Lessons from a User Experience Evaluation of NLP Interfaces

Eduardo Calò
Utrecht University
Utrecht, The Netherlands
e.calo@uu.nl

Lydia Penkert
Independent Researcher
Cologne, Germany
lydiapenkert@gmail.com

Saad Mahamood
trivago N.V.
Düsseldorf, Germany
saad.mahamood@trivago.com

Abstract

Human evaluations lay at the heart of evaluations within the field of Natural Language Processing (NLP). Seen as the "golden standard" of evaluations, questions are being asked on whether these evaluations are both reproducible and repeatable. One overlooked aspect is the design choices made by researchers when designing user interfaces (UIs). In this paper, four UIs used in past NLP human evaluations are assessed by UX experts, based on standardized human-centered interaction principles. Building on these insights, we derive several recommendations that the NLP community should apply when designing UIs, to enable more consistent human evaluation responses.

1 Introduction

Reproducible and repeatable evaluation lays at the heart of science. Increasingly for the field of Natural Language Processing (NLP), questions are being asked on whether the evaluations conducted by researchers are in fact reproducible and repeatable. Only a minority of published experiments can be reproduced, due to either non-working and non-functional code or resource limits, such as financial or time limits (Belz et al., 2021). Estimates range between $5 - 20\%$ of papers being repeatable without significant barriers if the original author(s) help is sought (Belz and Thomson, 2023).

The design of user interfaces (UIs) plays an important role in conducting effective and reliable human evaluations. This aspect is commonly overlooked by researchers, although it has been shown that giving task-adequate and usability-conforming UIs to evaluators increases the quality of the annotations gathered. However, researchers often design human evaluations quickly, overlooking the fact that the way a human evaluation is presented directly impacts the quality of the data they collect (Huynh et al., 2021). Flaws within UIs for collecting responses have been observed in past reproduction attempts of human evaluations (Belz and Thomson, 2023). Confusing UIs make it challenging for participants to give correct ratings due to an error-prone means of collecting responses (Thomson et al., 2024). Particularly, Sullivan Jr. et al. (2022) show that the choices made to design UIs critically impact the characteristics of rationales collected from participants: When given dragging affordance (i.e., the ability to drag to select more words at once), users select significantly more words than without it.

Given the importance of human evaluations in NLP and the increasing use of crowdsourced tasks (Shmueli et al., 2021), it is crucial to understand how researchers can apply standardized human-centered design (HCD) principles to the interfaces for human evaluations. By applying such principles, researchers will be able to create interfaces with a greater degree of usability for respondents and possibly solicit less error-prone responses. This might eliminate one source of reproducibility challenges and result in increasing the quality and reproducibility of NLP human evaluations.

To better understand these issues, we conducted an exploratory study in which we asked user experience experts to assess UIs used in past human evaluations. We present the results from the evaluation of these interfaces, summarize the general lessons we learned, and draw convenient recommendations that can be applied to designing UIs for human evaluations in NLP.

2 Background

2.1 Human Evaluation Practices in NLP

Human evaluations can either be intrinsic (i.e., evaluating properties of a given text) or extrinsic (i.e., evaluating the effectiveness of a given system) (Gatt and Krahmer, 2018). For intrinsic human evaluations, humans are involved in reading and rating texts, such as comparing generated

Findings of the Association for Computational Linguistics: NAACL 2025, pages 2915–2929
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

texts against human texts, for criteria such as quality, correctness, naturalness, understandability, etc. (Gkatzia and Mahamood, 2015; Belz et al., 2020). The process of humans providing their annotations for these evaluations can be seen as a psychological process (Pandey et al., 2022). Hence, human factors impact the quality of annotations during the annotation process, with attentional heuristics and high mental workload identified as influential factors. Additionally, information scientists have observed that annotation types affect human annotation quality through factors such as objectivity and descriptiveness (Cheng and Cosley, 2013). Consequently, the careful design of UIs to collect responses is of high importance if researchers are to avoid erroneous responses.

2.2 Human-centered Design for UIs

Human-centered design (HCD) aims to enhance the usefulness and usability of interactive systems by prioritizing the understanding of the needs of the users. By integrating principles from ergonomics, and usability knowledge and methods, HCD ensures that interactive systems are tailored to users' explicit needs, encompassing their goals, tasks, resources, and environments (UXQB e.V., 2022).

A key part of a successful human-centered design is usability, which enhances the system's effectiveness, efficiency, and user satisfaction within a defined context of use. Throughout the design process, design patterns and standardized interaction principles should be considered to ensure that the solutions are usable and meet users' needs.

For interactive systems, especially those utilized for repetitive tasks like annotation, efficiency is paramount to contribute to a positive user experience but also to ensure the quality of the outcome of the task itself.

ISO-9241-110 (2020) lists seven interaction principles that should be met when designing interactive systems, which we adopt in our paper:

- **Suitability for the user's tasks**: *the UI supports the users in the completion of their tasks.*
- **Self-descriptiveness**: *appropriate information is presented in the UI to make its capabilities and use immediately obvious.*
- **Conformity with user expectations**: *the UI's behavior is predictable based on the context of use and commonly accepted conventions in that context.*
- **Learnability**: *the UI supports the discovery of its capabilities, allows exploration, provides support, and minimizes the need for learning.*
- **Controllability**: *the user maintains control of the UI and the interactions' speed, sequence, and individualization.*
- **Use error robustness**: *the UI tolerates and assists the user in avoiding and recovering from errors.*
- **User engagement**: *functions and information are presented in an inviting and motivating manner.*

3 Methodology

3.1 Interface Selection

We selected four human evaluation UIs to assess. These UIs featured in papers that are part of the ReproHum project,[1] which attempts to investigate the reproducibility of human evaluations within NLP. With the original author(s) consent, the selection criteria for papers in ReproHum depends on the availability of sufficient details regarding materials (code, data, etc.) and evaluation procedures (Belz et al., 2023). After contacting the organizers of the project, we were given advice on which UIs would be of relevant interest for our evaluation.

For the purposes of our evaluation, we chose to focus only on papers that dealt with intrinsic evaluations and deliberately excluded evaluation interfaces that relied on either using text files or Excel spreadsheets. We did this for two reasons: (i) we wanted to focus only on interfaces that were used by crowdworkers. Since most crowdworkers are not experts, UI choices matter; (ii) shortcomings in the use of these modalities to receive user input have been reported (e.g., Ito et al., 2023). We randomly chose the following three papers and the interfaces therein to give us a snapshot of practices:

1. "It's not Rocket Science: Interpreting Figurative Language in Narratives" by Chakrabarty et al. (2022). We focus on the Amazon Mechanical Turk (MTurk)[2] interface used to rate the plausibility of machine- and human-generated idioms and similes from a given written fictional narrative (henceforth, **FL**).
2. "Data-to-text Generation with Macro Planning" by Puduppully and Lapata (2021). Our focus is to evaluate the MTurk interface used for fact validation, in which participants are given a set of tabular data and a set of gener-

[1] https://reprohum.github.io

[2] https://www.mturk.com/

ated sentences and asked to give the number of correct and/or incorrect facts (henceforth, **MLBF**). We also evaluate a second interface to measure the intrinsic quality of a generated output relative to another output (henceforth, **MLBC**). For these two evaluations, we restrict ourselves to the MLB (Major League Baseball) dataset (Puduppully et al., 2019) used by the authors.

3. "NeuralREG: An end-to-end approach to referring expression generation" by Castro Ferreira et al. (2018). The interface used in this paper is a bespoke implementation that asks users to rate three intrinsic text qualities (fluency, grammaticality, and clarity) of a generated summary text containing highlighted referring expressions relative to an input set of tabular data (henceforth, **REG**).

For the first two MTurk-based experiments, their respective HTML interfaces were modified to incorporate experiment data, as normally, these template holders are filled automatically by the MTurk platform. All interfaces[3] were hosted on a web server and made interactive to enable the evaluation to be as close as possible to the experience seen by the original evaluators.

3.2 Evaluation Procedure

For our evaluation, we recruited three user experience (UX) experts who are professional contacts of one of the authors. They have between 7 and 16 years of professional expertise and are experienced in conducting usability evaluations. One of the recruited experts has high familiarity with NLP, whilst the other two only have medium and low familiarity, respectively. However, since the UX experts were assigned to focus exclusively on possible UX issues, we do not believe that the level of NLP familiarity would have changed the outcome of their evaluations.

The experts were asked to evaluate each UI following the seven interaction principles for designing interactive systems (see §2.2) on a 3−point scale (*"not met"*, *"partially met"*, *"met"*). If the experts selected *"not met"* or *"partially met"*, they were asked to give the motivations for which the principle was not (fully) met. See Appendix A for the instructions given and the questions asked to the experts.[4] We randomized the order in which the interfaces were presented to avoid order bias.

Principle	**REG**	**FL**	**MLBC**	**MLBF**
Suitability	**2.000**	0.667	1.333	0.333
Self-descriptiveness	**0.667**	0.333	**0.667**	0.000
Conformity	**1.000**	0.333	0.000	0.000
Learnability	**2.000**	1.667	0.333	0.333
Controllability	1.000	0.667	0.000	**1.333**
Robustness	1.000	**1.667**	0.667	0.000
Engagement	0.667	**1.333**	0.000	0.000
Overall	**1.190**	0.952	0.429	0.286

Table 1: Rankings per principle and overall. Values in bold are of the interfaces that ranked first per principle and overall.

4 Results

To assess the consistency, we computed expert inter-annotator agreement (IAA) over all the interfaces and principles (Krippendorff's $\alpha = 0.339$). We also computed IAA per interface and per principle. See Table 3 and Table 4 in Appendix B for the detailed figures. Several findings are noteworthy, such as the extremely low agreement for FL among the interfaces and for Self-descriptiveness among the principles. In addition, there is moderate agreement for REG among the interfaces and for Conformity among the principles. Overall, IAA scores range from low to moderate, which is not surprising given the highly subjective nature of the task. Moreover, the fact that three UX experts have difficulty agreeing on the strengths and weaknesses of the evaluated interfaces shows that there are significant challenges in performing this type of evaluation using established interaction principles.

To see how the interfaces fared among each other, we ranked the interfaces both by principle and overall aspects. We mapped the categorical judgments given by the experts into numerical ratings (i.e., *"not met"*: 0, *"partially met"*: 1, *"met"*: 2, with intervals between the numerical ratings being equal) and then computed the rankings as the means of the numerical ratings (per principle and overall). See Table 1 for the figures. REG outperforms the other interfaces on many principles, while both MLB interfaces are the most deficient.

Furthermore, we performed a qualitative analysis of the comments we received from the experts when the principles were not (fully) met. One of

[3] See Appendix D for the screenshots of the interfaces.

[4] The raw annotations can be found at `https://doi.org/10.5281/zenodo.14730831`.

Principle	Recommendations
Suitability	• Add a submit button (see Limitations)
Self-descriptiveness	• Avoid confusing/subjective/judgmental/technical/redundant language • Avoid long instructions, but if needed explain/present them properly • Explain any part that may turn out to be unclear
Conformity	• Ensure uniformity in layout (e.g., length of the input fields) • Use proper/consistent colors (e.g., brightness, palette, etc.) • Organize/structure and position text in the right way • Use the appropriate type of question based on the data you want to collect
Learnability	• Provide the right amount of examples • Explain the terminology • Give feedback • Explain how to interact with the system
Controllability	• Provide users with the ability to revisit the instructions • Enable empty state revert
Robustness	• Clearly mark mandatory information • Provide proper error messages (e.g., not too early, not persistent, not generic) • Check input data in the backend • Check if unwanted interactions with UI/text may occur • Avoid default answers that may be misleading (e.g., default value of a slider)
Engagement	• Add a progress bar • Do not use aggressive language (e.g., all-caps) • Avoid heavy text/content/tables • Give positive feedback after completion

Table 2: Summary of the recommendations organized per principle.

the authors of the paper categorized the common trends in the comments to derive the recommendations (see §5). See Appendix C for some particular examples. In general, the analysis revealed several issues across different interfaces and principles.

Suitability is compromised by the absence of a submit button (FL, MLBF; see Limitations). **Self-descriptiveness** is hampered by the confusing placement of questions, the use of vague and subjective terms (REG), misleading information accompanying the choices (FL), long and technical instructions, with a lack of visual or textual hierarchy (MLBF), and redundant information (MLBC). **Conformity** is violated by a lack of uniformity in the layout (REG), odd color selection (REG, MLBF), inconsistent question formatting and positioning (FL), improper separation of sections, inappropriate use of free text fields, and non-standard information structuring (MLBC). **Learnability** suffers from an inadequate number of examples provided (FL), a lack of explanation of abbreviations and exercise feedback (MLBF), and the absence of a direct way to learn how to use the system (MLBC). **Controllability** issues arise from the impossibility for the users to return to the instructions (REG, MLBC), unclear indications of task completion (REG), the impossibility of reverting to questions' empty state (FL), and the disappearance of the options' labels after introducing the value (MLBF). **Robustness** is compromised by mandatory fields being unmarked (REG), bad handling of error messages (REG, MLBF, MLBC), input data not being checked after insertion (REG, MLBC), arguable choices in questions' default values (REG), the possibility of unwanted interaction with text (FL), and the wrong choice of question types (MLBC). **Engagement** suffers from the lack of progress indication (REG, MLBC), the use of aggressive language (FL, MLBF), the usage of heavy texts and tables (MLBF, MLBC), and the lack of positive feedback after task completion (MLBC).

5 Recommendations and Conclusion

Table 2 summarizes the main recommendations from our analyses. This exploratory study, despite a small sample, has revealed numerous flaws, evidencing the insufficient effort invested in designing UIs. The primary value of our study lies in the qualitative feedback, which serves as a strong indicator of the significant potential for improvement. Many of the issues we found could be readily addressed with minimal effort. Minor improvements in UI design can already have a substantial impact. Moreover, incorporating user considerations is something researchers should take into consideration (e.g., through piloting (Sripada et al., 2005; van Miltenburg et al., 2021), etc.). Such considerations might enable better and more consistent user responses, enhancing user satisfaction and potentially improving the reproducibility of the results.

Fortunately, steps towards blending human-computer interaction and NLP have been taken by the community (e.g., Blodgett et al., 2021, 2022, 2024; Luo, 2023; Soni et al., 2024). We hope that our recommendations will contribute to this aim and provide guidance for future development, enhancing the usability of interactive systems and possibly increasing the reliability of annotated data.

Limitations

The way we evaluated the interfaces (i.e., hosting HTML interfaces originally meant for MTurk on a web server) posed a constraint on how we could (not) present the submit button, resulting in multiple (unfairly negative) feedback from the experts on Suitability.

This study is exploratory in nature, as we focus on the evaluation of just four UIs. Despite the small sample size, we uncovered numerous issues. In future work, we would like to analyze more evaluation UIs in more papers concerning different NLP tasks. Furthermore, we intend to select one of the evaluated UIs, redesign it based on the recommendations from this study, and run new human evaluations comparing the original and redesigned versions, to assess the impact of a better UI design on the quality of the data collected.

Ethical Considerations

The three experts were not remunerated and voluntarily accepted to participate in the experiment after giving informed consent.

References

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2023. The 2023 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 35–48, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Su Lin Blodgett, Amanda Cercas Curry, Sunipa Dev, Michael Madaio, Ani Nenkova, Diyi Yang, and Ziang Xiao, editors. 2024. *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Mexico City, Mexico.

Su Lin Blodgett, Hal Daumé III, Michael Madaio, Ani Nenkova, Brendan O'Connor, Hanna Wallach, and Qian Yang, editors. 2022. *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington.

Su Lin Blodgett, Michael Madaio, Brendan O'Connor, Hanna Wallach, and Qian Yang, editors. 2021. *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Online.

Thiago Castro Ferreira, Diego Moussallem, Ákos Kádár, Sander Wubben, and Emiel Krahmer. 2018. NeuralREG: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1969, Melbourne, Australia. Association for Computational Linguistics.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Justin Cheng and Dan Cosley. 2013. How annotation styles influence content and preferences. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, HT '13, page 214–218, New York, NY, USA. Association for Computing Machinery.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Dimitra Gkatzia and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005 - 2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 57–60, Brighton, UK. Association for Computational Linguistics.

Jessica Huynh, Jeffrey P. Bigham, and Maxine Eskénazi. 2021. A survey of nlp-related crowdsourcing hits: what works and what does not. *CoRR*, abs/2111.05241.

ISO-9241-110. 2020. Ergonomics of human-system interaction — Part 110: Interaction principles.

Takumi Ito, Qixiang Fang, Pablo Mosteiro, Albert Gatt, and Kees van Deemter. 2023. Challenges in reproducing human evaluation results for role-oriented dialogue summarization. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 97–123, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Lin Luo. 2023. Influence of interface design driven by natural language processing on user participation. *Frontiers in Business, Economics and Management*, 11:63–66.

Rahul Pandey, Hemant Purohit, Carlos Castillo, and Valerie L. Shalin. 2022. Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. *International Journal of Human-Computer Studies*, 160:102772.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2023–2035, Florence, Italy. Association for Computational Linguistics.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text Generation with Macro Planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.

Nikita Soni, Lucie Flek, Ashish Sharma, Diyi Yang, Sara Hooker, and H. Andrew Schwartz, editors. 2024. *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*. ACL, TBD.

Somayajulu Sripada, Ehud Reiter, and Lezan Hawizy. 2005. Evaluation of an NLG system using post-edit data: Lessons learnt. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, Aberdeen, Scotland. Association for Computational Linguistics.

Jamar Sullivan Jr., Will Brackenbury, Andrew McNutt, Kevin Bryson, Kwam Byll, Yuxin Chen, Michael Littman, Chenhao Tan, and Blase Ur. 2022. Explaining why: How instructions and user interfaces impact annotator rationales when labeling text data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 521–531, Seattle, United States. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common Flaws in Running Human Evaluation Experiments in NLP. *Computational Linguistics*, pages 1–11.

UXQB e.V. 2022. CPUX-F Curriculum – Certified Professional for Usability and User Experience Foundation Level.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

A Instructions to Annotators

Annotators were asked to provide feedback on a Word document containing the instructions, the links to the interfaces, and the questions. Figure 1 shows the instructions that the experts received and Figure 2 the questions they were asked.

B Additional Experimental Results

Table 3 and Table 4 show the IAA per interface and per principle, respectively.

Interface	α
REG	0.500
FL	0.041
MLBF	0.167
MLBC	0.279

Table 3: Krippendorff's α per interface.

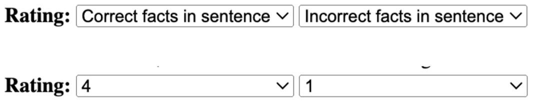
Principle	α
Suitability	0.298
Self-descriptiveness	0.057
Conformity	0.656
Learnability	0.298
Controllability	0.013
Robustness	0.500
Engagement	0.389

Table 4: Krippendorff's α per principle.

C Examples of Identified Areas for Improvement

In this section, we report some notable examples of flaws we found in the UIs.

In MLBF (Figure 3), the label description is placed within the drop-down options. In Figure 3 top, the default state is represented, while in Figure 3 bottom, the status after submitting a rating. This represents a controllability problem, as users are not able to see the label of the input field.

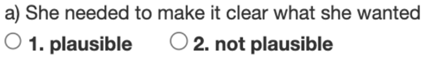


Rating: Correct facts in sentence | Incorrect facts in sentence

Rating: 4 | 1

Figure 3: MLBF - Controllability issue.

In FL (Figure 4), "plausible" is preceded by "1" and "not plausible" by "2". This represents a self-descriptiveness problem, as there is no apparent reason for the attribution of those numbers to the two options.

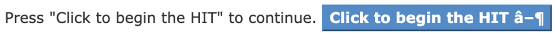


a) She needed to make it clear what she wanted

○ **1. plausible** ○ **2. not plausible**

Figure 4: FL - Self-descriptiveness issue.

In MLBC (Figure 5), redundant and duplicated information is present between the text on the left and the button label on the right. This represents a self-descriptiveness problem.

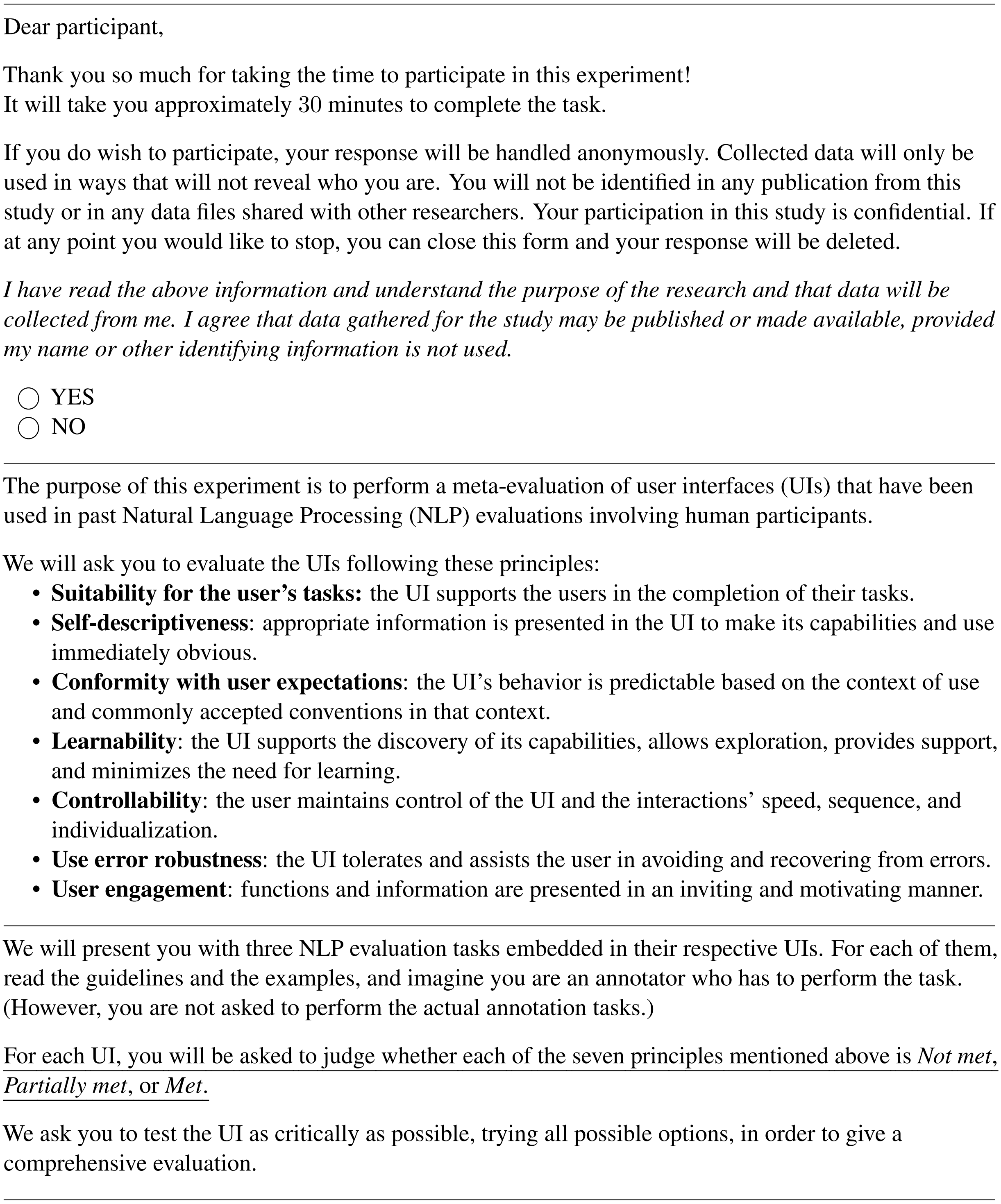


Press "Click to begin the HIT" to continue.

Figure 5: MLBC - Self-descriptiveness issue.

D Screenshots of the Interfaces

Figure 6 shows the FL interface. Figure 7 and Figure 8 show the MLBC instructions and task, respectively. Figure 9, Figure 10, and Figure 11 show the MLBF instructions, while Figure 12 and Figure 13 the MLBF task. Figure 14 and Figure 15 show the REG instructions and task, respectively.



Dear participant,

Thank you so much for taking the time to participate in this experiment!
It will take you approximately 30 minutes to complete the task.

If you do wish to participate, your response will be handled anonymously. Collected data will only be used in ways that will not reveal who you are. You will not be identified in any publication from this study or in any data files shared with other researchers. Your participation in this study is confidential. If at any point you would like to stop, you can close this form and your response will be deleted.

I have read the above information and understand the purpose of the research and that data will be collected from me. I agree that data gathered for the study may be published or made available, provided my name or other identifying information is not used.

○ YES
○ NO

The purpose of this experiment is to perform a meta-evaluation of user interfaces (UIs) that have been used in past Natural Language Processing (NLP) evaluations involving human participants.

We will ask you to evaluate the UIs following these principles:

- **Suitability for the user's tasks:** the UI supports the users in the completion of their tasks.
- **Self-descriptiveness**: appropriate information is presented in the UI to make its capabilities and use immediately obvious.
- **Conformity with user expectations**: the UI's behavior is predictable based on the context of use and commonly accepted conventions in that context.
- **Learnability**: the UI supports the discovery of its capabilities, allows exploration, provides support, and minimizes the need for learning.
- **Controllability**: the user maintains control of the UI and the interactions' speed, sequence, and individualization.
- **Use error robustness**: the UI tolerates and assists the user in avoiding and recovering from errors.
- **User engagement**: functions and information are presented in an inviting and motivating manner.

We will present you with three NLP evaluation tasks embedded in their respective UIs. For each of them, read the guidelines and the examples, and imagine you are an annotator who has to perform the task. (However, you are not asked to perform the actual annotation tasks.)

For each UI, you will be asked to judge whether each of the seven principles mentioned above is *Not met*, *Partially met*, or *Met*.

We ask you to test the UI as critically as possible, trying all possible options, in order to give a comprehensive evaluation.

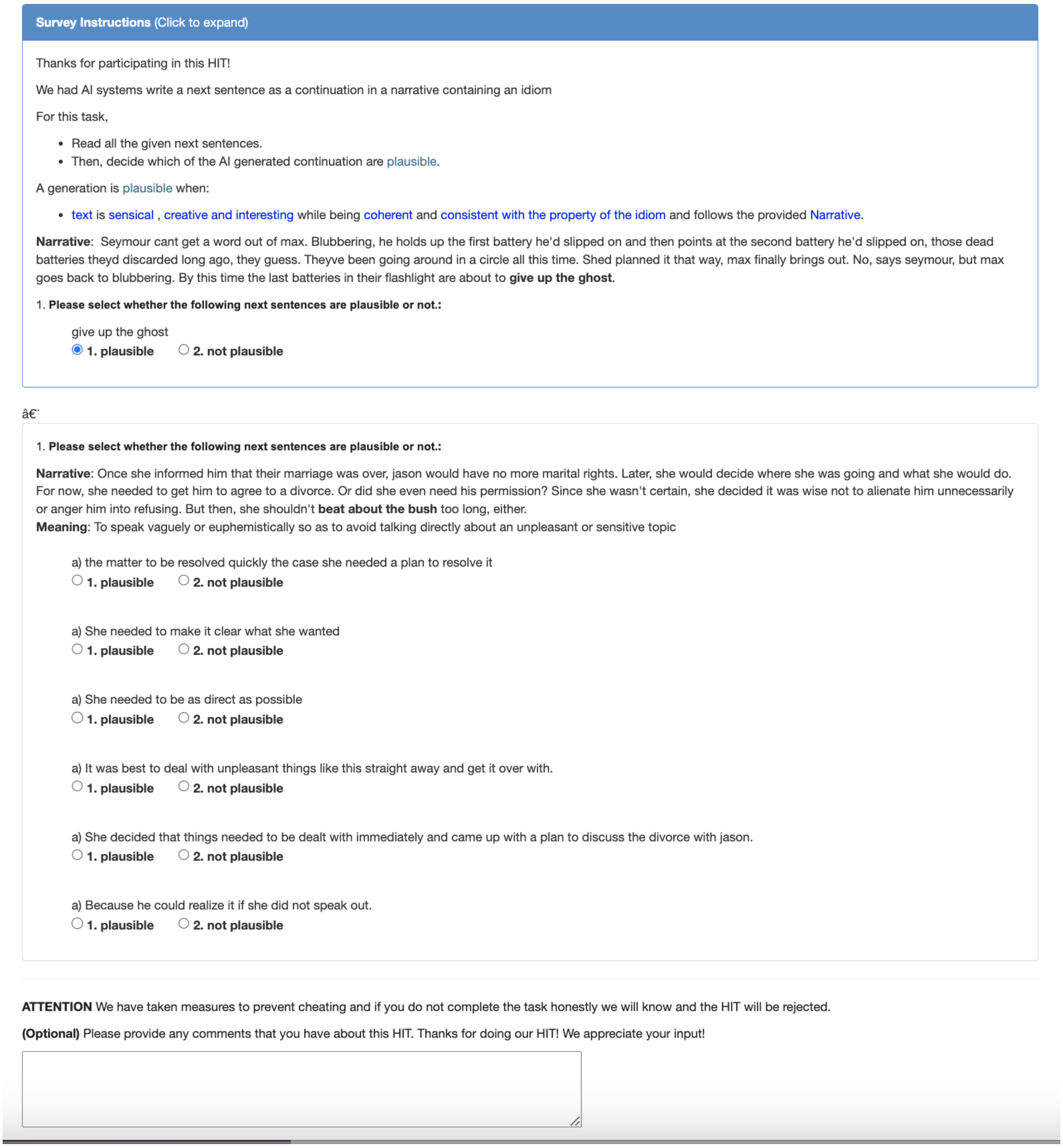
Figure 1: The instructions provided to the experts.

INTERFACE: `Link to the interface`

Are the following principles met?

- **Suitability**: *Not met*, *Partially met*, *Met*
  - If you answered *Not met* or *Partially met*, why do you think the principle is not (fully) met?
- **Self-descriptiveness**: *Not met*, *Partially met*, *Met*
  - If you answered *Not met* or *Partially met*, why do you think the principle is not (fully) met?
- **Conformity**: *Not met*, *Partially met*, *Met*
  - If you answered *Not met* or *Partially met*, why do you think the principle is not (fully) met?
- **Learnability**: *Not met*, *Partially met*, *Met*
  - If you answered *Not met* or *Partially met*, why do you think the principle is not (fully) met?
- **Controllability**: *Not met*, *Partially met*, *Met*
  - If you answered *Not met* or *Partially met*, why do you think the principle is not (fully) met?
- **Robustness**: *Not met*, *Partially met*, *Met*
  - If you answered *Not met* or *Partially met*, why do you think the principle is not (fully) met?
- **Engagement**: *Not met*, *Partially met*, *Met*
  - If you answered *Not met* or *Partially met*, why do you think the principle is not (fully) met?

Figure 2: The questions asked to the experts for each interface.



Survey Instructions (Click to expand)

Thanks for participating in this HIT!

We had AI systems write a next sentence as a continuation in a narrative containing an idiom

For this task,

- Read all the given next sentences.
- Then, decide which of the AI generated continuation are plausible.

A generation is plausible when:

- text is sensical , creative and interesting while being coherent and consistent with the property of the idiom and follows the provided Narrative.

Narrative: Seymour cant get a word out of max. Blubbering, he holds up the first battery he'd slipped on and then points at the second battery he'd slipped on, those dead batteries theyd discarded long ago, they guess. Theyve been going around in a circle all this time. Shed planned it that way, max finally brings out. No, says seymour, but max goes back to blubbering. By this time the last batteries in their flashlight are about to **give up the ghost**.

1. **Please select whether the following next sentences are plausible or not.:**

give up the ghost
● **1. plausible** ○ **2. not plausible**

â€¨

1. **Please select whether the following next sentences are plausible or not.:**

Narrative: Once she informed him that their marriage was over, jason would have no more marital rights. Later, she would decide where she was going and what she would do. For now, she needed to get him to agree to a divorce. Or did she even need his permission? Since she wasn't certain, she decided it was wise not to alienate him unnecessarily or anger him into refusing. But then, she shouldn't **beat about the bush** too long, either.
Meaning: To speak vaguely or euphemistically so as to avoid talking directly about an unpleasant or sensitive topic

a) the matter to be resolved quickly the case she needed a plan to resolve it
○ **1. plausible** ○ **2. not plausible**

a) She needed to make it clear what she wanted
○ **1. plausible** ○ **2. not plausible**

a) She needed to be as direct as possible
○ **1. plausible** ○ **2. not plausible**

a) It was best to deal with unpleasant things like this straight away and get it over with.
○ **1. plausible** ○ **2. not plausible**

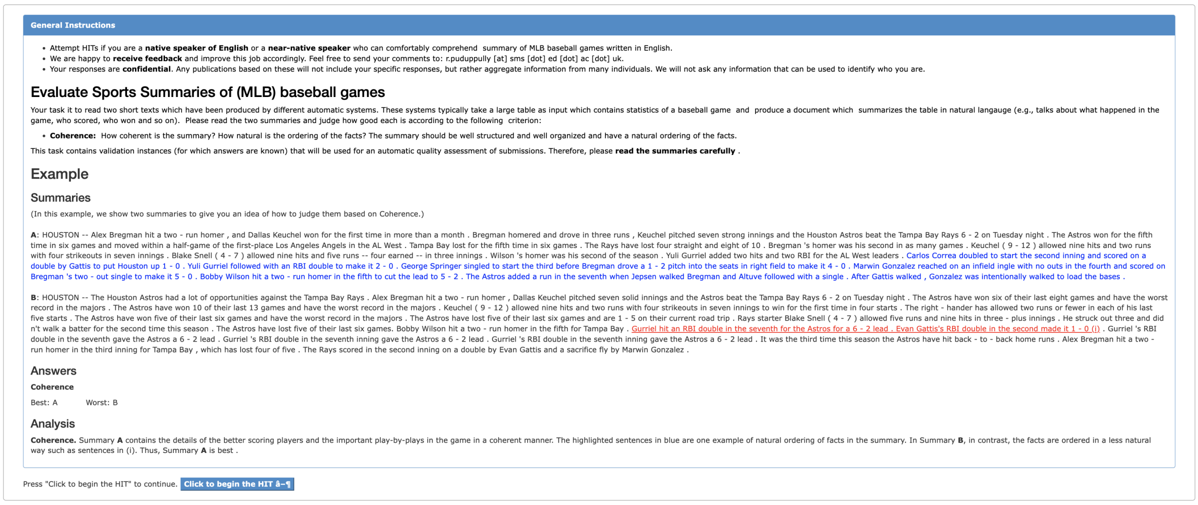
a) She decided that things needed to be dealt with immediately and came up with a plan to discuss the divorce with jason.
○ **1. plausible** ○ **2. not plausible**

a) Because he could realize it if she did not speak out.
○ **1. plausible** ○ **2. not plausible**

ATTENTION We have taken measures to prevent cheating and if you do not complete the task honestly we will know and the HIT will be rejected.

(Optional) Please provide any comments that you have about this HIT. Thanks for doing our HIT! We appreciate your input!

Figure 6: FL interface.



General Instructions

- Attempt HITs if you are a **native speaker of English** or a **near-native speaker** who can comfortably comprehend summary of MLB baseball games written in English.
- We are happy to **receive feedback** and improve this job accordingly. Feel free to send your comments to: r.puduppully [at] sms [dot] ed [dot] ac [dot] uk.
- Your responses are **confidential**. Any publications based on these will not include your specific responses, but rather aggregate information from many individuals. We will not ask any information that can be used to identify who you are.

Evaluate Sports Summaries of (MLB) baseball games

Your task it to read two short texts which have been produced by different automatic systems. These systems typically take a large table as input which contains statistics of a baseball game and produce a document which summarizes the table in natural langauge (e.g., talks about what happened in the game, who scored, who won and so on). Please read the two summaries and judge how good each is according to the following criterion:

- **Coherence:** How coherent is the summary? How natural is the ordering of the facts? The summary should be well structured and well organized and have a natural ordering of the facts.

This task contains validation instances (for which answers are known) that will be used for an automatic quality assessment of submissions. Therefore, please **read the summaries carefully** .

Example

Summaries

(In this example, we show two summaries to give you an idea of how to judge them based on Coherence.)

A: HOUSTON -- Alex Bregman hit a two - run homer , and Dallas Keuchel won for the first time in more than a month . Bregman homered and drove in three runs , Keuchel pitched seven strong innings and the Houston Astros beat the Tampa Bay Rays 6 - 2 on Tuesday night . The Astros won for the fifth time in six games and moved within a half-game of the first-place Los Angeles Angels in the AL West . Tampa Bay lost for the fifth time in six games . The Rays have lost four straight and eight of 10 . Bregman 's homer was his second in as many games . Keuchel (9 - 12) allowed nine hits and two runs with four strikeouts in seven innings . Blake Snell (4 - 7) allowed nine hits and five runs -- four earned -- in three innings . Wilson 's homer was his second of the season . Yuli Gurriel added two hits and two RBI for the AL West leaders . Carlos Correa doubled to start the second inning and scored on a double by Gattis to put Houston up 1 - 0 . Yuli Gurriel followed with an RBI double to make it 2 - 0 . George Springer singled to start the third before Bregman drove a 1 - 2 pitch into the seats in right field to make it 4 - 0 . Marwin Gonzalez reached on an infield ingle with no outs in the fourth and scored on Bregman 's two - out single to make it 5 - 0 . Bobby Wilson hit a two - run homer in the fifth to cut the lead to 5 - 2 . The Astros added a run in the seventh when Jepsen walked Bregman and Altuve followed with a single . After Gattis walked , Gonzalez was intentionally walked to load the bases .

B: HOUSTON -- The Houston Astros had a lot of opportunities against the Tampa Bay Rays . Alex Bregman hit a two - run homer , Dallas Keuchel pitched seven solid innings and the Astros beat the Tampa Bay Rays 6 - 2 on Tuesday night . The Astros have won six of their last eight games and have the worst record in the majors . The Astros have won 10 of their last 13 games and have the worst record in the majors . Keuchel (9 - 12) allowed nine hits and two runs with four strikeouts in seven innings to win for the first time in four starts . The right - hander has allowed two runs or fewer in each of his last five starts . The Astros have won five of their last six games and have the worst record in the majors . The Astros have lost five of their last six games and are 1 - 5 on their current road trip . Rays starter Blake Snell (4 - 7) allowed five runs and nine hits in three - plus innings . He struck out three and did n't walk a batter for the second time this season . The Astros have lost five of their last six games. Bobby Wilson hit a two - run homer in the fifth for Tampa Bay . Gurriel hit an RBI double in the seventh for the Astros for a 6 - 2 lead . Evan Gattis's RBI double in the second made it 1 - 0 (i) . Gurriel 's RBI double in the seventh gave the Astros a 6 - 2 lead . Gurriel 's RBI double in the seventh inning gave the Astros a 6 - 2 lead . Gurriel 's RBI double in the seventh inning gave the Astros a 6 - 2 lead . It was the third time this season the Astros have hit back - to - back home runs . Alex Bregman hit a two - run homer in the third inning for Tampa Bay , which has lost four of five . The Rays scored in the second inning on a double by Evan Gattis and a sacrifice fly by Marwin Gonzalez .

Answers

Coherence

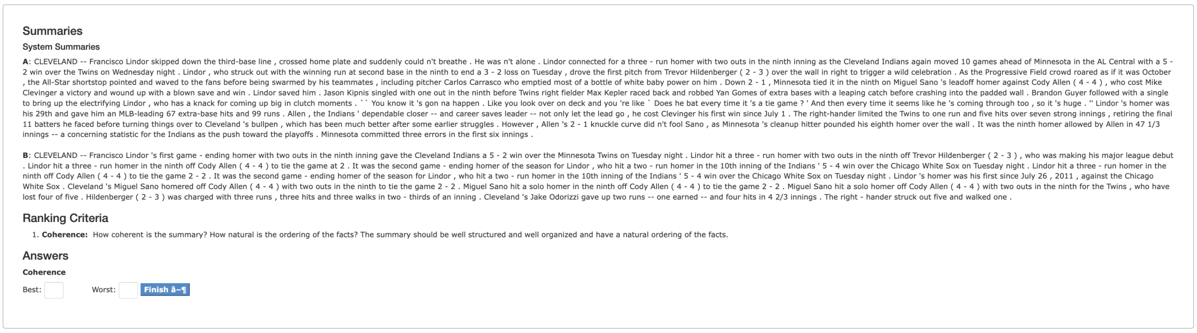
Best: A Worst: B

Analysis

Coherence. Summary **A** contains the details of the better scoring players and the important play-by-plays in the game in a coherent manner. The highlighted sentences in blue are one example of natural ordering of facts in the summary. In Summary **B**, in contrast, the facts are ordered in a less natural way such as sentences in (i). Thus, Summary **A** is best .

Press "Click to begin the HIT" to continue.

Figure 7: MLBC interface - instructions.



Summaries

System Summaries

A: CLEVELAND -- Francisco Lindor skipped down the third-base line , crossed home plate and suddenly could n't breathe . He was n't alone . Lindor connected for a three - run homer with two outs in the ninth inning as the Cleveland Indians again moved 10 games ahead of Minnesota in the AL Central with a 5 - 2 win over the Twins on Wednesday night . Lindor , who struck out with the winning run at second base in the ninth to end a 3 - 2 loss on Tuesday , drove the first pitch from Trevor Hildenberger (2 - 3) over the wall in right to trigger a wild celebration . As the Progressive Field crowd roared as if it was October , the All-Star shortstop pointed and waved to the fans before being swarmed by his teammates , including pitcher Carlos Carrasco who emptied most of a bottle of white baby power on him . Down 2 - 1 , Minnesota tied it in the ninth on Miguel Sano 's leadoff homer against Cody Allen (4 - 4) , who cost Mike Clevinger a victory and wound up with a blown save and win . Lindor saved him . Jason Kipnis singled with one out in the ninth before Twins right fielder Max Kepler raced back and robbed Yan Gomes of extra bases with a leaping catch before crashing into the padded wall . Brandon Guyer followed with a single to bring up the electrifying Lindor , who has a knack for coming up big in clutch moments . `` You know it 's gon na happen . Like you look over on deck and you 're like ` Does he bat every time it 's a tie game ? ' And then every time it seems like he 's coming through too , so it 's huge . '' Lindor 's homer was his 29th and gave him an MLB-leading 67 extra-base hits and 99 runs . Allen , the Indians ' dependable closer -- and career saves leader -- not only let the lead go , he cost Clevinger his first win since July 1 . The right-hander limited the Twins to one run and five hits over seven strong innings , retiring the final 11 batters he faced before turning things over to Cleveland 's bullpen , which has been much better after some earlier struggles . However , Allen 's 2 - 1 knuckle curve did n't fool Sano , as Minnesota 's cleanup hitter pounded his eighth homer over the wall . It was the ninth homer allowed by Allen in 47 1/3 innings -- a concerning statistic for the Indians as the push toward the playoffs . Minnesota committed three errors in the first six innings .

B: CLEVELAND -- Francisco Lindor 's first game - ending homer with two outs in the ninth inning gave the Cleveland Indians a 5 - 2 win over the Minnesota Twins on Tuesday night . Lindor hit a three - run homer with two outs in the ninth off Trevor Hildenberger (2 - 3) , who was making his major league debut . Lindor hit a three - run homer in the ninth off Cody Allen (4 - 4) to tie the game at 2 . It was the second game - ending homer of the season for Lindor , who hit a two - run homer in the 10th inning of the Indians ' 5 - 4 win over the Chicago White Sox on Tuesday night . Lindor hit a three - run homer in the ninth off Cody Allen (4 - 4) to tie the game 2 - 2 . It was the second game - ending homer of the season for Lindor , who hit a two - run homer in the 10th inning of the Indians ' 5 - 4 win over the Chicago White Sox on Tuesday night . Lindor 's homer was his first since July 26 , 2011 , against the Chicago White Sox . Cleveland 's Miguel Sano homered off Cody Allen (4 - 4) with two outs in the ninth to tie the game 2 - 2 . Miguel Sano hit a solo homer in the ninth off Cody Allen (4 - 4) to tie the game 2 - 2 . Miguel Sano hit a solo homer off Cody Allen (4 - 4) with two outs in the ninth for the Twins , who have lost four of five . Hildenberger (2 - 3) was charged with three runs , three hits and three walks in two - thirds of an inning . Cleveland 's Jake Odorizzi gave up two runs -- one earned -- and four hits in 4 2/3 innings . The right - hander struck out five and walked one .

Ranking Criteria

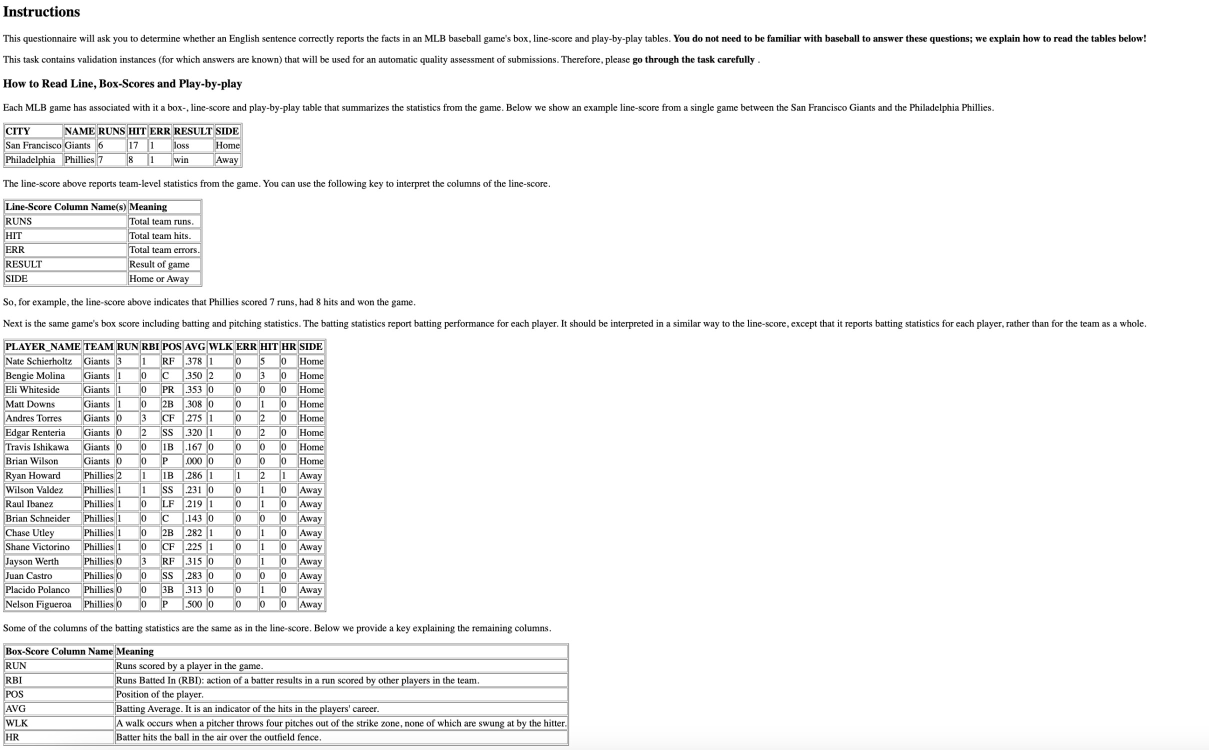
1. **Coherence:** How coherent is the summary? How natural is the ordering of the facts? The summary should be well structured and well organized and have a natural ordering of the facts.

Answers

Coherence

Best: □ Worst: □

Figure 8: MLBC interface - task.



Instructions

This questionnaire will ask you to determine whether an English sentence correctly reports the facts in an MLB baseball game's box, line-score and play-by-play tables. **You do not need to be familiar with baseball to answer these questions; we explain how to read the tables below!**

This task contains validation instances (for which answers are known) that will be used for an automatic quality assessment of submissions. Therefore, please **go through the task carefully** .

How to Read Line, Box-Scores and Play-by-play

Each MLB game has associated with it a box-, line-score and play-by-play table that summarizes the statistics from the game. Below we show an example line-score from a single game between the San Francisco Giants and the Philadelphia Phillies.

CITY	NAME	RUNS	HIT	ERR	RESULT	SIDE
San Francisco	Giants	6	17	1	loss	Home
Philadelphia	Phillies	7	8	1	win	Away

The line-score above reports team-level statistics from the game. You can use the following key to interpret the columns of the line-score.

Line-Score Column Name(s)	Meaning
RUNS	Total team runs.
HIT	Total team hits.
ERR	Total team errors.
RESULT	Result of game
SIDE	Home or Away

So, for example, the line-score above indicates that Phillies scored 7 runs, had 8 hits and won the game.

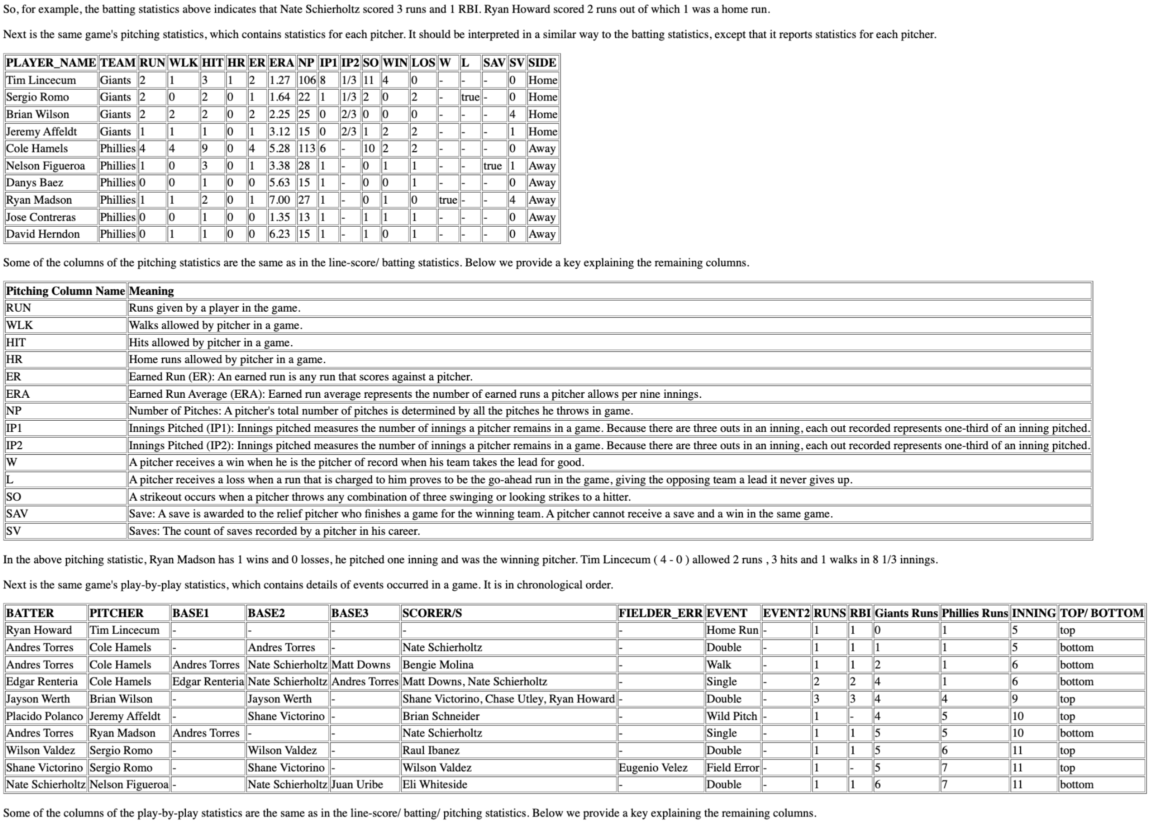
Next is the same game's box score including batting and pitching statistics. The batting statistics report batting performance for each player. It should be interpreted in a similar way to the line-score, except that it reports batting statistics for each player, rather than for the team as a whole.

PLAYER_NAME	TEAM	RUN	RBI	POS	AVG	WLK	ERR	HIT	HR	SIDE
Nate Schierholtz	Giants	3	1	RF	.378	1	0	5	0	Home
Bengie Molina	Giants	1	0	C	.350	2	0	3	0	Home
Eli Whiteside	Giants	1	0	PR	.353	0	0	0	0	Home
Matt Downs	Giants	1	0	2B	.308	0	0	1	0	Home
Andres Torres	Giants	0	3	CF	.275	1	0	2	0	Home
Edgar Renteria	Giants	0	2	SS	.320	1	0	2	0	Home
Travis Ishikawa	Giants	0	0	1B	.167	0	0	0	0	Home
Brian Wilson	Giants	0	0	P	.000	0	0	0	0	Home
Ryan Howard	Phillies	2	1	1B	.286	1	1	2	1	Away
Wilson Valdez	Phillies	1	1	SS	.231	0	0	1	0	Away
Raul Ibanez	Phillies	1	0	LF	.219	1	0	1	0	Away
Brian Schneider	Phillies	1	0	C	.143	0	0	0	0	Away
Chase Utley	Phillies	1	0	2B	.282	1	0	1	0	Away
Shane Victorino	Phillies	1	0	CF	.225	1	0	1	0	Away
Jayson Werth	Phillies	0	3	RF	.315	0	0	1	0	Away
Juan Castro	Phillies	0	0	SS	.283	0	0	0	0	Away
Placido Polanco	Phillies	0	0	3B	.313	0	0	1	0	Away
Nelson Figueroa	Phillies	0	0	P	.500	0	0	0	0	Away

Some of the columns of the batting statistics are the same as in the line-score. Below we provide a key explaining the remaining columns.

Box-Score Column Name	Meaning
RUN	Runs scored by a player in the game.
RBI	Runs Batted In (RBI): action of a batter results in a run scored by other players in the team.
POS	Position of the player.
AVG	Batting Average. It is an indicator of the hits in the players' career.
WLK	A walk occurs when a pitcher throws four pitches out of the strike zone, none of which are swung at by the hitter.
HR	Batter hits the ball in the air over the outfield fence.

Figure 9: MLBF interface - instructions (i).



So, for example, the batting statistics above indicates that Nate Schierholtz scored 3 runs and 1 RBI. Ryan Howard scored 2 runs out of which 1 was a home run.

Next is the same game's pitching statistics, which contains statistics for each pitcher. It should be interpreted in a similar way to the batting statistics, except that it reports statistics for each pitcher.

PLAYER_NAME	TEAM	RUN	WLK	HIT	HR	ER	ERA	NP	IP1	IP2	SO	WIN	LOS	W	L	SAV	SV	SIDE
Tim Lincecum	Giants	2	1	3	1	2	1.27	106	8	1/3	11	4	0	-	-	-	0	Home
Sergio Romo	Giants	2	0	2	0	1	1.64	22	1	1/3	2	0	2	-	true	-	0	Home
Brian Wilson	Giants	2	2	2	0	2	2.25	25	0	2/3	0	0	0	-	-	-	4	Home
Jeremy Affeldt	Giants	1	1	1	0	1	3.12	15	0	2/3	1	2	2	-	-	-	1	Home
Cole Hamels	Phillies	4	4	9	0	4	5.28	113	6	-	10	2	2	-	-	-	0	Away
Nelson Figueroa	Phillies	1	0	3	0	1	3.38	28	1	-	0	1	1	-	-	true	1	Away
Danys Baez	Phillies	0	0	1	0	0	5.63	15	1	-	0	0	1	-	-	-	0	Away
Ryan Madson	Phillies	1	1	2	0	1	7.00	27	1	-	0	1	0	true	-	-	4	Away
Jose Contreras	Phillies	0	0	1	0	0	1.35	13	1	-	1	1	1	-	-	-	0	Away
David Herndon	Phillies	0	1	1	0	0	6.23	15	1	-	1	0	1	-	-	-	0	Away

Some of the columns of the pitching statistics are the same as in the line-score/ batting statistics. Below we provide a key explaining the remaining columns.

Pitching Column Name	Meaning
RUN	Runs given by a player in the game.
WLK	Walks allowed by pitcher in a game.
HIT	Hits allowed by pitcher in a game.
HR	Home runs allowed by pitcher in a game.
ER	Earned Run (ER): An earned run is any run that scores against a pitcher.
ERA	Earned Run Average (ERA): Earned run average represents the number of earned runs a pitcher allows per nine innings.
NP	Number of Pitches: A pitcher's total number of pitches is determined by all the pitches he throws in game.
IP1	Innings Pitched (IP1): Innings pitched measures the number of innings a pitcher remains in a game. Because there are three outs in an inning, each out recorded represents one-third of an inning pitched.
IP2	Innings Pitched (IP2): Innings pitched measures the number of innings a pitcher remains in a game. Because there are three outs in an inning, each out recorded represents one-third of an inning pitched.
W	A pitcher receives a win when he is the pitcher of record when his team takes the lead for good.
L	A pitcher receives a loss when a run that is charged to him proves to be the go-ahead run in the game, giving the opposing team a lead it never gives up.
SO	A strikeout occurs when a pitcher throws any combination of three swinging or looking strikes to a hitter.
SAV	Save: A save is awarded to the relief pitcher who finishes a game for the winning team. A pitcher cannot receive a save and a win in the same game.
SV	Saves: The count of saves recorded by a pitcher in his career.

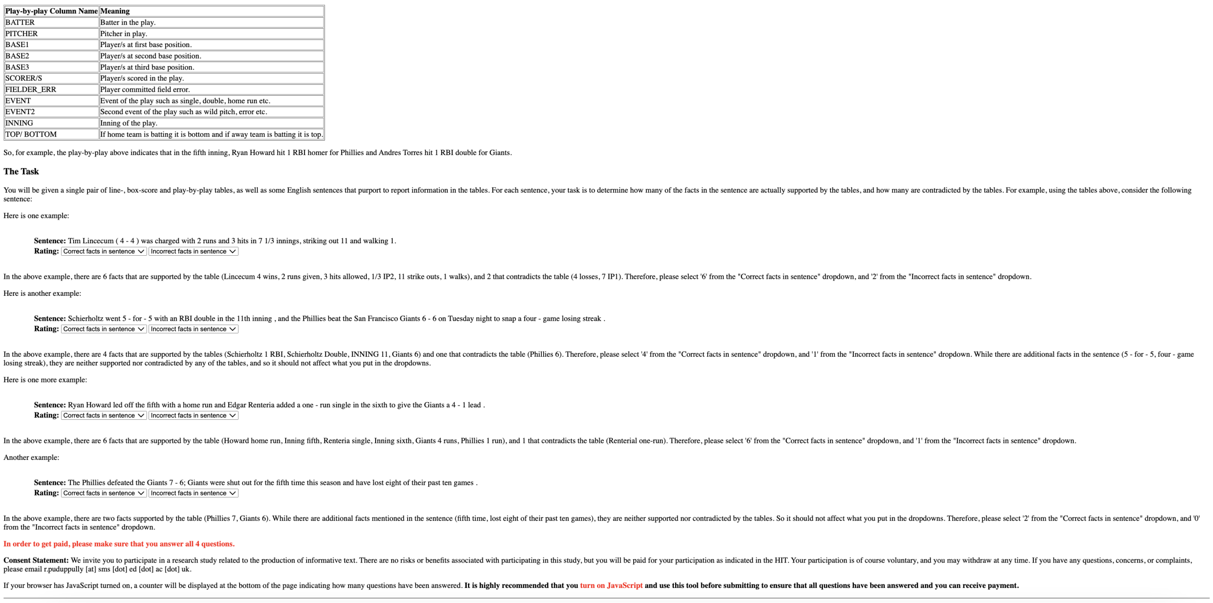
In the above pitching statistic, Ryan Madson has 1 wins and 0 losses, he pitched one inning and was the winning pitcher. Tim Lincecum (4 - 0) allowed 2 runs , 3 hits and 1 walks in 8 1/3 innings.

Next is the same game's play-by-play statistics, which contains details of events occurred in a game. It is in chronological order.

BATTER	PITCHER	BASE1	BASE2	BASE3	SCORER/S	FIELDER_ERR	EVENT	EVENT2	RUNS	RBI	Giants Runs	Phillies Runs	INNING	TOP/ BOTTOM
Ryan Howard	Tim Lincecum	-	-	-	-	-	Home Run	-	1	1	0	1	5	top
Andres Torres	Cole Hamels	-	Andres Torres	-	Nate Schierholtz	-	Double	-	1	1	1	1	5	bottom
Andres Torres	Cole Hamels	Andres Torres	Nate Schierholtz	Matt Downs	Bengie Molina	-	Walk	-	1	1	2	1	6	bottom
Edgar Renteria	Cole Hamels	Edgar Renteria	Nate Schierholtz	Andres Torres	Matt Downs, Nate Schierholtz	-	Single	-	2	2	4	1	6	bottom
Jayson Werth	Brian Wilson	-	Jayson Werth	-	Shane Victorino, Chase Utley, Ryan Howard	-	Double	-	3	3	4	4	9	top
Placido Polanco	Jeremy Affeldt	-	Shane Victorino	-	Brian Schneider	-	Wild Pitch	-	1	-	4	5	10	top
Andres Torres	Ryan Madson	Andres Torres	-	-	Nate Schierholtz	-	Single	-	1	1	5	5	10	bottom
Wilson Valdez	Sergio Romo	-	Wilson Valdez	-	Raul Ibanez	-	Double	-	1	1	5	6	11	top
Shane Victorino	Sergio Romo	-	Shane Victorino	-	Wilson Valdez	Eugenio Velez	Field Error	-	1	-	5	7	11	top
Nate Schierholtz	Nelson Figueroa	-	Nate Schierholtz	Juan Uribe	Eli Whiteside	-	Double	-	1	1	6	7	11	bottom

Some of the columns of the play-by-play statistics are the same as in the line-score/ batting/ pitching statistics. Below we provide a key explaining the remaining columns.

Figure 10: MLBF interface - instructions (ii).



Play-by-play Column Name	Meaning
BATTER	Batter in the play.
PITCHER	Pitcher in play.
BASE1	Player/s at first base position.
BASE2	Player/s at second base position.
BASE3	Player/s at third base position.
SCORER/S	Player/s scored in the play.
FIELDER_ERR	Player committed field error.
EVENT	Event of the play such as single, double, home run etc.
EVENT2	Second event of the play such as wild pitch, error etc.
INNING	Inning of the play.
TOP/ BOTTOM	If home team is batting it is bottom and if away team is batting it is top.

So, for example, the play-by-play above indicates that in the fifth inning, Ryan Howard hit 1 RBI homer for Phillies and Andres Torres hit 1 RBI double for Giants.

The Task

You will be given a single pair of line-, box-score and play-by-play tables, as well as some English sentences that purport to report information in the tables. For each sentence, your task is to determine how many of the facts in the sentence are actually supported by the tables, and how many are contradicted by the tables. For example, using the tables above, consider the following sentence:

Here is one example:

Sentence: Tim Lincecum (4 - 4) was charged with 2 runs and 3 hits in 7 1/3 innings, striking out 11 and walking 1.
Rating: Correct facts in sentence ⌄ Incorrect facts in sentence ⌄

In the above example, there are 6 facts that are supported by the table (Lincecum 4 wins, 2 runs given, 3 hits allowed, 1/3 IP2, 11 strike outs, 1 walks), and 2 that contradicts the table (4 losses, 7 IP1). Therefore, please select '6' from the "Correct facts in sentence" dropdown, and '2' from the "Incorrect facts in sentence" dropdown.

Here is another example:

Sentence: Schierholtz went 5 - for - 5 with an RBI double in the 11th inning , and the Phillies beat the San Francisco Giants 6 - 6 on Tuesday night to snap a four - game losing streak .
Rating: Correct facts in sentence ⌄ Incorrect facts in sentence ⌄

In the above example, there are 4 facts that are supported by the tables (Schierholtz 1 RBI, Schierholtz Double, INNING 11, Giants 6) and one that contradicts the table (Phillies 6). Therefore, please select '4' from the "Correct facts in sentence" dropdown, and '1' from the "Incorrect facts in sentence" dropdown. While there are additional facts in the sentence (5 - for - 5, four - game losing streak), they are neither supported nor contradicted by any of the tables, and so it should not affect what you put in the dropdowns.

Here is one more example:

Sentence: Ryan Howard led off the fifth with a home run and Edgar Renteria added a one - run single in the sixth to give the Giants a 4 - 1 lead .
Rating: Correct facts in sentence ⌄ Incorrect facts in sentence ⌄

In the above example, there are 6 facts that are supported by the table (Howard home run, Inning fifth, Renteria single, Inning sixth, Giants 4 runs, Phillies 1 run), and 1 that contradicts the table (Renterial one-run). Therefore, please select '6' from the "Correct facts in sentence" dropdown, and '1' from the "Incorrect facts in sentence" dropdown.

Another example:

Sentence: The Phillies defeated the Giants 7 - 6; Giants were shut out for the fifth time this season and have lost eight of their past ten games .
Rating: Correct facts in sentence ⌄ Incorrect facts in sentence ⌄

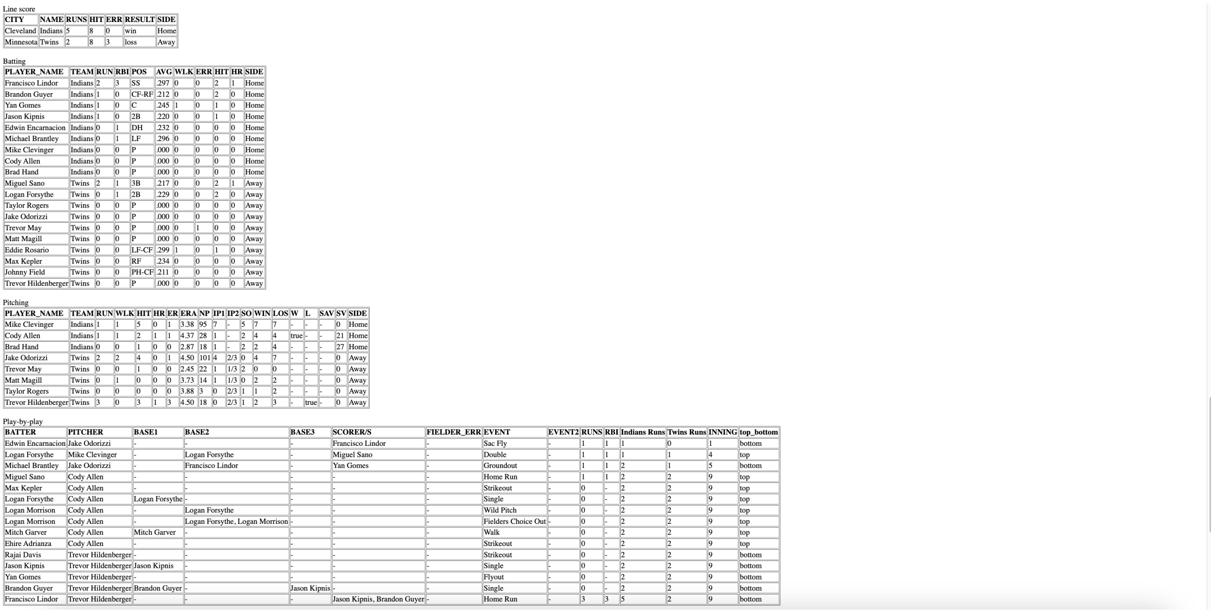
In the above example, there are two facts supported by the table (Phillies 7, Giants 6). While there are additional facts mentioned in the sentence (fifth time, lost eight of their past ten games), they are neither supported nor contradicted by the tables. So it should not affect what you put in the dropdowns. Therefore, please select '2' from the "Correct facts in sentence" dropdown, and '0' from the "Incorrect facts in sentence" dropdown.

In order to get paid, please make sure that you answer all 4 questions.

Consent Statement: We invite you to participate in a research study related to the production of informative text. There are no risks or benefits associated with participating in this study, but you will be paid for your participation as indicated in the HIT. Your participation is of course voluntary, and you may withdraw at any time. If you have any questions, concerns, or complaints, please email r.puduppully [at] sms [dot] ed [dot] ac [dot] uk.

If your browser has JavaScript turned on, a counter will be displayed at the bottom of the page indicating how many questions have been answered. **It is highly recommended that you turn on JavaScript and use this tool before submitting to ensure that all questions have been answered and you can receive payment.**

Figure 11: MLBF interface - instructions (iii).



Line score

CITY	NAME	RUNS	HIT	ERR	RESULT	SIDE
Cleveland	Indians	5	8	0	win	Home
Minnesota	Twins	2	8	3	loss	Away

Batting

PLAYER_NAME	TEAM	RUN	RBI	POS	AVG	WLK	ERR	HIT	HR	SIDE
Francisco Lindor	Indians	2	3	SS	.297	0	0	2	1	Home
Brandon Guyer	Indians	1	0	CF-RF	.212	0	0	2	0	Home
Yan Gomes	Indians	1	0	C	.245	1	0	1	0	Home
Jason Kipnis	Indians	1	0	2B	.220	0	0	1	0	Home
Edwin Encarnacion	Indians	0	1	DH	.232	0	0	0	0	Home
Michael Brantley	Indians	0	1	LF	.296	0	0	0	0	Home
Mike Clevinger	Indians	0	0	P	.000	0	0	0	0	Home
Cody Allen	Indians	0	0	P	.000	0	0	0	0	Home
Brad Hand	Indians	0	0	P	.000	0	0	0	0	Home
Miguel Sano	Twins	2	1	3B	.217	0	0	2	1	Away
Logan Forsythe	Twins	0	1	2B	.229	0	0	2	0	Away
Taylor Rogers	Twins	0	0	P	.000	0	0	0	0	Away
Jake Odorizzi	Twins	0	0	P	.000	0	0	0	0	Away
Trevor May	Twins	0	0	P	.000	0	1	0	0	Away
Matt Magill	Twins	0	0	P	.000	0	0	0	0	Away
Eddie Rosario	Twins	0	0	LF-CF	.299	1	0	1	0	Away
Max Kepler	Twins	0	0	RF	.234	0	0	0	0	Away
Johnny Field	Twins	0	0	PH-CF	.211	0	0	0	0	Away
Trevor Hildenberger	Twins	0	0	P	.000	0	0	0	0	Away

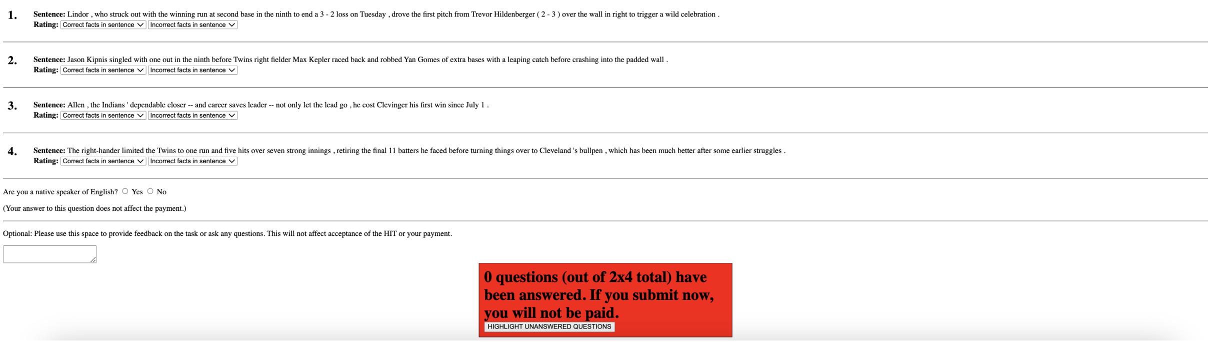
Pitching

PLAYER_NAME	TEAM	RUN	WLK	HIT	HR	ER	ERA	NP	IP1	IP2	SO	WIN	LOS	W	L	SAV	SV	SIDE
Mike Clevinger	Indians	1	1	5	0	1	3.38	95	7	-	5	7	7	-	-	-	0	Home
Cody Allen	Indians	1	1	2	1	1	4.37	28	1	-	2	4	4	true	-	-	21	Home
Brad Hand	Indians	0	0	1	0	0	2.87	18	1	-	2	2	4	-	-	-	27	Home
Jake Odorizzi	Twins	2	2	4	0	1	4.50	101	4	2/3	0	4	7	-	-	-	0	Away
Trevor May	Twins	0	0	1	0	0	2.45	22	1	1/3	2	0	0	-	-	-	0	Away
Matt Magill	Twins	0	1	0	0	0	3.73	14	1	1/3	0	2	2	-	-	-	0	Away
Taylor Rogers	Twins	0	0	0	0	0	3.88	3	0	2/3	1	1	2	-	-	-	0	Away
Trevor Hildenberger	Twins	3	0	3	1	3	4.50	18	0	2/3	1	2	3	-	true	-	0	Away

Play-by-play

BATTER	PITCHER	BASE1	BASE2	BASE3	SCORER/S	FIELDER_ERR	EVENT	EVENT2	RUNS	RBI	Indians Runs	Twins Runs	INNING	top_bottom
Edwin Encarnacion	Jake Odorizzi	-	-	-	Francisco Lindor	-	Sac Fly	-	1	1	1	0	1	bottom
Logan Forsythe	Mike Clevinger	-	Logan Forsythe	-	Miguel Sano	-	Double	-	1	1	1	1	4	top
Michael Brantley	Jake Odorizzi	-	Francisco Lindor	-	Yan Gomes	-	Groundout	-	1	1	2	1	5	bottom
Miguel Sano	Cody Allen	-	-	-	-	-	Home Run	-	1	1	2	2	9	top
Max Kepler	Cody Allen	-	-	-	-	-	Strikeout	-	0	-	2	2	9	top
Logan Forsythe	Cody Allen	Logan Forsythe	-	-	-	-	Single	-	0	-	2	2	9	top
Logan Morrison	Cody Allen	-	Logan Forsythe	-	-	-	Wild Pitch	-	0	-	2	2	9	top
Logan Morrison	Cody Allen	-	Logan Forsythe, Logan Morrison	-	-	-	Fielders Choice Out	-	0	-	2	2	9	top
Mitch Garver	Cody Allen	Mitch Garver	-	-	-	-	Walk	-	0	-	2	2	9	top
Ehire Adrianza	Cody Allen	-	-	-	-	-	Strikeout	-	0	-	2	2	9	top
Rajai Davis	Trevor Hildenberger	-	-	-	-	-	Strikeout	-	0	-	2	2	9	bottom
Jason Kipnis	Trevor Hildenberger	Jason Kipnis	-	-	-	-	Single	-	0	-	2	2	9	bottom
Yan Gomes	Trevor Hildenberger	-	-	-	-	-	Flyout	-	0	-	2	2	9	bottom
Brandon Guyer	Trevor Hildenberger	Brandon Guyer	-	Jason Kipnis	-	-	Single	-	0	-	2	2	9	bottom
Francisco Lindor	Trevor Hildenberger	-	-	-	Jason Kipnis, Brandon Guyer	-	Home Run	-	3	3	5	2	9	bottom

Figure 12: MLBF interface - task (i).



1. **Sentence:** Lindor , who struck out with the winning run at second base in the ninth to end a 3 - 2 loss on Tuesday , drove the first pitch from Trevor Hildenberger (2 - 3) over the wall in right to trigger a wild celebration .
Rating: Correct facts in sentence Incorrect facts in sentence

2. **Sentence:** Jason Kipnis singled with one out in the ninth before Twins right fielder Max Kepler raced back and robbed Yan Gomes of extra bases with a leaping catch before crashing into the padded wall .
Rating: Correct facts in sentence Incorrect facts in sentence

3. **Sentence:** Allen , the Indians ' dependable closer -- and career saves leader -- not only let the lead go , he cost Clevinger his first win since July 1 .
Rating: Correct facts in sentence Incorrect facts in sentence

4. **Sentence:** The right-hander limited the Twins to one run and five hits over seven strong innings , retiring the final 11 batters he faced before turning things over to Cleveland 's bullpen , which has been much better after some earlier struggles .
Rating: Correct facts in sentence Incorrect facts in sentence

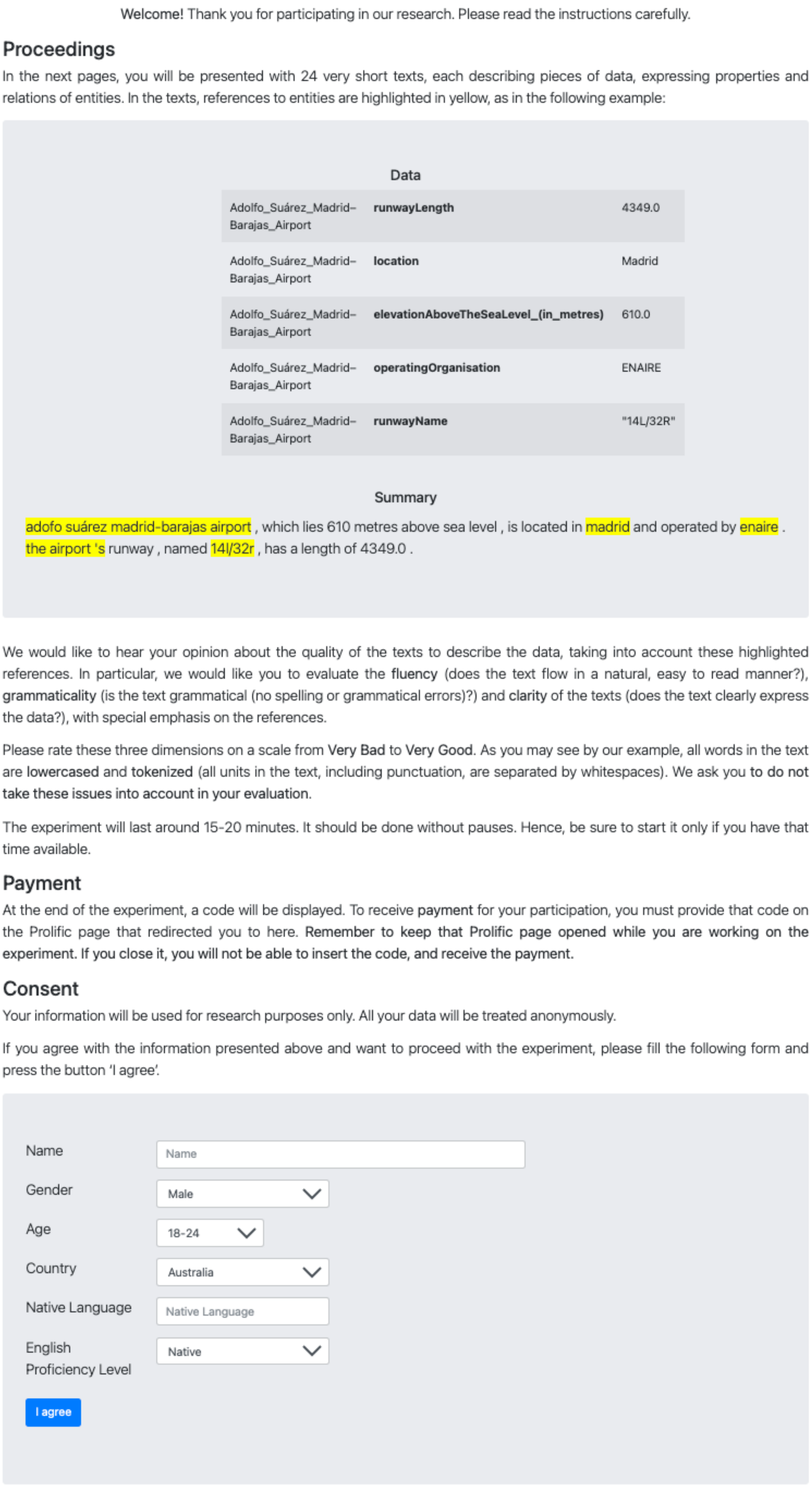
Are you a native speaker of English? ○ Yes ○ No

(Your answer to this question does not affect the payment.)

Optional: Please use this space to provide feedback on the task or ask any questions. This will not affect acceptance of the HIT or your payment.

0 questions (out of 2x4 total) have been answered. If you submit now, you will not be paid.

Figure 13: MLBF interface - task (ii).



Welcome! Thank you for participating in our research. Please read the instructions carefully.

Proceedings

In the next pages, you will be presented with 24 very short texts, each describing pieces of data, expressing properties and relations of entities. In the texts, references to entities are highlighted in yellow, as in the following example:

Data

Adolfo_Suárez_Madrid–Barajas_Airport	**runwayLength**	4349.0
Adolfo_Suárez_Madrid–Barajas_Airport	**location**	Madrid
Adolfo_Suárez_Madrid–Barajas_Airport	**elevationAboveTheSeaLevel_(in_metres)**	610.0
Adolfo_Suárez_Madrid–Barajas_Airport	**operatingOrganisation**	ENAIRE
Adolfo_Suárez_Madrid–Barajas_Airport	**runwayName**	"14L/32R"

Summary

adofo suárez madrid-barajas airport , which lies 610 metres above sea level , is located in madrid and operated by enaire . the airport 's runway , named 14l/32r , has a length of 4349.0 .

We would like to hear your opinion about the quality of the texts to describe the data, taking into account these highlighted references. In particular, we would like you to evaluate the **fluency** (does the text flow in a natural, easy to read manner?), **grammaticality** (is the text grammatical (no spelling or grammatical errors)?) and **clarity** of the texts (does the text clearly express the data?), with special emphasis on the references.

Please rate these three dimensions on a scale from **Very Bad** to **Very Good**. As you may see by our example, all words in the text are **lowercased** and **tokenized** (all units in the text, including punctuation, are separated by whitespaces). We ask you **to do not take these issues into account in your evaluation.**

The experiment will last around 15-20 minutes. It should be done without pauses. Hence, be sure to start it only if you have that time available.

Payment

At the end of the experiment, a code will be displayed. To receive **payment** for your participation, you must provide that code on the Prolific page that redirected you to here. **Remember to keep that Prolific page opened while you are working on the experiment. If you close it, you will not be able to insert the code, and receive the payment.**

Consent

Your information will be used for research purposes only. All your data will be treated anonymously.

If you agree with the information presented above and want to proceed with the experiment, please fill the following form and press the button 'I agree'.

Name: Name

Gender: Male

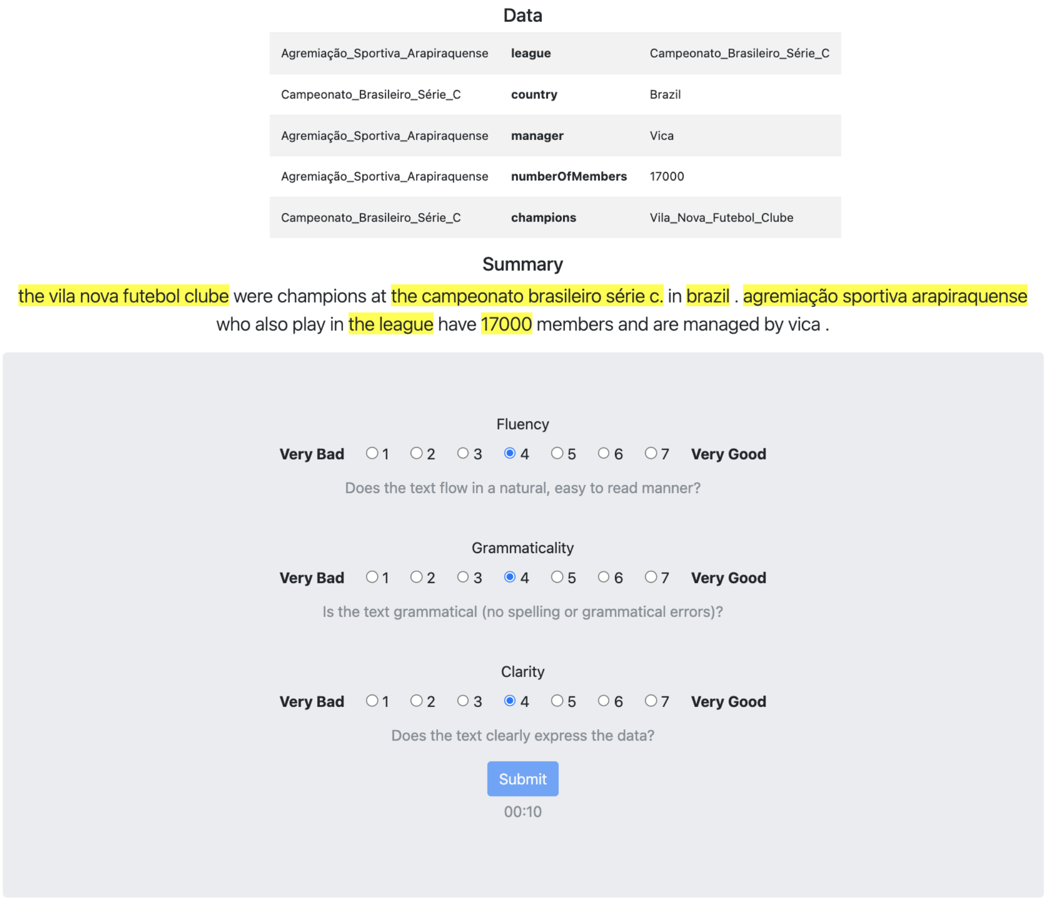
Age: 18-24

Country: Australia

Native Language: Native Language

English Proficiency Level: Native

Figure 14: REG interface - instructions.



Data

Agremiação_Sportiva_Arapiraquense	**league**	Campeonato_Brasileiro_Série_C
Campeonato_Brasileiro_Série_C	**country**	Brazil
Agremiação_Sportiva_Arapiraquense	**manager**	Vica
Agremiação_Sportiva_Arapiraquense	**numberOfMembers**	17000
Campeonato_Brasileiro_Série_C	**champions**	Vila_Nova_Futebol_Clube

Summary

the vila nova futebol clube were champions at the campeonato brasileiro série c. in brazil . agremiação sportiva arapiraquense who also play in the league have 17000 members and are managed by vica .

Fluency

Very Bad ○ 1 ○ 2 ○ 3 ● 4 ○ 5 ○ 6 ○ 7 **Very Good**

Does the text flow in a natural, easy to read manner?

Grammaticality

Very Bad ○ 1 ○ 2 ○ 3 ● 4 ○ 5 ○ 6 ○ 7 **Very Good**

Is the text grammatical (no spelling or grammatical errors)?

Clarity

Very Bad ○ 1 ○ 2 ○ 3 ● 4 ○ 5 ○ 6 ○ 7 **Very Good**

Does the text clearly express the data?

Figure 15: REG interface - task.