**British National Corpus Full Word Count**

**CMIST**

**University of Manchester**

**John Mcloughlin**

**September 2016**

Extract from corpus via xsltproc

Split into 50 input files (e.g. \*\*\*.txt; we used BNCa.txt, BNCb.txt, etc.) for loading into NVIVO. NVIVO can’t load the full corpus at once. Have to split. In NVIVO, create an empty project, e.g. BNCFull.nvp.

Use External Data > Import Internals menu option in NVIVO. (There is a short minimenu option as an alternative: In the blank white space of Sources > Internals, right-click the mouse and Import Documents.) Load fifty input files into NVIVO, by highlighting in Windows about five files at one time. NVIVO hangs if you load too much in at once. No error message is given. Eventually you have one NVIVO project with all fifty files in. Save it. It is a project, e.g. BNCFull.nvp. Each file is a ‘source’.

NVIVO cannot export more than 65536 rows to an Excel file. Need to run word count query on three files at a time to get export under this limit. This is 15 queries. Each query refers to 3 or 4 files, e.g. 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4.

Save the query outputs as 15 csv files from Excel. You have done queries on 3-4 files at a time, and exported the results using Right-Click in NVIVO to output the single xlsx file.

Load the CSV files into a fresh SQLite database table. URL <http://sqlite.org/> accessed Sept. 2016.

Re-enter NVIVO.

Run word count query on table to generate output. Save as CSV file.

Problem: Running the word count query on only three files at a time means that the full corpus cannot be compared all at once. This is a problem caused by NVIVO’s inability to export more than 65536 rows and inability to run the query on all of the corpus in one go.

To create bar charts, use grep ( Linux/ Mac ) or FINDSTR ( Windows ) to extract the required word from the CSV file. EG

grep teach BNC\_Full\_word\_count.csv > teach.csv

This will extract values for all words related to the stem ‘teach’ and create a new CSV file, teach.csv. Load this into Excel and create a bar chart. Save as BNC\_Full\_teach\_bar.xsls