Alternative routes to highlight cultural semantic associations of a given key word: further experiments

8.1 Introduction

In an attempt to find alternatives to the time-consuming task of coding a whole dataset of more than 1500 sentences, or a whole wordlist of more than 10,000 words, Chapter 7 explored three possible shortcuts to highlighting culture-based semantic associations of a key word. The first route applied manual semantic analysis to the most frequent 50/100/150/200/250/300 content words in the wordlist, by generating concordances for each word, reading through the concordance lines and matching each word to one or more of the semantic categories available. The second one used the four most frequent content words to extract sentences from the manually coded dataset and create a sampled sub-corpus. Finally, the third route was based on random selection of sentences from the manually coded dataset, to create a random sub-corpus.

Of the three routes tested, the most promising one was random sampling, as the results were very close to the results of the original datasets, at both qualitative and quantitative levels. Also the first route, based on analysis of the most frequent 300 words, returned very interesting results, in the light of the fact that the most frequent 300 words in the wordlists cover only about 3% of the words in the dataset. Route two, on the other hand, will be discarded because its results were similar to, though slightly lower than, random sampling.

Consequently, the current chapter aims to verify whether the results obtained in the previous chapter with the most frequent 300 words in the wordlist and with random sampling may be considered dependent on the datasets and/or coding methods used (R.Q. 5). To this aim, an automatic semantic tagging tool (Wmatrix) was applied to the English elicited chocolate and wine datasets, as well as to the English Web datasets on chocolate and wine created for the current project. As described in Chapter 5, the English Web datasets were assembled by extracting 10,000 sentences including the node words from the UKWAC corpus – a large general corpus created from the Web using spidering tools. The extracted sentences were then purged of duplicates, which led to the creation of two sub-corpora: the chocolate sub-corpus, with 8436 sentences and 286243 running words; and the wine sub-corpus, with 7343 sentence and 277006 words.[1]

[1] The word count reported here is Wmatrix's (see Chapter 5, Table 5_4).

Semantic tagging with Wmatrix differs from the manual tagging used in the present work in two ways: first, semantic tagging is word-based rather than sentence-based; second, the USAS tagset in Wmatrix includes more than 400 different tags, while the my coding scheme includes about 90 tags. Finally, it must be remembered here that this experiment could be accomplished for English only, because automatic tagging with Wmatrix does not apply to Italian.

8.2 Most frequent 50/100/150/200/250/300 content words

In Section 7.2 in Chapter 7, the analysis of the most frequent 300 content words in each of the elicited datasets retrieved about 65-70% of the semantic fields in the respective dataset, corresponding to almost 90% of the fields with high conventionalisation and over 86% of the semantic associations. From a quantitative perspective, Spearman's test showed a strong level of correlation, with *rho* ranging between 0.800 and 0.900 ($p < 0.01$). In particular, as regards the English datasets, the top 300 content words showed – for *chocolate* and *wine*, respectively – 68.18% and 70.59% of the semantic fields, with correlation values of 0.810 and 0.877.

Let us now see what happens if we adopt a different coding scheme, and also a different set of data.

The Wmatrix interface (see Chapter 5.3.2) – which automatically POS tags, and performs semantic analysis of the given data – was used to generate frequency wordlists of the *chocolate* and *wine* English elicited datasets (including 1886 and 1938 sentences, and 9967 and 10967 words[2], respectively), and the *chocolate* and *wine* English Web datasets (including 8436 and 7343 sentences and 286243 and 277006 words, respectively, once purged of duplicate sentences).[3] In Wmatrix's frequency lists, each entry in the word list is accompanied by its raw count and the semantic category assigned. Thus the semantic categories appearing in the most frequent 50/100/150/200/250/300 content words could easily be qualitatively and quantitatively compared to the semantic categories appearing in the whole dataset (i.e. the semantic frequency list of the whole dataset). My interest while performing this analysis and comparison was in content words; thus, all the words corresponding to grammatical categories (USAS tags Z4 through to Z99) were ignored.

The following sections summarize and comment the results obtained with the elicited data and the Web data, separately.

8.2.1 Elicited data

The results of the comparison between the most frequent 50/100/150/200/250/300 content words in the elicited word lists and the corresponding whole datasets are summarised in Tables 8_1 and 8_2 below.

[2] The word count reported here – for both elicited and Web data – is Wmatrix's and, as explained in Chapter 5, is characterized by the fact that some entries in the word list are multi-word-expressions.
[3] See Chapter 5.

	USAS tags (n.)	USAS tags (%)	tag increase	Spearman's rho ($p < 0.01$)
Top 50 words	39	14.94	+ 39 fields	0.610
Top 100 words	62	23.75	+ 23 fields	0.720
Top 150 words	81	31.03	+ 19 fields	0.798
Top 200 words	92	35.25	+ 11 fields	0.818
Top 250 words	106	40.61	+ 14 fields	0.852
Top 300 words	119	45.59	+ 13 fields	0.882
whole dataset	261	100		

Table 8_1. English Elicited *chocolate*: most frequent 300 words, tagged with Wmatrix

	USAS tags (n.)	USAS tags (%)	tag increase	Spearman's rho ($p < 0.01$)
Top 50 words	38	14.18	+ 38 fields	0.588
Top 100 words	58	21.64	+ 20 fields	0.714
Top 150 words	73	27.24	+ 15 fields	0.755
Top 200 words	91	33.96	+ 18 fields	0.768
Top 250 words	112	41.79	+ 21 fields	0.865
Top 300 words	123	45.90	+ 11 fields	0.886
whole dataset	268	100		

Table 8_2. English Elicited *wine*: most frequent 300 words, tagged with Wmatrix

Column one shows the number of most frequent (Top) content words considered; column two indicates the number of USAS tags retrieved at each threshold; column three shows the number of USAS tags retrieved at each threshold, as a percentage of the number of USAS tags present in the whole dataset;[4] column four highlights the number of new tags entering the list at each threshold; finally, column five shows the result of a quantitative comparison between the USAS tags at each threshold and the whole dataset.

A comparison between Tables 8_1 and 8_2 to Tables 7_1 and 7_3 in Chapter 7 – the latter illustrating the results of same type of analysis performed using manual coding – shows an interesting picture: although the percentage of semantic categories in the top 300 words is lower when USAS tagging is applied (about 44.5% vs. 68-70%), Spearman's test results are very similar, *rho* being always in the strong range.

8.2.2 Web data

The results of the comparison between the most frequent content words in the Web word lists and the corresponding whole datasets are summarised in Tables 8_3 and 8_4 below.

Since results at the 300th word showed constant gradual increases, but on the whole were still rather low, I extended the analysis to the 450th content word.

[4] Grammatical categories Z4-Z99 in the USAS tagset were excluded from the counts.

	USAS tags (n.)	USAS tags (%)	tag increase	Spearman's rho ($p < 0.01$)
Top 50 words	28	6.09	+ 28 fields	0.369
Top 100 words	45	9.78	+ 17 fields	0.396
Top 150 words	68	14.78	+ 23 fields	0.435
Top 200 words	87	18.91	+ 19 fields	0.478
Top 250 words	104	22.61	+ 17 fields	0.526
Top 300 words	116	25.22	+ 12 fields	0.542
Top 350 words	123	26.74	+ 7 fields	0.555
Top 400 words	136	29.57	+ 13 fields	0.578
Top 450 words	144	31.30	+ 8 fields	0.576
whole dataset	460	100		

Table 8_3. English Web *chocolate*: most frequent 300 words, tagged with Wmatrix

	USAS tags (n.)	USAS tags (%)	tag increase	Spearman's rho ($p < 0.01$)
Top 50 words	31	6.74	+ 31 fields	0.372
Top 100 words	49	10.65	+ 18 fields	0.487
Top 150 words	69	15.00	+ 20 fields	0.554
Top 200 words	84	18.26	+ 15 fields	0.604
Top 250 words	99	21.52	+ 15 fields	0.640
Top 300 words	113	24.57	+ 14 fields	0.639
Top 350 words	124	26.96	+ 11 fields	0.706
Top 400 words	136	29.57	+ 12 fields	0.730
Top 450 words	143	31.09	+ 7 fields	0.734
whole dataset	460	100		

Table 8_4. English Web *wine*: most frequent 300 words, tagged with Wmatrix

Not unexpectedly, the results obtained with the two Web corpora are not as good as those obtained with the elicited datasets. In fact, the percentage of semantic categories retrieved by the most frequent 300 content words with respect to the whole corpus is higher in the elicited datasets (about 44%), than in the Web datasets (about 25%), and so are correlation values (in the modest range for web corpora; in the strong range for elicited datasets). Extending the analysis to the 450th content word improved the results, in terms of percentage of categories retrieved, as well as correlation results.

Percentage-wise, the results are easily explained by the different sizes of the elicited and Web corpora. Indeed, 300 words correspond to about 3% of the elicited datasets, but less than 0.1% of the Web corpora. Correlation-wise, however, they confirm the hypothesis that the most frequent words in corpus are the most representative of the contents of the corpus.

8.3 Randomly sampled sub-corpora

The randomly sampled sub-corpora described in Section 7.4 proved highly representative of the whole dataset, by showing 79-94% of the semantic fields in the whole corpus, 97-100% of the highly conventionalised fields, and 94-98% of the cultural associations in the original datasets. Furthermore, Spearman's test ($p < 0.01$) highlighted very strong correlation between semantic field values in the randomly sampled sub-corpora and their corresponding datasets: for English *chocolate*, $r = 0.931$; for Italian *chocolate*, $r = 0.950$; for English *wine*, $r = 0.961$; and for Italian *wine*, $r = 0.935$.

Let us now see what happens if we adopt a different coding scheme, and also a different set of data.

8.3.1 Elicited data

The randomly sampled sub-corpora described in Chapter 7.4 were automatically tagged using Wmatrix and the USAS tagset, and the results were compared to those obtained with automatic tagging of the whole datasets they were extracted from. The results are summarised in Table 8_5.[5]

	USAS tags (n.)	**USAS tags (%)**	**Spearman's rho ($p < 0.01$)**
English *chocolate* random sample	179	66.79	0.867
English *chocolate* whole dataset	268	100	
English *wine* random sample	221	80.36	0.903
English *wine* whole dataset	275	100	

Table 8_5. Randomly sampled elicited sub-corpora, tagged with Wmatrix

As Table 8_5 shows, even by applying the USAS coding scheme, the randomly sampled sub-corpora proved highly representative of the corresponding original dataset, and much more so than the most frequent 300 content words in the wordlist. In fact, the randomly sampled corpora showed almost 67% and 80% of the USAS fields in the whole corpus, for English *chocolate* and English *wine* respectively, with very strong correlation results. However, as already noticed with the elicited data, the results in Table 8_5 are lower than those obtained with manual tagging (see Table 7_39 in Chapter 7), in particular as regards the percentage of semantic categories retrieved (66.79% vs. 84.09% for *chocolate*, 80.36% vs. 86.9% for *wine*); Spearman's Test results, on the other hand, are similar in two cases.

8.3.2 Web data

As a final experiment, the English *chocolate* and *wine* Web datasets were randomly sampled and the resulting sub-corpora were semantically tagged using Wmatrix. Given that, with small sized datasets such as the elicited ones, random samples in the 25-35% size range of the original dataset provided very good results, it is to be expected that for much larger datasets such as the Web ones 25% could be a more than suitable sampling limit. Thus, following the procedure described in Chapter 7, Section 7.4, which saw the use of a software programme for mathematical calculations to list a specific number of random positive integers within a given range, different for each corpus, two randomly sampled Web sub-corpora were created, including 2109 and 1836 sentences for *chocolate* and *wine* respectively. The results of automatic tagging, compared to automatic tagging of the corresponding whole Web datasets are summarised in Table 8_6.

[5] In Table 8_5 as well as in Table 8_6, grammatical categories Z4-Z99 in the USAS tagset were not excluded from the counts.

	USAS tags (n.)	USAS tags (%)	Spearman's rho ($p < 0.01$)
English *chocolate* random sample	416	91.03	0.972
English *chocolate* whole dataset	457	100	
English *wine* random sample	405	86.91	0.987
English *wine* whole dataset	466	100	

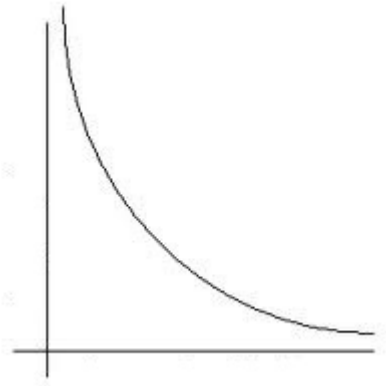
Table 8_6. Randomly sampled Web sub-corpora, tagged with Wmatrix

The randomly sampled sub-corpora, semantically analysed using the USAS tagset, proved highly representative of the Web corpus they were extracted from. In fact, they retrieved 86-91% of the semantic categories in the whole Web corpora and showed very strong range correlation results.

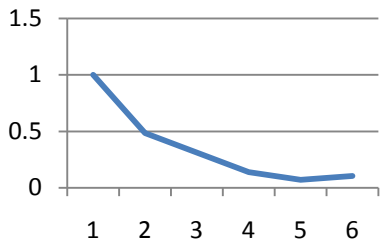
8.4 Mathematical progression of field increases: a Zipf-like curve?

In Chapter 7, when I first experimented with concordance reading and semantic classification of the most frequent 50/100/150/200/250/300 content words in the word list and noticed a dramatic decrease in the number of fields being retrieved, a Zipf-like distribution came to mind. Zipf's law, which has been found to describe the distribution of word frequencies in natural languages, such as English, but also in random text, declares that "the distribution of word frequencies […], if the words are aligned according to their ranks, is an inverse power law with the exponent very close to 1" (Wentan Li, 1992). This could be graphically represented as in Graph 8_1.

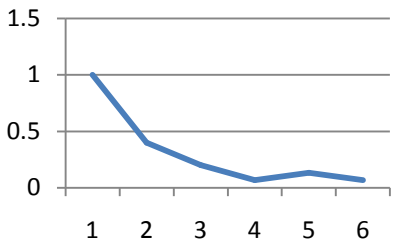
At this point in the research, after having experimented with a wider number of corpora and coding schemes, I have collected a reasonable number of examples of 'category increase progressions' which can be plotted and compared to each other, in order to decide whether Zipf's law can be called into play.



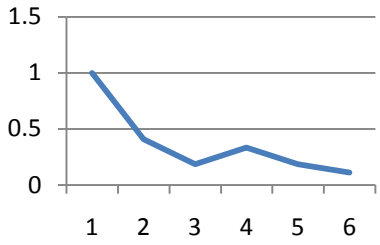
Graph 8_1. Graphic example of Zipf's distribution



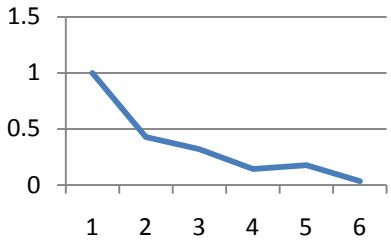
Graph 8_2. Data from Table 7_1



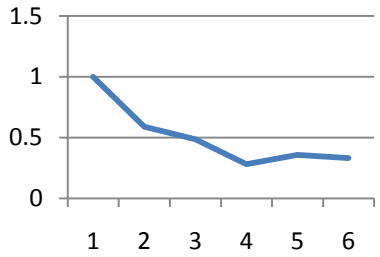
Graph 8_3. Data from Table 7_2



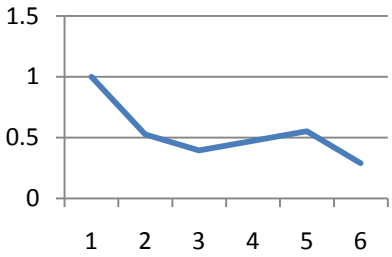
Graph 8_4. Data from Table 7_3



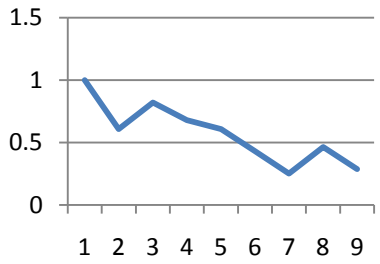
Graph 8_5. Data from Table 7_4



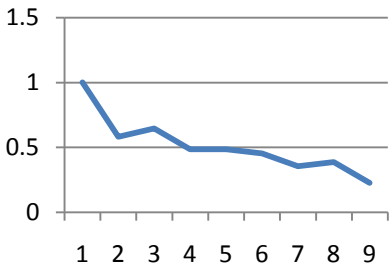
Graph 8_6. Data from Table 8_1



Graph 8_7. Data from Table 8_2



Graph 8_8. Data from Table 8_3



Graph 8_9. Data from Table 8_4

The 'field increase' data from Tables 7_1-7_4 (Chapter 7) and Tables 8_1-8_4 (Chapter 8) are plotted in Graphs 8_2-8_9. For an easier comparison, the data have been rescaled, by converting each set into percentages of the field increase value in rank 1.

These graphs suggests caution in making reference to Zipf's law. Indeed, only the curve in Graph 8_2 resembles that in Graph 8_1; in all the other graphs, one or more of the points clearly detaches from Zipf's curve. It should be said, however, that the number of data plotted is very small, and probably not enough for a final decision on this matter.

8.5 Conclusions

The current chapter repeated the experiments described in Chapter 7 – i.e. comparing a dataset to the most frequent content words in its wordlist, and to a sub-corpus randomly sampled from the same dataset, the latter being seen as possible shortcuts to the semantic analysis of the whole dataset – but used a different tagging scheme. When passing from manual coding to USAS tagging, differences were observed in the percentages of semantic categories retrieved. Indeed, with USAS tagging, both routes retrieved a smaller percentage of semantic categories. This is most certainly explained by the fact that the USAS dataset – "arranged in a hierarchy

with 21 major discourse fields expanding into 232 category labels" (Archer, Wilson, & Rayson, 2002, pp. 1-2), some of which are further subdivided into finer categories marked by a decimal point followed by a further digit, or "one or more 'pluses' or 'minuses' to indicate a positive or negative position on a semantic scale" (*ibid.*) – includes a higher number of categories and is much finer-grained than the tagset used for manual coding. Indeed, a similar phenomenon was noticed in Chapter 7 when comparing semantic field and conceptual domain results, i.e. a finer-grained scheme and a broader one.

Interestingly, however, although the percentage of semantic categories retrieved was lower when USAS tagging was applied (about 44.5% vs. 68-70% when analysing the most frequent words in the wordlist; 66.79% vs. 84.09% for *chocolate*, and 80.36% vs. 86.9% for *wine*, when analysing the random sub-corpus), in both cases Spearman's test results were very similar to those obtained with manual tagging, *rho* being always in the strong range.

Furthermore, the current chapter applied the USAS tagset to larger sets of data taken from the Web, and compared the whole dataset results to the most frequent content words in its wordlist, and to randomly sampled sub-corpora. The Web data showed that the results of the most-frequent-semantic-words analysis are dependent on corpus size. In fact, the top 300 word retrieved only about 25% of the semantic categories. Even extending to 450 the number of words considered, the percentage of categories retrieved was still very low (about 31%). Correlation values were in the medium range and showed constant linear increase, thus strengthening the hypothesis that the most frequent words in corpus are highly representative of the contents of the corpus.

On the other hand, the random sampling procedure seemed to be less sensitive to corpus size. The elicited random corpora – corresponding to 25-34% of the original datasets and including 3,527-4,603 running words, showed 66.8-80% of the semantic categories in the whole datasets and correlation values in the strong-very strong range (0.867 for *chocolate*, 0.903 for *wine*). The web random corpora – corresponding to 25% of the original datasets and including 73,780-89,901 running words, showed 86-91% of the semantic categories in the whole datasets and correlation values in the very strong range (0.972 for *chocolate*, 0.987 for *wine*).

Finally, in the light of the data in the present chapter as well those in Chapter 7, Section 8.4 investigated whether the distribution of semantic categories in the most frequent content items in the wordlist can be said to follow Zipf's law. My results invite caution in this respect. However, the distribution of semantic categories in the most frequent words in the wordlist is an issue which is worth of further investigation.