Developing a tool for fair and reproducible use of paid crowdsourcing in the digital humanities

Tuomo Hiippala and **Helmiina Hotti** and **Rosa Suviranta**
Department of Languages, University of Helsinki
Helsinki, Finland
{tuomo.hiippala, helmiina.hotti, rosa.suviranta}@helsinki.fi

Abstract

This system demonstration paper describes ongoing work on a tool for fair and reproducible use of paid crowdsourcing in the digital humanities. Paid crowdsourcing is widely used in natural language processing and computer vision, but has been rarely applied in the digital humanities due to ethical concerns. We discuss concerns associated with paid crowdsourcing and describe how we seek to mitigate them in designing the tool and crowdsourcing pipelines. We demonstrate how the tool may be used to create annotations for diagrams, a complex mode of expression whose description requires human input.

1 Introduction

Crowdsourcing is regularly used to create data for training and evaluating natural language processing and computer vision algorithms (Kovashka et al., 2016; Poesio et al., 2017). These fields often rely on paid crowdsourcing, which means that the work is distributed through online platforms and the performers are paid for their work. In the digital humanities, however, crowdsourcing is often associated with use of volunteers who are motivated by personal interests and altruism (Dunn and Hedges, 2013; Daugavietis, 2021). Conversely, paid crowdsourcing is viewed as ethically problematic (Terras, 2015) due to sweatshop wages (Fort et al., 2011) and other exploitative practices, such as invisible labour (Kummerfeld, 2021; Toxtli et al., 2021).

In this article, we present ongoing work on a tool for fair and reproducible use of paid crowdsourcing in the digital humanities. We argue that paid crowdsourcing offers a viable alternative to fields that fall under the umbrella of digital humanities, but are unlikely to attract a volunteer workforce. However, using paid crowdsourcing warrants attention to ethics. We demonstrate how ethical concerns may be addressed by incorporating mechanisms that discourage exploitative practices into the design of the tool and crowdsourcing pipelines.

2 Ethical issues related to crowdsourcing

As a portmanteau of *crowd* and *outsourcing*, the term crowdsourcing inherently evokes ideas of exploiting cheap labour in the global economy (Schmidt, 2013, p. 531). Paid crowdsourcing typically involves *requesters*, who post *tasks* on an online *platform*, which then distributes the tasks to *workers*. The platform thus acts as a mediator between the requesters and workers, and charges a commission from the requester. In natural language processing, crowdsourcing has become an established way of creating corpora due to the availability of a large pool of workers, short turnaround time and perceived cost efficiency (Fort et al., 2011).

Digital humanities have been cautious of paid crowdsourcing due to ethical issues related to low wages and workers' rights (Terras, 2015). Similar concerns have also been voiced in the field of natural language processing (Fort et al., 2011), which are increasingly supported by empirical evidence. Hara et al. (2018) show that only 4% of the workers on Amazon Mechanical Turk earn more than the federal minimum wage in the United States ($7.25 per hour). The average wage paid by the requesters amounts to $11.58 per hour, but requesters who pay less than the minimal wage outnumber those who pay fair wages (Hara et al., 2018, p. 7).

In addition to fair pay, recent research has highlighted issues arising from qualification labour, which refers to low- or non-paid work that workers must perform to qualify for tasks that pay more (Kummerfeld, 2021). Qualification labour emerges as a result of an information asymmetry between the requesters and workers. The requesters want to recruit high-performing workers by paying more, but higher wages also attract spammers who do not take the work seriously. Because the requesters cannot assess the quality of work in advance, they

Proceedings of the 6th LaTeCH-CLfL Workshop, pages 7–12
October 16, 2022. ©2022 International Conference on Computational Linguistics

are inclined to pay less, which drives away high-performing workers (Fort et al., 2011, p. 418). Making tasks only available to highly-qualified workers mitigates this problem, but to qualify for these tasks, the workers must perform approximately two months worth of non- or low-paying work (Kummerfeld, 2021).

Other forms of invisible labour on crowdsourcing platforms include the time spent searching for tasks, interacting with requesters and managing payments. Toxtli et al. (2021, p. 319) estimate that the median time spent on invisible labour accounts for 33% of active working time on crowdsourcing platforms. Because the workers are not compensated for this effort, invisible labour drives down their hourly wage. Additional forms of invisible labour include working on tasks that are rejected or expire, that is, the worker cannot complete the tasks within the timeframe set by the requester.

3 Crowdsourcing in digital humanities

Given the issues described above, it is not surprising that crowdsourcing in the digital humanities has mainly relied on volunteers who are motivated by personal interests and altruism (Dunn and Hedges, 2013; Daugavietis, 2021). Successful examples of volunteer-based crowdsourcing include platforms such as Zooniverse[1] and the *Transcribe Bentham* project (Causer et al., 2018), which have been able to attract a large body of motivated volunteers. This form of crowdsourcing in the digital humanities can also be conceptualised as a form of citizen science and peer production (Van Hyning, 2019).

However, some fields of study in the humanities may not be able to attract a sufficiently large body of volunteers. One such example is the emerging discipline of multimodality research, which studies how human communication relies on intentional combinations of expressive resources (see e.g. Bateman et al., 2017; Wildfeuer et al., 2020). As an emerging discipline, multimodality research is not widely known among the public at large, and its objects of study – everyday communicative situations and artefacts – are arguably less likely to attract the kind of attention needed for recruiting volunteers.

Multimodality research is currently undergoing a turn towards empirical research, which has been accompanied by calls for creating larger corpora to support this effort (Parodi, 2010; Thomas, 2014). Current multimodal corpora remain small, because creating multiple layers of cross-referenced annotations needed to capture multimodal phenomena requires time and resources (Bateman, 2014). Hiippala et al. (2021) have recently argued that the size of multimodal corpora can be increased by combining crowdsourced and expert annotations.

As researchers working in the field of multimodality research, our motivation to develop a tool for fair and reliable use of paid crowdsourcing arises from the prospect of building large multimodal corpora with multiple layers of rich annotation. At the same time, we acknowledge the ethical dimensions of using paid crowdsourcing and seek to address them in the design and use of the tool.

4 System design

4.1 Guiding principles for tool design and use

To mitigate the issues described in Section 2, we identify the following desiderata for developing and using the tool. First of all, the tool encourages the requesters to pay a fair wage to the workers (Fort et al., 2011; Hara et al., 2018). To do so, the tool asks requesters to estimate the time spent on a single task, which is used to calculate a task price that ensures that the workers are paid at least $12 per hour. We also encourage including explicit payment information in the instructions to reduce invisible labour related to wages (Toxtli et al., 2021).

To reduce invisible labour resulting from rejected or expired tasks, we emphasise the need for clear instructions and sufficient time to perform the tasks. Because crowdsourcing platforms attract a global workforce (Pavlick et al., 2014), we recommend the use of multimodal instructions that combine written language and visualisations to support workers who speak English as a foreign language. We also encourage the requesters to be transparent about their identity (Adda et al., 2013) and the purposes of their research to enable the workers make moral judgements about their willingness to participate (Schmidt, 2013). To reduce invisible labour from rejected work, we propose paying for work that contains human errors – e.g. a missing bounding box in an image segmentation task – and re-submitting these images for corrections.

To avoid hidden qualification work, we encourage using a combination of pedagogically-motivated training and paid examinations to train a workforce on the platform instead of using high-performing workers only (Kummerfeld, 2021). Pedagogically-motivated training refers to training

[1] https://zooniverse.org

tasks that teach the workers to perform the task. If the worker makes an error, they are provided with the correct answer and an explanation. The workers are later shown a similar task to assess their learning. Workers who pass the training are allowed to take a paid examination, which measures their performance. Those who pass the examination can then access to the actual tasks.

Implementing these desiderata into the tool design requires a modular structure, which allows constructing complex pipelines in a flexible manner, while simultaneously configuring the properties of individual tasks and their associated instructions and training data.

4.2 Technical description and architecture

The tool is written in Python 3.9 and designed for the Toloka[2] crowdsourcing platform, which has a well-documented and extensive API. Toloka also maintains a Python library for accessing the API, which we use for interacting with the platform.[3] The source code for the tool, which may be installed via the Python Package Index (PyPI), is available at: `https://github.com/thiippal/abulafia`

The architecture of the tool is based on three types of objects: tasks, actions and task sequences. Tasks allow creating individual crowdsourcing tasks and configuring payments, input/output data, quality control mechanisms and user interface. Actions, in turn, are used to manipulate the input/output data. These actions may include, for example, aggregating responses from multiple workers. Our tool implements the aggregation algorithms available in the Crowd-kit library for Python (Ustalov et al., 2021). To support reproducibility, both tasks and actions are configured using separate files that use the YAML markup language. The YAML configuration files are used for instantiating Python objects, which may be combined into task sequences to define and execute complex crowdsourcing pipelines.

5 System demonstration

In this section, we demonstrate how our tool can be used to crowdsource descriptions for a complex mode of communication, namely diagrams. Diagrams combine diverse expressive resources, such as natural language, photographs, illustrations, drawings, lines and arrows into a common discourse organisation (Hiippala and Bateman, 2022). Computational processing of diagrams is challenging, because their constituent parts are not fixed, but determined dynamically by the communicative goals set for the diagram (Hiippala et al., 2021). To exemplify, Figure 1 shows a diagram that uses written language and lines to pick out parts of an illustration, but we cannot know how the illustration should be decomposed without first considering the diagram as a whole, as the written labels determine how the depicted object should be decomposed into its constituent parts.

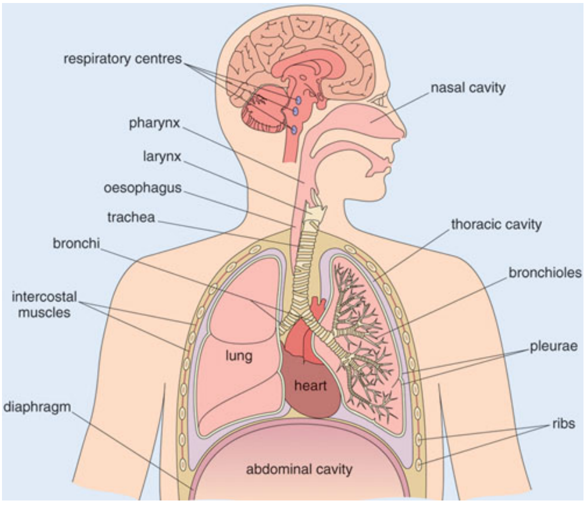

Figure 1: A primary school science diagram

To demonstrate how our tool may be used to decompose diagrams into analytical units, we define a pipeline with four steps. The pipeline aims to identify written labels and the parts they describe (see Figure 2). Each step consists of multiple tasks and actions. We first establish whether the diagram contains text (1), before asking the workers to outline all instances of written text (2). Next, we ask the workers to determine whether text elements refer to other parts of the diagram (3). Finally, we request the workers to outline the part(s) of the diagram referred to by the text (4).

Essentially, steps 1 and 3 consist of binary classification tasks (yes/no) in which agreement between the three workers is evaluated computationally. Steps 2 and 4, in turn, combine human verification with computational evaluation of agreement on the final decision between three workers (accept/reject).

As Figure 2 shows, each step combines a training with a paid examination, which is used to recruit the workforce needed for completing the step. Workers

[2] https://toloka.ai
[3] https://github.com/Toloka/toloka-kit

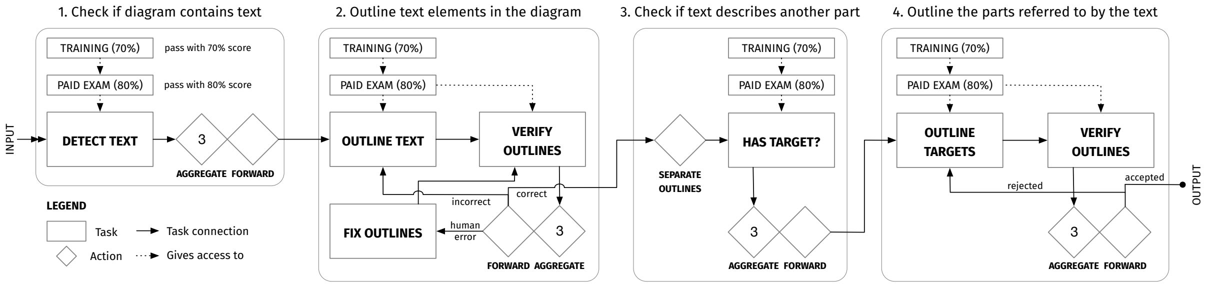

Figure 2: A crowdsourcing pipeline with four steps. See the legend in the lower left-hand corner for details.

Step	Task	Assignments	Total cost	% Re-annotated	Workers	Time
1	Detect text	300	$3.90	–	11	13 min
2	Outline text	103	$23.04	6.80%	32	38 min
2	Verify outlines	312	$44.46	–	42	34 min
2	Fix outlines	1	$0.38	–	1	2 min
3	Has target?	2980	$100.98	–	11	1 h 40 min
4	Outline target	1004	$255.90	14.94%	9	7 h 28 min
4	Verify outlines	3747	$184.07	–	64	4 h 26 min
	Total	8290	$612.73	1.86%	170	15 h 2 min

Table 1: Tasks, assignments, total cost, percentage of re-annotated assignments, number of workers and time spent

who pass the examination are also granted a skill that allows them to access similar tasks in the future. In each step, the AGGREGATE actions use the Dawid-Skene algorithm implemented in the Crowdkit library (Ustalov et al., 2021) to determine the most likely answer based on three responses from the workers. The FORWARD actions, in turn, determine where each assignment should be sent based on the result. Individual assignments are forwarded immediately upon completion.

We used 100 diagrams from the AI2D-RST corpus (Hiippala et al., 2021) as input to the pipeline. The pipeline and its configuration files can be found at: https://github.com/thiippal/latech-clfl-2022. We aimed to train at least 10 workers to perform each of steps 1–2 and 50 workers for each of steps 3–4 using paid examinations. The workers could take a paid examination if they passed the training with a 70% score. A score of 80% in the paid examination would grant access to the actual tasks. The total cost for paid examinations amounted to $183.71. This amount is excluded from the expenses in Table 1, which provides details on each task in the pipeline. The cost and time needed for training the workers depended largely on task type. Finally, we estimated the time needed to complete each assignment, and set the wage to $12 per hour.

6 Results and discussion

Based on the results of step 1, 96 out of 100 diagrams contained text elements. These 96 diagrams contained a total of 996 text elements, which were outlined and verified in step 2. 733 of these elements were classified as referring to another part of the diagram in step 3. Their targets were also outlined and verified, which yielded 784 annotations for non-textual elements in step 4.

Table 1 shows how crowdsourcing costs and time increase as the tasks become more demanding and the level of detail in the annotation increases. Whereas the tasks in steps 1–2 are fairly simple and describe entire diagrams, task complexity increases considerably for steps 3–4, because they target specific parts of the diagrams and require reasoning about their content and structure. This also increases the number of tasks needed for evaluating agreement between the workers, which is necessary for ensuring annotation quality.

Figure 3 shows example outputs from step 4. Whereas annotations for the diagram on the left are complete, annotations for the diagram on the right show considerable variation. In the right-hand diagram, stages 3, 5 and 6 feature rectangular bounding boxes which indicate that the numbers below refer to the text and illustration above. Zooming in on other stages shows that their outlines are drawn

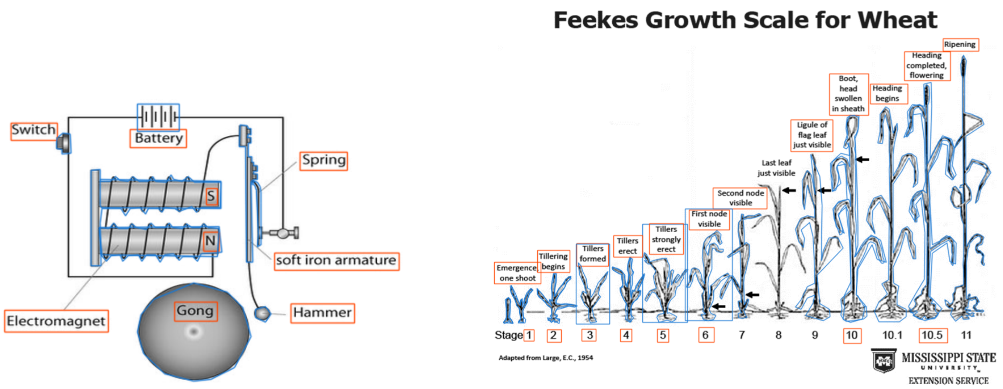

Figure 3: Two diagrams with crowdsourced annotations from step 4 of the pipeline. The diagrams have been converted into grayscale to highlight the annotations. The red bounding boxes indicate textual elements, whereas the blue boxes are used for the elements that the texts refer to. Note that bounding boxes for all text elements identified in step 2 are not visualised for the diagram on the right.

twice, as the workers have associated both written labels (above) and numbers (below) with the illustration. This shows how multiple workers who work on the same diagram make different inferences about the task and the diagram itself.

Furthermore, the annotations for stage 8 are missing altogether. This results from a false decision in step 3 of the crowdsourcing pipeline. Because three workers agreed that these written elements do not describe other elements in the diagram, they were not forwarded to step 4. These missing annotations could be created by adding a final verification step to the pipeline, which asks the workers to evaluate the completeness of the annotations.

Overall, the results suggest that paid crowdsourcing holds much potential for the digital humanities. As Table 1 showed, crowdsourced workers can create a large number of annotations in a relative short time. However, one must also account for the time needed for designing the pipeline, training materials and paid examinations, which are needed for ensuring quality results. In short, developing crowdsourcing pipelines is an iterative process of trial and error.

Our results may also be used to estimate the cost of creating similar annotations for all 1000 diagrams in the AI2D-RST corpus (Hiippala et al., 2021). Note, however, that the descriptions created above are partial, as they only target elements that consist of written text and the parts that they describe. Decomposing entire diagrams into analytical units by targeting other expressive resources such as arrows and lines would increase the costs considerably. In short, paid crowdsourcing is not cheap if used in an ethically responsible manner, but can be used to produce descriptions needed for building multimodal corpora.

Finally, researchers are responsible for applying paid crowdsourcing in a fair and ethical manner, which emphasises the need for transparency in relation to how crowdsourcing is used in academic research. However, not all issues outlined in Section 2 may be addressed by the requesters, as the platforms are ultimately responsible for designing the algorithms that distribute work and constrain the actions that workers and requesters can take. These are concerns that the research community should address together, as paid crowdsourcing has become a part of the research infrastructure in data-driven fields and beyond (cf. Fort et al., 2011).

7 Conclusion

In this article, we introduced a new tool for fair and reproducible use of paid crowdsourcing in the digital humanities. We showed how ethical issues associated with paid crowdsourcing can be mitigated by emphasising them in (1) tool development and (2) crowdsourcing pipeline design. We also demonstrated how the tool can be used to crowdsource descriptions of complex multimodal data. We conclude that paid crowdsourcing can be applied productively in the digital humanities, but its use warrants attention to ethical concerns at all stages of the process.

Acknowledgements

The development of the tool was supported by a $500 data grant from Toloka and by the Helsinki

Institute of Social Sciences and Humanities.

References

Gilles Adda, Joseph J. Mariani, Laurent Besacier, and Hadrien Gelas. 2013. Economic and ethical background of crowdsourcing for speech. In Maxine Eskenazi, Gina-Anne Levow, Helen Meng, Gabriel Parent, and David Suendermann, editors, *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*, pages 303–334. Wiley.

John A. Bateman. 2014. Using multimodal corpora for empirical research. In Carey Jewitt, editor, *The Routledge Handbook of Multimodal Analysis*, second edition, pages 238–252. Routledge, London and New York.

John A. Bateman, Janina Wildfeuer, and Tuomo Hiippala. 2017. *Multimodality: Foundations, Research and Analysis – A Problem-Oriented Introduction*. De Gruyter Mouton, Berlin.

Tim Causer, Kris Grint, Anna-Maria Sichani, and Melissa Terras. 2018. 'Making such bargain': Transcribe Bentham and the quality and cost-effectiveness of crowdsourced transcription. *Digital Scholarship in the Humanities*, 33(3):467–487.

Jãnis Daugavietis. 2021. Motivation to engage in crowdsourcing: Towards the synthetic psychological–sociological model. *Digital Scholarship in the Humanities*, 36(4):858–870.

Stuart Dunn and Mark Hedges. 2013. Crowd-sourcing as a component of humanities research infrastructures. *International Journal of Humanities and Arts Computing*, 7(1-2):147–169.

Karën Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. Association for Computing Machinery.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A. Bateman. 2021. AI2D-RST: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55(3):661–688.

Tuomo Hiippala and John A. Bateman. 2022. Introducing the diagrammatic semiotic mode. In *Diagrammatic Representation and Inference: 13th International Conference (Diagrams 2022)*, volume 13462 of *Lecture Notes in Computer Science*, Cham. Springer.

Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, and Kristen Grauman. 2016. Crowdsourcing in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 10(3):177–243.

Jonathan K. Kummerfeld. 2021. Quantifying and avoiding unfair qualification labour in crowdsourcing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–349, Online. Association for Computational Linguistics.

Giovanni Parodi. 2010. Research challenges for corpus cross-linguistics and multimodal texts. *Information Design Journal*, 18(1):69–73.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Massimo Poesio, Jon Chamberlain, and Udo Kruschwitz. 2017. Crowdsourcing. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 277–295. Springer, Dordrecht.

Florian A. Schmidt. 2013. The good, the bad and the ugly: Why crowdsourcing needs ethics. In *Proceedings of the 2013 International Conference on Cloud and Green Computing*, pages 531–535.

Melissa Terras. 2015. Crowdsourcing in the digital humanities. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A New Companion to Digital Humanities*, pages 420–438. Wiley, Oxford.

Martin Thomas. 2014. Evidence and circularity in multimodal discourse analysis. *Visual Communication*, 13(2):163–189.

Carlos Toxtli, Siddharth Suri, and Saiph Savage. 2021. Quantifying the invisible labor in crowd work. *Proceedings of the ACM on Human-Computer Interaction*, 5(319).

Dmitry Ustalov, Nikita Pavlichenko, Vladimir Losev, Iulian Giliazev, and Evgeny Tulin. 2021. A general-purpose crowdsourcing computational quality control toolkit for Python. In *The Ninth AAAI Conference on Human Computation and Crowdsourcing: Works-in-Progress and Demonstration Track*.

Victoria Van Hyning. 2019. Harnessing crowdsourcing for scholarly and GLAM purposes. *Literature Compass*, 16:e12507.

Janina Wildfeuer, Jana Pflaeging, John A. Bateman, Ognyan Seizov, and Chiao-I Tseng, editors. 2020. *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*. De Gruyter, Berlin, Munich and Boston.