

# Human Evaluation of NLP System Quality

INLG Tutorial, 24th September 2024

Unit 3: Quality Criteria and Evaluation Modes

[Link to Unit 3 Resources](#)

# Overview

## Unit 3: Quality criteria and evaluation modes

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Research question(s) and hypotheses
3. Quality criteria and evaluation modes
4. Using QC properties and evaluation modes:
  - Comparing evaluation measures in existing evaluations;
  - Creating evaluation measures for new evaluations.
5. Connection with formulating the research question
6. Unit summary and pointers to other units
7. References
8. Practical task

# Overview

## Unit 3: Quality criteria and evaluation modes

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Research question(s) and hypotheses
3. Quality criteria and evaluation modes
4. Using QC properties and evaluation modes:
  - Comparing evaluation measures in existing evaluations;
  - Creating evaluation measures for new evaluations.
5. Connection with formulating the research question
6. Unit summary and pointers to other units
7. References
8. Practical task

# Unit aims and learning outcomes

- The aims of Unit 3 are:
  - To introduce the concept of null hypothesis testing and relate it to the formulation of research questions.
  - To deepen understanding of the concepts of quality criteria and evaluation modes first introduced in Unit 2.
  - To introduce a taxonomy of quality criteria which can be used as a basis for designing standardised hence comparable evaluation measures.
  - To explain the connection between evaluation measures and the formulation of research questions, as separate from experimental design.
- After completion of the unit, participants will be able to:
  - Critically assess reports of hypothesis testing for validity of the basic approach.
  - Apply basic knowledge of hypothesis testing in selecting appropriate approach in own work.
  - Understand and apply the introduced quality criterion taxonomy in assessing existing work and designing new evaluations.

# Prerequisites and connections with other units

- Prerequisite(s) of Unit 3: Unit 3 can be studied as a stand-alone unit, but the terms and definitions introduced in Unit 2 are assumed.
- Unit 3 is a prerequisite of Units 4–8, and most particularly of Unit 4 which builds on the information about quality criteria and evaluation modes, and Unit 5 which assumes basic knowledge of null hypothesis testing.

# Overview

## Unit 3: Quality criteria and evaluation modes

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Research question(s) and hypotheses
3. Quality criteria and evaluation modes
4. Using QC properties and evaluation modes:
  - Comparing evaluation measures in existing evaluations;
  - Creating evaluation measures for new evaluations.
5. Connection with formulating the research question
6. Unit summary and pointers to other units
7. References
8. Practical task

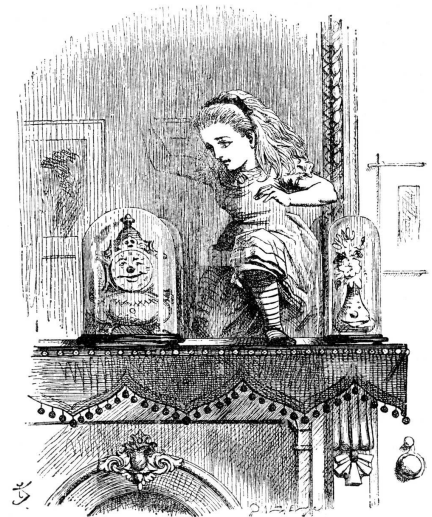
# Research questions, quality criteria and evaluation modes

- In this unit we take the first step in Phase I (Design) of developing a human evaluation in NLP: formulating the research question(s) and corresponding hypotheses.
- Assuming a very simple scenario for the moment, suppose we have created a new language generation system  $M_{new}$  and want to know if it performs better than an existing system  $M_{old}$ .
- Need to decide is what we mean by ‘better than’ – could be overall quality, but in most cases the evaluation task is easier if more specific.
- Suppose we have been working to improve the system’s grammaticality so our *quality criterion* is Grammaticality.
- Grammaticality assesses the *correctness* of the *form* of outputs *in their own right*, in terms of the quality criterion classification system we will introduce later in this unit.

E.g. we know these sentences are grammatically correct without considering their *meaning*, or anything other than the sentence:

*Colorless green ideas sleep furiously.*

*All mimsy were the borogoves, And the mome raths outgrabe.*



# Research questions, quality criteria and evaluation modes

- For a fully specified research question, we also need to decide *evaluation modes* as the answer can be expected to be different depending on which modes we choose:
  - Do we just want to know whether  $M_{new}$  outputs are more grammatical than  $M_{old}$  outputs (relative), or also quantify by how much (absolute)?
  - Are we interested more in users' perception of the system's grammaticality (subjective), or in measuring the degree to which its outputs conform to a given notion of grammar (objective)?
  - Do we want to assess outputs directly (intrinsic), or in terms of their effect on something external to the system, e.g. how many post-edits a user performs (extrinsic)?
- Recall from Unit 2:
  - Quality criterion + evaluation modes = evaluation measure;
  - Evaluation measure + experimental design = evaluation method.



# Research questions, quality criteria and evaluation modes

- Suppose we decide to assess our quality criterion Grammaticality in *absolute*, *subjective* and *intrinsic* evaluation modes (by far the most common combination of evaluation modes in NLP).
- This gives us [Absolute, Subjective, Intrinsic Grammaticality] as the *evaluation measure m*.
- This could, at a later stage in the Experiment Design phase, be decided to be assessed by asking evaluators to rate each system output individually on a scale of 1–5 (there are many other options).
- But that is part of how we choose to find an answer for our research question (Experiment Design, Unit 4), whereas the evaluation measure (quality criterion + evaluation modes) is part of the research question itself.

# Research questions, quality criteria and evaluation modes

- Next we need to **formulate the research question** – two common forms are:
  - A. Is  $M_{new}$  more [absolutely, subjectively and intrinsically grammatical] than  $M_{old}$ ?
  - B. Which of  $M_{new}$  and  $M_{old}$  is more [absolutely, subjectively and intrinsically grammatical]?
- The corresponding hypotheses that are tested by the evaluation experiment are:
  - A. **Null hypothesis:** there is no difference between  $M_{new}$  and  $M_{old}$  in terms of [absolute, subjective and intrinsic grammaticality].  
**Alternative hypothesis:**  $M_{new}$  is more [absolutely, subjectively and intrinsically grammatical] than  $M_{old}$
  - B. **Null hypothesis:** there is no difference between  $M_{new}$  and  $M_{old}$  in terms of [absolute, subjective and intrinsic grammaticality].  
**Alternative hypotheses:**
    - $M_{new}$  is more [grammatical in absolute, subjective and intrinsic terms] than  $M_{old}$
    - $M_{old}$  is more [grammatical in absolute, subjective and intrinsic terms] than  $M_{new}$

# Research questions, quality criteria and evaluation modes

- The choice of research question impacts the statistical power of the experiment and the types of statistical tests that can be applied – we will come back to this in Unit 5.
- NB: Answering research questions of type A can never show that  $M_{old}$  is better, only either that  $M_{new}$  is better or that we have no evidence to conclude that  $M_{new}$  is better.  
Answering research questions of type B can show either that  $M_{old}$  is better, or that  $M_{new}$  is better (or that we have no evidence supporting either conclusion).

# Research questions, quality criteria and evaluation modes

- More generally, in a typical evaluation scenario in NLP, we have a set of alternative approaches, implemented as systems  $M_i$ , to solving a task, and want to find out something about their relative performance at the task.
- Set of alternative methods (systems)  $M_i$  can include:
  - Baseline method(s)
  - Previously reported, independently developed systems
  - Newly proposed systems
  - Human topline(s)
- In order to answer a given research question, we need a test set  $s$  that can be used to obtain comparable system outputs  $o_i^s$  from all systems  $M_i$
- We also need an evaluation method  $E_m$  that can be used to obtain measured values  $v_i$  for each system  $M_i$  (represented by its test set outputs  $o_i^s$ ), and evaluation measure  $m$ , or:  $E_m: (o_i^s, s) \mapsto v_i$

# Research questions, quality criteria and evaluation modes

- We answer the research question using a standard approach called **null hypothesis testing** in a form mainly due to Fisher (1935).
- Recall that for both formulations A and B, the **null hypothesis** was that there is no difference in terms of  $E_m$  between (pairs of) systems in our set.
- We look for strong enough evidence to allow us to conclude, on the basis of test data set  $s$ , samples of system outputs  $o_i^s$  and values of  $E_m$  computed on them, that we are wrong in assuming the null hypothesis is correct, and to conclude instead that there is a difference.
- We test the strength of the evidence via a **statistical hypothesis test** involving the calculation of a **test statistic**.
- Most commonly, a **p-value** computed from the test statistic is taken to indicate the level of significance.
- The p-value expresses the probability under the null hypothesis of obtaining a test statistic value that is at least as extreme as the one actually observed.
- A small p-value means there is little overall chance of obtaining a result like the one observed if the null hypothesis (there is no difference between the systems in terms of  $E_m$ ) is true.

# Answering the research question

- It is standard to require a significance level of at least 0.05, i.e. to require the overall probability of obtaining such test static values to be less than 0.05.
- Our results can then either be significant (if  $p$  is less than 0.05) in which case we can draw the provisional conclusion that a pair of systems has different performance in terms of  $E_m$ .
- Or our results are not significant, in which case we can conclude nothing at all.
- A standard way of reporting the answer to our research question is as a partial ranking of systems indicating for each pair of systems the significance level of the difference between them.
- It's tempting (and researchers often can't resist) to treat absence of statistical significance as some sort of result, especially when comparing a system's performance with human performance, but that is emphatically not the right thing to do.
- We will return to test statistics and hypothesis testing, and how they relate to types of response values obtained in human evaluations, in Unit 5.
- In Unit 4, we look at Design Steps 3–9 which specify the remaining experiment properties needed for an evaluation method  $E_m$  that assesses evaluation measure  $m$ .
- In the next section in this unit, we look at how to put together evaluation measures  $m$ .

# Overview

## Unit 3: Quality criteria and evaluation modes

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Research question(s) and hypotheses
3. Quality criteria and evaluation modes
4. Using QC properties and evaluation modes:
  - Comparing evaluation measures in existing evaluations;
  - Creating evaluation measures for new evaluations.
5. Connection with formulating the research question
6. Unit summary and pointers to other units
7. References
8. Practical task

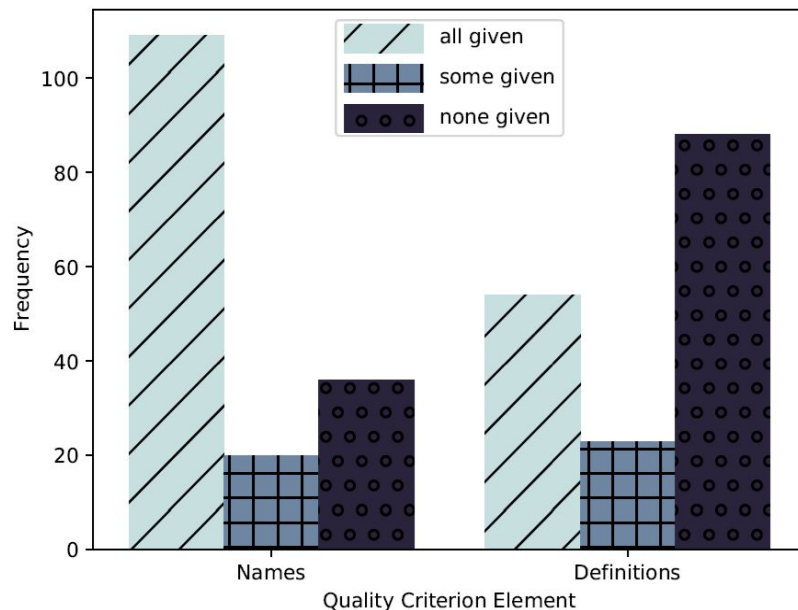
# Quality criteria and evaluation measures

- Recall from Unit 2:  
Quality criterion + evaluation modes = **evaluation measure**;  
Evaluation measure + experimental design = **evaluation method**.
- We will first look at quality criteria (including a first look at a QC taxonomy), then evaluation modes.
- Next, will look at two ways of using quality criteria properties and evaluation modes:
  - Comparing what different evaluations are assessing;
  - Devising an evaluation measure for a new evaluation.
- Finally, will introduce the QCET interactive taxonomy tool, and use it in a practical task.



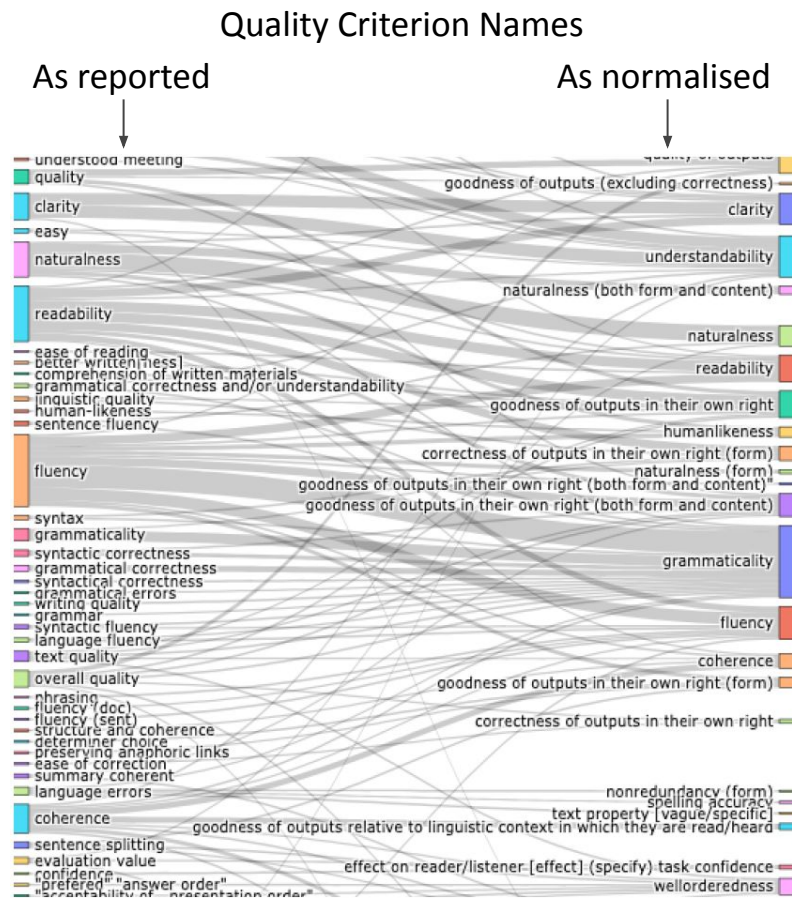
# Quality criteria

- Separation into quality criteria (QCs), measures and methods is in part motivated by the need to establish what is being evaluated in different evaluations (e.g. for comparing and reproducing results).
- This is otherwise impossible because a large number of different quality criterion names are being used with often different meaning (van der Lee et al., 2019; Howcroft et al., 2020).
- Howcroft et al. (2020) mapped QCs in all INLG/ ENLG papers with human evaluations published 2000–2019 to normalised QC names.
- This was all but impossible as the information that would make clear what is being evaluated (QC name, definition, question/ instructions put to evaluators) is often not published at all.



# Quality criteria

- Howcroft et al. (2020) found 200+ different QC names in 165 papers (478 evaluations).
- These were mapped to 71 normalised QC names intended to be atomic in the sense of evaluating single aspects of quality.
- E.g. *Fluency* was mapped to 15 different normalised QCs, *Readability* to 10, *Coherence* to 8.
- In many cases, a reported QC was mapped to multiple normalised QCs.
- Normalised QCs were structured into a taxonomy showing commonalities and differences between QCs (Belz et al., 2020; Howcroft et al., 2020).



# Taxonomy of quality criteria

- To the right is the QC taxonomy from Howcroft et al. (2020)
- Recall the example from earlier: Grammaticality assesses the *correctness* of the *form* of outputs *in their own right*.
- Terms in italics refer to the main three levels in the taxonomy that correspond to three **quality criterion properties**:
  - Type of quality assessed: *Correctness, Goodness, Feature*
  - Aspect of outputs assessed: *Form, Content, Both form and content*
  - Frame of reference relative to which system quality is assessed: Outputs *In their own right, Relative to the inputs, Relative to an external frame of reference.*
- Will go through each QC property in turn.

Quality of outputs	Correctness of outputs	Correctness of outputs in their own right	Correctness of outputs in their own right (form)	Grammaticality
			Correctness of outputs in their own right (content)	Spelling accuracy
			Correctness of outputs in their own right (both form and content)	
		Correctness of outputs relative to input	Correctness of outputs relative to input (form)	
			Correctness of outputs relative to input (content)	
			Correctness of outputs relative to input (both form and content)	
		Correctness of outputs relative to external frame of reference	Correctness of outputs relative to external frame of reference (form)	
			Correctness of outputs relative to external frame of reference (content)	
			Correctness of outputs relative to external frame of reference (both form and content)	
Quality of outputs	Goodness of outputs	Goodness of outputs in their own right	Goodness of outputs in their own right (form)	Speech quality
				Nonredundancy (form)
			Goodness of outputs in their own right (content)	Nonredundancy (content)
				Information content of outputs
				Coherence
				Wellorderedness
			Goodness of outputs in their own right (both form and content)	Readability
				Fluency
				Understandability
				Nonredundancy (both form and content)
				Clarity
		Goodness of outputs relative to input	Goodness of outputs relative to input (form)	
			Goodness of outputs relative to input (content)	Answerability from input
			Goodness of outputs relative to input (both form and content)	
		Goodness of outputs relative to external frame of reference	Goodness of outputs relative to linguistic context in which they are read/heard	Naturalness (form)
				Naturalness (content)
				Naturalness (both form and content)
			Appropriateness	Appropriateness (form)
				Appropriateness (content)
				Appropriateness (both form and content)
		Goodness of outputs relative to how humans use language	Human-likeness	Human-likeness (form)
				Human-likeness (content)
				Human-likeness (both form and content)
		Goodness of outputs relative to system use	Goodness as system explanation	
			Usability	
			User satisfaction	
			Ease of communication	
			Usefulness (nonspecific)	Usefulness for task/information need
		Goodness of outputs relative to grounding	Referent resolvability	
Feature-type criteria	Feature-type criteria assessed (basis of outputs in their own right)	Text Property (PROPERTY)	PROPERTY = [ conversational, informative, vague/specific, original, varied, visualizable, elegant, poetic, humorous, conveying a style or a sentiment, ... ]	Text Property [Complexity/simplicity (form)]
				Text Property [Complexity/simplicity (content)]
				Text Property [Complexity/simplicity (both form and content)]
		Detectability of controlled feature (PROPERTY)	PROPERTY = [ conversational, vague/specific, original, varied, visualizable, informative, humorous, tongue-twister, conveying a style or a sentiment, ... ]	
Feature-type criteria assessed (basis of outputs and external frame of reference)	Effect on reader/listener (EFFECT)	Inferability of speaker/author stance (OBJECT)	EFFECT = [ leaves, is interested, changes behavior, feels entertained, is amused, is engaged, bears in a specific emotional state, ... ]	
Feature-type criteria assessed (basis of outputs and external frame of reference)	Inferability of speaker/author trait (TRAIT)	Inferability of speaker/author trait (TRAIT)	TRAIT = [ personality type, identity of author/speaker, ... ]	

# Type of quality assessed (Belz et al., 2020)

- For *Correctness* and *Goodness* type criteria, it is normally clear which end of the scale is preferred regardless of evaluation context. E.g. one would normally want output texts to be more *fluent*, more *grammatical*, more *clear*.
- For *Feature* type criteria this does not hold; in one evaluation context, one end of the scale might be preferable, in another, the other, and in a third, the criterion may not apply.

E.g. *Funny* and *Entertaining* might be desirable properties for a narrative generator, but are inappropriate in a nursing report generator.

- Will now take a closer look at each of these highest-level QC categories:
  - **Correctness**
  - **Goodness**
  - **Feature**

# Type of quality assessed (Belz et al., 2020)

- **Correctness:** For correctness criteria it is possible to state, generally for all outputs, the conditions under which outputs are maximally correct (hence of maximal quality).  
E.g. for *Grammaticality*, outputs are (maximally) correct if they contain no grammatical errors; for *Semantic Completeness*, outputs are correct if they express all the content in the input.
- **Goodness:** For *Goodness* criteria, in contrast to *Correctness* criteria, there is no single, general mechanism for deciding when outputs are maximally good, only for deciding for two outputs which is better and which is worse.  
E.g. for *Fluency*, even if outputs contain no disfluencies, there may be other ways in which any given output could be more fluent.
- **Features:** For criteria X in this class, outputs are not generally better if they are more X. Depending on evaluation context, more X may be better or less X may be better.  
E.g. outputs can be more specific or less specific, but it's not the case that outputs are, in the general case, better when they are more specific.

# Aspect of outputs assessed (Belz et al., 2020)

- **Form of output:** Evaluations of this type aim to assess the form of outputs alone, e.g. *Grammaticality* is only about the form, a sentence can be grammatical yet be wrong or nonsensical in terms of content.
- **Content of output:** This type of evaluations aim to assess the content/meaning of the output alone, e.g. *Meaning Preservation* only assesses output content; two sentences can be considered to have the same meaning, but differ in form.
- **Both form and content of output:** Here, evaluations assess outputs as a whole, not distinguishing form from content. E.g. *Coherence* is a property of outputs as a whole, either form or meaning can detract from it.

# Frame of reference of assessment (Belz et al., 2020)

This QC property is about whether assessment of output quality involves a frame of reference in addition to the outputs themselves, i.e. whether the evaluation process also consults (refers to) anything else. We distinguish three cases:

- **Quality of output in its own right:** assessing output quality without referring to anything other than the output itself, i.e. no system-internal or external frame of reference.  
E.g. *Poeticness* is assessed by considering (just) the output and how poetic it is.
- **Quality of output relative to input:** the quality of an output is assessed relative to the input.  
E.g. *Answerability* is the degree to which the output question can be answered from information in the input.
- **Quality of output relative to a system-external frame of reference:** output quality is assessed with reference to system-external information, e.g. a knowledge base, or a sample of gold-standard outputs.  
E.g. *Factual Accuracy* assesses outputs relative to a source of real-world knowledge.

# Evaluation modes

Evaluation modes are orthogonal to quality criteria, i.e. any given quality criterion can in principle be combined with any of the modes:

1. **Objective vs. subjective:** Examples of *objective* assessment include any automatically counted or otherwise quantified measurements such as mouse-clicks, occurrences in text, etc. *Subjective* assessments involve ratings, opinions and preferences by evaluators.
2. **Absolute vs. relative:** whether evaluators are shown outputs from a single system during evaluation (*absolute*), or from multiple systems in parallel (*relative*), in the latter case typically ranking or preference-judging them.
3. **Extrinsic vs. intrinsic:** in *extrinsic* evaluation, system performance is assessed in terms of the system's effect on something external to the system, e.g. how it affects the performance of an embedding system or of a user at a task;  
in *intrinsic* evaluation, outputs are assessed only within the system context (includes e.g. assessment relative to inputs or to an expected standard).



# Overview

## Unit 3: Quality criteria and evaluation modes

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Research question(s) and hypotheses
3. Quality criteria and evaluation modes
4. Using QC properties and evaluation modes:
  - Comparing evaluation measures in existing evaluations;
  - Creating evaluation measures for new evaluations.
5. Connection with formulating the research question
6. Unit summary and pointers to other units
7. References
8. Practical task

# Comparing reported evaluation measures

Comparison of evaluations of QC named Coherence from four different papers, using the introduced QC properties and evaluation modes:

Definition of <i>Coherence</i> in paper	Quality criterion properties			Evaluation mode		
	Type of quality	Form/ content	Frame of reference	Obj. vs. subject.	Abs. vs. relative	Extr. vs. intrinsic
"[whether] the poem [is] thematically structured" (Van de Cruys, 2020)						
"measures if a question is coherent with previous ones" (Chai & Wan, 2020)						
"refers to the meaning of the generated sentence, so that a sentence with no meaning would be rated with a 1 and a sentence with a full meaning would be rated with a 5" (Barros et al., 2017)						
"measures [a conversation's] grammaticality and fluency" (Juraska et al., 2019)						

# Comparing reported evaluation measures

Comparison of evaluations of QC named Coherence from four different papers, using the introduced QC and evaluation mode properties:

Definition of Coherence in paper	Quality criterion properties			Evaluation mode		
	Type of quality	Form/ content	Frame of reference	Obj. vs. subject.	Abs. vs. relative	Extr. vs. intrinsic
"[whether] the poem [is] thematically structured" (Van de Cruys, 2020)	Goodness					
"measures if a question is coherent with previous ones" (Chai & Wan, 2020)	Goodness					
"refers to the meaning of the generated sentence, so that a sentence with no meaning would be rated with a 1 and a sentence with a full meaning would be rated with a 5" (Barros et al., 2017)	Correctness					
"measures [a conversation's] grammaticality and fluency" (Juraska et al., 2019)	Goodness					
	Correctness					

# Comparing reported evaluation measures

Comparison of evaluations of QC named Coherence from four different papers, using the introduced QC and evaluation mode properties:

Definition of Coherence in paper	Quality criterion properties			Evaluation mode		
	Type of quality	Form/content	Frame of reference	Obj. vs. subject.	Abs. vs. relative	Extr. vs. intrinsic
"[whether] the poem [is] thematically structured" (Van de Cruys, 2020)	Goodness	Content				
"measures if a question is coherent with previous ones" (Chai & Wan, 2020)	Goodness	Content				
"refers to the meaning of the generated sentence, so that a sentence with no meaning would be rated with a 1 and a sentence with a full meaning would be rated with a 5" (Barros et al., 2017)	Correctness	Content				
"measures [a conversation's] grammaticality and fluency" (Juraska et al., 2019)	Goodness	Form and Content				
	Correctness	Form				

# Comparing reported evaluation measures

Comparison of evaluations of QC named Coherence from four different papers, using the introduced QC and evaluation mode properties:

Definition of Coherence in paper	Quality criterion properties			Evaluation mode		
	Type of quality	Form/content	Frame of reference	Obj. vs. subject.	Abs. vs. relative	Extr. vs. intrinsic
"[whether] the poem [is] thematically structured" (Van de Cruys, 2020)	Goodness	Content	None			
"measures if a question is coherent with previous ones" (Chai & Wan, 2020)	Goodness	Content	External FoR			
"refers to the meaning of the generated sentence, so that a sentence with no meaning would be rated with a 1 and a sentence with a full meaning would be rated with a 5" (Barros et al., 2017)	Correctness	Content	None			
"measures [a conversation's] grammaticality and fluency" (Juraska et al., 2019)	Goodness	Form and Content	None			
	Correctness	Form				

# Comparing reported evaluation measures

Comparison of evaluations of QC named Coherence from four different papers, using the introduced QC and evaluation mode properties:

Definition of Coherence in paper	Quality criterion properties			Evaluation modes		
	Type of quality	Form/content	Frame of reference	Obj. vs. subject.	Abs. vs. relative	Extr. vs. intrinsic
"[whether] the poem [is] thematically structured" (Van de Cruys, 2020)	Goodness	Content	None	Subjective	Absolute	Intrinsic
"measures if a question is coherent with previous ones" (Chai & Wan, 2020)	Goodness	Content	External FoR	Subjective	Absolute	Intrinsic
"refers to the meaning of the generated sentence, so that a sentence with no meaning would be rated with a 1 and a sentence with a full meaning would be rated with a 5" (Barros et al., 2017)	Correctness	Content	None	Subjective	Absolute	Intrinsic
"measures [a conversation's] grammaticality and fluency" (Juraska et al., 2019)	Goodness	Form and Content	None	Subjective	Absolute	Intrinsic
	Correctness	Form				

# Comparing reported evaluation measures

- It can be difficult to find enough information in papers to decide what evaluation measure is assessed in a given reported evaluation, due to the paucity of information typically shared about evaluations in NLP.
- Use these sources of information in order of preference:
  - Question asked of evaluators
  - Instructions/definitions given to evaluators
  - Definition of QC provided in paper
  - Name of QC provided in paper

# Creating new evaluation measures

- The quality criteria properties and evaluation modes can also be used to systematically identify which aspects of quality we want to evaluate.
- **Type of quality:** Are we interested in the (i) correctness of outputs (where we know when a system output is maximally correct), (ii) their goodness where one end of the scale is always better than the other, or (iii) a feature where either end of the scale may be considered better depending on evaluation context.

If it's more than one of these, we need to assess multiple quality criteria.

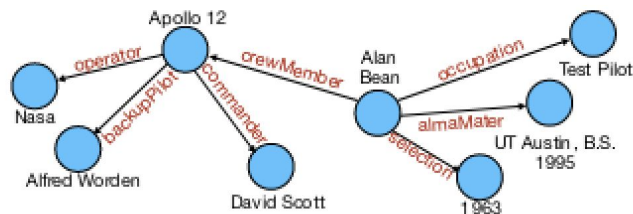
- **Frame of reference:** Are we interested in an aspect of quality that (i) can be assessed solely based on the output, (ii) relates output quality to the given input, or (iii) relates output quality to something external to the system.
- **Aspect of outputs:** What needs to be taken into account in assessing the aspect of quality we're interested in: (i) the form of outputs, (ii) the content/meaning of outputs, or (iii) both form and content.



# Creating new evaluation measures

For example, in a table-to-text generation task such as WebNLG:

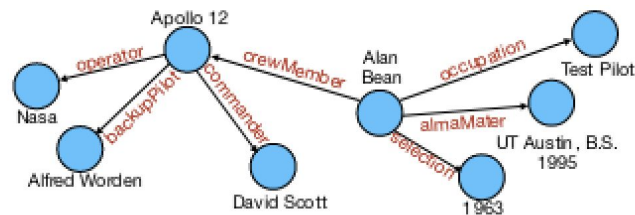
- We would be interested in whether the output text expresses all and only the content of the input triples (graph, table). We therefore select:
  - *Correctness* for type of quality, and
  - *Relative to Input* for frame of reference.
- To assess this kind of correctness, we only need to look at the content of the output, so we select:
  - *Content Only* for aspect of outputs.
- You could then opt to select *Input Coverage* (which has all the above properties) as the quality criterion (definition: *correctly expresses all and only the content of the input*).
- However, this will place a high cognitive load on evaluators for longer texts (such as in WebNLG), so you may decide to break it into two finer grained quality criteria such as *Absence of Additions*, and *Absence of Omissions*.



*Alan Bean graduated from UT Austin in 1955 with a Bachelor of Science degree. He was hired by NASA in 1963 and served as a test pilot. Apollo 12's backup pilot was Alfred Worden and was commanded by David Scott.*

# Creating new evaluation measures

- Let's use *Input Coverage* as an example, and select some suitable evaluation modes.
- Considering **absolute vs. relative**: could choose either, but for correctness-type criteria, *Absolute* assessment (each system separately) is comparatively straightforward and gives more information.
- **Intrinsic vs. extrinsic**: hard to conceive of an evaluation method that would fully assess *Input Coverage* extrinsically; *Intrinsic* would typically be chosen.
- **Objective vs. subjective**: in human evaluation, objective assessment of *Input Coverage* is also hard to conceive, and *Subjective* would be the default.



*Alan Bean graduated from UT Austin in 1955 with a Bachelor of Science degree. He was hired by NASA in 1963 and served as a test pilot. Apollo 12's backup pilot was Alfred Worden and was commanded by David Scott.*

# Overview

## Unit 3: Quality criteria and evaluation modes

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Research question(s) and hypotheses
3. Quality criteria and evaluation modes
4. Using QC properties and evaluation modes:
  - Comparing evaluation measures in existing evaluations;
  - Creating evaluation measures for new evaluations.
5. Connection with formulating the research question
6. Unit summary and pointers to other units
7. References
8. Practical task

# Incorporation into the research question

- Continuing with the example from the last section, we have chosen *absolute, subjective and intrinsic Input Coverage* as our evaluation measure.
- Analogously to our example research questions from the start of the unit, we then end up with this choice of research questions (if we're comparing two systems):
  - A. Is  $M_{new}$  better in terms of *absolute, subjective and intrinsic Input Coverage* than  $M_{old}$ ?
  - B. Which of  $M_{new}$  and  $M_{old}$  is better in terms of *absolute, subjective and intrinsic Input Coverage*?
- We have thus defined an evaluation measure  $m$  (let's call it *asi-IC*), but still need to create the evaluation method  $E_{asi-IC}$  to obtain measured values  $v_i$  for each system  $M_i$  (represented by its test set outputs  $o_i^s$ ) we wish to compare, or:

$$E_{asi-IC} \cdot (o_i^s, s) \mapsto v_i$$

# Moving on to experiment design

- *Input Coverage* is unlikely to be the only aspect of quality we would want to assess in a table-to-text task; *Fluency* is also typically used.
- Having selected our quality criteria and evaluation modes for each evaluation measure we wish to assess, and formulated our research questions, we move on to the rest of the experiment design phase, where we select rating instruments, response collection, interface design, and other **experiment properties** for each evaluation measure.
- This will then give us the fully specified evaluation methods  $E_{asi-IC}$  and  $E_{Fluency}$ .

# Overview

## Unit 3: Quality criteria and evaluation modes

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Research question(s) and hypotheses
3. Quality criteria and evaluation modes
4. Using QC properties and evaluation modes:
  - Comparing evaluation measures in existing evaluations;
  - Creating evaluation measures for new evaluations.
5. Connection with formulating the research question
6. Unit summary and pointers to other units
7. References
8. Practical task

# Unit summary

- Unit 3 focussed on evaluation measures, hypothesis testing, and the connection between them, in NLP evaluation.
- We construed evaluation measures as being composed of a quality criterion and three evaluation modes.
- Statistical hypothesis testing begins with formulating a research question incorporating the evaluation measure and a set of systems we wish to compare, typically asking for pairs of systems which is better in terms of the evaluation measure.
- We formulate the corresponding null hypothesis and at least one alternative hypothesis, again incorporating evaluation measure and systems.
- Using an appropriate test statistic and p-value computed on it, we test for the probability of observing results such as the ones obtained exceeding a given threshold, typically 0.05.
- We introduced quality criterion and evaluation mode properties and demonstrated two use cases with examples:
  - Comparing existing evaluations;
  - Devising new evaluation measures.

# Pointers to other units

- Unit 4 → the remaining steps in Phase I (Design) including rating instrument used to assess an evaluation measure.
- Unit 5 → more on hypothesis testing and significance testing.
- Unit 6 → implementation of rating instrument.



# Overview

## Unit 3: Quality criteria and evaluation modes

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Research question(s) and hypotheses
3. Quality criteria and evaluation modes
4. Using QC properties and evaluation modes:
  - Comparing evaluation measures in existing evaluations;
  - Creating evaluation measures for new evaluations.
5. Connection with formulating the research question
6. Unit summary and pointers to other units
7. References
8. Practical task

# References

Essential:

[Twenty Years of Confusion in Human Evaluation: NLG needs evaluation sheets and standardised definitions.](#)

D Howcroft, A Belz, M Clinciu, D Gkatzia, S Hasan, S Mahamood, S Mille, E. van Miltenburg, S. Santhanam, V. Rieser. INLG'20.

[Disentangling the Properties of Human Evaluation Methods: A Classification System to Support Comparability, Meta-Evaluation and Reproducibility Testing.](#)

A Belz, S Mille, D Howcroft. INLG'20.

[QCET: An Interactive Taxonomy of Quality Criteria for Comparable and Repeatable Evaluation of NLP Systems.](#)

A Belz, S Mille, Craig Thomson. INLG'24.

Further reading:

[The Human Evaluation Datasheet: A Template for Recording Details of Human Evaluation Experiments in NLP.](#)

A. Shimorina and A. Belz. 2022. HumEval'22.

[The logic of inductive inference.](#)

Fisher, Ronald A. *Journal of the royal statistical society* 98.1 (1935): 39-82.

# Overview

## Unit 3: Quality criteria and evaluation modes

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Research question(s) and hypotheses
3. Quality criteria and evaluation modes
4. Using QC properties and evaluation modes:
  - Comparing evaluation measures in existing evaluations;
  - Creating evaluation measures for new evaluations.
5. Connection with formulating the research question
6. Unit summary and pointers to other units
7. References
8. Practical task

# The QCET tool

- The Taxonomy of Quality Criteria for Evaluations (QCET) tool can be used to navigate the taxonomy tree of standardised quality criteria (extended from Howcroft et al., 2020).
- The tool allows users to:
  - Navigate the taxonomy tree by collapsing and expanding child nodes, where each node represents a quality criterion.
  - Prune branches of the taxonomy tree based on drop down options.
  - View full details of any quality criterion, including example questions in different evaluation modes.

# Example quality criteria

## Showing Node QCO-f-1

Parent

QCO-f : Correctness of outputs in their own right (form).

Node Details

### QCO-f-1 : Grammaticality

**Definition:**

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

**Suggested question to evaluators in subjective, intrinsic, absolute mode:**

*To what degree is this output free of grammatical errors, looking at its form only and ignoring its content/meaning?*

**Suggested question to evaluators in subjective, intrinsic, relative mode:**

*Which of these outputs has fewer grammatical errors, looking at its form only and ignoring its content/meaning?*

**Additional notes and information:**


*None.*

Children 

None


# Example quality criteria

## Q : Quality of outputs

The overall quality of an output. 

[show details](#)

### QC : Correctness of outputs

The degree to which an output is correct. 

[show details](#)

### QG : 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. 


[show details](#)

### QF : Feature-type criteria



[show details](#)

## Q : Quality of outputs

The overall quality of an output. 

[show details](#)

### QC : Correctness of outputs

The degree to which an output is correct. 

[show details](#)

#### QCO : Correctness of outputs in their own right

The degree to which an output is correct, considering only the output. 

[show details](#)

#### QCO-f : Correctness of outputs in their own right (form)

The degree to which an output is correct, considering only the output, and assessed on its form only. 

[show details](#)

##### QCO-f-1 : Grammaticality

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

[hide details](#)

**Suggested question to evaluators in subjective, intrinsic, absolute mode:**

*To what degree is this output free of grammatical errors, looking at its form only and ignoring its content/meaning?*


**Suggested question to evaluators in subjective, intrinsic, relative mode:**

*Which of these outputs has fewer grammatical errors, looking at its form only and ignoring its content/meaning?*

**Additional notes and information:**


None.

##### QCO-f-2 : Spelling accuracy

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

[show details](#)

## Q : Quality of outputs

The overall quality of an output. 

[show details](#)

### QF : Feature-type criteria



[show details](#)

### QFO : Feature-type criteria assessed looking at outputs in their own right



[show details](#)

#### QFO-1 : Text Property

The degree to which an output has a specific property (not a controlled feature). Open class criterion. 

[show details](#)

#### QFO-1-2 : Text Property [Complexity/simplicity]

The degree to which an output is complex/simple. 

[show details](#)


#### QFO-1-2-f : Text Property [Complexity/simplicity (form)]

The degree to which an output is expressed in complex/simple terms. 

[show details](#)




## QGO-b : 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. 


[show details](#)

### QGO-b-1 : Readability

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


[show details](#)

### QGO-b-2 : Fluency

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


[show details](#)

### QGO-b-3 : Understandability

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

[show details](#)

### QGO-b-4 : 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. 

[show details](#)

# Taxonomy of Quality Criteria For Evaluations (QCET) Tool

## Instructions

Click on the headers below to show the instructions.

The Underlying QC Taxonomy	▼
The QCET Tool	▼
Nodes	▼
Use cases	▼
Extensibility	▼

## Taxonomy Tree

Prune tree by level 1 QC classes (Correctness, Goodness, Features)

Show all Level 1 QC Classes ▼

Prune tree by level 2 QC classes (In its own Right, Relative to Input, Relative to External Frame of Reference)

Show all Level 2 QC Classes ▼

Prune tree by form vs. content QC classes (Form, Content, Both)

Show all Form/Content/Both QC Classes ▼

Q : Quality of outputs ⊕

The overall quality of an output. 

[show details](#)

# Practical Task

*Classifying reported evaluation methods with the QCET Taxonomy tool.*

For each of the evaluation measures below, use QCET to identify the (one or more) standard quality criteria that are evaluated in it, along with the applicable evaluation modes (absolute or relative, subjective or objective, intrinsic or extrinsic)

- **Fluency** in: Qian Yu, Lidong Bing, Qiong Zhang, Wai Lam, and Luo Si. 2020. [Review-based question generation with adaptive instance transfer and augmentation](#). In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 280–290. Association for Computational Linguistics.
- **Dialogue Efficiency** in: Yan Qu and Nancy Green. 2002. [A constraint-based approach for cooperative information-seeking dialogue](#). In Proceedings of the International Natural Language Generation Conference, pages 136–143, Harriman, New York, USA. Association for Computational Linguistics.

For completing the practical task, you should use the annotated copies of the above papers that can be found [in the Unit 3 Resources folder](#).