Title: Historic machines from 'prams' to 'Parliament': new avenues for collaborative linguistic research

Types of proposal: short paper

Authors:
Mia Ridge, British Library
Giorgia Tolfo, British Library
Kalle Westerling, British Library
Nilo Pedrazzini, The Alan Turing Institute/University of Oxford
Barbara McGillivray, King's College London and The Alan Turing Institute

Research in computational linguistics has made successful attempts at modelling word meaning at scale, but much remains to be done to put these computational models to the test of historical scholarship (see e.g. Beelen et al. 2021). More importantly, a lot of computational research looks at texts in a historical vacuum, 'synchronically', as linguists would say. In *Living with Machines*, an interdisciplinary research project that rethinks the impact of technology on the lives of ordinary people during the Industrial Revolution, we decided to address a fundamental question: what did people mean by ‘machine’ and how has this meaning changed over time?

This paper outlines how a simple research question like 'what was a machine?' can provide an opportunity to engage the public with our work while also generating data for analysis and new avenues of research in a radically collaborative way.

Turning to a diachronic perspective, we wanted to capture how changes in the usage of this word in nineteenth century texts can help us understand the role of machines in nineteenth century imaginations. An earlier crowdsourcing project used a working definition of machines as ‘devices or equipment not powered by people or animals’; the results of that project showed that this definition did not reflect usage in contemporary newspaper articles.

Accordingly, we designed the 'What's that machine?' citizen science tasks to find out what a 'machine' was in the 19th century as part of our linguistic and historical research. As engaging the public with our research is a key goal of the project, crowdsourcing, rather than internal annotation, was a natural fit; it also allowed us to tackle classification challenges at scale.

We set up two related 'What's that machine?' tasks on the Zooniverse platform: *Describe it!* and *Classify it!* The former asked the public to transcribe excerpts from newspaper articles that mentioned or defined a 'machine'. The latter asked people to look closely at the content of an article to assign meaning to the ‘machine’ mentioned based on categories derived from the *Oxford English Dictionary* (OED) definitions (Mechanical apparatus, Transport, War and military, Figurative, Other).

We developed these categories by reviewing the OED, where we found twenty-six senses and definitions that cover the use of this word from its appearance from the mid-16th century to now. Using the OED definitions as a starting point, we invited people to close-read a selection of newspaper articles published in the 1800s and match the use of the word 'machine' to the more appropriate category. Comments by participants on the Zooniverse 'Talk' boards show how the task provided opportunities for the public to understand what people thought of as 'machines' within our period of interest, while their classifications contribute to a large dataset of human-annotated historical newspapers at the lexical semantic level. This paper covers the process of developing, launching and analysing the results of this task, from formulating the research question to designing the crowdsourced annotation task, exploring, iterating, and finding ways to collaborate through new and radical participatory research practices.

When we began reviewing the results, we found that the vast majority of 'machines' were a few specific types of machines (especially sewing machines) listed within newspaper advertisements. Repurposing the existing 'subject sets' of images already uploaded to Zooniverse, we designed a dedicated task 'Ad or not?' task to identify newspaper ads. This task was designed as a simplified task specifically for inclusion in the Zooniverse app, which broadened our pool of participants and meant the task was completed very swiftly. The ability to exclude advertisements immediately improved our analysis and visualisations.

These tasks, based on our human ability to understand the meaning of words in context and the ability of computers to work at scale, are a great example of the benefits of collaboration. In the context of *Living with Machines* these tasks are useful for a) producing annotations that can be used as a lexicon or provide training data for building classifiers (thereby algorithmically expanding the machine learning vocabulary further); b) creating a modular and reusable end-to-end pipeline (exporting and processing texts such as newspaper articles, then exporting and processing the resulting crowdsourced annotations, facilitating their (re)use for other downstream tasks); c) providing a reflection on sampling tools and procedures that fit project research parameters, and d) offer a concrete opportunity of public engagement with lexical work.

Digitising historical periodicals is a complex task. The prevalence of advertisements within our initial dataset was the result of poor quality 'segmentation' processes applied to newspaper pages during the overall outsourced digitisation process. Excluding ads allowed us to refine the original classifications and re-run the original classification task in a revised version, to get a better sense of how the word machine was used in context in the period of interest.

Our research is ongoing. We plan to use this 'ad or not' data to train a machine learning model to detect context-specific ads within digitised historical newspapers, and to release the model for open usage by other researchers. Uses of this model include evaluating optical layout recognition (OLR) processes within newspaper corpora, and excluding (or including) advertisements from distant reading and crowdsourcing activities.

This form of public engagement will provide an opportunity to increase non-specialist's awareness of data science and computational linguistic methods, and to discuss the complexity of defining terms by sampling contemporaneous texts. As the project proceeds and introduces a more direct loop between machine learning and source data selection, we aim to contribute to the field of human computation, with future plans include building 'human in the loop' processes where the public evaluate and improve 'ad or not?' predictions.

Our work shows that designing for public engagement and linguistic research is a multidisciplinary collaboration from the earliest stages of research design through to final publication. This practical project at the intersection of disciplines around the arts and humanities is a vital contribution to the broader goal of proving the benefits of participatory research and paves the way for research on the use of words in context.

References

Beelen, Kaspar, Federico Nanni, Mariona C. Ardanuy, Kasra Hosseini, Giorgia Tolfo & Barbara McGillivray. 2021. When time makes sense: A historically-aware approach to targeted sense disambiguation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2751–2761. DOI: 10.18653/v1/2021.findings-acl.243.

"machine, n.". OED Online. December 2021. Oxford University Press. https://www.oed.com/view/Entry/111850 (accessed February 11, 2022).