# Reproducing a Manual Evaluation of the Simplicity of Text Simplification System Outputs

**Maja Popović, Sheila Castilho, Rudali Huidrom and Anya Belz**
ADAPT Centre
School of Computing
Dublin City University, Ireland
name.surname@adaptcentre.ie

## Abstract

In this paper we describe our reproduction study of the human evaluation of text simplicity reported by Nisioi et al. (2017). The work was carried out as part of the ReproGen Shared Task 2022 on Reproducibility of Evaluations in NLG. Our aim was to repeat the evaluation of simplicity for nine automatic text simplification systems with a different set of evaluators. We describe our experimental design together with the known aspects of the original experimental design and present the results from both studies. Pearson correlation between the original and reproduction scores is moderate to high (0.776). Inter-annotator agreement in the reproduction study is lower (0.40) than in the original study (0.66). We discuss challenges arising from the unavailability of certain aspects of the original set-up, and make several suggestions as to how reproduction of similar evaluations can be made easier in future.

## 1 Introduction

Against a background of growing interest in approaches to reproducibility assessment in general, and specific reproduction studies in particular, this paper reports a reproduction study of a human evaluation of text simplicity carried out as part of the ReproGen Shared Task 2022 on Reproducibility of Evaluations in NLG. We participated with a contribution in Track A, carrying out a reproduction study of the human evaluation of sentence simplicity reported by Nisioi et al. (2017), one of the five papers offered in the track.

In the original paper, nine automatic text simplification systems were evaluated by human annotators for four different criteria: Correctness and number of changes, Meaning Preservation, Grammaticality, and Simplicity. In this paper, we concentrate only on Simplicity. We first summarise the original study and describe the details of our reproduction study (Section 2. We then present the results from both studies (Section 3) in terms of the system-level Simplicity scores of the nine systems, and the inter-annotator agreement estimated as quadratic Cohen's Kappa. We also report Pearson's correlation coefficient between the original and the reproduction system scores.

We finish (Section 4) with a discussion of the differences between the two studies and the impact of missing information about the original set-up, and suggest how to make future human evaluations easier to repeat.

## 2 Experimental Design in Original and Reproduction Study

A commonly cited motivation for automatic text simplification (ATS) systems is that texts containing uncommon words or long and complicated sentences can be difficult to read and understand by people as well as difficult to analyze by machines. ATS is the process of transforming one text into another text which ideally has the same meaning, but is easier to read and understand by a wider audience and also easier to process with NLP tools. ATS systems can be rule-based or corpus-based, namely trained on parallel corpora consisting of original texts and their simplified versions.

For human evaluation of ATS systems, the usual quality criteria are Meaning Preservation (the degree to which the meaning of the original text is retained in the simplified output; analogous to Adequacy in MT), Grammaticality (whether the grammar of the generated output is good), and Simplicity (how difficult/simple the generated output is).

This paper focuses on simplicity evaluation in the form of comparing the automatically simplified output with the original text that was the input: the original sentence is presented together with its automatically simplified version, and the evaluators are asked whether the simplified version is simpler, equally simple/difficult, or more difficult than the original.

*Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 80 - 85
July 17-22, 2022 ©2022 Association for Computational Linguistics

### 2.1 Original experiment

The original paper (Nisioi et al., 2017) reported the first attempt of using neural networks for automatic text simplification. Two basic neural text simplification (NTS) system variants for the English language were developed, one relying only on internal word representations (which we refer to as NTS in tables and results below), and the other additionally using external word2vec representations (NTS-W2V). Each system variant was used to generate outputs in three different ways: (i) by beam search with size 5 (NTS-DEFAULT and NTS-W2V-DEFAULT), (ii) by re-ranking an n-best list using the automatic metric BLEU (Post, 2018) (NTS-BLEU and NTS-W2V-BLEU), and (iii) by re-ranking using the SARI metric (Xu et al., 2016) (NTS-SARI and NTS-W2V-SARI). These six system variants together with an additional three publicly available systems (for which outputs generated in previous work were available), referred to as PBSMT, SARI+PPDB and LIGHTLS in results tables and briefly explained in the next section, were manually evaluated in terms of the three criteria of Meaning Preservation, Grammaticality and Simplicity. In addition, BLEU and SARI scores were calculated.

The outputs from all nine systems, as well as scripts for both automatic evaluation metrics are publicly available.[1] Human sentence-level annotations are however not published, and only the system-level scores were reported in the paper.

#### 2.1.1 Evaluation Data

The developed NTS systems were evaluated on 359 publicly available sentences originating from English Wikipedia[2] and previously released by Xu et al. (2016). These sentences were simplified with the NTS system variants from Nisioi et al. (2017) as well as the three previous systems: PBSMT, a phrase-based SMT system with reranking (Wubben et al., 2012), SARI+PPDB, a paraphrase-based system proposed by Xu et al. (2016), and LIGHTLS, an unsupervised lexical simplification system based on word embeddings (Glavaš and Štajner, 2015).

For each of the nine systems, automatic scores were calculated on all sentences, whereas human evaluation was carried out on the first 70 sentences only. Since each sentence was simplified by 9 systems, 630 sentences were manually evaluated in total.

### 2.2 Evaluating simplicity

In both original and reproduction study, the manual evaluation of simplicity was performed by three non-native English speakers who were given the original sentence and an automatically generated simplification of it, one pair at a time. They were asked to assign a score to each pair according to the following guidelines:

- +2 if the simplified version is much simpler than the original,
- +1 if the simplified version is somewhat simpler than the original,
- 0 if they are equally simple/difficult,
- -1 if the simplified version is somewhat more difficult than the original, and
- -2 if the simplified version is much more difficult than the original.

The inter-annotator agreement reported by Nisioi et al. (2017) (in the form of quadratic Cohen's Kappa) was 0.66.

The reported aggregated system-level scores (mean sentence-level scores, shown in Table 1, Simplicity/original/score column) indicated that all variants of the newly proposed NTS model substantially outperform all of the comparator systems in terms of simplicity, i.e. generate outputs with a higher level of simplicity than the three previous state-of-the-art ATS systems.

### 2.3 Reproduction study

Our reproduction experiment was carried out on the same data as the original one, namely the first 70 sentences of the test set simplified by each of the nine systems. The evaluation was carried out by three non-native speakers, too, same as in the original evaluation. They received the same instructions as described in the original paper and in Section 2.1.

Further details about the original evaluation which may or may not have affected results and reproducibility were, however, not available.[3] Such details where we have information only for our reproduction include:

[1] https://github.com/senisioi/NeuralTextSimplification

[2] https://github.com/cocoxu/simplification/

[3] After contacting the authors of the original paper, the responses received were from authors not familiar with the details requested.

- *Native languages of evaluators*

  Reproduction: each evaluator had a different native language (Serbian, Brazilian Portuguese and Manipuri).

- *Evaluators' background*

  Reproduction: all the evaluators were computational linguistics researchers.

- *Evaluators' experience with TS and its evaluation*

  Reproduction: one evaluator had experience with TS evaluation and thus was familiar with the concept of simplicity, whereas the other two did not.

- *Whether the evaluators were able to ask any additional questions or only worked with the above guidelines*

  Reproduction: the two evaluators without experience needed a few additional instructions and examples in order to fully understand the concept of simplicity in this context, and to be able to separate it from meaning and grammar.

- *Number of sentences assessed by each evaluator*

  Reproduction: one evaluator (the one with the experience with TS evaluation) annotated all sentences whereas the other two evaluators annotated half of the sentences each.

  As with the other details in this list, we do not know how the sentences were distributed among the three evaluators in the original study.

- *Number of multiply annotated sentences used for IAA*

  Reproduction: each sentence was annotated by two evaluators, IAA is computed on the whole set.

  We do not know whether this was the case in the original experiment or only a subset of sentences was annotated by more than one evaluator. We also do not know whether any (or all) sentences were evaluated by all three evaluators.

It might also be worth noting that in our reproduction identical sentence pairs (where the output is identical to the input) were not presented to the evaluators but were immediately assigned the score 0. We do not know whether the same was the case in the original evaluation.

## 3 Results

### 3.1 Comparing the different ATS systems

The 'original' column in Table 1 presents the ranks and system-level reproduction scores obtained for the nine systems in the original study, and the 'reproduction' column presents the same for the reproduction study. It can be seen that overall, the three ATS systems from previous work, PBSMT, SARI+PPDB and LIGHTLS, have notably lower reproduction scores in both studies, so that the claim from the original paper that the proposed NTS systems generate outputs with higher levels of simplicity is confirmed.

As for comparing the individual NTS systems, the reproduction scores indicate that the NTS-w2v-SARI system (re-ranking with SARI scores) reaches the highest simplicity levels, as well as that the re-ranking is generally beneficial for both model variants. The original scores, on the other hand indicate that re-ranking with automatic metrics was of benefit to the NTS-w2v variant, but for the NTS variant, while re-ranking with BLEU (NTS-BLEU) led to a dramatic improvement in reproduction, re-ranking with SARI (NTS-SARI) actually dropped the reproduction score. In contrast, according to the reproduction scores, re-ranking with SARI had more of a beneficial effect than re-ranking with BLEU.

The last column in Table 1 shows the small-sample coefficient of variation (CV*) for each of the individual system-level reproduction score pairs across the two experiments as a quantified measure of degree of reproducibility (Belz et al., 2022). Lower CV* indicates better reproducibility. Here, the CV* scores show that some systems' human scores are more reproducible than others, but it is not immediately obvious why the human evaluators in the original and reproduction studies should have disagreed particularly about the two systems with the highest CV* (NTS-BLEU and PBSMT).

Pearson correlation coefficient between the original and the reproduction scores is 0.766, i.e. moderate to high. Spearman's rank correlation is slightly higher at 0.787.

### 3.2 Inter-annotator agreement (IAA)

The IAA in the original experiment was reported as quadratic Cohen's Kappa with a value of 0.66. We also calculated this coefficient for our reproduction, where and the value is lower, 0.40. Unfortunately, we cannot really interpret this discrepancy because,

| automatic text simplification system | Simplicity | | | | small-sample coefficient of variation (CV$^*$) ↓ |
|---|---|---|---|---|---|
| | original | | reproduction | | |
| | rank | score | rank | score | |
| NTS DEFAULT | (3) | 0.46 | (5) | 0.33 | 5.41 |
| NTS-SARI | (5) | 0.38 | (3/4) | 0.34 | 1.69 |
| NTS-BLEU | (1) | 0.92 | (3/4) | 0.34 | 22.0 |
| NTS-W2V-DEFAULT | (6) | 0.21 | (6) | 0.32 | 4.84 |
| NTS-W2V-SARI | (2) | 0.63 | (1) | 0.46 | 6.66 |
| NTS-W2V-BLEU | (4) | 0.40 | (2) | 0.36 | 1.68 |
| PBSMT | (9) | -0.55 | (7) | 0.08 | 35.6 |
| SARI+PPDB | (7) | 0.03 | (9) | 0.01 | 0.99 |
| LightLS | (8) | -0.01 | (8) | 0.03 | 1.98 |

Table 1: System-level Simplicity scores for the nine ATS outputs and system ranks according to these scores, together with CV$^*$s between scores in original and reproduction experiment. Note that CV$^*$ is computed on shifted scores, i.e. while the scores assigned by the human evaluators ranged from -2 to +2, before computing CV$^*$ they were shifted to range from 0 to 4.

as mentioned in Section 2.3, many of the details of the original experiment are missing, and we do not know what subset of sentences IAA was computed over in the original experiment, or how many individual scores per sentence. If the IAA values do reflect an actual difference, then one possible reason might be the experience of the evaluators with TS and familiarity with the notion of simplicity. In the reproduction study, only one evaluator was already familiar with it while the other two required additional explications. Furthermore, due to how sentences were assigned to evaluators, IAA is calculated only between the experienced and inexperienced annotators and not between the two inexperienced. These factors could generally contribute to a lower IAA. On the other hand, it is possible that all evaluators in the original experiment had experience with TS evaluation so that this is the reason of a higher IAA, however this is only a speculation.

Availability of the sentence-level scores from the original study would have helped to compare the scores for each sentence and potentially find patterns in sentences that make human evaluation more difficult to reproduce.

### 3.3 Comparison with reproduction of automatic scores

In order to illustrate quantitatively the differences that can arise between reproducing human and reproducing automatic evaluations, Table 2 presents the Simplicity and CV$^*$ scores for two NTS system variants, NTS-DEFAULT and NTS-W2V-DEFAULT, together with their automatic metric scores (BLEU and SARI). These results are compared and analysed more comprehensively elsewhere (Belz et al., 2022).

The 'original' column shows the results reported in the original paper, the 'repr1' column shows the results reported in an earlier reproduction paper (Cooper and Shardlow, 2020) at REPROLANG 2020[4], the 'repr2' and 'repr3' columns show the results reported by Belz et al. (2022) when using two different evaluation scripts for BLEU, and the 'repr4' column shows results from the human evaluation carried out in the present work.

It can be noted that, while CV$^*$ values for the SARI metric are 0 (perfectly reproduced) and for the BLEU metric are around 1 (reflecting slight differences in implementation and tokenisation), CV$^*$ values for human Simplicity scores are over 4, demonstrating that human evaluation was more difficult to reproduce.

## 4 Conclusions

This paper reported the results of a reproduction study of a human evaluation of text simplicity. The obtained scores confirm some of the findings of the original paper, however findings relating to whether or not re-ranking with BLEU or SARI helped were not aligned in the two studies, in some cases showing opposite effects. Pearson correlation between the studies was moderate to high at 0.766. The inter-annotator agreement was lower in the reproduction study, 0.40 vs. 0.66, but we do not know whether it was computed in a comparable way.

[4] https://lrec2020.lrec-conf.org/en/reprolang2020/

| | | evaluation round | | | | | |
|---|---|---|---|---|---|---|---|
| metric | output | original | repr1 | repr2 | repr3 | repr4 | CV* ↓ |
| BLEU ↑ | NTS default | 84.51 | 84.50 | 85.60 | 84.20 | – | 0.838 |
| (automatic) | NTS-w2v default | 87.50 | – | 89.36 | 88.80 | – | 1.314 |
| SARI ↑ | NTS default | 30.65 | 30.65 | 30.65 | – | – | 0 |
| (automatic) | NTS-v2w default | 31.11 | – | 31.11 | – | – | 0 |
| Simplicity ↑ | NTS default | 0.46 | – | – | – | 0.33 | 5.41 |
| (human) | NTS-v2w default | 0.21 | – | – | – | 0.32 | 4.84 |

Table 2: Comparing CV*s of automatic and human system-level scores for two ATS systems, NTS default and NTS-w2v default. The CV*s indicate that human evaluation is more difficult to reproduce (presumably exacerbated when many experimental details are missing).

A deeper analysis of these differences is unfortunately not possible because we lack too many details for the original set-up. Also, sentence-level human annotations which would be helpful are not published (while the models and the automatic evaluation scripts are).

It appears to be the case that there is a tendency for comprehensive details about the human evaluation process to be reported only in papers dealing with human evaluation itself, although even in these, the provided information is not often fully complete. In papers where human evaluation is not the focus but only a method to assess the system(s), usually only very shallow information is provided, if any. Moreover, it is often the case that the authors themselves perform evaluations, sometimes with no overlap, which makes it impossible to report IAA. Fully reporting such details is disincentivised as doing so may lead to more negative reviews. Human evaluation is time and resource-expensive and it is usually not possible to (i) evaluate large amounts of text, (ii) involve a large number of evaluators, or (iii) evaluate large portions of text by several evaluators for IAA, because all these factors increase cost further.

As in previous work (Howcroft et al., 2020; Belz et al., 2020)), we conclude that reporting more details about human evaluation experiments would be of benefit scientifically. Details of human evaluations should be provided in each paper, even if the conditions were not perfect (and they often are not). It is more scientifically rigorous as well as more useful to provide full details than not providing information for fear of negative review.

## Acknowledgments

All authors benefit from being members of the ADAPT SFI Centre for Digital Media Technology. The Centre is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106. Popović's work is fully supported by the Centre. Huidrom's work is fully supported by the Faculty of Engineering and Computing, DCU, via a PhD grant.

## References

Anya Belz, Simon Mille, and David Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *International Natural Language Generation Conference 2020 (INLG'20)*.

Anya Belz, Maja Popović, and Simon Mille. 2022. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 16–28, Dublin, Ireland. Association for Computational Linguistics.

Michael Cooper and Matthew Shardlow. 2020. CombiNMT: An exploration into neural text simplification models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5588–5594, Marseille, France. European Language Resources Association.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.