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

7.1 Introduction

Semantically coding full elicited sentences is not the only possible method for extracting cultural information from text. Corpus linguistic studies such as those by Leech and Fallon (1992), Muntz (2001), Oakes (2003), and Schmid (2003) – all summarised in Chapter 3 – have shown that wordlists are suitable tools for the analysis of cultural traits, and for cross-cultural comparisons. The procedure adopted by all these authors is based on (either manual or automatic) semantic analysis of the whole wordlist. However, manual semantic analysis of a complete wordlist is a highly complex and time-consuming task, while automatic analysis is only possible for those languages for which a semantic tagger exists – and, in the case of cross-cultural comparisons, for which the taggers in the different languages are based on the same semantic schemes.

Fleischer's theory (Fleischer, 1998, discussed in Chapter 2, Section 2.2.2), as well as the results obtained in Chapter 6, suggest the existence of a relationship between cultural associations, their level of conventionalisation and frequency of occurrence of the given associations. Consequently, as semantic associations are conveyed through words which, in turn, have a clear frequency distribution in the corpus, it seems reasonable to hypothesise that highly conventionalised cultural associations might appear through an analysis of the most frequent words in the corpus. Indeed, Pullman, McGuire, and Cleveland (2005, p.328) suggest using the most frequent words in a wordlist to identify the semantic categories for content analysis.

The current chapter explores the possibility of using only the most frequent words in the wordlist to highlight the same cultural traits that would emerge from the analysis of the whole corpus (R.Q. 3 in my Research Questions list, see Chapter 1 or Chapter 5). Such a possibility would represent a convenient shortcut to the desired results. In the current experiments, coding each dataset (composed of more than 1500 sentences) took me about a week and proved a rather challenging task, due to the efforts required for being consistent and coherent in the application of the coding scheme. I have not attempted manual coding of a whole wordlist, but I expect it to take about the same amount of time and effort. Coding a smaller portion of the dataset or wordlist would inevitably be less time-consuming, and less complex in terms of coding coherence and cohesion.

Three different routes will be explored in the current chapter, using the elicited datasets on *chocolate*, and *wine* described in Chapter 5 and semantically analysed in Chapter 6 at the level of semantic fields and conceptual domains – two hierarchical levels of semantic classification corresponding, respectively, to finer-grained and broader tagging schemes. The first route applies 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 uses the four most frequent content words to extract sentences from the manually coded dataset and create a sampled sub-corpus. Finally, the third route is based on random selection of sentences from the manually coded dataset, to create a random sub-corpus.

In all the cases, the results will be compared to the results of the whole datasets (see Chapter 6), the latter being used as control group.

7.2 Route one: using the most frequent words in the dataset

As a first experiment, I decided to apply manual semantic analysis to the most frequent 50/100/150/200/250/300 content words in the wordlist of in each elicited dataset.

7.2.1 Creating the wordlists

Using Wordsmith Tools 5 (Scott, 2008), frequency wordlists were created for each of the four elicited datasets described in Chapters 5 and 6. The datasets, two in English and two in Italian, are collections of sentences revolving around two given key words – *chocolate* and *wine* – and elicited from native speakers by means of questionnaires with sentence completion and sentence writing tasks. As explained in Chapter 5, Section 5.1.3, given that the first task in each questionnaire was a sentence completion exercise, the English and Italian datasets were saved in two different formats: Format 1 (F1) which includes the words given in the first six sentences; and Format 2 (F2), which does not include the given text. For generating wordlists, Format 2 (F2) of the datasets was used, in order to avoid quantitative biases in the frequency list, due to the given text in the sentence completions task.

Furthermore, stop-lists were applied, in order to automatically filter out highly frequent words which do not match any of the semantic categories considered in the Codebook, such as function words, as well as other non-desired words, such as the various forms of the key word itself, which were likely to appear among the most frequent items. In the current chapter, analyses are guided (and limited) by the semantic categories set in the Codebook. In fact, if while performing manual coding of the elicited datasets it was possible to update the coding scheme with any new semantic categories that the two coders deemed necessary, when looking at words in the wordlist this is no longer advisable, since the results of the wordlists will have to be compared to the manual semantic analysis of the elicited datasets (Chapter 6).

More specifically, a different stop-list was created for and applied to each dataset. The stop-lists used – adaptations of stop-lists created for computational linguistic projects[1] – include articles, prepositions, personal pronouns and adjectives, relative pronouns, conjunctions, adverbs of time and space, auxiliary verbs (in all their forms), modal verbs, and the various forms of the specific node word. Exceptions were made for those linguistic elements which matched one (or more) of the semantic categories considered in the coding scheme. Thus, the stop-lists do not include personal pronouns and adjectives *he, his, she, her, hers* (and their Italian counterparts), as they match semantic categories WOMEN and MEN, and modal verbs *want* and Italian *volere*, as these match semantic category DESIRE. Verbs *have* and *be*, which have a semantic meaning when indicating respectively possession or existence, were not treated as exceptions because the coding scheme considered does not include semantic categories matching those meanings.

7.2.2 Coding the most frequent content words in the wordlist

For each dataset, the most frequent content words in the frequency wordlist were individually matched to one or more of the semantic fields in the Codebook. For the specific reasons explained below, the following words were ignored:

- Thinking verbs (e.g.: *think – find – seem – know*) and declarative verbs (e.g. *say*), as they perform a modality/hedging function or a narrative function which are not relevant in the current semantic analysis.
- Words like *thing*, which are used to subsume a wide and unspecified range of referents.
- Verbs whose meaning depends on what follows (e.g.: *make – feel – come – use – go – come – give – put – help – keep – see – break*; e.g., make a cake = COOKING; makes me feel sick = HEALTH; makes me happy = HAPPINESS). This was in order to avoid duplicating the frequency of some semantic fields.
- The less frequent part of a compound word. The words that were part of a compound word were counted only once. For example, in compound words *ice-cream* and *fair-trade*, the root that appeared sooner in the frequency list (*cream; trade*) was kept, while the other one (*ice; fair*) was ignored. This was used to overcome the limits of a tool which does not recognize compound words and multiword units and was possible because I looked at all concordances.

Indeed, looking at concordances was necessary to overcome semantic issues, such as polysemy and homography, and also coding issues, such as distinguishing when words 'red'/*rosso* or 'white'/*bianco* were used to refer to a type of *wine* ('red *wine* is strong' or '*il vino rosso è più buono di quello bianco*' [red *wine* is nicer that white *wine*]), or to a colour ('*wine* can be white in colour'; '*quando penso al vino penso al colore rosso intenso*' [when I think of *wine* I think of a dark red colour]).

Concordances were generated for each word, and matching was done after reading through all the concordance lines. Consequently, for example, word *bicchiere* ('glass'), ranking fifth in the Italian *wine* wordlist, was matched to the following fields: QUANTITY, since 45% of concordance lines included the glass as a measure of quantity, as in *bevo mezzo bicchiere di vino al giorno* ('I drink half a glass of *wine* every day'), or *un bicchiere di vino basta per ubriacare* ('one glass of *wine* is enough

[1] The original English stop-list is available at http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop; the Italian one at http://snowball.tartarus.org/algorithms/italian/stop.txt.

to get you drunk'); and SERVING, as 3.5 % of concordance lines referred to the glass as the ideal serving object, as in *per bere il vino bisogna avere il bicchiere giusto* ('drinking *wine* requires the right type of glass'). It can be noticed, in this example, that a good 51.5 % of concordance lines was ignored: in fact, in all the remaining sentences word *bicchiere* appeared because it is the usual way to drink *wine* and did not seem to be used to indicate quantity. Cases belonging to this category included, for instance, *gradisci un bicchiere di vino?* ('would you like a glass of *wine*?'), where saying 'a glass of *wine*' is tantamount to saying 'some *wine*'.[2] Having to ignore concordance lines, however, was a very rare circumstance.

Other circumstances where those when a word in the list had a clear sense, but its meaning did not fit any of the semantic fields in the Codebook. These cases were classified as OTHER, and include, for instance, the following words: 'famous'; *particolare* ('specific', 'peculiar'), and *effetti* (plural noun 'effects'). For these cases, the possibility of creating a new category was considered, but disregarded, for two reasons: a practical one, connected to the fact that adding a new category would imply re-tagging the whole corpus; and a theoretical one, based on the idea that a semantic association which did not seem salient when reading the whole corpus would probably be a minor one, at least in terms of frequency.[3]

Finally, content words having specific evaluative meaning were classified as POSITIVE ASSESSMENT or NEGATIVE ASSESSMENT. The POSITIVE- and NEGATIVE-ASSESSMENT categories will be discussed separately from the other semantic fields, in dedicated sections.

This process of concordance reading and semantic classification went on till the limit of 300 useful words was reached. Indeed, it was noticed that at the 300th most frequent word, raw frequency was actually very low (2 or 3 hits), and the number of new fields being retrieved had dramatically decreased in a fashion that seemed very close to a Zipf-like trend, as Tables 7_1-7_4 show. (The mathematical progression of the data in the tables will be analysed in Chapter 8, in the light of a wider number of examples).

Finally, the semantic categories resulting from the analysis were compared to those in the whole elicited corpus, the latter being used as a control group.

7.2.3 Semantic fields analysis at different thresholds

The results of the analysis of semantic fields at different thresholds are provided in Tables 7_1-7_4. Column one shows the number of most frequent (Top) words considered; column two indicates the overall percentage of fields covered. Columns three and four show the percentage of highly conventionalised fields (H Cnv) and cultural associations (H+M Cnv) covered. Finally, the last column summarizes field increase in passing from one threshold to the next. Percentages are rounded to the second decimal.

[2] Had the speaker wanted to underline quantity, s/he would have used modifier 'one' instead of 'a'.

[3] Had verbs *have* and *be* been not included in the stop-list, they would have fallen in category OTHER and eventually disregarded.

Matched Words	Overall fields (%)	H Cnv (%)	H+M Cnv (%)	Field increase
TOP 50	32.95	48.57	45.76	+ 29 fields
TOP 100	48.86	74.29	66.10	+ 14 fields
TOP 150	57.95	82.86	77.97	+ 9 fields
TOP 200	62.50	88.57	83.05	+ 4 fields
TOP 250	64.77	91.43	86.44	+ 2 fields
TOP 300	68.18	91.43	86.44	+ 3 fields

Table 7_1. *Chocolate* English wordlist:
Semantic fields analysis at different thresholds

Matched Words	Overall fields (%)	H Cnv (%)	H+M Cnv (%)	Field increase
TOP 50	34.88	59.38	52.73	+ 30 fields
TOP 100	48.84	71.88	67.27	+ 12 fields
TOP 150	55.81	84.38	76.36	+ 6 fields
TOP 200	58.14	84.38	78.18	+ 2 fields
TOP 250	62.79	87.50	83.64	+ 4 fields
TOP 300	65.12	90.63	87.27	+ 2 fields

Table 7_2. *Chocolate* Italian wordlist:
Semantic fields analysis at different thresholds

Matched Words	Overall fields (%)	H Cnv (%)	H+M Cnv (%)	Field increase
TOP 50	31.76	51.43	46.15	+ 27 fields
TOP 100	44.71	65.71	63.46	+ 11 fields
TOP 150	50.59	74.29	69.23	+ 5 fields
TOP 200	61.18	85.71	82.69	+ 9 fields
TOP 250	67.06	91.43	88.46	+ 5 fields
TOP 300	70.59	94.29	94.23	+ 3 fields

Table 7_3. *Wine* English wordlist:
Semantic fields analysis at different thresholds

Matched Words	Overall fields (%)	H Cnv (%)	H+M Cnv (%)	Field increase
TOP 50	28.57	53.33	48.15	+ 28 fields
TOP 100	46.43	71.11	62.96	+ 12 fields
TOP 150	57.14	84.44	77.78	+ 9 fields
TOP 200	61.90	84.44	81.48	+ 4 fields
TOP 250	67.86	86.67	85.19	+ 5 fields
TOP 300	69.95	86.67	87.04	+ 1 field

Table 7_4. *Wine* Italian wordlist:
Semantic fields analysis at different thresholds

It must be clarified that these tables do not consider semantic field OTHER – used as a bin category for all those content words with no direct match to any of the Codebook categories – and semantic field ASSESSMENT. The latter, in fact, is completely different in nature from the other semantic fields, and will be treated separately (see Section 7.3.1.3).

To sum up, the most frequent semantic fields appeared as soon as in the top (i.e. most frequent) 50 words. Furthermore, analysis of the distribution of fields across respondents carried out using Molinari’s evenness index (see Chapter 6, Section 6.2.1) showed that most of the fields emerging in the top 300 words can be considered culturally determined, in Fleischer’s framework of reference. In fact, the top 300 content words – though covering only 65-70% of the total number of semantic fields in the Codebook – highlighted about 86-94% of the highly conventionalised fields, and 86-94% of high plus medium conventionalisation fields or ‘cultural associations’. An in-list comparison between the percentage of highly conventionalised fields and cultural associations, however, shows a lower percentage of the latter. This applies to all cases, except Italian *wine*, which is probably explained by the Italian *wine* dataset unique distribution of fields across conventionalisation levels (see Table 6_10 in Chapter 6).

Finally, the semantic fields emerging from the most frequent 300 words were quantitatively compared to the fields in the whole dataset. Percentage frequency of each of the words considered was distributed across the relevant semantic fields. This eventually led to establishing percentage values of the semantic fields emerging from the top 300 words (Tables 7_5-7_8). The latter were then correlated with field percentage mean values across respondents as they emerged from the analysis of the whole dataset (see Tables 6_1 and 6_2, in Chapter 6). Correlation was performed by applying Spearman’s Rank Correlation Coefficient (see Chapter 6, Section 6.2.2).

Semantic field	**%**	**Semantic field**	**%**	**Semantic field**	**%**
F-food	1.94	P-children	0.33	FE-relax	0.09
F-product/shape	1.74	C-gift	0.31	P-sharing/society	0.09
comparison	0.99	F-manufacturing	0.27	FE-seduction	0.08
FE-desire	0.85	FET-physical properties	0.26	H-dieting	0.08
FE-happiness	0.85	P-family	0.23	FE-guilt	0.07
F-drink	0.78	FET-price	0.22	FE-unpleasant	0.07
F-recipe	0.76	E-religion	0.21	FE-love	0.06
F-bakery/cooking	0.75	FET-sweet	0.20	FE-sex	0.06
F-composition	0.75	FE-comfort	0.17	FET-package	0.06
FET-taste/smell	0.74	FET-energy	0.17	H-body	0.06
G-geo locations	0.71	EN-tech	0.15	F-serving	0.05
H-beauty	0.70	FE-mood	0.15	FET-genuine	0.05
E-transaction	0.69	FE-senses	0.15	P-friendship	0.05
FET-quantity	0.60	E-time	0.14	EN-dirt	0.04
E-event	0.57	FET-colour	0.14	EN-house	0.04
P-people	0.53	FE-nice	0.13	G-spreading	0.03
P-women	0.51	H-medicine	0.13	P-age	0.03
FE-passion	0.49	LD-drugs & addiction	0.11	CUL-culture	0.02
P-men	0.43	E-language	0.10	E-history	0.02
FET-quality/type	0.42	L-existence	0.10	FE-bribing	0.02
CUL-artistic production	0.38	E-fair trade	0.09		
H-health	0.35	EN-animals	0.09		

Table 7_5. English *chocolate* wordlist: Semantic fields in top 300 content words

Semantic field	%	Semantic field	%	Semantic field	%
F-food	2.26	FE-mood	0.31	FE-happiness	0.11
FET-quality/type	1.59	F-drink	0.27	LD-drugs & addiction	0.11
FET-taste/smell	0.98	FET-colour	0.26	FE-seduction	0.10
F-bakery/cooking	0.89	H-dieting	0.26	FE-memory	0.09
F-product/shape	0.78	H-health	0.25	P-people	0.09
FE-desire	0.61	H-beauty	0.23	FET-genuine	0.08
FET-quantity	0.57	FE-senses	0.21	CUL-studying/intellect	0.06
F-recipe	0.56	H-body	0.21	P-friendship	0.06
FE-passion	0.49	FET-sweet	0.20	P-men	0.06
P-children	0.46	FET-physical properties	0.19	P-age	0.05
F-composition	0.45	P-women	0.19	FE-guilt	0.04
G-geo locations	0.44	E-language	0.18	C-party	0.03
E-event	0.40	FET-energy	0.18	CUL-culture	0.03
CUL-artistic production	0.39	C-gift	0.17	EN-dirt	0.03
F-manufacturing	0.38	E-transaction	0.14	EN-nature	0.03
FE-nice/pleasant/pleasure	0.37	H-medicine	0.14	FE-sex	0.03
comparison	0.36	L-existence	0.14	EN-house	0.02
P-family	0.35	E-history	0.12	FE-peace	0.01
E-time	0.32	G-spreading	0.12		

Table 7_6. Italian *chocolate* wordlist: Semantic fields in top 300 content words

Semantic field	%	Semantic field	%	Semantic field	%
FET-quality/type	3.14	F-bakery/cooking	0.35	P-sharing/society	0.11
F-drink	2.42	FE-happiness	0.33	FET-sweet	0.08
G-geo locations	1.16	P-friendship	0.29	E-language	0.07
FET-taste/smell	1.04	F-storage	0.28	CUL-culture	0.06
comparison	0.99	FE-passion	0.25	EN-dirt	0.06
F-serving	0.88	FE-posh	0.25	FE-nice/pleasant/pleasure	0.06
FET-quantity	0.84	E-event	0.24	FE-seduction	0.06
E-excessive drinking	0.81	H-medicine	0.23	L-existence	0.06
F-food	0.79	FE-relax	0.22	P-age	0.06
FET-price	0.75	C-gift	0.20	E-driving	0.04
F-product/shape	0.66	E-religion	0.20	FE-love	0.04
E-time	0.59	F-manufacturing	0.20	P-children	0.04
F-composition	0.57	P-men	0.20	CUL-artistic production	0.03
E-transaction	0.48	G-spreading	0.18	E-work	0.03
P-people	0.46	C-party	0.14	FE-memory	0.03
P-family	0.42	FET-genuine	0.13	FET-packaging	0.03
FE-desire	0.41	F-recipe	0.12	I-fantasy/magic	0.03
H-health	0.38	FET-colour	0.12	EN-nature	0.01
FET-physical properties	0.37	FE-comfort	0.11	FE-mood	0.01
P-women	0.37	H-body	0.11	FE-senses	0.01

Table 7_7. English *wine* wordlist: Semantic fields in top 300 content words

Semantic field	%	Semantic field	%	Semantic field	%
F-drink	1.80	FE-confidence	0.26	C-party	0.12
G-geo locations	1.19	FET-quality/type	0.26	FE-passion	0.10
FET-taste/smell	1.00	FET-quantity	0.24	FET-price	0.10
F-recipe	0.96	CUL-culture	0.22	C-gift	0.09
F-manufacturing	0.87	P-men	0.21	E-driving	0.07
F-food	0.82	FE-nice/pleasant/pleasure	0.21	FE-mood	0.07
P-friendship	0.78	G-spreading	0.21	EN-dirt	0.06
FET-genuine	0.74	P-family	0.20	FE-desire	0.05
E-language	0.53	CUL-artistic production	0.19	L-existence	0.05
E-event	0.51	E-work	0.19	FE-love	0.03
comparison	0.43	H-medicine	0.19	FE-seduction	0.03
H-health	0.42	E-transaction	0.18	FE-unpleasant	0.03
E-time	0.41	FE-happiness	0.18	H-body	0.03
F-bakery/cooking	0.40	FET-sweet	0.18	LD-drugs & addiction	0.03
FET-physical properties	0.35	E-religion	0.16	P-age	0.03
F-composition	0.33	FET-packaging	0.15	P-posh	0.03
F-storage	0.32	EN-house	0.14	FE-memory	0.01
F-serving	0.29	EN-nature	0.14	P-sharing/society	0.01
CUL-studying/intellect	0.27	FET-colour	0.13		
E-excessive drinking	0.27	P-children	0.13		

Table 7_8. Italian *wine* wordlist: Semantic fields in top 300 content words

Despite only about 60% of the total number of semantic fields in the dataset emerged from the top 300 content words in the wordlist, and despite field ranking is different in the two cases, Spearman’s test showed strong correlation. In fact, Spearman’s results for the English *chocolate* semantic fields was r = 0.810; for Italian *chocolate*, r = 0.881; for English *wine*, r = 0.877; and for Italian *wine*, r = 0.859. In all cases p was lower than 0.01.

7.2.4 Conceptual domains analysis

Analysis of the top 300 content words in the frequency list was performed also at the level of conceptual domains – a superordinate semantic classification – and results were compared to domains in the whole dataset (see Tables 6_4 and 6_11, in Chapter 6).

Table 7_9 shows percentage results in the wordlists. R stands for rank. Cnv shows the conventionalisation level of that domain in the whole dataset (Chapter 6). Bold signals the absence of that particular domain in the sampled sub-corpus. Domains are listed in alphabetical order.

Domain	***Chocolate* Eng.**			***Chocolate* It.**			***Wine* Eng.**			***Wine* It.**		
	%	**R**	**Cnv**	**%**	**R**	**Cnv**	**%**	**R**	**Cnv**	**%**	**R**	**Cnv**
Ceremony	0.31	11	L	0.51	9	H	0.34	9	L	0.18	10	M
Comparison	0.99	6	L	0.26	10	H	0.99	6	M	0.06	11	M
Culture	0.40	9	M	1.00	7	H	0.09	10	L	1.02	6	H
Environment	0.34	10	M	0.08	13	M	0.07	11	M	0.34	8	M
Events	2.24	5	H	1.08	6	M	3.44	3	M	2.15	3	M
Features	2.98	4	M	3.84	1	M	6.30	1	M	1.50	4	M
Feelings & emotions	4.57	3	M	3.41	3	M	2.35	5	M	2.17	2	H
Food	6.26	1	M	3.56	2	M	3.48	2	M	2.79	1	M
Geo	0.58	8	H	0.58	8	M	0.65	7	H	1.40	5	M
Health &Body	0.60	7	M	1.09	5	M	0.38	8	M	0.24	9	M
Imagination	**0.00**		M	0.03	14	M	0.03	13	NC	**0.00**		L
Life	0.10	13	M	0.14	11	L	0.06	12	M	0.05	12	H
Loss & damage	0.11	12	L	0.11	12	M	**0.00**		L	0.03	13	M
People	4.96	2	H	1.26	4	M	2.36	4	H	0.43	7	H
Sports	**0.00**		NC	**0.00**		L	**0.00**		NC	**0.00**		NC

Table 7_9. Conceptual domains in the *chocolate*, and *wine* datasets’ most frequent 300 content words

Tables 7_10-7_13 summarize how the conceptual domains emerged at various thresholds of the most frequent words in the wordlist, and how this compares to the whole elicited datasets.

Matched Words	**n. domains**	**domain %**	**domain increase**	**H Cnv (%)**	**H+M Cnv (%)**
TOP 50	8	53.33	+ 8 domains	100	63.64
TOP 100	11	73.33	+ 3 domains	100	72.73
TOP 150	13	86.67	+ 3 domains	100	90.91
TOP 200	13	86.67	+ 0 domains	100	90.91
TOP 250	13	86.67	+ 0 domains	100	90.91
TOP 300	13	86.67	+ 0 domains	100	90.91
whole dataset	15				

Table 7_10. *Chocolate* English wordlist:
Conceptual domain analysis at different thresholds

Matched Words	**n. domains**	**domain %**	**domain increase**	**H Cnv (%)**	**H+M Cnv (%)**
TOP 50	11	73.33	+ 11 domains	100	76.92
TOP 100	13	86.67	+ 2 domains	100	92.31
TOP 150	13	86.67	+ 2 domains	100	92.31
TOP 200	14	93.33	+ 1 domain	100	100
TOP 250	14	93.33	+ 0 domains	100	100
TOP 300	14	93.33	+ 0 domains	100	100
whole dataset	15				

Table 7_11. *Chocolate* Italian wordlist:
Conceptual domain analysis at different thresholds

Matched Words	**n. domains**	**domain %**	**domain increase**	**H Cnv (%)**	**H+M Cnv (%)**
TOP 50	8	57.14	+ 8 domains	100	80
TOP 100	9	64.29	+ 1 domain	100	80
TOP 150	10	71.43	+ 1 domain	100	90
TOP 200	13	92.86	+ 3 domains	100	100
TOP 250	13	92.86	+ 0 domains	100	100
TOP 300	13	92.86	+ 0 domains	100	100
whole dataset	14				

Table 7_12. *Wine* English wordlist:
Conceptual domain analysis at different thresholds

Matched Words	**n. domains**	**domain %**	**domain increase**	**H Cnv (%)**	**H+M Cnv (%)**
TOP 50	10	66.67	+ 10 domains	75	76.92
TOP 100	11	73.33	+ 1 domain	75	84.62
TOP 150	12	80.00	+ 1 domain	100	92.31
TOP 200	13	86.67	+ 1 domain	100	100
TOP 250	13	86.67	+ 0 domain	100	100
TOP 300	13	86.67	+ 0 domain	100	100
whole dataset	15				

Table 7_13. *Wine* Italian wordlist:
Conceptual domain analysis at different thresholds

Domain coverage ranges from about 86.7% of English *chocolate* and Italian *wine* to over 93% of Italian *chocolate* – values which are remarkably higher than the corresponding semantic field coverage, ranging from the 65% of Italian *chocolate* to almost 70.6% of English *wine*.

The top 300 content words in the wordlist, representing slightly more than 2.5% of the total number of running words in the datasets, showed all of the highly conventionalised domains in all the datasets, and all of the cultural associations (high plus medium conventionalisation domains) in all cases except English *chocolate*. The domains which are left out in the top 300 content words are always the ones with lowest conventionalisation (L or NC), except for English *chocolate* where one medium conventionalisation domain is left out.

Domain SPORTS is systematically absent from the *wine* and *chocolate* domain lists above, but this is no surprise, as SPORTS showed very few occurrences also in the whole datasets – so few that it ranked last in all domains lists, and that Molinari's evenness index could not be computed. The other domain which is frequently absent from the top 300 words domain lists is Imagination; in the analyses of the whole datasets, this domain ranked among the less frequent ones (12 out of 15 and 13 out of 14 in the two English datasets, and 14 out of 15 in the Italian ones), but showed different levels of conventionalisation depending on cases (medium level in the *chocolate* datasets, low level in the Italian *wine* dataset, unknown level in the English *wine* dataset). Consequently, presence/absence of a domain in the most frequent 300 content words seems to be possibly related to frequency of that domain in the whole dataset, as well as conventionalisation.

At quantitative level, Spearman's Rank Correlation Coefficient showed strong/very strong correlation between conceptual domains emerging from the top 300 content words in the frequency wordlist and the whole dataset. In fact, with $p < 0.01$, for English *chocolate* $r = 0.813$; for Italian *chocolate*, $r = 0.963$; for English *wine*, $r = 0.969$; and for Italian *wine*, $r = 0.924$.

7.2.5 Semantic field Assessment

The manual coding scheme, used in coding the whole datasets, included four types of assessment (Positive, Negative, Neutral and Undecided), and the four elicited datasets 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, as summarised in Table 6_15 in Chapter 6.

In the current experiment, as described in Section 7.3.1, content words having specific evaluative meaning were classified as Positive assessment or Negative assessment. In all the four elicited datasets, the most frequent 300 content words in the word list included words with evaluative meaning, as summarised in Table 7_14. The numerical values in the table indicate the overall percentage frequency of the items having that particular evaluative meaning and appearing among the most frequent 300 content words.

	Positive	**Negative**
English *chocolate*	2.99	0.57
Italian *chocolate*	1.37	0.06
English *wine*	1.02	0.42
Italian *wine*	1.45	0.09

Table 7_14. Assessment field results in the top 300 words

As was the case with the whole elicited datasets, positive evaluation predominates over negative evaluation.

Looking back at all the analyses in Section 7.2, the results achieved can be considered more than satisfactory, given that the most frequent 300 words in the wordlists cover only about 3% of the words in the datasets.

7.3 Route two: creating a sub-corpus by sampling using the most frequent lemmas in the dataset

As an alternative route, the top words in the frequency wordlist were used to extract a 'sample' subset of sentences from the whole corpus, thus creating a 'sampled sub-corpus' which was then analysed at sentence level. The reasoning subtending such an unusual sampling procedure was that, as Szalay and Maday (1973) suggested, semantic, mental associations are not isolated entities, but rather are connected in networks. Consequently, the semantic and mental associations of a lemma that associates frequently with the key word under investigation might be among the cultural associations of the key word itself.

One by one, the top words in the frequency wordlist (created following the procedure described in Section 7.1) were used to extract sentences from the corpus. Although the frequency list includes words, when looking for the corresponding sentences, they were treated as lemmas. Sentences including more than one instance of the given lemma, or more than one of the considered lemmas, were retrieved only once. More concretely, in the English elicited corpus on chocolate, for example, the most frequent word was ‘like’, so the first step in the creation of the sampled sub-corpus was extracting every sentence containing word ‘like’ or any of its inflected forms (‘likes’, ‘liked’, etc.); the second most frequent word was ‘eat’, and the second step was extracting all sentences containing ‘eat’ or any of its inflected forms (‘eating’, ‘eats’, ‘ate’, etc.), excluding those which had already been retrieved in the previous step; and so on and so forth.

This procedure was initially applied to the English *chocolate* dataset. As described in Chapter 5, Table 5_2, and Chapter 6, Table 6_1, this dataset includes 1886 sentences and 88 semantic fields. As summarised in Table 7_15, the most frequent word in the word list (*like*, as a verb, preposition and conjunction), treated as lemma, retrieved 141 sentences, corresponding to 49 semantic fields. The second most frequent word (verb *eat*) retrieved a further 134 sentences and provided 17 new semantic fields. The third most frequent word (verb *make*) contributed a further 199 sentences to the sub-corpus, corresponding to 5 new semantic fields. The next most frequent word was the third person singular form of lemma *make* (*makes*), and was therefore ignored. Next in the list came word *good*; this contributed a further 67 new sentences and 2 new fields. At this point it was clear that the number of new semantic fields retrieved was drastically dropping, regardless of the number of new sentences entering the corpus. However, the sub-corpus thus created, which included a total of 541 sentences (28.7% of the original dataset), was already able to show more than 80% of the semantic fields in the original dataset (see Table 6_1) and, most importantly, all of the fields with a high level of conventionalisation.

Consequently, I decided to stop the sampling procedure and consider the sub-corpus finished. A similar procedure was applied to the other elicited datasets available.

7.3.1 Semantic fields analysis at different thresholds

The results of the sampling procedure in terms of semantic fields are summarised in Tables 7_16 to 7_18. In all these tables, percentage values are rounded to the first decimal place. Column one shows the steps and the corresponding lemmas used for retrieving the sub-corpus sentences; column two indicates the overall percentage of fields covered by the retrieved sentences; columns three and four show the percentage of highly conventionalised fields (H Cnv) and cultural associations (H+M Cnv) covered. Finally, the last two columns summarize field and sentence increases in passing from one stage to the next in the retrieving process.

As in the previous section, the tables below do not consider semantic field ASSESSMENT, as it will be treated separately (see Section 7.3.2.3).

As mentioned in Section 7.3.2, the English *chocolate* sub-corpus (Table 7_15) includes a total of 541 sentences (28.7% of the original dataset), shows 83% of the semantic fields in the original dataset and, most importantly, 100% of the fields with a high level of conventionalisation and 94.92 of the cultural associations.

The Italian *chocolate* dataset originally included 1603 sentences and 86 fields. Its sampled sub-corpus includes 489 sentences (30.5% of original dataset) and 63 fields, corresponding to over 70% of the total number of fields in the original, almost 97% of the highly conventionalised fields, and over 94.5% of the cultural associations (see Table 7_16).

The *wine* datasets included, respectively, 1938 sentences and 84 fields for English, and 1573 sentences and 84 fields for Italian. After this sampling procedure, the English *wine* sub-corpus includes 672 sentences (34.7% of the original dataset) and 67 fields, corresponding to almost 80% of the total number of fields in the original, 97% of the highly conventionalised fields, and slightly more than 96% of the cultural associations (see Table 7_17). The Italian *wine* sub-corpus includes 412 sentences (26.2% of original dataset) and 61 fields, corresponding to slightly more than 70% of the total number of fields in the original, almost 96% of the highly conventionalised fields, and about 94.5% of the cultural associations (see Table 7_18).

Lemmas	**Overall fields (%)**	**H Cnv (%)**	**H+M Cnv (%)**	**Field increase**	**Sentence increase**
1: like	55.7	71.4	67.80	+ 49 fields	+ 141 sentences
2: like + eat	75.0	94.3	88.14	+ 17 fields	+ 134 sentences
3: like + eat + make	80.7	100	94.92	+ 5 fields	+ 199 sentences
4: like + eat + make +good	83.0	100	94.92	+ 2 fields	+ 67 sentences

Table 7_15. *Chocolate* English elicited sub-corpus:
Semantic fields analysis at different thresholds

Lemmas	**Overall fields (%)**	**H Cnv (%)**	**H+M Cnv (%)**	**Field increase**	**Sentence increase**
1: fare	66.3	93.8	89.10	+ 57 fields	+ 302 sentences
2: fare + fondente	67.4	96.9	91.00	+ 1 fields	+ 62 sentences
3: fare + fondente + piacere	69.8	96.9	92.73	+ 2 fields	+ 70 sentences
4: fare + fondente + piacere + molto	73.3	96.9	94.55	+ 3 fields	+ 55 sentences

Table 7_16. *Chocolate* Italian elicited sub-corpus:
Semantic fields analysis at different thresholds

Lemmas	**Overall fields (%)**	**H Cnv (%)**	**H+M Cnv (%)**	**Field increase**	**Sentence increase**
1: drink	64.0	85.7	84.62	+ 54 fields	+ 305 sentences
2: drink + red	73.8	94.3	94.23	+ 8 fields	+ 162 sentences
3: drink + red + good	77.4	97.1	95.15	+ 3 fields	+ 97 sentences
4: drink + red + good + like	79.7	97.1	96.15	+ 2 fields	+ 108 sentences

Table 7_17. *Wine* English elicited sub-corpus:
Semantic fields analysis at different thresholds

Lemmas	Overall fields (%)	H Cnv (%)	H+M Cnv (%)	Field increase	Sentence increase
1: fare	65.5	84.4	83.33	+ 55 fields	+ 180 sentences
2: fare + rosso	72.6	95.6	94.44	+ 6 fields	+ 87 sentences
3: fare + rosso + bianco	72.6	95.6	94.44	+ 0 fields	+ 53 sentences
4: fare + rosso + bianco + buon	72.6	95.6	94.44	+ 0 fields	+ 92 sentences

Table 7_18. *Wine* Italian elicited sub-corpus:
Semantic fields analysis at different thresholds

A comparative look at the summary tables above shows that the top four words in the frequency wordlist, treated as lemmas, provided sub-corpora whose size varies between 25% and 35% of the corresponding original dataset. Despite their limited size, the sub-corpora show over 95% of the highly conventionalised fields in the original datasets (corresponding to a maximum of one or two of the less frequent conventionalised fields being absent from each sub-corpus), and a slightly lower percentage of the cultural associations (always exceeding 94%). The number of sentences retrieved at each stage of the sampling procedure varies in a non linear fashion, yet a steady decrease can be seen in the number of new fields retrieved at each stage, to the point that field-wise it seemed useless to continue the process after the fourth semantic lemma.

Finally, each sub-corpus was treated as an autonomous set of data, and semantic field values were calculated as percentages of the total number of sentences in the sub-corpus. Tables 7_19-7_22 show the semantic fields retrieved in each sub-corpus, in decreasing order of frequency.

semantic field	%	semantic field	%	semantic field	%
F-food	13.68	FET-sweet	1.48	FET-price	0.37
H-body	9.98	FE-sex	1.29	FET-packaging	0.37
FE-happiness	8.32	FE-mood	1.29	F-storage	0.18
F-product/shape	7.95	C-gift	1.29	H-dieting	0.18
FET-quantity	6.65	E-transaction	1.11	E-religion	0.18
H-health	6.10	FE-passion	1.11	E-war	0.18
FET-quality/type	5.91	EN-animals	1.11	E-law	0.18
FET-taste/smell	5.91	H-medicine	0.92	E-holiday	0.18
F-composition	5.55	FE-nice/pleasant/pleasure	0.92	FE-senses	0.18
FE-desire	3.88	FE-guilt	0.92	FE-seduction	0.18
F-bakery/cooking	3.70	E-economy	0.74	FE-surprise	0.18
F-manufacturing	3.70	FE-relax	0.74	FE-peace	0.18
FE-unpleasant	3.70	P-family	0.74	FE-loneliness	0.18
E-event	3.33	L-existence	0.74	P-gay	0.18
E-time	3.33	comparison	0.55	P-royalty	0.18
G-geo locations	3.33	FE-love	0.55	P-posh	0.18
F-recipe	2.59	I-fantasy/magic	0.55	LD-theft	0.18
F-drink	2.40	FET-energy	0.55	C-party	0.18
H-beauty	2.22	E-fair trade	0.37	EN-house	0.18
P-women	2.03	E-work	0.37	EN-dirt	0.18
P-children	2.03	FE-memory	0.37	L-future	0.18
CUL-artistic production	2.03	FE-comfort	0.37	FET-physical properties	0.18
P-men	1.85	P-friendship	0.37	FET-colour	0.18
P-sharing/society	1.48	I-dream	0.37		
P-people	1.48	LD-drugs & addiction	0.37		

Table 7_19. English *chocolate*: Semantic fields in the 4-lemma sampled sub-corpus

semantic field	%	semantic field	%	semantic field	%
FET-quality/type	17.38	FET-physical properties	2.04	FE-seduction	0.61
F-bakery/cooking	8.38	H-dieting	1.84	FE-comfort	0.61
H-health	8.38	P-family	1.84	FET-colour	0.61
FET-taste/smell	6.95	FE-mood	1.64	FE-love	0.41
FET-quantity	5.73	P-people	1.64	FE-guilt	0.41
H-body	5.52	E-transaction	1.43	FE-relax	0.41
F-product/shape	4.70	FE-happiness	1.43	P-friendship	0.41
H-beauty	4.70	C-gift	1.43	EN-tech	0.41
P-children	4.70	FE-sex	1.23	FET-genuine	0.41
FE-nice/pleasant/pleasure	4.50	CUL-studying/intellect	1.23	S-sports	0.41
G-geo locations	4.50	FET-sweet	1.23	E-playing	0.20
F-recipe	4.29	FE-no reaction	1.02	E-language	0.20
comparison	4.09	P-age	1.02	E-economy	0.20
CUL-artistic production	3.89	I-dream	1.02	E-fair trade	0.20
H-medicine	3.68	EN-nature	1.02	E-history	0.20
F-food	3.48	EN-house	1.02	FE-loneliness	0.20
F-manufacturing	2.86	FET-energy	1.02	FE-persuasion	0.20
E-event	2.86	F-drink	0.82	P-men	0.20
FE-desire	2.86	P-women	0.82	LD-hiding	0.20
F-composition	2.66	E-time	0.61	C-party	0.20
FE-passion	2.25	FE-senses	0.61	EN-dirt	0.20

Table 7_20. Italian *chocolate*: Semantic fields in 4-lemma sampled sub-corpus

semantic field	%	semantic field	%	semantic field	%
FET-quality/type	20.24	F-bakery/cooking	1.34	FE-peace	0.30
E-excessive drinking	12.05	FE-happiness	1.34	LD-drugs & addiction	0.30
H-health	11.90	FE-passion	1.34	C-ceremonies	0.30
F-drink	10.57	P-age	1.34	EN-animals	0.30
FET-quantity	6.10	F-product/shape	1.19	CUL-artistic production	0.30
F-food	5.65	F-composition	1.04	FET-packaging	0.30
G-geo locations	5.51	EN-dirt	1.04	E-language	0.15
FET-taste/smell	4.91	F-manufacturing	0.89	E-transaction	0.15
F-recipe	4.02	H-body	0.89	E-law	0.15
P-women	3.87	E-religion	0.89	FE-nice/pleasant/pleasure	0.15
comparison	3.72	P-people	0.89	FE-sex	0.15
FE-unpleasant	3.57	F-serving	0.74	FE-mood	0.15
E-time	3.27	FE-love	0.74	FE-memory	0.15
P-men	2.68	C-gift	0.74	FE-surprise	0.15
P-sharing/society	2.38	FET-physical properties	0.74	FE-guilt	0.15
FE-desire	2.23	E-event	0.60	FE-freedom	0.15
FET-price	2.08	P-children	0.60	G-spreading	0.15
P-posh	1.79	L-existence	0.60	EN-nature	0.15
P-family	1.79	E-driving	0.45	EN-house	0.15
F-storage	1.64	FE-relax	0.45	FET-sweet	0.15
H-medicine	1.64	C-party	0.45	FET-genuine	0.15
P-friendship	1.64	E-holidays	0.30		
FET-colour	1.64	E-work	0.30		

Table 7_21. English *wine*: Semantic fields in 4-lemma sampled sub-corpus

semantic field	%	semantic field	%	semantic field	%
FET-quality/type	33.98	FE-happiness	2.18	LD-drugs & addiction	0.73
H-health	17.48	P-children	2.18	EN-dirt	0.73
FET-quantity	12.38	FET-genuine	2.18	FET-packaging	0.73
F-recipe	11.65	E-religion	1.94	H-dieting	0.49
F-food	8.25	E-event	1.94	H-body	0.49
H-medicine	6.31	FET-physical properties	1.94	E-history	0.49
F-storage	5.34	CUL-studying/intellect	1.70	FE-memory	0.49
P-friendship	5.10	FET-colour	1.70	FE-peace	0.49
G-geo locations	4.37	CUL-artistic production	1.46	FE-loneliness	0.49
F-drink	4.13	FET-price	1.46	P-men	0.49
E-driving	4.13	C-gift	1.21	C-party	0.49
F-manufacturing	3.40	EN-nature	1.21	L-existence	0.49
F-serving	3.16	E-transaction	0.97	FET-sweet	0.49
E-language	3.16	E-work	0.97	H-beauty	0.24
E-excessive drinking	3.16	FE-confidence	0.97	E-playing	0.24
FE-unpleasant	3.16	FE-desire	0.97	FE-seduction	0.24
comparison	2.91	FE-mood	0.97	P-women	0.24
F-bakery/cooking	2.67	FE-relax	0.97	P-age	0.24
FE-nice/pleasant/pleasure	2.67	P-posh	0.97	P-sharing/society	0.24
P-family	2.67	F-product/shape	0.73	LD-hiding	0.24
FET-taste/smell	2.67	FE-no reaction	0.73	C-ceremonies	0.24
F-composition	2.18	FE-passion	0.73	CUL-culture	0.24
E-time	2.18	FE-comfort	0.73		

Table 7_22. Italian *wine*: Semantic fields in 4-lemma sampled sub-corpus

A quantitative comparison between the sampled sub-corpora and their corresponding datasets was performed, by applying Spearman’s Rank Correlation Coefficient. Although the sampled corpora include only about 72-83% of the total number of semantic fields present in the corresponding datasets and show them in a different ranking order, Spearman’s test highlighted very strong correlation between the two paired sets of data. In fact, with $p < 0.01$, for English *chocolate* $r = 0.903$; for Italian *chocolate*, $r = 0.894$; for English *wine*, $r = 0.905$; and for Italian *wine*, $r = 0.919$.

7.3.2 Conceptual domains analysis

The sampled sub-corpora were analysed also at the level of conceptual domains, and results were compared to domains in the whole datasets (Tables 6_4 and 6_11, Chapter 6).

Table 7_23 shows conceptual domains as they appeared in the 4-lemma sampled sub-corpora. Values are expressed as percentages on the total number of sentences in the sub-corpus. R stands for rank. Cnv shows the conventionalisation level of that domain in the whole dataset (Chapter 6). Bold signals the absence of that particular domain in the sampled sub-corpus. Domains are listed in alphabetical order.

Tables 7_24-7_27 summarize how the conceptual domains emerged in the sampled sub-corpora, moving from 1 lemma to 4 lemmas, and how this compares to the whole elicited datasets.

Domain	***Chocolate* Eng.**			***Chocolate* It.**			***Wine* Eng.**			***Wine* It.**		
	%	**R**	**Cnv**	**%**	**R**	**Cnv**	**%**	**R**	**Cnv**	**%**	**R**	**Cnv**
Ceremony	1.48	9	L	1.64	11	H	1.49	10	L	1.94	10	M
Comparison	0.55	11	L	4.09	9	H	3.72	8	M	2.91	9	M
Culture	2.03	8	M	5.11	7	H	0.30	13	L	3.40	8	H
Environment	1.48	9	M	2.66	10	M	1.64	9	M	1.94	10	M
Events	9.98	6	H	5.93	6	M	18.30	3	M	19.17	4	M
Features	21.63	3	M	35.38	1	M	36.31	1	M	57.52	1	M
Feelings & emotions	24.40	2	M	18.40	4	M	11.01	6	M	15.78	5	H
Food	39.74	1	M	27.20	2	M	27.08	2	M	41.50	2	M
Geo	3.33	7	H	4.50	8	M	5.65	7	H	4.37	7	M
Health &Body	19.41	4	M	24.13	3	M	14.43	5	M	25.00	3	M
Imagination	0.92	10	M	1.02	12	M	**0.00**		NC	**0.00**		L
Life	0.92	10	M	**0.00**		L	0.60	11	M	0.49	12	H
Loss & damage	0.55	11	L	0.20	14	M	0.30	12	L	0.97	11	M
People	10.54	5	H	10.63	5	M	16.96	4	H	12.14	6	H
Sports	**0.00**		NC	0.41	13	L	**0.00**		NC	**0.00**		NC

Table 7_23. Conceptual domains in the *chocolate*, and *wine* sampled 4-lemma sub-corpora

Lemmas	**n. domains**	**domain %**	**domain increase**	**H Cnv (%)**	**H+M Cnv (%)**
1: like	14	93.33	+ 14 domains	100	100
2: like + eat	14	93.33	+ 0 domains	100	100
3: like + eat + make	14	93.33	+ 0 domains	100	100
4: like + eat + make +good	14	93.33	+ 0 domains	100	100
whole dataset	15	100			

Table 7_24. *Chocolate* English elicited sub-corpus: conceptual domains

Lemmas	**n. domains**	**domain %**	**domain increase**	**H Cnv (%)**	**H+M Cnv (%)**
1: fare	13	86.67	+ 13 domains	66.33	66.33
2: fare + fondente	14	93.33	+ 1 domain	100	100
3: fare + fondente + piacere	14	93.33	+ 0 domains	100	100
4: fare + fondente + piacere + molto	14	93.33	+ 0 domains	100	100
whole dataset	15	100			

Table 7_25. *Chocolate* Italian elicited sub-corpus: conceptual domains

Lemmas	**n. domains**	**domain %**	**domain increase**	**H Cnv (%)**	**H+M Cnv (%)**
1: drink	12	85.71	+ 12 domains	100	100
2: drink + red	13	92.86	+ 1 domain	100	100
3: drink + red + good	13	92.86	+ 0 domains	100	100
4: drink + red + good + like	13	92.86	+ 0 domains	100	100
whole dataset	14	100			

Table 7_26. *Wine* English elicited sub-corpus: conceptual domains

Lemmas	**n. domains**	**domain %**	**domain increase**	**H Cnv (%)**	**H+M Cnv (%)**
1: fare	13	86.67	+ 13 domains	100	100
2: fare + rosso	13	86.67	+ 0 domains	100	100
3: fare + rosso + bianco	13	86.67	+ 0 domains	100	100
4: fare + rosso + bianco + buon	13	86.67	+ 0 domains	100	100
whole dataset	15	100			

Table 7_27. *Wine* Italian elicited sub-corpus: conceptual domains

Domain coverage ranges from 86.7% of Italian *wine* to over 93% of both English and Italian *chocolate* – values which are remarkably higher than the corresponding semantic field coverage, ranging from 72.6% of Italian *wine* to slightly over than 83% of English *chocolate*. In all the sub-corpora, the 4 most frequent words in the wordlist, treated as lemmas, showed all of the high and medium conventionalisation domains. In the English sub-corpora, they also showed all of the low conventionalisation domains, leaving out only and all of the domains that were so poorly attested in the original dataset as to being unclassifiable in terms of conventionalisation. In the Italian sub-corpora, instead, the left-out domains are the unclassified ones (when present) and/or low conventionalisation ones.

Similarly to what happened in the most frequent words in the wordlist, the absent domains include domain SPORTS – absent from the English *chocolate*, English *wine* and Italian *wine* domain lists above – and domain IMAGINATION – missing in the English *wine* and Italian *wine* sub-corpora, and showing, respectively, NC and low conventionalisation. Finally, the Italian *chocolate* sub-corpus is missing the domain LIFE which showed low conventionalisation in the corresponding dataset. Consequently, presence/absence of a domain in the sampled sub-corpora seems to be related to both frequency and conventionalisation of that domain.

At quantitative level, Spearman's Rank Correlation Coefficient (for $p < 0.01$) showed very strong correlation between conceptual domains in the 4-lemma sampled sub-corpora and the corresponding datasets: for English *chocolate*, $r = 0.911$; for Italian *chocolate*, $r = 0.965$; for English *wine*, $r = 0.977$; and for Italian *wine*, $r = 0.977$.

7.3.3 Semantic field ASSESSMENT

The results of the ASSESSMENT field in the 4-lemma sampled corpora are summarised in Table 7_28 and in Figure 7_1, below.

In Table 7_28 the figures are percentages of the total number of sentences in the sub-corpus. Figure 7_1 shows the 4-lemma corpora on the left, and the corresponding elicited datasets on the right. The four colours, labelled 1-4, indicate respectively Positive, Negative, Neutral, and Undecided assessment.

The ASSESSMENT results in the 4-lemma sampled sub-corpora are only partially comparable to those in the whole elicited datasets. In fact, the Italian 4-lemma sampled corpora show a majority of positive sentences, a somehow smaller percentage of neutral sentences, followed by a yet smaller percentage of negative sentences, and a few undecided sentences, like the corresponding whole datasets (Table 6_15, Chapter 6). Similarities between the pairs of data, however, hold true only rank-wise, but not proportion-wise, as visible in Figure 7_1.

	Positive	**Negative**	**Neutral**	**Undecided**
English *chocolate*	54.71	25.69	18.67	0.92
Italian *chocolate*	55.06	17.23	25.66	2.06
English *wine*	51.64	22.77	18.30	7.29
Italian *wine*	64.32	14.56	18.20	2.91

Table 7_28. ASSESSMENT field results in the 4-lemma sampled sub-corpora

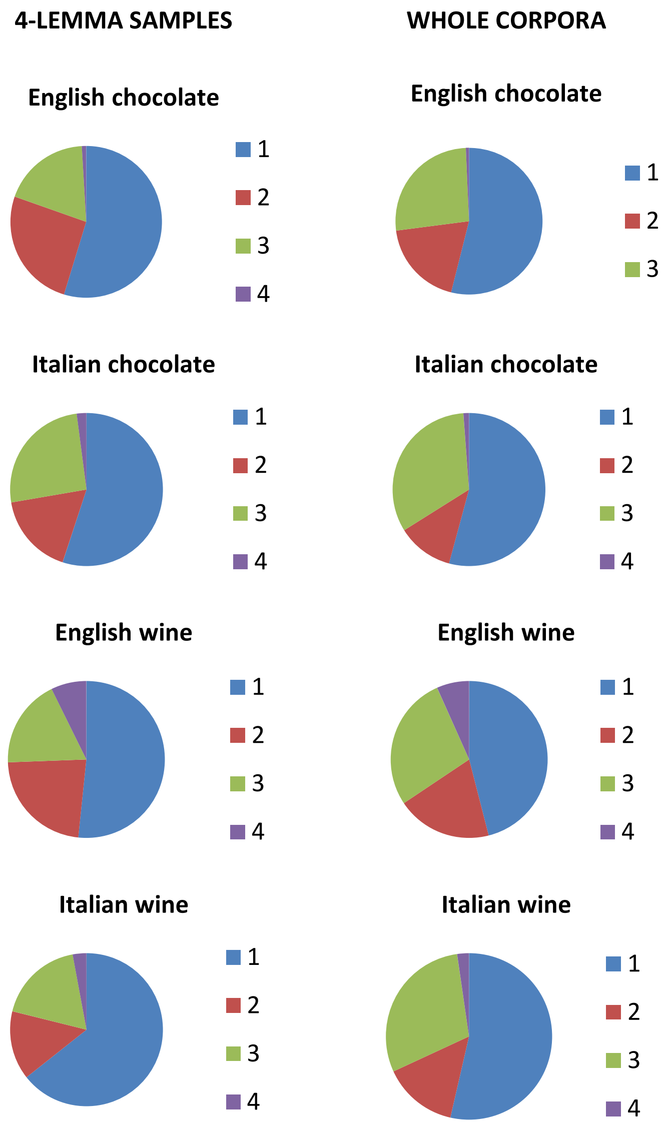

Figure 7_1. ASSESSMENT field results:
4-lemma sampled datasets vs. whole elicited datasets

In the English 4-lemma sampled sub-corpora, on the other hand, the percentage of negative sentences is higher than that of neutral sentences. The higher number of negative sentences in the English sub-corpora does not seem to be related in any way to the lemmas used for sampling. In fact, while in all the four cases, at least one of the lemmas is marked by a clear positive connotation (‘like’ and ‘good’ in the two English *chocolate* sub-corpora; ‘*piacere*’ in the Italian *chocolate* sub-corpus, and ‘*buon*’ in the Italian *wine* sub-corpus), none of the lemmas has an intrinsic negative connotation.

7.3.4 Conventionalisation level analysis and cross-cultural comparison

This section applies the conventionalisation-analysis-plus-t-test procedures described in Chapter 6 to the sampled sub-corpora, in order to assess the extent to which these smaller, but apparently rather representative sets of data could be suitable to establish the level of conventionalisation of semantic associations and to perform cross-cultural comparisons.

For each semantic field and conceptual domain, Molinari's evenness index was computed, and three levels of conventionalisation were distinguished using confidence intervals. The results are reported in Tables 7_29-7_32, in order of conventionalisation. The 99% confidence intervals were: 0.77-0.88 for English *chocolate*; 0.79-0.89 for Italian *chocolate*; 0.79-0.90 for English *wine*, and 0.81-0.90 for Italian *wine*. The evenness values are reported in column G2,1, accompanied by indication of their corresponding levels of conventionalisation (column Cnv).

Field	G2,1	Cnv	Field	G2,1	Cnv	Field	G2,1	Cnv
F-product/shape	0.63	H	P-family	0.73	H	FE-nice/pleasant/pleasure	1.00	L
F-manufacturing	0.71	H	G-geo locations	0.64	H	FE-love	1.00	L
F-food	0.70	H	L-existence	0.58	H	FE-memory	1.00	L
H-health	0.73	H	FET-quality/type	0.71	H	FE-comfort	1.00	L
H-body	0.74	H	FET-quantity	0.73	H	FE-relax	1.00	L
H-beauty	0.58	H	FET-sweet	0.62	H	P-children	1.00	L
E-economy	0.73	H	FET-taste/smell	0.74	H	P-friendship	1.00	L
E-transaction	0.76	H	F-bakery/cooking	0.78	M	P-sharing/society	1.00	L
E-event	0.65	H	F-drink	0.78	M	P-people	1.00	L
FE-unpleasant	0.61	H	F-composition	0.78	M	I-fantasy/magic	1.00	L
FE-desire	0.66	H	F-recipe	0.78	M	I-dream	1.00	L
FE-sex	0.61	H	E-time	0.77	M	LD-drugs &addiction	1.00	L
FE-happiness	0.71	H	FE-mood	0.77	M	C-gift	1.00	L
FE-passion	0.76	H	comparison	1.00	L	EN-animals	1.00	L
FE-guilt	0.75	H	H-medicine	1.00	L	CUL-artistic production	1.00	L
P-women	0.65	H	E-work	1.00	L	FET-energy	1.00	L
P-men	0.77	H	E-fair trade	1.00	L	FET-packaging	1.00	L
Domain	**G2,1**	**Cnv**	**Domain**	**G2,1**	**Cnv**	**Domain**	**G2,1**	**Cnv**
Food	0.61	H	Geography	0.64	H	Comparison	1.00	L
Health & Beauty	0.67	H	Life	0.59	H	Loss & damage	1.00	L
Events	0.61	H	Features	0.66	H	Ceremony	1.00	L
Feelings & Emotions	0.64	H	Imagination	0.75	M	Culture	1.00	L
People	0.60	H	Environment	0.78	M			

Table 7_29. English *chocolate* 4-lemma sub-corpus: Conventionalisation results

Field	G2,1	Cnv	Field	G2,1	Cnv	Field	G2,1	Cnv
comparison	0.71	H	I-dream	0.59	H	E-time	1.00	L
F-product/shape	0.78	H	C-gift	0.68	H	FE-senses	1.00	L
F-bakery/cooking	0.60	H	EN-nature	0.75	H	FE-love	1.00	L
F-composition	0.78	H	CUL-studying/intellect	0.76	H	FE-desire	1.00	L
F-recipe	0.71	H	FET-quality/type	0.60	H	FE-happiness	1.00	L
H-dieting	0.76	H	FET-quantity	0.73	H	FE-comfort	1.00	L
H-health	0.72	H	FET-sweet	0.76	H	FE-relax	1.00	L
H-medicine	0.68	H	FET-taste/smell	0.74	H	P-women	1.00	L
H-beauty	0.78	H	F-manufacturing	0.83	M	P-age	1.00	L
E-transaction	0.77	H	F-food	0.80	M	P-friendship	1.00	L
FE-no reaction	0.75	H	H-body	0.80	M	P-people	1.00	L
FE-sex	0.77	H	FE-nice/pleasant/pleasure	0.81	M	EN-house	1.00	L
FE-seduction	0.71	H	FE-passion	0.81	M	EN-tech	1.00	L
FE-mood	0.78	H	CUL-artistic production	0.85	M	FET-colour	1.00	L
P-children	0.72	H	FET-physical properties	0.79	M	FET-genuine	1.00	L
P-family	0.79	H	F-drink	1.00	L	FET-energy	1.00	L
G-geo locations	0.65	H	E-event	1.00	L	S-sports	1.00	L
Domain	**G2,1**	**Cnv**	**Domain**	**G2,1**	**Cnv**	**Domain**	**G2,1**	**Cnv**
Food	0.65	H	Ceremony	0.67	H	Events	0.81	L
Health & Beauty	0.65	H	Features	0.66	H	Culture	0.80	L
Feelings & Emotions	0.65	H	Comparison	0.71	M	Sports	1.00	L
Geography	0.65	H	People	0.70	M			
Imagination	0.59	H	Environment	0.68	M			

Table 7_30. Italian *chocolate* 4-lemma sub-corpus: Conventionalisation results

Field	**G2,1**	**Cnv**	**Field**	**G2,1**	**Cnv**	**Field**	**G2,1**	**Cnv**
comparison	0.78	H	FET-quality/type	0.66	H	E-work	1.00	L
F-drink	0.63	H	FET-physical properties	0.75	H	E-religion	1.00	L
F-food	0.51	H	FET-quantity	0.72	H	E-holidays	1.00	L
F-composition	0.76	H	FET-price	0.78	H	E-event	1.00	L
F-recipe	0.62	H	F-storage	0.81	M	FE-relax	1.00	L
H-health	0.66	H	H-medicine	0.81	M	P-children	1.00	L
E-excessive drinking	0.65	H	FE-desire	0.83	M	P-age	1.00	L
E-time	0.78	H	FE-happiness	0.79	M	LD-drugs & addiction	1.00	L
FE-unpleasant	0.61	H	P-posh	0.82	M	C-ceremonies	1.00	L
FE-love	0.75	H	P-sharing/society	0.84	M	C-party	1.00	L
FE-passion	0.76	H	FET-taste/smell	0.82	M	C-gift	1.00	L
P-women	0.62	H	F-product/shape	1.00	L	EN-animals	1.00	L
P-men	0.59	H	F-serving	1.00	L	EN-dirt	1.00	L
P-friendship	0.77	H	F-bakery/cooking	1.00	L	CUL-artistic production	1.00	L
P-people	0.76	H	F-manufacturing	1.00	L	L-existence	1.00	L
P-family	0.78	H	H-body	1.00	L	FET-colour	1.00	L
G-geo locations	0.53	H	E-driving	1.00	L	FET-packaging	1.00	L
Domain	**G2,1**	**Cnv**	**Domain**	**G2,1**	**Cnv**	**Domain**	**G2,1**	**Cnv**
Food	0.58	H	Geography	0.53	H	Ceremony	1.00	L
Health & Beauty	0.61	H	Features	0.70	H	Culture	1.00	L
Events	0.66	H	Comparison	0.78	M	Life	1.00	L
Feelings & Emotions	0.68	H	Environment	0.81	M			
People	0.61	H	Loss & Damage	1.00	L			

Table 7_31. English *wine* 4-lemma sub-corpus: Conventionalisation results

Field	**G2,1**	**Cnv**	**Field**	**G2,1**	**Cnv**	**Field**	**G2,1**	**Cnv**
comparison	0.78	H	P-friendship	0.80	H	E-work	1.00	L
F-product/shape	0.71	H	P-family	0.77	H	FE-unpleasant	1.00	L
F-bakery/cooking	0.81	H	CUL-studying/intellect	0.76	H	FE-desire	1.00	L
F-drink	0.76	H	FET-quality/type	0.65	H	FE-happiness	1.00	L
F-manufacturing	0.68	H	FET-physical properties	0.75	H	FE-mood	1.00	L
F-food	0.78	H	FET-quantity	0.70	H	FE-passion	1.00	L
F-composition	0.79	H	FET-colour	0.77	H	FE-memory	1.00	L
F-serving	0.64	H	FET-genuine	0.77	H	FE-relax	1.00	L
F-storage	0.78	H	FET-price	0.75	H	P-children	1.00	L
F-recipe	0.71	H	FET-packaging	0.71	H	P-posh	1.00	L
H-health	0.65	H	E-time	0.84	M	C-gift	1.00	L
H-medicine	0.78	H	G-geo locations	0.84	M	EN-nature	1.00	L
E-language	0.82	H	H-dieting	1.00	L	EN-dirt	1.00	L
E-religion	0.75	H	H-body	1.00	L	CUL-artistic production	1.00	L
E-excessive drinking	0.77	H	E-transaction	1.00	L	L-existence	1.00	L
FE-confidence	0.73	H	E-event	1.00	L	FET-taste/smell	1.00	L
FE-nice/pleasant/pleasure	0.76	H	E-driving	1.00	L			
Domain	**G2,1**	**Cnv**	**Domain**	**G2,1**	**Cnv**	**Domain**	**G2,1**	**Cnv**
Food	0.62	H	People	0.54	H	Geography	0.84	L
Health & Beauty	0.61	H	Features	0.68	H	Ceremony	1.00	L
Events	0.69	H	Comparison	0.78	M	Environment	1.00	L
Feelings & Emotions	0.63	H	Culture	0.81	M	Life	1.00	L

Table 7_32. Italian *wine* 4-lemma sub-corpus: Conventionalisation results

These results were compared to conventionalisation levels in the whole datasets (see Chapter 6), to establish the percentage of fields in the sub-corpus which coincides with the whole dataset conventionalisation results. Field-wise, comparison of the sub-corpus to the whole dataset showed the following percentages of correctly identified conventionalisation levels: 49% for English *chocolate*, and Italian *chocolate*; 45% for English *wine*; and 62% for Italian *wine*.

However the real focus of this work are cultural associations, which include fields with medium conventionalisation, as well as those with high conventionalisation. Consequently, if we disregard the distinction between high and medium conventionalisation, in the 4-lemma sub-corpora the following percentages of cultural associations were correctly indicated: 54.9% for English *chocolate*; 62.8% for Italian *chocolate*; 52.9% for English *wine*; and about 58% for Italian *wine*.

At the level of conceptual domains, the English sub-corpora showed 57.1% and 53.8% matches for *chocolate* and *wine*, respectively, while the Italian sub-corpora showed lower levels of matching: 30.8% for *chocolate* and 25% for *wine*. However, if we disregard the distinction between high and medium levels of conventionalisation, in the 4-lemma sub-corpora 100% of cultural associations were correctly indicated.

All things considered, this sampling method provided conventionalisation results which were only partially comparable to those of the whole datasets. A possible explanation for this will be put forward further on in the chapter, after comparing these result to those obtained with random sampling.

Finally, the English and Italian semantic associations in the sub-corpora were compared by means of Welch *t* test, in order to highlight the cases when the difference in means was statistically significant. T-test results were then triangulated with conventionalisation results, applying the procedure adopted in Chapter 6 to understand which differences could be safely attributed to culture and which to circumstantial elements, such as population sampling. The logical reasoning followed in Chapter 6 led to considering a difference in means as having cultural origins in the following cases: when the field with the higher mean also shows high level of conventionalisation; when the field with higher mean shows medium level of conventionalisation against a high level (H) or absence (NC) of conventionalisation in the other culture. All other cases are uncertain, and need confirmation from other population samples.

The results are summarised in Tables 7_33 and 7_34. While in Chapter 6 I considered only t-test results significant for P < 0.01, in the current experiments I extended the significance level to 0.05, as a consequence of the smaller size of the datasets analysed.

Field	**P (< 0.05)**	**T**	**ff**	**st.error of df**	**mean values English**	**Cnv**	**mean values Italian**	**Cnv**
comparison	0.0004	3.6667	70	0.077	0.03	L	**0.32**	**H**
F-bakery/cooking	0.0011	3.3486	100	0.125	0.23	M	**0.65**	**H**
F-food	0.0000	4.4769	146	0.127	**0.84**	**H**	0.27	M
E-time	0.0095	2.3599	147	0.069	**0.21**	M	0.05	L
FE-unpleasant	0.0012	2.8598	147	0.081	**0.23**	**H**	NC	NC
FE–nice/pleasant/pleasure	0.0001	4.4347	147	0.066	0.06	L	**0.35**	M
FE-happiness	0.0000	4.3674	118	0.094	**0.52**	**H**	0.11	L
P-children	0.0067	3.0555	147	0.081	0.12	L	**0.37**	**H**
P-sharing/society	0.0040	2.5248	147	0.037	0.09	L	NC	NC
FET-quality/type	0.0000	5.5560	90	0.176	0.37	**H**	**1.35**	**H**
H-dieting	0.0220	3.3486	100	0.125	NC	NC	**0.14**	**H**
H-health	0.0375	2.3440	67	0.056	0.38	H	**0.65**	**H**
H-medicine	0.0132	20.1016	128	0.127	0.06	L	**0.29**	**H**
H-beauty	0.0197	2.8718	147	0.018	0.14	H	**0.37**	**H**
P-men	0.0270	1.9746	147	0.051	**0.12**	**H**	NC	NC
P-age	0.0241	2.7045	147	0.029	NC	NC	**0.08**	L
EN-animals	0.0134	2.5249	85	0.028	**0.07**	L	0	L
CUL-artistic production	0.0180	2.5466	147	0.068	0.13	L	**0.30**	M
CUL–studying/intellect	0.0327	2.5547	147	0.037	NC	NC	**0.10**	**H**
FET-physical properties	0.0126	2.5627	68	0.051	0.01	**NC**	**0.14**	**M**

Domain	**P (< 0.05)**	**T**	**ff**	**st.error of df**	**mean values English**	**Cnv**	**mean values Italian**	**Cnv**
Comparison	0.0004	3.6667	70	0.077	0.03	L	**0.32**	**H**
Health & Body	0.0050	2.9133	147	0.220	1.23	H	**1.87**	**H**
Culture	0.0015	3.5500	147	0.076	0.13	L	**0.40**	L
Features	0.0000	4.8660	105	0.289	1.33	H	**2.73**	**H**

Table 7_33. *Chocolate* sub-corpora: T-Test results for semantic fields and conceptual domains

Field	**P** **(< 0.05)**	**T**	**ff**	**st.error** **of df**	**mean values** **English**	**Cnv**	**mean values** **Italian**	**Cnv**
F-drink	0.0000	6.0707	114	0.117	**0.81**	H	0.10	**H**
F-recipe	0.0077	2.9438	148	0.142	0.31	H	**0.73**	**H**
H-medicine	0.0089	2.8998	148	0.079	0.13	M	**0.35**	**H**
E-language	0.0020	3.7559	148	0.049	0.01	NC	**0.19**	**H**
E-excessive drinking	0.0000	5.9076	108	0.137	**0.92**	**H**	0.11	H
P-women	0.0001	4.1763	87	0.071	**0.30**	**H**	0	NC
P-age	0.0023	2.6399	148	0.039	**0.10**	L	NC	NC
P-sharing/society	0.0001	4.0953	87	0.044	**0.18**	**M**	0	NC
FET-taste/smell	0.0010	3.1137	148	0.079	**0.38**	M	0.13	L
F-manufacturing	0.0402	2.3399	148	0.067	0.07	L	**0.23**	**H**
F-storage	0.0141	2.7086	148	0.079	0.13	M	**0.34**	**H**
FE- unpleasant	0.0133	2.2269	148	0.086	**0.27**	**H**	0.08	L
FE- desire	0.0189	2.1534	148	0.054	**0.17**	M	0.05	L
P-men	0.0112	2.5746	130	0.067	**0.20**	**H**	0.03	NC
P-posh	0.0256	2.0200	148	0.052	**0.14**	M	0.03	L
P-people	0.0331	1.8157	148	0.038	**0.07**	**H**	NC	NC
CUL-studying/intellect	0.0327	2.6067	148	0.037	NC	NC	**0.10**	**H**
FET-quality/type	0.0235	2.4057	148	0.243	1.55	H	**2.13**	**H**
FET-genuine	0.0378	2.4564	148	0.041	0.01	NC	**0.11**	**H**

Domain	**P** **(< 0.05)**	**T**	**ff**	**st.error** **of df**	**mean values** **English**	**Cnv**	**mean values** **Italian**	**Cnv**
Events	0.0021	2.8939	148	0.215	**1.40**	**H**	0.77	H
People	0.0000	4.4252	128	0.209	**1.30**	**H**	0.37	H
Culture	0.0076	3.1475	148	0.049	0.02	L	**0.18**	M

Table 7_34. *Wine* sub-corpora: T-Test results for semantic fields and conceptual domains

Consequently, considering the 0.05 level of significance, in the 4-lemma sub-corpora, the following semantic fields would appear as distinctively more prominent for Italians than for the English, when talking about *chocolate*: COMPARISON; BAKERY/COOKING; DIETING; HEALTH; MEDICINE; BEAUTY; CHILDREN; STUDYING/INTELLECT; QUALITY/TYPE; and PHYSICAL PROPERTIES.

On the other hand, more prominent for the English than for Italians appear to be: FOOD; UNPLEASANT, HAPPINESS, and MEN. As regards conceptual domains, the following would appear as prevalent in Italian rather than in English: COMPARISON; HEALTH & BODY, and FEATURES. No domain emerges as predominantly English.

Table 7_34 below illustrates the situation with reference to key word *wine*. Considering the 0.05 level of significance, the following semantic fields would appear as distinctively more prominent for the Italians than for the English, when talking about *wine*: MANUFACTURING; STORAGE; RECIPE; MEDICINE; LANGUAGE; STUDYING/INTELLECT; QUALITY/TYPE; and GENUINE. On the other hand, more prominent for the English than for the Italians appear to be: DRINK; EXCESSIVE DRINKING; UNPLEASANT, WOMEN; MEN; SHARING/SOCIETY; and PEOPLE. As regards conceptual domains, domains EVENTS; AND PEOPLE appear as prevalent in English rather than in Italian. No domain emerges as predominantly Italian.

Unfortunately, these results are rather different from the ones obtained with the whole corpus, and described in Chapter 6, Section 6.2.2. This type of cross-cultural comparison is highly dependent on quantitative results, which, despite the high level of correlation attested in Section 7.3.2.1, are strongly connected to sample structure.

7.4 Route three: random sampling

The results obtained in Section 7.3.2, by sampling using the most frequent lemmas, seem to confirm the hypothesis that semantic mental (and cultural) associations are connected in networks. But how does this method compare to random sampling? This issue is faced in the following sub-sections. For each elicited dataset, a random sample will be created and compared to the results of 4-lemma sampling as well as those of the whole dataset. Kilgarriff (2001b) suggests generating several random samples and average the results, to guarantee maximal representativeness of the sample; in the current work multiple random sampling will be substituted with sampling on different data sets followed by assessment of the consistency of the results.

In order to proceed with random sampling in the elicited datasets, a software programme for mathematical calculations, Mathematica,[4] was set to list a specific number of random positive integers within a given range, different for each dataset. Indeed, I wanted the random sub-sets to match in size the 4-lemma sampled datasets. Consequently, for English *chocolate* 541 integers in the 1-1886 range were obtained; for Italian *chocolate*, 489 integers in the 1-1603; for English *wine*, 672 integers in the 1-1938; and, for Italian *wine*, 412 integers in the 1-1573 range. The random integers listed by the software were used to extract sentences from the elicited datasets.

The randomly sampled corpora were thus created and assessed following all the analytical steps used with the 4-lemma sampled corpora, and their respective results were compared.

7.4.1 Semantic fields analysis

The semantic fields retrieved by the randomly sampled sub-corpora are summarised in Tables 7_35-7_38, accompanied by the corresponding frequency calculated as a percentage of the total number of sentences in each sub-corpus. Semantic fields are listed in decreasing order of frequency.

[4] Copyright: Wolfram Research, Inc. (http://www.wolfram.com/mathematica/). Mathematica is a fully fledged software for symbolic calculation. Its built-in random number extraction function is based on an algorithm which produces a different sequence of pseudorandom choices whenever you run Mathematica, as a consequence of the fact that the initialization seed depends on the instant (day, hour, minutes, seconds) the function is called. Given a range of N integers, the probability that a specific integer number is extracted is 1/N, which means that all and any integers have the same probability of being extracted.

semantic field	%	semantic field	%	semantic field	%
F-product/shape	10.72	FE-unpleasant	1.66	L-existence	0.55
FET-quality/type	8.32	FE-comfort	1.66	E-language	0.37
FE-happiness	7.21	P-women	1.66	E-fair trade	0.37
F-food	7.02	P-men	1.48	FE-no reaction	0.37
FET-taste/smell	6.65	FET-colour	1.48	FE-sex	0.37
H-body	5.73	FE-nice/pleasant/pleasure	1.29	I-fantasy/magic	0.37
FE-desire	5.36	FET-sweet	1.29	LD-theft	0.37
H-health	4.25	FET-price	1.29	LD-hiding	0.37
E-event	4.25	FE-love	1.11	EN-tech	0.37
F-composition	4.07	FE-mood	1.11	comparison	0.18
G-geo locations	3.51	FET-packaging	1.11	F-storage	0.18
F-bakery/cooking	3.33	H-beauty	0.92	H-dieting	0.18
E-transaction	3.33	E-religion	0.92	E-economy	0.18
FET-quantity	2.59	FE-senses	0.92	E-law	0.18
F-manufacturing	2.40	FE-seduction	0.92	FE-surprise	0.18
FE-passion	2.40	LD-drugs & addiction	0.92	FE-bribing	0.18
P-children	2.40	EN-animals	0.92	P-gay	0.18
CUL-artistic production	2.22	FET-energy	0.92	P-royalty	0.18
F-recipe	2.03	FE-memory	0.74	P-sharing/society	0.18
H-medicine	2.03	FE-guilt	0.74	G-spreading	0.18
C-gift	2.03	FE-relax	0.74	C-ceremonies	0.18
E-time	1.85	P-people	0.74	C-party	0.18
EN-dirt	1.85	FET-physical properties	0.74	EN-nature	0.18
F-drink	1.66	E-work	0.55	EN-house	0.18
F-product/shape	10.72	P-family	0.55		

Table 7_35. English *chocolate*: Semantic fields in the randomly sampled sub-corpus

semantic field	%	semantic field	%	semantic field	%
FET-quality/type	12.27	H-beauty	2.04	P-age	0.61
F-food	7.36	FE-happiness	2.04	S-sports	0.61
F-product/shape	6.54	FET-colour	1.84	E-language	0.41
F-bakery/cooking	6.54	F-drink	1.64	FE-no reaction	0.41
FET-taste/smell	6.13	E-history	1.64	FE-peace	0.41
F-recipe	5.73	F-manufacturing	1.43	FE-loneliness	0.41
comparison	4.70	E-time	1.43	P-sharing/society	0.41
G-geo locations	4.70	P-family	1.43	EN-nature	0.41
FE-passion	4.09	CUL-studying/intellect	1.43	EN-house	0.41
FE-nice/pleasant/pleasure	3.68	FET-physical properties	1.43	E-fair trade	0.20
FE-mood	3.68	FE-seduction	1.23	E-war	0.20
CUL-artistic production	3.68	L-existence	1.23	FE-love	0.20
E-event	3.48	FET-genuine	1.23	FE-memory	0.20
FET-quantity	3.48	FET-energy	1.23	FE-bribing	0.20
H-medicine	3.07	FE-comfort	1.02	P-women	0.20
P-children	3.07	C-gift	1.02	P-men	0.20
H-health	2.86	EN-dirt	1.02	P-friendship	0.20
FE-desire	2.86	FE-sex	0.82	G-spreading	0.20
E-transaction	2.66	FE-relax	0.82	I-fantasy/magic	0.20
F-composition	2.45	P-people	0.82	I-dream	0.20
H-dieting	2.25	FET-sweet	0.82	EN-tech	0.20
LD-drugs & addiction	2.25	F-storage	0.61	CUL-culture	0.20
H-body	2.04	FE-senses	0.61		

Table 7_36. Italian *chocolate*: Semantic fields in the randomly sampled sub-corpus

semantic field	%	semantic field	%	semantic field	%
FET-quality/type	12.80	F-storage	1.79	E-economy	0.30
G-geo locations	7.44	E-time	1.64	FE-nice/pleasant/pleasure	0.30
H-health	6.85	F-product/shape	1.49	FE-mood	0.30
FET-taste/smell	6.55	E-transaction	1.49	FE-passion	0.30
FET-price	5.36	FET-packaging	1.49	FE-comfort	0.30
F-drink	4.91	F-manufacturing	1.34	G-spreading	0.30
F-food	4.76	P-age	1.19	LD-theft	0.30
FET-quantity	3.72	EN-dirt	1.04	LD-drugs & addiction	0.30
E-excessive drinking	3.13	E-religion	0.89	CUL-culture	0.30
FE-happiness	3.13	E-event	0.89	CUL-studying/intellect	0.30
F-composition	2.83	C-gift	0.89	FET-sweet	0.30
comparison	2.53	FET-colour	0.89	F-serving	0.15
H-medicine	2.53	E-language	0.74	E-driving	0.15
FE-unpleasant	2.38	E-work	0.60	E-war	0.15
FE-relax	2.38	E-holidays	0.60	E-history	0.15
P-men	2.38	FE-no reaction	0.60	FE-seduction	0.15
P-sharing/society	2.38	FE-love	0.60	FE-memory	0.15
FE-desire	2.23	C-party	0.60	FE-peace	0.15
F-recipe	2.08	CUL-artistic production	0.60	FE-freedom	0.15
FET-physical properties	2.08	L-existence	0.60	FE-confidence	0.15
F-bakery/cooking	1.93	FE-senses	0.45	EN-nature	0.15
P-women	1.93	P-children	0.45	EN-house	0.15
P-friendship	1.93	P-people	0.45	L-future	0.15
P-posh	1.93	C-ceremonies	0.45	FET-genuine	0.15
P-family	1.93	H-body	0.30		

Table 7_37. English *wine*: Semantic fields in the randomly sampled sub-corpus

semantic field	%	semantic field	%	semantic field	%
FET-quality/type	12.86	FET-physical properties	1.94	LD-drugs & addiction	0.73
G-geo locations	8.25	F-composition	1.70	C-party	0.73
H-health	6.55	FE-confidence	1.70	EN-nature	0.73
FET-quantity	5.83	FE-children	1.70	EN-house	0.73
F-manufacturing	5.58	C-gift	1.70	F-product/shape	0.49
F-food	5.58	FET-colour	1.70	FE-mood	0.49
FE-friendship	5.34	F-drink	1.46	FE-posh	0.49
F-recipe	5.10	comparison	1.21	G-spreading	0.49
FET-taste/smell	4.61	F-serving	1.21	EN-dirt	0.49
E-language	3.88	E-history	1.21	EN-tech	0.49
H-medicine	3.64	E-driving	1.21	L-existence	0.49
E-excessive drinking	3.64	E-time	1.21	H-dieting	0.24
FE-unpleasant	3.40	FE-love	0.97	FE-senses	0.24
FE-family	2.91	CUL-culture	0.97	FE-desire	0.24
F-bakery/cooking	2.67	FET-sweet	0.97	FE-sex	0.24
F-storage	2.67	FET-genuine	0.97	FE-passion	0.24
CUL-artistic production	2.67	FET-price	0.97	FE-competitiveness	0.24
E-religion	2.43	FET-packaging	0.97	FE-comfort	0.24
E-event	2.43	E-work	0.73	FE-freedom	0.24
FE-nice/pleasant/pleasure	2.43	FE-no reaction	0.73	FE-women	0.24
CUL-studying/intellect	2.43	FE-relax	0.73	FE-men	0.24
E-transaction	2.18	FE-peace	0.73	FE-royalty	0.24
FE-happiness	2.18	FE-sharing/society	0.73	S-sports	0.24

Table 7_38. Italian *wine*: Semantic fields in the randomly sampled sub-corpus

How do these results compare to the results obtained with the original elicited datasets? A summary of this comparison is provided in Table 7_39, below.

Randomly sampled corpus	Overall fields (%)	H Cnv (%)	H+M Cnv (%)	Spearman's Rho
English *chocolate*	84.09	97.14	94.92	0.931
Italian *chocolate*	79.07	96.88	96.36	0.950
English *wine*	86.90	97.14	98.08	0.961
Italian *wine*	94.05	100	98.15	0.935

Table 7_39. Randomly sampled sub-corpora: semantic field results

As Table 7_39 reports, the randomly sampled sub-corpora showed almost 100% of the highly conventionalised fields in the original datasets, and a slightly lower percentage of the cultural associations (always exceeding 96%). Furthermore, Spearman's test highlighted very strong correlation with the values in the original datasets.

As regards semantic fields, the random sub-corpora proved markedly more representative of the original datasets than the 4-lemma sampled sub-corpora. This is evident at all the six levels of analysis considered in the table.

7.4.2 Conceptual domain analysis

The randomly sampled corpora were analysed also at the broader level of conceptual domains, where they retrieved the domains reported in Table 7_40. Values are expressed as percentages of the total number of sentences in the sub-corpus. R stands for rank. Cnv shows the conventionalisation level of that domain in the whole dataset (see Chapter 6). Bold signals the absence of that particular domain in the sampled sub-corpus. Domains are listed in alphabetical order.

Domain	***Chocolate* Eng.**			***Chocolate* It.**			***Wine* Eng.**			***Wine* It.**		
	%	**R**	**Cnv**	**%**	**R**	**Cnv**	**%**	**R**	**Cnv**	**%**	**R**	**Cnv**
Ceremony	2.40	9	L	1.02	13	H	1.93	9	L	2.43	9	M
Comparison	0.18	14	L	4.70	9	H	2.53	8	M	1.21	10	M
Culture	2.22	10	M	5.32	7	H	1.19	11	L	6.07	8	H
Environment	3.51	8	M	2.04	11	M	1.34	10	M	2.43	9	M
Events	12.01	5	H	10.02	5	M	10.71	5	M	18.93	3	M
Features	24.40	3	M	28.43	2	M	33.33	1	M	30.83	1	M
Feelings & emotions	26.99	2	M	22.70	3	M	13.69	4	M	15.05	4	H
Food	31.42	1	M	32.31	1	M	21.28	2	M	26.46	2	M
Geo	3.70	7	H	4.91	8	M	7.74	7	H	8.74	7	M
Health & Body	13.12	4	M	12.27	4	M	9.67	6	M	10.44	6	M
Imagination	0.37	13	M	0.41	15	M	**0.00**		NC	**0.00**		L
Life	0.74	12	M	1.23	12	L	0.74	12	M	0.49	12	H
Loss & damage	1.66	11	L	2.25	10	M	0.60	13	L	0.73	11	M
People	7.39	6	H	6.95	6	M	14.58	3	H	11.89	5	H
Sports	**0.00**		NC	0.61	14	L	**0.00**		NC	0.24	13	NC

Table 7_40. Conceptual domains in the *chocolate*, and *wine* randomly sampled sub-corpora

At the level of conceptual domains, the four randomly sampled corpora retrieved over 92% of the domains present in the original datasets, and all of the high conventionalisation domains, as well as of the cultural associations. Finally, correlation results were always in the strongest range. These results are summarised in Table 7_41.

Randomly sampled corpus	**Overall domains (%)**	**H Cnv (%)**	**H+M Cnv (%)**	**Spearman's Rho**
English *chocolate*	93.33	100	100	0.982
Italian *chocolate*	100	100	100	0.968
English *wine*	92.86	100	100	0.995
Italian *wine*	93.33	100	100	0.992

Table 7_41. Randomly sampled sub-corpora: conceptual domains results

As was the case with semantic fields, the randomly sampled corpora are more representative of the original datasets than the 4-lemma sampled corpora, at the qualitative as well as quantitative levels.

7.4.3 *Semantic field Assessment*

The results of the Assessment field in the randomly sampled corpora are summarised in Table 7_42 and Figure 7_2. In the figure, the four colours, labelled 1-4, indicate respectively Positive, Negative, Neutral, and Undecided assessment.

The results of the four randomly sampled corpora are perfectly in keeping with those of the corresponding whole datasets (Table 6_15, Chapter 6), showing a majority of positive sentences, a somehow smaller percentage of neutral sentences, followed by a yet smaller percentage of negative sentences, and a few undecided sentences.

	Positive	Negative	Neutral	Undecided
English *chocolate*	55.64	19.96	23.66	0.74
Italian *chocolate*	53.99	10.22	33.74	2.04
English *wine*	46.13	20.09	26.49	7.29
Italian *wine*	52.91	15.78	30.10	1.21

Table 7_42. Assessment field results in the randomly sampled sub-corpora

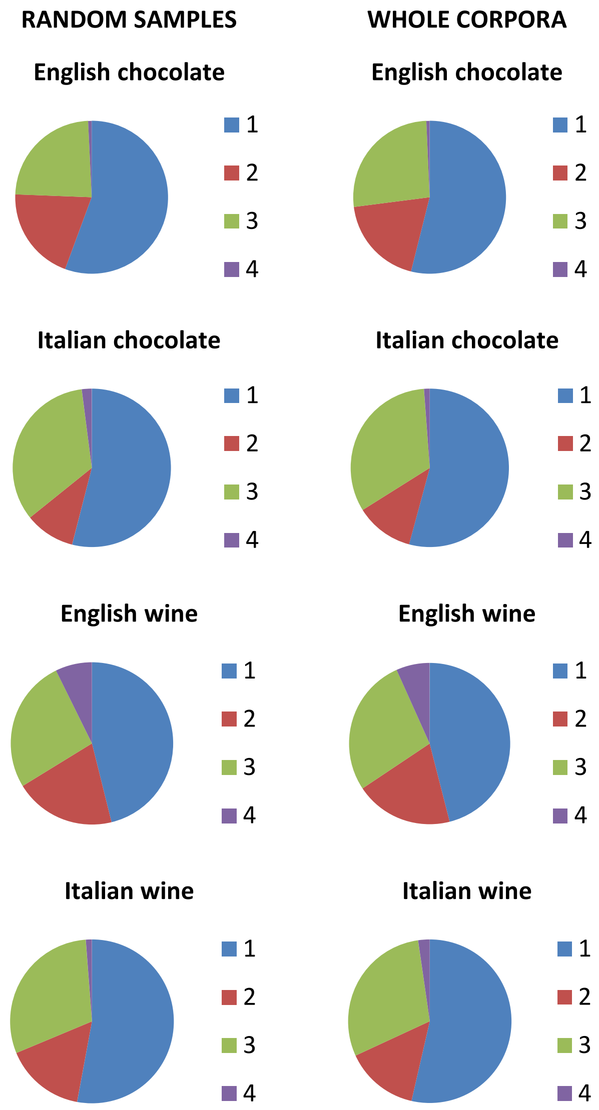

Figure 7_2. Assessment field results:randomly sampled datasets vs. whole elicited datasets

The high level of representativeness of the random corpora is evident not only rank-wise but also proportion-wise, as appears from Figure 7_2. In the Figure, the graphs on the left refer to the randomly sampled corpora, while those on the right to the corresponding elicited datasets.

7.4.4 Conventionalisation level analysis and cross-cultural comparison

For each semantic field and conceptual domain, Molinari's evenness index was computed, and three levels of conventionalisation were distinguished using confidence intervals. The results are reported in Tables 7_43-7_46, in order of conventionalisation. The 99% confidence intervals were: 0.74 - 0.98 for English *chocolate*; 0.74 - 1.00 for Italian *chocolate*; 0.71 – 1.00 for English *wine*, and 0.77 – 1.01 for Italian *wine*. The evenness values are reported in column G2,1, accompanied by indication of their corresponding levels of conventionalisation (column Cnv).

These results were compared to conventionalisation levels in the whole datasets (see Chapter 6, Tables 6_1, 6_2, 6_8 and 6_9 for semantic fields and 6_4 and 6_11 for conceptual domains).

Field	G2,1	Cnv	Field	G2,1	Cnv	Field	G2,1	Cnv
P-men	0.67	H	FET-sweet	0.77	M	FE-comfort	1.00	L
F-bakery/cooking	0.70	H	F-composition	0.78	M	FE-guilt	1.00	L
f-product/shape	0.70	H	P-children	0.78	M	FE-love	1.00	L
FET-taste/smell	0.70	H	E-event	0.78	M	FE-no reaction	1.00	L
P-family	0.71	H	FE-passion	0.78	M	FE-relax	1.00	L
H-health	0.72	H	FE-unpleasant	0.79	M	FE-sex	1.00	L
FE-desire	0.73	H	H-body	0.79	M	FE-seduction	1.00	L
FE-memory	0.73	H	E-time	0.80	M	FET-colour	1.00	L
FET-physical properties	0.73	H	EN-dirt	0.80	M	FET-energy	1.00	L
F-food	0.74	M	G-geo locations	0.80	M	FET-price	1.00	L
FE-happiness	0.74	M	C-gift	0.81	M	FET-quantity	1.00	L
FET-quality/type	0.75	M	F-recipe	0.81	M	H-medicine	1.00	L
E-religion	0.75	M	CUL-artistic production	0.82	M	I-fantasy/magic	1.00	L
EN-animals	0.75	M	E-transaction	0.84	M	LD-drugs & addiction	1.00	L
FE-senses	0.75	M	E-fair trade	1.00	L	LD-hiding	1.00	L
H-beauty	0.75	M	E-work	1.00	L	LD-theft	1.00	L
FE-mood	0.76	M	EN-tech	1.00	L	P-people	1.00	L
FET-packaging	0.76	M	F-drink	1.00	L			
FE-nice/pleasant/pleasure	0.77	M	F-manufacturing	1.00	L			

Table 7_43. English *chocolate* random sub-corpus: Conventionalisation results

								Cn
Field	**G2,1**	**Cnv**	**Field**	**G2,1**	**Cnv**	**Field**	**G2,1**	**v**
FET-sweet	0.58	H	H-medicine	0.79	M	FE-comfort	1.00	M
F-recipe	0.60	H	LD-drugs & addiction	0.80	M	FE-relax	1.00	M
F-product/shape	0.66	H	CUL-artistic production	0.80	M	FE-peace	1.00	M
P-children	0.69	H	H-body	0.80	M	FE-loneliness	1.00	M
FET-quality/type	0.69	H	H-beauty	0.80	M	P-age	1.00	M
FE-mood	0.70	H	FE-happiness	0.80	M	P-sharing/society	1.00	M
comparison	0.72	H	FE-passion	0.81	M	P-people	1.00	M
G-geo locations	0.72	H	H-health	0.83	M	P-family	1.00	M
FE-sex	0.73	H	E-event	0.84	M	EN-nature	1.00	M
F-food	0.74	M	FET-quantity/type	0.84	M	EN-house	1.00	M
C-gift	0.75	M	F-drink	1.00	M	CUL-studying/intellect	1.00	M
EN-dirt	0.75	M	F-manufacturing	1.00	M	L-existence	1.00	M
F-seduction	0.76	M	F-composition	1.00	M	FET-physical properties	1.00	M
H-dieting	0.77	M	F-storage	1.00	M	FET-colour	1.00	M
FE-nice/pleasant/pleasure	0.77	M	E-language	1.00	M	FET-genuine	1.00	M
E-transaction	0.78	M	E-history	1.00	M	FET-energy	1.00	M
FET-taste/smell	0.78	M	E-time	1.00	M	S-sports	1.00	M
FE-desire	0.78	M	FE-no reaction	1.00	M			
F-bakery/cooking	0.79	M	FE-senses	1.00	M			

Table 7_44. Italian *chocolate* random sub-corpus: Conventionalisation results

Field	G2,1	Cnv	Field	G2,1	Cnv	Field	G2,1	Cnv
G-geo locations	0.35	H	P-family	0.78	M	FE-love	1.00	M
P-men	0.59	H	H-health	0.78	M	FE-nice/pleasant/pleasure	1.00	M
FET-quality/type	0.62	H	F-drink	0.78	M	FE-mood	1.00	M
FE-unpleasant	0.64	H	FET-quantity	0.78	M	FE-passion	1.00	M
F-storage	0.66	H	FET-physical properties	0.78	M	FE-comfort	1.00	M
FE-relax	0.70	H	F-food	0.79	M	P-children	1.00	M
FE-happiness	0.71	H	P-sharing/society	0.79	M	P-age	1.00	M
E-excessive drinking	0.71	H	E-transaction	0.80	M	G-spreading	1.00	M
FE-senses	0.71	H	FET-packaging	0.80	M	LD-theft	1.00	M
P-people	0.71	H	P-women	0.82	M	LD-drugs & addiction	1.00	M
FET-price	0.74	M	F-composition	0.85	M	C-ceremonies	1.00	M
FET-taste/smell	0.75	M	F-product/shape	1.00	M	C-party	1.00	M
E-religion	0.76	M	F-manufacturing	1.00	M	C-gift	1.00	M
E-event	0.76	M	F-recipe	1.00	M	CUL-artistic production	1.00	M
P-friendship	0.77	M	H-body	1.00	M	CUL-culture	1.00	M
FE-desire	0.77	M	E-work	1.00	M	CUL-studying/intellect	1.00	M
EN-dirt	0.77	M	E-language	1.00	M	L-existence	1.00	M
comparison	0.78	M	E-economy	1.00	M	FET-colour	1.00	M
H-medicine	0.78	M	E-holidays	1.00	M	FET-sweet	1.00	M
F-bakery/cooking	0.78	M	E-time	1.00	M			
P-posh	0.78	M	FE-no reaction	1.00	M			

Table 7_45. English *wine* random sub-corpus: Conventionalisation results

Field	G2,1	Cnv	Field	G2,1	Cnv	Field	G2,1	Cnv
CUL-artistic production	0.65	H	E-event	0.80	M	FE-relax	1.00	M
FE-unpleasant	0.68	H	H-health	0.80	M	FE-peace	1.00	M
F-recipe	0.71	H	F-bakery/cooking	0.81	M	P-children	1.00	M
LD-drugs & addiction	0.71	H	P-family	0.82	M	P-posh	1.00	M
F-food	0.72	H	E-excessive drinking	0.83	M	P-sharing/society	1.00	M
F-manufacturing	0.72	H	E-language	0.84	M	G-spreading	1.00	M
P-friendship	0.72	H	FET-taste/smell	0.85	M	C-party	1.00	M
FET-sweet	0.73	H	F-drink	1.00	M	C-gift	1.00	M
comparison	0.75	H	F-serving	1.00	M	EN-nature	1.00	M
FET-quality/type	0.76	H	F-storage	1.00	M	EN-house	1.00	M
E-religion	0.77	M	H-medicine	1.00	M	EN-dirt	1.00	M
FE-nice/pleasant/pleasure	0.77	H	E-history	1.00	M	EN-tech	1.00	M
CUL-studying/intellect	0.77	H	E-driving	1.00	M	CUL-culture	1.00	M
F-composition	0.77	M	E-work	1.00	M	L-e1istence	1.00	M
FE-confidence	0.77	M	E-time	1.00	M	FET-physical properties	1.00	M
FET-colour	0.77	M	FE-no reaction	1.00	M	FET-genuine	1.00	M
G-geo locations	0.78	M	FE-love	1.00	M	FET-price	1.00	M
FET-quantity	0.78	M	FE-happiness	1.00	M	FET-packaging	1.00	M
E-transaction	0.79	M	FE-mood	1.00	M			

Table 7_46. Italian *wine* random sub-corpus: Conventionalisation results

Comparison between conventionalisation levels in the randomly sampled sub-corpus and in the whole dataset showed matching conventionalisation levels in highly variable percentages: 37,5% for English *chocolate*; 43.6% for Italian *chocolate*; 34.4% for English *wine*; and 25% for Italian *wine*. However the real focus of this work are cultural associations, which include fields with medium conventionalisation, as well as those with high conventionalisation. Consequently, if we disregard the distinction between H and M conventionalisation, in the randomly sampled sub-corpora the following percentages of cultural associations were correctly indicated: 93.9% for English *chocolate*; 98% for Italian *chocolate*; 78.7% for English *wine*; and about 89.3% for Italian *wine*.

At the level of conceptual domains, the *chocolate* randomly sampled sub-corpora showed 30.8% matches for English and 66.7% for Italian, while the *wine* sub-corpora showed 69.2% matches for English and 53.8% for Italian. However, if we disregard the distinction between H and M conventionalisation, in the randomly sampled sub-corpora the following percentages of cultural associations were correctly indicated, at the level of conceptual domains: 81.8% for English *chocolate*; 100% for Italian *chocolate*; 90% for English *wine*; and about 100% for Italian *wine*.

Thus, the randomly sampled corpora, proved slightly more representative of the original datasets than the 4-lemma sampled corpora also at the conventionalisation analysis. However, semantic fields or domains were identified as having the correct conventionalisation level or as being cultural association in highly variable percentages in the different sub-corpora and analytical situations.

Finally, the English and Italian semantic associations in the random sub-corpora were compared by means of Welch *t* test, in order to highlight the cases when the difference in means was statistically significant. T-test results were then triangulated with conventionalisation results, applying the procedure adopted in Chapter 6 to understand which differences could be safely attributed to culture and which to circumstantial elements, such as population sampling. The results are summarised in Tables 7_47 and 7_48.

While in Chapter 6 I considered only t-test results significant for P < 0.01, in the current experiments I extended the significance level to 0.05, as a consequence of the smaller size of the datasets analysed. Consequently, considering the 0.05 level of significance, in the random sub-corpora, the following semantic fields would appear as distinctively more prominent for Italians than for the English, when talking about *chocolate*: Comparison; Recipe; Dieting; History; Mood; Studying/intellect; Genuine. On the other hand, more prominent for the English than for Italians appear to be: Unpleasant; Quality; and Packaging. As regards conceptual domains, only Comparison would appear as prevalent in Italian rather than in English. No domain emerges as predominantly English.

Field	**P**	**T**	**ff**	**st.error**	**mean values**	**Cnv**	**mean values**	**Cnv**
	(< 0.05)			**of df**	**English**		**Italian**	
comparison	0.0001	4.2430	64	0.083	0.01	NC	**0.37**	**H**
CUL-studying/intellect	0.0071	2.7839	62	0.040	0.00	NC	**0.11**	**M**
E-history	0.0039	3.0030	61	0.042	0.00	NC	**0.13**	**M**
F-bakery/cooking	0.0026	3.0720	114	0.098	0.21	H	**0.51**	**M**
FE-guilt	0.0448	2.0241	85	0.023	**0.05**	L	0.00	NC
FE-happiness	0.0017	3.1821	141	0.091	**0.45**	M	0.16	M
FE-mood	0.0106	2.6136	83	0.083	0.07	M	**0.29**	**H**
FE-nice/pleasant/pleasure	0.0124	2.5497	88	0.081	0.08	M	**0.29**	M
FE-passion	0.0478	1.9948	123	0.084	0.15	M	**0.32**	M
FET-genuine	0.0131	2.5547	62	0.037	0.00	NC	**0.10**	**M**
FET-packaging	0.0331	2.1534	85	0.032	**0.07**	**M**	0.00	NC
FET-price	0.0074	2.7274	85	0.030	**0.08**	L	0.00	NC
FET-quality/type	0.0031	3.0273	106	0.144	**0.52**	**M**	0.95	H
FE-unpleasant	0.0060	2.8037	85	0.037	**0.10**	**M**	0.00	NC
F-recipe	0.0060	2.8196	79	0.113	0.13	M	**0.44**	**H**
H-body	0.0147	2.4612	146	0.080	**0.36**	M	0.16	M
H-dieting	0.0074	2.7641	66	0.059	0.01	NC	**0.17**	**M**
P-women	0.0179	2.3898	120	0.037	**0.10**	L	0.02	NC

Domains	**P**	**T**	**ff**	**st.error**	**mean values**	**Cnv**	**mean values**	**Cnv**
	(< 0.05)			**of df**	**English**		**Italian**	
Comparison	0.0001	4.2430	64	0.083	0.01	NC	**0.37**	**M**
Food	0.0489	1.9806	141	0.280	1.95	H	**2.51**	M
Culture	0.0051	2.8705	89	0.096	0.14	M	**0.41**	M
Features	0.0037	3.0411	147	0.227	1.52	M	**2.21**	M

Table 7_47. *Chocolate* random sub-corpora:
T-Test results for semantic fields and conceptual domains

Field	P (< 0.05)	T	ff	st.error of df	mean values English	Cnv	mean values Italian	Cnv
CUL-artistic production	0.0500	2.2715	150	0.059	0.04	M	**0.18**	**H**
CUL-studying/intellect	0.0219	2.3442	70	0.059	0.02	M	**0.16**	**H**
E-holidays	0.0449	2.0346	89	0.022	**0.04**	**M**	0.00	NC
E-language	0.0026	3.1037	80	0.065	0.06	M	**0.26**	M
F-drink	0.0004	3.2354	150	0.083	**0.37**	M	0.10	M
FE-confidence	0.0372	2.1252	67	0.048	0.01	NC	**0.11**	M
FE-desire	0.0037	2.9696	108	0.051	**0.17**	**M**	0.02	NC
FE-nice/pleasant/pleasure	0.0219	2.3442	70	0.059	0.02	M	**0.16**	**H**
FE-relax	0.0340	2.1430	128	0.060	**0.18**	**H**	0.05	M
FET-price	0.0000	4.4645	122	0.075	**0.40**	M	0.06	M
F-manufacturing	0.0033	3.0206	78	0.089	0.10	M	**0.37**	**H**
F-recipe	0.0409	2.0749	89	0.088	0.16	M	**0.34**	**H**
P-age	0.0041	2.9467	89	0.030	**0.09**	**H**	0.00	NC
P-friendship	0.0342	2.1482	95	0.098	0.14	M	**0.35**	**H**
P-men	0.0087	2.6755	102	0.060	**0.18**	**H**	0.02	NC
P-posh	0.0236	2.2901	129	0.049	**0.14**	M	0.03	M
P-sharing/society	0.0175	2.4041	137	0.054	**0.18**	M	0.05	M
P-women	0.0039	2.9450	115	0.044	**0.14**	**M**	0.02	NC

Domain	P (< 0.05)	T	ff	st.error of df	mean values English	Cnv	mean values Italian	Cnv
Events	0.0122	2.5454	120	0.180	0.80	M	**1.26**	M
Culture	0.0008	3.5083	76	0.090	0.09	L	**0.40**	M

Table 7_48. *Wine* random sub-corpora:
T-Test results for semantic fields and conceptual domains

Considering the 0.05 level of significance, the following semantic fields would appear as distinctively more prominent for the Italians than for the English, when talking about *wine*: MANUFACTURING; RECIPE; NICE/PLEASANT/PLEASURE; CONFIDENCE; FRIENDSHIP; ARTISTIC PRODUCTION; and STUDYING/INTELLECT. On the other hand, more prominent for the English than for the Italians appear to be: HOLIDAYS; DESIRE; WOMEN; MEN; and AGE. As regards conceptual domains, no domain emerges as predominantly Italian or English.

These results are rather different from the ones obtained with the whole corpus, and described in Chapter 6, Section 6.2.2, as well as from the ones in the 4-lemma sub-corpora.

7.5 Conclusions

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, the present chapter explored three possible shortcuts to highlighting culture-based semantic associations of a key word. The first method applied manual semantic analysis to the top 50/100/150/200/250/300 content words in the wordlist; the second one used the top 4 content words to create a sub-corpus which was manually analysed sentence by sentence; the third applied random sampling techniques to create a sub-corpus which was manually analysed sentence by sentence. The results of these experiments were compared – both qualitatively and quantitatively – to those in Chapter 6, and to each other. Tables 7_49 and 7_50 offer a comparative summary of the results, with reference to semantic fields and conceptual domains, respectively.

	Top 300 words				4-lemma sampling				Random sampling			
	Fields (%)	H Cnv (%)	H+M Cnv (%)	Rho	Fields (%)	H Cnv (%)	H+M Cnv (%)	Rho	Fields (%)	H Cnv (%)	H+M Cnv (%)	Rho
Chocolate - UK	68.18	91.43	86.44	0.810	83.0	100	94.92	0.903	84.09	97.14	94.92	0.931
Chocolate - IT	65.12	90.63	87.27	0.881	73.3	96.90	94.55	0.894	79.07	96.88	96.36	0.950
Wine - UK	70.59	94.29	94.23	0.877	79.7	97.10	96.15	0.905	86.90	97.14	98.08	0.961
Wine - IT	69.95	86.67	87.04	0.859	72.6	95.60	94.44	0.919	94.05	100	98.15	0.935

Table 7_49. Semantic fields: Summary of results

	Top 300 words				4-lemma sampling				Random sampling			
	Dom. (%)	H Cnv (%)	H+M Cnv (%)	Rho	Dom. (%)	H Cnv (%)	H+M Cnv (%)	Rho	Dom. (%)	H Cnv (%)	H+M Cnv (%)	Rho
Chocolate - UK	86.67	100	90.91	0.813	93.33	100	100	0.911	93.33	100	100	0.982
Chocolate - IT	93.33	100	100	0.963	93.33	100	100	0.965	100	100	100	0.968
Wine - UK	92.86	100	100	0.969	92.86	100	100	0.977	92.86	100	100	0.995
Wine - IT	86.67	100	100	0.924	86.67	100	100	0.977	93.33	100	100	0.992

Table 7_50. Conceptual domains: Summary of results

The top 300 content words retrieved 65-70% of the total number of semantic fields in the whole datasets, 86-94% of the highly conventionalised fields and an almost identical percentage of the cultural associations. The top four words in the frequency wordlist, treated as lemmas, provided sub-corpora whose size varied between 25% and 35% of the corresponding original dataset and showed 72.6-83% of the semantic fields in the datasets, corresponding to over 95% of the highly conventionalised fields in the original datasets, and 94-96% of the cultural associations. Finally, the randomly sampled corpora, identical in size to the 4-lemma ones, showed 79-94% of the semantic fields in the datasets, corresponding to 96-100% of the highly conventionalised fields and 94-98% of the cultural associations.

Results were systematically higher when considering a less fine-grained tagging scheme, i.e. when analysing conceptual domains, composed of a smaller number of higher and broader semantic categories. In fact, all the routes considered managed to show 100% of the highly conventionalised domains and of the cultural associations, with the only exception of the top 300 words in the *chocolate* English wordlist which retrieved 100% of the high conventionalisation domains, but only 91% of the cultural associations.

If we look at Spearman's test results, showing the quantitative level of correspondence to the contents of the whole datasets, the top 300 words in its wordlist showed levels of correlation in the 0.810-0.881 range (for $p < 0.01$) at the level of semantic fields and in the 0.813-0.969 range at the level of conceptual domains; the 4-lemma sampled sub-corpora showed a higher degree of correlation, with results in the 0.894-0.919 range for semantic fields and in the 0.911-0.977 range for conceptual domains; finally, the randomly sampled sub-corpora showed even higher degrees of correlation, their results being in the 0.931-0.961 range for semantic fields and in the 0.968-0.995 range for conceptual domains.

Finally, separate analysis of the ASSESSMENT category, showed qualitative and quantitative results that are perfectly comparable to those of the whole dataset only when the random sampling technique was applied.

Thus, all the methods managed to highlight an interesting percentage of the semantic fields present in each dataset. More importantly, however, they retrieved almost all of the highly conventionalised fields and cultural associations, and their quantitative results showed strong to very strong level of correlation with those of the corresponding elicited dataset. However, the most representative route proved to be the random sampling one, as it systematically showed higher results that the others at all levels of analysis, including separate analysis of semantic field ASSESSMENT.

Furthermore, only the two sampling procedures provided data which could be used to autonomously assess semantic fields and domains in terms of conventionalisation, as distribution of fields and domains across subjects was known. This could not be done in the analysis of the most frequent words in the wordlist (route 1), because of lack of distributional information. The results obtained were encouraging, with the random procedure looking slightly more promising, but not brilliant. This is most probably due to the fact that conventionalisation analysis is strongly dependent on corpus size. The original datasets, which I deemed suitable in size for this type of analysis, were themselves small corpora. Sub-sets corresponding to 25-35% of the original size are probably too small for a correct autonomous interpretation of the data.

Finally, the English and Italian semantic associations in the sub-corpora were compared by means of Welch *t* test, in order to highlight the cases where the difference in means was statistically significant. T-test results were then triangulated with conventionalisation results, applying the procedure adopted in Chapter 6. Unfortunately, the results obtained with the sub-corpora were rather different from the ones obtained with the whole datasets. Indeed, this type of cross-cultural comparison is highly dependent on quantitative results, which, in turn are strongly connected to sample structure.

To conclude, all the routes tested in this chapter seem suitable and useful as shortcuts to a qualitative analysis of cultural semantic associations of a given node word. In fact, they highlighted almost all of the most frequent and highly conventionalised fields and domains. At a quantitative level, however, the creation of randomly sampled sub-corpora seems more promising than the other two, as it did not only highlight constantly higher percentages of semantic fields and conceptual domains, but also showed higher levels of correlation to the values in the original datasets.

Furthermore, the results of routes one and two, both based on the most frequent semantic items in the dataset, either in the form of word or of lemma with annexed semantic associations, seem to confirm Fleischer’s theory that cultural associations are at least partly connected to frequency. However, the results obtained with the 4-lemma procedure are rather similar to those obtained with the random sampling ones, but are not as good as the latter. This leaves me with a reasonable doubt that sampling by the most frequent lemmas does nothing more than ordinary random sampling plus some skewing of the data.

For this reason, from now on in this work, the 4-lemma sampling procedure will be discarded.