Twenty years of confusion in Human Evaluation

NLG needs evaluation sheets and standardised definitions

David M. Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, & Verena Rieser

Evaluation is complex

One input

Many valid outputs

So automated evaluation is hard

```
name[Aromi], food[Chinese], customer rating[5
out of 5], area[city centre]
```

Aromi is a restaurant providing Chinese food. It is located in the city centre. Its customer rating is 5 out of 5.

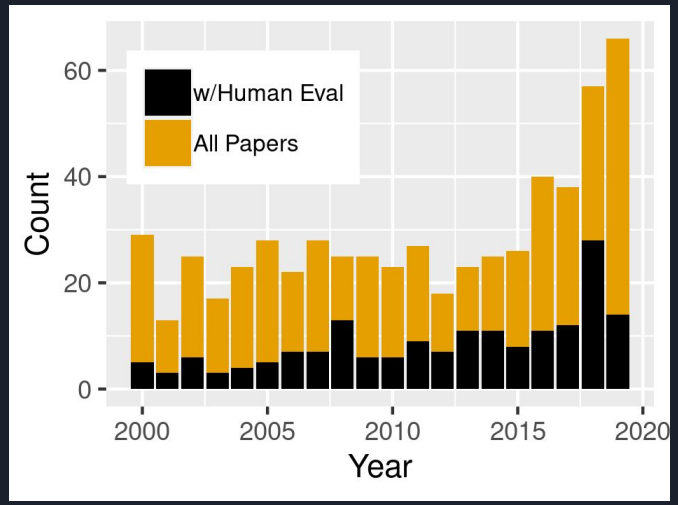
There's a Chinese place in the city center called Aromi with a 5-star rating.

Five out of five is Aromi, a centrally located Chinese restaurant.

You can't go wrong with Aromi, serving Chinese food in the city centre. 5/5

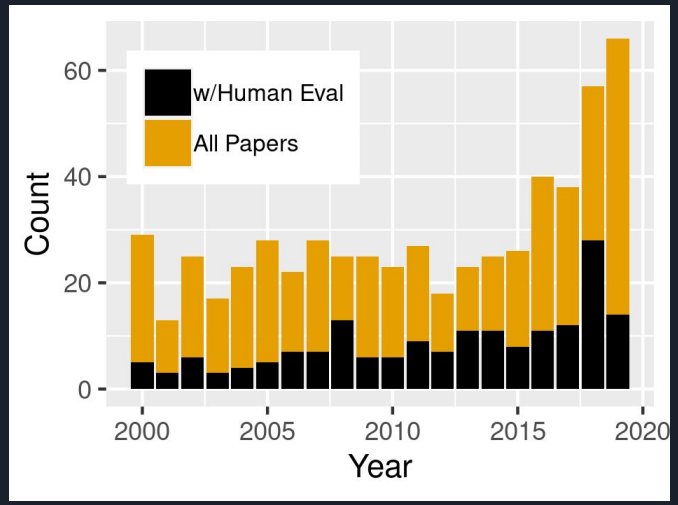
Human evaluation...

Human evaluation viewed as more reliable.

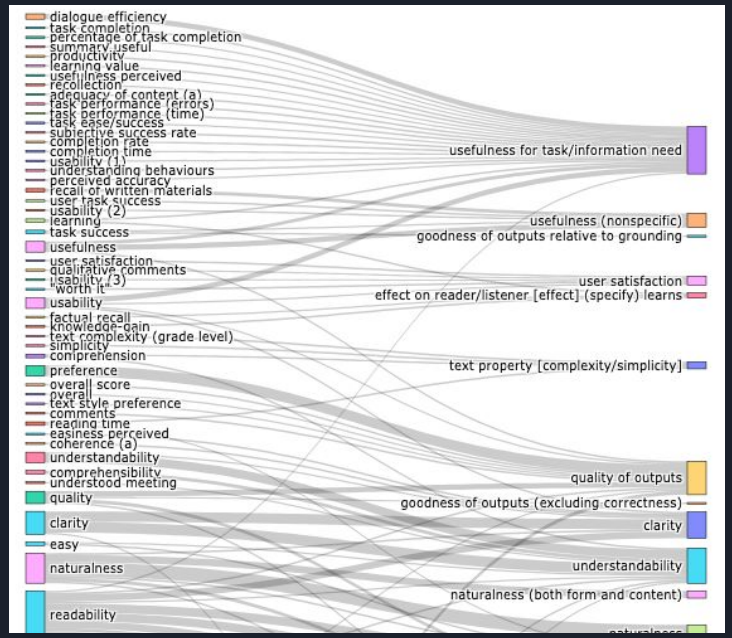


Human evaluation… is also complex!

Human evaluation viewed as more reliable.



Many names for the same thing!



What is the current state of human eval?

Initial Pool	Initial Selection	Final Selection
578 INLG/ENLG papers (2000-2019)	217 with likely human evaluations	165 confirmed to have human evaluations

Systematic review (cf. PRISMA, Moher et al. 2009)

Overview of Annotations: System Properties

Language	Input	Output	Task
English	raw/structured data	text: multiple sentences	dialogue turn generation

- Capturing basic facts about the systems being evaluated

Example annotation for Porayska-Pomsta & Mellish 2004

Overview of Annotations: Quality Criteria

Verbatim Criterion Name	Verbatim Definition	Normalised Criterion Name	Paraphrase of Definition
appropriateness	"appropriateness...determined against both contextual information and other linguistic possibilities available at the time the linguistic decisions are made...[and] based on socio-cultural factors and conventions established by a given speech community"	6a. Appropriateness (both form and content)	the extent to which a response is appropriate for a tutor to say to a student in response to a wrong answer in a tutoring dialogue

- What aspect of text quality do the authors aim to evaluate?
- What aspect of text quality do they actually evaluate?

Example annotation for Porayska-Pomsta & Mellish 2004

Overview of Annotations: Operationalisation

List/Range of response values	Size of rating instrument	Type of scale or rating instrument	Data type of collected responses	Form of Response Elicitation	Verbatim question, prompt, etc	Paraphrase of Question, Prompt, etc	Statistics
1,2,3,4,5	5	numerical rating scale	ordinal	direct quality estimation	not given	how appropriate is the response in the given situation?	mean, t-test

- What were humans asked to do or assess?
- How were they asked to do it?
- What kind of data came from it?

Example annotation for Porayska-Pomsta & Mellish 2004

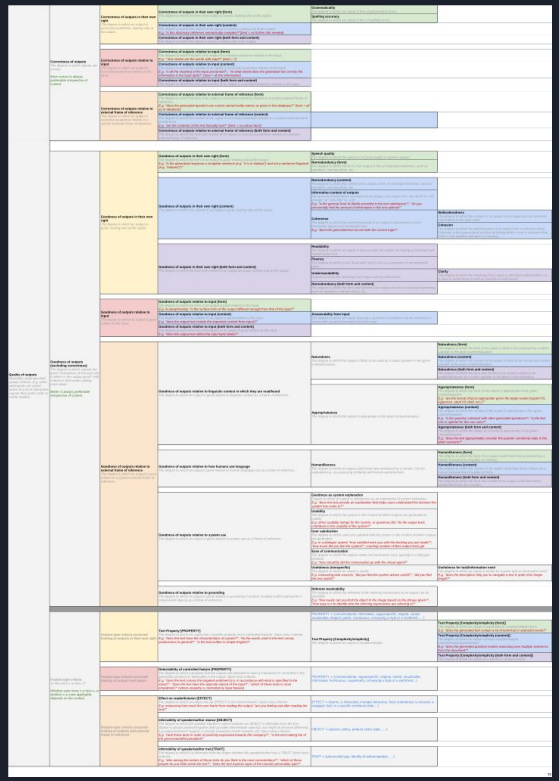
A glimpse of the annotation guidelines

Guidelines for system, criterion, & operationalisation

- Captured most frequent labels, e.g., system tasks or forms of elicitation response categorically

Example: Quality Criterion

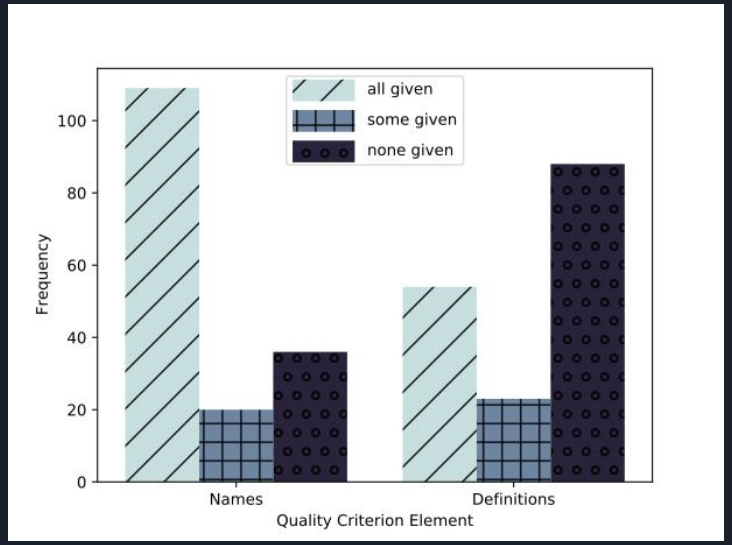
- Reduced 204 "verbatim" criteria
- To 71 normalised criteria (in color to right)
- Top-level based on Belz et al. 2020



Basic details about quality criteria are often missing.

Most papers **state what they intend to measure**, but…

Most papers **omit the definition** of the criterion they are evaluating!



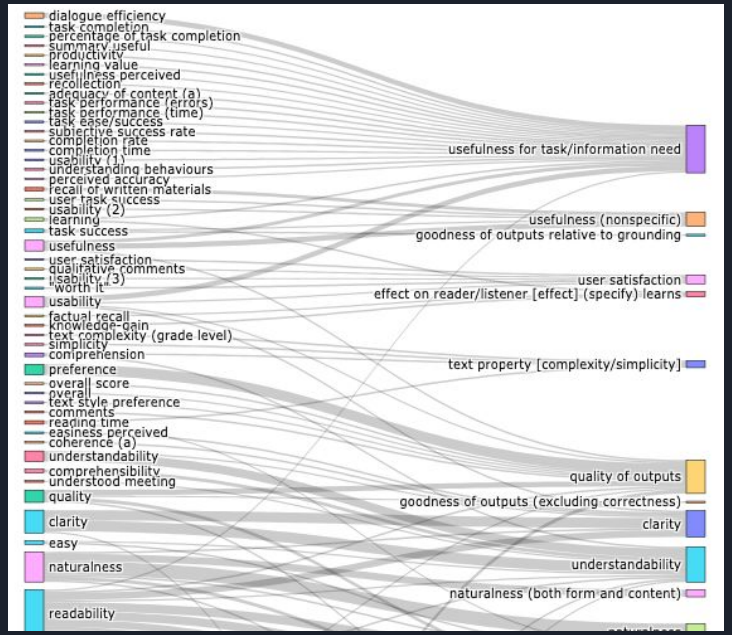
Confusion about what is being evaluated

The same name used to mean many things.

Example: **Fluency**

- Fluency
- Goodness of outputs in their own right
- Readability
- Goodness in their own right, grammaticality, and naturalness
- Coherence, humanlikeness, quality of outputs

Many names for the same basic criterion.



Some of our findings

- Prompts often combined criteria
  - 50 cases combining 2-4 criteria
- Most common elicitation methods:
  - Direct quality estimation (207)
  - Relative quality estimation (72)

- Five most frequent evaluation criteria:
  - Usefulness (39)
  - Grammaticality (39)
  - Quality of outputs (35) <- underspecified
  - Understandability (30)
  - Correctness rel. to input (29)

Key findings

1. Little shared practice in human evaluation
2. Reporting in papers rarely complete

Our Recommendations for Writing

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.

Take-aways

Conclusion:

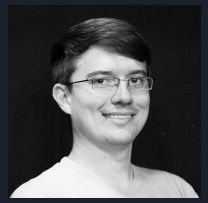
- **We need to standardise our experimental design and terminology** to make it easier to understand and compare the results of human evals.

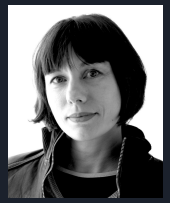
Resources:

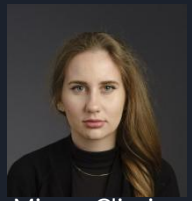
- A dataset based on 20 years of INLG publications using human evaluations
- New annotation scheme to determine what is reported in NLG papers
- Reporting recommendations based on our experiences.

`https://github.com/evalgenchal/20Y-CHEC`

The Team

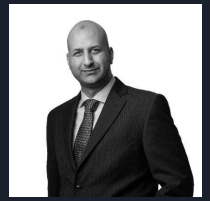

Dave Howcroft


Anya Belz


Miruna Clinciu

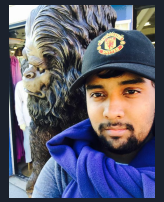

Dimitra Gkatzia


Sadid Hasan


Saad Mahamood


Simon Mille

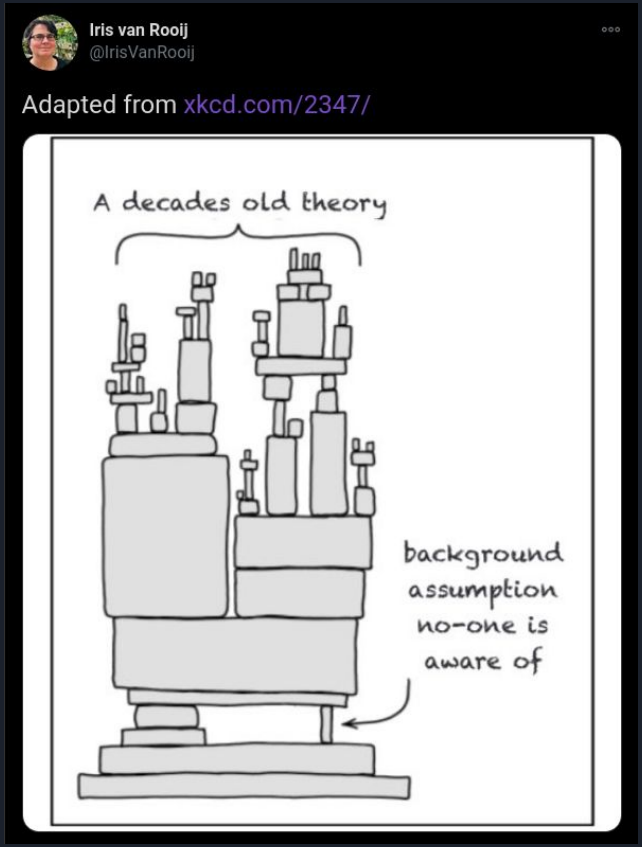

Emiel van Miltenburg


Sashank Santhanam


Verena Rieser

We thank you for your attention!

Bonus round!



We want to be aware of the assumptions.