A Reproduction Study of an Annotation-based Human Evaluation of MT Outputs

Maja Popović and **Anya Belz**
ADAPT Centre
School of Computing
Dublin City University, Ireland
{maja.popovic,anya.belz}@adaptcentre.ie

Abstract

In this paper we report our reproduction study of the Croatian part of an annotation-based human evaluation of machine-translated user reviews (Popović, 2020). The work was carried out as part of the ReproGen Shared Task on Reproducibility of Human Evaluation in NLG. Our aim was to repeat the original study exactly, except for using a different set of evaluators. We describe the experimental design, characterise differences between original and reproduction study, and present the results from each study, along with analysis of the similarity between them. For the six main evaluation results of Major/Minor/All Comprehension error rates and Major/Minor/All Adequacy error rates, we find that (i) 4/6 system rankings are the same in both studies, (ii) the relative differences between systems are replicated well for Major Comprehension and Adequacy (Pearson's > 0.9), but not for the corresponding Minor error rates (Pearson's 0.36 for Adequacy, 0.67 for Comprehension), and (iii) the individual system scores for both types of Minor error rates had a higher degree of reproducibility than the corresponding Major error rates. We also examine inter-annotator agreement and compare the annotations obtained in the original and reproduction studies.

1 Introduction

Interest in, and concern about, reproducibility is growing in Natural Language Processing (NLP). Reproducibility of human evaluations, however, has received next to no attention, and the ReproGen Shared Task[1] on Reproducibility of Human Evaluations in Natural Language Generation (NLG) addresses this lack. We participated in ReproGen with a contribution in Track B, the Reproduce Your Own Track. More specifically, we repeated the human evaluation of a mixed set of movie and product review translations produced by three leading Machine Translation (MT) systems, as reported by Popović (2020). In this paper, we summarise the original study in terms of the overall evaluation method (Section 2.1), the quality criteria underlying the annotations from which evaluation scores were derived (Section 2.2), the annotation process and instructions (Section 2.3), and the process by which reviews were selected and translated for the evaluation (Section 2.4). We then present Comprehension and Adequacy error rate results from the original and reproduction studies side by side, and look at how similar system rankings and individual scores are in the two studies (Section 3). Next we compare the inter-annotator agreement in the two studies (Section 4) using diverse metrics. We discuss and interpret results obtained in our reproduction study (Section 5), and draw some conclusions (Section 6).

2 Study Design

2.1 Evaluation Method

The core idea behind the annotation-based evaluation method proposed by Popović (2020) is that instead of assigning overall scores to each sentence,[2] or classifying each error into a predefined error scheme, evaluators mark up word spans in translated texts that contain given types of errors. Two error types, corresponding to the two quality criteria *Comprehensibility* and *Adequacy* (see also Section 2.2), were marked up at two levels of severity (*Major* and *Minor*). The method yields both overall error-rate scores (percentage of words that have been marked up for each error type), and a basis for further quantitative and qualitative analysis of errors and challenging linguistic phenomena. In contrast, current manual evaluation methods for

[1] https://reprogen.github.io/

[2] When we say 'sentence' we mean any sentence-like segment, which may consist of just one word or phrase.

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 293–300,
Aberdeen, Scotland, UK, 20-24 September 2021.

MT typically ask annotators either to assign overall per-sentence scores, or to rank two or more translations in terms of given quality criteria, i.e. information about any errors that motivate scores/rankings is not recorded. The method can be applied to any language generation task, genre/domain and language (pair), and can be guided by diverse error types (quality criteria).

2.2 Quality Criteria and Error Rates

The two quality criteria underlying error annotations were *Comprehensibility* and *Adequacy*, both commonly used in MT (ALPAC, 1966; White et al., 1994; Roturier and Bensadoun, 2011).

Comprehensibility: The degree to which a text can be understood. When evaluating Comprehensbility of a translated text, the source language text *is not* shown to evaluators. In terms of the classification system proposed by Belz et al. (2020), Comprehensibility captures the *goodness* of both the *form and content* of a *text in its own right*, and is assessed here by a *subjective, absolute, intrinsic* evaluation measure.

Adequacy (in MT): The degree to which a translation conveys the meaning of the original text in the source language. When evaluating adequacy of a translated text, the source language text *is* shown to evaluators. In terms of Belz et al.'s classification system, Adequacy captures the *correctness* of the *content* of a text *relative to the input*, and is assessed here also by a *subjective, absolute, intrinsic* evaluation measure.

Annotators were asked to mark up translations first for Comprehensibility, then for Adequacy, distinguishing two levels of severity for each: major errors (incomprehensible/not conveying the meaning of the source) and minor errors (difficult to understand due to grammar or stylistic errors/not an optimal translation choice for the given source). Six error rates were then calculated from the mark-up: *Comprehensibility-All*, *Comprehensibility-Major*, *Comprehensibility-Minor*, *Adequacy-All*, *Adequacy-Major*, and *Adequacy-Minor*. These are simply the percentage of words that are part of a text span that has been marked up in the given error category.

2.3 Annotation Process

Annotators first marked up all issues related to Comprehensibility in the translated text without access to the source text. Next, they marked up all issues related to Adequacy while also referring to the source text.

The translated texts were given to the evaluators in the form of a Google Doc, and they were asked to mark major issues with red colour and minor issues with blue colour. In addition to general definitions of Comprehensibility and Adequacy, the evaluators were given detailed guidelines which can be found in the original paper (Popović, 2020).

Evaluators were first given a small number of practice texts to annotate in order to familiarise themselves with the process and clarify any questions and uncertainty. In the original study, these texts were included in calculating the reported results. However, during this practice round in the original study the number and distribution of evaluators varied, which was not repeatable. Therefore, in the reproduction study, the practice texts are not included in calculating reported results, and the results from the original study included in this paper have been adjusted accordingly.

In both studies, each translated review was annotated by two evaluators. All evaluators in both studies were fluent in the source language and native speakers of the target language. However, the backgrounds of the two groups of evaluators are different. In the original study, all seven evaluators working on the Croatian translations were either students or researchers in computational linguistics. Six evaluators had some experience with human translation, and three had experience with machine translation. Three evaluators had a technical background. In contrast, all the evaluators in the reproduction study were translation students, so had the same background and the same or very similar levels of experience with translation.

2.4 Data

The original study involved translations in two similar target languages, Croatian and Serbian, while the reproduction study involved only the Croatian translations, partly for reasons of cost, and partly due to availability of evaluators.

28 English reviews from the Large Movie Review Dataset v1.0[3] (Maas et al., 2011) were selected, as well as 122 English reviews from the 14 categories[4] of the 2018 version of the Amazon

[3] https://ai.stanford.edu/~amaas/data/sentiment

[4] Beauty, Books, CDs and Vinyl, Cell Phones and Accessories, Grocery and Gourmet Food, Health and Personal Care,

	reviews	sentences
	116	894
Amazon MT outputs	68	557
Bing MT outputs	35	279
Google MT outputs	61	467
total MT outputs	164	1303

Table 1: Number of evaluated reviews and sentences.

Product Review dataset[5] (McAuley et al., 2015). In the selection process, overly long ($>$ 350 words) and overly short ($<$ 30 words) reviews were excluded, and an equal number of positive and negative reviews were selected to ensure balance in terms of sentiment polarity, while a balanced distribution between topics in Amazon reviews was also aimed for.

The selected (English) user reviews were then translated into Croatian using Google Translate, Bing and Amazon Translate, yielding a total of 450 Croatian translations of which 164 were arbitrarily selected as a manageable number for inclusion in the evaluation. The 164 selected translations correspond to 116 original English reviews which were mostly translated by one, and in some cases by two, of the MT systems, in order to increase diversity in translations (hence in error types). 68 of the translations were produced by Amazon Translate, 35 by Bing Translator and 61 by Google Translate. The reason for including fewer Bing translations was their notably lower quality. The number of reviews and sentences evaluated for each system can be seen in Table 1.[6]

The primary aim of the original study was not the comparative evaluation of multiple MT systems. Rather, the aim was to test a new evaluation scheme. The system-level error rates reported in the next section can therefore not be considered a fair assessment of the respective quality of the three systems involved. For this purpose, normally the same source texts translated by all systems would be evaluated. In the present context, different source texts translated by different systems were evaluated, chosen as explained above. Nevertheless, assessments of the reproducibility of the obtained human evaluation scores are valid regardless of this diversity in test sets, provided the latter are the same in the original and the reproduction study.

3 Evaluation Scores

Columns 2–7 in Table 2 show the overall evaluation scores obtained in the original and in the reproduction study in the form of error rates, i.e. percentages of words marked up as errors. The following tendencies can be observed in both studies: (i) four out of six system rankings (the exceptions being Major Comprehension and Minor Adequacy) are the same in both studies; (ii) error rates were higher for Comprehension than for Adequacy in all three error subcategories and for all systems, except that the Adequacy-Minor rate for Bing was higher than its Comprehension-Minor rate in the original study which also affected the corresponding All rate; (iii) Bing exhibits the highest error rates in all error categories except Comprehension-Minor and Adequacy-Minor in the reproduction study; and (iv) Google has slightly lower error rates than Amazon in all error categories except for Comprehension-Major and Adequacy-Major in the reproduction study, and Adequacy-Major in the original study).

The last three columns in Table 2 show the coefficient of variation (CV) for each of the individual error-rate scores across the two studies as our primary measure of degree of reproducibility (Belz, 2021). CV is a standard measure of precision in metrological studies of reproducibility.[7] The main general tendencies are as follows: (i) the Adequacy-All and Adequacy-Minor error rates (except for the Minor rate for Bing) have better reproducibility (CV is lower) than the corresponding Comprehension rates; and (ii) the Adequacy-Major error rates (except for the Major rate for Google) have worse reproducibility (CV is higher) than the corresponding Comprehension-Major rates.

Table 3 shows Pearson's r between the system-level Comprehension and Adequacy error rates in the original and the reproduction studies, for each of the All, Major and Minor subcategories. A clear pattern can be observed: correlation between system scores in the Minor categories is far worse than in the All and Major categories.

Since there are only three systems to calculate correlation on, we also calculated Pearson's r between sentence-level error counts and the results are presented in Table 4. The picture confirms the

Home and Kitchen, Movies and TV, Musical Instruments, Patio, Lawn and Garden, Pet Supplies, Sports and Outdoors, Toys and Games, Video Games.

[5] http://jmcauley.ucsd.edu/data/amazon/

[6] The annotated data sets resulting from both the original study and the reproduction study are publicly available under the Creative Commons CC-BY licence here: https://github.com/m-popovic/QRev-annotations

[7] We used the de-biased version of CV, for small samples, as proposed by Belz (2021).

System	Comprehension error rate (%)								
	original study			reproduction study			coefficient of variation (CV)		
	All	Major	Minor	All	Major	Minor	All	Major	Minor
All	21.9	9.2	12.7	29.6	13.4	16.2	29.81	37.06	24.15
Amazon	19.6	**7.6**	12.0	26.9	**10.2**	16.7	31.30	29.13	32.65
Bing	31.1	15.1	16.0	39.1	22.3	16.8	22.72	38.38	4.86
Google	18.3	**7.1**	11.2	26.5	**11.5**	15.0	36.498	47.17	28.92

System	Adequacy error rate (%)								
	original study			reproduction study			coefficient of variation (CV)		
	All	Major	Minor	All	Major	Minor	All	Major	Minor
All	21.1	8.2	12.9	24.8	12.3	12.5	16.07	39.88	3.14
Amazon	17.9	6.5	**11.4**	22.6	9.5	**13.1**	23.14	37.39	13.84
Bing	30.2	13.2	**17.0**	33.9	21.2	**12.7**	11.51	46.37	28.87
Google	17.5	7.0	10.5	21.4	9.7	11.7	19.99	32.24	10.78

Table 2: Error rates (percentages of words that are marked problematic) for Major/Minor Comprehensibility and Adequacy in Croatian translated texts in the two evaluation studies, shown for the three MT systems combined (All) and individually. CV between error rates in original and reproduction for each error category, using the de-biased version of CV proposed by Belz (2021). Bold indicates different system rank in original/reproduction studies.

System-level scores		
	Comprehension	Adequacy
All	0.9979**	0.9982**
Major	0.9882*	0.9986**
Minor	0.6663	0.3623

Table 3: Pearson correlation coefficients between system-level scores in the original and reproduction studies. ** = significant at $\alpha = 0.01$; * = significant at $\alpha = 0.05$.

Sentence-level scores		
	Comprehension	Adequacy
All	0.695**	0.720**
Major	0.580**	0.656**
Minor	0.403**	0.390**

Table 4: Pearson correlation coefficients between original sentence-level error counts in the original and reproduction studies. (All significant at $\alpha = 0.01$.)

system-level correlation results: while All and Major error counts correlate reasonably well for both error types (although slightly better for Adequacy than for Comprehension), the coefficients for the Minor error types are notably lower.

We will return to some of the above points in the discussion section (Section 5).

4 Inter-annotator Agreement

The original study reported inter-annotator agreement (IAA) in terms of F-score and normalised edit distance (definitions below). In this paper we also report Krippendorff's α for both original and reproduction study, following Kreutzer et al. (2020) who used it in a similar error marking study.[8]

[8]Cohen's kappa was not considered appropriate for either of the studies for the reasons explained in detail in the original paper (Popović, 2020).

Krippendorff's α: In order to quantify agreement by this method, error annotations were converted to a sentence-level quality score, namely the *number* of words marked up for error in a given sentence. For a perfect sentence, no words would be marked so this score would be zero. Using the standard definition,[9] we computed three separate α scores: (i) from just the Major error annotations, (ii) from just the Minor error annotations, and (iii) from both (corresponding to the All subcategory from previous sections).

F-score: To compute sentence-level F1-score, the starting point was the paired sequences *ev1* and *ev2* of word-level error labels (*Major*, *Minor* or *None*) assigned by the two annotators to a sentence. Precision was then computed as the labels from *ev1* also present in *ev2,* and Recall as labels from *ev2* also present in *ev1*. The F1-score was then calculated in the usual way, as the harmonic mean of Precision and Recall. Due to possible length differences in a pair of label sequences (due to insertion of X labels representing missing words), matches are defined as position-independent, which can result in overestimation of agreement.

To yield system-level scores, sentence-level scores are micro-averaged by aggregating matches and lengths.

Edit distance: The standard definition of edit distance[10] with insertions, deletions and substitutions all at cost=1 is applied to paired sequences *ev1* and *ev2* of word-level error labels (as above). Nor-

[9]https://en.wikipedia.org/wiki/Krippendorffs_alpha

[10]Also known as Levenshtein distance (Levenshtein, 1966).

malised edit distance scores are obtained by dividing the summed cost of edits by sequence length. However, while usually (in speech recognition and MT), normalisation is carried out by the length of the 'correct' reference string, here neither of the label sequences is (in)correct. Therefore, the edit-distance metric is symmetrised (in a similar way as the F-score is with Precision and Recall) by first computing edit distance of *ev1* against *ev2*, then of *ev2* against *ev1*, then summing over both and normalising by the sum of the lengths of *ev1* and *ev2*. The resulting measure does penalise differences in label position, thus compensating for the drawback of the position-independent F-score above.

To yield system-level scores, sentence-level scores are micro-averaged by aggregating edit distances and lengths.

Illustration of IAA metrics: Examples of two sentences annotated by two different annotators, along with counts obtained in computing the metrics, are shown in Table 5.

The first two rows show the annotated texts as described in Section 2.3, namely major errors in red/bold, and minor errors in blue/italics. The next two rows below show the extracted error label sequences that form the basis for measuring agreement. The abel sequences were used directly for calculating F-score and edit distance, whereas for Krippendorff's α, label counts were derived instead (rows 5, 6 and 7).

IAA scores computed with the above metrics for the original and reproduction studies are shown in Table 6. IAA is generally good in both studies in terms of all metrics. Furthermore, all metrics indicate a higher IAA for Adequacy than for Comprehensibility (although in some cases differences are very small). Another clear tendency for both studies is a notably lower Krippendorff's α for error annotations in the Minor categories than in the Major and All categories.

For Comprehension errors, IAA is better in the original study according to all metrics. For Adequacy errors, F-score, edit distance and α for Minor errors are also better in the original study, while α for All and Major errors are better in the reproduction study.

5 Discussion

In previous sections, we presented results and similarities/differences observable in them in objective terms. In this section, we discuss and interpret results, aiming to draw conclusions and to identify reasons for similarities and differences.

5.1 Differences in overall scores

As mentioned in Section 3, both studies show broadly similar tendencies in error rates, with some exceptions for Major Comprehension errors and Minor Adequacy errors, as follows. In terms of Major Comprehension error rates from the reproduction study, Amazon is slightly better than Google, while the original study indicates the opposite. As for Minor Adequacy errors, the original study clearly indicates that the Bing translations contain the largest number of errors, which is in line with other scores, too. In the reproduction study, annotators found fewer Minor Adequacy errors in Bing translations than in Amazon translations, however the number of Major Adequacy errors for Bing is much higher in the reproduction study than in the original one. Apparently the two groups of annotators perceived similar overall number of errors but different distributions between Major and Minor ones.

Taking into account the lower inter-annotator agreement for Minor errors, as well as the fact that in both studies the majority of evaluators reported that it was often difficult to distinguish between major and minor errors, the difference in Minor Adequacy errors is not very surprising. As for Major Comprehension errors, lower inter-annotator agreement and larger degree of subjectivity in assessing Comprehensibility may contribute to the slight difference in scores.

Both system-level and sentence-level error rates correlated better across the stwo studies for Major error types than for Minor error types. This also points in the direction of minor errors being generally harder to annotate reliably, something that will need to be addressed in future evaluations.

In terms of the pairwise degree of reproducibility captured by CV, individual pairs of Major error rates differed more between the two studies than individual pairs of Minor error rates. This is not a contradiction with other results: CV measures how far off each individual score is from its original counterpart, whereas Pearson's r measures covariance between *sets* of original and reproduction scores, i.e. how similar their relative ranks and the distances between scores are. In other words, in our results, Major error rates are on average further apart in absolute terms, but evince a more similar

text annotated by ev1	**Ne shvaćajte ih ako udarite u tešku torbu .**
text annotated by ev2	Ne **shvaćajte** ih ako udarite u **tešku torbu** .
error labels, ev1	Major Major Major Major Major Major Major Major Major
error labels, ev2	None Major None None None None Major Major None
major error counts, ev1 ev2	9 3
minor error counts, ev1 ev2	0 0
total error counts, ev1 ev2	9 3
F score (matching labels)	33.3 (3 label matches, total number of labels e1 = 9, e2 = 9)
edit (unmatched labels)	66.0 (6 label mismatches)
text annotated by ev1	*Nadmašio* me na svakom koraku i stalno *me iznenadila priča* .
text annotated by ev2	**Nadmašio me X** na svakom koraku i stalno me *X iznenadila* priča .
error labels, ev1	Minor None None None None None None Minor Minor Minor None
error labels, ev2	Major Major Minor None None None None None None Minor Minor None None
major error counts, ev1 ev2	0 2
minor error counts, ev1 ev2	4 3
total error counts, ev1 ev2	4 5
F score (matching labels)	83.3 (10 label matches. total number of labels e1=11, e2=13)
edit distance (unmatched labels)	25.0 (3 label mismatches)

Table 5: Illustration of IAA metrics: two sentences annotated for comprehension by two evaluators, error labels, error counts used for Krippendorff's α, F-score on labels and edit distance on labels. Bold/red stands for major errors, italics/blue for minor errors, and an X represents an omitted word.

IAA	Comprehension					Adequacy				
	↑ α			↑ F	↓ edit	↑ α			↑ F	↓ edit
	major	minor	all	score	dist.	major	minor	all	score	dist.
original	0.621	0.412	0.687	82.3	22.3	0.679	0.420	0.699	84.1	19.9
reproduction	0.467	0.363	0.636	76.4	30.0	0.734	0.394	0.723	83.0	23.2

Table 6: IAA scores for Comprehensibility and Adequacy: Krippendorff's α, F-score and normalised edit distance.

overall picture in relative terms, than Minor error rates, which does appear to support the conclusion that Minor errors are harder to agree on, and that the dividing line between Major and Minor errors is also hard to agree on.

5.2 Differences in inter-annotator agreement

Some tendencies in IAA are similar in both studies (Section 4). IAA was reasonably good in both studies in terms of all metrics. A contributing factor is likely to be that the annotators were not asked to perform any fine-grained error categorisation. Another clear tendency for both studies was a notably lower Krippendorff's α for minor errors: since these tend to be far less severe (not completely unintelligible, not entirely changing the meaning of the source text), it may be the case that judgments here reflect personal preferences more.

There were also notable differences between the two studies, including IAA being worse in the reproduction study than the original according to 8 out of 10 measures in Table 6. On the face of it, this is not as expected, given the apparently greater homogeneity of the second cohort of evaluators mentioned above. The two cohorts may have other characteristics not accessible to us that would explain the difference.

Moving on to comparing IAA across different error types, the reason for lower Krippendorff's α for Minor errors is probably the generally greater difficulty of agreeing on Minor error annotations mentioned in Section 5.1.

One possible explanation for all metrics indicating a higher IAA for Adequacy than for Comprehensibility is that Adequacy is guided by the original source text while Comprehensibility relies only on the translated text, possibly allowing more space for subjectivity in judgments.

However, to gain a more complete understanding of the above, future work needs to analyse differences in more detail. There are also potential improvements that can be made in the guidelines which could make a difference to IAA measures (for details see the original paper).

5.3 Mark-up Agreement between the Two Studies

To assess the similarity between the annotations produced in the original and reproduction studies, we paired all strings from the original study with all strings from the reproduction study, and then applied the F1 metric as described in Section 4 above, except that this time we used the *word* strings, not the label strings. Table 7 presents an example of a

	original study	reproduction study
annotations	Običnо **ventilator**, ali *neimpresioniran* Običnо **ventilator**, ali *neimpresioniran*	Običnо **ventilator**, ali *neimpresioniran* Običnо **ventilator**, ali **X** neimpresioniran
all errors prec / rec	ventilator neimpresioniran ventilator neimpresioniran prec = 3 matches / 4 words = 75	ventilator neimpresioniran ventilator X rec = 3 matches / 4 words = 75
major errors prec / rec	ventilator ventilator prec = 2 matches / 3 words = 66.7	ventilator ventilator X rec = 2 matches / 2 words = 100
minor errors prec / rec	neimpresioniran neimpresioniran prec = 1 match / 1 word = 100	neimpresioniran rec = 1 match / 2 words = 50

Table 7: Illustration of annotation overlap metric: example text annotated twice in the original study (left), and twice in the reproduction study (right). Red/bold = major error, blue/italics = minor error, X = omitted word.

	Comprehension	Adequacy
All errors	56.3	58.0
Major errors	54.2	56.4
Minor errors	39.3	36.6

Table 8: Overlap between words marked up in the two studies in terms of word-string F1 score.

sentence annotated in the two studies, all marked-up words, and the corresponding word-string Precision and Recall scores. The corresponding F1 values are shown in Table 8 for all error cateogories.

The main tendency is that the overlap for Minor errors is notably lower than for Major and All errors, providing further evidence that Minor errors are harder to agree on. As for Comprehension vs. Adequacy errors, overlap in Minor annotations is worse for Adequacy (than Comprehension), but overlap in Major and All annotations is worse for Comprehension, which aligns with results from Section 3 that the system rankings for Major Comprehension and Minor Adequacy were switched between the two studies.

6 Conclusion

In this paper, we reported results from a reproduction study of an annotation-based human evaluation of MT outputs where errors related to comprehensibility and meaning correctness were annotated in texts by marking up word involved in an error. We compared the corresponding Comprehension and Adequacy system-level error rates for the three MT systems assessed in the two studies, distinguishing subcategories All, Major and Minor for each. We found that 4 out of 6 system rankings were the same in both studies, but that the relative differences between systems are not well replicated for both types of Minor error rates (Pearson's 0.36 for Adequacy-Minor, 0.67 for Comprehension-Minor). However, the *individual* system scores for both types of Minor error rate had a higher degree of reproducibility (as measured by the coefficient of correlation, CV), than the corresponding Major error rates. Results also showed that Minor Adequacy and Major Comprehension annotations and system rankings differed more than other error categories.

The reproduction study reported here was a contribution to the ReproGen Shared Task in the 'Reproduce Your Own' Track, and as such we had the benefit of having full access to all resources and information from the original evaluation, a luxury not normally available when conducting a reproduction study of someone else's work. The main difference between properties of the original study and our reproduction was the characteristics of the cohort of evaluators who had slightly different backgrounds. There were pronounced similarities between the two studies, but also very clear differences, notably including in system rankings. All in all, while repeating the study was simply a matter of recruiting a new cohort of evaluators, obtaining the same results proved somewhat less simple.

Acknowledgements

Both authors benefit from being members of the ADAPT SFI Centre for Digital Media Technology which is funded by Science Foundation Ireland through the SFI Research Centres Programme, and co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106.

The original study (Popović, 2020) was partly funded by the European Association for Machine Translation (EAMT).

The reproduction study was funded by the ADAPT NLG research group.

References

ALPAC. 1966. Language and machines. Computers in translation and linguistics.

Anya Belz. 2021. Quantifying reproducibility in NLP and ML. *arXiv preprint arXiv:2109.01211*.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. Correct Me If You Can: Learning from Error Corrections and Markings. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT 20)*.

Vladimir Iosifovich Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707–710.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015)*, pages 43–52, Santiago, Chile.

Maja Popović. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Johann Roturier and Anthony Bensadoun. 2011. Evaluation of MT Systems to Translate User Generated Content. In *Proceedings of the MT Summit XIII*, Xiamen, China.

John White, Theresa O'Connell, and Francis O'Mara. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the 1994 Conference of Association for Machine Translation in the Americas*, pages 193–205.