkauchak/simplification/. Currently it is only available upon request by the authors.

Table 1
Overview of the evaluation dimensions, scales and raters (CW = crowd workers) and sources per dataset.

dataset	simplicity		meaning preservation		Source	# raters
	definition	scale	definition	scale		
QATS	-	1 (bad), 2 (ok), 3 (good)	-	1 (bad), 2 (ok), 3 (good)	EventS, EncBrit, LSLight	-
Hsplit	”Is the output simpler than the input?”	-2 to +2	Does the output preserve the meaning of the input?	1 to 5	TurkCorpus	3 experts
PWKP test		1 (no), 2 (maybe), 3 (yes)	Does the output add information, compared to the input? Does the output remove important information, compared to the input?	1 (no), 2 (maybe), 3 (yes)	PWKP	5 experts
ASSET	The simplified sentence is easier to understand than the original sentence.	0 (”strongly disagree”) to 100 (”strongly agree”)	The simplified sentence adequately express the meaning of the original, perhaps omitting the least important information.	0 (”strongly disagree”) to 100 (”strongly agree”)	TurkCorpus	ASSET: 15 CW & HL+SL: 12-35 CW
Fusion	How much simpler is sentence 2 than sentence 1	-2 (much less simple) to +2 (much simpler)	Sentence 2 preserves the meaning of sentence 1	1 to 5	Newsela	3 CW

III) by crowd workers or experts, an IV) on different item types, i.e., questions or statements, V) on different types of simplification pairs, i.e., manually or automatically simplified sentences, VI) on sources which are reused for text simplification, e.g., English Wikipedia and Simple English Wikipedia (in HSplit) or which are directly designed for text simplification (in ASSET, human-likert), VII) on sentence pairs with different aspirations in the simplicity level, e.g., the simplified sentence must be simpler or the simplified sentence can also be more complex. Hence, the following points make it difficult to compare judgements of text simplification systems reported in system papers. In the following, we will analyse if more problems in human evaluation exist. Therefore, we analyse if the annotators consistently understand the scales in each of the datasets.

To analyse the interpretation of the scale, we compare the ratings of simplification pairs in which no change was made from the original to the simplified sentence. These sentence pairs are further called *no-change pairs*. As complexity assessment is a subjective task, different ratings of the simplifications are expected. But if the simplified sentence is identical to the original sentence, the rating can be expected to be the same because not the absolute simplicity of the sentence is measured but the change/simplification which does not exist in this case. Hence, we use the no-change pairs of the datasets to check whether different interpretations of the rating scales exist. An overview of the proportion of no-change pairs per dataset and their size themselves are given in Table 2.

We will focus on the analysis of the evaluation dimensions of simplicity and meaning preservation. The interpretation of the *grammaticality* dimensions couldn’t be analysed as in all datasets grammaticality was only rated for the simplified sentence but not for the original sentence. In the analysis, we will verify the following hypotheses, which are based on the dataset and scale descriptions in the previous section.

Hypothesis 1: In HSplit and Fusion, the *simplicity rating* of no-change pairs are equal to the neutral element, i.e., 0.
The simplicity ratings in HSplit are judged on a scale ranging form -2 to +2 including the neutral element 0. Following the scale definition in [6], the neutral element of the scale indicates that the simplicity of the original and the simplified sentence of a pair are the same. Hence, we hypothesise that the simplicity ratings of no-change pairs in HSplit and Fusion are equal to 0. A score of -2 indicates a more complex simplified sentence and +2 a more easy simplified sentence compared to the original sentence.

Hypothesis 2: In ASSET, human-likert, and system-likert, the *simplicity ratings* of no-change pairs are equal to the lowest element of the scale, i.e., 0, as it indicates the worst simplification.
In ASSET, human-likert, and system-likert, the annotators rate the relative simplicity of simplification pair based on their level of agreement to a given statement. The scale ranges from 0 (strongly disagree) to 100 (strongly agree). Hence, the lowest value indicates a rejection of the statement, which is interpretable as the worst simplification. In the rating instruction[20], the question raises how to annotate the sentence pair if the original and the simplified sentence are exactly the same. The answer refers to the formulation of the dimension that some change should have been made. However, it does not indicate an expected behaviour of the annotators, e.g., do not judge an identical pair or judge with a specific value. Hence, we can only assume that the lowest score, i.e., 0, indicates that the simplified sentence is more complex than the original sentence as well as that the simplified sentence is as simple/complex as the original sentence. Following this interpretation, a score of 50 would indicate that the simplified sentence is more, roughly 50% more, simple than the original sentence.

Hypothesis 3: The *meaning preservation rating* is equal to the maximum element in QATS, HSplit, PWKP test, ASSET, human-likert, system-likert, and Fusion.
In no-change pairs, the meaning of the original sentence is exactly the same as in the simplified sentence. As meaning preservation measures in all corpora the extent to which the meaning is preserved in the simplified compared to the original sentence, we hypothesise the highest possible value for no-change pairs in the evaluation dimension of meaning preservation. The highest possible value for QATS and PWKP test[7] is 3, for HSplit 5, and for ASSET, human-likert and system-likert 100, respectively.
In contrast, the lowest possible values would indicate that the simplified sentence has a completely different meaning than the original sentence. Even if the scale has a middle element, this element does not have to indicate a neutral element as for the simplicity scale in HSplit. Following [9], it is also possible to express the indecision of the rater, which is more likely in this case.

Hypothesis 4: If different interpretations of the scales exist, the rater groups' ratings significantly differ for sentence pairs in which the original and the simplified sentences are not identical.

[7] In PWKP test, the meaning preservation score is based on the averaged reversed ratings of information gain and information loss (see [11]).

If at least one of the previous hypotheses can be disproved, the rating behaviour of the annotators will be analysed in more detail. The deviation in the scores of the hypothesis lead to the assumptions that the raters differently understood the rating scales. To evaluate the extent of the misunderstanding, we compare the ratings per sentence pair, including sentence pairs with changes, of different rating groups, e.g., preferring the highest and middle value of the scale. For example, if a rater group rated the simplicity of no-change pairs of ASSET with 50 and not with the assumed 0 score, we have a closer look at their simplicity ratings on pairs with a change. If a rater group prefers 50 for no-change pairs, they most likely annotate the pairs with a change differently than the rater group preferring 0. Hence, it is hypothesised that the ratings of such rater groups significantly differ from each other.

4. Results and Discussion

Each of the selected datasets contains some no-change pairs which are rated by a different number of annotators. An overview of the number of no-change pairs and annotators of no-change pairs per dataset is provided in Table 2. For the ratings of ASSET, human-likert, and system-likert, we normalised the human judgements by their individual mean and standard deviation, following the description in [7, 14]. In the following, we will analyse the raters' interpretation of the dimensions simplicity and meaning preservation to disprove or corroborate the hypotheses.

Table 2
Overview of the size of the datasets. Sentences correspond to the number of different original sentences in the datasets, whereas sentence pairs correspond to the number of different simplification pairs, e.g., produced by different systems or humans. No-change pairs are sentences in which the original and the simplified sentence are exactly the same. An annotation record is a rated score on one of the evaluation dimensions of one of the sentence pairs by one rater.

	QATS	HSplit	PWKP test	ASSET	system-likert	human-likert	Fusion
# sentences	139	70	100	100	100	100	319
# sentence pairs	631	1960	500	100	151	108	2920
# no-change sentence pairs	107	346	20	5	2	3	338
# annotation records	1893	7840	2000	4500	9273	9357	17520
# no-change annotation records	321	1384	80	225	90	126	2028
# no-change in %	16.96	17.65	4	5	0.97	1.35	11.58
# no-change pair raters		3	5	23	19	30	3

4.1. Simplicity Rating

In HSplit, the ratings of the experts are consistent and corroborate Hypothesis 1. All ratings of the 346 identical sentence pairs agree on the assumed neutral value of 0, except one annotator for three of overall 7840 annotation records (0.03%).

In Fusion, on average, only 6 of 338 no-change sentence pairs (1.88%) are not scored with the neutral value as assumed in Hypothesis 1. The overall average score of all no-change pairs' simplicity judgements of all three annotators are equal to -0.0026±0.05. Interestingly deviations in both directions exist, i.e., closer to more simple and closer to more difficult.

In ASSET, the annotators do not agree with their ratings for the no-change pair on the dimension of simplicity. For each pair, roughly half of the annotators decides on the minimum value, which is hypothesised, and roughly the other half on the middle value. One annotator per no-change pair rates simplicity with the highest possible score. In contrast to Hypothesis 2, the simplicity ratings in ASSET are not always equal to the lowest element.

Table 3
Overview of the simplicity ratings (normalised with their mean and standard deviation) of all annotators which rated more than one no-change pair of human-likert or system-likert. The first column contains an anonymised version of the worker ids (each worker is assigned to an id following the occurrence of its name in the dataset). The last two columns in both tables highlight if the annotators rate a similar score for the pairs (same) or not (differ).

worker_id	90	143	199	207	265	same	differ
0	5.56	-	3.37	-	5.29	x	
7	0.82	0.82	2.02	-	2.02	x	
8	52.98	1.19	55.67	-	2.16		x
12	1.72	1.72	5.59	-	3.65	x	
13	49.91	49.91	51.33	-	51.33	x	
17	47.70	39.89	81.71	-	-		x
19	49.59	49.59	-	-	49.91	x	
21	0.66	0.66	1.43	-	1.43	x	
24	1.07	1.07	1.97	-	1.97	x	
26	-	49.95	51.14	-	-	x	
30	0.73	0.73	2.20	-	2.20		x
32	9.67	96.02	94.16	-	-		x
35	98.51	-	85.63	-	50.85		x
36	50.02	-	51.08	-	51.08	x	
44	-	-	91.80	-	2.80		x
45	49.94	49.94	51.14	-	51.14	x	

Similar to the annotators' behaviour in ASSET, in human-likert and system-likert, the annotators can be split into three rating groups: preferring 0 for no-change pairs, preferring 50 or 100. Again, in contrast to Hypothesis 2, the simplicity ratings in human-likert and system-likert are not always equal to the lowest element. Further analysis is required to check whether these ratings are done by mistake or due to different scale interpretations (see subsection 4.3).

The results of these datasets show that some crowd-workers and experts interpret the simplicity scale as hypothesised. In contrast, the number of points and points ranging from negative to positive or only positive, seem to influence the interpretation of the scale: Both datasets with a scale ranging between -2 and +2 achieved a higher consistency than the scale ranging between 0 and 100. The different interpretations might be due to different understandings of the middle point of the scale [9].

4.2. Meaning Preservation Rating

For the dimension of meaning preservation, the human ratings in human-likert, system-likert and ASSET rather meet the values assumed in the hypotheses, i.e., close to 100. For all 5 identical pairs, more than 80% rate a score of the maximum category (80 to 100). But for some of the sentence pairs, one of the annotators rate the meaning preservation also either with a value between 0 and 19 or 40 and 59. Hence, a small proportion also interpreted the scale differently than hypothesised in Hypothesis 3. In comparison to the simplicity rating scale, the meaning preservation scale seems clearer to understand, which might be due to a clearer formulation of the scale item.

The annotators of HSplit again agree all in the same rating, here the maximum value, except for 8 out of 346 identical pairs (2.31%). Furthermore, in QATS all no-change pairs are rated with the highest possible value, which is "good". In Fusion, 15 of the 338 no-change pairs (4.43%) were rated with a different value than the highest value. The overall average score of all no-change pairs' meaning preservation judgements of all three annotators is equal to 4.98±0.12. Hence, Hypothesis 3 can also be approved for HSplit, QATS and Fusion.

In contrast, in PWKP test, the ratings are below the values hypothesised. Each of the annotators rated the no-change pairs with a score ranging on average from 2.275 to 2.525. 3 of the 5 annotators annotated half of the no-change pairs with the highest value, but another rater only selected it for 30% of the pairs. Hence, for PWKP test Hypothesis 3 is disproved. However, it must be considered that the alignment of PWKP test was reproduced. Hence, the results of PWKP must be interpreted with caution because the found effects might be due to a misalignment.

4.3. Consistent Interpretations

Roughly half of the annotators in ASSET, human-likert and system-likert either annotated simplicity with the lowest value and the other half with the middle value. As stated in Hypothesis 4, we analyse whether the annotators stick to their scale interpretation or not.

In system-likert and human-likert, 16 of 34 annotators rated more than one no-change pair on the simplicity dimension. 10 of the 16 annotators are consistent in their ratings (see Table 3), they rated the no-change pairs all either with a score between 0 and 19 or 40 and 59. Looking closer at the ratings, 5 of the 10 raters, decided on a score between 0 and 19 on all of their no-change pairs, as hypothesised in Hypothesis 2, and the other half on a score between 80 and 100. However, also 6 of 16 raters alternate between the lowest, middle or highest value, hence, they seem to have no clear scale interpretation.

In ASSET, 20 crowd workers annotated more than one no-change pair. 13 of them always annotated the same value for all simplicity ratings of their no-change pairs (see Table 4). Similar to system-likert and human-likert, the annotators are split into nearly equally sized groups preferring either the lowest or the highest value for simplicity. Overall, we, can confirm, that different simplicity scale interpretations occur in system-likert, human-likert and ASSET. The different understandings of the lowest value might be due to a not-intended misinterpretation of the middle value [9].

To further investigate the different scale interpretations also on simplification pairs with a change, we divided the raters into groups based on their preferred score on the no-change pairs, i.e., preference-1 and preference-50. The groups are compared sentence-wise on the evaluation dimension of simplicity.

In the averages of the simplicity rating of both groups, the different interpretations are also

Table 4
Overview of the simplicity ratings (normalised with their mean and standard deviation) of all annotators which rated more than one no-change pair of ASSET. The last two columns in both tables highlight if the annotators rate a similar score for the pairs (same) or not (differ).

worker_id	67	90	143	200	311	same	differ
0	2.46	-	2.46	-	-	x	
2	49.94	98.51	-	-	-		x
3	45.16	52.98	1.19	52.98	1.19		x
5	-	5.56	-	2.66	2.66	x	
6	0.82	0.82	0.82	0.82	0.82	x	
7	-	49.59	49.59	49.59	49.59	x	
8	0.66	0.66	0.66	0.66	0.66	x	
11	2.88	9.67	96.02	-	-		x
12	49.80	-	-	50.77	49.80	x	
13	1.07	1.07	1.07	1.07	1.07	x	
14	-	47.70	39.89	97.47	48.67		x
15	49.91	49.91	49.91	49.91	49.91	x	
18	49.74	-	49.74	-	-	x	
19	3.67	1.72	1.72	1.72	4.64	x	
20	49.95	-	49.95	5.36	5.36		x
23	-	49.76	-	49.76	49.76	x	
26	-	-	-	3.46	97.77		x
27	0.73	0.73	0.73	0.73	0.73	x	
28	49.94	49.94	49.94	-	-	x	
29	-	-	-	28.44	75.11		x

present. In system-likert and human-likert, the rater group preference-1 (n_{raters}=5, $n_{ratings}$=911, M=52.87±40.18) have an overall lower simplicity average than the rater group preference-50 (n_{raters}=5, $n_{ratings}$=634, M=63.77±33.88) on simplification pairs with a change. The same applies also to ASSET: preference-1 (n_{raters}=7, $n_{ratings}$=571, M=35.58±37.24), preference-50 (n_{raters}=6, $n_{ratings}$=292, M=44.43±33.22).

Comparing all sentence pairs with changes rated by both rater groups using a Mann-Whitney-U-test, the simplicity ratings are significantly differing between both groups in system-likert and human-likert (U=252213.0, $p \leq 0.01$) and ASSET (U=64127.0, $p \leq 0.01$). Hence, it seems that both groups interpret the simplicity scale differently, but apply their different interpretations to all rated pairs. Hypothesis 4 can be corroborated.

5. Conclusion and Future Work

Concerning the research questions asked, in the dataset analysed, human annotators (experts and crowd workers) mostly agree on one label, i.e., the highest value of the scale, in the judgements of meaning preservation. In contrast, the analysis has also shown that different scale interpretations exist for the evaluation dimension of simplicity in the dataset with crowd-sourced human ratings on a continuous scale. Some raters prefer the lowest value and some the middle value of the scale to indicate the same level of simplicity in no-change pairs. However, the values are not randomly seeded, a clear distinction between raters who annotate the lowest or neutral element on several no-change pairs is possible. This leads to the assumption that they did not rate the lowest or the middle element by mistake but understood the scale differently.

Following the analysis results, on the one hand, the interpretation of the simplicity scale is consistent when rated by experts or using a neutral element for simplicity. On the other hand, crowd-workers had different interpretations of the simplicity scale, i.e., either the lowest or the middle element of the scale indicate no change in simplicity. The scale and the annotations also could get clearer, e.g., by reformulating the definition or scale ending, the crowd-workers could get more certain by seeing more examples before the annotation, or one could rely only on (trained) experts.

In contrast, the expert ratings in HSplit, PWKP test and the crowd worker ratings in Fusion regarding all evaluation dimensions and the ratings in ASSET regarding *meaning preservation* are congruent with the values hypothesised. Overall, a deeper analysis of the interpretation of human rating scales in text simplification is required. Therefore, a user study could be conducted in which several sentence pairs with and without changes would be rated on different scales or with different instructions by crowd workers and experts.

Not only the different scale interpretations by human raters but also the different implementations of the scales of human evaluation limit the comparison of human evaluation of text simplification. Hence, best practices as e.g. published for natural language generation [21] are in high demand for text simplification. We hope that this paper increases the awareness of problems in text simplification evaluation and kicks off a discussion regarding these challenges, e.g., training of human annotators or showing examples to them, developing clear and precise statements or questions for evaluation dimensions, best number of points of a scale, e.g., 0 to 100 or -2 to +2, or rating of experts or crowd workers.

Acknowledgments

This research is part of the PhD-program ”Online Participation”, supported by the North Rhine-Westphalian (German) funding scheme ”Forschungskolleg”. We thank the anonymous and non-anonymous reviewers for their valuable feedback during the preparation of this paper.

References

[1] F. Alva-Manchego, C. Scarton, L. Specia, Data-driven sentence simplification: Survey and benchmark, Computational Linguistics 46 (2020) 135–187. URL: https://aclanthology.org/2020.cl-1.4. doi:10.1162/coli_a_00370.

[2] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415. URL: https://aclanthology.org/Q16-1029. doi:10.1162/tacl_a_00107.

[3] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia,

Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[4] S. Stajner, Automatic text simplification for social good: Progress and challenges, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 2637–2652. URL: https://aclanthology.org/2021.findings-acl.233. doi:10.18653/v1/2021.findings-acl.233.

[5] M. Maddela, F. Alva-Manchego, W. Xu, Controllable text simplification with explicit paraphrasing, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 3536–3553. URL: https://aclanthology.org/2021.naacl-main.277. doi:10.18653/v1/2021.naacl-main.277.

[6] E. Sulem, O. Abend, A. Rappoport, Simple and effective text simplification using semantic and neural methods, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 162–173. URL: https://aclanthology.org/P18-1016. doi:10.18653/v1/P18-1016.

[7] F. Alva-Manchego, L. Martin, A. Bordes, C. Scarton, B. Sagot, L. Specia, ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4668–4679. URL: https://aclanthology.org/2020.acl-main.424. doi:10.18653/v1/2020.acl-main.424.

[8] R. Likert, A technique for the measurement of attitudes, Archives of Psychology 22 (1932).

[9] S. Y. Y. Chyung, K. Roberts, I. Swanson, A. Hankinson, Evidence-based survey design: The use of a midpoint on the likert scale, Performance Improvement 56 (2017) 15–23. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pfi.21727. doi:https://doi.org/10.1002/pfi.21727. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/pfi.21727.

[10] L. Martin, S. Humeau, P.-E. Mazaré, É. de La Clergerie, A. Bordes, B. Sagot, Reference-less quality estimation of text simplification systems, in: Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA), Association for Computational Linguistics, Tilburg, the Netherlands, 2018, pp. 29–38. URL: https://aclanthology.org/W18-7005. doi:10.18653/v1/W18-7005.

[11] E. Sulem, O. Abend, A. Rappoport, Semantic structural evaluation for text simplification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 685–696. URL: https://aclanthology.org/N18-1063. doi:10.18653/v1/N18-1063.

[12] S. Narayan, C. Gardent, Hybrid simplification using deep semantics and machine translation, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 435–445. URL: https://aclanthology.org/P14-1041. doi:10.3115/v1/P14-1041.

[13] Z. Lin, X. Wan, Neural sentence simplification with semantic dependency information, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 13371–13379. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17578.

[14] T. Scialom, L. Martin, J. Staiano, Éric Villemonte de la Clergerie, B. Sagot, Rethinking automatic evaluation in sentence simplification, 2021. arXiv:2104.07560.

[15] J. T. Nadler, R. Weston, E. C. Voyles, Stuck in the middle: The use and interpretation of mid-points in items on questionnaires, The Journal of General Psychology 142 (2015) 71–89. URL: https://doi.org/10.1080/00221309.2014.994590. doi:10.1080/00221309.2014.994590. arXiv:https://doi.org/10.1080/00221309.2014.994590, pMID: 25832738.

[16] S. Štajner, M. Popović, H. Saggion, L. Specia, M. Fishel, Shared task on quality assessment for text simplification, in: Proceedings of the Workshop on Quality Assessment for Text Simplification (QATS), Association for Computational Linguistics, Portorož, Slovenia, 2016, pp. 22–37. URL: http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-QATS_Proceedings.pdf#page=28.

[17] F. Alva-Manchego, L. Martin, C. Scarton, L. Specia, EASSE: Easier automatic sentence simplification evaluation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 49–54. URL: https://aclanthology.org/D19-3009. doi:10.18653/v1/D19-3009.

[18] M. Schwarzer, Crowdsourcing Text Simplification with Sentence Fusion, Bachelor thesis, Pomona College, 2018. URL: https://cs.pomona.edu/classes/cs190/thesis_examples/Schwarzer.18.pdf.

[19] M. Schwarzer, T. Tanprasert, D. Kauchak, Improving human text simplification with sentence fusion, in: Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15), Association for Computational Linguistics, Mexico City, Mexico, 2021, pp. 106–114. URL: https://aclanthology.org/2021.textgraphs-1.10.

[20] F. Alva-Manchego, Automatic Sentence Simplification with Multiple Rewriting Transformations, Phd thesis, University of Sheffield, Sheffield, UK, 2020. URL: https://etheses.whiterose.ac.uk/28690/.

[21] C. van der Lee, A. Gatt, E. van Miltenburg, S. Wubben, E. Krahmer, Best practices for the human evaluation of automatically generated text, in: Proceedings of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics, Tokyo, Japan, 2019, pp. 355–368. URL: https://aclanthology.org/W19-8643. doi:10.18653/v1/W19-8643.