**British National Corpus Full Word Count**

**CMIST**

**University of Manchester**

**John Mcloughlin**

**September 2016**

1. Extract

Extract from corpus via xsltproc

1. Create Project

In NVIVO, create an empty project, e.g. BNCFull.nvp.

1. Split Output

NVIVO is unable to load large text files so it is necessary to split the extracted text into 50 files for loading into NVIVO. Split into 50 input files (e.g. \*\*\*.txt; we used BNCa.txt, BNCb.txt, etc.) for loading into NVIVO.

1. Load

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’.

1. Run Word Frequency Query

Once the word frequency query has been run, the output can be saved to an Excel spreadsheet. NVIVO cannot export more than 65536 rows to an Excel file due to an unknown NVIVO restriction. It is necessary 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.

1. Extract Word Count Totals

Each of the 15 CSV files will contain word counts. It is now necessary to summarise them. It may be possible to do so in Excel, but we chose to use the free database SQLite ( URL <http://sqlite.org/> accessed Sept. 2016 ).

Invoke SQLite from a Terminal ( Linux/ Mac ) or Command Prompt ( Windows ).

sqlite3.exe

Run the following SQLite commands to load the CSV data ( SQLite prompt is sqlite> )

sqlite> .mode csv

This tells SQLite that the data are comma separated. Next, load data from CSV file:

.import bnc\_full\_1.csv wctab1

loads data from Excel file bnc\_full\_1.csv and creates a table called wctab1. SQLite uses the header details in the CSV file to create the table column names. Run this import command for all 15 CSV files to create tables wctab1 to wctab15.

Create a new table to hold the total word counts. Use this SQL:

sqlite> CREATE TABLE wordcount(  
 "Word" TEXT,  
 "Length" TEXT,  
 "Count" TEXT,  
 "Weighted\_Percentage" TEXT,  
 "Similar\_Words" TEXT  
);

A table similar to each of wctab1 to wctab15. Now load data from wctab1 to wctab15 into new table wordcount. Use this SQL:

sqlite> insert into wordcount select \* from wctab1 ;

The semi colon at the end tells SQLite that the command is ready to run. Press RETURN. Repeat this step for each of the 15 tables wctab1 to wctab15. Now generate the word count output with this SQL:

sqlite> .output BNC\_Full\_word\_count.csv

sqlite> select word, sum(count) from wordcount group by word ;

This will generate the a list of words and the total counts for that word and save it to the file BNC\_Full\_word\_count.csv

Save the SQLite session ( if required ):

sqlite> .save BNCFWC.sqlite

Saves the SQLite session and tables to the named file.

1. Create Bar Charts

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

Run this command from a Terminal ( Linux/ Mac ) or Command Prompt ( Windows ). 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

1. NVIVO Limit Problems

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.