Studentsourcing annotation for Early Modern English

Universität Konstanz

George Walkden • george.walkden@uni-konstanz.de
http://walkden.space • Twitter: @gwalkden

Early Modern English is a language with only incipient standardization and **substantial variation**, therefore presenting challenges to automated methods of annotation. Together with the 23 **participants in a Masters-level seminar** at the University of Konstanz I built a small **publicly available corpus** (47,481 words) of English 1500–1800. On the whole, this worked well, suggesting that **the method could be deployed more widely**.

Studentsourcing

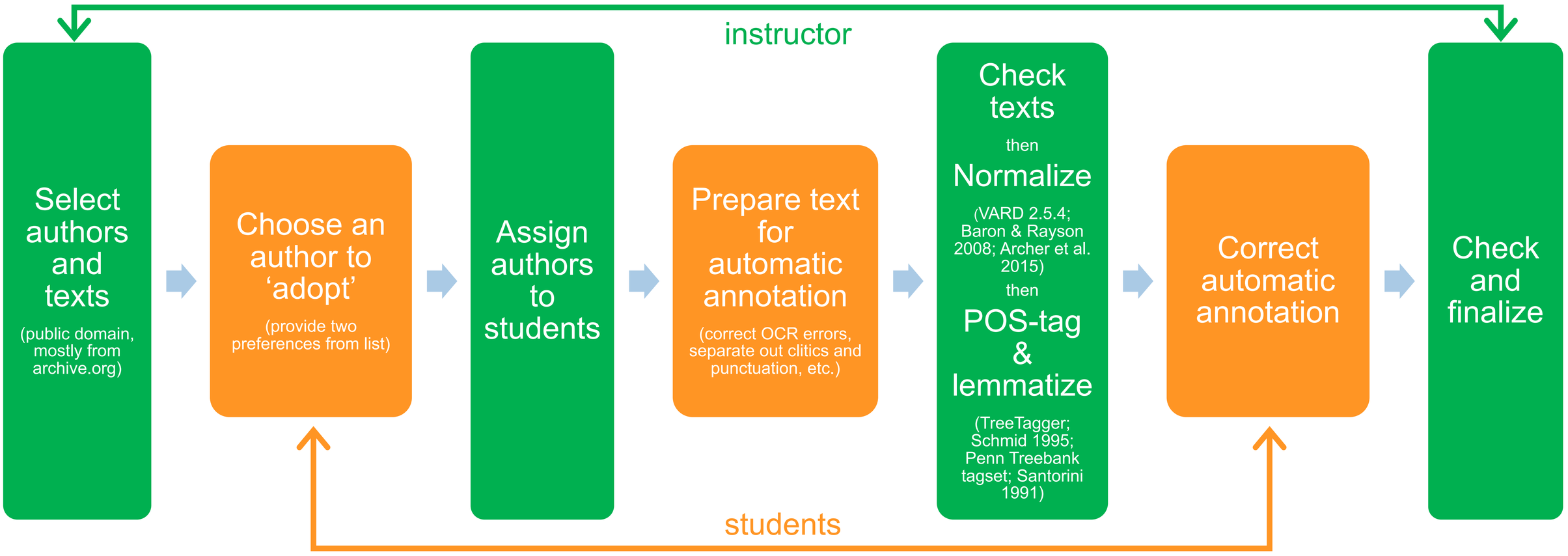
Variant of **crowdsourcing**, but students are the crowd. Successfully deployed at scale in biology (Hernandez et al. 2018). Corpus creation within the curriculum is not widespread, though see Lüdeling et al. (2008), Krause et al. (2012), and Zeldes (2017) for pioneering precedents.

Three of the writers included in the Emerging Voices corpus. **Left**: Teresia Phillips (1709–65), courtesan and bigamist. **Centre**: Ignatius Sancho (c. 1729–80), composer and freed slave. **Right**: Anne Askew (1521–46), poet and martyr.

Composition of the corpus

The corpus consists of 24 text samples of 2,000 words (+/- 200), each by a different author: 8 per century, 15 female, 9 male. The texts are **egodocuments** (Presser 1958): 'a broad category comprising several forms of autobiographical texts, including autobiographies, memoirs, diaries, travel journals, and personal letters' (Mascuch et al. 2016). Efforts have been made to deliberately **overrepresent** writers underrepresented in existing corpora: women, writers from outside England, people of colour, writers not from the upper classes.

Corpus construction process



Pluses and minuses

+ **Meaningful** assessment: not term papers ending up in the waste paper basket
+ **Well received**: it was appreciated 'that we're actually producing sth relevant for further research' (student evaluation comment)
+ More efficient **use of time** than normal assessment
+ High-quality manually checked **end product**
+ Skills involved are to some extent **transferable**

– Requires **more instructor input** than the average course – probably ca. 2x
– Small **maximum group size** (scalable only with support)
– Early Modern English is **challenging** for students
– Lack of appropriate annotation **tools** (?)
– Is using students in this way **exploitative**? (Keralis 2018)
– Difficult to give appropriate **training** and clear **objectives**

Outlook and future possibilities

- Use of **semi-automated** annotation tools
- More **levels** of annotation (syntax, discourse, coreference)
- Overlapping consensus-based annotation
- Quantify annotation **quality**?
- Use **XML** or **TEI** XML?
- Find better ways to **credit** students for work (see the manual online for full credits)

Further suggestions welcomed!

Download the corpus: walkden.space/voices

References: **Archer**, Dawn, et al. 2015. Guidelines for normalising Early Modern English corpora: Decisions and justifications. *ICAME Journal* 39, 5–24. **Baron**, Alistair, and Paul **Rayson**. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. *Proceedings of the Postgraduate Conference in Corpus Linguistics 2008*. **Hernandez**, Simon, et al. 2018. *Tiny Earth: a research guide for studentsourcing antibiotic discovery*. **Keralis**, Spencer. 2018. Disrupting labor in the Digital Humanities. In Kim & Stommel (eds.), *Disrupting the Digital Humanities*, 273–294. **Krause**, Thomas, et al. 2012. Multiple tokenizations in a diachronic corpus. Paper presented at *Exploring ancient languages through Corpora*, Oslo. **Lüdeling**, Anke, et al. 2008. Das Lernerkorpus Falko. *Deutsch als Fremdsprache* 2, 67–73. **Mascuch**, Michael, et al. 2016. Egodocuments and history: a short account of the *longue durée*. *The Historian* 78, 11–56. **Presser**, Jacques. 1958. Memoires als geschiedbron. *Winkler Prins Encyclopedie* 7, 208–210. **Santorini**, Beatrice. 1991. Part-of-speech tagging guidelines for the Penn Treebank project. **Schmid**, Helmut. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop, Dublin, Ireland*. **Zeldes**, Amir. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources & Evaluation* 51, 581–612.