Software for the Johnson Dictionary Project

Alex Collier

RDUES
University of Birmingham
England

ABSTRACT

This paper firstly describes the tag set which has been implemented for the encoding of the text of the Dictionary, covering the choice of tag labels and their hierarchical inter-relationships which ultimately identify the hierarchical structure of the Dictionary itself. The various automatic validation procedures, which have identified several consistent errors on the part of the keyboarders, will also be touched upon.

Following on from this, I will say something about the most typical structures which Johnson used in creating his definitions. This analysis proves very easy to do once the tags are in place, since it is merely a case of finding the most frequently used tag sequences.

Since the Project has been concerned with the preparation of an electronic version of two different Editions, it will be interesting to discover whether there are any significant differences between the Editions. This may be in terms of vocabulary use in the definitions or in the frequency and distribution of citations from particular authors. It will also be possible to compare Johnson's defining vocabulary with a balanced corpus of modern English in order to identify typical lexical characteristics of the definition style.

Finally, a brief resume of the functionality of the final product will be presented.

The Text

The Johnson Project at the University of Birmingham has been in progress for the past four years. The text now entered encompasses the complete text of both the First (1755) and Fourth (1773) Editions. The text has been marked up in order to distinctly identify the various parts (headword, etymology, part-of-speech etc) of the dictionary entry's structure, making it possible to process the text with a free-text retrieval package such as is currently in use with other online dictionaries and highly structured documents. In all twenty-two different tags have been used, the meaning of which is set out below.

@01 headword

@02 part of speech

@03 etymology or source of word

@04 definition

@05 citation

@06 author [of citation]

@07 title [of cited work]

@08 location [of cited passage, if given]

@09 Note on usage, spelling or pronunciation [e.g. This word is now obsolete]

@10 area of discourse [e.g. In law or A sea term]

@11 cross-reference

@12 citation/definition [e.g. @01 WOODROOF. @12 An herb. @06 Ainsworth]

@13 sub-headword @14 sub-headword collocation

@15 sub-headword definition [e.g. @13 To WORK @14 out. @15 To effect by toil.]

@16 authorial comment [e.g. I know not well the meaning here]

@17 headword collocation [e.g. @01 STAR @17 of Bethlehem]

@18 citation/etymology [e.g. @18 from waghelen, to shake, German. @06 Skinner]

@19 alternative headword or variant form [e.g. @01 ROISTER, @19 or roisterer. @02 n.s. @03 [from the verb.] @04 A turbulent, brutal, lawless, blustering fellow.]

@22 Alternative name [e.g. @01 RIBWORT. @02 n.s. @22 [plantago.] @04 A plant.]

@23 Grammatical information [e.g. This word has no plural.]

@24 Latin equivalent [e.g. ERELONG. adv. [from ere and long] Before a long time had elapsed. @24 Nec longum tempus.]

All tags start at the beginning of the line and no end tags are employed since each field ends where the next begins. The one exception to this is the etymology field, which often contains cross-references or notes on pronunciation or usage.

In addition to structural markup, typographical features such as font change have also been included in the encoded text, but no attempt has been made to assign the function performed by the change in form.

The Dictionary as Corpus

The dictionaries are currently held as text files, pending the finalisation of the facilities in the final product. The results presented here are based on the analysis of the text from the point of view of a corpus researcher. This analysis has been carried out for the most part using standard tools available under the Unix (tm) operating system, plus the tools which we employ on an everyday basis for the analysis of corpus data.

When looking at a corpus of natural language for the first time it is useful to discover the components of the language, which would normally be achieved by creating a wordlist of the types (unique forms of words) and their related frequency of occurrence. This was therefore the first task which was undertaken - looking initially only at the markup. In the case of the Dictionary the types are already a known entity, since we put them there. The number of types is thus twenty-two. The following figure shows how many times each of these tags occurred.

01	42788
02	40910
03	35409
04	65110
05	109562
06	108108
07	52971
08	12614
09	5968
10	888
11	933
12	4025
13	1241
14	1239
15	1139
16	169
17	467
18	306
19	87
22	279
23	377
24	23

Having answered the question 'What are the words of the language', we can now move on to look at how the words are combined into 'sentences'. It is quite easy at this stage to pull out all sequences of tags. If this is done and all the combinations are collated, we get 41,728 sequences of tags, which consist of 10,820 unique sequences (types). Of these, over 9,000 occur only once. The following is a list of the first few most frequently-used sequences:

Frequency	Sequence
3990	1 2 3 4 5 6
3870	1 2 3 4
3728	1 2 3 4 5 6 7
982	1 2 4 5 6
903	1 2 3 12 7
873	1 2 3 4 5 6 5 6
640	1 2 3 4 5 6 7 8
609	1 2 4 5 6 7
595	1 2 3 4 5 6 7 5 6
567	1 2 3 12 6
501	1 2 3 4 5 6 7 5 6 7
462	1 2 3 4 5 6 5 6 7
349	1 2 12 6
311	1 2 4
309	1 2 3 4 5 6 5 6 5 6
306	1 2 3 4 5 7
278	1 2 3 4 9 5 6
256	1 2 3 4 9 5 6 7
183	1 2 3 4 5 6 7 5 6 5 6
171	1 2 3 4 5 6 4 5 6

Johnson's most favoured sequence is therefore:

Headword
Part of Speech
Etymology
Definition
Citation
Author

which is used for nearly one in ten of his entries.

An example of this sequence is the entry for ‘abdicate’:

@01 To <<a>>A<<’>>BDICATE.
@02 <cf2>v.a.<cf1>
@03 [Lat. <cf2>abdico.<cf1>]
@04 To give up right; to resign; to lay down an office.
@05 Old Saturn, here, with upcast eyes,
Beheld his <cf2>abdicated<cf1> skies.
@06 <cf2>Addison.<cf1>

The above list of twenty sequences accounts for nearly 50 per cent of all the entries.

Johnson’s Use of Language

In addition to treating the codes which we have used to map the structure of Johnson’s entries as a ‘language’ with ‘words’ and ‘sentences’ it is very interesting to examine Johnson’s own use of language, that is, his definitions.

By compiling a wordlist of all the text of the definitions it becomes possible to compare Johnson’s defining vocabulary with a corpus of modern-day English in order to highlight the particular characteristics of the Johnsonian lexical choice. We find that Johnson uses 22,776 types (unique words) in composing his definitions, the text of which runs to 378,748 tokens. Of those types, over 4,000 are not found in the Birmingham Collection of English Texts (c. 20 million tokens). Some of the most frequently-used examples are as follows:

publick	180
inclose	73
cloaths	69
musick	58
contrariety	49
cloath	46
superiour	45
errour	41
inclosed	39
encrease	36
domestick	32
harrass	32
liquour	31
physick	31
traffick	29
publickly	28
inclosure	27
frolick	25
inferiour	24
croud	23

These words all exhibit spelling variances typical of the 16th to 18th century period but no longer found because they are obsolete (eg ick -> ic; our -> or) or have been normalised: (en <-> in). As well as examining words unique to Johnson, we can look at types which he uses far more frequently than would be

expected in a modern corpus. To list just a few of the more than 14,000:

Type	Ratio	Frequency
benefice	1221.12	25
terrour	1172.28	48
concretion	1025.74	21
anciently	976.90	40
blockhead	879.21	18
declivity	830.37	17
subtilty	781.52	16
deject	732.67	15
betokening	683.83	14
solicitation	683.83	14

The second column gives a score of how many times more frequently the word occurs in Johnson than we would expect based on its occurrence in the modern corpus. It is fascinating to find the word 'blockhead' here, used 18 times, each time as a direct synonym for the following headwords: 'asshead', 'block', 'blunderer', 'bull-head', 'buzzard', 'clodpoll', 'clotpoll', 'dolt', 'dulhead', 'dullard', 'half-wit', 'jobbernowl', 'jolthead', 'loggerhead', 'numskull', 'oaf', 'snipe' and 'sot'. Likewise 'terrour' occurs in 48 different definitions, including direct synonyms such as 'dread', 'horrour' (sic) and 'trepidation', derived forms ('affright', 'blast', 'grim'), as well as antonyms, eg 'dreadless', 'fearlessly' and 'peace' ('freedom from terrour', as Johnson calls it). There is much room here for further investigation, but unfortunately little room for further examples in this context.

The Editions

Since the aim of the Project has been to encompass two major editions of the Dictionary, we are now in a position to compare the use of language in the definitions of the two works.

This is accomplished merely by comparing the wordlists we made earlier from the text of the definitions. This gives us an interesting insight into the changes which took place in the language, or at least in Johnson's version of the language, in the years intervening the two Editions. We find that, for instance, 'harrassing' has become 'harassing', that 'hopelesness' has given way to 'hopelessness' and that 'molosses' has changed into 'melasses'.

The Final Product

We are now at the stage of finalising the functionality of the product. It will be distributed on CD-ROM and will provide access to both the Editions simultaneously so that comparisons can be made between them. This will be implemented via MS-Windows. The product will offer many of the facilities found in free-text retrieval packages (eg full text search, proximity search) as well as features specific to dictionary texts such as searching for a target in a particular field or displaying a headword list.

One problem encountered when searching for citations from a particular author or source is that a great many abbreviations are employed. In addition to the form 'Shakespeare', for example, we find 'Sh.', 'Sha.', 'Shak.', 'Shake.', 'Shakep.', 'Shakes.', 'Shakesp.', 'Shakespear' and 'Shaksp.'. In order to facilitate the process of locating all instances of such diversely-spelled authors and titles we are planning to include ancillary databases of equivalents which would function in the same way as the thesaurus feature found in some text retrieval software.