Human Evaluation of NLP System Quality

INLG Tutorial, 24th September 2024

Unit 2: Development and Components of Human Evaluations

Link to Unit 2 Resources

Overview

Unit 2: Development and components of human evaluations

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Standard terminology and definitions
3. Components of a (ready-to-run) human evaluation
4. Steps in creating and running a human evaluation
5. Example human evaluation in terms of the standard components
6. Unit summary and pointers to other units
7. References

Overview

Unit 2: Development and components of human evaluations

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Standard terminology and definitions
3. Components of a (ready-to-run) human evaluation
4. Steps in creating and running a human evaluation
5. Example human evaluation in terms of the standard components
6. Unit summary and pointers to other units
7. References

Unit aims and learning outcomes

- The aims of Unit 2 are:
    - To introduce core standard terminology used throughout the tutorial.
    - To examine the components and processes common to all human evaluations, introducing a standard framework comprising:
        - A standard process diagram for human evaluations, and
        - A standard decomposition of the steps in creating and running a human evaluation.
- After completion of the unit, participants will be able to understand and:
    - Use standard terminology and definitions relating to human evaluation.
    - Apply the standard process diagram in designing evaluations with this structure.
    - Follow the four phases and any iterations over them in creating and running a human evaluation.

Prerequisites and connections with other units

- Prerequisite(s) of Unit 2: Unit 1 is helpful but not required.
- Unit 2 is a prerequisite of Units 3–8, as it introduces standard terminology and concepts, as well as the structural and procedural framework, used by later units.

Overview

Unit 2: Development and components of human evaluations

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Standard terminology and definitions
3. Components of a (ready-to-run) human evaluation
4. Steps in creating and running a human evaluation
5. Example human evaluation in terms of the standard components
6. Unit summary and pointers to other units
7. References

Terminology and definitions

- Using standard terminology with shared definitions is important across all of science.
- Terminologies in NLP/ML and computer science are more in flux than in other fields, leading to difficulties building on prior work.
- Here, we introduce more loosely explained terms we use in the tutorial, and general, more formally defined terms.
- **Evaluation**: in an NLP context, the assessment of *system quality* by automatic or manual assessment of system outputs or other aspects of system behaviour.
- **Human evaluation**: evaluation involving some form of human assessment or interaction.
- **System quality**: the level of correctness or goodness of *system outputs*, or the degree to which they achieve a given target feature, in terms of a given *quality criterion*.
- **System outputs**: most commonly, the (literal) outputs produced by a system when given an input; can be a sequence of outputs as in dialogue, or contextualised e.g. including interface captures.

Terminology and definitions

- A single **evaluation** *M* is a measurement in terms of measurand *m* performed on object *O* at time *t* under set of conditions *C*, returning a measured value *v*:

  $M: (m, O, t, C) \mapsto v$

  In NLP/ML, measurand ≅ *quality criterion*, object ≅ system, conditions ≅ *experiment properties*.
- An **evaluation experiment**, or simply **experiment**, is a coordinated set of evaluations, typically for multiple comparable systems and system outputs.
- **Quality criterion**: *what* is evaluated; a criterion in terms of which system quality is assessed; quality criteria are agnostic about *how* they are evaluated.
- **Evaluation modes**: absolute vs. relative, subjective vs. objective, intrinsic vs. extrinsic; need to be specified to turn a quality criterion into an *evaluation measure* that can be implemented; orthogonal to quality criteria, i.e. any given quality criterion can be combined with any modes.

Terminology and definitions

- **Experimental design** is the full specification of how to obtain a quantitative or qualitative response value for a given *evaluation measure*, yielding a fully specified *evaluation method*.

- In sum:

  Quality criterion + evaluation modes = **evaluation measure**;

  Evaluation measure + experimental design = **evaluation method**.

Overview

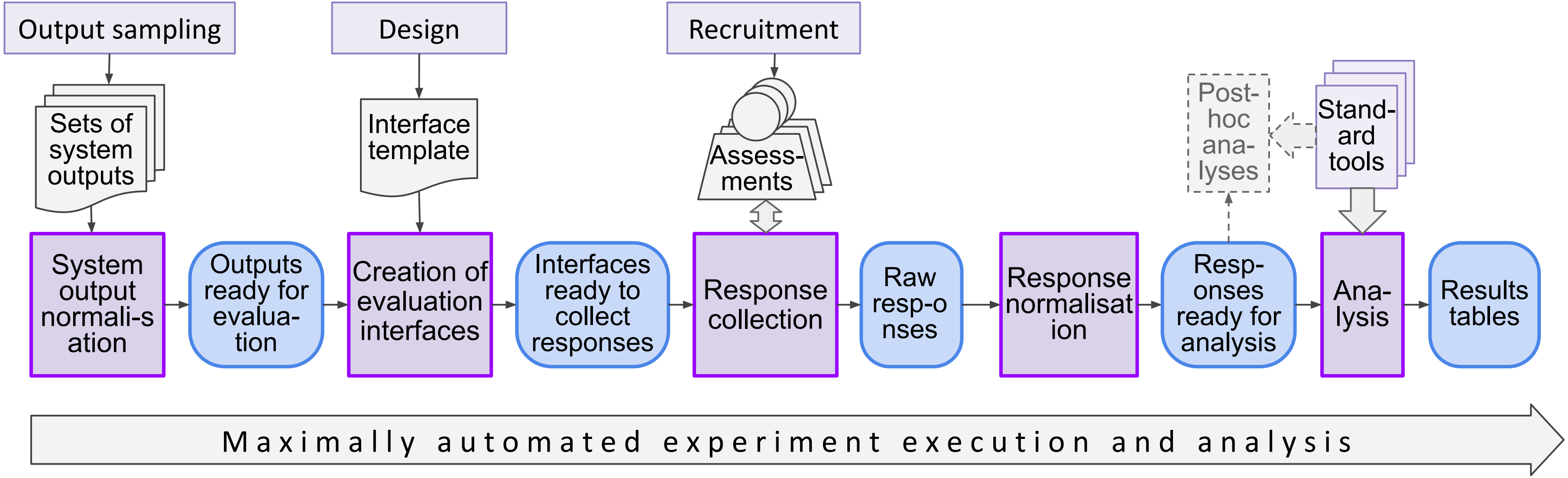
Unit 2: Development and components of human evaluations

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Standard terminology and definitions
3. Components of a (ready-to-run) human evaluation
4. Steps in creating and running a human evaluation
5. Example human evaluation in terms of the standard components
6. Unit summary and pointers to other units
7. References

Component vs. development perspective

- In considering what it takes to put together an evaluation experiment we will take two perspectives:
  - What the component processes are in an evaluation experiment ⟶ **process diagram** showing components of a human evaluation.
  - What phases we need to go through, and what tasks we need to perform in each phase, in developing the evaluation ⟶ **diagram of development phases**
- The process diagram will show the (initially empty) ‘containers’ that need to be filled step by step in the four development phases.
- We will then look briefly at what happens in each development phase (this unit).
- And then in more detail at options and good practice for each component process (Units 3–7).
- In the practical session, we will work through exercises focusing on the latter parts of a fully instantiated evaluation pipeline.

Components of a (ready-to-run) human evaluation



Output sampling

Design

Recruitment

Sets of system outputs

Interface template

Assessments

Post-hoc analyses

Standard tools

System output normali-sation

Outputs ready for evaluation

Creation of evaluation interfaces

Interfaces ready to collect responses

Response collection

Raw resp-onses

Response normalisation

Resp-onses ready for analysis

Analysis

Results tables

Maximally automated experiment execution and analysis

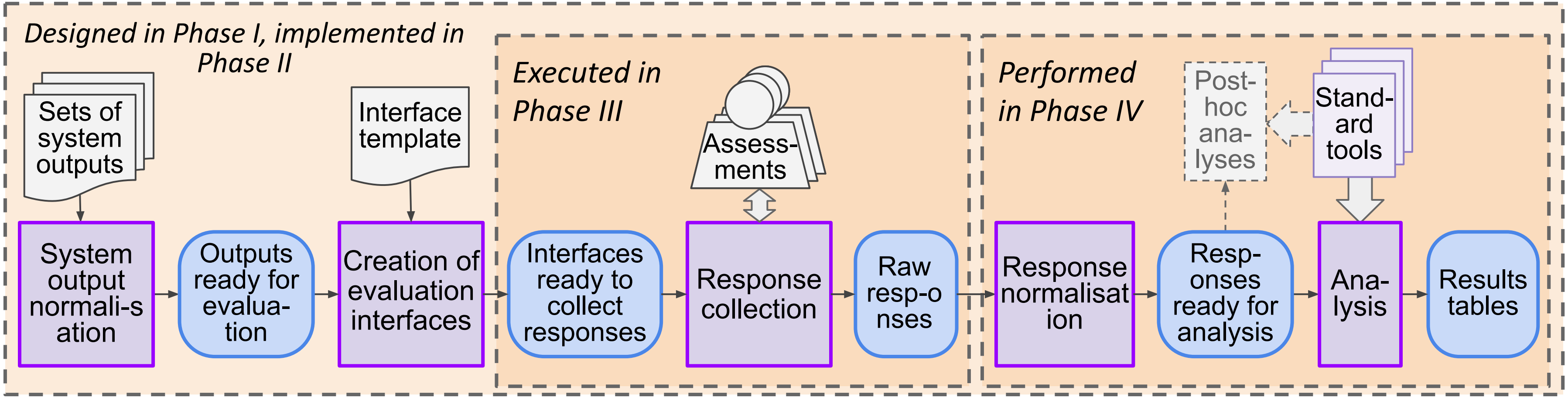
🟪 One of five core component processes in a human evaluation.

🟦 An output generated by one of the five core processes, becoming an input accepted by the next process.

▭ An input or resource used by one of the core processes, but created outside of the main pipeline.

▢ A contributory process that creates one of the inputs/resources.

Components of a (ready-to-run) human evaluation



Component processes and development phases of a human evaluation

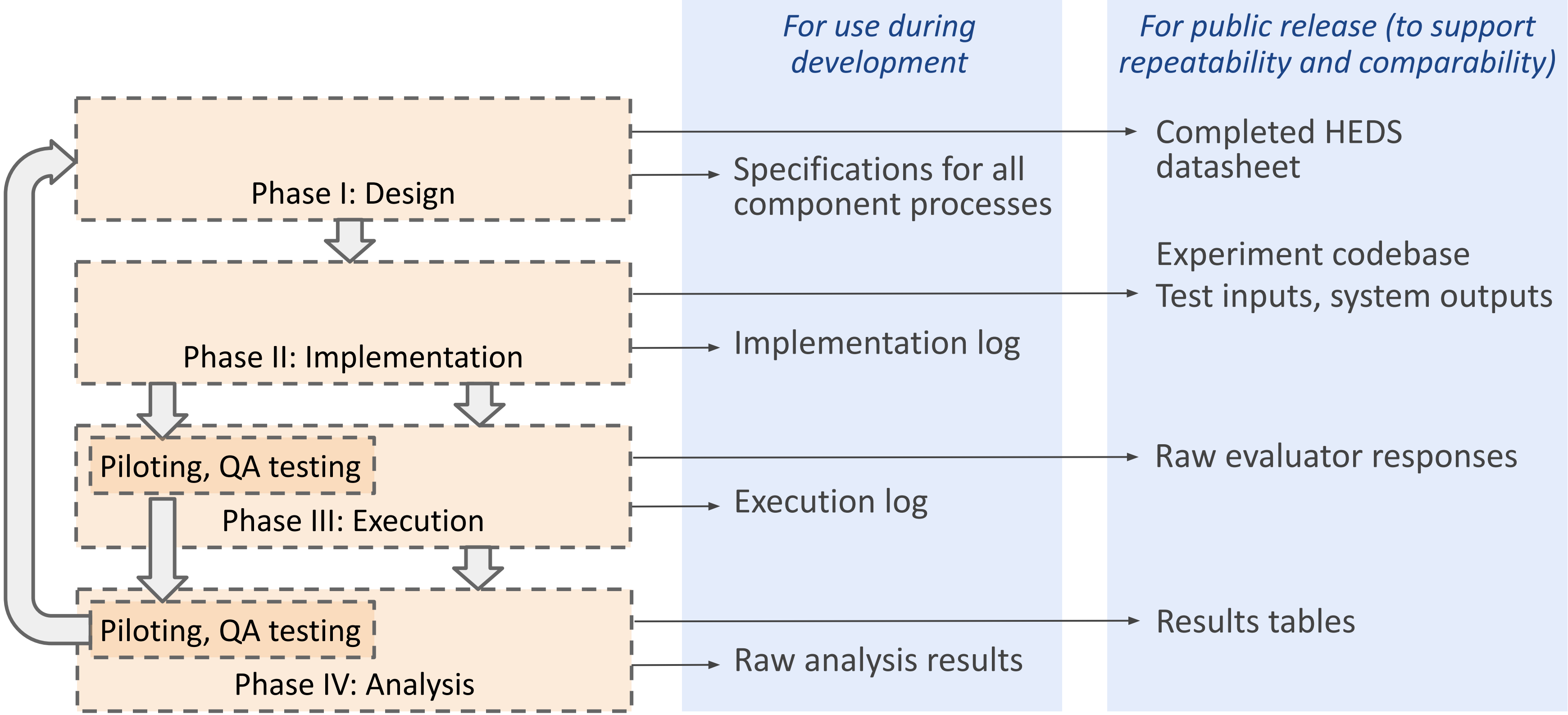
- Safest to fully automate all processes, including pipelining them together where possible.
- Store generated data structures in standard formats and locations.
- All component processes and external resources/inputs specified in Phase I (Design), implemented in Phase II (Implementation).
- Experiment run in Phase III (Execution); results analysed in Phase IV (Analysis).

Overview

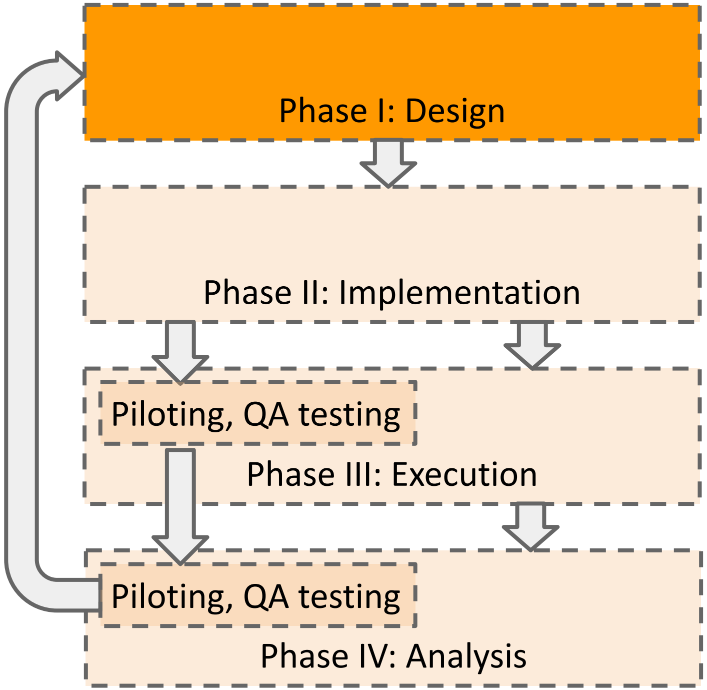
Unit 2: Development and components of human evaluations

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Standard terminology and definitions
3. Components of a (ready-to-run) human evaluation
4. Steps in creating and running a human evaluation
5. Example human evaluation in terms of the standard components
6. Unit summary and pointers to other units
7. References

Steps in creating and running a human evaluation



Steps in creating and running a human evaluation

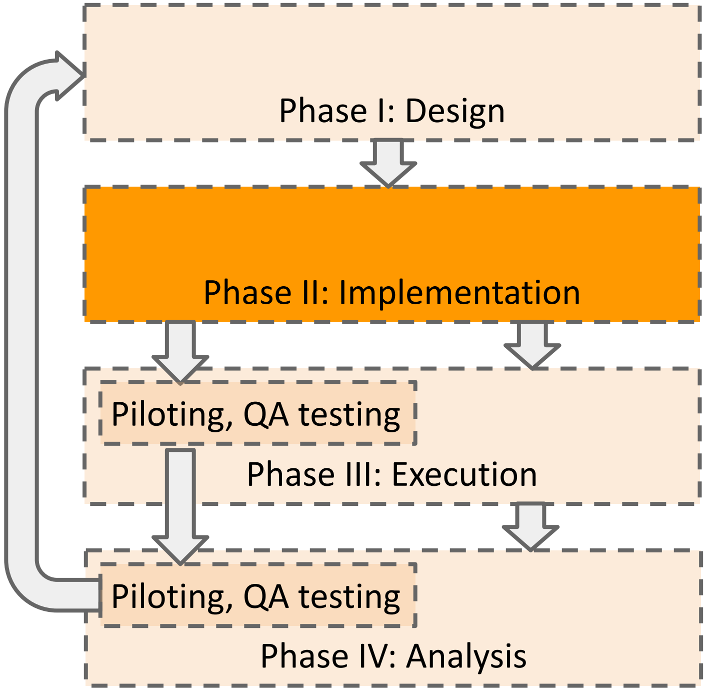


Phase I – Design

1. Research question(s) and hypotheses; Selection of systems; Quality criteria & evaluation modes
2. Number of outputs & evaluators
3. Output sampling
4. Rating instrument
5. Evaluator type & characteristics
6. Evaluator recruitment, training
7. Conditions during experiment
8. Quality assurance
9. Analysis
10. Impact assessment
11. Ethical review

Completing human evaluation datasheet (HEDS).

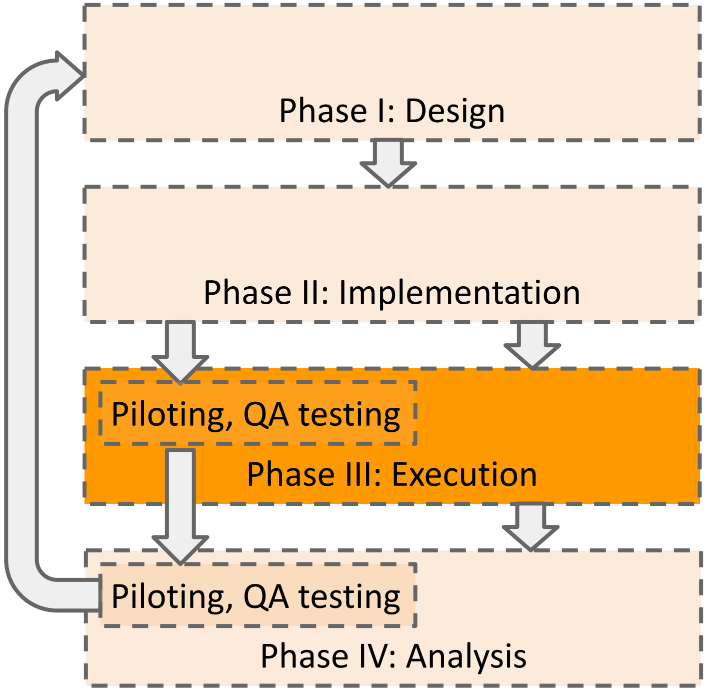
Steps in creating and running a human evaluation



Phase II – Implementation

- Implementation of code (or protocol) for:
  a. System output sampling.
  b. Output normalisation.
  c. Evaluation interface instantiation.
  d. Response collection, including evaluator monitoring.
  e. Evaluator training.
  f. Response normalisation.
  g. Aggregation and analysis of results.
- Create wrapper script(s) to call the five component processes, pipelining them where possible.
- Perform code testing, use code review, and adopt other good coding practices.
- Update human evaluation datasheet (HEDS).

Steps in creating and running a human evaluation



Phase III – Execution

Pre-final execution – iterate as necessary:

a. Perform interface robustness testing.
b. Run response collection as pilot experiment.
c. Test pilot responses for inter and intra-annotator agreement.
d. Collect feedback from pilot evaluators.
e. If feasible, test for reproducibility.
f. Collate and implement improvements.

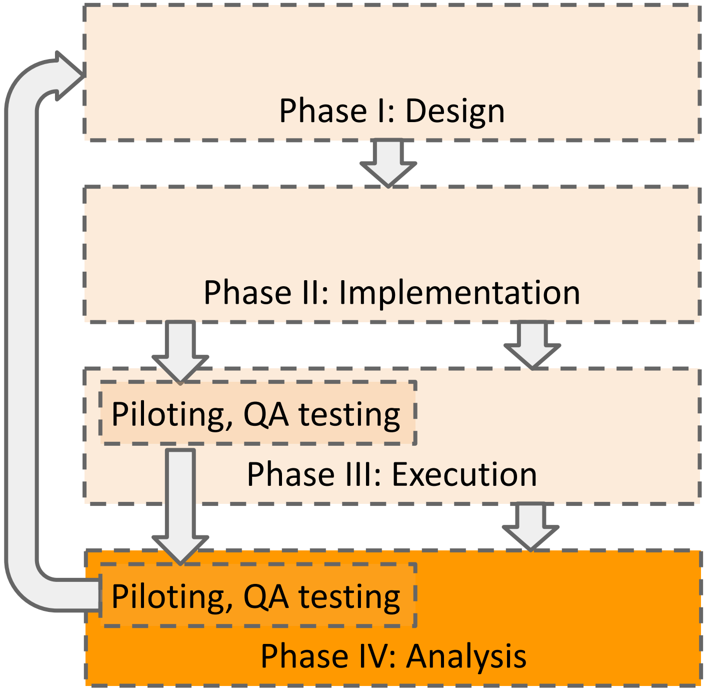
Final execution:

g. Complete preregistration with final HEDS sheet.
h. Run response collection (in pipeline) with full number of evaluators/items.

Post-final execution:

i. E.g. for reproducibility testing.

Steps in creating and running a human evaluation



Phase IV – Analysis

a. Run response normalisation and aggregation/analysis exactly as preregistered.

b. If needed, run additional posthoc tests, including multiple test corrections as needed.

c. Create new scripts to generate any additional tables needed.

d. Report results in two separate parts, always clearly stating which is which:

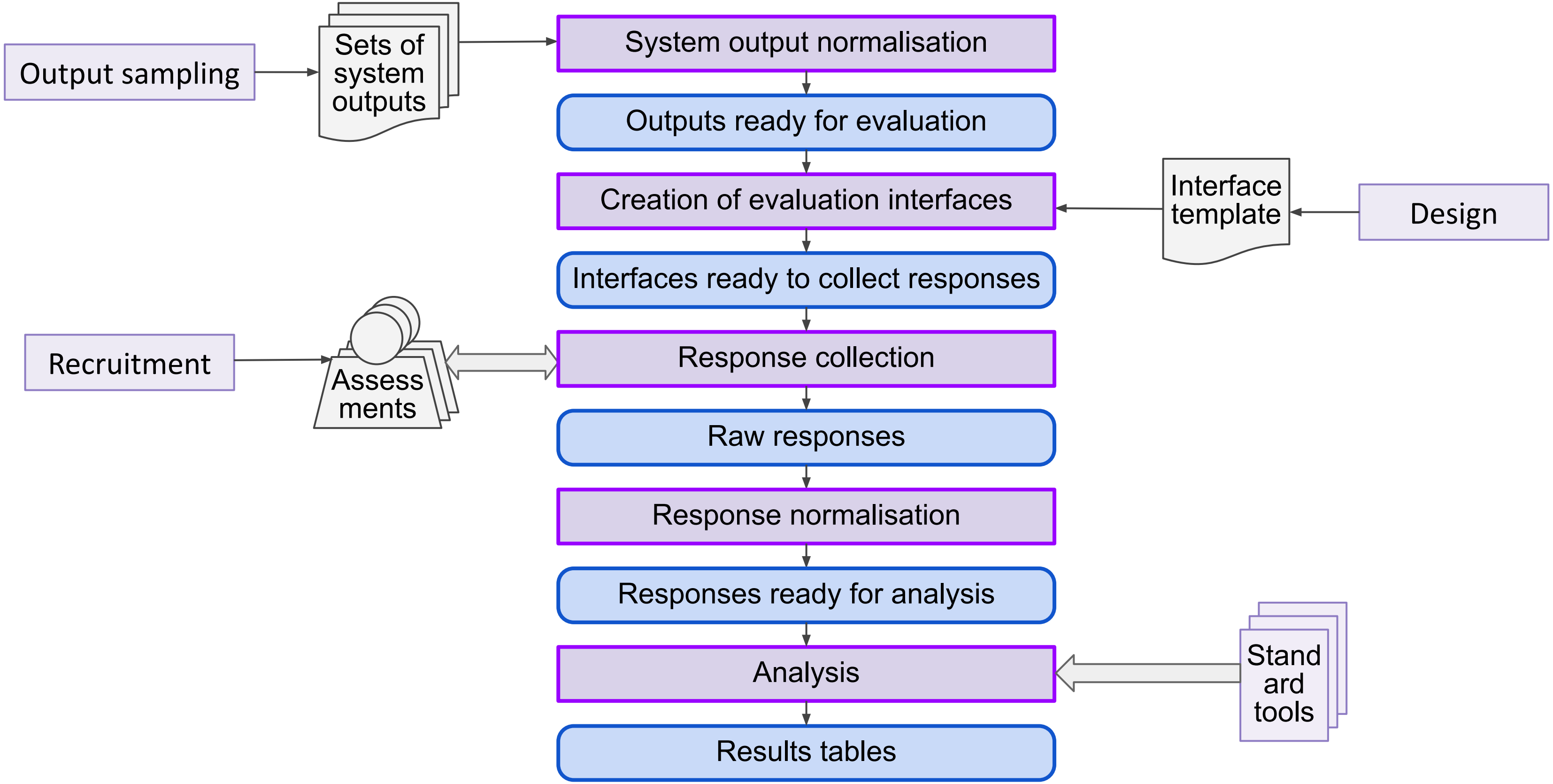
   - Preregistered results
   - Post-hoc results

Overview

Unit 2: Development and components of human evaluations

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Standard terminology and definitions
3. Components of a (ready-to-run) human evaluation
4. Steps in creating and running a human evaluation
5. Example human evaluation in terms of the standard components
6. Unit summary and pointers to other units
7. References

Example human evaluation – WebNLG 2023



Output sampling

Sets of system outputs

System output normalisation

Outputs ready for evaluation

Creation of evaluation interfaces

Interface template

Design

Interfaces ready to collect responses

Recruitment

Assess ments

Response collection

Raw responses

Response normalisation

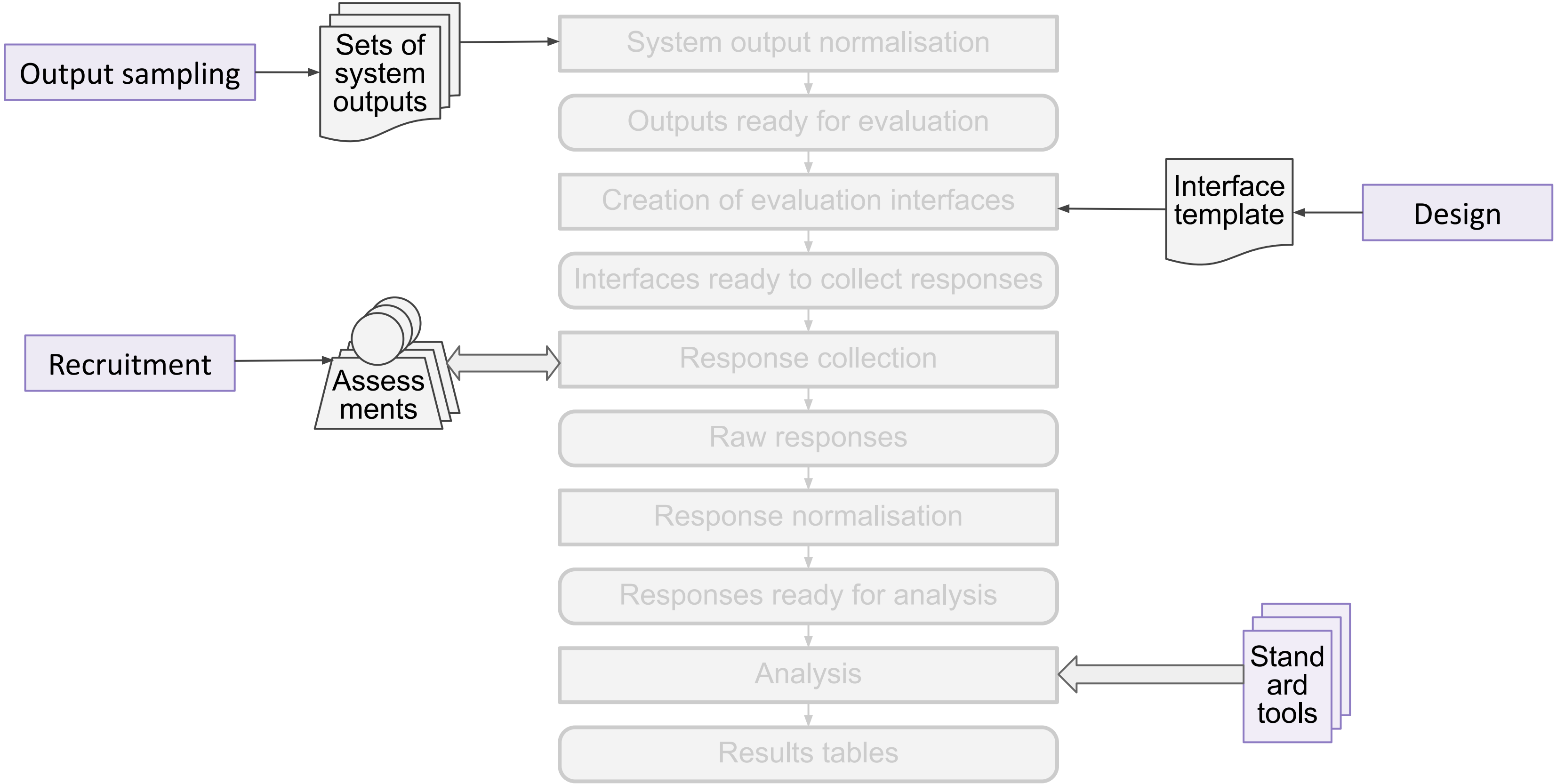
Responses ready for analysis

Analysis

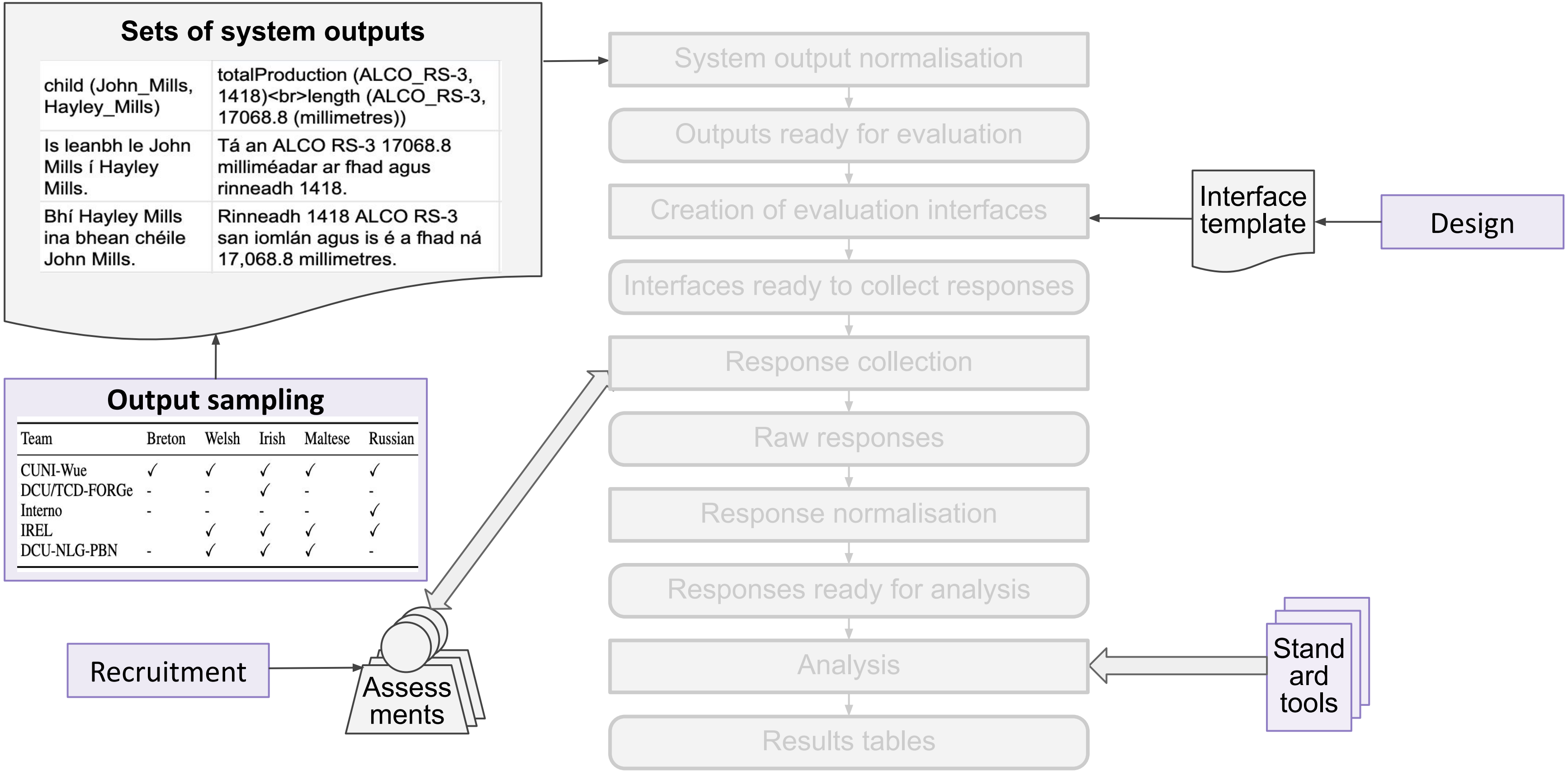
Stand ard tools

Results tables

Example human evaluation – WebNLG 2023



Example human evaluation – WebNLG 2023



Sets of system outputs

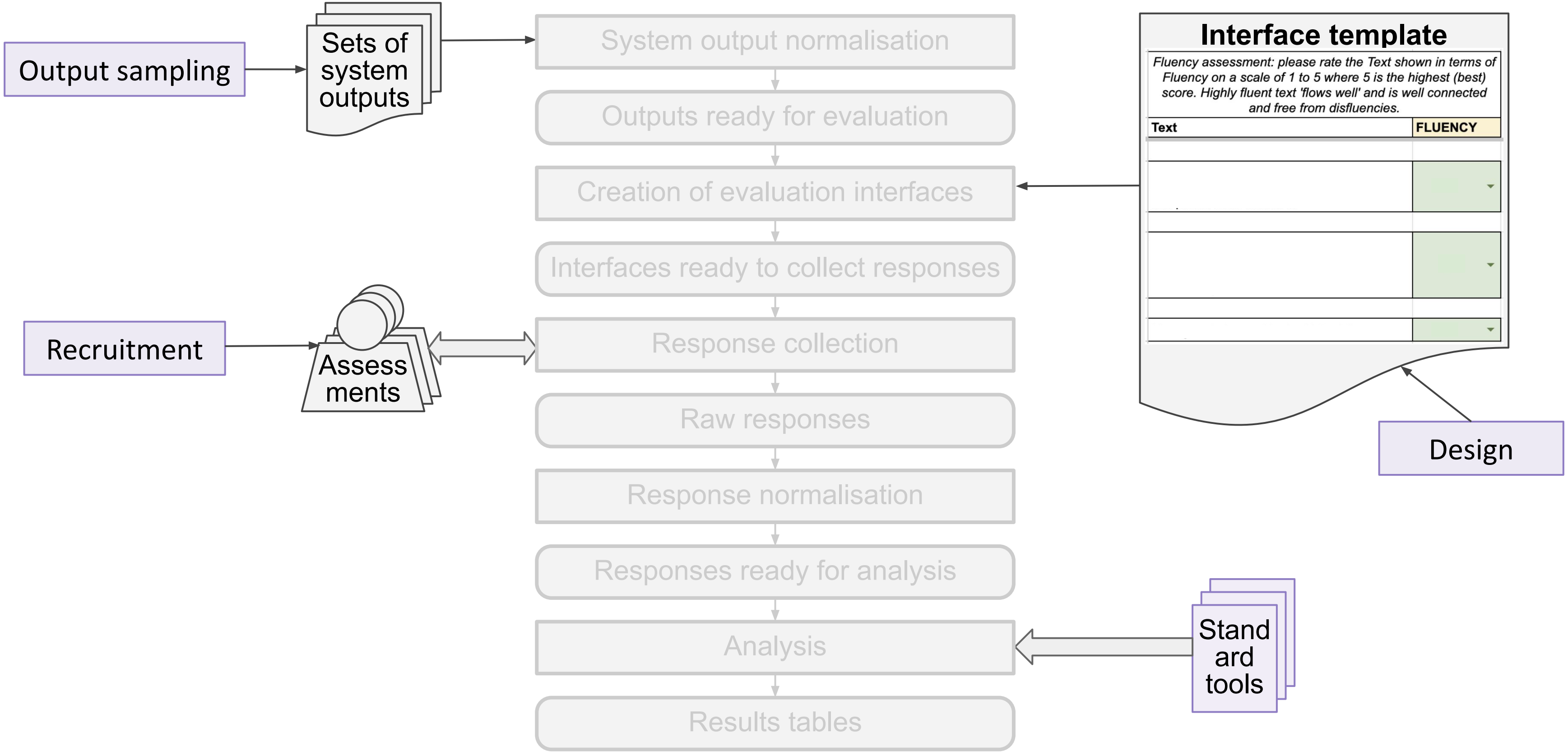
child (John_Mills, Hayley_Mills)	totalProduction (ALCO_RS-3, 1418) length (ALCO_RS-3, 17068.8 (millimetres))
Is leanbh le John Mills í Hayley Mills.	Tá an ALCO RS-3 17068.8 milliméadar ar fhad agus rinneadh 1418.
Bhí Hayley Mills ina bhean chéile John Mills.	Rinneadh 1418 ALCO RS-3 san iomlán agus is é a fhad ná 17,068.8 millimetres.

Output sampling

Team	Breton	Welsh	Irish	Maltese	Russian
CUNI-Wue	✓	✓	✓	✓	✓
DCU/TCD-FORGe	-	-	✓	-	-
Interno	-	-	-	-	✓
IREL		✓	✓	✓	✓
DCU-NLG-PBN	-	✓	✓	✓	-

Recruitment
Assessments
System output normalisation
Outputs ready for evaluation
Creation of evaluation interfaces
Interfaces ready to collect responses
Response collection
Raw responses
Response normalisation
Responses ready for analysis
Analysis
Results tables
Interface template
Design
Standard tools

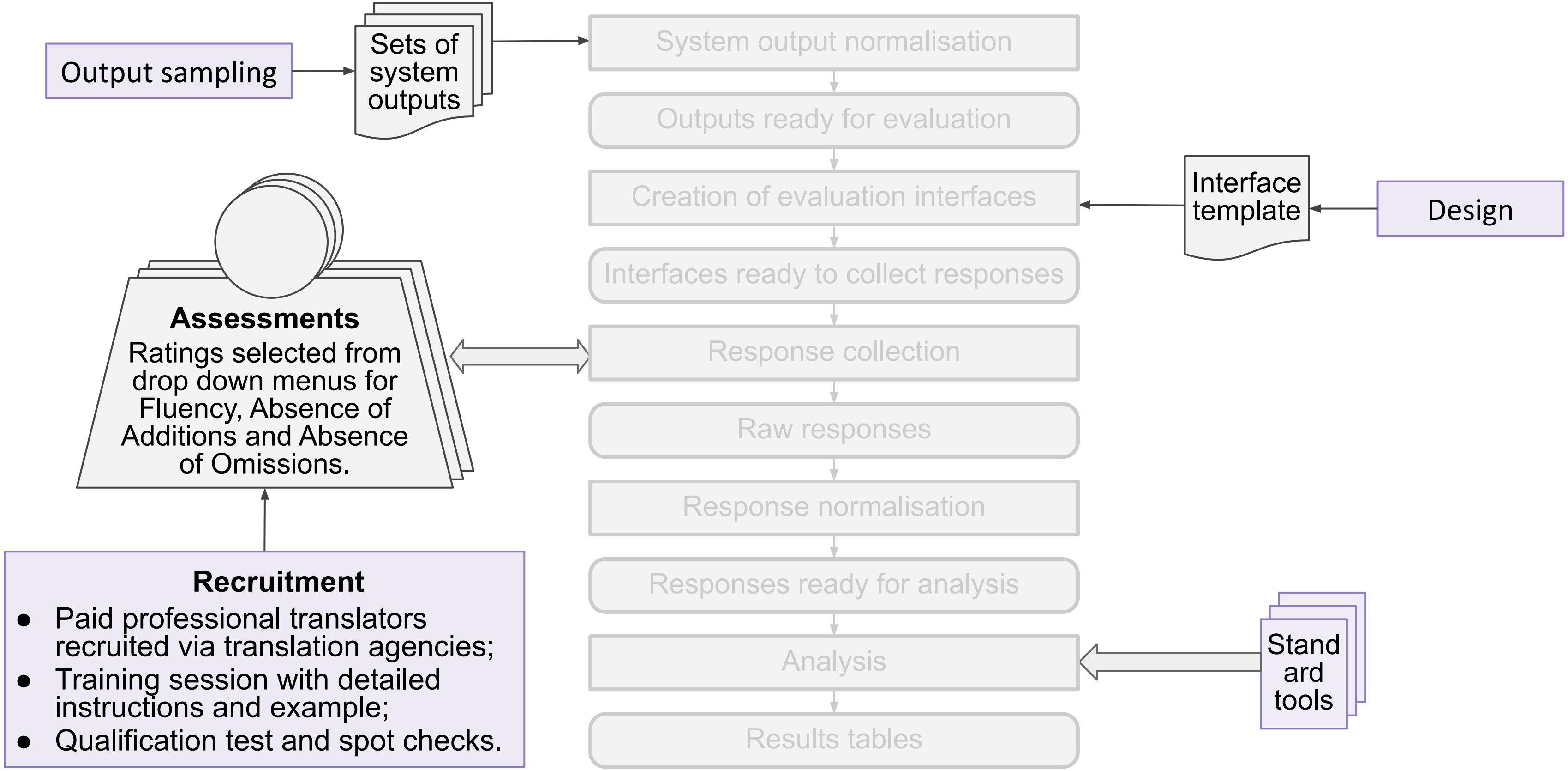
Example human evaluation – WebNLG 2023



Fluency assessment: please rate the Text shown in terms of Fluency on a scale of 1 to 5 where 5 is the highest (best) score. Highly fluent text 'flows well' and is well connected and free from disfluencies.

Text	FLUENCY

Example human evaluation – WebNLG 2023



Output sampling

Sets of system outputs

System output normalisation

Outputs ready for evaluation

Creation of evaluation interfaces

Interface template

Design

Interfaces ready to collect responses

Assessments

Ratings selected from drop down menus for Fluency, Absence of Additions and Absence of Omissions.

Response collection

Raw responses

Response normalisation

Responses ready for analysis

Analysis

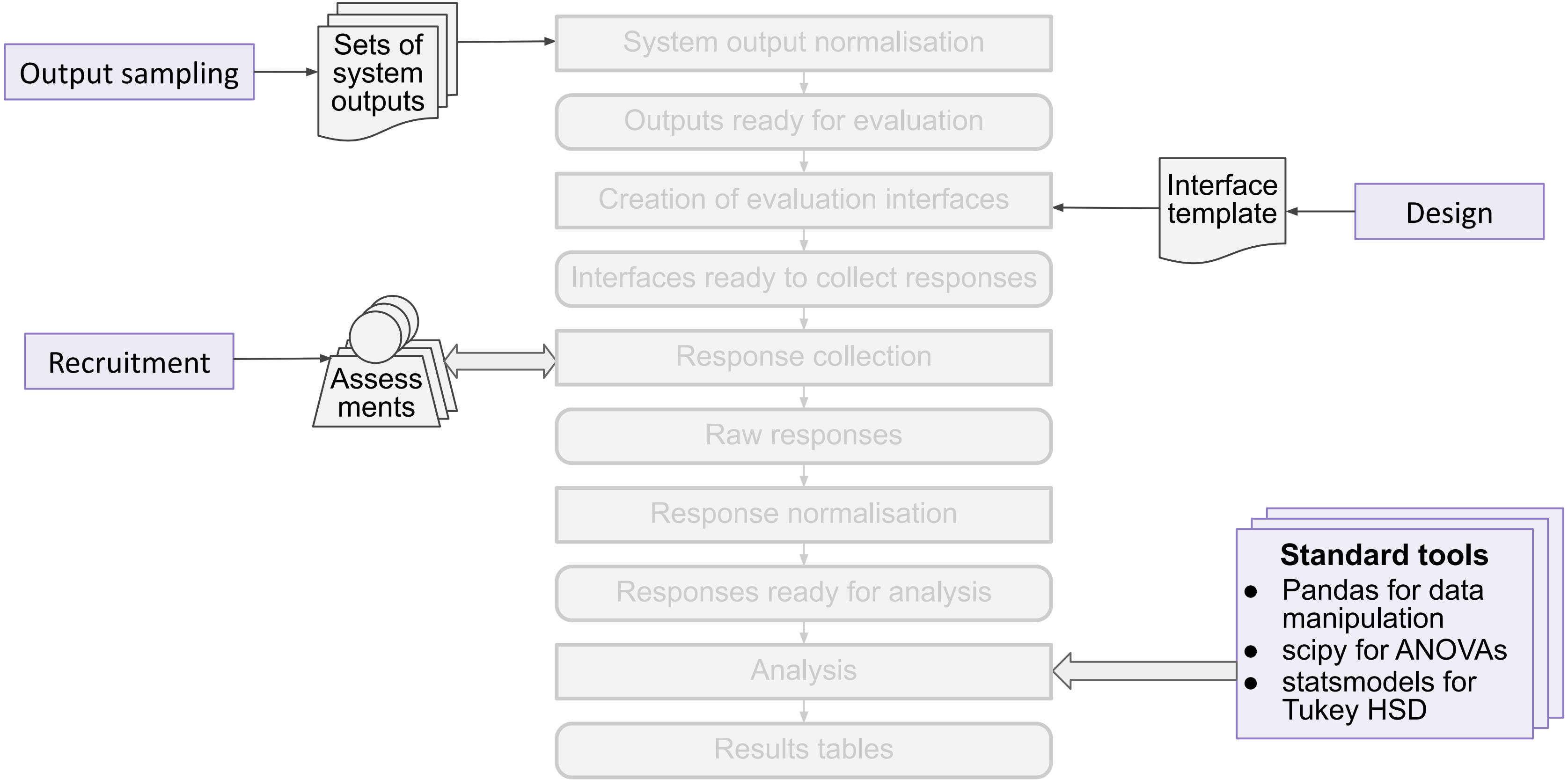
Standard tools

Results tables

Recruitment

- Paid professional translators recruited via translation agencies;
- Training session with detailed instructions and example;
- Qualification test and spot checks.

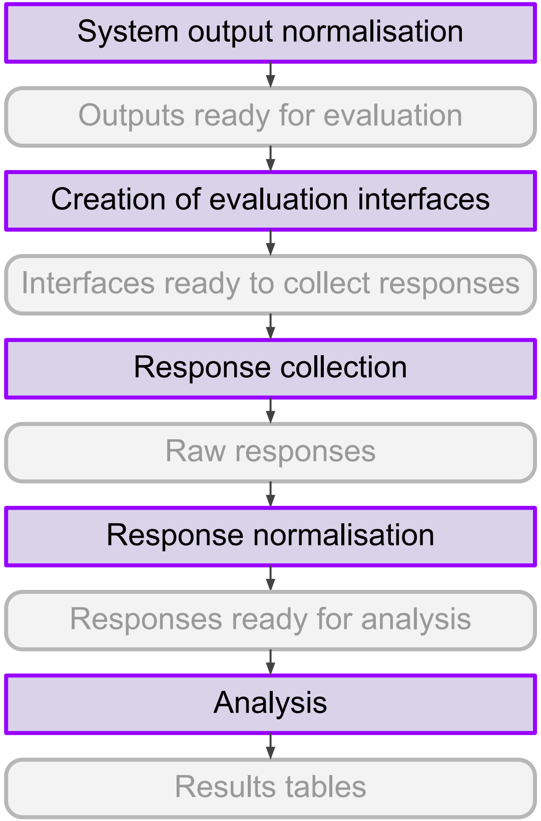
Example human evaluation – WebNLG 2023



Standard tools

- Pandas for data manipulation
- scipy for ANOVAs
- statsmodels for Tukey HSD

Example human evaluation – WebNLG 2023



System output normalisation

Outputs ready for evaluation

Creation of evaluation interfaces

Interfaces ready to collect responses

Response collection

Raw responses

Response normalisation

Responses ready for analysis

Analysis

Results tables

Output text were left as submitted; inputs were converted from "subject | predicate | object" to "predicate (subject, object)".

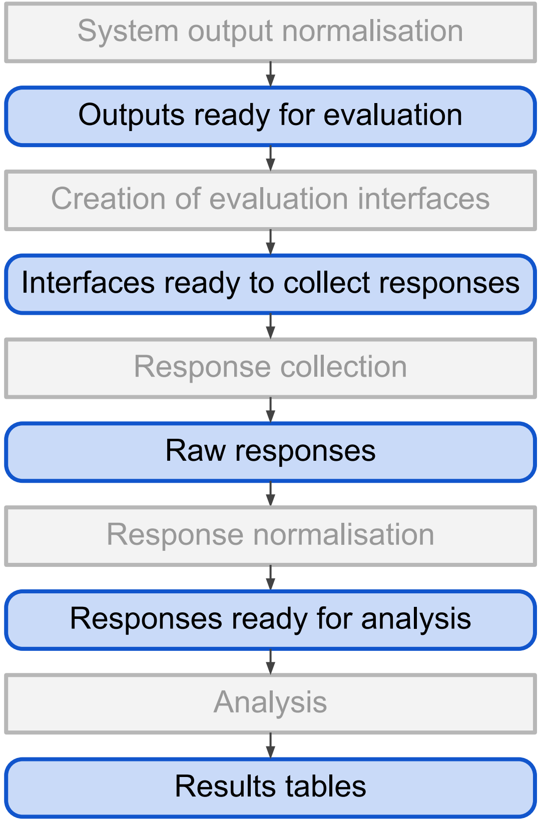
The Google sheet templates were instantiated with system inputs/outputs via Google Cloud API using Latin-square allocation. Each evaluator had one sheet.

Evaluators were given access to the instantiated Google sheets, and completed them in their own time.

No normalisation required as all scores were selected from drop-down menus. Scores were extracted with a script that kept track of evaluators, systems and evaluation criteria.

Python scripts applied to extracted scores to compute ANOVA test, Tukey HSD test, Pearson and Spearman correlations with standard libraries.

Example human evaluation – WebNLG 2023



System output normalisation

Outputs ready for evaluation

Creation of evaluation interfaces

Interfaces ready to collect responses

Response collection

Raw responses

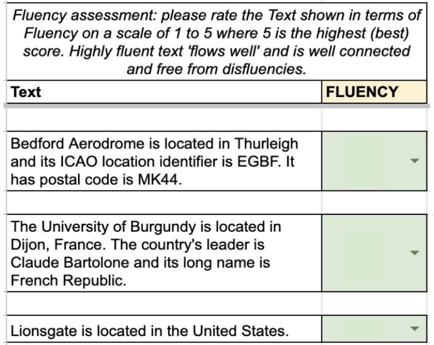
Response normalisation

Responses ready for analysis

Analysis

Results tables

Evaluator	Sample 0	Sample 1
E1	Bedford Aerodrome is located in Thurleigh and its ICAO location identifier is EGBF. It has postal code is MK44.	The University of Burgundy is located in Dijon, France. The country's leader is Claude Bartolone and its long name is French Republic.
E2	…	…

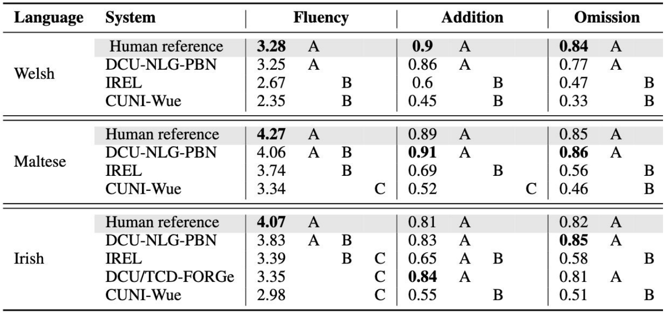


Fluency assessment: please rate the Text shown in terms of Fluency on a scale of 1 to 5 where 5 is the highest (best) score. Highly fluent text 'flows well' and is well connected and free from disfluencies.

Text	FLUENCY
Bedford Aerodrome is located in Thurleigh and its ICAO location identifier is EGBF. It has postal code is MK44.	
The University of Burgundy is located in Dijon, France. The country's leader is Claude Bartolone and its long name is French Republic.	
Lionsgate is located in the United States.	

Sample ID	Team 1	Team 1 evaluator	Team 2	Team 2 evaluator	Team 3	Team 3 evaluator
1	4	E1	4	E2	3	E3
2	3	E2	4	E3	5	E1
3	3	E3	5	E1	4	E2

Sample ID	System A	System B
0	5	4
1	2	5



Language	System	Fluency				Addition				Omission		
Welsh	Human reference	**3.28**	A			**0.9**	A			**0.84**	A	
	DCU-NLG-PBN	3.25	A			0.86	A			0.77	A	
	IREL	2.67		B		0.6		B		0.47		B
	CUNI-Wue	2.35		B		0.45		B		0.33		B
Maltese	Human reference	**4.27**	A			0.89	A			0.85	A	
	DCU-NLG-PBN	4.06	A	B		**0.91**	A			**0.86**	A	
	IREL	3.74		B		0.69		B		0.56		B
	CUNI-Wue	3.34			C	0.52			C	0.46		B
Irish	Human reference	**4.07**	A			0.81	A			0.82	A	
	DCU-NLG-PBN	3.83	A	B		0.83	A			**0.85**	A	
	IREL	3.39		B	C	0.65	A	B		0.58		B
	DCU/TCD-FORGe	3.35			C	**0.84**	A			0.81	A	
	CUNI-Wue	2.98			C	0.55		B		0.51		B

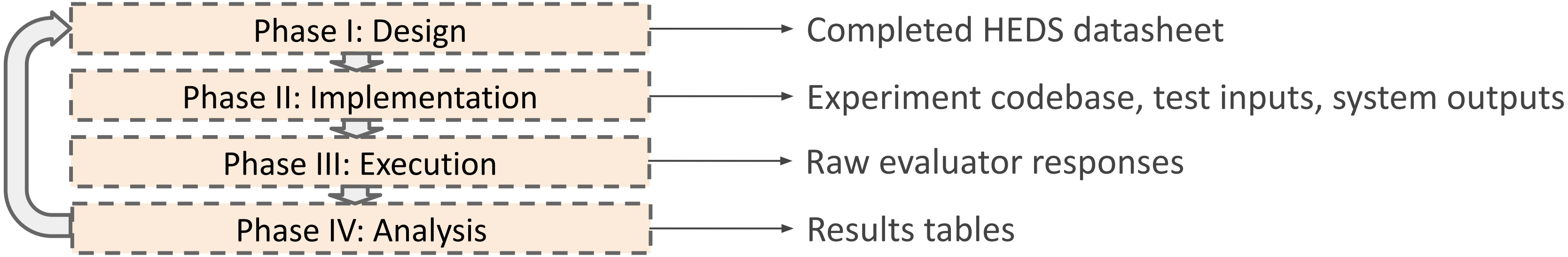
Overview

Unit 2: Development and components of human evaluations

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Standard terminology and definitions
3. Components of a (ready-to-run) human evaluation
4. Steps in creating and running a human evaluation
5. Example human evaluation in terms of the standard components
6. Unit summary and pointers to other units
7. References

Unit summary

- Standard terminology and definitions are important for comparability with and building on prior work.
- Human evaluations can be construed as comprising five core consecutive processes:
    1. System output normalisation
    2. Creation of evaluation interfaces
    3. Response collection
    4. Response normalisation
    5. Analysis
- A good approach to developing human evaluation experiments proceeds in 4 phases:



Phase I: Design → Completed HEDS datasheet

Phase II: Implementation → Experiment codebase, test inputs, system outputs

Phase III: Execution → Raw evaluator responses

Phase IV: Analysis → Results tables

Pointers to other units

Unit 3 → research question(s) and hypotheses; selection of quality criteria and evaluation modes (Phase I).

Unit 4 → rating instrument, response collection, interface design, evaluator recruitment, ethical considerations; HEDS (Phase I).

Unit 5 → all aspects of analysis (Phase IV).

Unit 6 → all aspects of implementation (Phase II).

Unit 7 → all aspects of execution (Phase III).

Overview

Unit 2: Development and components of human evaluations

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Standard terminology and definitions
3. Components of a (ready-to-run) human evaluation
4. Steps in creating and running a human evaluation
5. Example human evaluation in terms of the standard components
6. Unit summary and pointers to other units
7. References

References

Essential reading:

Disentangling the Properties of Human Evaluation Methods: A Classification System to Support Comparability, Meta-Evaluation and Reproducibility Testing. A Belz, S Mille, D Howcroft. International Natural Language Generation Conference 2020 (INLG'20).

Further reading:

The Human Evaluation Datasheet: A Template for Recording Details of Human Evaluation Experiments in NLP. Anastasia Shimorina and Anya Belz. 2022. 2nd Workshop on Human Evaluation of NLP Systems (HumEval).

QCET: An Interactive Taxonomy of Quality Criteria for Comparable and Repeatable Evaluation of NLP Systems. A Belz, S Mille, Craig Thomson. INLG 2024, to appear.

Belz, A. (2022). A metrological perspective on reproducibility in NLP. *Computational Linguistics*, *48*(4), 1125-1135.