Text Hackathon: Extracting Knowledge from Big Digital Texts

(Centre for Textual Studies, De Montfort University, 10-12th November 2017)

Jonathan Culpeper, Lancaster University, UK

Introductory readings

For more detailed overviews of corpus linguistics, I recommend Biber et al. (1998), McEnery and Wilson (2001) and Mcenery and Hardie (2011). For more practically oriented accounts, I recommend two excellent books: Adolphs (2006) and the more extensive McEnery et al. (2006). For a cutting-edge and up-to-date handbook of thoughts about and activity in the field, the best volume is Biber and Reppen (2015).

Corpus analysis programs

These fall into three camps:

(1) Those that you download to your computer, and then load your own corpora/data:

- *AntConc* (http://www.laurenceanthony.net/software.html): Easy to use but well specified, free, and has supporting video guides for the user.
- *Wordsmith Tools* (http://www.lexically.net/wordsmith/): Similar to the above but does even more ... but is not free.

(2) Those that you access via a web browser, and then upload your own corpora/data:

- *WMatrix* (http://ucrel.lancs.ac.uk/wmatrix/): Unlike the above, you don't download a program to your computer but upload your corpus/data to WMatrix web browser. Covers a number of the key functions of the programs in (1), but its distinctive advantage is that it will also automatically tag your corpus/data for part-of-speech and semantic field, and then you can look at grammatical or semantic patterns.
- *SketchEngine* (https://www.sketchengine.co.uk/): This is not free, but has an impressive array of corpora pre-loaded, powerful analysis tools, and also the ability for you to upload your own corpora/data.

(3) Those that you access via a web browser, and have the corpora/data already in them:

- *CQP-edition of BNCweb* (http://bncweb.lancs.ac.uk/): Obviously, contains the British National Corpus (BNC). (Soon to be complemented by the new BNC2014).
- *CQPweb* (https://cqpweb.lancs.ac.uk/): Almost a hundred corpora (depending on restrictions) covering many genres, Englishes, historical texts and different languages. Includes Early English Books Online (EEBO-TCP).
- *http://corpus.byu.edu/* A platform containing a selection of very large and current corpora. You can also download the corpora.

References

Adolphs, S. (2006). *Introducing electronic text analysis*. London: Routledge.

Biber, D., & Reppen, Randi. (2015). *The Cambridge handbook of English corpus linguistics* (Cambridge handbooks in language and linguistics).

McEnery, T., & Hardie, Andrew. (2011). *Corpus linguistics : Method, theory and practice* (Cambridge textbooks in linguistics). Cambridge: Cambridge University Press.

McEnery, T., & Wilson, Andrew. (2001). *Corpus linguistics : An introduction* (2nd ed., Edinburgh textbooks in empirical linguistics). Edinburgh U.P.

McEnery, T., Xiao, Richard, & Tono, Yukio. (2006). *Corpus-based language studies : An advanced resource book* (Routledge applied linguistics series). London: Routledge.