Crowd Post-Correction of HTR Output in a Pedagogical Context: The Case of the Paris Bible Project 'Correct-a-thon'

Estelle Guéville, Yale University
David Joseph Wrisley, NYU Abu Dhabi

Crowd transcription has caught on in recent years and has been bolstered by the design of different interfaces to match the objectives of content co-creation (Hedges & Dunn 2018). Examples of crowdsourced transcription outreach in the Anglophone world include Transcribe Bentham!, the Smithsonian Transcription Center, and Zooniverse, each focusing on activating large cultural collections. The rise of customizable, cloud-based crowd-transcription platforms such as From the Page has led to a diversity of projects in the GLAM sector, as well as to expert, or semi-expert, academic crowdsourcing events of limited duration, involving "the work of skilled volunteers and collaborators transcribing documents and even creating metadata for materials” (Bastida, 2022). In an overview of crowd-transcription projects from 2011 to 2020, Brumfield discusses the “cross-pollination” of crowd tasks involving archival sources, as well as the emergence of transcription for engaging discipline-specific, expert communities. He also notes an “emerging consensus on 'free labor' and ethics,” recognising that “crowdsourcing projects cannot succeed without guidance, support, and intervention by institutional staff" (Brumfield, 2020).

In January 2023, our own transcription event took place in Besançon, France. It grew out of the Paris Bible Project [parisbible.github.io] which aims to understand the production and diffusion of the mass-produced biblical manuscripts in medieval Europe using digital methods. Inspired by two events organized for medievalists to transcribe multiple witnesses of the same medieval textual tradition (Morreale, 2020 and 2021), ours was a week-long event developed in collaboration with master’s students at the Université de Franche-Comté. It was not designed as a kick-off event for an institutional transcription project, but aimed to model forms of collaborative research-based inquiry in collaboration with HTR (Healey, 2005).

In our paper, we argue that since there is a large variety of expert and semi-expert communities interested in the future of the "transcribe-a-thon," we need to design such events both for these communities as well as for the potential reuse of their data. We also argue that, while such community events are often sponsored by research-led projects, they should not only be held to improve the quality of data, but they should also be framed as opportunities to learn elements of data literacy, particularly when the "academic crowd" is made up of students.

Brumfield suggested that the rise of artificial intelligence would change the crowdsourced transcription practices of projects in the 2020s (Brumfield, 2020). His observation has most definitely been borne out in the case of HTR platforms such as Transkribus, e-Scriptorium or Calfa, since the creation of ground truth for specialized HTR models requires a significant amount of expert human labor. The National Archives of Finland, for example, had a team of volunteers create ground truth of court records from different jurisdictions in Finland (Kallio, 2017), with the objective of training a model that could transcribe the 600,000 images from the collection. Other events focus instead on crowd post-correction of HTR output (HTREC, 2022). We even called our own event a "correct-a-thon," by which we meant that the focus of the event was on thinking about how HTR systems can (and cannot) transcribe the full complexity of writing systems such as medieval manuscripts. Other projects carried out in the classroom have been developed with a similar critical perspective in mind (Schlagdenhauffen, 2020), but they have benefited from greater contact hours.

Academic crowdsourcing depends on a set of skills and expert knowledge (Ridge, 2020). As such, it can be a demanding endeavor both for those leading and those participating, especially when events are short in duration. In our case, we aimed to produce data that would be reusable, but guidance and careful design were necessary elements for that to be the case. We also wanted our initiative to provide an opportunity for experiential learning, grounded in an ethical digital pedagogy and values-based teaching, in which we discussed with the students the biases of computer vision for paleography. It was important to us that the correct-a-thon provide participants with specific, transferable digital skills and more generally, an opportunity to learn more about codicology, palaeography and the materiality of manuscripts. With this content-specific focus, we also wanted our participants to learn about research processes and how they might integrate digital tools into their own research while working together in a team. Furthermore, as a crowdsourcing event, it was important that the participants were duly acknowledged for their contributions, "balancing credit and privacy" (Brumfield, 2020). We chose two specific forms of academic recognition: the publication of results on a scholarly blog as well as a collective publication of ground truth, indexed in HTR United, providing students as well with an opt-out from inclusion of their names in the project if they so desired (Chagué *et al*, 2021; Romein *et al*, 2022).

In order to be sure that these educational goals were achieved, it is important to remember that any such "transcribe-a-thon" or "correct-a-thon" is based on a specific workflow which must be tailored to the particular participants (Dariah Campus 2020). In our paper, we identify some of the most important ways that an event of limited duration can be designed as such. In particular, the correct-a-thon did not focus on all aspects of the set-up of the HTR process, but omitted (admittedly important) questions of document ingestion, layout analysis, and model training. Instead, we focused on post-correct using specific transcription norms, on variation in a single type of script and on model retraining (Guéville & Wrisley, 2020). As such, rather than being only an exercise in ground truth creation, the event led to a much deeper set of conversations about the challenges of HTR with medieval manuscripts and how data creation is intimately linked to research questions. This meant, of course, that there was a tradeoff between progress on the transcription and the scholarly engagement of the master's students. This "slower" progress on the data creation may be disappointing to some, but we feel that projects involving academic crowdsourcing, must ethically and pedagogically, focus on the communities which participate in them.

References

Bastida, A. (2022). “FromThePage vs Transkribus.” April 26, 2022. https://content.fromthepage.com/fromthepage-vs-transkribus/.

Brumfield, B. (2020). ”The Decade in Crowdsourcing Transcription”. January 9, 2020 https://content.fromthepage.com/decade-in-crowdsourcing/.

Chagué, A., Clérice, T., Romary, L. (2021). HTR-United : Mutualisons la vérité de terrain !. *DHNord2021 - Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*, MESHS, Nov 2021, Lille, France. https://hal.science/hal-03398740.

Dariah Campus. (2019). "Sharing the Experience: Workflows for the Digital Humanities" https://campus.dariah.eu/event/sharing-the-experience-workflows-for-the-digital-humanities.

Guéville, E., and Wrisley, D.J.. (2020). "Rethinking the Abbreviation: Questions and Challenges of Machine Reading Medieval Scripta.” *Dark Archives 20/20*, 9 September 2020.

Healey, M. (2005). "Linking Research and Teaching: Exploring Disciplinary Spaces and the Role of Inquiry-Based Learning. In : Barnett, R. (ed) Reshaping the University: New Relationships Between Research, Scholarship and Teaching. Open University Press, Berkshire/New York, p. 67-78.

Hedges, M. and Dunn, S. (2018). Academic Crowdsourcing in the Humanities. Elsevier. https://doi.org/10.1016/C2015-0-04363-5.

Kallio, M. (2017). “Deliverable 8.8, layout analysis and crowdsourcing”, Available at: https://read.transkribus.eu/wp-content/uploads/2017/12/Deliverable_8.8.pdf

Morreale, L. (2020). "Image du Monde Challenge: Transcribing *The View of the World* by Gossuin de Metz.” https://imagedumonde.wordpress.com/.

Morreale, L. (2021). "La Sfera Challenge: An International Competition to Transcribe Goro Dati's *La Sfera*." https://lasferachallenge.wordpress.com/.

Ridge, M. (2020). Crowdsourcing in Cultural Heritage; a Practical Guide to Designing and Running Successful Projects. In: Schuster K., Dunn S. (eds) Routledge International Handbook of Research Methods in Digital Humanities. Routledge, Abingdon, p. 461–480.

Romein, A. et al, 2022. “Exploring Data Provenance in Handwritten Text Recognition Infrastructure: Sharing and Reusing Ground Truth Data, Referencing Models, and Acknowledging Contributions“ https://doi.org/10.5281/zenodo.7267245.

Schlagdenhauffen, R. (2020). “Optical Recognition Assisted Transcription with Transkribus: The Experiment concerning Eugène Wilhelm’s Personal Diary (1885-1951)”. 2020. https://hal.science/hal-02520508v2.

Venice Centre for Digital and Public Humanities (VeDPH). HTREC 2022: Improving the HTR Output of Greek Papyri and Byzantine Manuscripts. https://www.aicrowd.com/challenges/htrec-2022.