HUM19UK Corpus FAQ

1. What is in the HUM19UK corpus?

HUM19UK is the **H**uddersfield, **U**trecht, **M**iddelburg corpus of 19th century British fiction (novels only). It was created between 2016-2019 as a collaborative project between the University of Huddersfield (UK), Utrecht University (the Netherlands) and University College Roosevelt in Middelburg (the Netherlands).

The corpus contains 100 complete novels written by 100 authors (50 male/50 female) over 100 years, with roughly 10 novels per decade. It totals 13 million words.

Find the detailed contents of the corpus in the PDF file titled “HUM19UK Corpus Contents” among the downloads.

2. How did we decide on the contents of the HUM19UK corpus?

Creating a representative sample of 19th century British prose fiction (in novel form) was not without challenges. Other corpora that aim to represent prose fiction from a particular time period, such as the Brown family of corpora, have used random sampling techniques whereby titles are sampled from a list of publications from a particular year or decade. This method was not practical for us because, first of all, we were unable to obtain a definitive list of British novels published during the 19th century from which to sample. Secondly, even if we had such a list of publications it is likely that it (and therefore any sample derived from it) would contain rare and hard to find novels. Therefore, any random sample would have been restricted to those novels that were accessible. A further, more significant restriction was access to machine-readable digital versions of novels. Since we did not have the resources to create our own machine-readable versions of texts, we relied completely on various on-line sources that provided such versions of novels. This means that the population of 19th century novels from which we could sample from was actually the population available online in electronic form. However, producing a list of 19th century novels available online would be a rather difficult task.

Instead, we adopted a different approach that involved hand-picking texts for inclusion in our corpus using some guiding criteria designed to make the corpus representative and balanced. Our aim was to create a balanced corpus representative of 19th century fiction that contained complete novels with publication dates spread across the whole time period. In order to achieve this representativeness, balance and spread we aimed to include roughly one text per year across the 100 years of the 19th century (therefore creating a corpus comprising 100 texts in total), and that each text should be written by a different author, with a 50/50 male/female split across the corpus. Also important for us was that authors who although not very well known now were well-read during the 19th century itself were included in the corpus as well as well-known authors and texts from the literary canon.

The corpus was constructed in a cyclical fashion, which means we created different versions of the corpus before producing the final version that is available for download on this website. The different versions of the corpus evolved as we re-thought important factors such as representativeness, balance and spread in light of feedback from colleagues at the University of Huddersfield and at PALA and IALS conferences.

The final content of the corpus is the result of the interaction between our decisions about the structure of the corpus (100 texts by 100 different authors spread across 100 years) and accessibility to trustworthy machine-readable versions of texts.

3. How did we clean and tag the corpus?

The published version of the HUM19UK corpus contains machine-readable versions of novels that have been cleaned and annotated. By cleaned we mean that, where necessary, we removed from the text any illustrations, captions, reviews, transcriber notes, and introductions and epilogues by anyone other than the text's author.

The annotation we added is minimal, but we hope of some use. All chapter headings and numbers, as well as any other divisions in the text, such as books or parts have been placed in tags. Where multiple volumes of one text appear in the corpus, volumes are tagged as divisions. Where the electronic version of the text we used contained page numbers we also placed these in tags so that they will be ignored by corpus software. We also added a small header to each text which included: novel title; author's name; author's gender; year of first publication; and source of the machine-readable version of the text.

Additionally, we did not remove any sections of the text that although not part of the story told in the novel may be relevant to its interpretation, such as prefaces by the author, epigraphs and content pages (where present in the transcription). Instead, we enclosed these in angle brackets (i.e. < >) so that they will be ignored by most corpus tools but can be extracted if required for analysis.

4. What has been done to enable use of different parts of the corpus?

The file name of each corpus text is its year of publication. This should allow you to easily cluster texts per decade, should you want to use only one or a selection of decades of our reference corpus.

The tags for the author's gender should enable you to easily cluster texts according to gender.

Chapter tags allow you to extract all first chapters, last chapters, introductions, or any other combination of chapters across some or all texts.

5. What are the sources of the machine-readable texts included in HUM19UK?

- Project Gutenberg: www.gutenberg.org
- Chadwyck Healey: http://collections.chadwyck.co.uk
- Celebration of Women Writers: https://digital.library.upenn.edu/women/
- Chawton House: https://chawtonhouse.org/the-library/library-collections/womens-writing-in-english/novels-online/
- Public Library UK: http://www.public-library.uk/

6. Who were involved in the creation of the HUM19UK corpus?

University of Huddersfield:

Fransina Stradling, Dr. Brian Walker (now independent scholar), Prof. Dan McIntyre (now at Uppsala University, Sweden), Elliott Land, Dr. Hazel Price (Now at Salford University, UK)

Utrecht University/University College Roosevelt:

Prof. Michael Burke

7. Who do I contact for further information?

Please email Fransina Stradling on fransina.stradling2@hud.ac.uk with any questions you may have.

8. How do I cite the corpus?

When you cite the HUM19UK Corpus in your work, please use the following reference:

HUM19UK, Version 1. 2019. University of Huddersfield, Utrecht University, University College Roosevelt, Middelburg. https://www.linguisticsathuddersfield.com/hum19uk-corpus