A Review of Human Evaluation for Style Transfer

Eleftheria Briakou
ebriakou@cs.umd.edu

Sweta Agrawal
sweagraw@cs.umd.edu

Ke Zhang
kzhang@dataminr.com

Joel Tetreault
jtetreault@dataminr.com

Marine Carpuat
marine@cs.umd.edu

What is style?

“style is an intuitive notion involving the manner in which something is said”

McDonald and Pustejovsky. 1985

Style Transfer in NLP

INFORMAL TO FORMAL

INPUT: Gotta see both sides of the story

OUTPUT: You have to consider both sides of the story.

POSITIVE TO NEGATIVE

INPUT: The screen is just the right size.

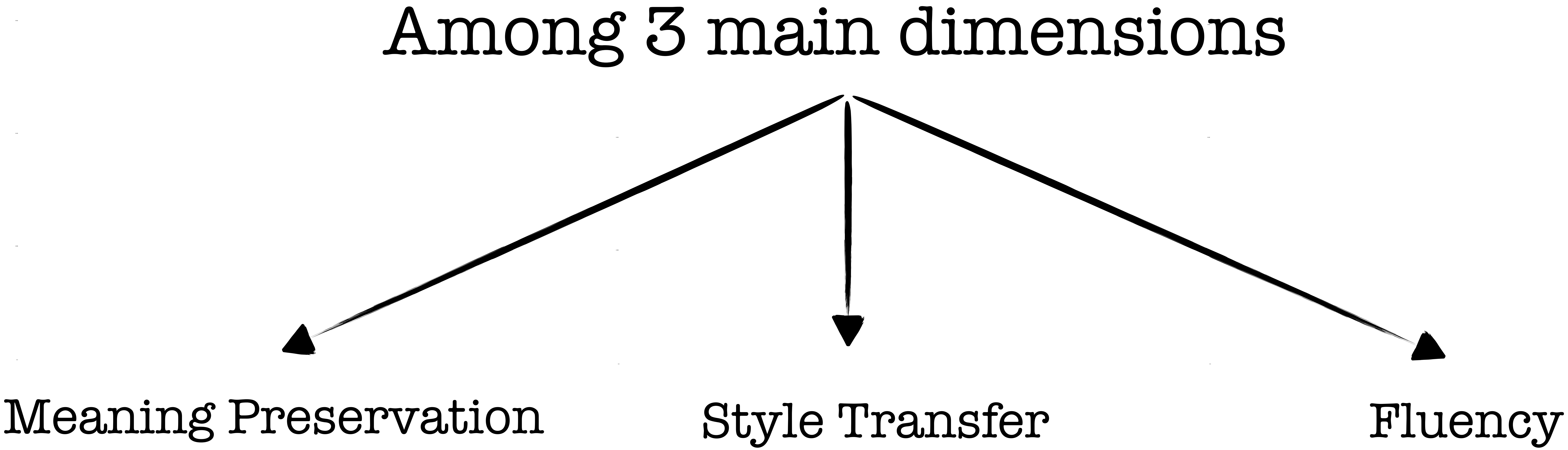
OUTPUT: The screen is too small.

MODERN TO SHAKESPEAREAN

INPUT: Bring her out to me.

OUTPUT: Call her forth to me.

Evaluation of Style Transfer in NLP



Among 3 main dimensions

Meaning Preservation

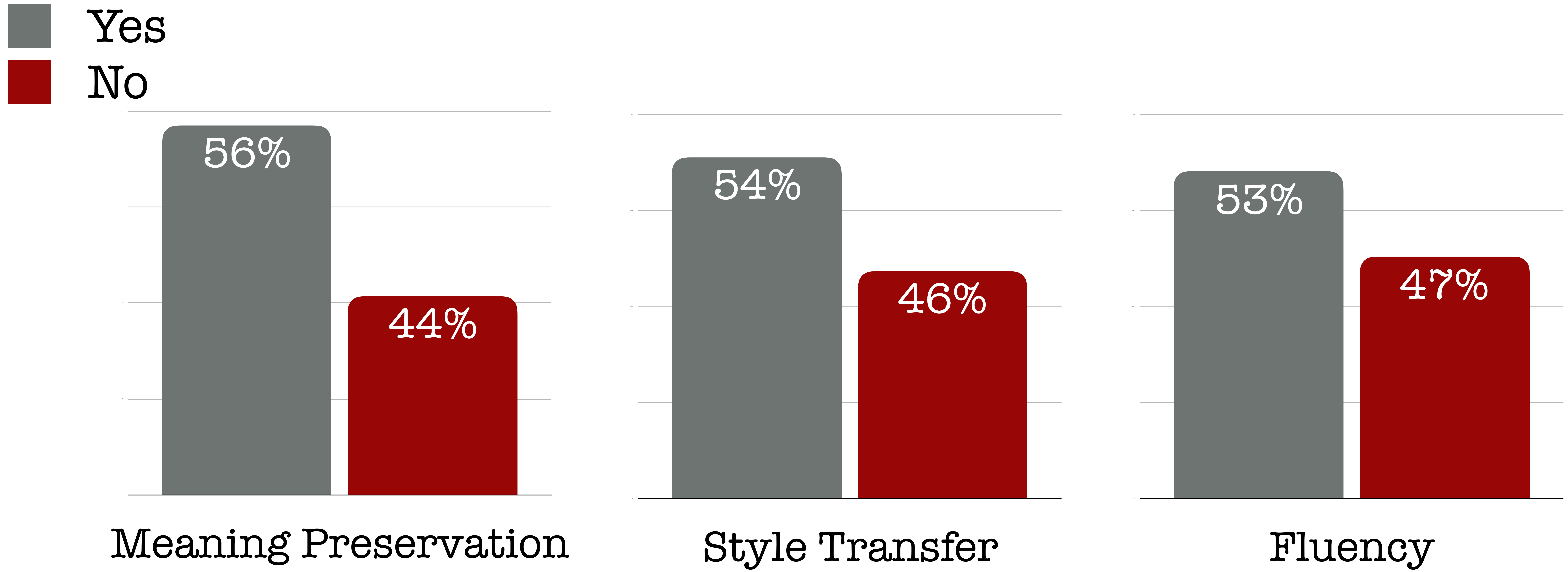
Style Transfer

Fluency

How often do we rely on human evaluations?

69 out of 97 Style Transfer papers resort to human evaluation

How often do we rely on human evaluations?



69 out of 97 Style Transfer papers resort to human evaluation

Our structured review

Structured Review Findings

- **Underspecification:** human annotation design attributes are underspecified in paper descriptions

- **Availability & Reliability:** most papers not release the human ratings and do not give details that can help assess their quality

- **Lack of standardization:** inconsistent annotation protocols across papers

Structured Review Findings

Underspecification: human annotation design attributes are underspecified in paper descriptions

Impact: hampers reproducibility and replicability

Availability & Reliability: most papers not release the human ratings and do not give details that can help assess their quality

Lack of standardization: inconsistent annotation protocols across papers

Structured Review Findings

Underspecification: human annotation design attributes are underspecified in paper descriptions

Impact: hampers reproducibility and replicability

Availability & Reliability: most papers not release the human ratings and do not give details that can help assess their quality

Impact: hurts research on evaluation

Lack of standardization: inconsistent annotation protocols across papers

Structured Review Findings

Underspecification: human annotation design attributes are underspecified in paper descriptions

Impact: hampers reproducibility and replicability

Availability & Reliability: most papers not release the human ratings and do not give details that can help assess their quality

Impact: hurts research on evaluation

Lack of standardization: inconsistent annotation protocols across papers

Impact: hampers comparisons across systems

Paper Selection

✦ **Paper list of Jin et al, 2021**

- contains more than 100 papers and is publicly available
- (https://github.com/ fuzhenxin/Style-Transfer-in-Text)

✦ **Filtering list**

- include papers that employ ST evaluation

✦ **Final list**

- 97 papers
- 86 of the from NLP & AI top-tier venues
- 11 pre-prints

Paper Selection

✦ **Paper list of Jin et al, 2021**

- contains more than 100 papers and is publicly available
- (https://github.com/ fuzhenxin/Style-Transfer-in-Text)

✦ **Filtering list**

- include papers that employ ST evaluation

✦ **Final list**

- 97 papers
- 86 of the from NLP & AI top-tier venues
- 11 pre-prints

Review Structure

QLOBAL CRITERIA

- Task(s)
- Presence of human annotation
- Annotator's details
- Annotator's compensation
- Quality control
- Agreement statistics
- Evaluated systems
- Size of evaluated instance set
- Size of annotations per instance
- Sampling method
- Annotations' availability

DIMENSION-SPECIFIC CRITERIA

FOR EACH DIMENSION

Howcroft et al, 2020

- Presence of human evaluation
- Quality criterion name
- Form of response elicitation
- Details on collected responses
- Size of rating instrument

Review Structure

QLOBAL CRITERIA

- Task(s)
- Presence of human annotation
- Annotator's details
- Annotator's compensation
- Quality control
- Agreement statistics
- Evaluated systems
- Size of evaluated instance set
- Size of annotations per instance
- Sampling method
- Annotations' availability

DIMENSION-SPECIFIC CRITERIA

FOR EACH DIMENSION

Howcroft et al, 2020

- Presence of human evaluation
- Quality criterion name
- Form of response elicitation
- Details on collected responses
- Size of rating instrument

Review Structure

QLOBAL CRITERIA

- Task(s)
- Presence of human annotation
- Annotator’s details
- Annotator’s compensation
- Quality control
- Agreement statistics
- Evaluated systems
- Size of evaluated instance set
- Size of annotations per instance
- Sampling method
- Annotations’ availability

DIMENSION-SPECIFIC CRITERIA

FOR EACH DIMENSION

Howcroft et al, 2020

- Presence of human evaluation
- Quality criterion name
- Form of response elicitation
- Details on collected responses
- Size of rating instrument

Review Structure

QLOBAL CRITERIA

- ✦ Task(s)
- ✦ Presence of human annotation
- ✦ Annotator’s details
- ✦ Annotator’s compensation
- ✦ Quality control
- ✦ Agreement statistics
- ✦ Evaluated systems
- ✦ Size of evaluated instance set
- ✦ Size of annotations per instance
- ✦ Sampling method
- ✦ Annotations’ availability

DIMENSION-SPECIFIC CRITERIA

FOR EACH DIMENSION

Howcroft et al, 2020

- ✦ Presence of human evaluation
- ✦ Quality criterion name
- ✦ Form of response elicitation
- ✦ Details on collected responses
- ✦ Size of rating instrument

Review Structure

QLOBAL CRITERIA

- Task(s)
- Presence of human annotation
- Annotator's details
- Annotator's compensation
- Quality control
- Agreement statistics
- Evaluated systems
- Size of evaluated instance set
- Size of annotations per instance
- Sampling method
- Annotations' availability

DIMENSION-SPECIFIC CRITERIA

FOR EACH DIMENSION

Howcroft et al, 2020

- Presence of human evaluation
- Quality criterion name
- Form of response elicitation
- Details on collected responses
- Size of rating instrument

Review Structure

QLOBAL CRITERIA

- ✦ Task(s)
- ✦ Presence of human annotation
- ✦ Annotator’s details
- ✦ Annotator’s compensation
- ✦ Quality control
- ✦ Agreement statistics
- ✦ Evaluated systems
- ✦ Size of evaluated instance set
- ✦ Size of annotations per instance
- ✦ Sampling method
- ✦ Annotations’ availability

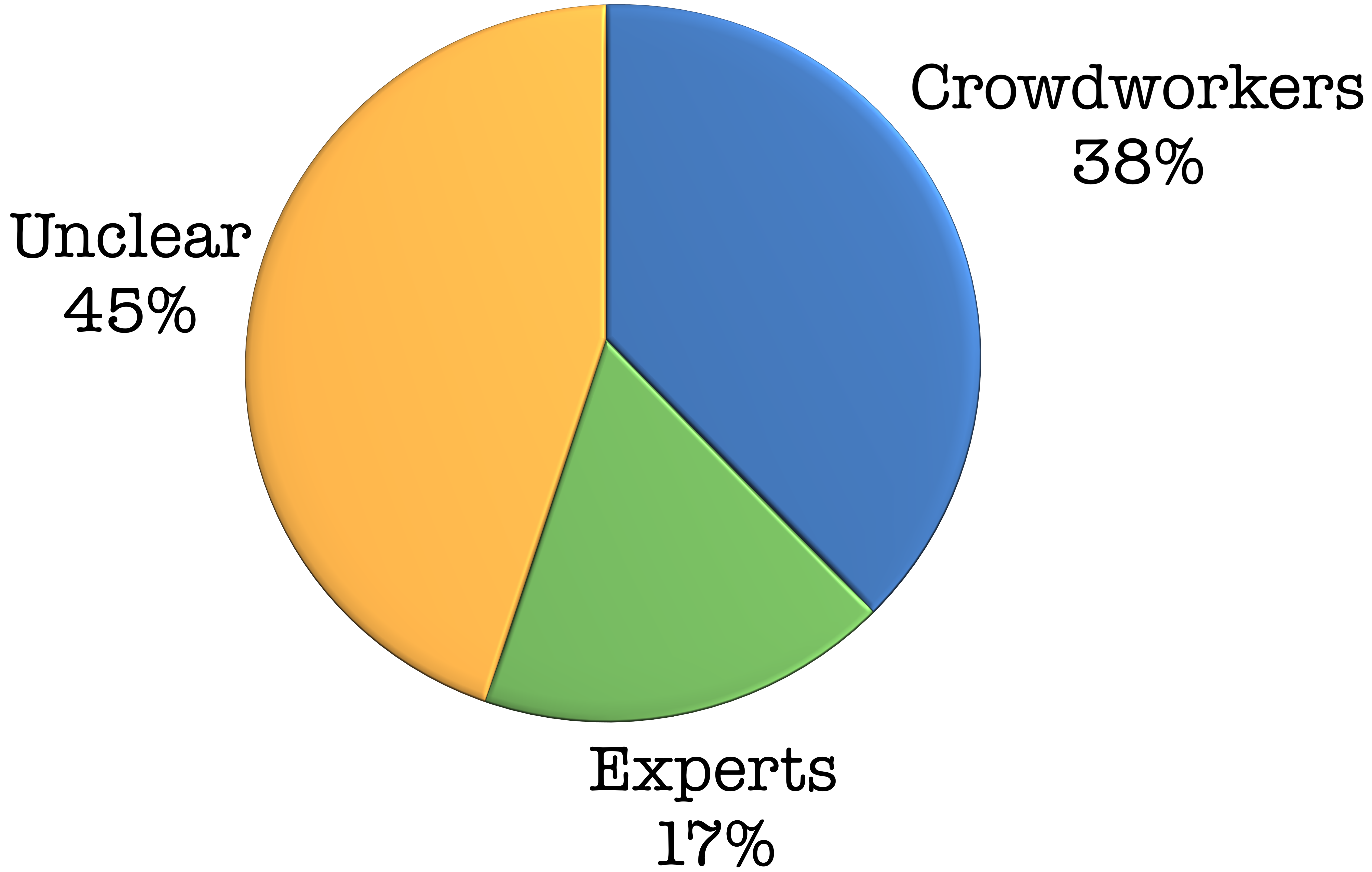
DIMENSION-SPECIFIC CRITERIA

FOR EACH DIMENSION

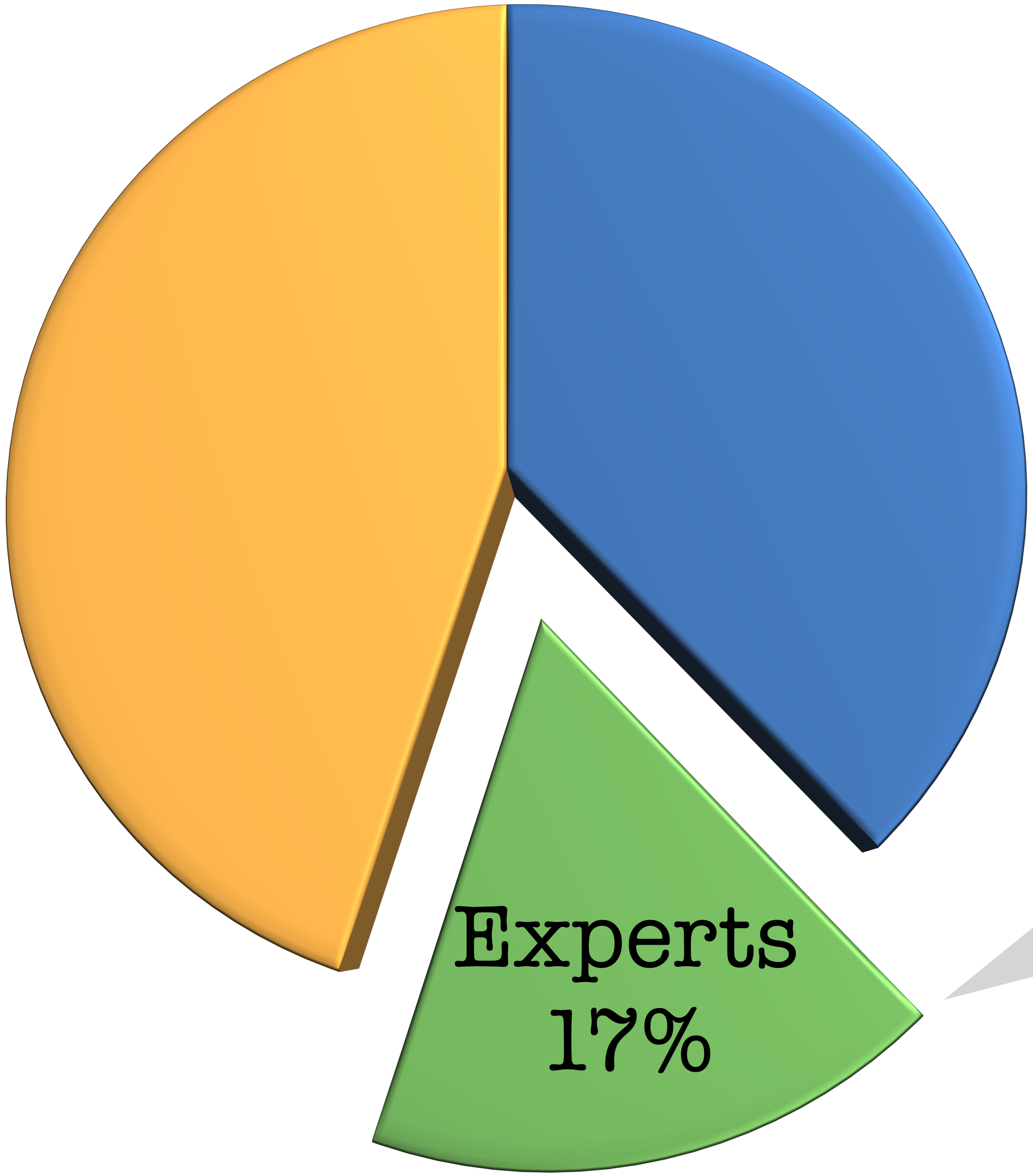
Howcroft et al, 2020

- ✦ Presence of human evaluation
- ✦ Quality criterion name
- ✦ Form of response elicitation
- ✦ Details on collected responses
- ✦ Size of rating instrument

Who are the annotators?



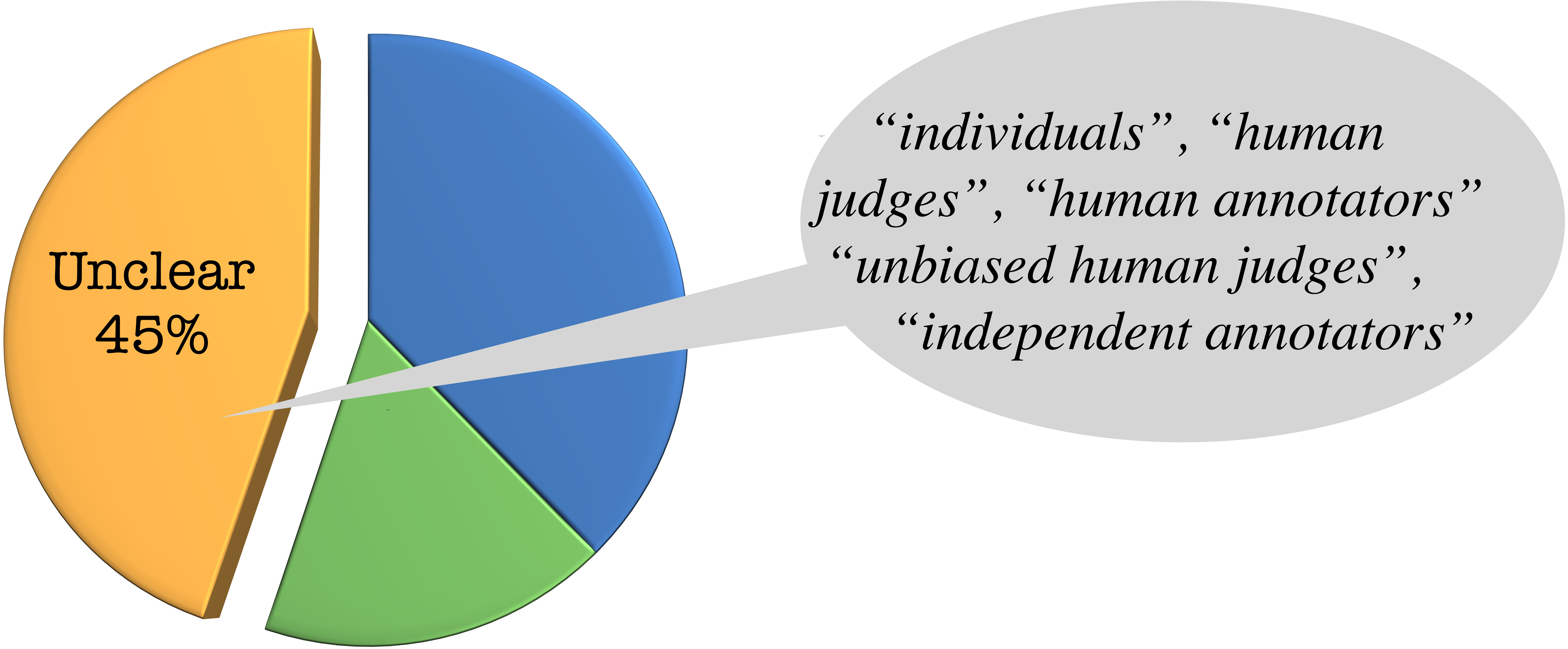
Who are the annotators?



“bachelor degree”

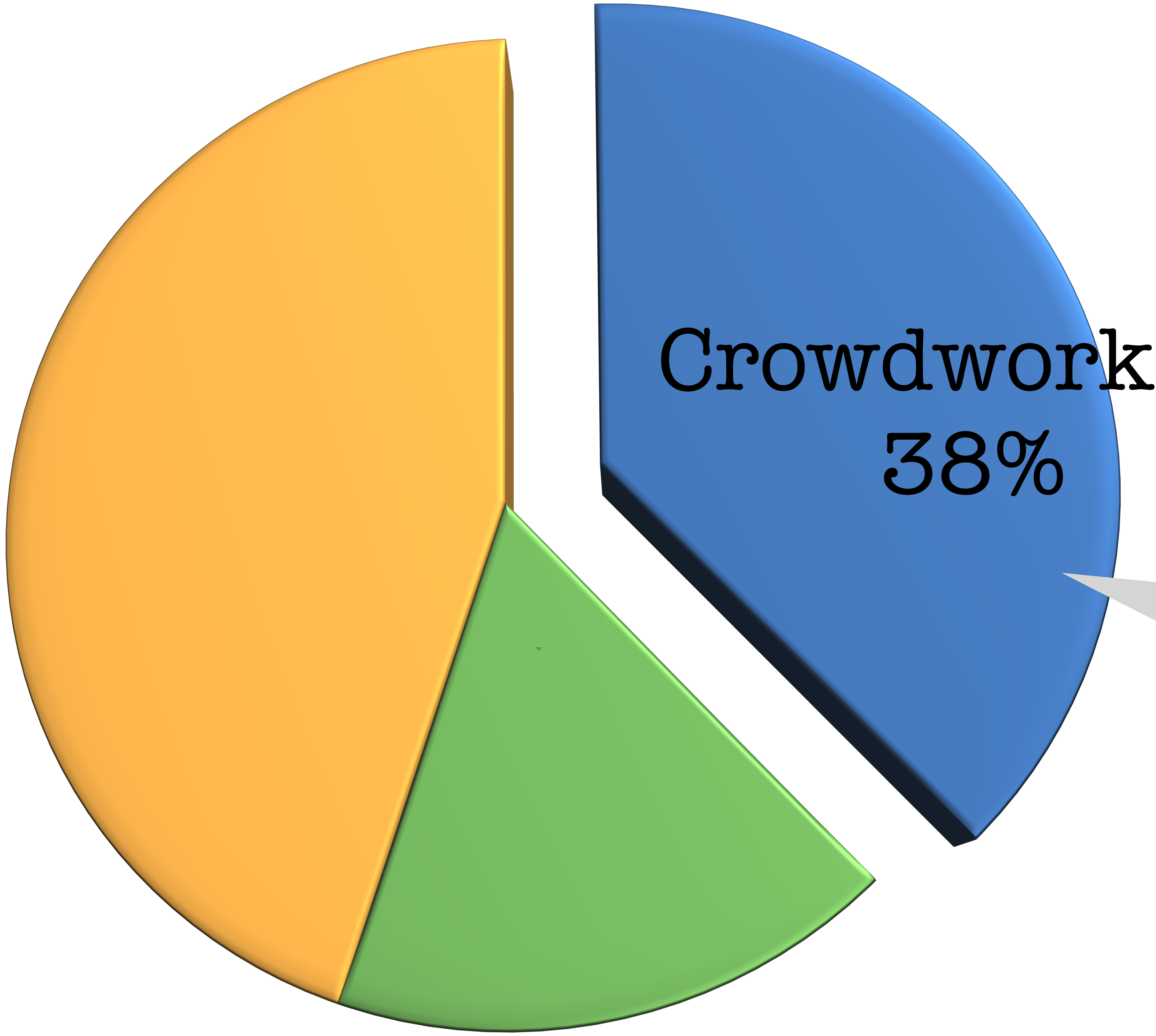
“independent of the authors’ research group”, “annotators with linguistic background” “well-educated volunteers”, “graduate students in computational linguistics” “major in linguistics” “linguistic background”, “authors”

Who are the annotators?



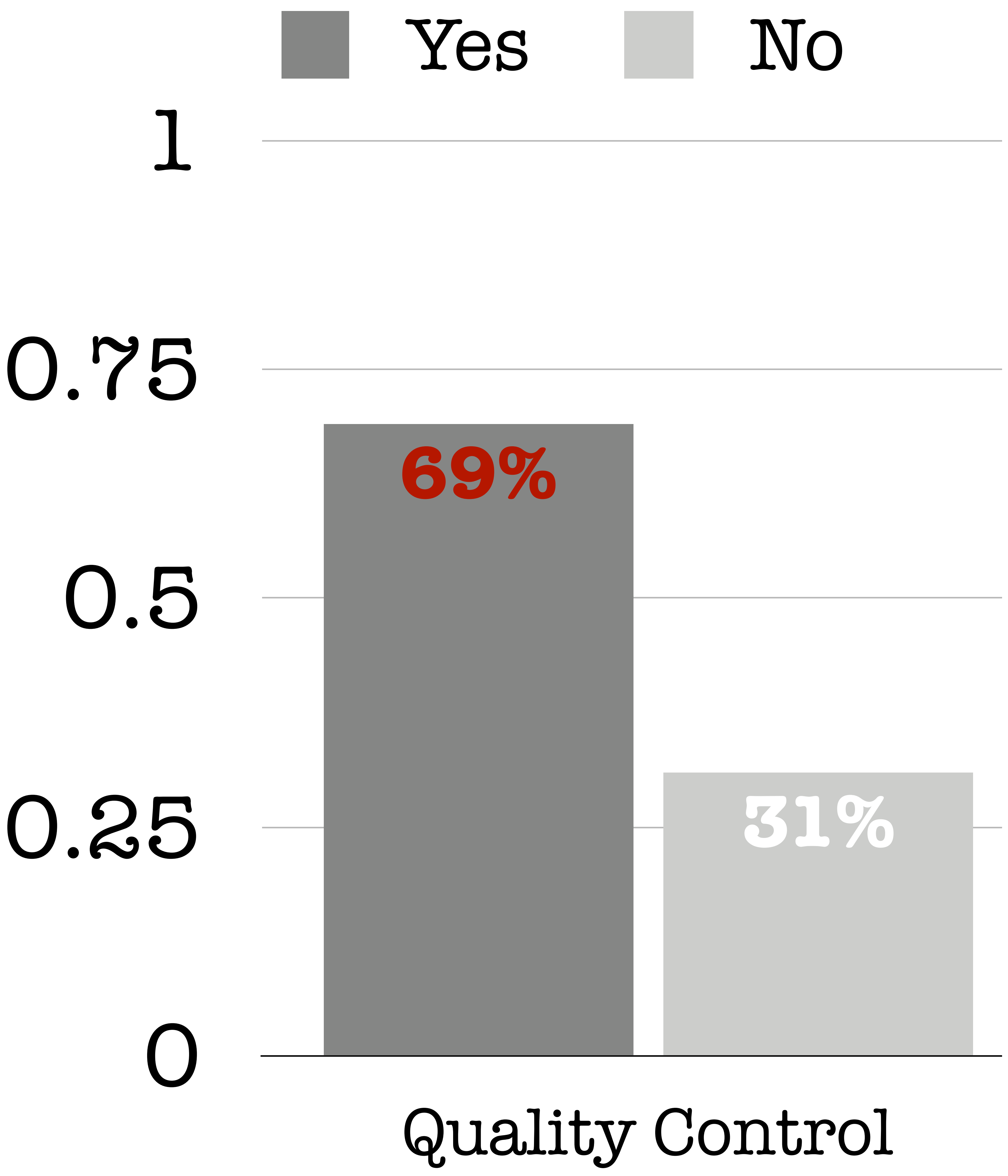
“individuals”, “human judges”, “human annotators” “unbiased human judges”, “independent annotators”

Who are the annotators?



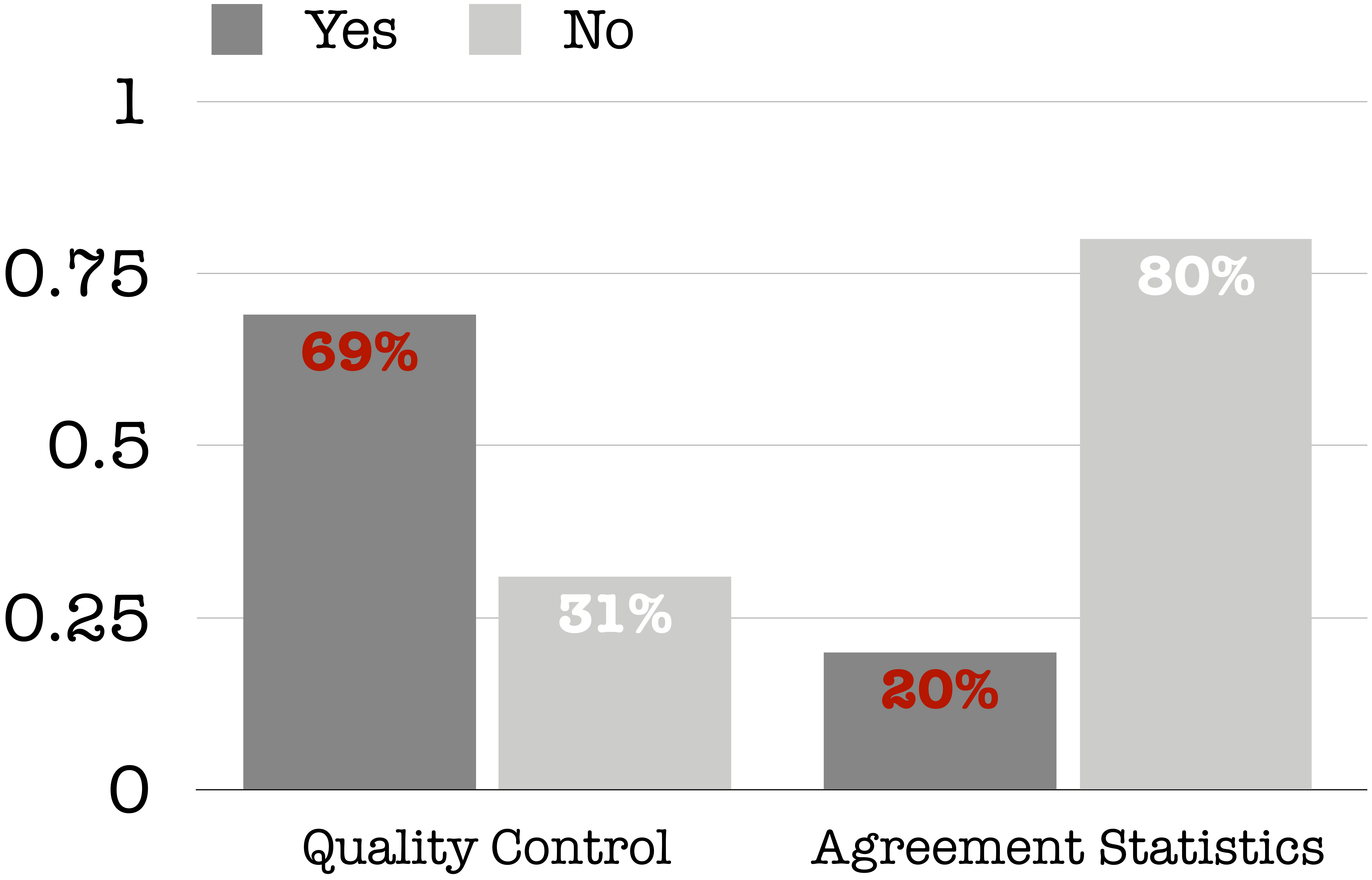
"Amazon Mechanical Turk workers", "crowd-workers"

How reliable are annotators?



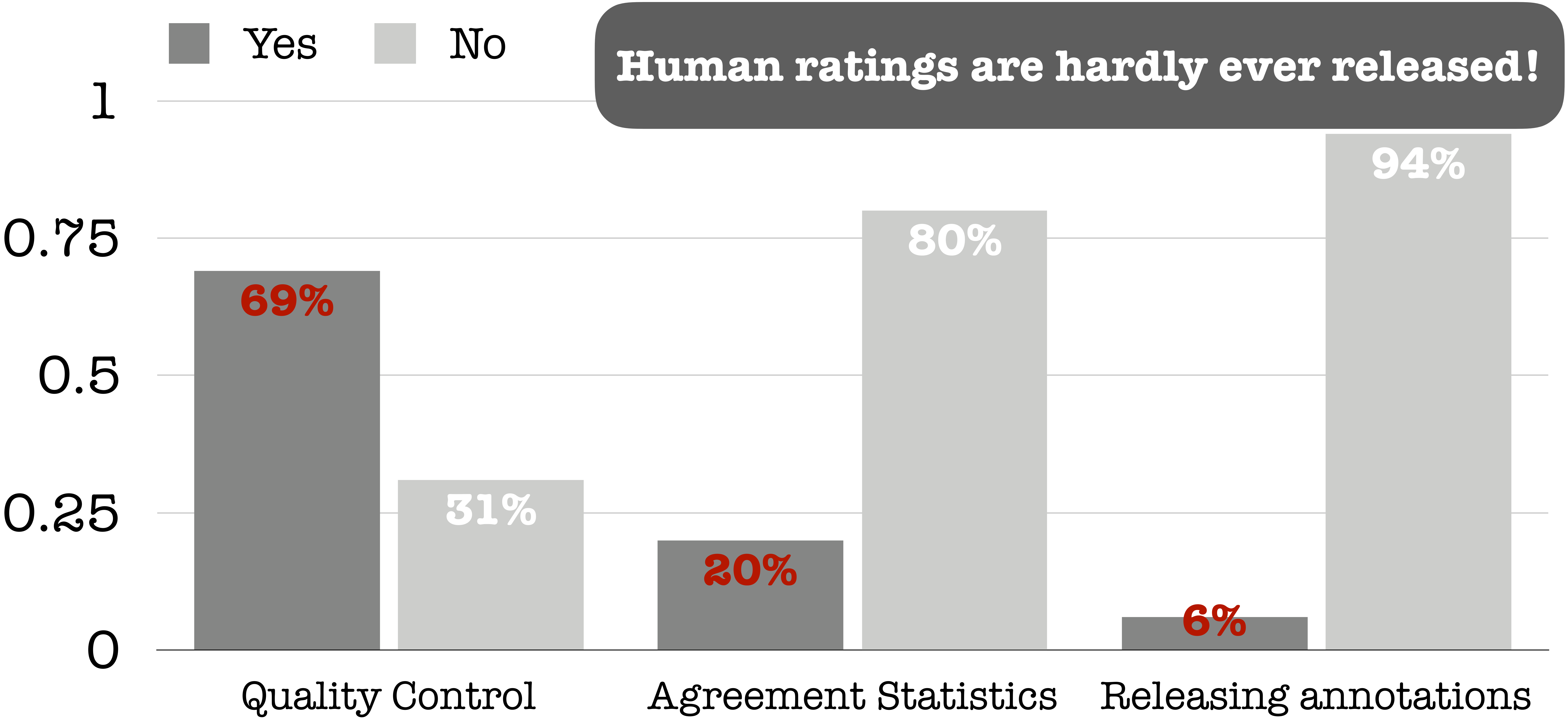
The majority of crowd-sourced evaluations do not employ quality control!

How reliable are annotators?



The vast majority of evaluations do not report agreement statistics

How reliable are annotators?



Human ratings are hardly ever released!

Quality criterion names

🟦 **Style Transfer**

🟨 **Meaning Preservation**

🟩 **Fluency**

Quality criterion names

🟦 **Style Transfer** 🟨 **Meaning Preservation** 🟩 **Fluency**

attribute compatibility, *formality*, *politeness level*, *sentiment*, *style transfer intensity*, *attractive captions*, *attribute change correctness*, *bias*, *creativity*, *highest agency*, *opposite sentiment*, *sentiment*, *sentiment strength*, ***similarity to the target attribute***, ***style correctness***, ***style transfer accuracy***, ***style transfer strength***, ***stylistic similarity***, *target attribute match*, *transformed sentiment degree*

Quality criterion names

🟦 Style Transfer

🟨 **Meaning Preservation**

🟩 Fluency

content preservation, *meaning preservation*, ***semantic intent***, ***semantic similarity***, *closer in meaning to the original sentence*, *content preservation degree*, *content retainment*, *content similarity*, ***relevance***, ***semantic adequacy***

Quality criterion names

🟦 **Style Transfer**

🟨 **Meaning Preservation**

🟩 **Fluency**

fluency, grammaticality, naturalness*, gibberish language, language quality*

Quality criterion names

🟦 **Style Transfer**

🟨 **Meaning Preservation**

🟩 **Fluency**

Inconsistent terminology across papers...

- Leads to different **interpretations** by annotators
- Makes it harder to understand exactly **what is being evaluated**
- **Hampers comparison** of evaluation methods **across papers**

How is each attribute evaluated?

Direct rating type: each system output is assessed in isolation for that dimension

Example

Rate the sentence on a 7- point discrete scale of –3 (Very informal) to 3 (Very formal).

How is each attribute evaluated?

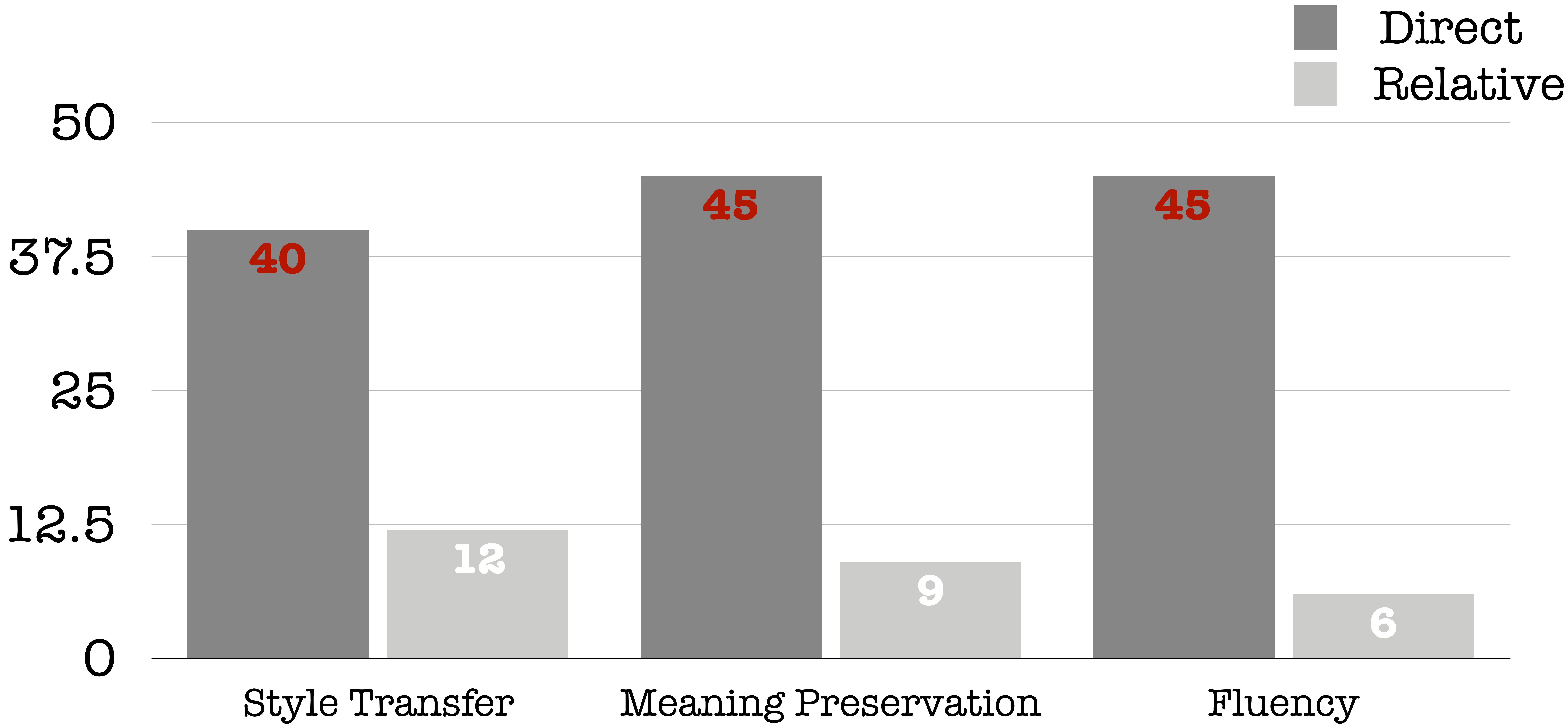
Direct rating type: each system output is assessed in isolation for that dimension

Relative rating type: two or more system outputs are compared against each other

Example

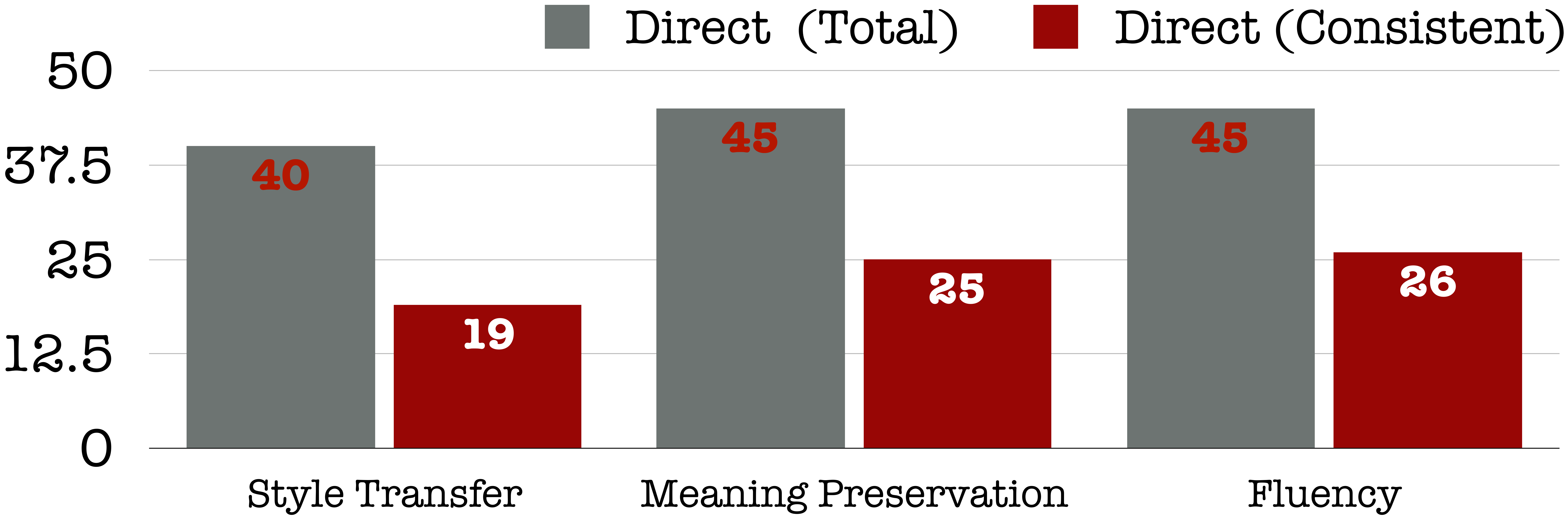
Rank the sentences in order of their formality (from most informal to most formal).

Direct evaluation is the most frequent rating type across attributes



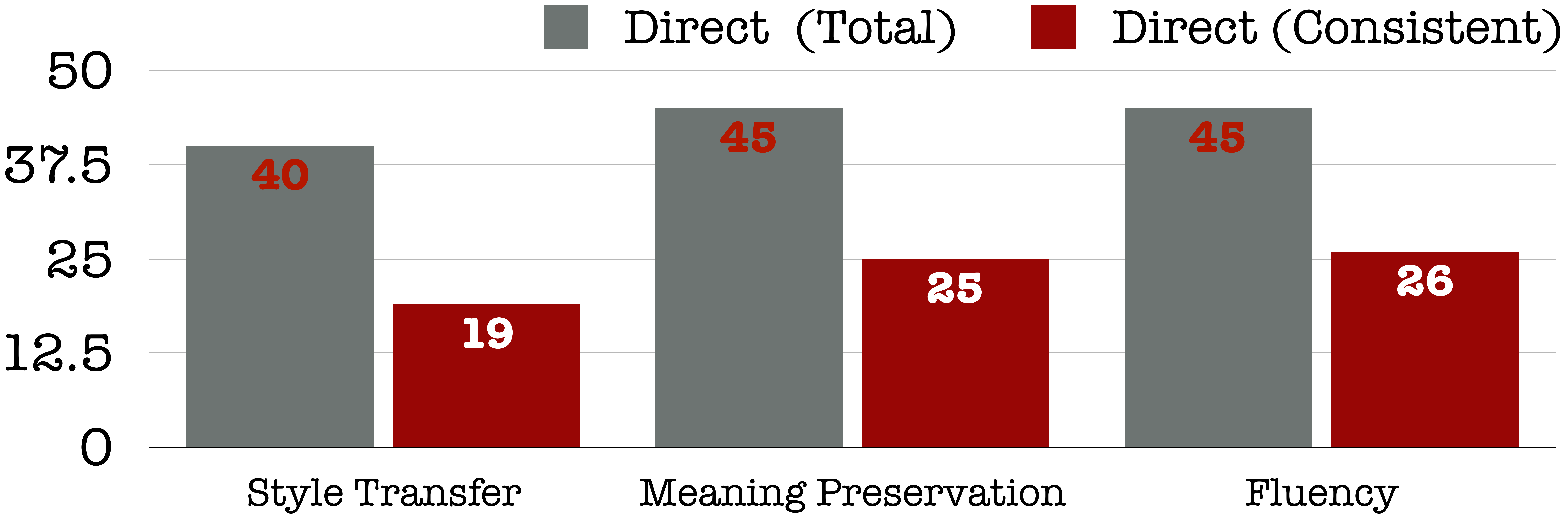
Are direct evaluations consistent across papers?

5-point scale is the most consistent framework
Yet discrepancies are found at least
among 40% of reviewed papers



Are direct evaluations consistent across papers?

5-point scale is the most consistent framework
Yet discrepancies are found at least among 40% of reviewed papers



Structured Review Findings

Underspecification: human annotation design attributes are underspecified in paper descriptions

Impact: hampers reproducibility and replicability

Availability & Reliability: most papers not release the human ratings and do not give details that can help assess their quality

Impact: hurts research on evaluation

Lack of standardization: inconsistent annotation protocols across papers

Impact: hampers comparisons across systems

Our Recommendations

☑ Describe evaluation protocols

☑ Release annotations

☑ Standardize evaluation protocols

Our Recommendations

☑ Describe evaluation protocols

☑ Release annotations

☑ Standardize evaluation protocols

- details on the procedures followed for recruiting annotators
- annotator's compensation
- inter-annotator agreement statistics
- number of annotations per instance
- number of systems evaluated
- number of instances annotated
- selection method of the annotated instances
- detailed description of evaluated frameworks

Our Recommendations

☑ Describe evaluation protocols

☑ Release annotations

☑ Standardize evaluation protocols

- Reliability & Reproducibility
- Enable development of better automatic metrics
- Shed light into difficulty of the task

Our Recommendations

☑ Describe evaluation protocols

☑ Release annotations

☑ Standardize evaluation protocols

- Fair comparisons across papers
- Clear documentation (following datasheets)

Our Recommendations

- ☑ Describe evaluation protocols
- ☑ Release annotations
- ☑ Standardize evaluation protocols

https://github.com/Elbria/ST-human-review