Another PASS: A Reproduction Study of the Human Evaluation of a Football Report Generation System

Simon Mille
Universitat Pompeu Fabra
Barcelona, Spain
simon.mille@upf.edu

Thiago Castro Ferreira
Federal University of Minas Gerais
Belo Horizonte, Brazil
thiagocf05@ufmg.br

Anya Belz and **Brian Davis**
ADAPT Research Centre
Dublin City University
Dublin 9, Ireland
{anya.belz,brian.davis}@adaptcentre.ie

Abstract

This paper reports results from a reproduction study in which we repeated the human evaluation of the PASS Dutch-language football report generation system (van der Lee et al., 2017). The work was carried out as part of the ReproGen Shared Task on Reproducibility of Human Evaluations in NLG, in Track A (Paper 1). We aimed to repeat the original study exactly, with the main difference that a different set of evaluators was used. We describe the study design, present the results from the original and the reproduction study, and then compare and analyse the differences between the two sets of results. For the two 'headline' results of average Fluency and Clarity, we find that in both studies, the system was rated more highly for Clarity than for Fluency, and Clarity had higher standard deviation. Clarity and Fluency ratings were higher, and their standard deviations lower, in the reproduction study than in the original study by substantial margins. Clarity had a higher degree of reproducibility than Fluency, as measured by the coefficient of variation. Data and code are publicly available.[1]

1 Introduction

Recent years have seen growing interest in, and concern about, reproducibility across the Natural Language Processing (NLP) field. The ReproGen Shared Task on Reproducibility of Human Evaluations in Natural Language Generation (Belz et al., 2020a) was the first shared task to focus on reproducibility of human evaluations (rather than metrics). We report on our participation in ReproGen, where our contribution was in Track A, the Main Reproducibility Track. More specifically, we repeated the human evaluation study reported by van der Lee et al. (2017). In this paper, we describe how we approached this task, present the results obtained, and compare our results with those reported in the original paper, using different methods of analysis.

2 Summary of the Evaluated System

PASS (Personalized Automated Soccer texts System) is a modular data-to-text system that produces Dutch summaries of football matches and is a partial re-implementation of the GoalGetter system (Theune et al., 2001). Like GoalGetter, PASS is a template and rule-based system. Unlike GoalGetter, PASS (i) tailors the tone of football reports for supporters of one of the clubs in a match, (ii) has a modular architecture, and (iii) uses templates informed by the MEmo FC (Multilingual Emotional Football Corpus) corpus (Braun et al., 2016).

Data and Language Sources: Automatically scraped football match data from Goal.com,[2] subsequently stored in XML-format, is used as input data, and the MEmo FC corpus as reference data.

System Architecture: The PASS[3] architecture is a data-to-text pipeline consisting of the following modules: (1) the ***governing module*** (used in slightly different versions for different report parts) processes topics one by one, and interacts with the other modules as necessary; (2) the ***topic collection module*** extracts topics from the match data and orders them; (3) the ***lookup module*** retrieves all matching template categories for a given match event and their corresponding templates from a

[1] https://github.com/ThiagoCF05/ReproGen2021-vanderLee

[2] https://www.goal.com/

[3] https://github.com/TallChris91/PASS

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 286–292,
Aberdeen, Scotland, UK, 20-24 September 2021.

database; (4) the ***between-text variety module*** removes templates that were used in the last match report to ensure variety; (5) the ***ruleset module*** checks whether constraints associated with a given template category are met; (6) the ***template selection module*** selects templates from the remaining categories in a weighted random fashion; (7) the ***template filler module*** fills empty template slots with the relevant information from the match data; (8) the ***text collection module*** combines the text produced for the different report parts in the right order; (9) the ***information variety module*** removes repeated information; and (10) the ***reference variety module*** replaces repeated referring expressions.

3 Study Design

We aimed to keep all aspects of study design the same to the extent that was possible. In the sections below, we consider different aspects of study design and describe common features and differences, before summarising same/different properties in Section 3.5.

3.1 Evaluated Texts

The evaluations used ten pairs of alternative system outputs randomly selected[4] from the reports for all football matches from one season of one Dutch league (see top of Figure 1 for an example pair). In each pair, both reports are generated by PASS for the same match, but one report is tailored for supporters of one team, the other for supporters of the other team. Each pair of reports was evaluated by each of the 20 participants.

The questionnaires presented pairs of match reports to evaluators side by side (see Figure 1). Both the order of matches and of report variants for each match was identical in the original and the reproduction study. Sides are not randomised: the report on the left is always for the team in the top answer of the first question in the questionnaire, and the report on the right is always for the team in the bottom answer.[4] This may have made it easier for participants to guess the intended readership, hence contributed to the very high stance identification rates in Table 1.

3.2 Evaluation Criteria

Evaluators were first asked to identify the stance of each text, by completing the statement *Deze tekst is bedoeld voor fans van* ('this text is intended for fans of').[5] Then the quality of the texts was evaluated according to two main criteria, namely *Fluency* and *Clarity*, each of which was assessed via (dis)agreement with two statements: (S1) *Deze tekst is in correct Nederlands geschreven* ('This text is written in correct Dutch' and (S2) *Deze tekst is gemakkelijk leesbaar* ('This text is easy to read') in the case of Fluency; and (S3) *De boodschap van deze tekst is mij geheel duidelijk* ('The message of this text is very clear to me') and (S4) *Tijdens van het lezen van deze tekst begreep ik meteen wat er stond* ('While reading this text, I immediately understood what it said') in the case of Clarity. We used the same Dutch statements as the original study.

All four statements ask the evaluators to consider the text in its own right, that is, texts are not evaluated relative to inputs or an external frame of reference. In terms of the quality criteria properties proposed by Belz et al. (2020b), S3 and S4 have the same properties, whereas S1 and S2 do not. S1 falls into the *Correctness* category (i.e. it is possible to define conditions under which the quality criteria are maximally good), while S2 is in the *Goodness* category (i.e. it is not possible to define such conditions). Another difference between S1 and S2 is that the former considers the form of the text only (independently of the meaning), and the latter takes into account both the form and the meaning.

S3 and S4 have the same basic properties as S2, the three mapping to the specific quality criteria of Understandability, Clarity and Readability, respectively, according to the taxonomy proposed by Howcroft et al. (2020, Appendix D) which incorporates the properties from Belz et al. (2020b) as the top three levels of the taxonomy. S1 maps to Grammaticality. In the taxonomy, Clarity (understandability without effort) is a sub-criterion of Understandability (irrespective of effort), a detail which we return to in the results section (Section 4).

3.3 Evaluation Questionnaire

In the original study, pairs of alternative match reports were presented to evaluators side by side, on the same single page as the evaluation questions (a copy of a page from the original questionnaire is shown on the left of Figure 1). The introduction and ten text pairs were given to evaluators printed out

[4] Information provided via email by the authors of the original paper.

[5] Questions and all other text in the questionnaires were in Dutch. We have provided our own translations.

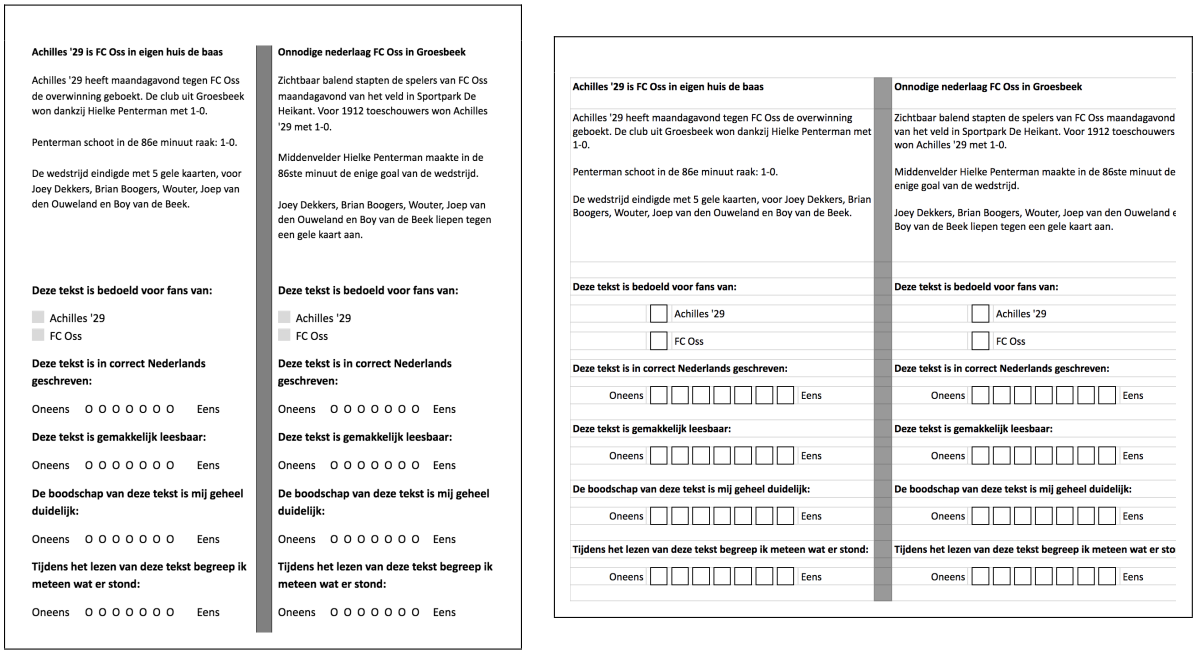

Figure 1: Sample evaluation page from questionnaire in original (left) vs. reproduction study.

on paper. We were unable to do this due to COVID-19 pandemic restrictions, and used online electronic forms instead. The side-by-side text presentation meant we could not use the more commonly used online survey platforms. We opted for a Google Sheet (shown on the right of Figure 1), where only the checkboxes were editable, which made the side-by-side presentation of text pairs possible.

For each of the four quality statements, answers were collected on a 7-point Likert scale (lowest agreement rating 1, highest 7), where the lower (left-hand) side was labeled *Oneens* ('Disagree'), and the higher (right-hand) side *Eens* ('Agree').

Our version of the questionnaire is not identical in every respect, most notably the checkboxes are squares rather than circles, and the alignments and text distribution are slightly different. It cannot entirely be ruled out that such differences affect results, but it seems unlikely.

The recruited participants were provided with two short sets of instructions: (i) the original (Dutch) rating instructions used by van der Lee et al. (2017), in the first tab of the evaluation spreadsheet (updated only to correctly reflect the researchers and institutions involved in the reproduction study), and (ii) additional, specific instructions (in English) relating to the use of the spreadsheet format, in an email that also contained a link to the form.[6]

3.4 Evaluators

In the original study, participants were "all recruited on the campus of the Radboud Universiteit (Nijmegen, The Netherlands). More specifically, [the first author] recruited all participants in the Huygensgebouw of the university, where the faculty of natural sciences, mathematics, and information science is located."[7] Role (student, staff, etc.) and subject area of evaluators was not recorded, but the authors deem it likely that they were students/staff in the faculty subjects (natural sciences, maths, information science) as the faculty's Huygensgebouw building is somewhat isolated on the campus.

We recruited our evaluators remotely (due to COVID-19 pandemic restrictions) via Dutch university research groups known to us, and additionally via personal connections to current and former

[6]Message sent to the participants: "Thank you for accepting to take part in the PASS system evaluation experiment! Below you will find a link to your spreadsheet where your responses will be collected. The sheet contains 12 tabs; we kindly ask you to read carefully the Intro tab and then answer the questions in the following 11 tabs (checkboxes for Page 1 to Page 10, checkbox and free text for Closing). Important notes: (i) The sheet is assigned to you only, and none of the checkboxes or other answers should be filled in when you open it. In the unlikely event that a participant already edited the sheet, please contact us so we can assign you another sheet; (ii) For Pages 1 to 10, please only use lowercase "x" in the checkboxes; you are expected to fill in exactly 10 checkboxes per Page (5 for each text, corresponding to the 5 questions); (iii) Additional instructions are provided (in Dutch) in the Intro tab; (iv) Please complete the evaluation by [DATE]."

[7]Correspondence with the first author of van der Lee et al. (2017).

	van der Lee et al. (2017)	this paper	% in/decrease	CV*
% correctly identified stance	91%	96.75%	+6.32%	6.107
χ^2 for stance identification	233.33†	349.77	–	–
p for χ^2	< 0.001	< 0.00001	–	–
mean Clarity	**5.64**	**6.30**	+11.17%	13.193
stdev	0.88	0.627	-28.75%	–
mean Fluency	**5.36**	**6.14**	+14.18%	16.372
stdev	0.79	0.616	-22.03%	–

Table 1: The results reported in the original paper, alongside the corresponding numbers from our reproduction study. χ^2 is calculated on the contingency table for guessed vs. actual intended stance. CV* is calculated on scores on shifted scales (see in text). † χ^2 is affected by missing values in the original questionnaires.

students and staff in the natural sciences and computer science. This did give us a different cohort of evaluators (e.g. higher average age of 36.8, vs. 20.6 in the original study; some evaluators known to us) and this may be one of the contributing factors to differences in results.

As in the original study, evaluators were not paid or compensated in any other way, and we did not control for demographic balance.

3.5 Summary of Recorded Study Properties

In order to assess reproducibility, and more particularly to be able to compare the degree of reproducibility of different sets of studies, it is important to capture in exactly which respects (in terms of which properties) the reproduction study differs from the original study (Belz, 2021). Below we list the properties in terms of which we *know* whether our reproduction study and the original by van der Lee et al. (2017) were either different or the same, using the basic starter set of properties from Belz (2021), in turn based on Howcroft et al. (2020) and Belz et al. (2020b) (note that system properties don't apply, because the same set of outputs is reused in the present context, rather than regenerated from same inputs):

1. Name and definition of measurand (quality criterion): same.
2. Evaluation modes: same.
3. Method of response elicitation: same.
4. Method for aggregating or otherwise processing raw participant responses: same.
5. Code used to compute and analyse results: different (but only very basic measures were calculated, such as mean and standard deviation).
6. Test set: same.
7. Any preparatory steps such as preprocessing of text taken: same.
8. Procedure of applying measurement method: same.
9. Response collection method: different (paper form in original study, online form in reproduction study, slightly different layout).
10. Quality assurance method(s): different (none in original study which has missing values; checking for completeness and removing questionnaires with all same values in reproduction study).[8]
11. Instructions to evaluators: in evaluation form same, in email different (see Section 3.3).
12. Evaluation interface: different, see Figure 1.

4 Results from Original and Reproduction Study

As described in Section 3.2, the questionnaire contained five rating statements for each text (which we briefly gloss here as 'intended readership', 'correct Dutch', 'easily readable', 'message clear', and 'understood while reading'), but van der Lee et al. (2017) report three scores, for (i) 'intended readership,' (ii) 'correct Dutch' and 'easily readable' combined into a single Fluency score, and (iii) 'message clear' and 'understood while reading' combined

[8] Information provided in direct communication by the authors of the original paper.

		Original study	Reproduction study	CV*
Clarity	S3 avg	5.75 (0.915)	6.36 (0.563)	12.031
	S4 avg	5.52 (0.906)	6.23 (0.686)	14.605
	Both	**5.64**	**6.2975**	13.193
Fluency	S1 avg	5.34 (0.798)	6.22 (0.564)	18.303
	S2 avg	5.41 (0.864)	6.06 (0.661)	13.711
	Both	**5.36**†	**6.14**	16.372

Table 2: Mean scores for the four separate rating statements (Si = ith statement in the questionnaire). Standard deviation in brackets (not corrected for small sample size). CV$*$ calculated on scores on shifted scales (see in text). In our reproduction study, all pairwise differences between S1, S2, S3, S4 are statistically significant at $\alpha = 0.01$ according to a 2-tailed paired t-test, except for the difference between S1 and S4. S1, S2, S3, S4 are also all positively correlated with each other, Pearson's r ranging from 0.36 for S1/S4 to 0.74 for S3/S4. † the mismatch in the average for 'Both' is due to missing values in the original evaluation.

into one Clarity score. The final scores for Fluency and Clarity were calculated by averaging all scores for all texts (both statements and both stance variants) in each case.

We collected 21 evaluations in total, one of which was excluded because the ratings for all questions and all texts were exactly identical in it, which we interpreted as a misunderstanding of the task.[9]

Table 1 shows all results and statistics reported by van der Lee et al. (2017) in Column 2, and the corresponding figures from our reproduction study in Column 3. We also show percentage increases/decreases from original study to reproduction study (Column 4), and the de-biased coefficient of variation (CV*) where appropriate (Column 5), following Belz (2021). The coefficient of variation is the standard deviation over the mean, and is a standard measure of precision used in metrological studies to capture degree of reproducibility. In the implementation we used (Belz, 2021), it is corrected for small sample size. CV* is our primary measure for quantifying the reproducibility of the evaluation scores reported by van der Lee et al. (2017) (stance identification accuracy, mean Fluency and mean Clarity). Note that we shifted all evaluation scales (originally 1..7) to 0..6 prior to computing percentage change and CV*, for fair comparison with the other ReproGen reproduction studies.[10]

As can be seen from the table, all three main evaluation scores went up in our reproduction study: intended stance was correctly identified in 96.75% of cases (compared to 91% in the original study); mean Clarity was 6.3 (compared to 5.64); and mean Fluency was 6.14 (compared to 5.36). Standard deviation for both mean Fluency and mean Clarity went down (better), and the chi-squared value for stance identification and its significance both increased (better).

CV* for Fluency was 16.372 for a mean of 4.75, unbiased sample standard deviation of 0.691 with 95% CI (-3.263, 4.645), and sample size 2. CV* for Clarity was 13.193, for a mean of 4.969, unbiased sample standard deviation of 0.583 with 95% CI (-2.7502, 3.916), and sample size 2. See Belz (2021) for full explanation of this way of reporting CV.

Confidence intervals for (unbiassed) standard deviation (the enumerator in CV*) are large because of the small sample size and corrections incorporated for it. Larger sample sizes increase confidence that the CV for the sample accurately reflects the CV in the general population, and it is important to be clear about level of confidence.

The main conclusions we can draw from the CV* figures is that (i) stance identification is very similar in the two studies, and that (ii) Clarity has a better degree of reproducibility than Fluency.

As mentioned in Section 3.2, according to the standardised quality criteria proposed by Howcroft et al. (2020), S4 is Clarity (understandability without effort), and S3 is Understandability (irrespective of effort); it so happens that Clarity is a subcriterion of Understandability. There are no such parent-child relations between other pairs of S1, S2, S3 and S4, and the taxonomy makes no predictions whether scores for them will be higher or

[9] If we included all 21 evaluations, the average Fluency and Clarity scores would be slightly higher, and degree of reproducibility (CV*) slightly worse.

[10] Both % change and CV in general underestimate variation for scales with a lower end greater than 0.

lower, relative to each other, in the same evaluation. However, the taxonomy does predict that S4 scores (Clarity, or 'understandability without effort') will be lower than S3 scores (Understandability, or 'understandability irrespective of effort') in the same evaluation, because a text that is understandable with effort is also understandable irrespective of effort, but not vice versa. In Table 2, the average S3 score is 6.36, while the average S4 score is indeed lower, at 6.23. This was also the case in the original evaluation where average S3 = 5.752 and average S4 = 5.518. The differences in question were statistically significant at $\alpha = 0.01$ in our reproduction study (see also Table 2).

The taxonomy also predicts that reproducibility (CV^*) and standard deviation will be worse for S4 than S3, which again is borne out in the case of both original and reproduction evaluation by the figures in Table 2. Regarding the other CV^* figures in the last column of Table 2, this is highest (worst) by some margin for S1 ('Grammaticality'), and lowest (best) for S3 ('Understandability'), closely followed by S2 ('Readability') and S4 ('Clarity'). This may come as a surprise as it might be expected that Correctness-type evaluation measures (such as S1) are more reproducible than Goodness-type evaluation measures (S2, S3, S4), which on the face of it involve less clear-cut judgments (e.g. there are no maximally good outputs).

5 Conclusion

In this paper, we reported work which aimed to repeat the human evaluation experiment reported by van der Lee et al. (2017) as closely as possible. We characterised the properties which we know to be either the same or different in our reproduction study (compared to the original study), and presented scores from the reproduction study side by side with scores reported in the original paper (Table 1). We computed percentage increase/decrease, and coefficient of variation as the measure of degree of reproducibility. We found that on the whole, our reproduction study rated the PASS system more highly than the original, with considerably less variation among raters. Furthermore, stance identification accuracy had the highest degree of reproducibility, and mean Clarity scores had a higher degree of reproducibility than mean Fluency.

Note that we have not speculated about the likely reasons for the differences between the two sets of results. We know that those properties marked as different in Section 3.5 are all *possible* reasons. Out of these, it would seem likely that a sizeable part of the difference is down to the different cohorts of evaluators: older, known to us, mostly from computer science backgrounds in the reproduction study, vs. younger, random passers-by recruited in the science building of a university in the original study.

The human evaluation studied here is about as simple as such evaluations get: just one system was evaluated, on three quality criteria and 10 output pairs, each evaluated by the same 20 raters. The coefficient of variation gives a measure of degree of reproducibility that is comparable across measures and across studies, so we can e.g. make the (relative) assessment that Clarity was found to have a higher degree of reproducibility than Fluency. However, the measure does not enable us to make an (absolute) assessment whether either one of them had *good* reproducibility. In order to do this, we would have to know what normally counts as good reproducibility in similar circumstances in NLP. Since NLP currently has very few reproduction studies, and none that report coefficients of variation for human evaluations, such assessments are not possible at this point in time. They will become possible over time if more studies start to report CV (or other measures of precision) for reproduction studies.

Acknowledgments

Mille's work on this study was supported by the European Commission under the H2020 program contract numbers 786731, 825079, 870930 and 952133, and Castro Ferreira's by the Brazilian agency CAPES under Post-doctoral grant No. 88887.508597/2020-00.

References

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

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2020a. ReproGen: Proposal for a shared task on reproducibility of human evaluations in NLG. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 232–236, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz, Simon Mille, and David M. Howcroft. 2020b. 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.

Nadine Braun, Martijn Goudbeek, and Emiel Krahmer. 2016. The multilingual affective soccer corpus (MASC): Compiling a biased parallel corpus on soccer reportage in English, German and Dutch. In *Proceedings of the 9th International Natural Language Generation conference*, pages 74–78, Edinburgh, UK. 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.

Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2017. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104, Santiago de Compostela, Spain. Association for Computational Linguistics.

M. Theune, E. Klabbers, J. R. De Pijper, E. Krahmer, and J. Odijk. 2001. From data to speech: a general approach. *Natural Language Engineering*, 7(1):47–86.