Chapter 10

Semantic associations of *chocolate*, and *wine* in general Web corpora

10.1 Introduction

The current chapter explores the possibility of using general Web corpora to highlight cultural semantic associations of the given node words (R.Q. 5 in my Research Questions list; see Chapter 1 or Chapter 5), by applying the manual coding adopted for the elicited data, and comparing the Web results to those of the elicited data.

As we have seen in Chapter 2, elicited data are not the only source of cultural information. Cultural and cross-cultural analyses have also been based on corpora, either general (e.g.: Leech & Fallon, 1992; Schmid, 2003) or specific (e.g.: Manca 2008); the use of Web corpora for cultural analysis, however, is still rather limited, despite their potential (see Chapter 3, Section 3.4).

In a pilot study to the current work, Bianchi (2007) compared a specialised Web corpus on *chocolate* created by manually selecting texts from the Internet – selection being based on three criteria: variety of sources; presumed production by native speakers; and presence of the key concept – to a large general corpus of about 10 million words created according to 'traditional' methods and criteria. Both corpora were in Italian. In each of the two corpora, concordances for lemma *cioccolato* were retrieved and classified in terms of semantic fields and conceptual domains of the node word.[1] The specialised corpus provided 1612 sentences with the node word; the general corpus, despite its size, only 849 sentences. Semantic analysis results showed a higher number of both semantic fields and conceptual domains in the specialised corpus (64 vs. 44, and 15 vs. 12 respectively). However, quantitative (i.e. frequency) differences between the two corpora were not statistically significant at the Mann-Whitney test, and decreased when moving from semantic fields to conceptual domains, that is to say from a more to a less fine-grained analysis. Finally, the differences in the number of semantic fields and conceptual domains were explained by the significantly different number of sentences retrieved in each corpus, as well as by relevant differences in the time-span covered by the two corpora (before 2001 for the general corpus; around 2003 for the specialised corpus).

[1] The terminology used in Bianchi (2007) is slightly different from the one adopted in the current work: 'semantic fields' were then called 'semantic contexts', while 'conceptual domains' were called 'conceptual fields'.

In the marketing field, Aggarwal, Vaidyanathan and Venkatesh (2009) used Google's application program interface (API) to retrieve from the Web sentences that included specific brand names. Subsequently they derived each brand's online positioning by using mutual information values of the adjectives accompanying brand names.

In all the cases above, though with different methods, the analyses were performed on the whole set of data, either in the form of its wordlist or the sentences including the node word. Interestingly, however, Chapters 7 and 8 have shown that alternative, shorter routes based on the most frequent words in the wordlist or on sampling procedures could be used to retrieve almost all of the semantic associations present in a corpus. In particular, the random sampling procedure proved to be the most suitable one with large corpora.

The current chapter analyses the semantic associations of *chocolate* and *wine* in the English and Italian Web datasets described in Chapter 5, Section 5.2.2.3 and compares them to the results of the elicited data (Chapter 6). The datasets were extracted from two large, general Web corpora (UKWAC and ITWAC), by automatically retrieving 10,000 sentences which included the key words under investigation, and purging the retrieved sentences of duplicates. This led to the creation of the following four datasets: the English *chocolate* Web dataset of 8436 sentences; the Italian *chocolate* Web dataset of 8352 sentences; the English *wine* Web dataset of 7343 sentences; and the Italian *wine* Web dataset of 8239 sentences.

For a precise comparison, the coding scheme used in Chapter 6 needs to be manually applied to the Web datasets. However, their size, which is about four times larger than that of the elicited datasets, makes manual semantic analysis time-consuming and prone to the risk of inconsistency. Consequently, manual coding will be applied to a sub-corpus created by random sampling, and the results of manual coding will be compared to the results of the elicited data (see Chapter 6), the latter being used as control groups.

10.2 Sampled Web sub-corpora: creation and coding

As already noted in Chapter 8, 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. Consequently, I decided that 25% could be a more than suitable sampling limit for the sampling of the much larger Web data.

For each of the four Web datasets, a sampled sub-corpus was created following the random sampling procedure used in Chapters 7 and 8. A software programme for mathematical calculations, Mathematica, was set to list a specific number of random positive integers within a given range, different for each corpus (2109 integers in the 1-8436 range for English *chocolate*; 2088 integers in the 1-8352 range for Italian *chocolate*; 1836 integers in the 1-7343 range for English *wine*; and 2060 integers in the 1-8239 range for Italian *wine*). The random numbers thus obtained were used to extract sentences from the Web datasets.

This produced four sub-corpora, each having a size corresponding to 25% of the original Web dataset: the English *chocolate* random Web sub-corpus, including 2109 sentences; the Italian *chocolate* random Web sub-corpus of 2088 sentences; the

English *wine* random Web sub-corpus of 1836 sentences; and the Italian *wine* random Web sub-corpus of 2060 sentences.

The sub-corpora thus created were manually coded at sentence level, by applying the semantic coding scheme described in the Codebook (see the Appendix), and were compared to the elicited datasets (see Chapter 6). Both qualitative and quantitative comparisons was performed, at the level of semantic fields and conceptual domains, the latter being superordinate, broader categories.

10.3 Inter-culture analysis

At the level of semantic fields, the randomly sampled Web sub-corpora provided the results in Table 10_1. In the table, the first column specifies the sub-corpus; the second column shows, percentage-wise, how many of the semantic fields in the corresponding elicited dataset were retrieved by the Web sub-corpus; the third and fourth columns show the percentage of high conventionalisation fields (H Cnv) and of semantic associations (H+M Cnv) covered by the fields in the Web sub-corpus; finally, column five reports the results of a quantitative comparison between the Web sub-corpora and the corresponding elicited datasets performed by applying Spearman’s Rank Correlation Coefficient. Percentage values are rounded to the first decimal.

	Overall fields (%)	**H Cnv (%)**	**H+M Cnv (%)**	**Spearman’s Rho ($p < 0.01$)**
English *chocolate* random sub-corpus	97.7	97.1	98.3	0.587
Italian *chocolate* random sub-corpus	91.9	100	98.2	0.541
English *wine* random sub-corpus	94.0	94.3	94.2	0.587
Italian *wine* random sub-corpus	92.9	97.8	96.3	0.593

Table 10_1. Random Web sub-corpora: Semantic fields

As the table illustrates, the English *chocolate* Web sub-corpus shows 97.7% of the semantic fields in the English *chocolate* elicited dataset and, most importantly, over 97% of the fields with a high level of conventionalisation and over 98% of the cultural associations. The Italian *chocolate* Sampled Web sub-corpus includes almost 92% of the total number of fields in the Italian *chocolate* elicited dataset, 100% of the highly conventionalised fields and almost over 98% of the cultural associations. The English *wine* Web sub-corpus retrieved 94% of the total number of fields in the English *wine* elicited dataset, and over 94% of the highly conventionalised fields and cultural associations. Finally, the Italian *wine* sub-corpus showed almost 93% of the total number of fields in the Italian *wine* elicited dataset, almost 98% of the highly conventionalised fields and over 96% of the semantic associations.

From a qualitative perspective, the picture emerging from the random sampling experiment above is highly satisfactory, as the randomly sampled Web sub-corpora retrieved about 92-98% of the fields present in the elicited data and, more importantly, 94-100% of the fields with high conventionalisation, and 94-98% of the cultural associations. From a quantitative one, however, correlation results were always in the modest range.

At the level of conceptual domains comparisons provided excellent results at both qualitative and quantitative levels. Indeed, all four Web random sub-corpora showed 100% of the conceptual domains, and correlation results were all in the strong range, or higher (with p < 0.01, r = 0.946 for English *chocolate*; r = 0.939 for Italian *chocolate*; r = 0.870 for English *wine*; r = 0.892 for Italian *wine*). This type of result was expected, given what had already been noticed about coding scheme granularity, in the previous chapters.

Finally, the random Web sub-corpora retrieved a limited number of fields and domains which are not present in the corresponding elicited datasets. These are listed in Table 10_2, along with the number of extra fields and domains retrieved by the random Web sub-corpora, and the difference in size between each randomly sampled sub-corpus and its corresponding elicited counterpart (columns Extra sentences and Extra size). The name of the extra fields/domains is also shown, along with the corresponding rank in decreasing order of frequency.

		Extra fields / domains		**Extra sentences**	**Extra size**
	n.	**field / domain**	**rank**	**n.**	**%**
Chocolate English Web random sub-corpus	Fields: 5	FET-genuine	21	223	+ 11.8
		CUL-studying/intellect	34		
		F-serving	37		
		FE-competitiveness	41		
		P-age	45		
		Total field ranks	*49*		
	Domains: 0				
Chocolate Italian Web random sub-corpus	Fields: 6	E- law	7	233	+ 14.5
		FET- price	28		
		F- serving	31		
		P-posh	41		
		E-work	43		
		E-holidays	44		
		Total field ranks	*49*		
	Domains: 0				
Wine English Web random sub-corpus	Fields: 7	E-history	30	-102	-5.3
		EN-tech	31		
		P-royalty	37		
		FE-competitiveness	41		
		S-sports	41		
		FET-energy	44		
		FE-loneliness	46		
		Total field ranks	*48*		
	Domains: 1	Sports	13		
		Total domain ranks	*13*		
Wine Italian Web random sub-corpus	Fields: 4	P-people	42	487	+ 31
		E-war	43		
		EN-animals	43		
		LD-theft	43		
		Total field ranks	*43*		
	Domains: 0				

Table 10_2. Summary of extra fields and domains retrieved by the random Web sub-corpora

The Web sub-corpora are 11%-31% larger that their elicited counterparts, with the noticeable exception of English *wine* which is actually smaller by about 5%. Interestingly, the latter sub-corpus shows the highest number of extra fields (with as many as 7) and even one extra domain. On the other hand, the Italian *wine* random sub-corpus retrieved the smallest number of extra fields (only 4), despite it is 31% larger than its elicited counterpart. Such a picture suggests that the presence of extra fields and domains in the sampled sub-corpora is not due to differences in corpus size. So, what could be the reasons for the constant presence of extra fields in the Web sub-corpora? As was the case in Bianchi (2007), the time when the Web corpora and the elicited datasets were collected may still play a role: the wacky corpora were developed between 2005 and 2007 (Baroni, Bernardini, Ferraresi, & Zanchetta, 2008), while the questionnaires were distributed in 2009. Furthermore, a look at the actual contents of the corpora[2] might provide us with further hints.

The elicited corpora include mostly short and easy sentences written by individuals in a given context (the questionnaire and the place where it was distributed) and seen along with their co-text (the sentence preceding and following the one undergoing tagging). In their answers, the respondents talk about the given node word, making reference to themselves, family, or friends. Finally, a few well-known set phrases or proverbs are sometimes reported.

Conversely, the Web sub-corpora include sentences extrapolated from wider text, the latter being no longer visible. Indeed, coding the Web corpora proved more difficult than coding the elicited data, as it was frequently necessary to go over the same sentence more than once, before one could be sufficiently sure of its meaning. Quite frequently, in the Web sub-corpora, the node word is not the topical element of the sentence, as, for example, when recipes are provided and *chocolate* is only one of the many ingredients, or when a place or event which accidentally included the presence of *chocolate* is described. Finally, many sentences were characterised by a distinctive marketing or legal flavour, which suggests that a relatively large part of the Web corpora consists of advertising text written by manufacturers, dealers, or restaurants, as well as governmental decrees. This might explain the extra fields LAW, PRICE and WORK (in the Italian *chocolate* Web corpus), GENUINE and COMPETITIVENESS in the English *chocolate* one, as well as TECH, and COMPETITIVENESS in the English *wine* one.

10.3.1 Semantic field ASSESSMENT

The results of the analysis of the semantic field ASSESSMENT in the randomly sampled Web sub-corpora are reported in Tables 10_3 and 10_4, and graphically illustrated in Figure 10_1, in direct comparison with the results in the elicited datasets. In the tables, the numerical values are percentages of the total number of sentences in each sub-corpus.

The results of the semantic field ASSESSMENT in the four Web sub-corpora seem in keeping with those in their corresponding elicited datasets (Table 7_15 in Chapter 7). In fact, the latter showed a majority of positive sentences, a somehow smaller number of neutral sentences, followed by a yet smaller number of negative sentences, and a few undecided sentences. This same ranking can be seen in Table 10_4, where positive assessment precedes neutral assessment, which in turn precedes negative as well and undecided assessment results.

[2] Manual tagging implied reading the datasets sentence by sentence and provided a general idea of the corpus contents, though at an intuitive level.

	Positive	Negative	Neutral	Undecided
English Web random sub-corpus	48.84	8.58	36.75	5.83
Italian Web random sub-corpus	46.62	7.46	41.78	4.14
English elicited dataset	53.92	19.03	26.35	0.69
Italian elicited dataset	54.21	11.85	32.75	1.19

Table 10_3. *Chocolate*: Semantic field ASSESSMENT

	Positive	Negative	Neutral	Undecided
English Web random sub-corpus	60.51	5.17	28.16	6.15
Italian Web random sub-corpus	47.54	3.45	45.55	3.45
English elicited dataset	46.00	19.60	27.69	6.70
Italian elicited dataset	53.59	14.49	29.62	2.29

Table 10_4. *Wine*: Semantic field ASSESSMENT

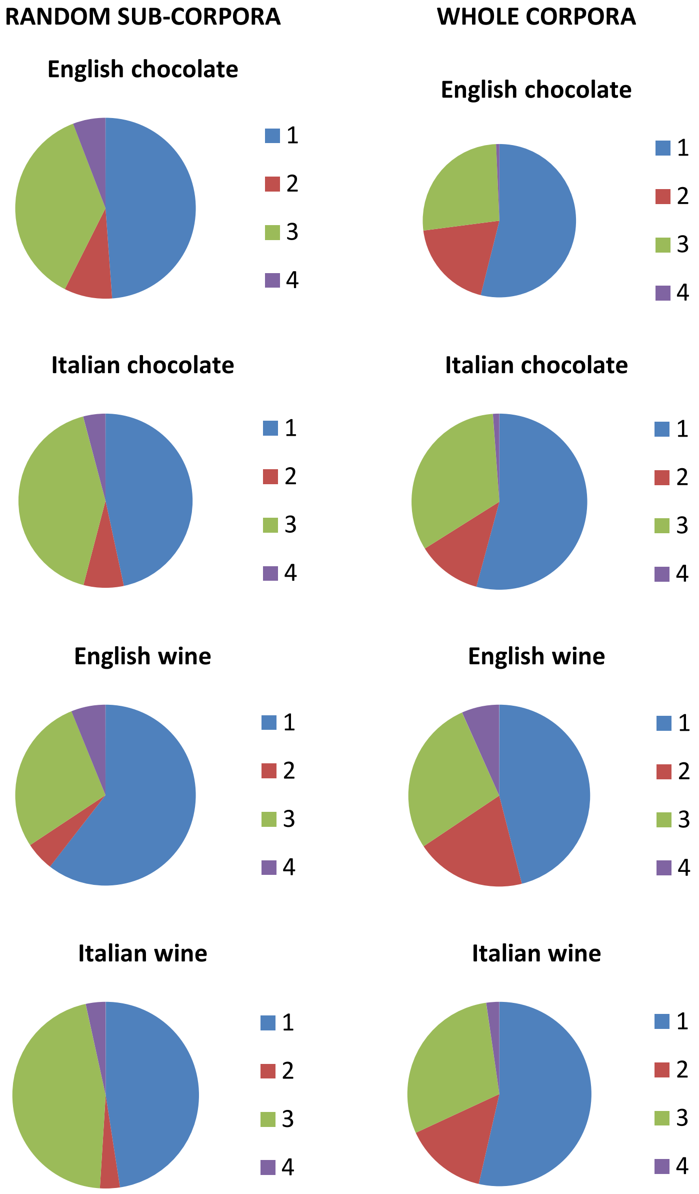

Figure 10_1. ASSESSMENT field results:
random sub-corpora vs. whole elicited datasets

Interestingly, however, the Web sub-corpora systematically show percentages of negative assessment which are remarkably lower than those in the elicited datasets. This is probably connected to the already noticed marketing flavour of the Web data.

10.4 Cross-cultural comparison

In Chapter 6, cross-cultural comparisons were performed between the English and Italian elicited datasets about *chocolate* and *wine*. The comparative procedure adopted consisted in quantitative correlation analysis using Spearman's Rank Correlation Coefficient, followed by Welch *t* Test for Independent Samples, introduced to try and understand where the cultural differences lied. At the level of semantic fields, Spearman's test showed strong positive correlation between the English and Italian datasets, with Spearman's Rho equal to 0.719 ($p < 0.01$) and to 0.735 ($p < 0.01$) for *chocolate* and *wine*, respectively. The t-test suggested that, at the level of semantic fields, the Italians seem to distinguish themselves from the British for their more frequent matching of *chocolate* to the following concepts: BAKERY/COOKING; RECIPE; DIETING; MEDICINE; BEAUTY; HISTORY; NICE/PLEASANT/PLEASURE; CHILDREN; FAMILY; STUDYING/INTELLECT; QUALITY/TYPE; GENUINE. On the other hand, more prominent for the English than for Italians appeared to be: WOMEN, and PRICE. As regards *wine*, the following semantic fields emerged as distinctively more prominent for Italians than for the English: BAKERY/COOKING; EVENT; WOMEN; NATURE; ARTISTIC PRODUCTION; QUALITY/TYPE; QUANTITY; GENUINE; PRICE. On the other hand, more prominent for English than for Italians were: PRODUCT/SHAPE; DRINK; MANUFACTURING; RECIPE; LANGUAGE; CONFIDENCE; DESIRE, NICE/PLEASANT/PLEASURE; MEN, FRIENDSHIP; POSH; SHARING/SOCIETY; PEOPLE; and STUDYING/INTELLECT.

At the level of conceptual domains, cross-cultural comparisons highlighted very few differences (r = 0.939 for p < 0.01 for *chocolate*; r = 0.942 for p < 0.01 for *wine*). T-test results were not always easy to interpret, but seemed to highlight domain CULTURE as the only conceptual domain that clearly distinguishes the Italians from the English in thinking about *chocolate*, and domains FEELINGS and CULTURE as the only conceptual domains that clearly distinguish the Italians from the English as regards *wine*.

Let us now see how the Web corpora fare in a cross-cultural comparison.

The English Web random sub-corpora were compared to their Italian Web random counterparts, at the level of semantic fields and conceptual domains. Quantitative comparisons were performed by applying Spearman's Rank Correlation Coefficient. Spearman's results showed strong correlations at the level of semantic fields (for $p < 0.01$, $r = 0.846$ for *chocolate* and $r = 0.894$ for *wine*) and very strong correlations at the level of conceptual domains (for p < 0.01, r = 0.964 for *chocolate* and r = 954 for *wine*). These results are in keeping with the ones obtained with the elicited data, where Spearman's rho was equal to 0.719 and 0.735 for *chocolate* and *wine*, respectively, at the level of semantic fields and to 0.939 e 0.942 at the level of conceptual domains.

Unfortunately, T-test analysis could not be applied to my Web data, as the Web sentences could not be grouped according to subject/author or website and were to be considered as individual instances from different authors/websites.

10.5 Concluding remarks

In line with previous studies which used corpora in cultural analyses, and as a follow-up to a preliminary experiment which suggested the possible use of general Web corpora to highlight cultural associations of a given node word, the current chapter applied manual tagging to four datasets created from general Web corpora following a random sampling procedure. The qualitative and quantitative results were compared to those obtained with elicited data, at the level of semantic fields, as well as conceptual domains.

In all the four cases, the sampled Web corpora retrieved over 90% of the semantic fields with high conventionalisation and of the cultural associations attested in the corresponding elicited datasets. However, the corpora also retrieved most of the low conventionalisation fields, along with a few extra fields whose conventionalisation level is not known, although one could speculate that – being those fields totally absent in the elicited corpora – they could be classified as having low conventionalisation. The same could be said for conceptual domains, as the Web sub-corpora retrieved all of the domains in the corresponding elicited datasets, which means 100% of the domains with high, medium or low conventionalisation; furthermore, the English *wine* sampled Web corpus retrieved also one extra domain (the only Codebook domain which had not been attested in the English *wine* elicited dataset).

The ASSESSMENT field matched, in ranking, the results of the elicited datasets, with positive assessment preceding neutral assessment, which in turn preceded negative as well and undecided assessment results. Interestingly, however, the Web sub-corpora systematically showed percentages of negative assessment which are remarkably lower than those in the elicited datasets, a result which is at least partly connected to the ‘marketing flavour’ of large part of the texts in the Web corpora – the latter being also a probable explanation for about 30% of the semantic fields present in the Web corpora, but absent in the corresponding elicited datasets.

Finally, correlation results were all in the modest range for semantic fields, and in the strong range, or higher, for conceptual domains – a similar improvement in correlation results when passing from a fine-grained to a broader coding scheme having been systematically attested in all the comparisons performed in the previous chapters.

Consequently, comparisons between the Web corpora under analysis and the elicited data suggest that large general Web corpora can be considered representative of the cultural associations of a node word. In fact, randomly sampled Web subsets of only 1800-2000 sentences, included all the relevant cultural associations of the node word. Furthermore, when the coding scheme adopted was broad and included few categories, the general Web corpora appeared to be representative not only at a qualitative level, but also at a quantitative one.

Unfortunately, as noticed in Chapter 6, we cannot rely on frequency alone to establish conventionalisation. Only the very highest ranks in the frequency list are systematically occupied by low conventionalisation fields, and only the very lowest ranks are systematically occupied by high conventionalisation ones. Any other position in the list can hardly tell us something about conventionalisation level.

Consequently, if we had only Web data, and no control elicited data, we would have to assess the conventionalisation level of each field/domain by applying an evenness index, as done in Chapter 6, in order to establish which of the retrieved semantic fields/conceptual domains can be safely considered cultural associations. A fundamental pre-requisite for applying the evenness computation is the possibility to group the Web sentences according to subject/author or website. This – along with T-test analyses for cross-cultural comparisons – could not be done in the current work, because at the time when the Web data were retrieved, the Sketch Engine did not provide information about the website each text was taken from. The updated version of the Sketch Engine, however, does provide this type of information, and its users can now benefit from the possibility to assess the distribution of concordance lines across Web sites (i.e. authors).

Finally, no marked and systematic differences can be seen between the results of the English data vs. those of the Italian data (see Tables 10_1-10_4). Consequently, although I cannot altogether exclude that some of the texts in the English Web corpus were written by non-British natives or that some of the marketing texts in the corpus were created for a foreign audience, authorship and readership, which as – we saw in Chapter 3 – might be problematic issues when using English (but not the Italian) Web data do not seem to have had much influence on the results.