Barriers and enabling factors for error analysis in NLG research

Emiel van Miltenburg, Tilburg University, The Netherlands. c.w.j.vanmiltenburg@tilburguniversity.edu*
Miruna Clinciu, Edinburgh Centre for Robotics, Heriot-Watt University, University of Edinburgh, UK
Ondřej Dušek, Charles University, Prague, Czechia
Dimitra Gkatzia, Edinburgh Napier University, UK
Stephanie Inglis, Arria NLG, Aberdeen, UK
Leo Leppänen, University of Helsinki, Finland
Saad Mahamood, trivago N.V., Düsseldorf, Germany
Stephanie Schoch, University of Virginia, USA
Craig Thomson, University of Aberdeen, UK
Luou Wen, Independent Researcher, UK

Abstract Earlier research has shown that few studies in Natural Language Generation (NLG) evaluate their system outputs using an error analysis, despite known limitations of automatic evaluation metrics and human ratings. This position paper takes the stance that error analyses should be encouraged, and discusses several ways to do so. This paper is based on our shared experience as authors as well as a survey we distributed as a means of public consultation. We provide an overview of existing barriers to carrying out error analyses, and propose changes to improve error reporting in the NLG literature.

1 Introduction

Error analysis is a formalised procedure through which researchers identify and categorise errors in system output. In the context of Natural Language Generation (NLG), error identification often means manually annotating the output text, ideally with multiple annotators (van Miltenburg et al., 2021a). The results of this analysis are often presented in a table, ranking the error categories by their frequency. This goes beyond the more common practice of providing some (strategically) hand-picked examples of 'cherries' (showing good model performance) and 'lemons' (showing the opposite).[1]

While error analysis is relatively labour-intensive, it has some important advantages over commonly used evaluation metrics (see Celikyilmaz et al. 2020 for an overview) or human ratings (Howcroft et al., 2020; van der Lee et al., 2021). These metrics only provide overall scores, and they do not explain what aspects of the output show room for improvement. Error analysis *does* provide this information, and as such it is an essential step towards tackling issues with the output. Based on an error analysis, one might for example establish a benchmark that targets common weaknesses of NLG systems. (See Van Miltenburg et al. 2021a for further discussion.) Moreover, error analyses provide a healthy dose of skepticism with regard to system performance, and as such help avoid the *fallacy of AI functionality* (Raji et al., 2022)[2]. Finally, it is simply not possible to automatically evaluate *all* aspects of NLG output (Raji et al., 2021). Error analysis is flexible enough to identify

*This project was led by the first author. The remaining authors are presented in alphabetical order.

[1] For reference on this terminology, see https://en.wikipedia.org/wiki/Cherry_picking and https://en.wikipedia.org/wiki/Lemon_law

[2] Briefly, the fallacy of AI functionality is the assumption that AI systems work as advertised, and can readily be deployed to carry out the task they were trained to perform, without any strong evidence to back up this claim. Although neural NLG systems may achieve high scores through automatic metrics on community leaderboards, they may still make surprising mistakes that keep them from being useful. These mistakes can be detected through manual inspection.

the issues that are most salient to the human eye.

Despite the usefulness of error analyses, Van Miltenburg et al. (2021a) have shown in their survey of INLG papers published in 2010, 2015, and 2020 that relatively few NLG papers included them (about 11% of the papers surveyed). Gehrmann et al. (2022) provide a similarly low number (about 23% of papers published at ACL, INLG, or EMNLP 2021). It is unclear why most authors do not report error analyses in their work, or how we might encourage authors to carry out an error analysis. We aim to provide clarity on both counts.

Based on earlier work by Van Miltenburg et al. (2021a) and our own experiences as NLG researchers, we identified nine different factors that might influence authors in their decision (not) to carry out an error analysis. We then carried out a public consultation in the form of a survey among NLG researchers to ask for their opinions on error analysis and to identify additional barriers and enabling factors for carrying out an error analysis. This way, we obtained a validated list to discuss in this position paper, where we take the stance that error analysis should be promoted.

Our findings suggest that NLG researchers generally appreciate error analyses they see in the work of others, but they are held back from carrying out an error analysis themselves for various reasons. We discuss the aspects that could enable the reporting of error analyses and argue for meaningful changes to the publication process, so that future researchers may reap the benefits of a research culture where error analyses are rewarded. The code and data for this research project are freely available online.[3]

2 Related work

2.1 Evaluation of NLP & NLG systems

Evaluation is a hot topic that is garnering more attention in both NLG and NLP research communities. There is increasing recognition that current automatic and human evaluation practices are insufficient (Gehrmann et al., 2022). This has resulted, recently, in several evaluation-focused workshops, such as *Eval4NLP*, *EvalNLGEval*, *HumEval*, and *GEM*. This shows a high interest in topics that specifically address the question of evaluation. These workshops are being organised on top of well-established academic conferences and events.

We believe there are several (interconnected) factors that have led to evaluation receiving this increased level of attention:

'Superhuman' performance Tasks are becoming saturated more quickly, with systems performing at or above what has been defined as a human level of performance under the given evaluation setup (Kiela et al., 2021). Current benchmarks have been criticized from two main angles. (1) The decontextualized setup of these tasks tends to make benchmarks less natural, which puts human judges at a disadvantage (Läubli et al., 2020). (2) More generally, it is questionable whether many of these computational tasks suitably model the broad language tasks that they claim to model (Raji et al., 2021).

Uninformative metrics There is also an awareness of poor correlation between human judgements and automatic metrics (Reiter and Belz, 2009; Novikova et al., 2017; Clinciu et al., 2021), as well as the need to move beyond a single number to evaluate the performance of a given system with diverse sets of evaluation suites (Mille et al., 2021).

Unequal comparisons Recent advances, such as the Transformer architecture (Vaswani et al., 2017), provided NLP practitioners with new and undoubtedly powerful tools for building NLG systems and metrics. However, these novel advances have not yet led to a flourish of commercial neural NLG systems, which remain largely symbolic (Dale, 2020).[4] Neural systems are prone to hallucination; they include extraneous and often factually inaccurate content (Ji et al., 2022) that metrics either miss or were never designed to detect (Thomson and Reiter, 2021). Dušek et al. (2020) show that, compared to non-neural data-driven, rule-based, or template-based models, sequence-to-sequence models typically score higher on word-overlap metrics such as BLEU or METEOR, and human ratings for naturalness, but lower in human ratings of overall quality.

Taken together, all of the above factors indicate that our evaluation tasks, metrics, and procedures likely need to be improved so that we can meaningfully compare different systems with each other as well as to humans, simple baselines, or other measures of acceptable performance.

As Gehrmann et al. (2022) note, there are many known issues with evaluation practices in NLG, and many proposals have been made to improve the situation. Gehrmann et al. (2022) looked at the adoption rates of different evaluation techniques, and they show that many current best practices (including error analysis) are not being followed. A recent interview of NLG practitioners (Zhou et al., 2022) showed that authors tend to prioritise certain types of quality criteria (such as correctness, grammaticality, usefulness, etc.) without a shared full understanding of what these criteria mean, something also observed by Howcroft et al.

[3] https://github.com/evanmiltenburg/ErrorAnalysisSurvey

[4] With the exception of machine translation, which may or may not be counted as an NLG task (depending on who you ask).

(2020) and Belz et al. (2020). There are also open questions as to which criteria are sufficient to demonstrate that a system is suitable for purpose.

2.2 Meta-science

This paper is an exercise in *meta-science*. By this term, we mean researchers studying and reflecting on the way scientific research is carried out and subsequently reported. Many people associate meta-science with the open science movement. Following the replication crisis in psychology and other fields, researchers have made different proposals to make our results more open and reproducible (Munafò et al., 2017). In NLP, we have seen initiatives to improve our reporting practices (Dodge et al., 2019) and to pre-register studies before carrying them out (van Miltenburg et al., 2021b).

Next to openness and reproducibility (Belz et al., 2021), one can also look at the incentive structures that exist in the scientific community, and that may boost some kinds of research, while discouraging other kinds of work. 'The incentives' constitute a broad header, which includes *citations* (what kinds of papers get cited?), *awards* (what kinds of papers get recognized through best paper awards?), *community standards* (what is seen as a valuable contribution?), and so on. Next to these, there are also restrictions such as *paper length* (how long should papers be?) which disincentivise authors to write lengthy discussions, and thus form barriers to carry out specific kinds of research. This paper looks at the structural properties of the NLP research culture that influence authors' decisions (not) to carry out error analyses.

This is not the first study looking at publication incentives in NLP. Rogers and Augenstein (2020) discuss our reviewing process and publication culture, and Van Miltenburg et al. (2021a) discuss different incentives that may en/discourage the inclusion of error analyses. Of those incentives, Gehrmann et al. (2022) identify accountability to reviewers as the main driver to improve the evaluation quality in published NLG research. This paper aims to find out to what extent these factors influence authors' decisions.

There is also work outside NLP that studies how to make researchers show desirable behaviour. For example, Ali-Khan et al. (2017) looked into incentives to take part in open science, and Singh et al. (2014) did the same for engagement in public policy. Given the number of variables involved in academic publishing, this is a multifaceted problem with different schools of thought on peer review improvement. Waltman et al. (2022) argue that there are four different perspectives on how to improve peer review (focusing on Quality & Reproducibility, Democracy & Transparency, Equity & Inclusion, Efficiency & Incentives). These categories of schools of thought provide a useful framework for thinking about the implications of any changes to the review process. For example, the idea to require or reward error analyses as part of the review process aligns with the Quality & Reproducibility school, but may go against the principles of the Efficiency & Transparency school, since it further burdens the reviewers (who already show signs of fatigue).

Regardless of your meta-scientific position, any proposal to improve the field should start by asking the relevant stakeholders about their experiences and ideas. We did this through a questionnaire, which is described in the next section.

3 Method

We asked NLG researchers and practitioners for their opinions about error analysis, as well as factors that affect the likelihood of including one in their work. We purposefully did not posit any hypotheses, since our aim is to describe the current perceptions of error analysis, and to sketch a path towards greater adoption of it in NLG research.

Survey Our survey opens with an information letter describing our study and its main goals, followed by an informed-consent form. Participants were allowed to skip all questions except for the informed consent. Upon their consent, participants were asked some general demographic questions, followed by questions about the following topics (see Appendix C for details):

1. Experience reading error analyses
2. Experience carrying out error analyses
3. Barriers and enabling factors to carry out error analyses
4. Necessity and usefulness of error analyses
5. Reporting practices
6. Other comments

Population of interest Our survey targets researchers and practitioners interested in NLG research. To maximize our reach, we spread our survey through Discord, Slack, Twitter, and the Corpora[5] and SIGGEN[6] mailing lists (SIGGEN is the Special Interest Group for ACL researchers working on Natural Language Generation). The SIGGEN community is not very large. For the 2020 SIGGEN board member elections, there were 428 eligible members (i.e., people subscribed to the SIGGEN list, after filtering out any duplicates). Of these, only 92 members cast a ballot.[7] This puts an upper bound on

[5] https://mailman.uib.no/listinfo/corpora

[6] https://www.jiscmail.ac.uk/cgi-bin/wa-jisc.exe?A0=SIGGEN

[7] As reported through the SIGGEN mailing list, by Jose M. Alonso (SIGGEN board member at the time of writing): https://www.jiscmail.ac.uk/cgi-bin/wa-jisc.exe?A2=SIGGEN;5f3966e0.2012

Experience in NLG		Affiliation	
No response	13	No response	12
Less than 2 years	13	Academia	51
2 - 5 years	23	Industry	8
6 - 10 years	5	Other	1
11 or more years	13		
I don't work in NLG	5		

Table 1: Demographics for our participants.

the number of responses we might reasonably expect to receive (particularly since voting takes less effort than filling in a survey).

Participants We received 72 responses (consenting to the survey and answering at least one question). Of those who indicated their affiliation, 51 were academics, eight were from industry, and one selected "other". Table 1 provides a general overview of the demographics. Because of the limited number of respondents per category, we did not carry out any subgroup analyses.

Analysis We performed a *quantitative* analysis of the responses to our closed questions. In addition, we performed a *qualitative* analysis of the open question responses, inspired by other qualitative approaches, such as thematic analysis (Braun and Clarke, 2006) and grounded theory (e.g., Strauss and Corbin 1994). We first read the responses for each question, to get a general sense of the answers. Then, we apply *open coding*: we organise the responses using short, descriptive labels (known as *codes*). The coding was done independently by one or two of the authors for each section. We used these codes to develop coherent themes that are reflected in the answers (*axial coding*). In turn, these themes are used to form a narrative about barriers and limitations, and enabling factors and benefits of error analyses.

The goal of obtaining a high inter-annotator agreement (or inter-coder reliability) is often criticized by qualitative researchers because it assumes the positivist idea that an objective interpretation of the data is both possible and desirable (Terry et al., 2017). If the focus on inter-annotator agreement is too strong, we may lose track of insights that cannot be captured by a strictly defined taxonomy. Instead, we can embrace researcher subjectivity in our quest to gain a deeper understanding of the perspectives of our respondents. Through discussions among ourselves, we ensure that the final narrative is both consistent with and supported by the coded responses. For a related discussion in NLP, see Basile et al. (2021) and the *Perspectivist Data Manifesto* (https://pdai.info), where the authors argue against aggregated datasets that hide any disagreements between annotators.

Pilot and positionality We acknowledge that our own position on the subject of error analysis is not neutral: all authors are in favor of promoting it. However, since we are all researchers in NLG, we did fill in a preliminary version of the survey, along with some colleagues outside of our project, resulting in 12 complete responses. This enabled us to test the questions, determine the duration of the survey, and substantiate our own stance towards error analysis. In lieu of a preregistration (since this is not a confirmatory study, see van Miltenburg et al. 2021b), we made sure to analyse our responses before the deployment of our survey, and committed the report to GitHub, so that it would be time-stamped. None of the authors filled in the final survey, so we can compare the final results to our own responses.

IRB approval Before carrying out our study, we obtained ethical approval from the Institutional Review Board (IRB) at the lead author's university. See §8 for more details on our ethical considerations.

4 Results

Our results are generally organised by the topics identified above in Section 3, but there are several themes (such as the importance of resources such as time and money) that recur throughout this section.

4.1 Experience reading error analyses

Of the 49 participants that answered this part of the survey, the majority (33) recalled having read an error analysis in an NLG paper. Most respondents found reading an error analysis at least moderately useful, and no respondent found it not useful. We also asked these participants what they found useful about the error analyses they have read. Their answers will be discussed in Section 4.4.

Sixteen participants indicated that they have not previously read a published error analysis. We asked these participants whether they found it surprising they had not seen any published error analyses. Seven participants responded to this question. Of these respondents, three participants agreed with this statement. One surprised respondent reasons that NLG errors are evident to daily users of NLG systems, while another observed that without understanding errors properly "it is quite hard to correctly develop a system", contrasting to a blind hyperparameter optimization effort for neural nets.

Participants who did not find the lack of published error analyses surprising highlighted that error analysis is time-consuming, tedious, and that the lack of standards for error analyses prevents useful comparisons even if the analysis is conducted. We also anticipated that these issues would form barriers to the broader adoption of error analyses, and will return to them in Section 4.3.

4.2 Experience running error analyses

A total of 37 respondents answered a question regarding whether they had ever carried out an error analysis, with 25 indicating they had and 12 indicating they had not. The respondents who had carried out an error analysis indicated in their free-text answers that the primary challenge and difficulty in carrying out an error analysis is resources. By this, they chiefly meant time, but the responses also mention tooling, scale, annotators and other similar factors. Error analyses were also seen as difficult to conduct, both in terms of developing a high-quality categorization scheme and in ensuring high inter-annotator agreement. The latter aspect plays into the resource cost, as iterative development is needed to ensure high inter-annotator agreement. This is further exacerbated by the lack of a standard methodology.

Experienced participants Among the 23 respondents that had carried out an error analysis, 13 participants reported having felt that there had not been enough resources or reference material for them to carry out an error analysis. At the same time, almost all of the participants (22 out of 23) that have conducted an error analysis would consider conducting another error analysis again in the future.

When asked why they were likely to carry out an error analysis in the future, the respondents generally indicate a belief in the analyses being useful. Some explicitly state that analyses allow for improved results in the future and provide insights beyond those provided by standard evaluation metrics. Some of the other respondents viewed error analyses as required, some for intrinsic reasons, with one answer being unclear with regard to whether the requirement is an intrinsic one or an extrinsic one. A few responses highlight that their ability to conduct error analyses is limited by resources or collaborator views on their necessity. One respondent viewed error analyses as unnecessary for academic publishing, but as a standard operating procedure for industry work.

Other participants For the participants that have not carried out an error analysis (12), seven have considered doing so, or plan to do so in the future, with only four respondents reporting never having even considered conducting one. Asked for the reason why they had not carried out an error analysis, a few respondents had simply not considered conducting an error analysis. Some lacked the resources, most commonly time, to do so. Multiple respondents indicated that they were conducting, or had conducted, research into rule-based NLG, and as such had ensured their systems did not make any errors before evaluating them.

When queried whether they would be willing to carry out an error analysis, seven respondents would consider conducting an error analysis, four respondents were uncertain, and one respondent answered with 'probably not.' We conclude that our community could potentially publish more error analyses (after all: most are willing to do so), given the right publishing environment. This brings us to the next section.

4.3 Barriers and enabling factors

Quantitative results. Before carrying out this survey, we identified nine factors that may influence the authors' decision (not) to carry out an error analysis. These factors were based on work by Van Miltenburg et al. (2021a), and our experiences as NLG researchers:

1. Page limits: if there is not enough space to present an error analysis, authors may be hesitant to include it or prioritise other aspects of their work.
2. Error taxonomy: if there is no established error taxonomy, authors may find it hard to categorize errors in the output of their system.
3. Annotation tools: if there are annotation tools dedicated to error analysis, it would make the process easier.
4. Crowdsourcing template: if there is no template, there is a higher barrier to carry out an error analysis, because the authors need to design a task by themselves.
5. Appreciation from reviewers: if reviewers do not ask for error analyses, or they do not reward them enough, authors are less tempted to carry out an error analysis.
6. Availability of annotators: if there are no annotators (other than the authors themselves), then carrying out an error analysis may be considered too much work to carry out alone.
7. Time: error analysis can be time-consuming. If researchers don't have enough time to carry out an error analysis, they will not do it.
8. Money: if researchers do not have the money to hire annotators/crowd workers, they need to carry out the full error analysis themselves.
9. Collaborators: error analysis may be considered too much work to be carried out alone.

Figure 1 provides an indication of which factors

I would be more likely to carry out an error analysis in a conference/journal paper if…	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
... there was a higher page limit.	3	3	9	12	4
... there would be an existing error taxonomy that I could use.	1	2	6	11	12
... there would be dedicated annotation tools for error analysis that I could use.	1	4	7	10	10
... there would be a crowdsourcing template for carrying out error analyses.	1	4	8	11	8
... reviewers paid more attention to error analyses.	0	2	6	9	15
... there were an available pool of annotators or crowd workers.	3	3	6	13	7
... I had more time.	0	4	2	10	16
... I had more money.	1	4	3	9	15
... I had more collaborators.	0	3	7	10	12

Figure 1: Heat map table showing our participants' (dis)agreement with nine statements about factors that make them more likely to carry out an error analysis. Numbers are absolute, i.e., counts of participants (dis)agreeing. Darker cells contain higher numbers.

make it more likely for our participants to carry out an error analysis. For all nine factors, the results skew positive, with participants recognising all the identified factors act as barriers to completing error analyses. Three of these stand out: time, money, and recognition from reviewers seem to be the most important. These results are also confirmed by the qualitative results.

Qualitative results. We further surveyed participants regarding other barriers that prevent them from carrying out an error analysis and what factors would instead enable them. The participants confirmed that resources are the premier barrier: time (including the time that could be allocated for improving the NLG systems), funds, tools to help with error analyses including a taxonomy of errors, access to experts that could help with error annotation as well as lack of system outputs in literature which could be used for comparison. Similarly, Zhou et al. (2022) also found that time limitations, especially for industry teams, constrained the use of qualitative or participatory evaluation approaches. As expected, access to these resources was identified as an enabler that helps researchers focus their effort on performing error analyses.

A number of participants mentioned that the current research culture does not reward such analyses, which prevents them from performing and reporting them. In fact, most participants identified culture change towards error analysis as an important factor for adopting it. Specifically, the participants proposed making error analysis a requirement for papers and explicitly recognising it in review forms; this should highlight its importance both for research and industrial/commercial applications.

15 participants responded that they are more likely to include an error analysis in a journal article, motivated by the benefits of publishing in a journal article, such as a higher page limit, increased time to publish, and higher demands on details. However, 14 participants responded that is equally likely to include an error analysis in a journal article, as well as in a conference publication as NLG research is heavily conference-focused.

When asked if there are currently enough resources to support error analysis, the majority of respondents to this question reported that error analysis resources are still missing (20), while a few participants stated that there are some resources available (10). Participants suggested that a well-documented error analysis taxonomy and procedures and standards, as well as annotation tools, are missing. Also, funding plays an important role in performing error analysis.

4.4 Necessity & usefulness

Quantitative results. Figure 2 shows the participants' attitude towards error analyses. The respondents overwhelmingly agree that error analyses are useful and provide insight into system performance. At the same time, we find that our participants have mixed feelings about carrying out an error analysis themselves. When asked whether they find it enjoyable or boring/tedious, there is a slight majority agreeing with both statements. Although some respondents responded positively to only one of the two statements, nine participants somewhat agreed with error analysis being both "enjoyable" and "tedious." Based on this observation, we might say that carrying out an error analysis is like eating broccoli or Brussels sprouts; we all know it is good for you (and there certainly are long-term health benefits), but not everyone enjoys the taste, and it may be difficult to finish your plate.[8]

Should both journal and conference papers include error analyses? Developing our questionnaire, we ex-

[8]Continuing the analogy: in our experience, it is generally more enjoyable to eat (annotate) together, than having dinner alone, even if you're not having the same meal.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
There should be more error analyses in the NLG literature	0	1	1	10	19
Error analyses are a valuable part of a paper.	0	0	2	4	25
Carrying out an error analysis is enjoyable.	0	7	6	14	3
Carrying out an error analysis is boring/tedious.	3	4	6	17	0
Error analyses are necessary to fully evaluate the performance of an NLG system.	1	0	1	5	23
Knowing what errors a system makes is helpful for future research.	0	0	0	9	21
Knowing what errors a system makes is helpful for practitioners/NLG in industry.	0	0	1	5	24
If you publish at a conference, and you present an NLG system as one of your main contributions, you should include an error analysis.	0	0	5	13	12
If you publish in a journal, and you present an NLG system as one of your main contributions, you should include an error analysis.	0	0	2	10	18

Figure 2: Heat map table showing the distribution of responses to a question where participants were asked to indicate their (dis)agreement with nine statements about the desirability/usefulness of error analyses. Numbers are absolute, i.e., counts of participants (dis)agreeing. Darker cells contain higher numbers.

pected that there would be a difference in standards between journals and conferences; journal papers might be seen as definitive products of research, while conference papers are still work-in-progress. The preliminary nature of conference papers might make our participants more lenient. Surprisingly, the majority of our participants agreed for both journal and conference papers that they should include an error analysis (if applicable). Admittedly, the agreement is less strong for conference papers than for journal papers, but these results do show that error analysis is important to readers of NLG papers.

Qualitative results. We asked the participants who have read error analyses in the past about the usefulness of those error analyses. By far the most common answer was that error analysis could help identify remaining challenges and direct future work, both at a high level, and in terms of improving individual systems. Several responses also mentioned researchers' bias and noted that a thorough error analysis is better than cherry-picked examples more commonly seen in a qualitative analysis section. Some respondents indicated that error analysis was a good complement to imperfect metrics, and could detect overlooked errors. The usage of error analysis to gauge whether a system was suitable for its purpose was also mentioned, along with gaining a better understanding of system limitations.

We received 27 responses in total to our question on what kinds of papers error analyses may be useful for. Most replies (16) mentioned experimental papers or papers presenting a new system. Five more respondents even implied that *all* papers should include error analysis; this probably still applies mostly to experimental papers as they are the most common type. Nine respondents mentioned various specific sub-fields or system types (e.g. end-to-end systems, dialogue systems). Three participants mentioned evaluation-related papers specifically. We also received multiple general remarks arguing in favour of error analysis and/or complaining about the lack thereof in current works.

4.5 Reporting practices

What should be included in reports containing an error analysis? Common themes underlying the responses were reporting practices that could enable replicability, reliability, and usefulness of both methodology and results. Table 2 provides an overview of the responses that were given in our pilot study, the main survey, or both.

Of the 16 respondents who answered this question, seven focused on reporting descriptive details such as the annotator training process, annotation process, and actual annotator details, expressing that this would better enable replicability of results as well as enable comparisons across studies via replicable methodology. This also includes reporting details that ensure the reliability of the methodology and results, such as reporting inter-annotator agreement and evidence of annotator quality or sampling method (specifically arguing for statistically-driven sampling).[9,10] At the same time, one participant warned against over-formalising error analyses.

Seven respondents explicitly argued for reporting

[9] In a recent publication, Shimorina and Belz (2022) provide a useful template for reporting these details.

[10] Also see Popović and Belz 2022 for a discussion of reporting scores and agreement for error annotation tasks.

Source	Recommendation
B	Provide the annotation guidelines, with an explanation of how these were created (e.g. as an appendix).
R	Provide details on how annotators were trained.
B	Provide details about the background of the annotators.
R	Provide inter-annotator agreement scores, to assess the reliability of the annotation process.
R	If using an existing error taxonomy, ensure it is appropriate for your system.
R	If possible, provide a comparison between different systems.
A	If comparing different systems, use appropriate statistics (e.g. Chi-square tests comparing the distribution of particular kinds of errors).
R	Provide a reflection on the potential sources of the errors.
R	Provide correlation scores between different types of errors to see which ones co-occur.
B	Provide details on how the outputs were sampled (e.g. stratified sampling).
A	Provide actual examples of system output.

Table 2: List of reporting practices suggested in the responses to our questionnaire by either the current authors (A), our respondents (R), or both (B).

practices related to error taxonomies and compared systems. The goal here is to increase the usefulness of the analyses for both aiding researchers and understanding systems: reporting of (potentially customized) error categories with definitions, justifications, and limitations to enable use in other works, and explicitly reporting system comparisons and observations (such as identifying commonalities across systems and the system impacts or correlations of errors).

Two participants also left suggestions in the 'other comments' field. One noted that "Error analysis should focus on language features, text genre characteristics and adequacy to the task, not a mere statistical analysis." The other participant highlighted the importance of sentence structure and the manual labour that goes into an error analysis. We may interpret this in light of the fact that humans can pick up nuances that (thus far) NLP systems have not been able to detect.

5 Discussion

5.1 Incentives and social dynamics

As noted in Section 4.3, most participants thought a culture change is necessary to make error analysis a common practice. One promising idea in this direction seems to be to explicitly reward researchers with badges for exemplary behaviour, such as preregistering confirmatory studies and publishing research code and data.[11] This idea has been proven to work in psychology (Kidwell et al., 2016), where open science practices increased among published papers after the introduction of badges displayed alongside each paper.[12] Building on the badges from the ACM (2020), NAACL 2022 also offered reproducibility badges.[13] Over 25% of accepted submissions earned at least one badge. Relatedly, as program chairs of COLING 2018, Derczynski and Bender (2021) introduced awards for specific parts of papers (best evaluation, most reproducible, best challenge, best error analysis) instead of having an overall best paper award. On top of that, only papers with published code and data were eligible for any of these awards. Following this initiative, the conference saw about one—third of all papers with full code. One other innovation from Derczynski and Bender (2021) was to introduce *paper types*: categories of papers with associated review forms that are tailored to the kind of contribution that authors want to make.[14] These review forms are public, so authors can prepare their work accordingly. Having specific review forms may nudge authors to include different kinds of information in their submissions, which they perhaps would not have included otherwise.[15]

It is still hard to gauge the impact of these initiatives on the NLP community, but at least open science badges help make our community norms and values explicit. However, following Yarkoni (2018), we have to acknowledge that scholarly behavior is also just a matter of personal responsibility. If you believe that it is important to highlight the limitations of your approach, then the time and effort needed to carry out an error analysis should be included in the planning of your project.

The carrot and the stick Incentives generally come in two forms: the carrot and the stick. The initiatives discussed above are an example of the former, rewarding authors for good behavior. What about the latter? Can we require authors to carry out an error analysis, *or else...*? This is not without precedent. NLP conferences have recently started requiring the inclusion of *Limita-*

[11] See: https://www.cos.io/initiatives/badges

[12] Though see Crüwell et al. 2023 for a critical evaluation of the *open data badge* policy in the *Psychological Science* journal.

[13] See: https://naacl2022-reproducibility-track.github.io

[14] NEJLT also uses the same paper types. See: https://www.nejlt.org/authorinfo/

[15] We are not aware of any studies that look into the effects of reviewing forms on the form or content of the submitted work. Future research could study e.g. the content of NLP papers before and after introducing (new criteria on) checklists for conference submissions.

tions and *Ethical Considerations* sections for all papers where such sections are appropriate (i.e., most NLP papers). Moreover, one might argue that an evaluation of an NLG system is not complete without an error analysis, especially given the unreliable nature of automatic metrics and the reductive nature of summary scores. It is simply good scholarship to provide an error analysis.

When should error analyses be required? Almost all of our respondents agreed that journal submissions should include an error analysis, and the majority of our respondents also agreed that the same should hold for conference papers. In hindsight, it is probably not the *venue* that counts, but the *state of completion* of the project. If you report on a finished project, then the final publication is the end product, regardless of the venue. At this point, the project should be fully documented, including an overview of all the limitations of the end product. This prevents *technical debt* (Sculley et al., 2015) from building up in the NLG community.[16]

Based on our observations, we would like to posit the following rule: *if* a paper presents a final result (as opposed to work-in-progress), *and* the paper presents both an automatic and a human evaluation, *then* the paper should also contain an error analysis.

Getting there A priori, the carrot is preferable to the stick. Without any hard requirements, there is more room for exceptions, i.e. papers that do not fit the traditional mould of NLG publications. Furthermore, encouragement policies are less likely to run into resistance from the community, compared to hard requirements. We do not necessarily need *everyone* to provide an error analysis; if we can encourage a critical mass of researchers to provide error analyses, then this will just grow to become the norm.

5.2 Making space for error analyses

Although page limits do not seem to be the main barrier for carrying out error analyses, it is also clear that additional content takes up space. We have recently seen this with limitations and ethical considerations sections, which for many conferences are now allowed to be put on an additional page following the conclusion (even though ethical considerations are an integral part of research design). EMNLP also features a reproducibility checklist, the authors of which suggest that researchers may want to provide important technical details in the appendix.[17] From these observations, it seems that our community is struggling to put all relevant information in the four-to-eight pages that are currently allotted to conference papers. *The medium is the message* (McLuhan, 1964); if conference papers remain the main publication venue for NLP research, then it is important that our values are reflected in the submission types. All relevant information should fit in the main body of the paper. We discuss two options to improve the situation.

Option 1: increase paper length The first option is to simply increase paper length (e.g. moving from 4/8 pages for short/long papers to 5/10 pages), or to add another length tier (resulting in papers of either 4, 8, or 12 pages).[18] This creates additional space to include relevant information, without introducing any new requirements. Over time, we should see the community converge on the type and amount of content that is required for papers in each tier to be publishable. The main attraction of this proposal is its simplicity, requiring little to no extra administration. The downside of this proposal is that it is unconstrained, so without any additional requirements it is not clear whether authors would actually carry out more error analyses.

Option 2: reserve space for error analyses Continuing the previous section (§5.1), the *ACL main conferences in NLP have not just required authors to include *limitations* and *ethical considerations* sections; they have also given authors additional space to provide these sections. Typically this space is provided *after* the conclusion, to ensure that authors do not cheat the page limit by using the additional space for other purposes. One way to stimulate error reporting would be to do the same for error analyses as well. On the one hand, this initiative adds more administrative burden, and it prevents authors from integrating the relevant content into the narrative of the paper (at least at submission time), but it does guarantee that authors actually include an error analysis, and it helps to normalise the idea that every paper should have sections detailing limitations, ethical considerations, and error analyses.

5.3 Error taxonomies & standardization

Recent work in the NLG community has aimed to provide an overview of our evaluation practices, and move towards standardising our terminology and assessment materials (Belz et al., 2020; Howcroft et al., 2020). There have been similar efforts in the areas of Explainable AI (Nauta et al., 2022) and Intelligent Virtual Agents (Fitrianie et al., 2019, 2020). The majority of our respondents indicated that they would be more likely to carry out an error analysis if there were an existing taxonomy of

[16] Epstein et al. (2018) make a similar point, but using a different framing than Sculley et al. (2015). They talk about the *AI knowledge gap*, where studies on new systems are published faster than studies characterizing the behaviour of those new systems.

[17] See: https://2020.emnlp.org/blog/2020-05-20-reproducibility

[18] Of course there are many other possibilities, including the option to let go of page limits altogether, or to only set an upper bound for conference submissions (based on the reviewing timeline).

errors that they could use. However, is it even possible to establish a standardised error taxonomy for NLG output? As one participant noted: it is "better to use a sensible characterization of errors that actually occur [...] than trying to shoehorn them into an existing taxonomy."

Several taxonomies have been proposed for different NLG/NLP tasks and some are used for evaluation by annotation, an approach that readily lends itself to error analysis. For machine translation, Popović (2020) asked annotators in separate experiments to mark comprehensibility and adequacy errors, also distinguishing major errors (those which alter the meaning) from minor errors (grammar or style). Freitag et al. (2021) asked annotators to mark up to five of the most severe errors within a segment, these were then assigned both a category and a severity. Costa et al. (2015) proposed a linguistically motivated and hierarchical taxonomy, and He et al. (2021) proposed a taxonomy and then used it to create the TGEA annotated dataset. For factual accuracy in data-to-text generation, Thomson and Reiter (2020) asked annotators to mark non-overlapping spans of text and assign them one of six categories. For prompted generation, Dou et al. (2022) asked annotators to mark all errors from a wide range of categories,[19] allowing multiple overlapping annotations and with some subjectivity between categories (*Encyclopedic* for one person could be *Needs Google* for another). These taxonomies could be used as-is, or they can be developed further to provide a more detailed analysis.[20]

NLG is difficult to define as a field (Gatt and Krahmer, 2018) and despite sharing some commonality (the generation of text), the purpose of any generated text is key to how we interact with it (Evans et al., 2002). This makes it difficult to form a "one size fits all" definition of NLG and, similarly, an error taxonomy. However, there are some high-level considerations when selecting or adapting a taxonomy:

Evaluation criterion: Humans are known to miss some errors when reading (Huang and Staub, 2021), and whether their annotations for one criterion might affect their subsequent reading and annotation of the remaining text is unknown. Asking annotators to consider multiple criteria simultaneously could compound this problem, increasing both disagreement and the volume of missed errors. In line with more general best practices for NLG evaluation (van der Lee et al., 2021), annotators should consider one criterion at a time.

Annotator agreement: Very low inter-annotator agreement might be indicative of an annotation procedure issue, but disagreement between annotators does not necessarily mean that some of the annotations must be flawed (Popović, 2021). Thomson and Reiter (2021) noted that even within a single criterion, two annotators could provide sets of errors that only partially overlap, yet can both be considered valid representations of the same complex underlying problem. In addition to calculating agreement, annotators could check each other's annotations and indicate whether they consider them one valid way of describing the underlying problems Thomson et al. (2023).

Distinct categories: Principles from both taxonomy and close-response survey design are also relevant to annotation; categories should be mutually exclusive and as exhaustive as is practical (Fowler and Cosenza, 2008). If there are too many categories (making it hard for annotators to keep all distinctions in mind), it may be beneficial to use more coarse-grained taxonomy.

Error instance vs cause: Hallucination is commonly considered a core error type in NLG but Van Miltenburg et al. (2021a) argue that errors should not be defined in the first instance by the process that caused them. An error in generated text can be defined in terms of how it fails to meet its purpose, a grammatical error, factual mistake, etc. The reason for this failure can then (optionally) be determined. Process errors should be recorded separately from text errors, i.e., we could mark an error as being an incorrect named entity, then indicated that this was caused by hallucination. Different types of hallucination, such as intrinsic versus extrinsic (Ji et al., 2022), can be considered at this second stage.

Error severity: Different errors may have a different impact on readers (van Miltenburg et al., 2020b). Similarly to error causes, severity can be assessed after the error is identified and categorised (Popović, 2020; Freitag et al., 2021), although this may be done immediately as part of recording the error. In such cases, annotators are following a sequential procedure where they first find the error span and assign a category, then consider how severe the error is.

Although there are still many (context-dependent) decisions for authors to make about the design of a suitable error analysis, these considerations do constrain the space of possible approaches. Moreover, it should be possible for researchers to agree on a standard error analysis taxonomy and format for common NLG tasks. These could be decided upon during the development of new tasks, or with new iterations of existing shared

[19] Grammar and Usage, Off-Prompt, Redundant, Self-Contradiction, Incoherent, Bad math, Encyclopedic, Commonsense, Needs Google, Technical Jargon.

[20] For more examples, Huidrom and Belz (2022) provide a further survey of existing error taxonomies, which they plan to use to develop a taxonomy of semantic errors in NLG output.

tasks, e.g. WebNLG (Castro Ferreira et al., 2020) or the surface realization shared task (Mille et al., 2020).

Another useful step may be the development of guidelines for what the output should look like. This is mostly a problem for neural data-driven NLG systems, which are commonly trained and evaluated on crowd-sourced data, where annotators are asked to write an output text for a given input. If the guidelines for writing those texts are underspecified, then there will (1) be a high degree of variation in the human-authored texts (see, e.g., van Miltenburg et al. 2017),[21] and (2) the decision of what the output should look like is essentially delegated to the crowd, meaning that the standard for comparison is only extensionally defined by the training corpus (van Miltenburg et al., 2020a; Schlangen, 2021). Without any clearly defined standards, it is more difficult to judge the quality of automatically generated output. With standards in place, it is also possible to define deviations from the norm, which we can then more easily flag as errors.

Finally, any taxonomy is better than no taxonomy at all. If there is no existing set of error categories, then we encourage authors to develop a taxonomy of their own. Once established, error taxonomies can have a big impact on future work in two ways:

1. They facilitate future error analyses and make it easier to compare different systems,
2. They may steer future research by highlighting specific issues in system output that should be resolved.

5.4 Resources: time, money, and tools

Time and money were considered by our respondents to be the main barriers to carrying out error analyses. These two factors are also clearly correlated: time-consuming tasks can be outsourced by paying someone else to do them, and vice-versa. You can save money by doing everything yourself. So what if you have neither time nor money to spend on error analysis?

Using student annotators. The go-to option for cheap annotation in academia is to have students carry out the work. We do not think it is ethical to have students annotate large amounts of data for free, but at least small batches of error analysis could be incorporated in education. We suggest the following guidelines for ethical data collection:

1. The exercise should support the end-goals of the course.
2. The amount of items to annotate should not be excessive. Once the learning goals have been achieved, it is not necessary to continue to exercise.
3. The data should be anonymised such that it is not possible to identify which student contributed the annotations.
4. Students should have the opportunity to opt-out of their data being used for research purposes (without this having any negative effect on their grades). Or even better: use an opt-in procedure where students may (anonymously) submit their results.
5. As a corollary of the previous points: grades should not be contingent on data quality.
6. Researchers should check with their colleagues or their institutional review board (IRB) whether this form of data collection is appropriate, given the power differential between teachers and students.

In short: 'free' annotation should not come at the cost of students' well-being. It requires dedication, and an up-front investment to responsibly integrate the exercise in an educational context.

(Lack of) time is an illusion. Many researchers have internalized the corporate values of *speed* and *efficiency*, prioritizing them over the slow contemplation that has traditionally been the hallmark of academia (Berg and Seeber, 2018). As a result, it often *feels* like we are just living from deadline to deadline, without any time to sit down and thoroughly analyze our results. But this is a *choice*; there are other options! In his (2018) COLING keynote, Min-Yen Kan promoted the idea of 'slow research' in NLP, as a counterpart to the fast-paced style of research that has grown popular in recent years. We would argue that a publication with a slow, deliberate error analysis may over time be more impactful than a paper lacking such in-depth information. (One might respond that slower research risks being scooped, but this overlooks the fact that error analyses and other time-consuming methods are substantial contributions in and of themselves.)

Of course, fast-paced research is there for a reason; many researchers believe they are expected to live up to the aphorism that they should *publish or perish*. Not publishing enough papers may reduce your chances of success in academia.[22] But, again following Yarkoni (2018), we shouldn't sacrifice good scholarship based on these incentives. At this point we should

[21] This variation is not necessarily bad (users may sometimes appreciate diversity), but it has been shown for use cases such as professional weather forecasting that users appreciate consistency in the output (Sripada et al., 2004). Either way, we do need to ensure that the texts are congruent with the purpose of the task. If the purpose is not made clear to the crowd-workers, the human-authored texts may be sub-optimal with regard to the communicative situation that the NLG system is embedded in.

[22] And as Rahal et al. (2023) note: "Quality research needs good working conditions." With more permanent positions, researchers may find themselves better able to focus on long-term research goals.

ask ourselves: how long does an error analysis *really* take? Granted, an extensive error analysis can be quite labour-intensive, but we should not let perfect be the enemy of good. Including a systematic error analysis of any kind is already much better than randomly picking some cherries and lemons to include in the appendix.

Just do it yourself. As with any annotation task, it is important to at least carry out some portion of the analysis yourself. There is no replacement for getting familiar with the output of your system, or with the process of identifying potential errors. This *dogfooding*[23] ensures that the task is feasible, and decreases the odds of overlooking important properties of the generated data. Although the majority of our participants found error analyses to be boring/tedious, there are clear benefits to this method, and an equal majority found the process to be enjoyable as well. As Sambasivan et al. (2021) note, data work is considered to be much less glamorous than modeling, but it is essential that we do it anyway.

Trade-offs are inevitable. Some NLG tasks are more time-consuming to evaluate than others. For example, manually assessing the quality of longer texts (e.g. summaries, stories, or news articles) takes longer than the assessment of shorter texts (e.g. image captions, product descriptions). In a multilingual setting, evaluation is also going to be more involved: one may want to have a universal set of error categories that work across different languages, or a large enough sample size for outputs in each language under consideration. Given time and money constraints, it may not be feasible to carry out a large-scale error analysis. As noted above: any error analysis is better than none, but the authors also need to be clear about their considerations and the limitations of their analysis. Example trade-offs include:

1. Coverage versus specificity: Carry out an in-depth analysis of a specific subset of the outputs, or a more superficial analysis of all the outputs?
2. Coverage versus reliability: Annotate more outputs with fewer annotators per output, or fewer outputs with more annotators?

There is no one-size-fits-all recommendation with regard to the trade-offs that authors should make. This process is guided by the research question, hypotheses, and the claims that the authors would like to make about their system. The strength of the error analysis influences the extent to which any claims about system performance can be substantiated.

[23] For lack of a better term, although *dogfooding* is typically used to refer to developers using their own software rather than just inspecting the results. See: https://en.wikipedia.org/wiki/Eating_your_own_dog_food

Optimisation and tools It may be possible to develop tools to carry out error analyses more efficiently. For example, after developing a dedicated app or mobile website, error analyses could be carried out *on the go* in brief sessions (e.g., waiting for the bus, or on the train). This is an interesting avenue for future research, although following Section 5.3 one might wonder whether it is feasible to develop universal tools for supporting error analysis, given the challenges of standardisation.

5.5 Collaboration

The majority of our participants indicated that they would be more likely to carry out an error analysis if they had more collaborators. How can we address this issue?

Shared tasks One proposal is to copy successful evaluation practices from other subfields of NLP. The Workshop on Machine Translation (WMT) asks all of its participants to rate a collection of translations "proportional to the number of tasks they entered" (Barrault et al., 2020).[24] This approach has been proposed in the NLG community as well, for the GEM shared task (Gehrmann et al., 2021, p. 109). Next to providing ratings, participants of shared tasks could also conduct error analyses. Once the outputs of all systems are submitted, the participants could analyse a subset of the outputs of all systems using an agreed-upon error taxonomy and annotation methodology. This has at least three distinct advantages: (1) Authors would be intimately familiar with the different kinds of mistakes that systems could potentially make, (2) system labels would be hidden so that participants are not biased in their judgments, (3) each shared task would produce richly annotated datasets (potentially further enriched with human and automatic evaluation scores).

Sharing resources Researchers in Psychology have proposed StudySwap (Chartier et al., 2018): a dedicated platform to share resources, such as equipment, participants, expertise, and so on.[25] The NLG/NLP community lacks such a platform. Of course, researchers may informally help each other out, but this is always easier for established researchers with a bigger network. It is tempting to suggest a centralised platform for collaborative NLP/NLG research, but this may not be feasible to sustain.

[24] These judgments are further complemented by those from crowd-workers, and a dedicated pool of linguists.

[25] Unfortunately the platform is currently dormant, but it has resulted in fruitful collaborations in the past.

6 Limitations of this study

Because our participants are volunteers, we run the risk of possible self-selection bias: only people that are interested in error analysis may have taken the time to respond to our survey. This means that our survey may overestimate the support for error analysis in our community. This issue is inherent to any voluntary survey. (For example, Jakobsen and Rogers (2022) report this limitation as well). Given this limitation, we are still able to make existential claims about the barriers that exist for researchers wanting to carry out and publish error analyses; at least some researchers are held back by the barriers listed above.

Another limitation is that our sample size is relatively small, with 72 participants. As we discussed above, this is not very surprising, given the limited size of the NLG community. Our participants were also allowed to skip as many questions as they liked in our survey. As a result, several questions were answered by less than half of our 72 participants. This may be seen as a limitation of our study, because a small group of researchers may not be representative of the larger research community. But our study does serve its original purpose: to consult other researchers about potential barriers and enabling factors for the use of error analysis in NLG, and to ensure that our list of barriers and enabling factors does not have any glaring omissions.

Two participants indicated that they were not familiar with the concept of error analysis before this study. One of them also noted that, because of this, they would have liked to see an "I don't know" option for the Likert scale questions (although it was possible to leave these questions blank).

7 Conclusion & Future Work

We have carried out a survey among NLG researchers and practitioners. Our respondents were generally positive about error analysis, but they did see multiple barriers to the general adoption of this approach. By removing or minimizing these barriers (as discussed in Section 4.3) and motivating authors to include error analyses in their work (section 5.1), we may see greater adoption of error analysis in the future.

In the future, we would like to focus on developing tools and resources, such as error taxonomies, annotation tools, and clear guidelines that would help to encourage more routine and robust error analyses. In addition to development of resources, there also needs to be a structural change in the incentives around research publication that encourages prospective authors to conduct such analyses. More work is still needed to help enable error analyses by researchers and practitioners, but we are optimistic about the future of evaluation within NLG.

8 Ethical considerations

8.1 Positionality and transparency

We are aware that our position as authors is not neutral: we are all proponents of error analysis, and many of us have enough job stability to not have to worry about publishing as much. This gives us the time and space needed to publish longer studies, potentially with detailed error analyses. We have attempted to explicitly capture our opinions about error analysis before distributing our survey. This information is also available through our GitHub repository, both in raw form as well as in a short report.

8.2 IRB approval

Before carrying out our study, we obtained IRB approval from the lead author's university. This process separately considers the treatment of our participants, and the treatment of our research data. Our considerations for the IRB are detailed below.

8.2.1 Participants

Invitations: We sent out the invitation to take part in our study through social media and two mailing lists (SIGGEN and Corpora). These mailing lists are explicitly set up for the purpose of sending each other news (e.g. about upcoming conferences) and questions. People voluntarily subscribe to these mailing lists, and the invitation for our study falls within the expected use of those lists.

Information letter and informed consent: Our study starts with an information letter, describing the goal of the study, the expected duration, and potential risks/benefits of the study. The letter provides the names of the researchers involved, as well as an email address to contact for more information. The information letter is followed by a separate informed consent form, which specifies explicitly what participants agree to, when they take part in our study. They are also reminded of their rights: participation is fully anonymous, and participants are always free to quit the survey or withdraw their consent at any time, without any negative consequences.

Demographics and survey length: We aimed to minimize the amount of data collected about each participant. We only collected their general affiliation (Academia, Industry, Other) and their amount of experience (expressed in broad ranges, so as not to make people identifiable by the exact number of years). The rest of the survey has been streamlined to reduce the

burden as much as possible, and should be doable in about 10-15 minutes.

8.2.2 Data

IP-addresses: By default, our survey platform (Qualtrics) is set to store the IP addresses of all participants. Because this may be identifying information, we turned this setting off.

Data management: Because the data is fully anonymous, and participants have consented to the publication of the data, we are free to publish the responses to our survey. Before doing so, we checked the responses to the open questions for any identifying information that may need to be removed to protect the identity of our participants. All code and data have been shared through GitHub, and submitted along with this paper, thus providing maximal transparency.

8.3 Intended use of our results

Our proposals should be seen as part of the broader and ongoing discussions on publication and peer review in NLP (Rogers and Augenstein, 2020), and the state and quality of evaluations in NLG (Howcroft et al., 2020). As such, our proposals are not final, but are meant to be discussed further.

Although our policy proposals are grounded in the responses from the general NLG community, we do not know whether they are broadly supported by the community. Workshop and conference chairs may experiment with minor changes, but bigger changes may need to be put to a vote.

Acknowledgements

We would like to thank Emma Manning, who helped prepare this study and provided feedback on the survey questions. We would also like to thank members of SIGGEN, Corpora List, and everyone else who responded to our survey. We also appreciate the suggestions made by the anonymous reviewers of this paper (both for NEJLT, and for an earlier version we submitted to EMNLP). This project was supported by multiple different grants. Ondřej Dušek's work was funded by the European Union (ERC, Grant agreement No. 101039303 NG-NLG). Leo Leppänen's work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). Dimitra Gkatzia's work is supported under the EPSRC projects NLG for low-resource domains (EP/T024917/1) and CiViL (EP/T014598/1). Craig Thomson's work was supported under an EPSRC NPIF studentship grant (EP/R512412/1) and the ReproHum EPSRC grant (EP/V05645X/1).

References

ACM, Association for Computing Machinery. 2020. Artifact Review and Badging, Version 1.1. Online policy document, retrieved June 2022.

Ali-Khan, Sarah E, Liam W Harris, and E Richard Gold. 2017. Point of View: Motivating Participation in Open Science by Examining Researcher Incentives. *Elife*, 6:e29319.

Barrault, Loïc, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Basile, Valerio, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. In *Conference of the Italian Chapter of the Association for Intelligent Systems*.

Belz, Anya, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. 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.

Belz, Anya, 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.

Berg, Maggie and Barbara K. Seeber. 2018. *Challenging the Culture of Speed in the Academy*. University of Toronto Press, Toronto.

Braun, Virginia and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Castro Ferreira, Thiago, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem,

and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of Text Generation: A Survey. *CoRR*, abs/2006.14799.

Chartier, Christopher R., Amy Riegelman, and Randy J. McCarthy. 2018. Studyswap: A platform for interlab replication, collaboration, and resource exchange. *Advances in Methods and Practices in Psychological Science*, 1(4):574–579.

Clinciu, Miruna-Adriana, Arash Eshghi, and Helen Hastie. 2021. A study of automatic metrics for the evaluation of natural language explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2376–2387, Online. Association for Computational Linguistics.

Costa, Ângela, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. 2015. A linguistically motivated taxonomy for Machine Translation error analysis. *Mach. Transl.*, 29(2):127–161.

Crüwell, Sophia, Deborah Apthorp, Bradley J. Baker, Lincoln Colling, Malte Elson, Sandra J. Geiger, Sebastian Lobentanzer, Jean Monéger, Alex Patterson, D. Samuel Schwarzkopf, Mirela Zaneva, and Nicholas J. L. Brown. 2023. What's in a badge? a computational reproducibility investigation of the open data badge policy in one issue of psychological science. *Psychological Science*. PMID: 36730433 (First published online February 2, 2023; issue information not known yet).

Dale, Robert. 2020. Natural language generation: The commercial state of the art in 2020. *Natural Language Engineering*, 26(4):481–487.

Derczynski, Leon and Emily M. Bender. 2021. Towards Better Interdisciplinary Science: Learnings From COLING 2018. Technical report, IT University of Copenhagen. TR-2021-208. Available through: https://en.itu.dk/Research/Technical-Reports/Technical-Reports-Archive/2021/TR-2021-208.

Dodge, Jesse, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Dou, Yao, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 7250–7274, Dublin, Ireland.

Dušek, Ondřej, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of End-to-End Natural Language Generation: The E2E NLG challenge. *Comput. Speech Lang.*, 59:123–156.

Epstein, Ziv, Blakeley H. Payne, Judy Hanwen Shen, Abhimanyu Dubey, Bjarke Felbo, Matthew Groh, Nick Obradovich, Manuel Cebrián, and Iyad Rahwan. 2018. Closing the AI knowledge gap. *CoRR*, abs/1803.07233.

Evans, Roger, Paul Piwek, and Lynne Cahill. 2002. What is NLG? In *Proceedings of the International Natural Language Generation Conference*, pages 144–151, Harriman, New York, USA. Association for Computational Linguistics.

Fitrianie, Siska, Merijn Bruijnes, Deborah Richards, Amal Abdulrahman, and Willem-Paul Brinkman. 2019. What are We Measuring Anyway?: - A Literature Survey of Questionnaires Used in Studies Reported in the Intelligent Virtual Agent Conferences. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, IVA 2019*, pages 159–161, Paris, France.

Fitrianie, Siska, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. 2020. The 19 Unifying Questionnaire Constructs of Artificial Social Agents: An IVA Community Analysis. In *IVA '20: ACM International Conference on Intelligent Virtual Agents*, pages 21:1–21:8, Virtual Event, Scotland, UK.

Fowler, Floyd J. and Carol Cosenza. 2008. Writing effective questions. In *International Handbook of Survey Methodology*, pages 136–160. Lawrence Erlbaum Associates, New York, NY, US.

Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Gatt, Albert and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.*, 61:65–170.

Gehrmann, Sebastian, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Gehrmann, Sebastian, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *CoRR*, abs/2202.06935.

He, Jie, Bo Peng, Yi Liao, Qun Liu, and Deyi Xiong. 2021. TGEA: An error-annotated dataset and benchmark tasks for TextGeneration from pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6012–6025, Online. Association for Computational Linguistics.

Howcroft, David M., 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.

Huang, Kuan-Jung and Adrian Staub. 2021. Why do readers fail to notice word transpositions, omissions, and repetitions? A review of recent evidence and theory. *Lang. Linguistics Compass*, 15(7).

Huidrom, Rudali and Anya Belz. 2022. A survey of recent error annotation schemes for automatically generated text. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 383–398, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jakobsen, Terne Sasha Thorn and Anna Rogers. 2022. What Factors Should Paper-Reviewer Assignments Rely On? Community Perspectives on Issues and Ideals in Conference Peer-Review. *CoRR*, abs/2205.01005.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. *CoRR*, abs/2202.03629.

Kan, Min-Yen. 2018. "research fast and slow". Keynote presented at COLING 2018, Santa Fe, NM, USA. Slides available through http://bit.ly/kan-coling18.

Kidwell, Mallory C., Ljiljana B. Lazarević, Erica Baranski, Tom E. Hardwicke, Sarah Piechowski, Lina-Sophia Falkenberg, Curtis Kennett, Agnieszka Slowik, Carina Sonnleitner, Chelsey Hess-Holden, Timothy M. Errington, Susann Fiedler, and Brian A. Nosek. 2016. Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, 14(5):1–15.

Kiela, Douwe, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Läubli, Samuel, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A Set of Recommendations for Assessing Human-Machine Parity in Language Translation. *J. Artif. Intell. Res.*, 67:653–672.

van der Lee, Chris, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Comput. Speech Lang.*, 67:101151.

McLuhan, Marshall. 1964. *Understanding Media: The Extensions of Man*. McGraw-Hill. ISBN 81-14-67535-7.

Mille, Simon, Anya Belz, Bernd Bohnet, Thiago Castro Ferreira, Yvette Graham, and Leo Wanner. 2020. The third multilingual surface realisation shared task (SR'20): Overview and evaluation results. In *Proceedings of the Third Workshop on Multilingual Surface Realisation*, pages 1–20, Barcelona, Spain (Online). Association for Computational Linguistics.

Mille, Simon, Kaustubh D. Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Gangal, Mihir Kale, Emiel van Miltenburg, and Sebastian Gehrmann. 2021. Automatic Construction of Evaluation Suites for Natural Language Generation Datasets. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*.

van Miltenburg, Emiel, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021a. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

van Miltenburg, Emiel, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 21–30, Santiago de Compostela, Spain. Association for Computational Linguistics.

van Miltenburg, Emiel, Chris van der Lee, Thiago Castro-Ferreira, and Emiel Krahmer. 2020a. Evaluation rules! on the use of grammars and rule-based systems for NLG evaluation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 17–27, Online (Dublin, Ireland). Association for Computational Linguistics.

van Miltenburg, Emiel, Chris van der Lee, and Emiel Krahmer. 2021b. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, Online. Association for Computational Linguistics.

van Miltenburg, Emiel, Wei-Ting Lu, Emiel Krahmer, Albert Gatt, Guanyi Chen, Lin Li, and Kees van Deemter. 2020b. Gradations of error severity in automatic image descriptions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 398–411, Dublin, Ireland. Association for Computational Linguistics.

Munafò, Marcus R, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John Ioannidis. 2017. A Manifesto for Reproducible Science. *Nature human behaviour*, 1(1):1–9.

Nauta, Meike, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2022. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *CoRR*, abs/2201.08164.

Novikova, Jekaterina, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Popović, Maja. 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.

Popović, Maja. 2021. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online. Association for Computational Linguistics.

Popović, Maja and Anya Belz. 2022. On reporting scores and agreement for error annotation tasks. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 306–315, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rahal, Rima-Maria, Susann Fiedler, Adeyemi Adetula, Ronnie P. A. Berntsson, Ulrich Dirnagl, Gordon B. Feld, Christian J. Fiebach, Samsad Afrin Himi, Aidan J. Horner, Tina B. Lonsdorf, Felix Schönbrodt, Miguel Alejandro A. Silan, Michael Wenzler, and Flávio Azevedo. 2023. Quality research needs good working conditions. *Nature Human Behaviour*.

Raji, Inioluwa Deborah, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*.

Raji, Inioluwa Deborah, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of ai functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 959–972, New York, NY, USA. Association for Computing Machinery.

Reiter, Ehud and Anja Belz. 2009. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Comput. Linguistics*, 35(4):529–558.

Rogers, Anna and Isabelle Augenstein. 2020. What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262, Online. Association for Computational Linguistics.

Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Schlangen, David. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.

Sculley, D., Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Shimorina, Anastasia 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.

Singh, Gerald G, Jordan Tam, Thomas D Sisk, Sarah C Klain, Megan E Mach, Rebecca G Martone, and Kai MA Chan. 2014. A More Social Science: Barriers and Incentives for Scientists Engaging in Policy. *Frontiers in Ecology and the Environment*, 12(3):161–166.

Sripada, Somayajulu G., Ehud Reiter, Ian Davy, and Kristian Nilssen. 2004. Lessons from deploying nlg technology for marine weather forecast text generation. In *Proceedings of the 16th European Conference on Artificial Intelligence*, ECAI'04, page 760–764, NLD. IOS Press.

Strauss, Anselm and Juliet Corbin. 1994. Grounded Theory Methodology: An Overview. In N. K. Denzin and Y. S. Lincoln, editors, *Handbook of qualitative research*, pages 273–285. Sage Publications, Inc.

Terry, Gareth, Nikki Hayfield, Victoria Clarke, and Virginia Braun. 2017. Chapter 2: Thematic Analysis. In C. Willig and W. Rogers, editors, *The SAGE Handbook of Qualitative Research in Psychology*, pages 17–36. SAGE Publications Ltd.

Thomson, Craig and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

Thomson, Craig and Ehud Reiter. 2021. Generation challenges: Results of the accuracy evaluation shared task. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 240–248, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Thomson, Craig, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Computer Speech & Language*, 80:101482.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Waltman, Ludo, Wolfgang Kaltenbrunner, Stephen Pinfield, and Helen B Woods. 2022. How to Improve Scientific Peer Review: Four Schools of Thought.

Yarkoni, Tal. 2018. No, It's Not the Incentives, It's You. Published on *[citation needed]*, personal blog of Tal Yarkoni.

Zhou, Kaitlyn, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications. *CoRR*, abs/2205.06828.

A Information letter

What is this study about?

This research project aims to understand the status of error analysis in NLG. We aim to answer three questions:

- What do researchers think about error analysis?
- In what circumstances are researchers willing and able to carry out an error analysis?
- What are the barriers to carrying out an error analysis?

This study builds on an earlier position paper about error analysis, which shows that relatively few NLG papers provide an error analysis, and which provide a how-to guide for carrying out error analyses. You can read the paper here.

What does participating in the study entail?

For this study, we ask you to answer a short series of questions. We expect this to take about 10 minutes. Most of these questions are multiple-choice, but there are also some open questions. Your answers will be completely anonymous, and it is impossible for us to trace back the answers to you.

Disadvantages, consequences & risks

- You will be asked to answer a series of questions, which takes time. We tried to make the questionnaire as short as possible, so as to minimise any possible inconvenience.
- Although we tried to prevent any question from offending any participants, it may still be the case that you take offense to some of the questions. In this case, feel free to leave a comment at the end of the survey, or to contact either us or the ethics committee directly. Contact details are at the bottom of this page.
- Some questions might be controversial. We record minimal personal information, so that you are free to speak your mind, without any consequences. The only personal information we collect is whether you work in industry or in academia, and how experienced you are.
- We do not foresee any other risks connected to your taking part in this study.

Advantages

There are no direct advantages to taking part in this study. The indirect advantage is that your contribution will help us understand how NLG researchers feel about error analysis, and we aim to publish a full report through one of the many open-access venues in our field (e.g. INLG).

Rights

Under the main applicant's University's code of ethics, you are entitled to a number of rights:

- Your participation is completely voluntary, and you have the right to decline to participate and withdraw from the research once participation has begun, without any negative consequences, and without providing any explanation.
- You have the right, in principle, to request access to and rectification, erasure, restriction of or object to the processing of personal data. For more information, please see: URL. Do note that, because all data is fully anonymised, it may be impossible for us to delete or alter your responses.
- Your participation is fully confidential, meaning that your answers will be fully anonymised. We have configured Qualtrics such that it will also not collect your IP address.
- Your consent to participate only lasts for the duration of the study, and may be withdrawn at any time.

What does consent mean?

By consenting, you indicate that you are voluntarily taking part in this study, and that you allow for your data to be processed. This means that:

- You agree that your answers may be used to publish a research article on this topic.
- The data will be stored on the computers of the research team, with both local (hard drive) and online (protected cloud drive) backups.
- The data will be made public upon completion of this study.
- You acknowledge that there is no financial compensation for taking part in this study.

The actual consent form is on the next page.

Contact details

This study has been approved by the Research Ethics and Data Management Committee (REDC) of the DEPARTMENT. If you have any questions about this study, you may contact the principal investigator via email: EMAIL. If you have any remarks or complaints regarding this research, you may also contact the REDC via: EMAIL.

Full list of the researchers involved: NAMES

B Informed consent form

This is the consent form for our study about the status of error analysis in NLG. Full details about this study were provided on the previous page. If you want to read this information again, you can go back to the previous page. If anything is still unclear about this study, please contact: EMAIL.

Consent

By consenting, you indicate that you have read the description on the previous page, that you are voluntarily taking part in this study, and that you allow for your data to be processed. This means that:

- You agree to your responses being anonymously recorded.
- Your answers will be used to study the status of error analysis in NLG, and may be used in future publications pertaining to this topic.
- The data will be shared with our research team, with both local (hard drive) and online (protected cloud drive) backups. This data will be stored indefinitely, and made public upon completion of our research. Note again that none of your answers can be traced back to you.
- You acknowledge that there is no financial compensation for taking part in this study.

Note that you may still withdraw your consent after completing this form, without any negative consequences. We will delete all incomplete forms from our study.

Do you consent?

Do you agree to take part in this study? If you consent, please indicate this below by clicking "Yes". If you click "No", you will be directed to the end of this questionnaire. You may also close this page to stop participating in this study.

○ Yes, I consent.
○ No, I do not consent.

C Survey questions

These are all the questions we have asked our participants to answer. Due to the display logic, participants always see a subset of the questions, based on their earlier answers. We have reproduced this display logic below with conditional statements (*if * was selected for question *:*). If the statement is true, then the question immediately following the statement is displayed. Otherwise, questions with false conditionals are hidden.

Start of survey

1. Are you in academia or in industry? (If you have a dual affiliation, please respond with your dominant affiliation in mind.)

○ Academia
○ Industry
○ Other

2. How many years have you been working in NLG?

○ Less than 2 years
○ 2-5 years
○ 6-10 years
○ 11 or more years
○ I don't work in NLG

Definition of "error analysis"
Before continuing, we need to agree on the definition of error analysis. For the purposes of this questionnaire:

- We define "error analysis" as a formalised procedure (similar to annotation) in which errors in the output of an NLG system are identified and categorised, after which the frequencies for the different kinds of errors are reported.
- Error analyses are different from "error mentions", which give an impression of the kinds of errors that are made by an NLG system, but are less formal and don't quantify the distribution of errors.

Example
Below is an excerpt from Table 3 of Barros & Lloret (2015, ENLG). The authors "manually analysed all the generated sentences and classified these errors attending to frequent grammatical errors and frequent drafting errors." The table shows how often each type of error occurs in their data.

Error types	Number of sentences
Grammatical concordance: Nominal	2
Verbal	7
Non words semantic relations	36
Missing main verb	7
Incorrect syntactic order	38

3. Do you remember reading any NLG papers that include an error analysis?

○ Yes
○ No

If positive answer to question 3:
4. Did you find the error analyses to be useful?

○ Not at all useful
○ Slightly useful

○ Moderately useful
○ Very useful
○ Extremely useful

If *not at all useful* was *not* selected for question 4:
5. What did you find useful about the error analyses you've seen?
(Open question)

If *not at all useful* was selected for question 4:
6. Why didn't you find the error analyses to be useful?
(Open question)

If negative answer to question 3:
7. Is it surprising to you that you haven't seen any published error analyses?

○ Yes, because …
○ No, because …

8. Have you ever carried out an error analysis?

○ Yes
○ No

If positive answer to question 8:
9. What did you find challenging or difficult about carrying out an error analysis?
(Open question)

If positive answer to question 8:
10. Did you feel like there were enough resources/reference material for you to carry out an error analysis?

○ Yes
○ No

If positive answer to question 8:
11. Do you think you'll carry out an error analysis again in the future?

○ Definitely not
○ Probably not
○ Might or might not
○ Probably yes
○ Definitely yes

If positive answer to question 8:
12. Could you explain your answer to the previous question?
(Open question)

If negative answer to question 8:
13. Have you ever considered carrying out an error analysis?

○ Never
○ Once or twice
○ Regularly
○ I'm planning to carry out an error analysis in the future

If negative answer to question 8:
14. What is the reason you haven't carried out an error analysis?
(Open question)

If negative answer to question 8:
15. Are you willing to carry out an error analysis?

○ Definitely not
○ Probably not
○ Might or might not
○ Probably yes
○ Definitely yes

16. For what kinds of papers do you think error analyses may be useful?
(Open question)

17. I would be more likely to carry out an analysis in a conference/journal paper if…
(Closed question with multiple statements. Answer options: Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree)

- There was a higher page limit.
- There would be an existing error taxonomy that I could use.
- There would be dedicated annotation tools for error analysis that I could use.
- There would be a crowdsourcing template for carrying out error analyses.
- Reviewers paid more attention to error analyses.
- There were an available pool of annotators or crowd workers
- I had more time.
- I had more money.
- I had more collaborators.

18. Are there any other barriers that prevent you from carrying out an error analysis?
(Open question)

19. Please indicate whether you agree or disagree with the following statements
(Closed question with multiple statements. Answer options: Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree)

- There should be more error analyses in the NLG literature
- Error analyses are a valuable part of a paper.
- Carrying out an error analysis is enjoyable.
- Carrying out an error analysis is boring/tedious.

• Error analyses are necessary to fully evaluate the performance of an NLG system.
• Knowing what errors a system makes is helpful for future research.
• Knowing what errors a system makes is helpful for practitioners/NLG in industry.
• If you publish at a **conference**, and you present an NLG system as one of your main contributions, you should include an error analysis.
• If you publish in a **journal**, and you present an NLG system as one of your main contributions, you should include an error analysis.

20. I am . . . likely to include an error analysis in a journal article than/as I would be for a conference publication.

○ More
○ Less
○ Equally

21. Please explain your answer to the previous question
(Open question)

22. Are there currently enough resources to support error analysis?

○ Yes
○ No, I am still missing: . . .

23. Besides resources, are there any other factors that would make it more likely for you to carry out an error analysis?
(Open question)

We believe that it is essential for authors of error analyses to include a table with the distribution of errors in the output of their system. This data should be based on a formalised annotation procedure, with at least two annotators, so that the paper can also report inter-annotator agreement to gauge the reliability of the analysis.

24. What else would you recommend that authors should include in an error analysis?
(Open question)

25. This is the final question. Is there anything you would like to add or comment on?