Chapter 5

The current study: materials and method

5.1 Introduction

The experimental part of the work will address the following general questions:
1. Looking at two elicited datasets on *chocolate* and *wine*, to what extent do these concepts have similar cultural mental associations in both Britain and Italy?
2. What analytical tools and methods are most suitable for this type of analysis?
3. Can semantic analysis of corpora created from unelicited texts and from general Web corpora in particular provide information about cultural specificities, as much as semantic analysis of elicited data does?

General question n. 1 will be operationalized in two steps, or Research Questions:
R.Q. 1: What are the semantic associations of *chocolate*, and *wine* in the Italian and English cultures?
R.Q. 2: What are the differences between the Italian and English cultures with reference to *chocolate*, and *wine*?

General question n. 2 will be operationalized in the following steps:
R.Q 3: Could we identify the cultural associations of the two words without coding the entire dataset?
R.Q. 4: Could we identify the cultural associations of the two words using an automatic semantic tagger?

Finally, general question n. 3 will be operationalized in the following research question:
R.Q. 5: Could we identify the cultural associations of the two words using a general (Web) corpus?

The case studies' topics – *chocolate* and *wine* – were selected following a series of considerations. First of all, it seemed reasonable to bank on the experience gained with the supervision of the students' work on *chocolate* and start the new work from this topic: *chocolate* had shown to be a promising area for cross-cultural comparison, and – rather importantly – a specific coding scheme was already available. This type of topic – a consumable – also seemed of possible interest in marketing and consumer research. The second topic – *wine* – was chosen for similar reasons: it is a consumable and hence could possibly interest marketing researchers; it seemed a promising topic for cross-cultural comparison, since Italy has a long tradition in wine making, while the UK has none and is historically a 'beer country' (on the expected difference between Italy and the UK, see Chapter 6, Section 6.1); we are still in the realm of

'food and drink' and the *chocolate* coding scheme could be easily tested on and adapted to the new topic. Furthermore, both *chocolate* and *wine* have clear, though varied referents in both Italy and the UK, which facilitates collecting and analysing elicited as well as Web data.

It must be said that the original project plan intended to deal also with some third topic of an abstract nature. Subsequently, in the light of the number of datasets to be created and analysed (4 for each topic) and the amount of quantitative analysis to be carried out on each dataset, the idea of a third case study was discarded. In fact, a complete analysis and description of a third case study would have taken me to exceed the time and space limits imposed by Lancaster University for a PhD work.

However, I believe that two topics could be considered a minimal acceptable number of case studies, given the rather complex research design of the current work. The overall research design and its rationale is introduced in the following paragraphs.

Cultural mental associations can be highlighted and analysed within a single culture, but they become more prominent when different cultures are compared. On the other hand, assessment of the most suitable data sources and analytical methods can be better achieved with inter-language comparisons. For these reasons, all case studies will include a series of inter-cultural analyses, as well as cross-cultural ones.

Furthermore, two common points can be seen in marketing research methods and the cultural studies quoted in the previous chapters: 1. the use of elicited data; and 2. the use of analytical methods based on manual semantic coding. Elicited data, however, are limited in extension and time-consuming to collect. Consequently one of my research hypotheses is that elicited data could be replaced with non-elicited data from large general Web corpora – easily collectable in large quantities. This hypothesis is supported by Bianchi (2007; see Chapter 4, Section 4.3), who compared the psychological associations (or EMUs) to *chocolate* in a specialised corpus created around the node word and using the Web as source for text retrieval, and in a general corpus (CORIS) of about 100 million words created according to more 'traditional' methods and criteria, such as sampling and representativeness (Rossini Favretti, Tamburini, & De Santis, 2002). Her results showed that the two corpora, though constructed with different criteria and purposes in mind, include samples which could be considered as coming from the same population.

Elicited data is normally coded manually. Manual coding is a highly time-consuming task, and the more the data, the more coding becomes frustrating and prone to errors. Once elicited data are replaced with (ample) corpus data, however, performing manual coding may become awkward and should ideally be substituted with automatic coding.

For these reasons a core element in my research design is comparison of Web data to elicited data, the latter being used as a control situation. The Web data will be analysed starting from frequency word lists, and considering a variable number of the most frequent items, in an attempt to find a shortcut to cultural features that does not require (manually) tagging the whole Web corpus.

A secondary element is comparison between manual and automatic coding. This element is secondary because it could be performed only in the English datasets. The latter underwent automatic tagging with Wmatrix, as well as manual tagging. For Italian, no automatic semantic tagger comparable to the Wmatrix one is available.

In the experimental part of my research, I adopted a fixed procedure and applied it in two case studies, respectively focusing on *chocolate* and *wine* in British and Italian minds. All the case studies described in the current work take advantage of:

- specifically created sets of elicited data;
- specifically created Web datasets, generated from the same general Web corpora;
- the same analytical procedures.

For an easier reading of the various case studies and in order to avoid tedious repetitions, the present chapter describes the materials and methods that are common to all of them. This includes a description of the questionnaires used for collecting the elicited data, the software used to access the Web corpora end extract specific datasets, and the semantic automatic tagger used for the British data.

5.2 Materials

5.2.1 Elicited data

5.2.1.1 The questionnaires

The elicited data were collected specifically for the purpose of this study, by means of questionnaires with sentence completion and sentence writing tasks. The questionnaires' organization was inspired by Hair, Bush and Ortinau (2009, p. 186; see Chapter 4) and by Wilson and Mudraya (2006; see Chapter 2).

In passing, it was noted that the sentence completion task helped collecting at least a minimum amount of data from all and any of the respondents. In fact, a small number of respondents, who were presumably little inspired by the key words of each questionnaire, limited themselves to completing the given sentences.

The questionnaires, which also featured a picture illustrating the node word, began with the following completion sentences (or their respective translations into Italian):

Chocolate	Wine
1. Whenever I think of chocolate I ...	1. Whenever I think of wine I ...
2. Chocolate reminds me of ...	2. Wine reminds me of ...
3. The picture on the top leads me to ...	3. The picture on the top leads me to ...
4. Chocolate can ...	4. Wine can ...
5. I would use chocolate to ...	5. I would use wine to ...
6. It's common knowledge that chocolate ...	6. It's common knowledge that wine ...

This task was followed by a request to write 20 sentences using the node word given. The limit of twenty was inspired by the Twenty Statement Test (Grace & Cramer, 2003) – a sentence-writing test used in psychology to study self-identity – of which Wilson and Mudraya (2006) adopted a reduced version (including only 10 sentences).

5.2.1.2 The respondents

The *chocolate* and *wine* questionnaires were distributed together, and to the same groups of respondents. The English questionnaires were first circulated via e-mail among British natives living in England (friends or friends' friends residing in the London and Cambridge areas); unfortunately only about 20 people replied. Subsequently, paper versions were distributed manually among British students in the University campus at Lancaster. The Italian questionnaires were circulated exclusively by e-mail, among friends, colleagues and a large number of students from the Universities of Salento (Lecce, Southern Italy) and Pavia (Northern Italy), where I worked at the time.

In the accompanying e-mail message, or when asking a person whether they accepted to fill in the questionnaires for research purposes, it was made clear that they could only do so if they were native speakers and lived in England or Italy. Thus, I managed to reach a total of 90 and 63 native speakers of English and Italian, respectively. Based on knowledge of the age of the people to whom I sent the questionnaires, I can estimate that the English respondents fall in the 18 to 60 age range and a little more than two thirds of them are university students (aged 18-25); the Italian respondents can be estimated to be in the 18-70 age range, with a little less than two thirds of them being university students in the 18-25 age group.

Social variables were not specifically collected because my elicited data will eventually be compared to Web data which cannot be controlled for that sort of variables. This may be considered a limitation to the study, and will have to be born in mind when drawing (cross)cultural conclusions.

Although some parallelism could be seen in the English and Italian sampled populations (a majority of university students; data collected in both Northern and Southern areas of the two countries), the sampling procedure adopted falls into the category of convenience sampling and was considered acceptable because, as we have seen in Chapter 4, it is customary practice in exploratory marketing studies.

5.2.1.3. The elicited datasets

Almost all the *chocolate* and *wine* respondents completed the sentence-completion task, while in the sentence-writing task some wrote less that 20 sentences, or even no sentence at all. Table 5_1 provides a detailed summary of the number of sentences volunteered by the respondents. In the Table, the first row lists the number of sentences entered by respondents, while C stands for *chocolate* survey and W for *wine* survey.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
C IT							1					1														3	55	3	
C UK							3	2	1	1	3	1	4	1	2	1	2	3	1	2					2	3	55		
W IT	1						1				1														1	4	53	1	1
W UK	4				2		4	1		2	1	1		2	4	2	4	1	3			1	2	1	14	7	39		

Table 5_1. Sentence distribution across *chocolate* and *wine* respondents

As the table illustrates, five respondents (1 Italian and 4 English ones) refused to participate in the *wine* survey, but took part in the *chocolate* one. Of the others, only two English respondents did not finish the sentence completion task and contributed to the survey with less than 6 sentences. What is really noticeable from the table is that, while the English wrote a variable number of sentences going from 6 to 26, with more than 30% of them contributing with less than 24 sentences, only 3% of the Italian respondents wrote less than 24 sentences, and five respondents even exceeded the required number. This is easily explained by the way the questionnaires were collected. As detailed in section 5.2.1.2, only 20 English native speakers replied to my questionnaires by e-mail, while the remaining 70 were 'recruited' on Lancaster University campus. On the other hand, the 63 Italian respondents were all volunteer participants who filled in the e-mail questionnaires.

Using the data thus gathered, four elicited datasets were created, as detailed in Tables 5_2 and 5_3. As the first task in each questionnaire was a sentence completion exercise, each of the datasets was 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. F1 was used when performing manual coding of the whole set of elicited data (see Section 5.3.1.2); F2 when performing manual coding of the wordlists and – for English only – automatic tagging of the data (see Section 5.3.2). Indeed, a quick look at the frequency wordlists had shown that the top positions were occupied by words given in sentence completion tasks. Consequently format F2 seemed the most suitable one to avoid frequency biases due to the presence of given text, when tagging individual words rather than sentences.

Furthermore, as regards the creation of wordlists, two different tools were used in the current study, under different circumstances: Wordsmith Tools (Scott, 2008), for cross-language comparisons, and Italian inter-language comparisons; and Wmatrix (Rayson, 2008), for English inter-language comparisons based on automatic tagging (see Section 5.3.2). The former tool, like most others of the same family, can only count individual words, while the latter detects multi-word units, such as cheer_up, chocolate_bar, and cocoa_beans[1] (see Section 5.3.2) and treats them as individual words. Hence marked differences in the word counts, as shown in Table 5_2.

	Chocolate	***Wine***
Total n. of respondents	87	91
Total n. of sentences	1886	1938
Mean n. of sentences	21.7 (SD = 6.58)	21.3 (SD = 6.57)
Mean sentence length	6.95 (SD 4.01)	7.29 (SD 4.62)
Running words (format F1)	12946	13740
Running words (format F2) – Wordsmith Tools	10576	11611
Running words (format F2) - Wmatrix	9967	10967

Table 5_2. Elicited data summary – English

[1] These examples are taken from the Wmatrix frequency list of the elicited chocolate corpus.

	Chocolate	*Wine*
Total n. of respondents	63	62
Total n. of sentences	1603	1573
Mean n. of sentences	25 (SD = 3.14)	25.4 (SD = 3.25)
Mean sentence length	8.35 (SD 3.59)	8.59 (SD 3.61)
Running words (format F1)	13447	13153
Running words (format F2)	11754	11607

Table 5_3. Elicited data summary – Italian

As the tables clearly show, the Italian respondents were more diligent than the English ones in accomplishing the required tasks and wrote on average 25 sentences each (with a standard deviation around 3), against the 21 sentences (and standard deviation around 6) of the English. Furthermore, the Italian sentences were usually slightly longer than the English ones.[2] These two factors explain why the English and Italian datasets are comparable in size, despite the smaller number of Italian respondents.

A few of the sentences in the elicited data (15 for *chocolate*, 21 for *wine*) were connected to the questionnaire or the situation, rather than to the node word (e.g.: *Sorry I have revision to do*; *I feel daft writing about chocolate*; *I don't know as much about chocolate as I do about wine*), or were ambiguous in their reference to the node word or pertinence to the purpose of the survey (e.g.: *Wine begins with w*; *There is no wine in winegums*), but it was decided not to remove them from the elicited corpora. In fact, deleting sentences of this type from the elicited data, but not from the Web corpora would have been pointless, if not altogether methodologically wrong. At the same time it would be impossible to identify (and remove) 'irrelevant' sentences from the Web corpora, given their size and the fact that in some cases the pragmatic context of the original texts might be unintelligible.

5.2.2 The Web datasets

The Web datasets used in the current research were extracted from two large, general corpora (UKWAC and ITWAC) created in the WACKY project,[3] and accessed using the Sketch Engine, an on-line interface which provides access to a series of large corpora in several languages and offers concordancing and other linguistic query tools.

The general Web corpora, the interface used to access them and the extracted datasets are detailed in the following paragraphs.

5.2.2.1 UKWAC and ITWAC

In all the experiments, primary source of Web data were the English and Italian WACKY corpora, namely UKWAC (Baroni & Kilgarriff, 2006; Baroni, Bernardini, Ferraresi, & Zanchetta, 2008) and ITWAC (Baroni, Kilgarriff, Pomikálek, & Rychlý, 2006; Baroni & Ueyama, 2006; Baroni, Bernardini, Ferraresi, & Zanchetta,

[2] Mean sentence length and sentence length SD were calculated using the Wordlist feature in WordSmith Tools v.6.

[3] A project headed by Silvia Bernardini and Marco Baroni and carried out with the help of several international names including Stefan Evert, Serge Sharoff, William Fletcher, and Adam Kilgarriff. See the following website: http://wacky.sslmit.unibo.it/doku.php.

2008). They are both large general corpora created from the Web using spidering tools. UKWAC includes about two billion running words; ITWAC almost 1.5 million words. Both corpora have been lemmatised and POS tagged with Tree-Tagger.[4]

UKWAC and ITWACK were created following a specific procedure, described in Baroni, Kilgarriff, Pomikálek, and Rychlý (2006). First of all, two separate sets of seeds were selected: the first set included randomly paired words extracted from a newspaper corpus (2000 mid-frequency words); the other, from a vocabulary list for language learners (about 653 content words). The lists of the retrieved URLs were reviewed in order to keep only one (randomly selected) URL for each domain. The URLs which remained were fed to the Heritrix crawler,[5] specifying parameters that excluded retrieval of non html-format documents and limited searches to the country-specific domain of each corpus (e.g.: *.it* for the Italian corpus). From the retrieved html documents, the following were filtered out: document under 5KB or above 200KB; duplicate documents, along with the original;[6] pages containing a low rate of content words (low presence of content words being an indicator of noise); and pages containing words relating to pornography (as the latter were considered another indicator of noise). The remaining pages were stripped of boilerplate – i.e. of all those elements of a Web page which are the same across many pages – using the heuristic of the Hyppia project BTE tool,[7] based on html tag density (high density indicates boilerplate; low density indicates content-rich sections). Finally, near-duplicates were eliminated, using "a simplified version of the 'shingling' algorithm (Broder *et al.*, 1997)" (*ibid*., 2006, p. 3) and considering near duplicates those pages that shared at least two 5-grams of the 20 5-grams extracted from each document. Subsequently, the documents were lemmatised using Tree-Tagger, and the corpus was further 'cleaned' of cues such as number of words not recognised by the lemmatiser, proportion of words with upper-case initial letters, proportion of nouns, and proportion of sentence markers.

UKWAC and ITWAC were compared to relatively large corpora which are widely used as reference corpora in linguistic analysis. UKWAC was compared to the British National Corpus (Baroni, Bernardini, Ferraresi, & Zanchetta, 2008), and ITWAC to *la Repubblica* corpus, collecting 16 years of daily news (Baroni & Ueyama, 2006, sec. 4.1).[8] Comparisons showed that each Web corpus includes most of the vocabulary of the corresponding reference corpus. In the case of UKWAC, the corpus was able to "provide rich, up-to-date language data on even relatively infrequent words" (Baroni, Bernardini, Ferraresi, & Zanchetta, 2008, Sec. 3.1). Hence my believing that the WaCky corpora could be suitable material for the semantic

[4] On POS tagging and lemmatization, see Chapter 3, sections 3.5.1. For more detailed information on Tree-Tagger and the tagsets used for tagging UKWAC and ITWAC, see http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

[5] http://crawler.archive.org

[6] The authors, in fact, noted that 'typically, such documents came from the same site and were warning messages, copyright statements and similar, of limited or no linguistic interest" (Baroni & Kilgarriff, 2006, p. 2).

[7] http://www.smi.ucd.ie/hyppia/

[8] As the authors explicate: "Despite its being single-source, this is widely used as an Italian reference corpus thanks to its size and the variety of newspaper contents" (Baroni, Bernardini, Ferraresi, & Zanchetta, 2008, Sec. 3.1).

analyses that will be performed in the current work. Finally, both corpora showed differences from the reference corpora in terms of register, and UKWAC also in terms of text types.

5.2.2.2 The Sketch Engine

The two Web corpora described above were accessed using the Sketch Engine (www.sketchengine.co.uk; Kilgarriff, Rychly, Smrz, & Tugwell, 2004), an on-line interface which provides access to dozens of large corpora in several languages and offers concordancing and other linguistic query tools. This interface also provides the possibility for users to create their own corpora using WebBootCaT – a suite of scripts for bootstrapping corpora and terms from the Web, starting from a list of 'seeds', i.e. "terms that are expected to be typical of the domain of interest" (Baroni & Bernardini, 2003, par. 5), as input –, or upload already assembled corpora and query them.

The Sketch Engine concordancing feature displays lines in KWIC format, but if desired a sentence view can be activated. Furthermore, if the corpus used is lemmatised and POS tagged, advanced search parameters can be set to look for a lemma instead of a word form and/or a specific grammatical category. The retrieved concordance lines or sentences can be saved on your local machine in a type of text-only format readable with a simple text editor.

Other features are available in the on-line interface, such as the creation of wordlists and word sketches – lists of collocates organised according to grammatical relation with node word –, but none of these features was used in the current study. The Sketch Engine was here only used to access WACKY corpora and extract sentences around the node words of interest.

5.2.2.3 Creating the project datasets

The Sketch Engine interface was set to access the UKWAC and ITWAC corpora alternatively, and concordances were generated for each of the project node words. In particular, as the corpora used are lemmatised and POS tagged, the concordance interface was set to look for lemmas and all POS forms for the English and Italian node words *chocolate/cioccolato*, and *wine/vino*. Subsequently, the interface was set to save 10,000 full sentences. This led to the creation of two datasets (one in English and one in Italian) for each node word. It was immediately noticed, however, that the retrieved data included several duplicated sentences. This was the case with the sentences that included more than one occurrence of the node word, which appeared in the retrieved data as many times in a row as the occurrences of the node word. Consequently, the datasets were manually purged of all duplicated sentences.

Table 5_4 details number of sentences and running words in the retrieved datasets, before and after purging them of duplicated sentences.

		Chocolate	Wine
English as retrieved	Sentences	10000	10000
	Running words - Wordsmith	426815	365312
English without duplicates	Sentences	8436	7343
	Running words - Wordsmith	302545	290122
	Running words - Wmatrix	286243	277006
Italian as retrieved	Sentences	10000	10000
	Running words - Wordsmith	487305	503451
Italian without duplicates	Sentences	8352	8239
	Running words - Wordsmith	310422	324640

Table 5_4. The Web datasets from the WACKY corpora

As for the English elicited data, two separate word counts of the English Web data are reported in the table, one calculated with Wordsmith Tools, and one with Wmatrix (see Section 5.3.2 for an explanation of Wmatrix's word counts). The Wmatrix wordlists, however, were used only for inter-language comparisons based on automatic tagging.

5.3 Procedure

5.3.1 Manual coding

The manual coding task was performed following the steps suggested by Neuedorf (2002). These include the creation of an initial Codebook, followed by several cyclical phases of coder training, coding and discussion, followed by codebook revision.

Manual coding was applied, sentence by sentence, when coding the elicited datasets and the Web datasets; furthermore it was used when coding, word by word, the most frequent items in the elicited wordlists.

5.3.1.1 Origin and development of the Codebook

Before starting the coding process of the elicited data, a Codebook was drafted which includes a detailed description of the coding scheme (with examples) and of its origin, and instructions on how to apply the coding scheme in the task at hand. The coding scheme is based on a two-layered classification that includes semantic fields and conceptual domains, two hierarchical levels of semantic analysis.

The coding scheme described in the Codebook originates in a preliminary experiment of manual coding of Web data focusing on the node word *chocolate* in Italian (Bianchi, 2007, briefly described in Chapter 4) and in English (Cogozzo, 2005). The original codes – developed by two graduate students under my supervision – were applied, discussed and reviewed twice before including them in the Codebook, version 1.

The annotation of the *Chocolate* and *Wine* English elicited datasets was separately performed by myself and another coder – an Italian graduate student with excellent competence in English – and began following Codebook version 1. During annotation, the two coders met twice to discuss the need for further semantic fields and/or conceptual domains. When a new semantic field was agreed upon and added to the list, each coder reviewed the sentences s/he had already tagged. Thus, the coding

scheme grew from 15 conceptual domains and 83 semantic fields to the 16 conceptual domains and 92 semantic fields listed in Table 1 in the Appendix, and the Codebook was updated to version 2.

The coding scheme described in Codebook v.2 was used for the manual tagging of the UKWAC *chocolate* and *wine* corpora and for comparing the English *chocolate* and *wine* elicited datasets to their corresponding UKWAC datasets and to automatic semantic tagging.

Manual coding of the Italian *chocolate* and *wine* datasets was accomplished at a later stage by the same coders and following the procedure described in the previous paragraphs. It commenced by using Codebook v.2 and eventually led to updating the Codebook to version 3 (in the Appendix) which includes 16 conceptual domains and 96 semantic fields (see Table 5_5 in Section 5.3.2, or Table 2 in the Appendix). This coding scheme was then used in the manual tagging of the ITWAC *chocolate* and *wine* corpora, in comparing the Italian *chocolate* and *wine* elicited datasets to their corresponding ITWAC datasets, and – after reviewing all the English datasets on *chocolate* and *wine* – in all cross-cultural comparisons.

5.3.1.2 Manually coding whole datasets

When manually coding whole datasets (be they elicited or retrieved from the Web), coding was always performed by myself and a second coder who had received specific training in the use of the coding scheme. A second coder was necessary to guarantee reliability of the coding system.[9] Coding was done manually and required the coders to assign one or more semantic fields (chosen among the ones given) to whole sentences on the basis of their assessment of the semantic fields that were explicitly or implicitly mentioned in the given sentence.

In the elicited datasets, the unit of data collection was the questionnaire, while the unit of analysis was the sentence. In the Web datasets, units of data collection and units of analysis was always the sentence.

Decisions about the most suitable categories to assign to each sentence were usually triggered by specific words in the sentence (e.g. *Very good chocolate may be expensive* = PRICE; *Chocolate is good for your health* = HEALTH), but also by context (e.g. *So is Bulgarian wine* can only be understood in connection to the sentence that precedes it: *Chilean wine is good*), and/or general knowledge of the world (e.g. *I eat chocolate before sitting an exam* = ENERGY, because it's common knowledge that an exam is a hard task that drains your energy). In cases of disagreement between the two coders (on average about 3%), the suggestions of both were accepted. This solution was made possible by the fact that the task accepted that an unlimited number of codes be assigned to each sentence. Consequently, if Coder A though that sentence *Chocolate salami: made of extra dark chocolate with roasted hazels* was Composition, and Coder B thought it was Recipe, both tags were matched to the sentence.

At different stages in the coding process, the two coders met to discuss the need for further semantic fields and/or conceptual domains. When the need for new

[9] Inter-coder reliability, also called reproducibility, is one of the three forms of reliability used in content analysis, along with stability of coding by the same coder, and accuracy which can be described as correspondence of the text classification with standard norms (Weber, 1990).

semantic fields or conceptual domains was agreed upon, each coder reviewed the sentences s/he had already tagged, and the Codebook was updated.

5.3.1.3 Manually coding wordlists

Frequency wordlists were generated from the elicited data, and the most frequent items in the wordlists were coded manually by myself. Coding was repeated twice to determine stability, i.e. one of the three forms of reliability described in Weber (1990). The steps used to create and code the wordlists are described in Chapters 6, 7 and 8.

5.3.1.4 The coding scheme

As hinted at in the previous sections, the coding scheme is based on a two-layered, hierarchical classification that includes semantic fields – lower level, finer grained categories – and conceptual domains – higher level, broader categories. Multi-layered classifications like the one I used are not an uncommon event in content analysis (see for example Guerrero, Claret, Verbeke *et al.*, 2010, reviewed in Chapter 4; and the semantic categories in the USAS tagset, described in Section 5.3.2). A list of the semantic fields and conceptual domains used in the *wine* and *chocolate* studies is provided below (Table 5_5).

Conceptual domains	**Semantic fields**
Food [F]	Product/shape; Bakery/cooking; Manufacturing; Food; Composition; Recipe; Drink; Storage; Serving
Health & Body [H]	Dieting; Health; Medicine; Body; Beauty
Events [E]	Playing; Language/etymology; Economy; Religion/mythology; War; History; Law; Event; Transaction; Fair Trade; Time; Work; Driving; Excessive drinking; Holidays
Feelings & Emotions [FE]	No reaction; Unpleasant; Senses; Love; Desire; Nice/Pleasant/Pleasure; Sex; Happiness; Seduction; Mood; Passion; Competitiveness; Memory; Surprise; Loneliness; Freedom; Persuasion; Guilt; Comfort; Relax; Peace; Bribing; Confidence
People [P]	Women; Men; Gay; Children; Posh; Friendship; Royalty; Sharing/society; People; Family; Age
Geography [G]	Geographical locations; Spreading
Imagination [I]	Fantasy/magic; Dream
Loss & Damage [LD]	Theft; Drugs and addiction; Hiding
Ceremonies [C]	Ceremonies; Party; Gift
Environment & Reality [EN]	Nature; Animals; House; Dirt; Technology
Culture [CUL]	Artistic production; Culture; Studying/intellect
Life [L]	Future; Existence
Features [FET]	Quality/type; Colour; Sweet; Genuineness; Energy; Taste/Smell; Quantities; Price; Packaging; Physical properties
Sports [S]	Sports
Comparison [COM]	Comparison
Assessment	Assessment

Table 5_5. *Chocolate* and *Wine*: Summary of semantic fields and conceptual domains

Column one lists the conceptual domains (16 in all); the letters in squared brackets are initials which will be used in the current work to refer to domains when space does not allow mentioning the full name (e.g.: in tables). Column two lists the semantic fields

(96 in all). Further details on the coding scheme, including a definition of each semantic field and examples of sentences can be found in the Appendix (Table 3).

There is certainly a level of arbitrariness in the choice and naming of these categories, but this does not represent a problem in so far as they were applied systematically to all the data under investigation. Furthermore, explanations were provided in the Codebook to assist the coders in understanding the boundaries of each category. In fact, when creating a classification an important feature is that there is no overlapping between categories.

Semantic fields and conceptual domains were inspired by the data, and grew in number as more and new datasets were analysed.

5.3.2 Automatic semantic tagging

Automatic semantic tagging was also applied and compared to the manual one. Automatic tagging was achieved using Wmatrix (Rayson, 2008), a fully-automated and user-friendly on-line interface developed at the Lancaster's University Centre for Computer Corpus Research on Language (UCREL) for performing semantic tagging on text files in English. Unfortunately, however, no automatic semantic tagger comparable to the Wmatrix one exists for Italian. Consequently, only the English elicited and Web datasets could be analysed automatically.

The English elicited and Web datasets underwent automatic semantic tagging, using Wmatrix (Rayson, 2008).

In Wmatrix, semantic tagging is preceded by POS tagging and lemmatisation. POS tagging is performed using Claws - Constituent Likelihood Automatic Word-tagging System (Garside & Smith, 1997) and its standard CLAWS 7 tagset.[10] This probabilistic tagger, developed at UCREL and used for tagging the BNC,[11] reaches an accuracy of 96-98 % (Rayson, Archer, Piao, & McEnery, 2004). The semantic tagging component (described in Wilson & Rayson, 1993; Rayson, Archer, Piao, & McEnery, 2004; Archer, Rayson, Piao, & McEnery, 2004) includes a single word lexicon of 42,000 entries, and multi-word expression (MWE) templates, with 18,400 entries in all. Furthermore, it includes context rules and disambiguation algorithms for the selection of the correct semantic category. This semantic tagging process performs with a 92% accuracy rate (Piao, Rayson, Archer, & McEnery, 2004, quoted in Archer Rayson, Piao, & McEnery, 2004).

The semantic categories used in the system were originally based on the Longman Lexicon of Contemporary English (LLOCE) (McArthur, 1981), though some changes were subsequently made (Rayson, Archer, Piao, & McEnery, 2004). The current ontology includes 21 fields (Table 5_6), subdivided into 232 categories with up to three subdivisions, for a total of 453 tags.

Originally developed for automatic content analysis of elicited data, such as in-depth survey interviews (Wilson, 1993; Wilson & Rayson, 1993), the USAS tagset has been used with interesting results in several corpus linguistic studies on a range of different topics, from stylistic analysis of prose literature to the analysis of doctor-

[10] List of tags available at: http://ucrel.lancs.ac.uk/claws7tags.htm.
[11] See Chapter 3, Note 13.

patient interaction, and from translation to cross-cultural comparisons (see http://ucrel.lancs.ac.uk/usas).

A - General & Abstract Terms	N - Numbers & Measurement
B - The Body & the Individual	O - Substances, Materials, Objects & Equipment
C - Arts & Crafts	P - Education
E - Emotional Actions, States & Processes	Q - Linguistic Actions, States & Processes
F - Food & Farming	S - Social Actions, States & Processes
G - Government & the Public Domain	T - Time
H - Architecture, Building, Houses & the Home	W - The World & Our Environment
I - Money & Commerce	X - Psychological Actions, States & Processes
K - Entertainment, Sports & Games	Y - Science & Technology
L - Life & Living Things	Z - Names & Grammatical Words
M - Movement, Location, Travel & Transport	

Table 5_6. Semantic fields in the UCREL Semantic Analysis System tagset

At the end of the tagging process, Wmatrix publishes the output in several different formats, including a semantic frequency list. Furthermore, it offers features for generating a 'traditional' keyword list and a semantic keyword list, using the BNC as reference corpus.[12] The semantic frequency list produced by Wmatrix lists the USAS categories present in the dataset, in order of frequency. The semantic keywords list shows the key USAS categories in the dataset, compared to those in the reference corpus.

5.4 Research design

The present section illustrates the core research design adopted in the current study. This design was systematically applied to each of the key words selected for analysis (*chocolate*, and *wine*).

The elicited and Web datasets and the wordlists were compared to each other in several ways, in order to highlight the dominant EMUs in the given cultures and assess the advantages and disadvantages of the different analytical methods. Qualitative as well as quantitative analyses will be performed, at the level of both semantic fields and conceptual domains. By qualitative analyses I mean comparing the datasets in terms of presence/absence of the given fields and domains. By quantitative analyses I mean applying statistical calculations. A range of statistics will be used, including Spearman's Rank Correlation Coefficient, Molinari's evenness index, and Welch's T-Test. The statistics used will be described in Chapter 6, on the first occurrence of their usage.

This design will unfold in the chapters of this work as summarized below.

Chapter 6 will address R.Q.s 1 and 2. The chapter will describe the analytical method adopted for highlighting semantic associations, illustrate the results of the semantic analysis of the *chocolate*, and *wine* elicited datasets, and compare the Italian and English cultures along these two themes.

Chapter 7 will address R.Q. 3 and explore alternative routes to retrieve the semantic associations of *chocolate*, and *wine* in the Italian and English cultures without coding the whole dataset.

[12] See Chapter 3, Note 13.

Chapter 8 will verify the results obtained in Chapter 7 by testing the most promising alternative routes on the Web datasets and using an automatic coding system.

Chapter 9 will address R.Q. 4 and compare the results obtained by manual tagging in the Chapter 6 to those obtained using Wmatrix. Since Wmatrix does not treat Italian and no semantic tagger based on a similar coding scheme exists for this language, the chapter will analyse only the English elicited datasets.

Chapter 10 will address R.Q. 5 and analyse the semantic associations of *chocolate* and *wine* in the general Web corpora. To this aim, the manual coding procedure adopted for the elicited data will be applied and the results obtained with the Web corpora will be compared to those of the elicited data.

Finally, Chapter 11 will summarise the analytical and methodological results obtained, and suggesting possible expansions to the current research.