Workshop: Counting Words, Texts or Features

TEXT FORMATS

a) Locate a Word *.doc* (or write a very short one if necessary) and save it a) as Windows text, b) as web-page, c) as rich text file (*.RTF*), d) as XML. Close Word (otherwise it will selfishly deny any other programs access to the text file!).

b) Open the four in Notepad. (*File* | *Open* with file-type set to `*.*`.) Examine the differences and compare your results with a neighbour.

c) Now open them in Internet Explorer (*File* | *Open* | *Browse* | *All files*). Examine the differences and compare your results with a neighbour.

Implications

………………………………………………………………………………………………

………………………………………………………………………………………………

………………………………………………………………………………………………

WHAT YOU ACTUALLY GET IN YOUR TEXT

d) Use the Character Analyser utility in WordSmith 5 to count the characters.

   i. Explain any differing results.

   ii. Are there any apostrophes or dashes (as opposed to hyphens) in your text? If so, what does the Character Analyser think they are – what code have they been given and what does Character Map (*Start* | *Program Files* | *Accessories* | *System Tools*) say these are?

e) Make a word-list using the WordList tool. Does its word count match that of Word? If not, can you discover why not?

Implications

………………………………………………………………………………………………

………………………………………………………………………………………………

………………………………………………………………………………………………

HANDLING TEXTS WITH WORDSMITH

f) Study a text from the BNC to identify suitable tags. Prepare a tagfile and make a wordlist which distinguishes between the same word-form as one POS and another e.g. *bear* as noun and *bear* as verb.

g) Make a set of 10 different wordlists, using increasing sizes of corpus text, e.g. 1,000 words, 5,000 words, 10,000 words, 20,000 words etc. Create a graph using Excel which shows the proportions of hapax legomena in the 10 lists.

h) Find a way of counting the number of paragraphs, headings and sections in a sample text. Use your own mark-up for this (*notepad* is the best tool).

i) Find ten different text-types (advert, sermon, letter etc.) on the Internet or in our corpus resources. Determine what (if anything) is missing from these as in the gravestone example in the lecture.

Implications

……………………………………………………………………………………………

……………………………………………………………………………………………

……………………………………………………………………………………………

Reference

Scott, M. & C. Tribble, 2006. *Textual Patterns: keyword and corpus analysis in language education*, Amsterdam: Benjamins. P53.28.S42 (long & short loan + electronic access), Chapters 1& 2.