

Crowdsourcing and text evaluation

TOOLS, PRACTICES, AND NEW RESEARCH DIRECTIONS

Dave Howcroft (@_dmh), IR&Text @ Glasgow, 20 January 2020

Crowdsourcing

Recruiting experimental **subjects** or data **annotators** through the web, especially **using services like** Prolific Academic, FigureEight, or **Mechanical Turk** (but also social media).

Tasks Tools Platforms Practices

Tasks

- ▶ Judgements
 - ▶ Grammaticality
 - ▶ Fluency / naturalness
 - ▶ Truth values / accuracy
- ▶ Experiments
 - ▶ Pragmatic manipulations
 - ▶ Self-paced reading
- ▶ Data Collection
 - ▶ Label parts of text for meaning
 - ▶ Clever discourse annotations
 - ▶ Classifying texts (e.g. sentiment)
 - ▶ Corpus elicitation
 - ▶ WoZ Dialogues
 - ▶ Real-time collaborative games
- ▶ Evaluation
 - ▶ Combining all of the above...

Linguistic judgements

- Recruit subjects on AMT, Prolific
- Judge naturalness only (above) or naturalness and accuracy (below)

(Howcroft et al. 2013; my thesis)

Instructions

Please read the following dialogue, focusing on the system response:

User: Recommend a cheap restaurant.

System: Amy's Bread with cafe restaurant in Midtown. has best the overall quality selected among, with food food quality and decent service. Cost cost dollars.

The text in this example has many more problems. The first 'sentence' probably means that Amy's Bread is a cafe in Midtown, but it's not grammatically correct and we have to work to interpret it. Similarly, the rest of the text has a variety of errors in word order, repeated, and missing words. This text is pretty bad (almost word salad!) but still partially understandable.

Is the English in the system response correct?

This makes no sense! Its "word salad"

Major problems; very difficult to understand

Minor problems or awkward phrasing; mostly understandable.

Perfectly correct English

Which (if any) of the following facts are missing from the system response?
Please sort the remaining facts so they match their order in the text.

Missing

‡ (Amy's Bread) quality	best	X
‡ (Amy's Bread) price	12	X
‡ (Amy's Bread) neighborhood	Midtown West	X
‡ (Amy's Bread) cuisine	Cafes	X
‡ (Amy's Bread) food quality	excellent	X
‡ (Amy's Bread) service	decent	X

Here is the dialogue again, repeat the system response.

User: Recommend a cheap restaurant.

System: Amy's Bread with cafe restaurant in Midtown. has best the overall quality selected among, with food food quality and decent service. Cost cost dollars.

Does Da Andrea have good food quality?

☐ Yes
 ☐ No

Very Unnatural

Unnatural

Somewhat Unnatural

Neither Natural nor Unnatural

Somewhat Natural

Natural

Very Natural

Does the system response include any extra details?

☒ Yes
 ☐ No

User: Tell me about these West Village restaurants.

System: Da Andrea's price is 28 dollars, and Gene's price is 33 dollars. Da Andrea has very good food quality. Gene's has just good food quality .

Does Da Andrea have good food quality?

☐ Yes
 ☐ No

Very Unnatural

Unnatural

Somewhat Unnatural

Neither Natural nor Unnatural

Somewhat Natural

Natural

Very Natural

Meaning annotation

- Student project @ Uni Saarland
- Write sentences and annotate
- Based on "semantic stack" meaning representation used by Mairesse et al. (2010)

Markiere (nur!) die Tokens, die die Information enthalten!

`reject(time(21:45))`

Der Film kommt nicht um **21:45**

Annotiere die verbleibenden Wörter!

Der Film kommt nicht **um** 21:45

inform reject (blank) name time date (blank) alternative

Annotiere die verbleibenden Wörter!

Der Film kommt nicht um 21:45

inform reject (blank) name time (blank) alternative

W

Bist du dir der Annotation sicher?

Der Film kommt nicht	um	21:45
		21:45
	time	time
reject	reject	reject

Abschicken

Clever annotations

- Subjects recruited on Prolific Academic
- Read sentences in context
- Select the best discourse connective

(Scholman & Demberg 2017)

Explanations

The parts in grey provide the background for the sentences in black, which have a logical connection between them. Your task will be to "drag and drop" a connecting phrase from the list of candidate phrases to the green box in the text. Please choose the linking phrase that best reflects the meaning of the connection between the black sentences.

Please drag the best-suited connective into the green target box below.

Candidate connectives

because as a result more specifically in addition even though nevertheless by contrast none of these

He's attacked the concept of "building tenure," one of the most disgraceful institutions in American public schools. It means it is virtually impossible to fire or even transfer incompetent principals. **Once they are in the building, they stay //**

as an illustration one South Bronx principal kept his job for 16 years, despite a serious drinking problem and rarely showing up for work. He was finally given leave when he was arrested for allegedly buying crack.

Submit Add another connective

Eliciting corpora

Image-based

- Recruit from AMT
- Write text based on images

(Novikova et al. 2016)

Paraphrasing

- Recruit from Prolific Academic
- Paraphrase an existing text

(Howcroft et al. 2017)



1 / 4

Bond Street's price is 51 dollars, and it has good service, with excellent food quality. This Japanese , Sushi restaurant has very good decor. It has the best overall quality among the selected restaurants.

Please re-write the original text in your own words. Make sure you include all the same information as the original.

Next

Pragmatic manipulations

- Recruit subjects on AMT
- Subjects read a reported utterance in context
- Subjects rate the plausibility or likelihood of different claims


Greg frequently travels by air, to see family and attend conferences.

Last week he flew to a conference, and met up with Helen, an old colleague he occasionally traveled with. They went to breakfast together, and started talking about their travel.

Greg said to Helen: "I flew here. I got into business class!"


How often do you think Greg usually gets into business class, when flying on a plane?

Never Sometimes Always




How often do you think Greg usually carries his cell phone on board with him, when flying on a plane?

Never Sometimes Always




How often do you think Greg usually travels by airplane?

Never Sometimes Always



How often do you think Greg and Helen usually meet up?

Never Sometimes Always



Next

Dialogue

- Human-Human Interactions
- WoZ interactions
- Human-System Interactions
- Used both for elicitation and evaluation

Pictured: [ParlAI](#), [slurk](#), [visdial-amt-chat](#)

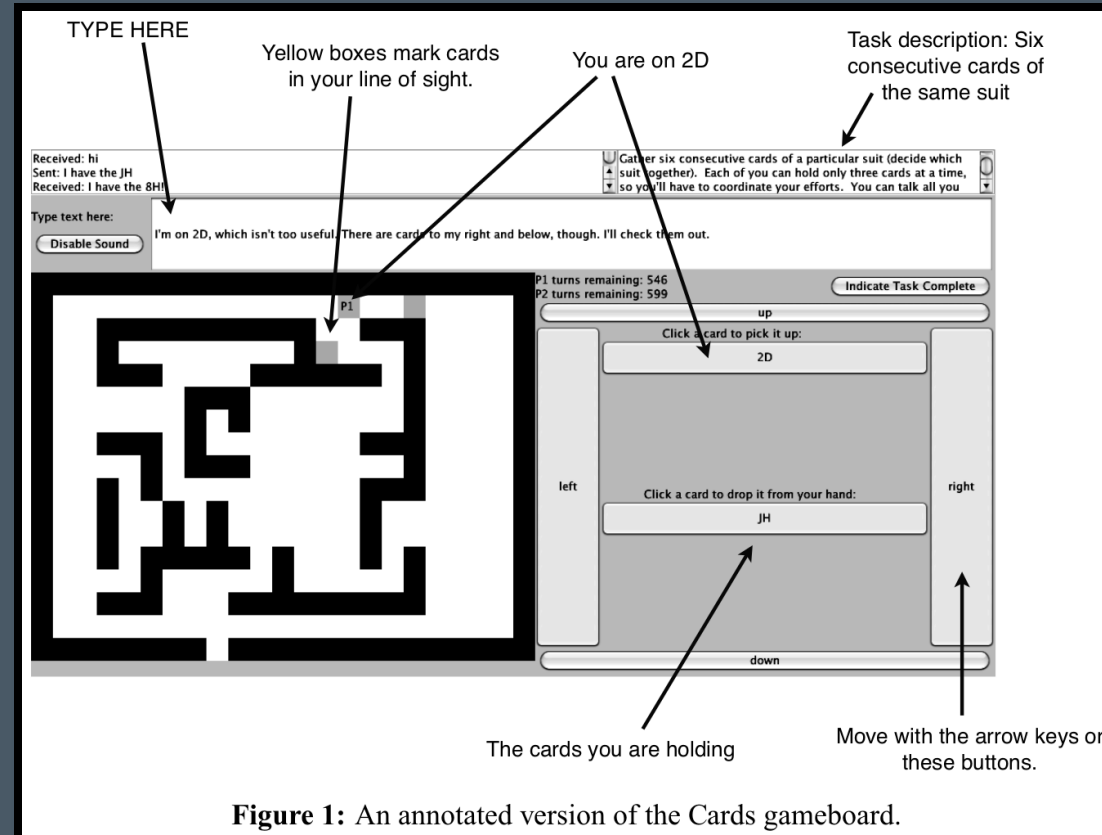
The collage consists of three screenshots illustrating dialogue systems:

- Top-left screenshot:** A 'Live Chat' interface for a task. The task description states: 'In this task, you will need to ask a question about a paragraph, and then provide your own answer to it. If you are ready, please click "Accept HIT" to start this task.' The paragraph text is: 'QA Collector: In the United States, heating, ventilation and air conditioning (HVAC) systems account for 30% (4.65 EJ/yr) of the energy used in commercial buildings and nearly 50% (10.1 EJ/yr) of the energy used in residential buildings.' The interface includes a 'Please provide a question given the paragraph' prompt, a 'You: How much of the energy do HVAC systems account for' prompt, and a 'QA Collector: Thanks. And what is your question?' response. There is a 'Please enter here...' input field and buttons for 'Submit HIT' and 'Return HIT'.
- Top-right screenshot:** A 'Slurk - Chatroom' window titled 'Test Room'. It shows a chat history with messages from 'minimal bot' and 'You'. The messages are: 'minimal bot has joined the room. Say "Hello!" :)', 'minimal bot: Hello minimal bot!', 'minimal bot: B has joined the room. Say "Hello!" :)', 'minimal bot: Hello B!', 'You: Hello everyone!', 'You: new_image_public', and 'You: This is a nice picture. Very nicely composed.' There is a large image of a landscape with a field and a tree. The window also shows 'Latency: 5 ms' and 'Users: You, A, minimal bot, B'.
- Bottom screenshot:** A 'Caption: The man is riding his bicycle on the sidewalk' task. It shows a chat history with messages from '3. Fellow Turkur:' and '4. You:'. The messages are: '3. Fellow Turkur: It has black wheels and handlebars. I cannot see the body of the bike that well.', '4. You: Is anyone else riding a bike?', '4. Fellow Turkur: No he is the only one.' There is a 'Type Message Here:' input field and a 'Send' button.

Real-time collaborative games

- Recruit subjects on AMT
- Together they have to collect playing cards hidden in a 'maze'
- Each can hold limited quantity
- Communicate to achieve goal

<http://cardscorpus.christopherpotts.net/>

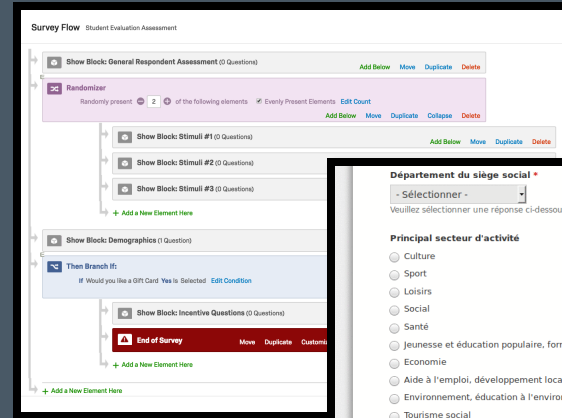


Evaluation

- ▶ Combines judgements, experiments, and data collection

Tools

- ▶ Built-in resources
- ▶ Qualtrics, SurveyMonkey, etc
- ▶ Google, MS, Frama forms
- ▶ LingoTurk
- ▶ REDCap
- ▶ ParlAI, slurk, visdial-amt-chat
- ▶ Your own server...




Quickly find research participants you can trust.

Launch your study to tens of thousands of trusted participants in minutes. Recruit niche or representative samples on-demand. Prolific builds the most powerful and flexible tools for online research. Sign up for free.



Research

Collect high quality responses from people around the world within minutes. [Learn more](#)

[SIGN UP TO RESEARCH](#)

Participate

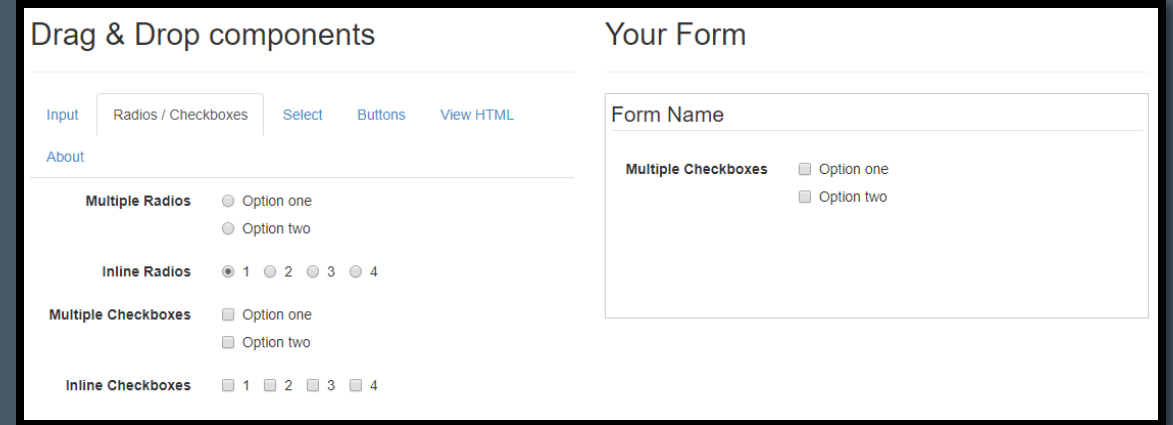
Take part in engaging research, earn cash, and help improve human knowledge. [Learn more](#)

[SIGN UP TO PARTICIPATE](#)

Built-in tools

Mechanical Turk and FigureEight both provide tools for **basic survey design**

- ▶ Designed for HITs
- ▶ Often quite challenging to use

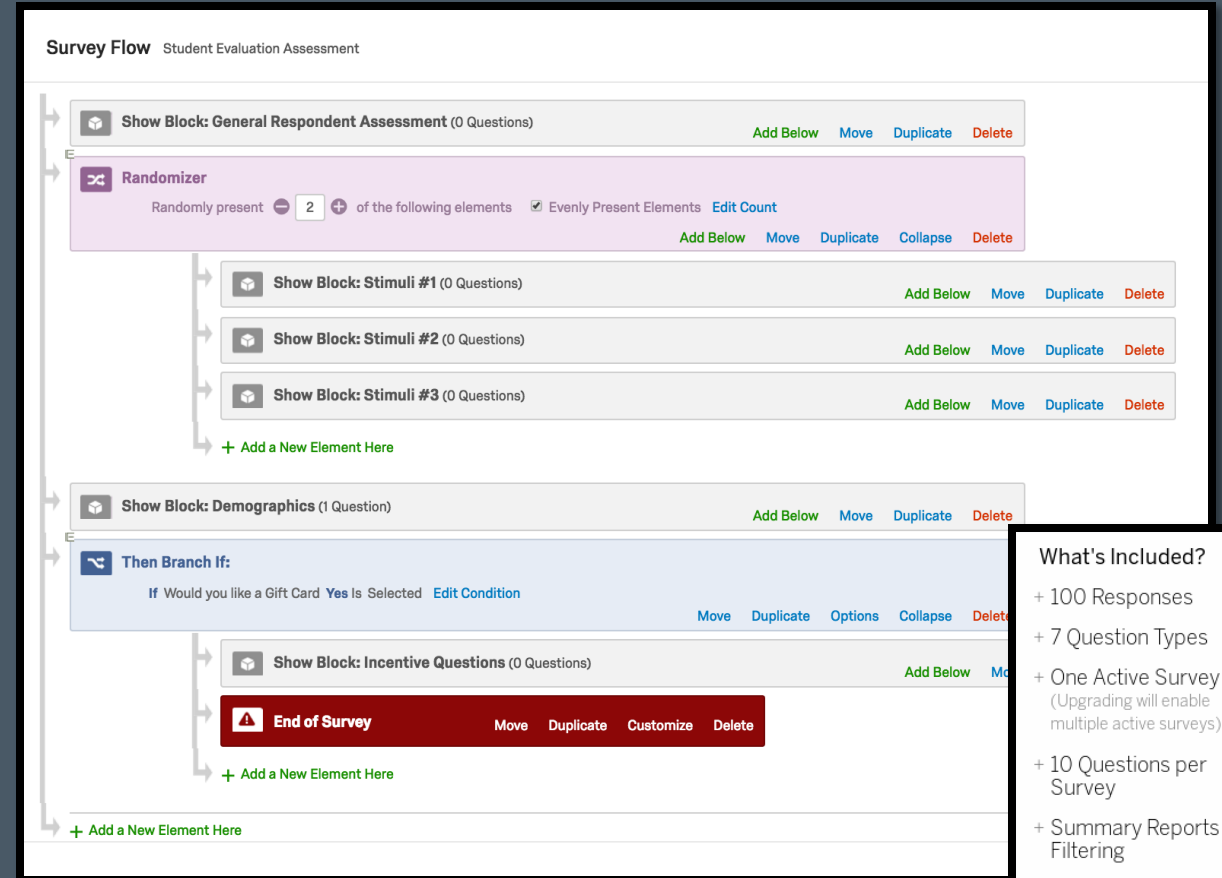


The screenshot shows the 'Drag & Drop components' interface for creating a survey form. On the left, under the 'Radios / Checkboxes' tab, there are four component types: 'Multiple Radios' (with two radio options), 'Inline Radios' (with four radio options), 'Multiple Checkboxes' (with two checkbox options), and 'Inline Checkboxes' (with four checkbox options). On the right, under 'Your Form', there is a 'Form Name' field and a preview of the selected 'Multiple Checkboxes' component, showing two options: 'Option one' and 'Option two'.

<https://blog.mturk.com/tutorial-editing-your-task-layout-5cd88ccae283>

Qualtrics

- ▶ A leader in online surveys
- ▶ Enterprise survey software available to students and researchers
- ▶ Sophisticated designs possible
- ▶ Cost: thousands / yr (@ lab/institution level)
 - ▶ Unless free is good enough



What's Included?

- + 100 Responses
- + 7 Question Types
- + One Active Survey (Upgrading will enable multiple active surveys)
- + 10 Questions per Survey
- + Summary Reports & Filtering
- + Survey Logic
- + Online Reporting (Upgrading will enable CSV/SPSS export)
- + Doesn't Expire

SurveyMonkey

- ▶ A leader in online surveys
- ▶ Sophisticated designs possible
- ▶ Responsive designs
- ▶ Cost: monthly subs available
 - ▶ Discounted for researchers
 - ▶ Unless free is good enough

Personal Plans		Business Plans		
	PREMIER	BEST VALUE ADVANTAGE	STANDARD	BASIC
	\$70 / month \$99 / month Billed \$840 annually	\$23 / month \$32 / month Billed \$276 annually	\$26 / month \$99 / month Billed monthly SAVE with Annual	\$0 Always free
	SIGN UP	SIGN UP	SIGN UP	SIGN UP
Collapse all ▼				
Survey Capabilities				
Number of surveys	UNLIMITED	UNLIMITED	UNLIMITED	UNLIMITED
Questions per survey	UNLIMITED	UNLIMITED	UNLIMITED	10 questions per survey
Number of responses	UNLIMITED	UNLIMITED	1,000 responses per month*	View 100 responses per survey
Get responses via web, social or email	•	•	•	•
Number of collectors	UNLIMITED	UNLIMITED	UNLIMITED	Use 3 collectors per survey
Pop up online surveys	•	•	•	•
Track email responses	•	•	•	•
Mobile apps for iOS and Android	•	•	•	•
Recurring surveys				

Support any project, team, or organization

Collaboration

Share surveys and data [across teams](#) without having to share passwords

Shared asset library

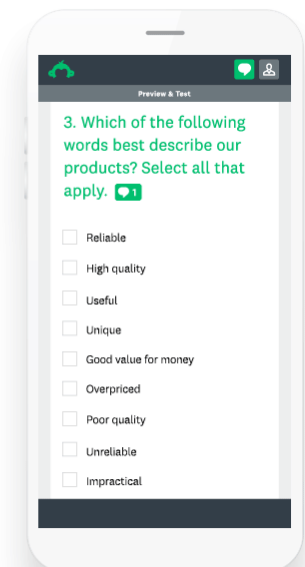
Share resources across your organization (guidelines, logos, templates etc.)

Sophisticated features

Survey logic including advanced branching, conditional questions and page [skip logic](#), AB tests, advanced piping, and more.

Integrations

Enrich your data by [connecting your surveys](#) to existing applications you already use



FramaForms

- ▶ Open alternative to Forms in GDocs, Office365, etc
- ▶ Based in France, part of a larger free culture and OSS initiative

<https://framaforms.org/>

The screenshot displays the FramaForms web interface. The main form area is titled "Département du siège social" with a red asterisk indicating a required field. Below the title is a dropdown menu labeled "- Sélectionner -" with a small downward arrow. A text prompt "Veuillez sélectionner une réponse ci-dessous" is positioned below the dropdown. Underneath, the section "Principal secteur d'activité" contains a list of radio button options: Culture, Sport, Loisirs, Social, Santé, Jeunesse et éducation populaire, formation, Economie, Aide à l'emploi, développement local, Environnement, éducation à l'environnement et au développement durable, Tourisme social, and Solidarité internationale. On the right side, a sidebar titled "Ajouter un champ" (Add a field) provides a grid of icons for various form elements: Champ texte (text field), Zone de texte (text area), Courriel (email), Nombre (number), Boutons radios (radio buttons), Cases à cocher (checkboxes), Liste de sélection (list selection), Grille (grid), Date, Heure (time), Fichier (file upload), Caché (hidden), Balisage (markup), Groupe de champs (group of fields), and Saut de page (page break).

FramaForms

- ▶ Open alternative to Forms in GDocs, Office365, etc
- ▶ Based in France, part of a larger free culture and OSS initiative

<https://framaforms.org/>



The screenshot displays the FramaForms configuration interface. On the left, there is a form titled "Département du siège social" with a dropdown menu set to "- Sélectionner -" and a prompt "Veuillez sélectionner une réponse ci-dessous". Below this is a section "Principal secteur d'activité" with a list of radio button options: Culture, Sport, Loisirs, Social, Santé, Jeunesse et éducation populaire, formation, Economie, Aide à l'emploi, développement local, Environnement, éducation à l'environnement et au développement durable, Tourisme social, and Solidarité internationale. On the right, a panel titled "Ajouter un champ" (Add a field) contains a grid of icons for various form elements: Champ texte, Zone de texte, Courriel, Nombre, Boutons radios, Cases à cocher, Liste de sélection, Grille, Date, Heure, Fichier, Caché, Balisage, Groupe de champs, and Saut de page.

LingoTurk

- ▶ Open source server for managing online experiments
- ▶ Used for a variety of tasks already
 - ▶ Corpus elicitation
 - ▶ Annotation
 - ▶ Experimental pragmatics
 - ▶ NLG system evaluation

(demo Uni Saarland server)

Public
Repo: <https://github.com/FlorianPusse/Lingoturk>

Set publishing options account balance: \$0.0

Reward per Question: (in \$)
0.35

Time Activated (days):
7

Maximum Assignments:
3

Keywords (komma seperated):
script,scripts,activity,activities,linking,aligning, English,events,description,descriptions,m

Blocked Workers (.csv):

Choose destination:

Previews:

Instructions to workers

Please go through all the 4 examples below to understand the

You will be presented with two descriptions of an everyday activity e.g. eating in a fast food restaurant. Each description consists of a sequence of sentences depicting events which are part of the activity. Two sentences will be highlighted, one in the first description and another in the second description. Please indicate whether or not the two sentences are similar and represent the same event. For example "leave restaurant" in the first description and "drive out of restaurant" in the second description represent the same

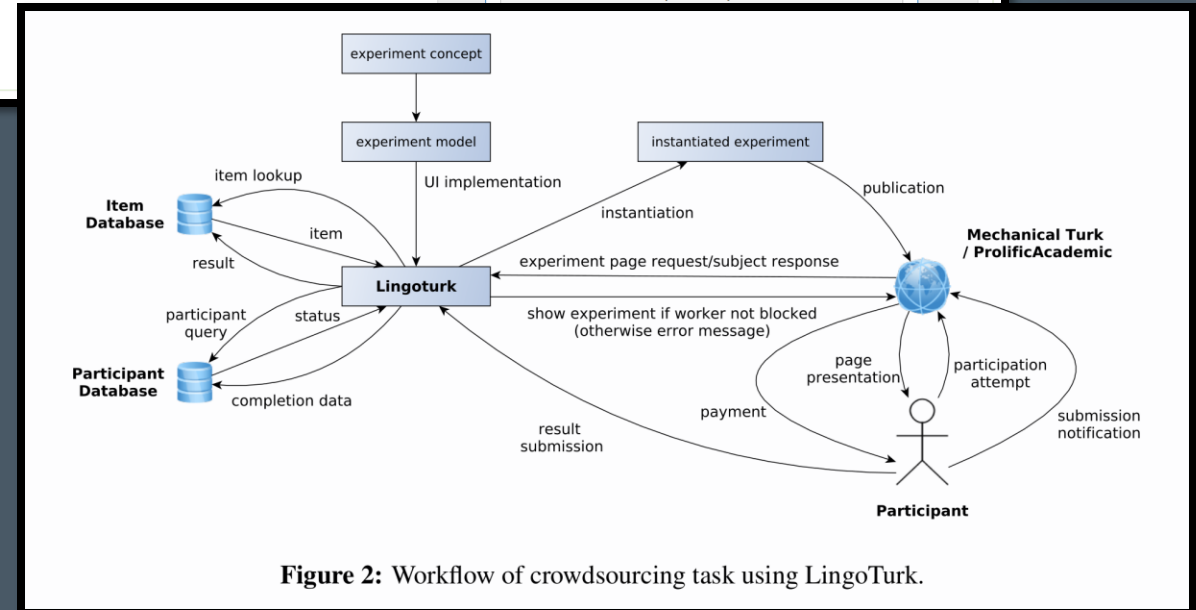


Figure 2: Workflow of crowdsourcing task using LingoTurk.

REDCap

- ▶ Server for running survey-based studies
- ▶ Free for our non-profits

Links to demos

- ▶ <https://projectredcap.org/software/try/>

Demo of all question types

- ▶ <https://redcap.vanderbilt.edu/surveys/?s=iTF9X7>

EXAMPLE SURVEY (your survey title goes here)

Resize font:   [Returning?](#)

These are your survey instructions that you would enter for your survey participants. You may put whatever text you like here, which may include information about the purpose of the survey, who is taking the survey, or how to take the survey.

This survey here is using a single weblink for all respondents, which can be posted on a webpage or emailed out from your email application of choice. **By default, all survey responses are collected anonymously** (that is, unless your survey asks for name, email, or other identifying information).

If you wish to track individuals who have taken your survey, you may upload a list of email addresses into a Contact List within REDCap Survey, in which you can have REDCap Survey send them an email invitation, which will track if they have taken the survey and when it was taken. This method still collects responses anonymously, but if you wish to identify an individual respondent's answers, you may do so by also providing an Identifier in your Contact List. Of course, in that case you may want to inform your respondents in your survey's instructions that their responses are not being collected anonymously and can thus be traced back to them.

The third method for collecting responses is to enter them manually yourself, which is especially helpful when you have received the respondents' survey responses in paper format.

Section 1 (This is a section header with descriptive text. It only provides informational text and is used to divide the survey into sections for organization. If the survey is set to be displayed as "one section per page", then these section headers will begin each new page of the survey.)

You may capture a respondent's signature electronically.

 [Add signature](#)

You may create MULTIPLE CHOICE questions and set the answer choices for them. You can have as many answer choices as you need. This multiple choice question is rendered as RADIO buttons.

- ☐ Choice One
☐ Choice Two
☐ Choice Three
☐ Etc.

[reset](#)

You may also set multiple choice questions as DROP-DOWN MENUS.

This is a TEXT BOX, which allows respondents to enter a small amount of text. A Text Box can be validated, if needed, as a number, integer, phone number, email, or zipcode. If validated as a number or integer, you may also set the minimum and/or maximum allowable values.

This question has "number" validation set with a minimum of 1 and a maximum of 10.

This type of multiple choice question, known as CHECKBOXES, allows for more than one answer choice to be selected, whereas radio buttons and drop-downs only allow for one choice.

- ☐ Choice One
☐ Choice Two
☐ Choice Three
☐ Select as many as you like

You can create YES-NO questions.

- ☐ Yes
☐ No

This question has vertical alignment of choices on the right.

[reset](#)

And you can also create TRUE-FALSE questions.

- ☐ True ☐ False

This question has horizontal alignment.

[reset](#)

DATE questions are also an option. If you click the calendar icon on the right, a pop-up calendar will appear, thus allowing the respondent to easily select a date. Or it can be simply typed in.

  Today Y-M-D

The FILE UPLOAD question type allows respondents to upload any type of document to the survey that you may afterward download and open when viewing your

 [Upload file](#)

Platforms

Prolific Academic

- ▶ Aimed at academic and market research
- ▶ Extensive screening criteria
- ▶ No design interface (recruitment only)
- ▶ 33% fee
- ▶ 10s of thousands of participants

More like traditional recruitment

<https://www.prolific.ac>

Mechanical Turk

- ▶ Aimed at "Human Intelligence Tasks"
- ▶ Limited screening criteria
- ▶ Limited design interface
- ▶ 40% fee
- ▶ 100s of thousands of participants

More like hiring temp workers

<https://www.mturk.com>

Best Practices

Ethics Oversight

- ▶ Requirements vary: check your uni
 - ▶ e.g. user studies on staff and students may be exempt while crowdsourcing is not
- ▶ Regardless of status, report presence/absence of ethical oversight in papers

Compensation

- ▶ General consensus: pay **at least minimum wage** in your jurisdiction
- ▶ Estimate time before hand
 - ▶ Pilot to improve estimate
- ▶ Bonus payments if necessary

Reporting your results

- ▶ How many subjects did you recruit?
 - ▶ Where did you recruit them?
 - ▶ What do we need to know about them (demographics)?
-
- ▶ Did you obtain an ethics review?
 - ▶ How did you collect informed consent?
 - ▶ How did you compensate subjects?

Reporting your results

- ▶ How many subjects did you recruit?
- ▶ Where did you recruit them?
- ▶ What do we need to know about them (demographics)?
- ▶ Did you obtain an ethics review?
- ▶ How did you collect informed consent?
- ▶ How did you compensate subjects?

3 Crowd Sourcing Ratings

To collect human judgements from a diverse group of speakers of US English, we used Amazon's Mechanical Turk service (AMT) to run two experiments. In the first experiment, subjects rated the naturalness of 174 passages used in Walker et al.'s (2007) study. As detailed in Section 5, this validation experiment confirmed that the judge-

ments collected on AMT correlate with those of the raters in Walker et al.'s (2007) study. Our second experiment collected ratings on 300 passages realized with modifications for better contrast expression (WITHMODS) and 300 passages without these modifications (NOMODS), both realized using OpenCCG. While this does not admit a direct comparison to the realizations produced by Walker et al. (2007), this controls for differences between the generators other than the variable of interest: the contrastive enhancements. In addition to these materials, five passages from the SRC were seen by all subjects to control for anomalous subject behavior.

Reporting your results

- ▶ How many subjects did you recruit?
- ▶ Where did you recruit them?
- ▶ What do we need to know about them (demographics)?
- ▶ Did you obtain an ethics review?
- ▶ How did you collect informed consent?
- ▶ How did you compensate subjects?

3.1 Survey Format

Each survey used demographic questions to determine the native speaker status of the subject. Instructions for completing comprehension questions and rating realizations followed the demographic questions.³ Each subject saw fifteen stimuli, each consisting of a sample user query and the target passage as in Figure 5. After reading the stimulus, the subject

prehesion question rated the natural point Likert scale *very natural*. At subject could offer responses, or ask questions of the researchers. The average completion time across all experiments was about ten minutes.

Passage selection is detailed in §3.3 and §3.4.

3.2 Quality Control

We used three strategies to filter out low-quality responses from AMT subjects.

Comprehension Questions A template-based yes-or-no question (exemplified in Figure 5) followed each passage. Subjects who answered less

Reporting your results

- ▶ How many subjects did you recruit?
- ▶ Where did you recruit them?
- ▶ What do we need to know about them (demographics)?
- ▶ Did you obtain an ethics review?
- ▶ How did you collect informed consent?
- ▶ How did you compensate subjects?

Subject Demographics Sixty-eight subjects responded to these 180 surveys initially. Subjects were allowed to complete up to six distinct surveys. One subject's data was excluded for non-native status and another's was excluded on the basis of uniform ratings (as detailed in §3.2). To compensate for the eight surveys completed by these subjects and ten surveys mistakenly administered in draft format, we recollected data for 18 of the 180 surveys. This resulted in a final pool of 80 subjects with an average (std. dev.) age 37.15 (13.5) years. Forty identified as female, thirty-nine identified as male, and one identified as non-gendered.

Because subjects in the validation study completed the survey in about 10 minutes on average with a standard deviation of about 5 minutes, we scaled the pay to \$2.00 per survey in this experiment. Since subjects could participate in this experiment multiple times, they could receive up to \$12.00 for their contribution.

Resources

Crowdsourcing Dialogue

- ▶ <https://github.com/batra-mlp-lab/visdial-amt-chat>
- ▶ <https://github.com/clp-research/slurk>
- ▶ <https://parl.ai/static/docs/index.html>
- ▶ <https://github.com/bsu-slim/prompt-recorder> (recording audio)

Tutorials

- ▶ Mechanical Turk: <https://blog.mturk.com/tutorials/home>

References

Howcroft, Nakatsu, & White. 2013. [Enhancing the Expression of Contrast in the SPaRky Restaurant Corpus](#). *ENLG*.

Howcroft, Klakow, & Demberg. [The Extended SPaRky Restaurant Corpus: designing a corpus with variable information density](#). *INTERSPEECH*.

Mairesse, Gašić, Jurčiček, Keizer, Thomson, Yu, & Young. 2010. [Phrase-based Statistical Language Generation using Graphical Models and Active Learning](#). *ACL*.

Novikova, Lemon, & Rieser. 2016. [Crowd-sourcing NLG Data: Pictures Elicit Better Data](#). *INLG*.

Scholman & Demberg. 2017. [Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task](#). *Proc. of the 11th Linguistic Annotation Workshop*.

Shifting Gears...

Does the way we use these
tools make sense?

Human Evaluation Criteria

Fluency

- ▶ Clarity
- ▶ Fluency
- ▶ Grammaticality
- ▶ Naturalness
- ▶ Readability
- ▶ Understandability
- ▶ ...

Adequacy

- ▶ Accuracy
- ▶ Completeness
- ▶ Informativeness
- ▶ Relevance
- ▶ Similarity
- ▶ Truthfulness
- ▶ Importance
- ▶ Meaning-Preservation
- ▶ Non-Redundancy
- ▶ ...

Operationalizing the Criteria

Grammaticality

- ▶ 'How do you judge the overall quality of the utterance in terms of its grammatical correctness and fluency?'
- ▶ 'How would you grade the syntactic quality of the [text]?'
- ▶ 'This text is written in proper Dutch.'

Readability

- ▶ 'How hard was it to read the [text]?'
- ▶ 'This is sometimes called "fluency", and ... decide how well the highlighted sentence reads; is it good fluent English, or does it have grammatical errors, awkward constructions, etc.'
- ▶ 'This text is easily readable.'

Sample sizes and statistics

van der Lee et al. (2019)

- ▶ 55% of papers give sample size
- ▶ "10 to 60 readers"
- ▶ "median of 100 items used"
 - ▶ range from 2 to 5400

We do not know what the expected effect sizes are or what appropriate sample sizes are for our evaluations!

Improving Evaluation Criteria

Validity begins with good definitions

- ▶ discriminative & diagnostic

Reliability is an empirical property

- ▶ Test-retest consistency
- ▶ Interannotator agreement
- ▶ Generalization across domains
- ▶ Replicability across labs

Developing a standard

- ▶ Survey of current methods
- ▶ Statistical simulations
- ▶ Organizing an experimental shared task
- ▶ Workshop with stakeholders
- ▶ Release of guidelines+templates

Objective Measures: Reading Time

In NLG Evaluation:

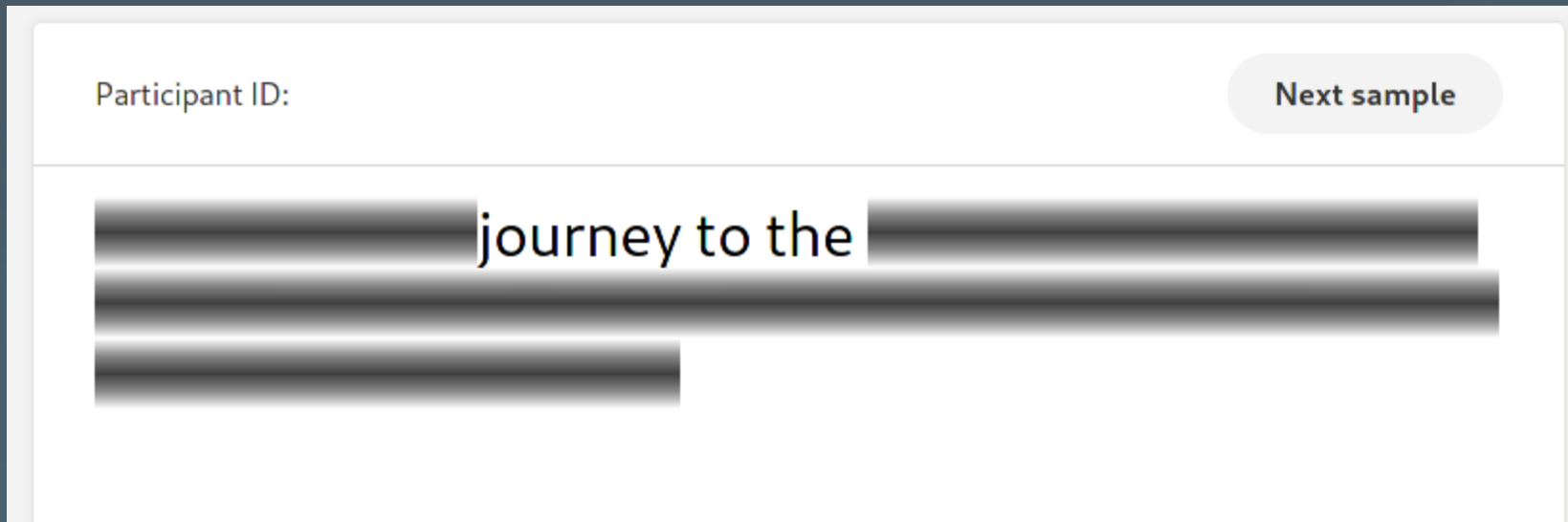
- ▶ Belz & Gatt 2008 – RTs as extrinsic measure
- ▶ Zarrieß et al. 2015 – sentence-level RTs

In psycholinguistics

- ▶ eye-tracking & self-paced reading
- ▶ understanding human sentence processing

Reading times can indicate fluency/readability

Objective Measures: Reading Time



The screenshot shows a web-based interface for measuring reading time. At the top left, there is a label "Participant ID:" followed by a text input field. At the top right, there is a button labeled "Next sample". Below these elements, there is a large text area containing the phrase "journey to the" followed by a long horizontal bar that has been highlighted by a mouse cursor, indicating the text being read.

Mouse-contingent reading times

Better evaluations → better proxies

Evaluations involving humans are expensive.

- ▶ So folks use invalid measures like BLEU

With better evaluations (↑validity, ↑reliability)

- ▶ Better targets for automated metrics

Better automated metrics → better objective functions!

Conclusion

Crowdsourcing

- ▶ Interesting tasks abound
- ▶ Tools to make life easier
- ▶ Best practices for conduct and reporting

Slides available at:

https://davehowcroft.com/talk/2020-01_glasgow/

Improving NLG Evaluation

For survey methods

- ▶ Better validity and reliability
- ▶ Statistical simulations
- ▶ Community efforts
 - ▶ Shared task & workshop

For objective methods

- ▶ Mouse-contingent reading times

Bringing it together

- ▶ Seeking better automated proxies