Data Collection – Books

What is the BNC2014?

The British National Corpus 2014 is a major project led by Lancaster University to create a 100 million word corpus (a large collection of 'real life' language) of modern-day British English. This corpus will be used by researchers to understand more about how language works and how it is evolving. Educators, dictionary compilers and the interested public will also be able to access the corpus to find usage examples of modern British English in different genres.

You can find more about the BNC2014 project at http://cass.lancs.ac.uk/.

Focussing on books

What do we need?

- Books written by **British** authors
- Genre:
  - Fiction: poetry and prose (general, children, teens, fantasy, crime, romance, etc.)
  - Non-academic non-Fiction (hobbies, history, popular science, travel, technical…)
- Date of publication:
  - Fiction: **first** published in or after 2010
  - Non-fiction: books published before 2010 are acceptable as long as the selected edition was published in or after 2010

How much to get from each book?

Ideally, we want a number between 20,000 and 50,000 running words from each book.

It is difficult to estimate the number of pages, as it will depend on the book layout and font. On average, a single page contains around 400 words. To reach 50,000 words, you would in principle need to scan 120 single pages.

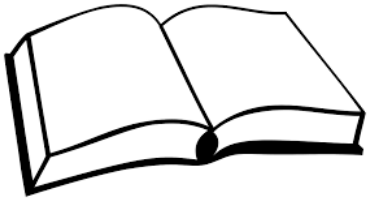
Scan can either the beginning, the middle or the end of the book; vary this between individual books.

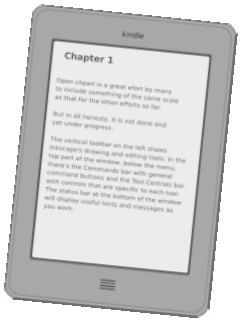
Corpus and copyright:

- Corpus data is collected for research (non-commercial) purposes.
- Only samples of data are taken.
- Following changes to the law in June 2014 (CDPA s. 29A), it is now permissible to copy works to which one has legal access (e.g. via the Library) to enable data and text mining for non-commercial research. This covers creating a corpus of language materials.

Task 1: Select the books to be scanned.

Both hard copies and kindle versions can be scanned.





Step ❶: Find a book fits the criteria specified in "What we need?":

- **Lancaster student/staff**: You will find a wide range of books at the university library's "Leisure Reading" section (A Floor – next to "Reserved books") and the English Literature section on the C floor.
- **Lancaster City library is an even better option.** It has an extensive collection, with a wide range genres, including many we are short of: children and teen fiction, poetry and non-fiction books.
- **If you are not in Lancaster**: do use your local or university libraries.
- **And of course, your own books are an excellent option, including kindles.**

To select a book, please check:

- Whether the author(s) is **British**
  - ➢ You can find a list of British authors at http://cass.lancs.ac.uk/wp-content/uploads/2018/04/authors.pdf
  - ➢ If the author is not listed, check the book blurb
  - ➢ Google is also an alternative

- Preference is for books published in or after 2010. **Non-fiction books** published before 2010 are acceptable as long as the selected edition was published in or after 2010. You will find the date of publication in the first pages.

Task 2: Scan the book

What do we want?

☑ **A full set of all scanned pages from one single book, all in one single file, with no missing bits.**

☒ No bibliographical references

☒ No index pages or glossaries

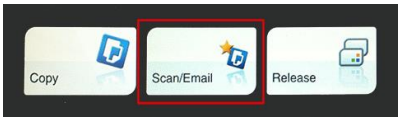
☒ No picture-only pages

How to scan:

The instructions here are for those at Lancaster University. Please adapt them if you are scanning somewhere.

To carry out this task, you will use one of the photocopy machines on campus. A map of LU printer locations is attached. Alternatively, please refer to website: http://www.lancaster.ac.uk/iss/info/IThandouts/printing/printer-locations-map.pdf

1. Place the book page(s) face down on the glass.
   ☺ **Tip**: Make sure the text is displayed horizontally. This is to minimize errors in the final stages, when we will have to convert the image into a Word file.
2. Login to the printer. When the "**Account Confirmation**" screen shows, press OK to continue.
3. Press the **Scan/Email** option.



4. To email the scanned document to yourself, select the **Me** option and confirm your email address.

5. Press the "Scan Size" option in the menu bar in the lower part of the screen to adjust the format and size of the page accordingly

6. Press the "Metric Sizes" option and choose the format and page size that best match those of the book you are scanning.

For most books, the best option is A4 landscape.

7. As you want to scan multiple pages, press on "Application" option (in the lower right corner) and make sure that the "Separate Scan" option is set as "ON".

8. Press "Start" to scan the page(s).

9. When scanning is complete, you will see the following message on the screen: *Load the next original and press [Start]*. So follow these instructions.

> *****IMPORTANT*****
>
> To reach around 50,000 words from a single book, you are likely to need about 50 double pages (two single pages of the book). **WE HIGHLY RECOMMEND YOU TO SCAN THEM IN SETS OF 10 (20 single pages).** This is to avoid ending up with a very heavy file to transmit to your email. The university system is very likely to block it and you will have to scan it all over again.

10. Once you have scanned all 10 pages you want, press "Finish". You then press "Start" to begin transmission.
11. If you need to scan more pages of the same book, repeat the steps above. **NOTE: Make sure you configure the settings again (steps 5 and 6). It usually resets once you have finished.**

Task 3: Write the metadata

Once you have concluded the scanning procedures, write down the following information about the book. You will need to provide these details when you submit your scanned files.

- Title:
- Author(s)
- Year of first publication:
- Year and edition of scanned publication:
- Genre: poetry, fiction (prose), non-fiction (prose)
- If possible, subgenre:

Task 4: Submit your scanned files

Your scanned files will be all sent to your email, in pdf format.

1. Save all files in your drive
2. Rename the files with a suitable name to indicate the book title. If you have more than one file for the same book, please number them in the correct order. For example "Kings_Speech_1".
3. Go to the project website (http://cass.lancs.ac.uk/?page_id=2462). Scroll down and click on "Book Collection". **Note**: You will need to login to a google account to upload the files. Alternatively, you can send us your files by email (writtenbnc2014@gmail.com). **Please remember to attach the metadata.**

What will happen next: we will do this task for you

To be included in the corpus, we need to convert the image into a text file. This step requires specific software so we are not able to do it in this session.

We also need to:

- Remove any picture or template (i.e. title of the book at the top of the page)
- Correct OCR errors and join words divided in syllables
- Add the metadata to each text