Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions

David M. Howcroft[1] ✉, **Anya Belz**[2], **Miruna Clinciu**[1], **Dimitra Gkatzia**[3], **Sadid A. Hasan**[4], **Saad Mahamood**[5], **Simon Mille**[6], **Emiel van Miltenburg**[7], **Sashank Santhanam**[8], and **Verena Rieser**[1]

[1]The Interaction Lab, MACS, Heriot-Watt University, Edinburgh, Scotland, UK
[2]University of Brighton, Brighton, England, UK
[3]Edinburgh Napier University, Edinburgh, Scotland, UK
[4]CVS Health, Wellesley, MA, USA
[5]trivago N.V., Düsseldorf, Germany
[6]Universitat Pompeu Fabra, Barcelona, Spain
[7]Tilburg Center for Cognition & Communication, Tilburg University, Tilburg, Netherlands
[8]Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA
✉ Corresponding author: D.Howcroft@hw.ac.uk

Abstract

Human assessment remains the most trusted form of evaluation in NLG, but highly diverse approaches and a proliferation of different quality criteria used by researchers make it difficult to compare results and draw conclusions across papers, with adverse implications for meta-evaluation and reproducibility. In this paper, we present (i) our dataset of 165 NLG papers with human evaluations, (ii) the annotation scheme we developed to label the papers for different aspects of evaluations, (iii) quantitative analyses of the annotations, and (iv) a set of recommendations for improving standards in evaluation reporting. We use the annotations as a basis for examining information included in evaluation reports, and levels of consistency in approaches, experimental design and terminology, focusing in particular on the 200+ different terms that have been used for evaluated aspects of quality. We conclude that due to a pervasive lack of clarity in reports and extreme diversity in approaches, human evaluation in NLG presents as extremely confused in 2020, and that the field is in urgent need of standard methods and terminology.

1 Introduction

Evaluating natural language generation (NLG) systems is notoriously complex: the same input can be expressed in a variety of output texts, each valid in its own context, making evaluation with automatic metrics far more challenging than in other NLP contexts (Novikova et al., 2017; Reiter and Belz, 2009). Human evaluations are commonly viewed as a more reliable way to evaluate NLG systems (Celikyilmaz et al., 2020; Gatt and Krahmer, 2018), but come with their own issues, such as cost and time involved, the need for domain expertise (Celikyilmaz et al., 2020), and the fact that the experimental setup has a substantial impact on the reliability of human quality judgements (Novikova et al., 2018; Santhanam and Shaikh, 2019).

Moreover, there is little consensus about how human evaluations should be designed and reported. Methods employed and details reported vary widely, issues including missing details (e.g. number of evaluators, outputs evaluated, and ratings collected), lack of proper analysis of results obtained (e.g. effect size and statistical significance), and much variation in names and definitions of evaluated aspects of output quality (van der Lee et al., 2019; Amidei et al., 2018). However, we currently lack a complete picture of the prevailing consensus, or lack thereof, regarding approaches to human evaluation, experimental design and terminology.

Our goal in this work, therefore, is to investigate the extent of the above issues and provide a clear picture of the human evaluations NLG currently employs, how they are reported, and in what respects they are in need of improvement. To this end, we examined 20 years of NLG papers that reported some form of human evaluation, capturing key information about the systems, the quality criteria employed, and how these criteria were operationalised in specific experimental designs.

The primary contributions of this paper are (1) an annotation scheme and guidelines for identifying characteristics of human evaluations reported in NLG papers; (2) a dataset containing all 165 INLG/ENLG papers with some form of human evaluation published in 2000–2019, annotated with

Proceedings of The 13th International Conference on Natural Language Generation, pages 169–182,
Dublin, Ireland, 15-18 December, 2020.

the scheme, and intended to facilitate future research on this topic; (3) analyses of our dataset and annotations, including analysis of quality criteria used in evaluations, and the similarities and differences between them; and (4) a set of recommendations to help improve clarity in reporting evaluation details.

2 Paper Selection

We selected papers for inclusion in this study following the PRISMA methodology (Moher et al., 2009) recently introduced to NLP by Reiter (2018) in his structured review of the validity of BLEU.

As summarised in Table 1, we began by considering all 578 papers published at the main SIGGEN venue(s): the International Natural Language Generation Conference (INLG) and the European Workshop on Natural Language Generation (ENLG), which were merged in 2016.

While many papers on NLG are published in other venues, including the *ACL conferences, EMNLP, AAAI, IJCAI, etc., focusing on INLG and ENLG provides a simple selection criterion which at the same time ensures a set of papers representative of what researchers specialising in NLG were doing across this time period. We screened the 578 papers looking for mention of a human evaluation, first by skimming for relevant section headings and then by searching in the PDFs for 'human', 'subject', and 'eval'. This left 217 papers.

During annotation (Section 3), we retained only papers that reported a human evaluation in the following sense: an experiment involving assessment of system outputs in terms of an explicitly or implicitly given quality criterion, either via (1) conscious assessment of outputs in terms of the criterion by evaluators (e.g. (dis)agreement with quality statement, direct and relative assessment, qualitative feedback); or (2) counts and other measurements of outputs and user interactions with them (e.g. user-text and user-system interaction measurements, task performance measurements).

We decided to allow evaluations matching the above conditions even if they did not evaluate system generated texts. This allowed the inclusion of papers which, e.g., assess wizard-of-oz or corpus texts to inform the design of an NLG system.

Figure 1 shows the distribution of the 165 papers meeting these conditions across publication years. The general increase of papers with human evaluations since 2012 aligns with the evaluation

Stage	Source	Count
1	INLG / ENLG papers 2000–2019	578
2	Likely with human evaluations	217
3	Confirmed human evals (full dataset)	165

Table 1: Number of papers at each selection stage.

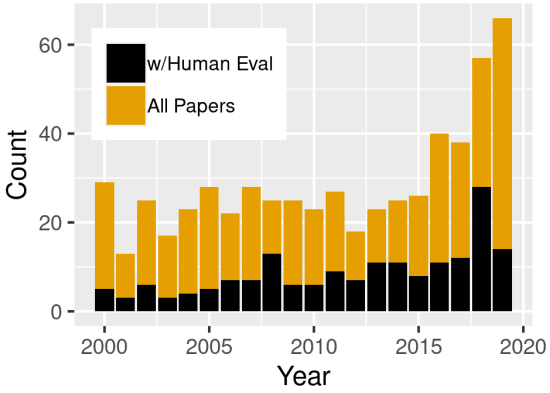

Figure 1: Number of INLG/ENLG papers per year with human evaluation (black) and overall (full bar).

trends found by Gkatzia and Mahamood (2015), who also reported an increase in the proportion of papers with intrinsic human evaluations between 2012–2015 compared to 2005–2008. However, only 28.54% of the papers in our sample contained a human evaluation compared to 45.4% reported by Gkatzia and Mahamood (2015).

3 Paper Annotation

In order to quantitatively study the evaluations in our dataset, we needed a systematic way of collecting information about different aspects of evaluations. Therefore, we developed an annotation scheme to capture different characteristics of evaluations, allowing us to investigate how human evaluations have been designed and reported in NLG over the past two decades, in particular what conventions, similarities and differences have emerged.

Below, we summarise our approach to studying aspects of quality assessed in evaluations (Section 3.1), present the final annotation scheme (Section 3.2), describe how we developed it (Section 3.3), and assessed inter-annotator agreement (IAA) (Section 3.4).[1]

[1]The dataset of annotated PDFs, annotation spreadsheet, annotation scheme, code, and guidelines resulting from the work are available in the project repository: https://evalgenchal.github.io/20Y-CHEC/

3.1 Aspects of quality

Researchers use the same term to describe the aspect of quality they are evaluating with sometimes very different meaning. Annotating (and later analysing) only such terms as are used in our papers would have restricted us to reporting occurrences of the terms, without any idea of where the same thing was in fact evaluated. We would not have been able to report even that, say, Readability is the *n*th most frequently evaluated aspect of quality, because not all papers in which Readability results are reported mean the same thing by it.

We wanted to be able to quantitatively study both usage of terms such as Readability, and the meanings associated with them in different papers. Side-stepping the question of whether there is a single, 'true' concept of say Readability that evaluations could aim to assess, we simply tried to determine, on the basis of *all* the information provided in a paper, which sets of evaluations assessed aspects of quality similar enough to be considered the same (see Section 3.2.2). This resulted in similarity groups which we assigned normalised names to, yielding a set of common-denominator terms for the distinct aspects of quality that were assessed, regardless of what terms authors used for them.

Below we refer to evaluated aspects of quality as **quality criteria** and the terms used to refer to different criteria as **quality criteria names**. Any name and definition capturing an aspect of quality can be a quality criterion. We do not wish to imply that there exists a set of 'true' quality criteria, and leave open in this paper the question of how such quality criteria relate to constructs with similar names researched in other fields such as linguistics and psycholinguistics.

3.2 Annotation scheme

The annotation scheme consists of seven closed-class and nine open-class attributes that capture different aspects of human evaluation methods and fall into three categories: (1) four *System* attributes which describe evaluated NLG systems, (2) four *Quality criterion* attributes which describe the aspect(s) of quality assessed in evaluations, and (3) eight *Operationalisation* attributes which describe how evaluations are implemented. Definitions and examples for all attributes can be found in the annotation guidelines in the Supplementary Material.

3.2.1 System attributes

The four attributes in this category cover the following properties of systems: ***language*** (as per ISO 639-3 (2019)), ***system input*** and ***system output*** (*raw/structured data*, *deep* and *shallow linguistic representation*, different types of text (*sentence*, *documents* etc.)), and ***task*** (e.g. *data-to-text generation*, *dialogue turn generation*, *summarisation*).

The most challenging aspect of selecting values for the system attributes was the lack of clarity in many papers about inputs/outputs. Where the information was clearly provided, in some cases it proved difficult to decide which of two adjacent attribute values to select; e.g. for ***system output***, single vs. multiple sentences, and for ***system input***, structured data vs. deep linguistic representation.

3.2.2 Quality criterion attributes

The attributes in this category are ***verbatim criterion name*** and ***verbatim criterion definition*** (both as found in the paper), ***normalised criterion name*** (see below), and ***paraphrased criterion definition*** (capturing the annotator's best approximation of what was really evaluated in the paper).

As mentioned above, to make it possible to report both on usage of quality criterion names, and on similarities and differences between what was really evaluated, we devised a set of normalised quality criterion names that would allow us to see how many distinct quality criteria are currently being used, and relate these to results from our other analyses. The normalised criterion names were determined by performing bottom-up clustering and renaming of values selected for the attributes ***verbatim criterion definition***, ***paraphrased criterion definition***, ***verbatim question/prompt*** and ***paraphrased question/prompt*** (see Section 3.2.3).

We counted 478 occurrences of (verbatim) quality criterion names in papers, mapping to 204 unique names. The clustering and renaming process above produced 71 criterion names which we consider truly distinct and which represent our set of normalised quality criteria. This means that in our analysis, 71 distinct evaluation criteria have been used in the last 20 years in NLG, not 204.

Some of the normalised criteria are less specific than others, and can be further specified to yield one of the other criteria, implying hierarchical relationships between some criteria. For example, a criterion might measure the overall *Correctness of the Surface Form* of a text (less specific), or it might more specifically measure its *Grammatical-*

ity or *Spelling Accuracy*. Using the classification system for human evaluations proposed by Belz et al. (2020) to provide the top two levels and some branching factors, we developed the hierarchical relationships between quality criteria into a taxonomy to help annotators select values (Appendix E). The set of normalised quality criteria names and definitions is provided in Appendix D.

Common issues we encountered in selecting values for the ***normalised quality criterion*** attribute were underspecified or unclear quality criterion definitions in papers, missing definitions (279 out of 478), missing prompts/questions for the evaluators (311/478), and missing criterion names (98/478). The more of this is missing in a paper, the more difficult it is to see beyond the information provided by authors to form a view of what is actually being evaluated, hence to choose a value for the normalised criterion name attribute.

3.2.3 Operationalisation attributes

The eight attributes in this category record different aspects of how responses are collected in evaluations: the ***form of response elicitation*** (*direct*, vs. *relative quality estimation*, *(dis)agreement with quality statement*, etc.), the ***verbatim question/prompt*** used in the evaluation and included in the paper, a ***paraphrased question/prompt*** for those cases where the paper does not provide the verbatim question/prompt, the ***data type of the collected responses*** (*categorical*, *rank order*, *count*, *ordinal*, etc.), the ***type of rating instrument*** from which response variable values are chosen (*numerical rating scale*, *slider scale*, *verbal descriptor scale*, *Likert scale*, etc.), the ***size of rating instrument*** (number of possible response values), the ***range of response values*** and any ***statistics*** computed for response values.

We found that for most papers, determining the type and size of scale or rating instrument is straightforward, but the large majority of papers do not provide details about the instructions, questions or prompts shown to evaluators; this was doubly problematic because we often relied on such information to determine what was being evaluated.

3.3 Annotation scheme development

The annotation scheme was developed in four phases, resulting in four versions of the annotations with two IAA tests (for details of which see Section 3.4), once between the second and third version of the scheme, and once between the third and fourth. From each phase to the next, we tested and subsequently improved the annotation scheme and guidelines. Annotations in all versions were carried out by the first nine authors, in roughly equal proportions.

	1st	2nd IAA test		
attribute	5 solo	9 solo	4 duo	5 best
input	0.38	0.36	0.38	0.31
output	0.43	0.17	0.28	0.30
task	0.17	0.50	0.53	0.71
elicit. form	0.11	0.27	0.47	0.24
data type	0.24	0.52	0.64	0.73
instrument	0.08	0.39	0.51	0.64
criterion	0.20	0.12	0.19	0.25

Table 2: Krippendorff's alpha with Jaccard for closed-class attributes in the 1^{st} and 2^{nd} IAA tests. Numbers are not directly comparable (a) between the two tests due to changes in the annotation scheme; (b) within the 2^{nd} test due to different numbers of annotators.

In the first phase, most of the 165 papers in our final dataset (Table 1) were annotated and then double-checked by two different annotators using a first version of the annotation scheme that did not have formal guidelines.

The double-checking revealed considerable differences between annotators, prompting us to formalise the annotation scheme and create detailed instructions, yielding Version 1.0 of the annotation guidelines. IAA tests on new annotations carried out with these guidelines revealed low agreement among annotators (see Table 2, 1^{st} IAA test), in particular for some of the attributes we were most interested in, including ***system task***, ***type of rating instrument***, and ***normalised quality criterion***.

We therefore revised the annotation scheme once more, reducing the number of free-text attributes, and introducing automated consistency checking and attribute value suggestions. Using the resulting V2.0 scheme and guidelines, we re-annotated 80 of the papers, this time pairing up annotators for the purpose of agreeing consensus annotations. We computed, and Table 2 reports, three sets of IAA scores on the V2.0 annotations: for all nine annotators separately ('9 solo'), for the 4 consensus annotations ('4 duo'), and for the 5 annotators whose solo annotations agreed most with everyone else's, shown in the '5 best' column. There was an overall improvement in agreement (substantial in the case of some attributes), but we decided to carry out one final set of improvements to definitions and instructions in the annotation guidelines

(with minimal changes to attribute names and values), yielding version 2.1 which was then used for the final annotation of all 165 papers in our dataset, on which all analyses in this paper are based.

3.4 Inter-Annotator Agreement

Papers for IAA tests: For each IAA test we manually selected a different arbitrary set of 10 NLG papers with human evaluations from ACL 2020.

Preprocessing: We cleaned up attribute values selected by annotators by normalising spelling, punctuation, and capitalisation. For the first annotation round which allowed empty cells, we replaced those with 'blank.' We also removed papers not meeting the conditions from Section 2.

Calculating agreement: The data resulting from annotation was a 10 (papers) $\times$ n (quality criteria identified by annotator in paper) $\times$ 16 (attribute value pairs) data frame, for each of the annotators. The task for IAA assessment was to measure the agreement across multiple data frames (one for each annotator) allowing for different numbers of criteria being identified by different authors.

We did this by calculating Krippendorff's alpha using Jaccard for the distance measure (recommended by Artstein and Poesio 2008). Scores for the seven closed-class attributes are shown in Table 2 for each of the two IAA tests (column headings as explained in the preceding section).

The consensus annotations ('duo') required pairs of annotators to reach agreement about selected attribute values. This reduced disagreement and improved consistency with the guidelines, the time it took was prohibitive.

For the attributes ***task***, ***data type***, and ***type of rating instrument*** (shortened to 'instrument' in the table), we consider the '5 best' IAA to be very good (0 indicating chance-level agreement). For ***system input*** and ***output***, IAA is still good, with the main source of disagreement the lack of clarity about text size/type in textual inputs/outputs. Replacing the different text size/type values with a single 'text' value improves IAA to 0.41 and 1.00 for inputs and outputs, respectively. The remaining issues for inputs are to do with multiple inputs and distinguishing structured data from deep linguistic representations, which prompted us to merge the two data input types.

Low agreement for normalised quality criteria is in part due to the lack of clear information about what aspect of quality is being assessed in papers, and the difficulty of distinguishing quality criteria from evaluation modes (see previous section). But cases where annotators mapped a single criterion name in the paper to multiple normalised criterion names were also a big factor because this substantially raises the bar for agreement.

4 Analysis and Results

In this section, we present results from analyses performed on the annotations of the 165 papers in our dataset. The dataset and code for analysis are available in the project repository.

The 165 papers in the dataset correspond to 478 individual evaluations assessing single quality criteria, i.e. 2.8 per paper. For the quality criterion attributes (Section 3.2.2) and the operationalisation attributes (Section 3.2.3) it makes most sense to compute occurrence counts on the 478 individual evaluations, even if that slightly inflates counts in some cases. For example, if multiple criteria are evaluated in the same experiment, should we really count multiple occurrences for every operationalisation attribute? But the alternatives are to either count per paper, leaving the question of what to do about multiple experiments in the same paper, or to count per experiment, leaving the problem of variation within the same experiment and also that it is not always clear whether separate experiments were carried out. For these reasons we opted to compute statistics at the individual-evaluation level for the quality-criterion and operationalisation attributes.

For the system attributes (Section 3.2.1), we report paper-level statistics. We do sometimes find more than one system type (with different language, input, output or task) being evaluated in a paper, but for those cases we add all attribute values found for the paper. Below we first report paper-level statistics for the system attributes (Section 4.1), followed by evaluation-level statistics for quality-criterion and operationalisation attributes (Section 4.2).

4.1 Paper-level statistics

Unsurprisingly, our analysis shows that the most frequent system language in our dataset is English, accounting for 82.14% of papers pre-2010, and 75.39% post-2010. Appendix A provides a detailed overview of results for this attribute.

In terms of the system task attribute, our analysis reveals that before 2010, *data-to-text generation* and *dialogue turn generation* were the most

Form	Count
direct quality estimation	207
relative quality estimation	72
(dis)agreement with quality statement	48
classification	38
task performance measurements	35
qualitative feedback	20
evaluation through post-editing/annotation	18
unclear	15
user-system interaction measurements	10
counting occurrences in text	8
user-text interaction measurements	6
other	1

Table 3: Counts of values selected for ***form of response elicitation***.

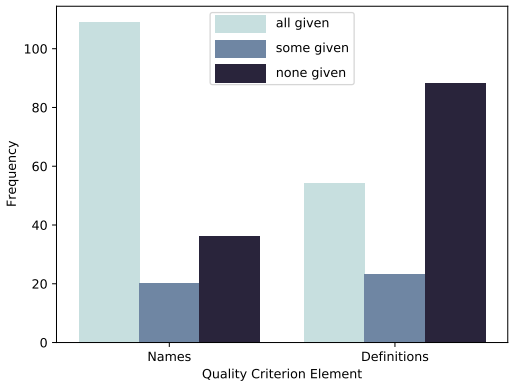

Figure 2: How many papers explicitly name and define all, some, or none of the quality criteria they evaluate.

common tasks, whereas post-2010 the most common tasks are *data-to-text generation*, *summarisation* and *dialogue turn generation*. The biggest increases are for question generation (0 pre-2010, 9 post-2010), end-to-end generation (1 increasing to 8), and summarisation (1 going up to 11).[2] For the system output attribute, we found that a big majority of systems output single or multiple sentences. Appendix B and C show task and output frequencies in more detail.

4.2 Evaluation-level statistics

4.2.1 Operationalisation attributes

Table 3 provides an overview of the most frequent values selected for the ***form of response elicitation*** attribute. We found that *direct quality estimation* where outputs are scored directly one at a time, was most common (207 times), followed by *relative quality estimation* where multiple outputs are ranked (72 times).[3]

To select values for this criterion, we relied on a combination of descriptions of the general experimental design, prompts/questions and instructions given to evaluators. We found that instructions to evaluators were almost never provided, example prompts/questions rarely, and even details of rating scales etc. were often missing.

What was usually clear was the ***type of scale or other rating instrument*** and its size and labels. From this, values for other operationalisation attributes such as ***form of response elicitation***, ***data type of collected responses*** and ***range of response values*** could usually be deduced, but as can be seen from Table 3, for 15 individual evaluations (5 papers) even the response elicitation methods were unclear.

4.2.2 Quality Criterion Names & Definitions

In this section, our aim is to look at the criterion names and definitions as given in papers, and how they mapped to the normalised criterion names. As shown in Figure 2 at the paper level, not all papers name their quality criteria and worryingly, just over half give no definitions for any of their quality criteria. As noted in Section 3, where explicit criterion names and/or definitions were missing in papers, we used the remaining information provided in the paper to determine which aspect of quality was evaluated, and mapped this to our set of normalised quality criteria.

Table 4 shows how often each normalised criterion occurs in our annotations of the 478 individual evaluations in the dataset. We can see that *Usefulness for task/information need, Grammaticality*, and *Quality of outputs* are the most frequently occurring normalised quality criterion names. *Fluency* which is one of the most frequent criterion names found in papers, ranks only (joint) seventh.

Table 5 shows 10 example criterion names as used in papers, and how we mapped them to our normalised criterion names. For example, Fluency was mapped to 15 different (sets of) normalised names (reflecting what was actually evaluated), including many cases where multiple normalised criterion names were selected (indicated by the prefix '*multiple (n)*').

It is not straightforward to interpret the information presented in Table 5. Objectively, what it shows is that we chose a much larger number of quality criteria to map certain original quality

[2]The increase in summarisation may be due to an increase in summarisation papers submitted to INLG, the increase in end-to-end generation in part to changing terminology.

[3]For explanations of attribute values see annotation guidelines in Supplementary Material.

Criterion Paraphrase	Count
usefulness for task/information need	39
grammaticality	39
quality of outputs	35
understandability	30
correctness of outputs relative to input (content)	29
goodness of outputs relative to input (content)	27
clarity	17
fluency	17
goodness of outputs in their own right	14
readability	14
information content of outputs	14
goodness of outputs in their own right (both form and content)	13
referent resolvability	11
usefulness (nonspecific)	11
appropriateness (content)	10
naturalness	10
user satisfaction	10
wellorderedness	10
correctness of outputs in their own right (form)	9
correctness of outputs relative to external frame of reference (content)	8
ease of communication	7
humanlikeness	7
appropriateness	6
understandability	6
nonredundancy (content)	6
goodness of outputs relative to system use	5
appropriateness (both form and content)	5

Table 4: Occurrence counts for normalised criterion names.

criteria names to than others. Fluency has been mapped to by far the largest number of different normalised criteria. This in turn means that there was the largest amount of variation in how different authors defined and operationalised Fluency (because we determined the normalised criteria on the basis of similarity groups of original criteria). In other words, the papers that used Fluency divided into 15 subsets each with a distinct understanding of Fluency shared by members of the subset. 15 is a large number in this context and indicates a high level of disagreement, in particular combined with the presence of many *multiple* sets.

Conversely, a criterion like *Clarity* has a high level of agreement (despite also being high frequency as shown in Table 4). Figure 3 shows a graphical representation of some of our mappings from original to normalised quality criteria in the form of a Sankey diagram, and illustrates the complexity of the correspondences between the two.

4.2.3 Prompts/questions put to evaluators

Prompts and questions put to evaluators (e.g. *how well does this text read?*) often try to explain the aspect of quality that evaluators are supposed to

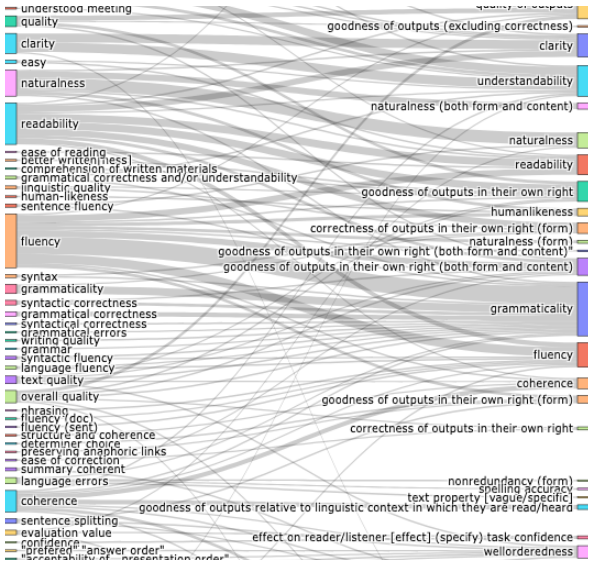

Figure 3: Part of Sankey diagram of evaluation criteria names from NLG papers between 2000 & 2019 (left) mapped to normalised criteria names representing our assessment of what was actually measured (right).

be evaluating using descriptors other than the criterion name, and can end up explaining one criterion in terms of one or more others (e.g. for Fluency, *how grammatical and readable is this text?*). We found fifty cases where the prompt/question references multiple normalised criteria (two and more), with a mean of 2.48 (min = 2, max = 4, median = 2, stdev = 0.64). Table 6 lists pairs of criteria referenced in the same prompt/question, ordered by pair-level frequency. For example, there were four prompts/questions that referenced both Fluency and Grammaticality. There is evidence that questions combining multiple quality criteria cause more variation in the responses, because different participants may weigh the importance of one of the quality criteria differently in their response; such complex quality criteria may best be measured using multiple items rather than a single question (van der Lee et al., 2019).

5 Discussion & Recommendations

Perhaps the most compelling evidence we found in our analyses in this paper is that (i) there is very little shared practice in human evaluation in NLG, in particular with respect to what to name the aspects of quality we wish to evaluate, and how to define them; and (ii) the information presented in NLG papers about human evaluations is very rarely complete. The latter can be addressed through better reporting in future work (see below). The

ORIGINAL CRITERION	MAPPED TO NORMALISED CRITERIA	Count
fluency	fluency; goodness of outputs in their own right; goodness of outputs in their own right (form); goodness of outputs in their own right (both form and content; grammaticality; humanlikeness); readability; [*multiple (3)*: goodness of outputs in their own right (both form and content), grammaticality, naturalness (form)]; [*multiple (2)*: goodness of outputs in their own right (form), grammaticality]; [*multiple (3)*: fluency, grammaticality]; [*multiple (2)*: grammaticality, readability]; [*multiple (2)*: fluency, readability]; [*multiple (3)*: goodness of outputs in their own right (both form and content), grammaticality, naturalness (form)]; [*multiple (3)*: coherence, humanlikeness, quality of outputs]; [*multiple (2)*: goodness of outputs in their own right (both form and content), grammaticality]	15
readability	fluency; goodness of outputs in their own right; goodness of outputs in their own right (both form and content); quality of outputs; usefulness for task/information need; readability; [*multiple (2)*: coherence, fluency]; [*multiple (2)*: fluency, readability]; [*multiple (2)*: readability, understandability]; [*multiple (3)*: clarity, correctness of outputs in their own right (form), goodness of outputs in their own right]	10
coherence	appropriateness (content); coherence; correctness of outputs in their own right (content); goodness of outputs in their own right (content); goodness of outputs relative to linguistic context in which they are read/heard; wellorderedness; [*multiple (2)*: appropriateness (content), understandability]; [*multiple (2)*: fluency, grammaticality]	8
naturalness	clarity; humanlikeness; naturalness; naturalness (both form and content); [*multiple (2)*: naturalness (both form and content), readability]; [*multiple (2)*: grammaticality, naturalness]	6
quality	goodness of outputs in their own right; goodness of outputs in their own right (both form and content); goodness of outputs (excluding correctness); quality of outputs; [*multiple (3)*: correctness of outputs relative to input (content), Fluency, Grammaticality]	5
correctness	appropriateness (content); correctness of outputs relative to input (content); correctness of outputs relative to input (both form and content); correctness of outputs relative to input (form)	4
usability	clarity; quality of outputs; usefulness for task/information need; user satisfaction	4
clarity	clarity; correctness of outputs relative to input (content); understandability; [*multiple (2)*: clarity, understandability]	4
informativeness	correctness of outputs relative to input (content); goodness of outputs relative to input (content); information content of outputs; text property (informative)	4
accuracy	correctness of outputs relative to input; correctness of outputs relative to input (content); goodness of outputs relative to input (content); referent resolvability	4

Table 5: Quality criterion names as given by authors mapped to normalised criterion names reflecting our assessment of what the authors actually measured. 'Count' is the number of different mappings found for each original criterion name.

former is far less straightforward to address.

One key observation from our data is that the same quality criterion names are often used by different authors to refer to very different aspects of quality, and that different names often refer to the same aspect of quality. We further found that more than half of the papers failed to define the criteria they evaluated, and about a quarter omitted to name the criteria being evaluated.

Our analysis has emphasised the need for better reporting of details of evaluations in order to help readers understand what aspect of quality is being evaluated and how. It took the first nine authors of the paper 25–30 minutes on average even in the final round of annotations to annotate a single paper, a measure of how hard it currently is to locate information about evaluations in papers.

Based on this experience we have put together a list of what we see as *reporting recommendations* for human evaluations presented in Table 7. The aim is to provide authors with a simple list of what information to include in reports of human evaluations at a minimum. The next step will be to develop the recommendations in Table 7 into a Human Evaluation Checklist giving full details of what to include in reports of human evaluation experiments, to complement existing recommendations for datasets and machine learning models, their intended uses, and potential abuses (Bender and Friedman, 2018; Gebru et al., 2018; Mitchell et al., 2019; Pineau, 2020; Ribeiro et al., 2020), aimed at making "critical information accessible that previously could only be found by users with great effort" (Bender and Friedman, 2018).

6 Conclusion

We have presented our new dataset of 165 papers each annotated with 16 attribute values that encode different aspects of the human evaluations reported in them. We described the carefully developed and validated annotation scheme we created for this

QUALITY CRITERION 1	QUALITY CRITERION 2	Count
Grammaticality	Goodness of outputs in their own right (both form and content)	6
Fluency	Grammaticality	4
Clarity	Goodness of outputs in their own right	4
Clarity	Correctness of outputs in their own right (form)	4
Goodness of outputs in their own right	Correctness of outputs in their own right (form)	4
Readability	Understandability	3
Appropriateness	Goodness of outputs relative to input (content)	3
Grammaticality	Naturalness (form)	2
Naturalness	Grammaticality	2
Fluency	Readability	2
Appropriateness (content)	Correctness of outputs relative to input (content)	2
Appropriateness	Information content of outputs	2
Information content of outputs	Goodness of outputs relative to input (content)	2
Readability	Grammaticality	2
Other	Other	60

Table 6: Quality criteria most frequently combined in a single prompt/question put to evaluators.

	SYSTEM
task	**What problem are you solving (e.g. data-to-text)?** How does it relate to other NLG (sub)tasks?
input/output	**What do you feed in and get out of your system?** Show examples of inputs and outputs of your system. Additionally, if you include pre and post-processing steps in your pipeline, clarify whether your input is to the preprocessing, and your output is from the post-processing, step, or what you consider to be the 'core' NLG system. In general, make it easy for readers to determine what form the data is in as it flows through your system.
	EVALUATION CRITERIA
name	**What is the name for the quality criterion you are measuring (e.g. grammaticality)**?
definition	**How do you define that quality criterion?** Provide a definition for your criterion. It is okay to cite another paper for the definition; however, it should be easy for your readers to figure out what aspects of the text you wanted to evaluate.
	OPERATIONALISATION
instrument type	**How are you collecting responses?** Direct ratings, post-edits, surveys, observation? Rankings or rating scales with numbers or verbal descriptors? Provide the full prompt or question with the set of possible response values where applicable, e.g. when using Likert scales.
instructions, prompts, and questions	**What are your participants responding to?** Following instructions, answering a question, agreeing with a statement? *The exact text you give your participants is important for anyone trying to replicate your experiments.* In addition to the immediate task instructions, question or prompt, provide the full set of instructions as part of your experimental design materials in an appendix.

Table 7: Reporting of human evaluations in NLG: Recommended minimum information to include.

purpose, and reported analyses and visualisations over the annotations.

Our analyses shed light on the kinds of evaluations NLG researchers have conducted and reported over the past 20 years. We have found a very high level of diversity of approaches, and fundamental gaps in reported details, including missing definitions of the aspect of quality being evaluated in about two-thirds of papers, and absence of basic details such as language, system input/output, etc.

We have proposed normalised quality criteria names and definitions to help us understand which evaluations actually evaluate the same thing. These are not intended as a set of standardised evaluation criteria that can be taken off the shelf and used. Rather, they are a first step in that direction. For a standardised set it would be desirable to ground evaluation criteria in related and much researched constructs in other fields. For example, there is a long history of studying readability (Chall, 1958; De Clercq et al., 2014).

Our single main conclusion is that, as a field, we need to standardise experimental design and terminology, so as to make it easier to understand and compare the human evaluations we perform.

Acknowledgments

Howcroft and Rieser's contributions were supported under EPSRC project MaDrIgAL (EP/N017536/1). Gkatzia's contribution was supported under the EPSRC project CiViL (EP/T014598/1). Mille's contribution was supported by the European Commission under the H2020 contracts 870930-RIA, 779962-RIA, 825079-RIA, 786731-RIA.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 307–317, Tilburg University, The Netherlands. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Anya Belz, Simon Mille, and David Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey.

Jeanne S. Chall. 1958. *Readability: An Appraisal of Research and Application*. The Ohio State University, Columbus, OH, USA.

Orphée De Clercq, Véronique Hoste, Bart Desmet, Philip Van Oosten, Martine De Cock, and Lieve Macken. 2014. Using the Crowd for Readability Prediction. *Natural Language Engineering*, 20(03):293–325.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2018. Datasheets for Datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, Stockholm, Sweden.

Dimitra Gkatzia and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005-2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 57–60.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pages 220–229, Atlanta, GA, USA. ACM Press.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *BMJ*, 339.

Jekaterina Novikova, 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.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Joelle Pineau. 2020. The Machine Learning Reproducibility Checklist (v2.0, Apr.7 2020).

Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics*, pages 1–8.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94, Tokyo, Japan. Association for Computational Linguistics.

Appendices

A System language

LANGUAGE	Before 2010	Since	Total
English	46	95	141
German	2	5	7
Japanese	4	3	7
Spanish	1	3	4
Chinese	1	3	4
Dutch	1	3	4
Other (13 languages)	1	14	15

Table 8: Language frequencies before and after 2010.

B System task

TASK	Before 2010	Since	Total
data-to-text generation	14	34	48
dialogue turn generation	7	14	21
summarisation (text-to-text)	1	11	12
referring expression generation	4	7	11
end-to-end text generation	1	8	9
question generation	0	9	9
feature-controlled generation	4	5	9
surface realisation (slr to text)	3	5	8
deep generation (dlr to text)	4	4	8
paraphrasing / lossless simplification	2	6	8
Other (15 tasks)	20	17	37

Table 9: Task frequencies before and after 2010.

C System Output

Output	Count
text: multiple sentences	68
text: sentence	40
text: documents	20
text: subsentential units of text	13
text: variable-length	10
no output (human generation)	7
raw/structured data	3
text: dialogue	3
shallow linguistic representation (slr)	2
deep linguistic representation (dlr)	1
speech	1
text: other (please specify): templates	1

Table 10: Counts for system output attribute.

D Alphabetical list of quality criterion names and definitions

Answerability from input: The degree to which an output (typically a question or problem) can be answered or solved with content/information from the input.

Appropriateness: The degree to which the output is appropriate in the given context/situation.

Appropriateness (both form and content): The degree to which the output as a whole is appropriate in the given context/situation. E.g. "does the text appropriately consider the parents' emotional state in the given scenario?"

Appropriateness (content): The degree to which the content of the output is appropriate in the given context/situation. E.g. "is the question coherent with other generated questions?"

Appropriateness (form): The degree to which the form of the output is appropriate in the given context/situation. E.g. "are the lexical choices appropriate given the target reader?"

Clarity: The degree to which the meaning of an output is absorbed without effort, i.e. is easy to understand as well as possible to understand.

Coherence: The degree to which the content/meaning of an output is presented in a well-structured, logical and meaningful way. E.g. "does the generated text accord with the correct logic?"

Cohesion: The degree to which the different parts of an output form a cohesive whole. Cohesion is the grammatical and lexical linking within a text or sentence that holds a text together and gives it meaning.

Correctness of outputs: The degree to which outputs are correct. Evaluations of this type ask in effect 'Is this output correct?' with criteria in child nodes adding more detail.

Correctness of outputs in their own right: The degree to which an output is correct/accurate/true, looking only at the output.

Correctness of outputs in their own right (both form and content): The degree to which both the form and content of an output are correct, looking only at the output.

Correctness of outputs in their own right (content): The degree to which the content of an output is correct, looking only at the output. E.g. "is this dictionary reference semantically complete?" (best = no further info needed).

Correctness of outputs in their own right (form): The degree to which the form of an output is correct, looking only at the output.

Correctness of outputs relative to external frame of reference: The degree to which an output is correct/accurate/true relative to a system-external frame of reference.

Correctness of outputs relative to external frame of reference (both form and content) : The degree to which the form and content of an output is correct/accurate/true relative to a system-external frame of reference.

Correctness of outputs relative to external frame of reference (content): The degree to which the content of an output is correct/accurate/true relative to a system-external frame of reference. E.g. "are the contents of the text factually true?" (best = no untrue facts).

Correctness of outputs relative to external frame of reference (form): The degree to which the form of an output is correct/accurate/true relative to a system-external frame of reference. E.g. "does the generated question use correct named entity names as given in this database?" (best = all as in database).

Correctness of outputs relative to input: The degree to which an output is correct/accurate/true relative to the input.

Correctness of outputs relative to input (both form and content): The degree to which the form and content of an output is correct/accurate/true relative to the input.

Correctness of outputs relative to input (content): The degree to which the content of an output is correct/accurate/true relative to the input. E.g. "is all the meaning of the input preserved?", "to what extent does the generated text convey the information in the input table?" (best = all the information).

Correctness of outputs relative to input (form): The degree to which the form of an output is correct/accurate/true relative to the input. E.g. " how similar are the words to the input?" (best = same).

Detectability of controlled feature [PROPERTY]: The degree to which a property that the outputs are intended to have (i.e. because it's controlled by input to the generation process) is detectable in the output. Open class criterion; PROPERTY can be a wide variety of different things, e.g. conversational, meaningful, poetic, vague/specific, etc.

Ease of communication: The degree to which the outputs make communication easy, typically in a dialogue situation. E.g. "how smoothly did the conversation go with the virtual agent?"

Effect on reader/listener [EFFECT]: The degree to which an output has an EFFECT in the listener/reader. Open class criterion; EFFECT can be a wide variety of different things, e.g. inducing a specific emotional state, inducing behaviour change, etc. E.g. measuring how much the user learnt from reading the output; "are you feeling sad after reading the text?"

Fluency: The degree to which a text 'flows well' and is not e.g. a sequence of unconnected parts.

Goodness as system explanation: Degree to which an output is satisfactory as an explanation of system behaviour. E.g. "does the text provide an explanation that helps users understand the decision the system has come to?"

Goodness of outputs (excluding correctness): The degree to which outputs are good. Evaluations of this type ask in effect 'Is this output good?' with criteria in child nodes adding more detail.

Goodness of outputs in their own right: The degree to which an output is good, looking only at the output.

Goodness of outputs in their own right (both form and content): The degree to which the form and content of an output are good, looking only at the output.

Goodness of outputs in their own right (content): The degree to which the content of an output is good, looking only at the output.

Goodness of outputs in their own right (form): The degree to which the form of an output is good, looking only at the output. E.g. "is the generated response a complete sentence?"

Goodness of outputs relative to external frame of reference: The degree to which an output is good relative to a system-external frame of reference.

Goodness of outputs relative to grounding: The degree to which an output is good relative to grounding in another modality and/or real-world or virtual-world objects as a frame of reference.

Goodness of outputs relative to how humans use language: The degree to which an output is good relative to human language use as a frame of reference.

Goodness of outputs relative to input: The degree to which an output is good relative to the input.

Goodness of outputs relative to input (both form and content): The degree to which the form and content of an output is good relative to the input. E.g. "does the output text reflect the input topic labels?"

Goodness of outputs relative to input (content): The degree to which an output is good relative to the input. E.g. "does the output text include the important content from inputs?"

Goodness of outputs relative to input (form): The degree to which the form of an output is good relative to the input. E.g. in paraphrasing: "is the surface form of the output different enough from that of the input?"

Goodness of outputs relative to linguistic context in which they are read/heard: The degree to which an output is good relative to linguistic context as a frame of reference.

Goodness of outputs relative to system use: The degree to which an output is good relative to system use as a frame of reference.

Grammaticality: The degree to which an output is free of grammatical errors.

Humanlikeness: The degree to which an output could have been produced by a human.

Humanlikeness (both form and content): The degree to which the form and content of an output could have been produced/chosen by a human.

Humanlikeness (content): The degree to which the content of an output could have been chosen by a human (irrespective of quality of form).

Humanlikeness (form): The degree to which the form of an output could have been produced by a human (irrespective of quality of content).

Inferrability of speaker/author stance [OBJECT]: The degree to which the speaker's/author's stance towards an OBJECT is inferrable from the text. E.g. "rank these texts in order of positivity expressed towards the company."

Inferrability of speaker/author trait [TRAIT]: The degree to which it is inferrable from the output whether the speaker/author has a TRAIT. Open-class criterion; TRAIT can be a wide variety of different things, e.g. personality type, identity of author/speaker, etc. E.g. "who among the writers of these texts do you think is the most conscientious?"

Information content of outputs: The amount of information conveyed by an output. Can range from 'too much' to 'not enough', or 'very little' to 'a lot'. E.g. "is the general level of details provided in the text satisfactory?", "do you personally find the amount of information in the text optimal?"

Multiple (list all): use only if authors use single criterion name which corresponds to more than one criterion name in the above list. Include list of corresponding criteria in brackets.

Naturalness: The degree to which the output is likely to be used by a native speaker in the given context/situation.

Naturalness (both form and content): The degree to which the form and content of an output is likely to be produced/chosen by a native speaker in the given context/situation.

Naturalness (content): The degree to which the content of an output is likely to be chosen by a native speaker in the given context/situation.

Naturalness (form): The degree to which the form of an output is likely to be produced by a native speaker in the given context/situation.

Nonredundancy (both form and content): The degree to which the form and content of an output are free of redundant elements, such as repetition, overspecificity, etc.

Nonredundancy (content): The degree to which the content of an output is free of redundant elements, such as repetition, overspecificity, etc.

Nonredundancy (form): The degree to which the form of an output is free of redundant elements, such as repetition, overspecificity, etc.

Quality of outputs: Maximally underspecified quality criterion. E.g. when participants are asked which of a set of alternative outputs they prefer (with no further details).

Readability: The degree to which an output is easy to read, the reader not having to look back and reread earlier text.

Referent resolvability: The degree to which the referents of the referring expressions in an output can be identified.

Speech quality: The degree to which the speech is of good quality in spoken outputs.

Spelling accuracy: The degree to which an output is free of spelling errors.

Text Property [PROPERTY]: The degree to which an output has a specific property (excluding features controlled by an input parameter). Open class criterion; PROPERTY could be a wide variety of different things: conversational, informative, etc. E.g. "does the text have the characteristics of a poem?"

Text Property [Complexity/simplicity]: The degree to which an output is complex/simple.

Text Property [Complexity/simplicity (both form and content)]: The degree to which an output as a whole is complex/simple.

Text Property [Complexity/simplicity (content)]: The degree to which an output conveys complex/simple content/meaning/information. E.g. "does the generated question involve reasoning over multiple sentences from the document?"

Text Property [Complexity/simplicity (form)]: The degree to which an output is expressed in complex/simple terms. E.g. "does the generated text contain a lot of technical or specialist words?"

Understandability: Degree to which the meaning of an output can be understood.

Usability: The degree to which the system in the context of which outputs are generated is usable. E.g. user-system interaction measurements, or direct usability ratings for the system.

Usefulness (nonspecific): The degree to which an output is useful. E.g. measuring task success, or questions like "did you find the system advice useful?"

Usefulness for task/information need: The degree to which an output is useful for a given task or information need. E.g. "does the description help you to select an area for buying a house?"

User satisfaction: The degree to which users are satisfied with the system in the context of which outputs are generated. E.g. in a dialogue system "how satisfied were you with the booking you just made?"

Wellorderedness: The degree to which the content of an output is well organised and presents information in the right order.

E Taxonomy of Quality Criteria

Figure 4 shows the 71 quality criteria (plus some filler nodes, in grey) structured hierarchically into a taxonomy. For the top three levels of branches in the taxonomy we used the quality criterion properties from Belz et al. (2020): (i) goodness vs. correctness vs. features; (ii) quality of output in its own right vs. quality of output relative to input vs. quality of output relative to an external frame of reference (yellow, red, orange); (iii) form of output vs. content of output vs. both form and content of output (green, blue, purple).

Note that the taxonomy is not necessarily complete in this state; it contains all and only those 71 distinct criteria that resulted from our survey.

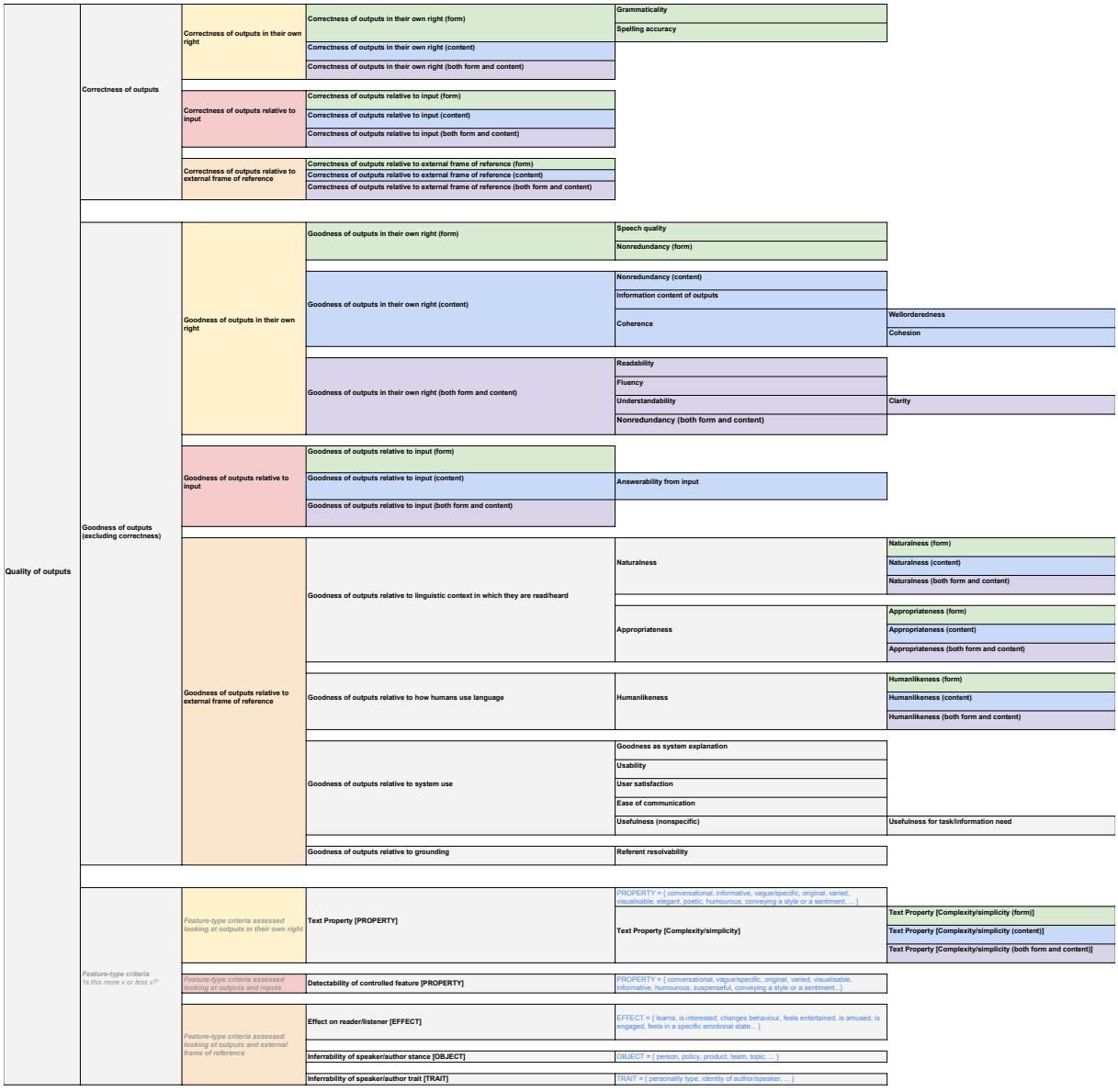

Figure 4: Taxonomy of normalised quality criteria; greyed out criterion names = not encountered, and/or included for increased completeness of taxonomy.