The 2023 ReproNLP Shared Task on Reproducibility of Evaluations in NLP: Overview and Results

Anya Belz
ADAPT/DCU, Ireland
and University of Aberdeen, UK
anya.belz@adaptcentre.ie

Craig Thomson
University of Aberdeen
Aberdeen, UK
c.thomson@abdn.ac.uk

Abstract

This paper presents an overview of, and the results from, the 2023 Shared Task on Reproducibility of Evaluations in NLP (ReproNLP'23), following on from two previous shared tasks on reproducibility of evaluations in NLG, ReproGen'21 and ReproGen'22. This shared task series forms part of an ongoing research programme designed to develop theory and practice of reproducibility assessment in NLP and machine learning, all against a background of an interest in reproducibility that continues to grow in the two fields. This paper describes the ReproNLP'23 shared task, summarises results from the reproduction studies submitted, and provides comparative analysis of the results.

1 Introduction

Reproducibility continues to be a topic dividing and troubling the Natural Language Processing (NLP) community (Belz et al., 2021a, 2023a). Despite a growing body of work on the topic, we still do not understand well enough what makes evaluations easier or harder to reproduce, and reproduction studies often reveal alarmingly low degrees of reproducibility not only for human evaluations but also for automatically computed metrics (Belz et al., 2023a).

With this fourth reproduction-focused shared task in NLP, following REPROLANG'20 (Branco et al., 2020), ReproGen'21 (Belz et al., 2021b) and ReproGen'22 (Belz et al., 2022), our aim is to continue to add to the body of reproduction studies in NLP and machine learning (ML) in order to increase the data points available for investigating reproducibility, and to begin to identify properties of evaluations that are associated with better reproducibility.

We start in Section 2 with a description of the organisation and structure of the shared task, followed by details of Track C and the participating teams (Section 3). Next, we present per-experiment results for each experiment in Track C, in terms of the reproduction task, degree of reproducibility assessments, and confirmation of findings (Section 4). We next look at the quality criteria assessed by evaluations and the properties of the ReproNLP evaluation studies in standardised terms as facilitated by HEDS datasheets, and explore if any properties appear to have an effect on degree of reproducibility (Section 5). We conclude with some discussion (Section 6) and a look to future work (Section 7).

2 ReproNLP 2023

ReproNLP 2023[1] consisted of three tracks. Tracks A and B were identical to the tracks in predecessor event ReproGen 2022: Track A a shared task in which teams try to reproduce the same previous evaluation results, Track B an 'unshared task' in which teams attempt to reproduce their own previous evaluation results.

Track C forms part of the ReproHum project[2] and the studies reproduced in it were selected according to criteria of suitability and balance to form part of a larger coordinated multi-lab multi-test reproduction study, as described in detail elsewhere (Belz et al., 2023a). The three tracks in overview were as follows:

A ***Main Reproducibility Track:*** For a shared set of selected evaluation studies, participants repeat one or more studies, and attempt to reproduce the results, using published information plus additional information and resources provided by the authors, and making commonsense assumptions where information is still incomplete.

[1] All information and resources relating to ReproNLP are available at https://repronlp.github.io/.

[2] https://reprohum.github.io/

B ***RYO Track:*** *R*eproduce *Y*our *O*wn previous evaluation results, and report what happened. Unshared task.

C ***ReproHum Track:*** For one or more of the set of papers selected for ReproHum Round 0, and for the specific experiments selected only, repeat one or more studies, and attempt to reproduce the results, using information provided by the ReproNLP organisers only.

There were no submissions for Tracks A and B this year. For the ReproHum Track (C), the specific experiments that are listed and described below were the subject of two reproduction studies each in the ReproHum project, and were also open to ReproNLP'23 participants. The original authors agreed to us using their experiments in the ReproHum project as well as in ReproNLP, and provided very detailed information about the experiments. The experiments, with many thanks to the authors for supporting ReproHum and ReproNLP, are:

1. Vamvas and Sennrich (2022): *As Little as Possible, as Much as Necessary: Detecting Over and Undertranslations with Contrastive Conditioning*. 1 human evalution study (of 2 in paper); English to German; 2 evaluators; 1 quality criteria; 1 system; approx. 1000 outputs; reproduction target: primary scores.
2. Lin et al. (2022): *Other Roles Matter! Enhancing Role-Oriented Dialogue Summarization via Role Interactions*. 1 human evaluation study; Chinese; 3 evaluators; 3 quality criteria; 200 outputs per system; 4 systems; reproduction target: primary scores.
3. Lux and Vu (2022): *Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features*. 1 human evaluation; German; Student evaluators; 1 quality criterion; 12 outputs per system; 2 systems; reproduction target: primary scores.
4. Chakrabarty et al. (2022): *It's not Rocket Science: Interpreting Figurative Language in Narratives*. 2 human evaluation studies (of 4 in paper); English; MTurk; 1 quality criterion; 25 outputs per system, 5/8 systems (including human reference texts); reproduction target: primary scores.
5. Puduppully and Lapata (2021) A: *Data-to-text Generation with Macro Planning*. First human evaluation (relative); English; MTurk; 3 quality criteria; 20 outputs (summaries) per system; 5 systems, reproduction target: primary scores.
6. Puduppully and Lapata (2021) B: *Data-to-text Generation with Macro Planning*. Second human evaluation (absolute); English; MTurk; 2 quality criteria; 80 outputs (sentences) per system; 5 systems; reproduction target: primary scores.

For Track C, the ReproHum project team gathered all code and other resources needed for repeating the study, and acted as a go-between in those cases were there were additional questions from the reproducing teams; this was to avoid using more of the original authors' time than was absolutely necessary. Authors of reproduction papers were also asked to complete a HEDS datasheet.[3] (Shimorina and Belz, 2022).

We issued a call for participation in one or more tracks, and made available broad guidelines[4] to participating teams about how to report reproduction results, and provided light-touch review with comments and feedback on papers. In addition, for Track C, the ReproHum team and partners agreed a common approach to reproduction which ReproHum participants were expected to follow.

3 ReproHum Track (C) in Detail

3.1 Paper Selection

The papers in Track C, or rather the six specific experiments from the five papers in Track C, were selected by a systematic process to achieve balanced and diverse distribution over three properties. The process is described in full detail in a previous paper, coauthored by all participants at the ReproHum partner labs (Belz et al., 2023a).

The three properties and their associated value ranges are shown in Table 1 in the column headings. The cells show property-value counts split across the three most common NLP tasks evaluated and an Other category. The counts are for the larger set of 20 experiments which we deemed to have sufficiently clear properties for reproduction, and from which we selected the subset of six for ReproNLP Track C.

[3] https://forms.gle/MgWiKVu7i5UHeMNQ9
[4] https://repronlp.github.io

Task	Num. Evaluators		Cognitive Complexity			Training and/or Expertise		
	small	not small	low	medium	high	neither	either	both
Dialogue	1	0	0	1	0	0	1	0
Generation	6	5	4	5	2	4	5	2
Summarisation	3	1	2	1	1	1	3	0
Other	2	2	1	0	3	2	0	2

Table 1: Counts of control property values by NLP task for 20 experiments (from 15 papers) with clear properties, from which the ReproNLP Track C experiments were selected to cover as many property combinations as possible.

3.2 Common Approach to Reproduction

In order to ensure comparability between studies, we agreed the following common-ground approach to carrying out reproduction studies:

1. Plan for repeating the original experiment in a form that is as far as possible identical to the original experiment, ensuring you have all required resources in place, then apply to research ethics committee for approval.
2. If participants were paid during the original experiment, determine pay in accordance with the ReproHum common procedure for calculating fair pay (Belz et al., 2023a).
3. Following ethical approval start the reproduction study following the steps below. Contact the ReproHum team with any questions rather than the original authors, as they have already provided us with all the resources and information they have. Don't communicate with other ReproHum teams about their reproduction studies. This is to avoid inadvertently affecting outcomes.
4. Complete HEDS datasheet.
5. Identify the following types of results reported in the original paper for the experiment:
   (a) Type I results: single numerical scores, e.g. mean quality rating, error count, etc.
   (b) Type II results: sets of numerical scores, e.g. set of Type I results .
   (c) Type III results: categorical labels attached to text spans of any length.
   (d) Qualitative conclusions/findings stated explicitly in the original paper.[5]
6. Carry out the allocated experiment exactly as described in the HEDS sheet.
7. Report the results in the following form:
   (a) Description of the original experiment.
   (b) Description of any differences in your repeat experiment.
   (c) Side-by-side presentation of all results (8a-d above) from original and repeat experiments, in tables.
   (d) Report quantified reproducibility assessments as follows:
      i. Type I results: Small-sample oefficient of variation CV* (Belz, 2022).
      ii. Type II results: Pearson's r, Spearman's ρ.
      iii. Type III results: Multi-rater: Fleiss's κ; Multi-rater, multi-label: Krippendorff's α.
      iv. Conclusions/findings: Side-by-side summary of conclusions/findings that are / are not confirmed in the repeat experiment.

[5]We now call these Type IV results.

3.3 Participants and Submissions

Table 2 provides an overview of the NLP labs that participated in Track C, alongside the papers from which they reproduced an experiment.

4 Per-Experiment Results

By design, each of the six experiments in Track C was repeated by two ReproHum partner labs, and in this section we take a look at how results achieved in the two repeat experiments compare to each other and to results from the original experiment, for each of the six experiments.

4.1 Vamvas and Sennrich (2022) *As Little as Possible, as Much as Necessary: Detecting Over and Undertranslations with Contrastive Conditioning*

4.1.1 Reproduction task

The reproduction task for this experiment was to repeat one human evaluation (of two in the paper) of an English-to-German MT post-processing system that checks translations for content additions and omissions as compared to the source text (a

Original paper	Experiment for reproduction						Labs
	#exps	language(s)	#ev-ors	#qc	#sys	#out-s	
Vamvas and Sennrich (2022)	1 (of 2)	En to Ger	2	1	1	1000	(a) **ADAPT/Tech Univ Dublin** (b) **UFAL/Charles Uniersity**
Lin et al. (2022)	1	Chinese	3	3	4	200	(a) **WICT/Peking Uniersity** (b) **Utrecht Uniersity**
Lux and Vu (2022)	1	German	34	1	2	12	(b) **ZHAW (Zurich)** (a) **Darmstadt University**
Chakrabarty et al. (2022)	2 (of 4)	English	MTurk	1	4	25	(a) **Groningen University** (b) **Trivago**
Puduppully and Lapata (2021) A	1	English	MTurk	2	5	20	(a) **Uni Illinois Chicago** (b) **TiCC/Tilburg**
Puduppully and Lapata (2021) B	1	English	MTurk	3	5	80	(a) **Napier University** (b) **Uni Santiago de Compostela**

Table 2: Overview of reproduced papers, experiments, and the 12 labs participating in ReproNLP 2023 (#=number of, ev-ors=evaluators, qc=quality criteria, sys=systems, out-s=outputs.

form of semantic consistency checking). The evaluation involved two evaluators, one quality criterion, one system, and about 1000 system outputs per evaluator.

Each evaluator was shown about 800 system outputs randomly sampled from development and test data, where outputs are word-spans of over/undertranslation errors (aka additions and omissions) detected in translations. The evaluation interface showed source text, translation and the detected error span. The evaluation task was to judge whether the error span marked up by a system was in fact a bad translation, or whether it was ok (there was a second step which was not a reproduction target).

4.1.2 Notable issues

Plátek et al. (2023) (Reproduction 2) used the evaluation tool/interface provided by the original authors as a Docker image, whereas Klubička and Kelleher (2023) (Reproduction 1) who had trouble running it used a Google spreadsheet which made for a very different interface, e.g. without repeated questions.

The script used by the original authors for producing results was found to have a bug in it. Klubička and Kelleher (2023) used only a corrected version of the script provided by the authors, whereas Plátek et al. (2023) corrected the script themselves and produced results with both the buggy and the corrected versions.

4.1.3 Degree of Reproducibility

The table below shows overtranslation (OT) and undertranslation (UT) precision scores. OT precision is the proportion of word spans annotated as an overtranslation (containing incorrectly added content) which were correct. UT precision is the same for undertranslations. The human evaluation was for the proposed system only. The following table shows the word-span-level OT and UT precision scores from the Original human evaluation (which used the script with the bug), Repro 1 (corrected script), and Repro 2 (which used both buggy and non-buggy versions of the script); the last two columns show two three-way CV* scores, one including results obtained with the buggy version of Repro 2, the other with the non-buggy version.

	Orig	Repro 1	Repro 2		CV* (n=3)	
	(+bug)	(-bug)	+bug	-bug	+bug	-bug
OT Prec	0.0742	0.0948	0.0678	0.0691	21.85	20.96
UT Prec	0.3941	0.3529	0.2209	0.2256	34.28	33.12

We can see that the buggy and non-buggy versions of Repro 2 produced very similar precision scores (even though there are notable differences in the raw counts). At the same time, the (buggy) original results are closer to the non-buggy Repro 1 results, all of which makes for a confusing picture.

We do know from the raw counts that the two corrected versions of the script do not produce the same (corrected) counts for the original experiment. This combined with the fact that we do not have buggy results for Orig and Repro 1 *as reported by Repro 1*, means we do not have sufficient comparability to draw conclusion from this pair of reproductions. In the table above, we use the buggy original results as reported by the Repro 2 authors because we do not have raw counts for the original results, whereas the Repro 2 authors calculated them with the script they corrected themselves. Moreover, they report both buggy and non-buggy results for

their reproduction.

All in all, it is hard to interpret the three-way CV* numbers above, given the above obervations, which is why we have greyed them out here, and do not include them in the comparative overview of results in Table 4.

Unlike for the other experiments below, we do not report correlations between score sets as there are only two scores in each set.

4.2 Lin et al. (2022) *Other Roles Matter! Enhancing Role-Oriented Dialogue Summarisation via Role Interactions*

4.2.1 Reproduction task

For this experiment, the task was to repeat one human evaluation of a Chinese role-oriented dialogue summariser. There were three evaluators, three quality criteria (Informativeness, Non-redundancy, and Fluency; an Overall aggregated metric was also reported), four systems (two baseline systems without role interaction, PGN-multi and BERT-multi, and two tested systems, PGN-both and BERT-both), and 200 outputs per system. The dataset was CSDS (Lin et al., 2021), a Chinese customer service dialogue summarisation dataset, from the test set of which 100 dialogues were randomly sampled for the human evaluation. Evaluators were asked to rate each sentence in a summary on a scale from 0 to 2 for each of the three quality criteria.

4.2.2 Notable issues

An interesting aspect of this pair of reproductions is that the original study triple-evaluated the first 10 evaluation items in order to assess IAA, which meant there are three scores for each of these items, compared to one score for the remaining 90. Rather than excluding the first ten items from aggregated results, the original authors decided to use the scores from the 'most experienced' evaluator only, discarding the others.

This was impossible to repeat as the assessment of experience was not explained (experience in terms of what?), and both reproducing teams (Gao et al., 2023; Ito et al., 2023) report results for keeping the first 10 scores of each of the evaluators, as well as for the mean of all three evaluators. The different variants reveal interesting differences in results and system rankings purely as the result of essentially arbitrary preferences for one evaluator over others.

4.2.3 Reproducibility

The following table shows 3-way reproducibility assessments for the original experiment (Lin et al., 2022), Repro 1 (Gao et al., 2023), and Repro 2 (Ito et al., 2023) in terms of CV* values (each computed over the three corresponding scores from the original, Repro 1 and Repro 2 experiments) for each of the four systems and each of the three quality criteria plus the overall aggregated measure (user=user-oriented, agent=agent-oriented, m=multi, b=both):

	CV* (n=3)							
	Inform		Non-Red		Fluency		Overall	
	user	agent	user	agent	user	agent	user	agent
PGN-m	5.89	5.91	5.67	1.28	11.1	15.37	6.01	6.54
PGN-b	5.72	4.61	3.53	0	12.5	12.07	6.58	5.72
BERT-m	2.14	13.29	3.76	5.95	6.74	6.77	1.75	5.72
BERT-b	6.22	13.66	0	2.41	6.93	7.61	3.72	6.98

Non-Redundancy has particularly good reproducibility, in fact the best reproducibility in ReproNLP 2023 of any quality criteria (see Table 4). CV* for for all system/measure combinations ranges from excellent to good for the most part.

4.2.4 Correlations

Table 3 shows Pearson's correlations between the PGN-* and BERT-* systems in (i) the original study compared to reproduction 1, (ii) the original study compared to reproduction 2, and (iii) reproduction 1 compared to reproduction 2, for each of the two modes user-oriented and agent-oriented. Correlations are > 0.9 for the user-oriented mode for all three criteria, for the agent-oriented mode for Informativeness, and (just) between Orig and Repro 2 for Fluency/agent-oriented.

Repro 1 has strikingly strong *negative* correlations for Fluency/agent-oriented mode, as well as weak to moderate correlations for Non-redundancy/agent-oriented. It is unclear why, but Repro 1 and agency-oriented mode are both associated with lower correlations. Finally, the Overall scores correlate less well with each other, especially when Repro 1 is involved.

4.2.5 Confirmation of findings

If we take the main findings to be the relative performance of the methods evaluated, and the reported ranks for the methods as the means of verification, then the following picture emerges. Ito et al. (2023) are unable to confirm the overall finding that the proposed approach really does improve the Fluency and Non-redundancy of summaries, while Gao et al. (2023) confirm the effectiveness of the proposed

	Informativeness		Non-Redundancy		Fluency		Overall	
	user-orient.	agent-orient.	user-orient.	agent-orient.	user-orient.	agent-orient.	user-orient.	agent-orient.
(i) Pearson's Orig v Repro 1								
PGN-*, BERT-*	0.943	1	0.948	0.486	0.908	-0.728	0.105	0.328
(ii) Pearson's Orig v Repro 2								
PGN-*, BERT-*	0.927	0.986	0.932	0.883	0.933	0.995	0.753	0.683
(iii) Pearson's Repro 1 v Repro 2								
PGN-*, BERT-*	0.984	0.984	0.999	0.263	0.96	-0.765	0.466	0.801

Table 3: Pearson's correlations between original study (Lin et al., 2022), reproduction 1 (Gao et al., 2023), and reproduction 2 (Ito et al., 2023), n=4, for each of the four quality criteria, for each of the two modes user-oriented and agent-oriented.

approach in terms of the Overall metric, but document slightly worse performance of the proposed method compared to the standard approach.

4.3 Lux and Vu (2022) *Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features*

4.3.1 Reproduction task

The experiment that was the reproduction target from this paper was a human evaluation of a German text-to-speech (TTS) system. Evaluators were students, 34 responses were collected, one quality criterion (Naturalness) was assessed, using 6 audio outputs each for four system variants: the proposed approach and a baseline each combined with two different TTS systems (Tacotron and FastSpeech). The primary score for each system was the percentage of times that the system was preferred (counts of no preference were also collected).

4.3.2 Notable issues

One issue with this pair of reproductions was that the original authors had reported and confirmed that the order of audio files[6] had been randomised in the original experiment on Google Forms. However, at the time of reproduction there was no option to randomise the order of Google Form questions while at the same time preserving the connection between audios and evaluation response. We provided both reproducing teams with the same random order of items. Each participant in both reproductions was shown items in this order.

Another interesting issue arose: while both reproducing teams (Hürlimann and Cieliebak, 2023; Mieskes and Benz, 2023) found very low reproducibility in terms of CV* and Pearson's r, Hürlimann and Cieliebak (2023) found much better reproduciblity when the system labels were swapped (i.e. when treating Tacotron as FastSpeech and vice versa). However, even if such an accidental transposition is assumed, the preference percentages reported by Mieskes and Benz (2023) in their reproduction study still do not confirm the original results, as we will see below.

4.3.3 3-way degree of reproducibility

The following table shows percentages of times that each baseline and proposed system version (*-base, *-prop) was preferred and where there was no preference (*-equal),[7] alongside three-way CV* values for scores from the three experiments (Original, Reproduction 1 (Hürlimann and Cieliebak, 2023), Reproduction 2 (Mieskes and Benz, 2023)):

Preferred system	Preference strength (% preferred			CV* (n=3)
	Orig	Repro 1	Repro 2	
FS-base	31.3	12.0	13.1	70.48
FS-prop	25.3	50.0	40.5	39.46
FS-equal	43.4	38.0	46.4	12.21
Taco-base	11.0	29.3	22.5	54.02
Taco-prop	52.0	29.3	25.7	48.87
Taco-equal	37.0	41.4	51.8	21.41

From this we can see that there is very little agreement (CV* is very high) among the three experiments, except for the *-equal percentages; Pearson's r values (Section 4.3.5) also confirm this. If instead we switch FS and Taco scores around in the two repeat evaluations (as indicated by the shading in the table) we get substantially improved reproducibility, again except for the *-equal percentages which remain similar:

[6] To be precise, the audio files were converted to audio-only video files.

[7] Note that the numbers in the three tables in this section may differ very slightly from those reported by Mieskes and Benz (2023) and (Hürlimann and Cieliebak, 2023), because we normalised percentages to add up to 100 excluding any skipped items.

Preferred system	Preference strength (% preferred			CV* (n=3)
	Orig	Repro 1 T	Repro 2 T	
FS-base	31.3	29.3	22.5	20.36
FS-prop	25.3	29.3	25.7	10.06
FS-equal	43.4	41.4	51.8	14.82
Taco-base	11.0	12.0	13.1	10.67
Taco-prop	52.0	50.0	40.5	15.81
Taco-equal	37.0	38.0	46.4	15.6

4.3.4 2-way degree of reproducibility

If we look at pairwise CV* (n=2) we can see that after transposition, Repro 1 T matches the original experiment much more closesly than Repro 2 T:

Preferred system	CV* of each Repro* with Orig (n=2)			
	Repro 1	Repro 2	Repro 1 T	Repro 2 T
FS-base	88.88	81.74	6.58	32.62
FS-prop	65.41	46.06	14.61	1.56
FS-equal	13.23	6.66	4.7	17.59
Taco-base	90.55	68.45	8.67	17.38
Taco-prop	55.68	67.49	3.91	24.79
Taco-equal	11.19	33.23	2.66	22.47

While it seems likely that some mixup has happened in the audio files that makes the transposed results match the original experiment better than the non-transposed results, we don't know exactly what has happened, and in fact we don't know for sure which scores belong to which system.

Something that might go some way towards explaining what has caused Repro 1 (T) to be a better match for the original scores is that in Repro 1, 157 evaluators were used, whereas the original used 34 and Repro 2 used 37, as more evaluators means better reliability (better representativeness of the sample relative to the population).

4.3.5 Correlations between score sets

The pairwise r coefficients (between the combined FastSpeech and Tacotron scores) below confirm that Repro 1 T tracks the original percentages more closely than Repro 2 T:

	O v R1	O v R2	R1 v R2	O v R1T	O v R2T	R1T v R2T
Pearson's	0.001	0.259	0.845	0.989	0.83	0.845

While there is no correlation at all between Orig and Repro 1, there is a mild positive correlation between Orig and Repro 2. Orig vs. the transposed Repro 1 (R1T) results is very strongly correlated (0.99), while Orig vs. R2T, and R1T vs. R2T are not much less strong. (We include the identical r for both Repro 1 vs. Repro 2 and Repro 1 T vs. Repro 2 T for ease of reference.)

4.3.6 Confirmation of findings

In terms of findings (Type IV results), on the basis of the non-transposed results, both reproducing teams are unable to confirm the original findings. On the basis of transposed results, Hürlimann and Cieliebak (2023) obtain the same system ranks in all cases (albeit in one case with a very small margin), showing Taco-prop > Taco-base, *but* FS-prop < FS-base (second table above). However, in Repro 2 T (created for this paper above) the proposed approach is found to be better in both FastSpeech and Tacotron.

4.4 Chakrabarty et al. (2022) *It's not Rocket Science: Interpreting Figurative Language in Narratives*

4.4.1 Reproduction task

The task here was to repeat two human evaluation studies (of four in the paper) of an English prompted text generator. The evaluation was carried out on MTurk, there was one quality criterion (Plausibility) evaluated in absolute mode, 25 outputs per system, and four systems addressing two tasks, namely continuation after idiom, and continuation after simile.

25 narratives ending in either an idiom or a simile were randomly sampled for each task. Each narrative was paired with (a) human-written continuations (5 for the similes, 3 for the idioms), and (b) automatically generated continuations, one by the baseline GPT2-XL model, one by a context-enhanced model, and one by a 'literal-enhanced' model. Each continuation was categorised as either plausible or not by evaluators.

4.4.2 3-way degree of reproducibility

The table below is a three-way comparison of percentages of plausible continuations for each of the four systems, separately for continuations after Idioms, and after Similes, obtained in the three experiments (Repro 1 is by Li et al. (2023), Repro 2 by Mahamood (2023)).[8] Three-way CV* values for the three experiments are shown in the last column:

Type	Model	% of plausible continuations			CV* (n=3)
		Orig	Repro 1	Repro 2	
Idioms	GPT2-XL	56	76	58	21.26
	+Context	68	92	83.33	18.32
	+Literal	48	68	66.66	22.45
	Human	80	68	80.55	11.38
Similes	GPT2-XL	60	68	64	7.64
	+Context	68	72	48	25.08
	+Literal	**76**	80	64	13.88
	Human	88	68	84	16.17

All CV* values are medium good, with GPT2-XL/Similes better on average.

[8] The number in red/bold was recalculated by Li et al. (2023) as 60; the original paper reports 76.

4.4.3 2-way degree of reproducibility

The 3-way CV* scores showed a medium degree of reproducibility, and a first indication that Repro 2 tracks the Orig scores more closely than Repro 1. This is supported by the pairwise CV* scores, except for +Context/Similes where Repro 1 is closer:

Type	Model	CV* of each Repro* with Orig (n=2)	
		Repro 1	Repro 2
Idioms	GPT2-XL	30.21	3.5
	+Context	29.91	20.2
	+Literal	34.38	32.45
	Human	16.17	0.68
Similes	GPT2-XL	12.46	6.43
	+Context	5.7	34.38
	+Literal	28.49	5.11
	Human	25.56	4.64

4.4.4 Correlations between score sets

The pairwise Pearson's r values show clearly that Repro 2 tracks the Orig scores much more closely than Repro 1, with which Orig has no correlation for idioms, and a medium *negative* correlation for Similes (note that none of the r values reach significance at $\alpha = 0.05$):

	Idioms			Similes		
	Orig	Repro 1	Repro 2	Orig	Repro 1	Repro 2
Orig	1	0.13	0.76	1	-0.5	0.68
Repro 1	0.13	1	0.38	-0.5	1	-0.32
Repro 2	0.76	0.38	1	0.68	-0.32	1

4.4.5 Confirmation of findings

In terms of main findings (Type IV results), the following picture emerges. The ranks determined by Orig, Repro 1 and Repro 2 are all different, for both Idioms and Similes. Repro 2 achieves closer similarity of ranks with Orig. Repro 1 has completely different ranks from Orig for Idioms and Similes.

4.5 Puduppully and Lapata (2021) A: *Data-to-text Generation with Macro Planning*

4.5.1 Reproduction task

In this experiment, five data-to-text methods (3 neural systems, one template, and human (gold) reference texts) were evaluated by relative human evaluations involving three quality criteria (Grammaticality, Coherence, and Conciseness), and 20 items from the Rotowire dataset (Wiseman et al., 2017). Pairs of systems were compared, with 10 combinations per input record, for a total of 200 evaluation items.

Each evaluation item was shown to 3 distinct workers on Amazon Mechanical Turk; there was no limit in the number of items a worker could complete. Evaluators were asked to select the best summary within the pair. Best-worst scaling was then applied (Louviere et al., 2015) to provide per-system scores ranging from -100 to 100.

4.5.2 Notable Issues

The authors of the original study performed attention checks whereby participants, if they failed, were excluded from future tasks (but the work they had done so far was retained). No process for these checks, or details of which output pairs were involved in a check were recorded. Following discussion with the original author, we created a method for systematic attention checks that was then used in both reproductions.

4.5.3 3-way degree of reproducibility

The table below shows the best-worst scores and CV* for the Grammaticality criterion:[9]

best-worst score (Grammaticality)				CV* (n=3)
System	Orig	Repro 1	Repro 2	
Gold	38.33	14.17	9.17	15.81
Templ	-61.67*	-23.33*	17.08*	62.23
ED+CC	5.00	-8.33	-19.58	16.28
RBF	13.33	9.17	-9.58	14.30
Macro	5.00	8.33	2.92	3.16

From this we can see that whilst CV* was low (good) for the Macro system, and moderate for others, the Templ (template) system score varied greatly between experiments and has a very high (bad) CV* value. In fact, the Templ system came out worst overall for the original experiment and Repro 1, yet best overall for the other Repro 2.

The next table shows results for Coherence, in the same format:

best-worst score (Coherence)				CV* (n=3)
System	Orig	Repro 1	Repro 2	
Gold	46.25*	12.50	-0.42	24.66
Templ	-52.92*	-20.00*	25.42	57.13
ED+CC	-8.33	-7.50	-15.00	5.60
RBF	4.58	9.17	-10.42	12.39
Macro	10.42	5.83	0.42	5.80

The same issue with the template system is observed, with CV* for other systems being low to moderate. Finally, the same is also seen for Conciseness:

[9] Note that because the measure used for assessing it ranges $-100..+100$, CV can't be applied directly. We have therefore shifted scores to the range $0..200$, which is acceptable here as we have an interval (with fixed endpoints).

best-worst score (Conciseness)				CV* (n=3)
System	Orig	Repro 1	Repro 2	
Gold	30.83	5.83	-1.67	18.63
Templ	-36.67	-5.83	43.75*	49.39
ED+CC	-4.58	-5.00	-25.83	16.84
RBF	3.75	0.83	-14.58	12.45
Macro	6.67	4.17	-1.67	5.08

In all above tables, the asterisk indicates that the system was significantly different from the Macro system.

4.5.4 Correlations between score sets

Spearman's rank correlation (ρ) for each study pair looks as follows for the three quality criteria, with the caveat that the sample size is small:

Grammaticality	Orig	Repro 1	Repro 2
Orig	1	0.975	-0.205
Repro 1	0.975	1	-0.100
Repro 2	-0.205	-0.100	1

Coherence	Orig	Repro 1	Repro 2
Orig	1	0.900	-0.100
Repro 1	0.900	1	-0.300
Repro 2	-0.100	-0.300	1

Conciseness	Orig	Repro 1	Repro 2
Orig	1	1	-0.051
Repro 1	1	1	-0.051
Repro 2	-0.051	-0.051	1

As expected, this shows near perfect alignment of system ranks between Orig and Repro 1, but no correlation at all between Repro 2 and either of the other two.[10]

4.5.5 Confirmation of findings

We saw in the preceding section that the original study and Repro 1 have close rank correlations. This was also reported by the Repro 1 authors (Arvan and Parde, 2023) who reported an overall ρ of 0.83 when concatenating scores for the three criteria.

In terms of statistical significance, no study (original or reproduction) found any difference, for any criteria, between the proposed (Macro) system and either of the other neural systems. Some differences were seen between Macro and either the human reference or the template, but whether these differences were significant varied greatly between experiments. Like van Miltenburg et al. (2023), we are unable to explain why there are such fundamental differences between their reproduction on the one hand, and Orig and Repro 1 on the other, e.g. why the template system is judged best for all criteria in their reproduction whilst being worst in the other studies. This difference has a large impact on both CV* and Spearman's ρ.

[10]This is so striking a finding that we will investigate it further in future work, something that wasn't possible in the short time we had to write this report.

4.6 Puduppully and Lapata (2021) B: *Data-to-text Generation with Macro Planning*

4.6.1 Reproduction task

In this experiment, an absolute human evaluation of the same data-to-text system as in the last section was performed to obtain the mean number of facts in the output text that are (i) supported by the input (#Supp) and (ii) contradicted by the input (#Cont). For this, 20 input records from the Rotowire dataset and corresponding verbalisations (summaries) generated by the same five systems as in Section 4.5 were selected. From each summary, 4 sentences were selected as evaluation items, for a total of 400 evaluation items. Reproduction 1 was carried out by Watson and Gkatzia (2023), Reproduction 2 by González-Corbelle et al. (2023).

Experiments were carried out on Amazon Mechanical Turk, participants were shown the four sentences from a given summary on a form and asked to provide counts for both #Supp and #Cont on the same form. Three participants scored each sentence. Other than the above, there was no restriction on the total number of tasks each participant could undertake.

4.6.2 3-way degree of reproducibility

The following table shows the mean #Supp counts for the original experiment and the two reproductions, alongside three-way CV* values:

System	Orig	Repro 1	Repro 2	CV* (n=3)
Gold	3.63	4.000	3.36	10.72
Templ	7.57*	6.3167*	6.27*	13.42
ED+CC	3.92	5.100	4.42	16.16
RBF	5.08*	4.9458	4.31	10.52
Macro	4.00	4.5458	4.08	8.56

For all systems, CV* is moderate, indicating some consistency between the three studies. The below table shows the same for #Cont counts:

System	Orig	Repro 1	Repro 2	CV* (n=3)
Gold	0.07	1.525	0.66	119.01
Templ	0.08	1.3583	0.90	101.57
ED+CC	0.91*	1.9042	1.95*	45.24
RBF	0.67*	1.7583	1.22	54.70
Macro	0.27	1.5333	0.55	103.39

In both the above tables, the asterisk indicates that the system was significantly different from the Macro system at $\alpha = 0.05$.

For #Cont counts, we see *much* higher (worse) values for CV* for all systems. Since the experi-

ment design only has participants provide a count for supported or contradicted facts, rather than annotating error spans in the text, it is not easy to determine whether there are differences between facts annotated as Supp and as Cont that might explain this very substantial difference.

However, we do know that there were far more Supp facts than there were Cont facts found (roughly 20–30 times as many), which would make the former far more stable than the latter.

This may be compounded by the fact that facts are overwhelmingly numeric in nature in this dataset, and it is particularly difficult to achieve acceptable agreement among evaluators regarding what counts as a numeric fact (Thomson et al., 2023). When annotating individual errors in system outputs for the same dataset, Thomson et al. noted that participants had to be specifically instructed as to what should be classed as a number, since ordinals, cardinals, determiners, and number-based phrases would otherwise be considered numeric by some annotators but not others.

4.6.3 Correlations between score sets

Shown below are the Pearson correlations between the studies for both the count of supporting facts (#Supp) and the count of contradicted facts (#Cont):

#Supp	Orig	Repro 1	Repro 2
Orig	1.000	0.912	0.942
Repro 1	0.912	1.000	0.989
Repro 2	0.942	0.989	1.000

#Cont	Orig	Repro 1	Repro 2
Orig	1.000	0.958	0.887
Repro 1	0.958	1.000	0.826
Repro 2	0.887	0.826	1.000

This shows strong correlations between all experiments, obscuring the fact that the raw counts in the reproduction studies being, in many cases, an order of magnitude higher than in the original study. Repro 2 has lower correlation with both Orig and Repro 1.

4.6.4 Confirmation of findings

The original study found there to significantly more supported facts (#Supp) in the template system compared with the proposed (Macro) system. Both reproduction studies confirm this. It also found significantly more supported facts in the RBF system compared to Macro, although this was not confirmed by either reproduction. For contradicted facts (#Cont), the original study showed the Macro system to have significantly fewer than the two other neural systems (ED+CC and RBF). Reproduction 1 found no significant differences, and Reproduction 2 confirmed Macro to have significantly fewer than ED+CC only.

5 Results by Quality Criterion

Table 4 provides an overview of the six ReproNLP experiments in terms of the quality criteria (measurands) assessed in the evaluations and the properties of the evaluation design (Shimorina and Belz, 2022). The first column identifies the studies and criteria, the last column shows the corresponding mean criterion-level CV*. The remaining columns show seven properties of each study/criterion, as per the HEDS datasheets; column headings identify HEDS question number (for brief explanation of each see table caption). Note that for property 3.2.1 (number of evaluators) we don't always have the information for both reproductions.

Note we are not including CV* for (Vamvas and Sennrich, 2022) because of the issues noted above. The experiment originally reported by Lin et al. (2022), and reproduced by Gao et al. (2023) and Ito et al. (2023), stands out for having good reproducibility for all three criteria assessed (all below 10), Non-redundancy having particularly low CV* (2.83). If we assume transposition of system outputs has indeed accidentally occurred, then the Naturalness evaluation from Lux and Vu (2022) is only slightly worse (14.55).

The evaluation from Chakrabarty et al. (2022) has the next best degrees of reproducibility, mean CV* for Plausibility after Idiom and Plausibility after Simile both being medium (in the 15-20 range). The assessments of Grammaticality, Coherence and Conciseness for the experiment from Puduppully and Lapata (2021) (A) have slightly worse reproducibility at just above 20 for all three criteria.

Finally, the second experiment from Puduppully and Lapata (2021) (B) has good reproducibility for the mean number of facts supported by the input (#Supp), but the worst reproducibility by far for the mean number of facts contradicted by the input (#Cont).

For comparison, in the ReproGen'22 studies, annotation-based evaluation (4.3.8=Anno) was clearly associated with lower reproducibility. Evaluations which involved assessment of content alone (4.1.2=Cont) also tended to have worse reproducibility. Assessing evaluation items relative to a system input (4.1.3=RtI) was also associated with

ReproNLP 2023									
Orig Study // *Repro 1* / *Repro 2*, **measurands**	**3.1.1**	**3.2.1**	**4.3.4**	**4.3.8**	**4.1.1**	**4.1.2**	**4.1.3**	**scores /item**	**mean CV***
Vamvas and Sennrich (2022) // *Klubička and Kelleher (2023)* / *Plátek et al. (2023)*									
Correctly Identified Omissions	~1000	2	Yes,No	Cl/Lab	Corr	Both	RtI	1-2	N/A
Correctly Identified Additions	~1000	2	Yes,No	Cl/Lab	Corr	Both	RtI	1-2	N/A
Lin et al. (2022) // *Gao et al. (2023)* / *Ito et al. (2023)*									
Informativeness	100	3	0,1,2	DQE	Feature	Cont	iiOR	1	7.18
Non-Redundancy	100	3	0,1,2	DQE	Good	Cont	iiOR	1	2.83
Fluency	100	3	0,1,2	DQE	Good	Form	iiOR	1	9.89
Lux and Vu (2022) // *Hürlimann and Cieliebak (2023)* / *Mieskes and Benz (2023)*									
Naturalness (speech)	12	34/157/37	A,B,Tie	RQE	Good	Form	iiOR	34/157/37	41.08
Naturalness (speech) transposed	12	34/157/37	A,B,Tie	RQE	Good	Form	iiOR	34/157/37	14.55
Chakrabarty et al. (2022) // *Li et al. (2023)* / *Mahamood (2023)*									
Plausibility (continuation idiom)	150	4/?/35	Yes,No	Cl/Lab	Good	Both	RtI	3	18.35
Plausibility (continuation simile)	200	7/?/45	Yes,No	Cl/Lab	Good	Both	RtI	3	15.69
Puduppully and Lapata (2021) A // *Arvan and Parde (2023)* / *van Miltenburg et al. (2023)*									
Grammaticality	200	206/262/?	A,B	RQE	Corr	Form	iiOR	3	22.36
Coherence	200	206/262/?	A,B	RQE	Good	Cont	iiOR	3	21.12
Conciseness	200	206/262/?	A,B	RQE	Good	Both	iiOR	3	20.48
Puduppully and Lapata (2021) B // *Watson and Gkatzia (2023)* / *González-Corbelle et al. (2023)*									
Mean # Supported Facts	400	131/167/144	0-20	Count	Corr	Content	RtI	3	11.88
Mean # Contradicted Facts	400	131/167/144	0-20	Count	Corr	Content	RtI	3	84.78

Table 4: Summary of some properties of ReproNLP experiments, alongside mean CV* (n=3). 3.1.1 = number of items assessed per system; 3.2.1 = number of evaluators in original/reproduction experiment; 4.3.4 = List/range of possible responses; 4.3.8 = Form of response elicitation (DQE: direct quality estimation, RQE: relative quality estimation, Cl/Lab: classification/labelling, Count: counting occurrences in text); 4.1.1 = Correctness/Goodness/Features; 4.1.2 = Form/Content/Both; 4.1.3 = each output assessed in its own right (iiOR) / relative to inputs (RtI) / relative to external reference (EFoR); scores/item = number of evaluators who evaluate each evaluation item.

lower reproducibility for three of the studies (where comparison of outputs to inputs was far more complex than a straightforward is-it-simpler decision as in e.g. (Nisioi et al., 2017). Finally, correctness assessment (4.1.1=Corr) was also associated with lower reproducibility. For those of these properties that were present in ReproGen'21, the tendencies were the same.

6 Discussion

In terms of general tendencies found in ReproNLP reproductions, there were quite a few issues (see Notable Issues sections above) that made carrying out a repeat experiment difficult. These were discussed in detail in a previous paper (Belz et al., 2023a).

In some cases, there were striking differences between the two paired reproduction studies: for example, Repro 2 for Chakrabarty et al. (2022) achieved much closer results to the original study than Repro 1 in terms of both pairwise CV* and Pearson's, and while Repro 1 for (Puduppully and Lapata, 2021) (A) achieved very similar results to the original study, Repro 2 results had very little in common with either the original study or Repro 1. This very clearly highlights the importance of carrying out more than one reproduction study to get a rounded picture of an evaluation's degree of reproducibility.

None of the reproductions produced the same system ranks for all quality criteria evaluated, although in some cases it was close. Given that sys-

tem ranks are the single most important result from the above types of evaluations, this is concerning.

In terms of patterns emerging about what properties make an evaluation more or less reproducible, we can glean two tendencies from the properties examined in Table 4: (i) there is some indication that Goodness-type criteria[11] are associated with better degree of reproducibility than Correctness-type criteria (see column 4.1.1 in Table 4); and (ii) sets of experiments that use the same number of evaluators (see column 3.2.1 in Table 4) tend to have better reproducibility than those that have different numbers.

7 Conclusions

Our intention in Track C had been to create a situation where we would have more than one reproduction of the same original study to analyse, in order to obtain truer estimates of the original study's reproducibility. Moreover, all three studies were supposed to be identical for as close as possible to ideal comparability. Two main problems arose: (a) the flaws, errors and bugs reported previously (Belz et al., 2023a,b) were in some cases fixed differently by reproducing authors, leading to different raw results; (b) reproducing authors in some cases chose different results to reproduce and compare, resulting in non-comparability; and (c) reproducing authors did not always manage to stick as closely as we had intended to the original experimental details, e.g. using different interfaces, revealing that the experiment was a reproduction, and most significantly, using very different numbers of evaluators. The latter is particularly significant, because it appears to be associated with worse reproducibility (see preceding section).

Our next step will be to fully standardise analysis and other scripts, and ask reproducing authors to both provide the same fully standardised set of results (something we did not have time for within the ReproNLP schedule). This will then provide the basis for more detailed analysis to be carried out and reported in future work.

We will also run another round of paired reproductions in the ReproHum project, using a different set of experiments for which we have corrected any issues prior to sharing them with the reproducing partners and where we are relaxing the strict-repetition requirement somewhat. We will again open up reproductions to any additional reproducing teams in ReproNLP 2024.

[11] From HEDS (Shimorina and Belz, 2022): "Goodness: select this option if, in contrast to correctness criteria, there is no single, general mechanism for deciding when outputs are maximally good, only for deciding for two outputs which is better and which is worse. E.g. for Fluency, even if outputs contain no disfluencies, there may be other ways in which any given output could be more fluent."

Acknowledgments

We thank the authors of the original papers that were up for reproduction in ReproNLP 2023. And of course the authors of the reproduction papers, without whom there would be no ReproHum project and no ReproNLP shared task.

Our work was carried out as part of the ReproHum project on Investigating Reproducibility of Human Evaluations in Natural Language Processing, funded by EPSRC (UK) under grant number EP/V05645X/1. In particular, we thank our numerous collaborators from NLP labs across the world who carried out the reproductions in Track C as part of the first batch of coordinated reproductions in the ReproHum project.

The ReproNLP work also benefits from the work being carried out in association with the ADAPT SFI Centre for Digital Media Technology which 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.

References

Mohammad Arvan and Natalie Parde. 2023. Human evaluation reproduction report for data-to-text generation with macro planning. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*.

Anya Belz. 2022. A metrological perspective on reproducibility in NLP. *Computational Linguistics*, 48(4):1125–1135.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021a. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021b. The reprogen shared task on reproducibility of human evaluations in nlg: Overview and results. *INLG 2021*, page 249.

Anya Belz, Anastasia Shimorina, Maja Popovic, and Ehud Reiter. 2022. The 2022 reprogen shared task

on reproducibility of evaluations in nlg: Overview and results. *INLG 2022*, page 43.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M Alonso-Moral, Mohammad Arvan, Jackie Cheung, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubička, Huiyuan Lai, Chris van der Lee, Emiel van Miltenburg, Yiru Li, Saad Mahamood, Margot Mieskes, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Pablo Mosteiro Romero, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023a. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023b. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with REPROLANG2020. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5539–5545, Marseille, France. European Language Resources Association.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Mingqi Gao, Jie Ruan, and Xiaojun Wan. 2023. A reproduction study of the human evaluation of role-oriented dialogue summarization models. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*.

Javier González-Corbelle, Jose M. Alonso-Moral, and A. Bugarín-Diz. 2023. Some lessons learned reproducing human evaluation of a data-to-text system. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*.

Manuela Hürlimann and Mark Cieliebak. 2023. Reproducing a comparative evaluation of german text-to-speech systems. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*.

Takumi Ito, Qixiang Fang, Pablo Mosteiro, Albert Gatt, and Kees van Deemter. 2023. Challenges in reproducing human evaluation results for role-oriented dialogue summarization. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*.

Filip Klubička and John D. Kelleher. 2023. Humeval'23 reproduction report for paper 0040: Human evaluation of automatically detected over- and undertranslations. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*.

Yiru Li, Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Same trends, different answers: Insights from a replication study of human plausibility judgments on narrative continuations. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*.

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. Csds: A fine-grained chinese dataset for customer service dialogue summarization. *arXiv preprint arXiv:2108.13139*.

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other roles matter! enhancing role-oriented dialogue summarization via role interactions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2558, Dublin, Ireland. Association for Computational Linguistics.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Florian Lux and Thang Vu. 2022. Language-agnostic meta-learning for low-resource text-to-speech with articulatory features. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6858–6868, Dublin, Ireland. Association for Computational Linguistics.

Saad Mahamood. 2023. Reproduction of human evaluations in: 'it's not rocket science: Interpreting figurative language in narratives'. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*.

Margot Mieskes and Jacob Georg Benz. 2023. h da@reprohum – reproduction of human evaluation and technical pipeline. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*.

Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Martijn Goudbeek Debby Damen, Chris van der Lee, Frédéric Tomas, and Emiel Krahmer. 2023. How reproducible is best-worst scaling for human evaluation? a reproduction of 'data-to-text generation with macro planning. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*.

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 (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Ondřej Plátek, Mateusz Lango, and Ondřej Dušek. 2023. With a little help from the authors: Reproducing human evaluation of an mt error detector. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Computer Speech & Language*, 80.

Jannis Vamvas and Rico Sennrich. 2022. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 490–500, Dublin, Ireland. Association for Computational Linguistics.

Lewis Watson and Dimitra Gkatzia. 2023. Unveiling nlg human-evaluation reproducibility: Lessons learned and key insights from participating in the ReproNLP challenge. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.