**Culture, corpora and semantics** is a methodological investigation in the use of elicited data and Web data in the analysis of cultural specificities starting from semantic elements. After considering and discussing several theoretical and analytical approaches to culture in linguistics, anthropology, psychology, and marketing research, a specifically developed method of analysis and cross-cultural comparison is applied to elicited data on *chocolate* and *wine*, gathered through free sentence-completion and sentence-writing tests on English and Italian respondents. The results obtained are discussed within the framework of cultural systems theories and used as control reference for further methodological investigations. In particular, the elicited data are qualitatively and quantitatively compared to non-elicited sentences on *chocolate* and *wine* from general Web corpora in English and Italian. Furthermore, in order to find an alternative route which could avoid the complex and time-consuming process of manually coding a large dataset, some alternative routes are tested, based on the creation of sub-corpora using sampling procedures and analysis of a limited number of the most frequent words in the dataset's wordlist. Finally, an automatic semantic tagger is tested on the elicited data, in order to assess the extent of its possible application in cultural analysis. Comparisons between the Web corpora and the elicited data suggest that large general Web corpora can be considered representative of the cultural associations to a node word and could thus be used in cultural analysis or in exploratory marketing research. Finally, in the light of the results of the various methodological tests, the work discusses general issues, such as the relationship between word frequency and cultural relevance, and tagset granularity.

**Francesca Bianchi** is researcher and lecturer of English language and linguistics at the University of Salento (Lecce, Italy). She has published several papers in corpus linguistics and more generally in the use of new technologies in research and foreign language teaching.

ISBN 978-88-8305-093-0
http://siba-ese.unisalento.it/index.php/culturecorpora

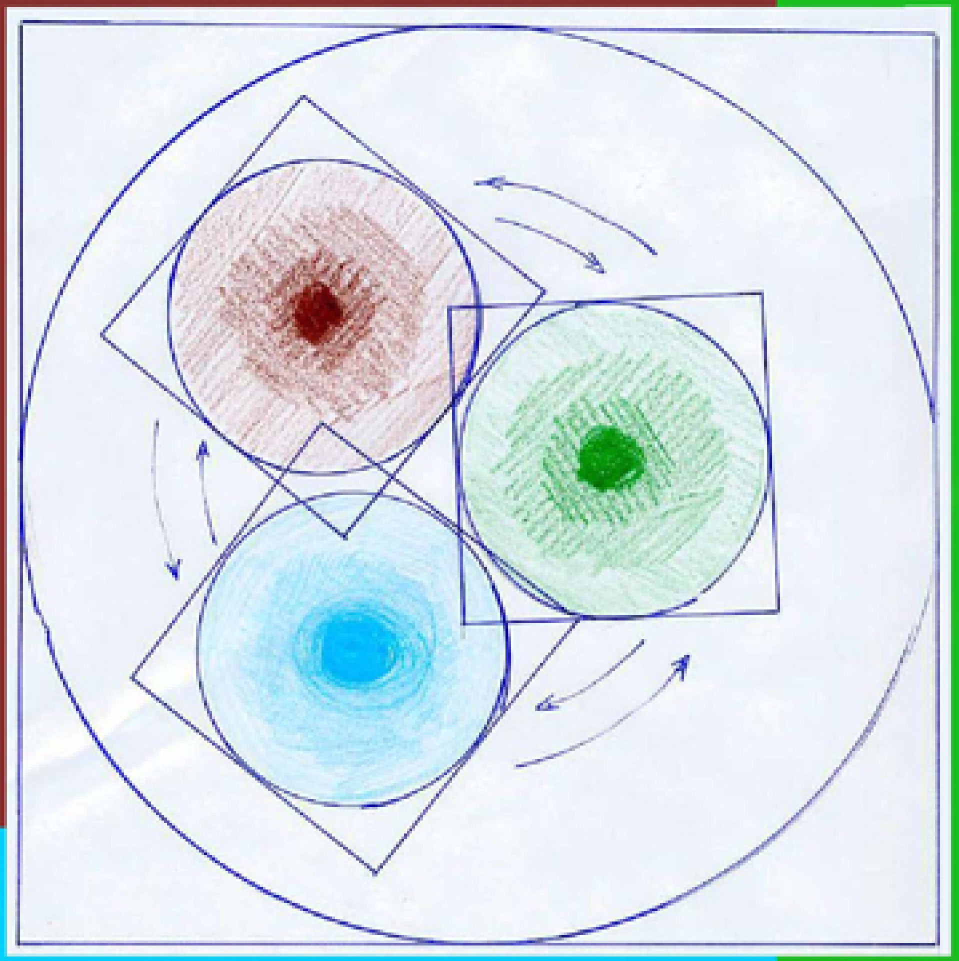
Francesca Bianchi

CULTURE, CORPORA AND SEMANTICS

# CULTURE, CORPORA AND SEMANTICS

## Methodological issues in using elicited and corpus data for cultural comparison

Francesca Bianchi



UNIVERSITÀ DEL SALENTO

# CULTURE, CORPORA AND SEMANTICS

## Methodological issues in using elicited and corpus data for cultural comparison

Francesca Bianchi

UNIVERSITÀ DEL SALENTO

2012

Coordinamento SIBA
UNIVERSITÀ DEL SALENTO

http://siba2.unisalento.it

eISBN 978-88-8305-093-0 (electronic version)

http://siba-ese.unisalento.it

# Contents

## List of Tables

## List of Figures

## List of Graphs

# Preface

This book grows out of the work of my PhD research, presented and discussed at the University of Lancaster, UK, in May-July 2012. In this work the relationship between text, semantics and culture is addressed by assessing various computational procedures of semantic analysis. More specifically, the analysis of two words in British English – *chocolate* and *wine* – and their denotationally comparable terms in Italian (*cioccolato/a*, *cioccolatino/i*) provides the opportunity to test different types of data, sampling procedures, coding methods, and a set of cultural theories in the identification of the cultural associations of those terms.

As the subtitle of the book clarifies, the goal of the present work is methodological, namely the development of a viable corpus linguistics method for distinguishing cultural associations of a given word from personal mental associations. To this end, an interdisciplinary approach was adopted. The theoretical framework for this work draws on several disciplines that study culture through language, though from different perspectives, namely corpus linguistics, cultural studies, marketing, anthropology and psychology, with a focus on their shared elements relevant to the goal of the present research. This was considered necessary in order to make the method applicable outside linguistics. However, the book presents a linguistic piece of research and addresses a perspective audience of linguists.

The work accomplishes two main goals. First, from a cultural perspective, it selects a cultural framework – cultural systems theories – that lends itself to computational semantic analysis, and develops a computational procedure for distinguishing the mental associations anchored in culture from those which are not.

Second, from a methodological perspective, the quantitative comparisons performed between the entire datasets (both elicited and Web-based) on the one hand, and smaller samples of the data on the other, show, in this particular context, to what extent findings based on smaller data samples are generalisable to the whole database the samples come from, thus adding useful pieces of information to our general knowledge in corpus linguistics.

In sum, this book, makes a foray into a multidisciplinary approach to the study of corpora, culture and semantics and provides researchers involved in (cross)cultural analysis with theoretical as well as practical ideas for a user-friendly corpus analysis of cultural associations.

## Acknowledgments

This book owes a debt of gratitude to several people.

First and foremost, I wish to thank Andrew Wilson, my PhD supervisor, for drawing cultural systems theories to my attention and for his precious comments and suggestions during the various stages of my research. Without his constant support and brilliant tutoring, this work would not have been possible.

I would also like to thank Paul Baker and Marina Bondi, the members of the viva panel, for their detailed review of my thesis. Their insightful observations have informed the revision of my work, and helped me achieve a more organic discussion of the various topics addressed here.

I also feel indebted to Carlo Alberto Ratti, physicist and now also good friend, for his precious support in dealing with statistical calculations, and for showing sincere interest in my work.

My thanks also go out to Pierpaolo Mocavero, former student at the Faculty of Languages in Lecce (Italy), for having endured the exhausting task of manually coding the chocolate and wine data, as second coder.

Finally, I thank all my family and friends for bearing with me during my long months, not to say years, of incessant work. A special place in this group is reserved to Sara Gesuato, knowledgeable colleague as much as good friend, for her precious advice on how to face and survive academic pressure, and her constant support.

I would also like to express my gratitude in advance to those readers who will kindly inform me of the inaccuracies of this work.

Chapter 1

# Introduction

## 1.1 An interdisciplinary perspective

Culture is a complex and variegated social and semiotic construct composed of explicit and implicit patterns of behaviour, ideas, and values which are acquired through processes of diachronic and synchronic transmission and socialization, and which are shared by members of a given group, however defined (e.g. professional category membership, shared interests, common interactional practices, national identity). Different cultures develop and share different features, but their members are frequently unaware of this diversity: while some cultural aspects may be visible in everyday life through language or other manufactured products, there are others which are not evident, even to members of the culture itself. Revolving around these core elements, various theories of culture have developed, in keeping with the specific perspectives of different scientific disciplines.

Indeed, culture is a key element in several disciplines, including literature, art, archaeology, philosophy, anthropology, psychology, semiotics, and more recently linguistics, translation studies, and marketing. Most of these disciplines take language, and in particular semantics, as a starting point for their cultural analyses, but adopt different analytical methods, tools and even types of data.

The current work starts from the belief that searching for common ground among the various research traditions that study culture through language cannot but be beneficial to the development of scientific knowledge and is likely to open up new opportunities for linguistics which may find suitable concrete applications in additional academic fields as well as in everyday life.

Among the disciplines which may take advantage of cultural information there is one which plays a leading role in the 21st century: marketing research. In fact, before launching a product on a market, "it is important to understand how [consumers] perceive products, how their needs are shaped and influenced and how they make product choices based on them." (van Kleef, van Trijp, & Luning, 2005, p. 181). Consumer perception and needs are determined, at least to some extent, by their cultural values and beliefs which, in their turn, can be assessed by semantic analysis of language.

With all this in mind, the current work reviews relevant works in linguistics, culture research and marketing research, in order to identify theoretical and methodological common ground between the three areas. Furthermore, it selects those theories, methods and tools which are best suited to a corpus linguist, develops them into an organic framework, and applies them to eight sets of data, in order to test the validity of the new method.

## 1.2 Aims of the project and Research Questions

The general aim of this project is to contribute to our understanding of cultural systems, and of the relationship between text, semantics, and culture. To this aim the work will:

- outline selected models of culture which can provide a useful theoretical framework for the current work and which can be tested on empirical data;
- take stock of existing lines of cultural research within and outside the field of linguistics;
- develop a methodological procedure for highlighting cultural associations of a given key word – i.e. the mental associations it brings to mind in the given country – which starts from corpus data and could be easily and readily applied in cross-cultural studies and marketing projects;
- assess the contribution that semantic analysis of corpora from non-elicited data (in the form of general Web corpora) may provide to cross-cultural comparison and possibly also to marketing research.

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?

## 1.3 Outline of the current work

The current work is logically divided into a theoretical part (Chapters 2 to 4), which creates the theoretical framework that inspired all the subsequent analyses and experiments, and an experimental section (Chapters 5 to 10), analysing the data and describing several methodological experiments. The work is rounded off in Chapter

11 with a summary of the results obtained, a discussion of the materials and methods used, and an overview of the limitations of the current work and possible directions for future research. A brief outline of the contents of each chapter is provided in the following sections.

Chapter 2 provides an introduction to culture and illustrates a few selected theories of culture that lend themselves to quantitative analyses of the semantic features of language, and which will form the reference framework for the current study. Furthermore, the chapter will provide a selected overview of interdisciplinary scientific papers suggesting semantic approaches to the study of culture, in a search for powerful quantitative methods to apply to corpus data.

Chapter 3 offers an introduction to corpora and corpus linguistics, to be used as an organic methodological framework within which to understand the materials and methods used in the current research. The chapter, however, is not intended as a complete list of all possible topics connected to corpora and corpus analysis, but rather a discussion of selected topics that are relevant to the current work.

Chapter 4 provides an overview of the types of materials and methods most frequently used in marketing research, with particular reference to those connected with textual data, and reviews selected marketing and consumer studies where content analysis of data is performed. The studies have been selected because of their similarities with the materials and methods used in my preliminary experiments and/or in the final design of the work. Finally, Section 4.3 briefly describes my preliminary experiments, outlines some theoretical and procedural features common to cultural studies, corpus linguistics and marketing research, and explains how these conflate into the current project.

Chapter 5 describes the materials and methods used in the experimental section. This includes: a description of the questionnaires used for collecting the elicited data; the resulting elicited datasets on chocolate and wine; the WACKY Web corpora and the software used to access them end extract specific datasets; the resulting Web datasets; and the software used for automatic semantic tagging of the British data. Finally, the chapter schematically outlines the research design adopted.

Chapter 6 highlights the semantic associations of *chocolate*, and *wine* in the Italian and English cultures (R.Q. 1) and compares them (R.Q. 2). To this aim, following the widely used habit of analysing elicited data in fields such as the social sciences, marketing, and also linguistics, analyses will be based on four sets of elicited data, specifically collected and manually coded. The analytical procedure adopted, though inspired by existing literature, is specific to the current work. The results of the analysis in Chapter 6 will be used as reference results for all the subsequent experiments.

Chapter 7 addresses R.Q. 3 and explores alternative routes to retrieve the semantic associations of *chocolate*, and *wine* in the Italian and English cultures without coding the whole dataset. In particular, the following three routes – inspired by theoretical considerations as well as attested analytical habits – were explored: 1. manual semantic analysis to the most frequent 50/100/150/200/250/300 content words in the wordlist, 2. using the four most frequent content words to extract sentences from the manually coded dataset and creation of a sampled sub-corpus; 3. random selection of sentences from the manually coded dataset and creation of a random sub-corpus.

Chapter 8 verifies the results obtained in Chapter 7 by testing the most promising alternative routes on different sets of data (i.e. the Web datasets) and using an automatic coding system.

Chapter 9 assesses the possibility of using an automatic semantic tagger to establish cultural associations of the given node words (R.Q. 4); more concretely, the chapter compares the results obtained by manual tagging in the previous chapters to those obtained using Wmatrix, the automatic semantic tagger developed at the University of Lancaster. 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

Finally, Chapter 10 addresses R.Q. 5 and explores the possibility of using general Web corpora to highlight cultural semantic associations of the given node words, by applying the manual coding procedure adopted for the elicited data and comparing the obtained results to the elicited data results.

As already mentioned, Chapter 11 concludes the work by summarising the analytical and methodological results obtained, and suggesting possible expansions to the current research.

Chapter 2

# Culture

## 2.1 Introduction

Culture has long been a debated issue in several different disciplines, such as literature, art, archaeology, philosophy, anthropology, semiotics, and more recently linguistics, translation studies, and marketing. This is in itself proof of and reason for the complexity of this phenomenon which is subject to continuous development and as such lends itself to endless argumentation.

Despite a different perspective on culture taken by each scientific discipline, and the peculiarities of individual theories, some common ideas seem to be shared by the scientific community at large. Most of those elements are summarised in the following definition of culture by Kroeber and Kluckhohn (1952, p. 181):

> "Culture consists of patterns, explicit and implicit, of and for behaviour acquired and transmitted by symbols, constituting the distinctive achievement of human groups, including their embodiment in artefacts; the essential core of culture consists of traditional (i.e. historically developed and selected) ideas and especially their attached values. Culture systems may, on the one hand, be considered as products of action, on the other hand, as conditioning elements of future action."

Indeed, a first important point shared by all cultural perspectives is seeing culture as a social event, based on learning and transmission of information. This has been recognised by several scholars, including, for example, scientist Cavalli-Sforza (1996), and anthropologists Geertz (1979) and Hall (1989). Information can be transmitted diachronically or synchronically. While several anthropologists such as Geertz, Kroeber and Kluckhohn, focus almost exclusively on the diachronic development of culture, other researchers recognize the existence of synchronic forces that produce cultural variation and development; among the latter are Cavalli-Sforza (1996), but also Lotman and Fleischer whose theories will be discussed later on in this chapter. Furthermore, information – and consequently culture – can be acquired consciously and/or unconsciously (Cavalli-Sforza, 1996; Hall, 1982). A frequently quoted theory in this respect is that by Hall (1982), who distinguishes between three levels of culture, which he calls 'technical, 'formal, and 'informal', depending on the level of awareness at which information is transmitted. The technical level of culture is characterised by the objective, denotative, monoreferential view of the world which is at the basis of scientific communication. This level of culture is explicit and formally taught. The formal level of culture includes traditions, customs, rules and procedures, i.e. well-established, conventional aspects of culture which we tend not to notice in everyday life, until they are flouted. Conventional, formal aspects of culture can and are indeed taught and learnt. The informal level of culture, instead, can neither be taught nor learned; it is passed on and acquired unconsciously, or 'out-of-awareness'. This is the level of values, value orientations,[1] beliefs, and judgments. It is at this informal level where people normally react in everyday life and communication. These three levels are metaphorically compared to an iceberg. The technical level of culture corresponds to the tip of the iceberg, which is the only constantly visible part, but also the smallest section of the iceberg. Immediately below that, there is an area, corresponding to the formal level of culture, which may or may not be visible depending on situational or contextual factors; therefore, this area is sometimes above and sometime below the water line, or the limit of consciousness. Finally, there is the biggest part of the iceberg, which is constantly hidden in water and not visible to the eye; this part corresponds to the informal level of culture. This is the level that supports and sustains the whole iceberg; no other part could exist without it. And indeed, values, value orientations and beliefs determine traditions, customs, rules and procedures, which on their turn become explicit in the concrete objects and facts of everyday life (such as music, art, food and drink, dress, architecture, institutions, visible behaviour, and, last but by no means least, language).

Two other important aspects of culture that are largely shared by the scientific community are the semiotic nature of cultural communication – which can boast a long history which reached its peak with the Moscow-Tartu school, as we shall see later on in this chapter – and the idea that culture is the key to interpreting all human action and thought (see for example Hall, 1989; Geertz, 1998; and Lotman, 1980a, 1994).[2]

Finally an aspect of culture that is of paramount relevance to the current work is its connection with language. The existence of a strong and direct link was hypothesised by Sapir (1929; 1949) and Whorf (1956), who believed that the world influences language, language influences culture and thought, and these influence the way we see the world. This sort of circular mechanism has been interpreted by followers of Sapir and Whorf in two slightly different ways that go under the name of 'strong version' and 'weak version' of the Sapir-Whorf hypothesis. According to the strong version, language predetermines human thought; our mental processes and consequently our ideas about the world are constrained by and limited to the possibilities offered by the language we speak. In the weak version, on the other hand, thoughts are influenced by language, but not determined by it. Despite the fact that some of the examples that Sapir and Whorf brought as evidence of their hypothesis

[1] The term 'value orientations' was first used by Kluckhohn (Kluckhohn & Strodtbeck, 1961).

[2] There might be strong differences, however, in the way this 'key' is conceived. Hall (1989), for example, defines culture as the medium we live in, the models or templates we use to interpret the world and to act within it. This contrasts strongly with Geertz's (1998) argument that culture should be interpreted as a series of control mechanisms, including projects, prescriptions, rules, and instructions, aimed at driving human behaviour.

have been disconfirmed by recent research, and that the Sapir-Whorf hypothesis has been challenged, their theory is still quoted by researchers in cultural studies.[3]

However, even though the Sapir-Whorf hypothesis (either in its strong or weak version) is not universally shared, yet, the idea of a direct link between language and culture is widely accepted. Cavalli-Sforza (1996, p. 247), for example, considers language as one aspect of culture, and linguistic evolution as an important element in cultural evolution. Halliday defines culture as a social reality in which meaning is determined by people, with their statuses, roles, and an shared values and knowledge (i.e. the context of situation); language is one of the semiotic systems composing culture and "*actively symbolizes* the social system, representing metaphorically in its patterns of variation the variation that characterizes human cultures" (Halliday, 1978, p. 3).[4] Fairclough (2003) believes that language, intended as discourse,[5] moulds culture and society.[6] In his "Manifesto for critical discourse analysis" (*ibid.*, pp. 202-211), he writes:

> "We can see social life as interconnected networks of social practices of diverse sorts (economic, political, cultural, family etc.). […] Every practice is an articulation of diverse social elements within a relatively stable configuration, always including discourse. Let us say that every practice includes the following elements: activities; subjects, and their social relations; instruments; objects; time and place; forms of consciousness; values; discourse.[7] These elements are dialectically related […]" (Fairclough, 2003, p. 205).

Finally, Hall (1982) sees language as a result of culture and an area where culture can be seen and analysed. The examples above are just a few of the names hypothesizing close relationship of language and culture and show how this belief is spread across research fields as different from each other as genetics, functional linguistics and discourse analysis, and anthropology.

A theory which merges all the above mentioned aspects in a single – though rather complex – organic framework has been developed by Fleischer (1998). His theory lends itself to a quantitative analysis of the semantic features of language and for this reason will form the main reference framework for the current study. His theory was markedly inspired by Lotman's studies as well as by general systems'

[3] See, for example, Wierzbicka (1991). In line with Sapir (1929) and the so-called Sapir-Whorf hypothesis, she assumes that, within each given culture, language both shapes and reflects reality, and the key to the understanding of this link lies in semantics. Therefore, 'cultural key words' – as she calls them – mirror the values and experience of the people in the given culture and are a major tool of culture perpetration; their meanings, however, are culture-specific and consequently impossible to understand for outsiders.

[4] Emphasis in the original.

[5] Considering the study of social discourse as the aim of cultural analysis is not only part of Fairclough's (and Critical Discourse Analysis's) perspective, but also of several other researchers in cultural studies, including for example Geertz (1998), Fliescher (see Section 2.3.1), and Halliday (1978).

[6] To be precise, Fairclough's attention is not on culture as such, but rather on society and ideology. Ideology, which can be defined as "a set of values and ideas advocated by the social dominant groups that guide actions and regulate the relationship of power and are expressed in conventional discourse" (Wu Rongquan, 2001, p. 617), is considered by Geertz (1998) as a cultural system. Indeed, values and ideas are at the core of both ideology and culture. What differs is the perspective that researchers adopt in commenting them.

[7] In the original, these elements are listed within a text box.

theories, both of which will be introduced before Fleischer's theory, in the hope that this might help disentangle Fleischer's complexity.

Finally, the current chapter will provide a selected overview of scientific papers with semantic approaches to the study of culture, in a search for quantitative methods to apply to my data.

## 2.2 Lotman and the semiotic perspective

In the 1960s, in Russia, a group of scholars guided by Yury Lotman – the so-called Moscow-Tartu Semiotic School – started looking at culture from a semiotic perspective, inspired by the structuralist linguistics of de Saussure and Bauden as developed by the Prague school, Jakobson and Trubetzkoj in particular (Lotman, 1994). Their theories, applied to the study of a vast range of cultural and artistic phenomena, have greatly inspired modern semiotics and European scholars in several disciplines. In this section, a few introductory words will be spent on Lotman's view of semiotics and symbols, before we see how Lotman conceptualises culture, language and the world and how these relate to each other.

As Lotman (1997) himself declares, semiotics is the study of symbols. An intrinsic characteristic of symbols is their having unitary meaning and precise delimitations, which makes them easily distinguishable from the surrounding context. In this respect they are similar to texts. Like texts – which are situated in time and provide a synchronic view of a tension between present and future (Lotman, 1993) – symbols have a diachronic dimension and provide the basis for cultural memory. Thus, they transport cultural information from one layer of culture to another one. In the production-reception process, what is chosen by an author because of its symbolic value is interpreted by the receiver through cultural reminiscence. It derives that semiotics could also be seen as a science that studies the nature and transmission of information and, consequently, culture, from a theoretical and historical perspective (Lotman, 1994).

Semiotic space appears as an intersection between several texts at different levels. Text is not reality, but it is the material we need to reconstruct reality. When reality is coded in texts, some elements are favoured and selected for memorisation, while others are discarded and will be treated as non-existent (Lotman & Uspenskij, 1975). Beyond semiotic space there is 'reality', which includes several partially interrelated languages. These two 'layers', taken together, represent what Lotman calls 'the semiotics of culture' (Lotman, 1993).

As we will see in the following paragraphs, Lotman sees culture as a wide semiotic system, having structural rules and acting on a non-cultural background (Lotman & Uspenskij, 1975). Within this system, several semiotic systems coexist, each of them being a different realisation of culture. Therefore, the study of culture should be framed within a general theoretical perspective that studies the mechanics of semiotic systems in general; studying culture ('the semiotics of culture') means studying functional relations between individual semiotic systems. Lotman's ideas will eventually lead him to the conceptualisation of what he calls 'the semiosphere' (Lotman, 1985a, 1985b, 1985c, 1985d), a whole semiotic space that motivates and substantiates individual semiotic acts and their reciprocal interactions. Figure 2_1 is a personal attempt to provide a graphical representation of the core elements of Lotman's theory, in its synchronic dimension.

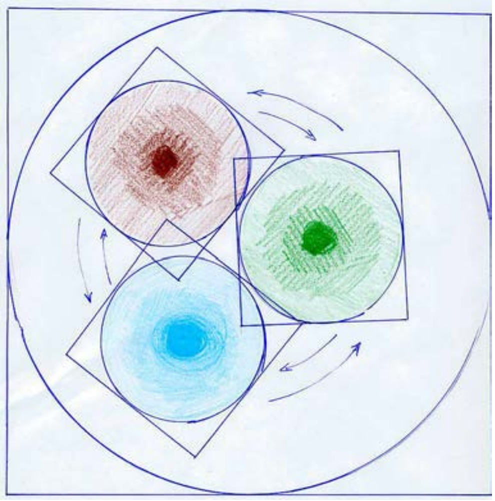


Figure 2_1. Lotman's semiosphere

Squares represent language, while circles stand for culture. The outer square and circle encompassing smaller ones represent natural language and culture, respectively. Smaller circles are the different semiotic systems, each coinciding with a corresponding language.

Culture (the main circle in Figure 2_1) is a complex system, synchronically composed of several semiotic (sub)systems (the smaller circles), such as the arts, literature, science and technology, but also any other type of human activity. Each of these subsystems is considered and analysed as a language of its own (the smaller squares in Figure 2_1). Therefore, culture as a system is expressed through language, or better through several different languages (Lotman, 1993). Each type of language produces (or creates) a different image of the world (Lotman, 1994).

Natural language (the main square in Figure 2_1) is the primary modelling system, while semiotic systems based on natural language, such as myths, folklore, religion, and art, are secondary modelling systems (Lotman, 1993). Though this distinction is necessary in order to highlight the specific features of each of these systems, in real life natural language and culture are closely intertwined (Lotman & Uspenskij, 1975): all languages exist and develop within specific cultures, and all cultures depend on the structure of natural languages (see Figure 2_1). Taking inspiration from the structure of language, the primary scope of culture is to provide a structural organisation of the world.

Another important function of culture is to preserve, transmit and create information (Lotman, 1980d, 1993). Preservation and transmission are possible because culture has a diachronic perspective: culture is the overall memory of humanity, where by memory Lotman means the ability that some systems have to pile up and store information (Lotman, 1980b, 1980c). This entails that culture is primarily a social type of event (Lotman & Uspenskij, 1975; Lotman, 1980d). According to circumstances, the researcher can take into consideration culture in general, the culture of a specific geographical area (e.g. the British culture) or of a specific period (e.g. the Renaissance), or the culture of a particular social group. As such, culture is not a repository of ready-made ideas and texts, but a living mechanism of collective conscience and the intellectual shape of life as it is developing on the Earth. It is a mechanism that creates texts, and texts are the realisation of the potential of culture (Lotman, 1980a). Indeed, humankind is immersed in semiotic space and the human intellect can only act within culture (Lotman, 1980a, 1994, p. 105). Therefore, culture constantly evolves (Lotman, 1980b) and its evolution takes place through alternating phases of gradual growth and explosion (Lotman, 1993).

Creation of new information is possible because of the co-existence of several languages (Lotman, 1980d), as it is linked to communication and transmission of information (see circular arrows in Figure 2_1). There could be no culture without language and communication. Lotman (1993) makes a clear distinction between 'language' and 'code'. The former is natural and includes its historical representation; the latter is artificial, and is created through agreements. If addresser and addressee used the same code, there would be perfect understanding between them and very limited transfer of information. In real life, however, addresser and addressee use separate languages, because each individual represents a separate system (smaller circles in Figure 2_1) and is extrasystemic with respect to any other individual (*ibid*.).

Human memory includes both collective and individual elements (Lotman, 1980c, 1994): while collective memory represents a common core that facilitates mutual understanding, individual memory hampers understanding, but represents the motor of communication (Lotman, 1994). Therefore, the languages used by two individuals have a limited area of overlap. Dialogue between people could be graphically represented as two partially overlapping areas, and communication derives from constant tension between increasing and reducing the area of overlap (see Figure 2_1).

It is thus clear that, for Lotman, culture is made of languages. Each language is a semiotic system of its own. The world is an extrasystemic element which enters language in the form of content (Lotman, 1993).[8] Language creates a world of its own. A major problem, then, is the adequacy or correspondence between the world created through language and the physical existing (extrasystemic) one. More than one language is needed to reflect upon extrasystemic reality, as different languages provide different images of the world. Therefore, the existence of several different languages marks the natural beginning of culture, as well as any other systems, but through time the aspiration towards one universal language leads to the creations of what Lotman (*ibid.*) calls 'second reality'. Second reality is created by culture.

Culture is, therefore, a dynamic system. A major source of dynamism is continuous attraction of extrasystemic elements towards the system, and *viceversa*. At any given point in time, systemic elements are considered existing and correct, while extrasystemic elements are tantamount to being incorrect and nonexistent. However, what seems extrasystemic with respect to one system may be systemic in another. Describing what is systemic will also indicate what is extrasystemic (Lotman, 1980b). Furthermore, every system is characterised by core and periphery (*ibid*.). The system

[8] This goes against any traditional literary/linguistic oppositions between form and content. For Lotman (1993), opposition is between language, which includes content and form, and the world. A similar standing is taken also by Geertz (1996, p. 25), who declares that, in cultural analysis, it would be impossible to draw a dividing line between the methods used to represent content and the content itself.

is stronger towards the core and becomes weaker as we depart from the core (see the use of colour in Figure 2_1). Movements from periphery to core and *viceversa* cannot be avoided and are characteristic of diachronic development. Complex systems, like culture, are also characterised by tension between understanding and non-understanding, and in any given moment we are more towards one or the other position (*ibid*.). Finally, the several systems existing within culture develop at different paces. However, quicker ones may speed up the development of slower systems. The evolution or development of these systems takes place through alternating explosive and gradual phases. Explosive phases lead to innovation, while phases of gradual growth guarantee continuity to the system (*ibid*.; Lotman, 1993).

It is interesting to notice here – thought this is not surprising in structuralist thinking – how Lotman's thought is constantly characterised by the juxtaposition and co-presence of opposites. Culture emerges only when compared to non-culture, and systemic elements become evident because extra-systemic elements exist. Evolution takes place because of constant tension between what is systemic and what is extrasystemic, between understanding and non-understanding, between maintenance of a given level and amount of information and the creation of new information, between symmetry and asymmetry (Lotman, 1985a). Communication derives from tension between increasing and reducing the area of overlap between different languages (Lotman, 1994). Development takes place through alternating phases of gradual growth and explosion. Gradual growth is aimed at maintaining what is given (homeostasis), while explosion disrupts what is given and establishes a new reality (development) (Lotman, 1985c, 1993).

To sum up, Lotman's semiotic theory postulates the existence of an inextricable link between culture, humankind and the world, where culture is seen as a living mechanism of collective conscience, made of several semiotic systems and expressed through language. Language, on its turn, merges culture and reality, thus creating a world of its own in which reality is filtered through cultural perception. Communication derives from constant tension between increasing and reducing the area of overlap of addresser's and addressee's languages.

Lotman's ideas were exploited and reviewed by several researchers and in particular they clearly inspired some systemic approaches to the study of culture. One of these is Fleischer's systems theory, which is summarised in the following section.

### 2.3 Systems theory

Systems theory, also called systems analysis (Lowe & Barth, 1980, p. 568), is a transdisciplinary field that studies systems and their properties. In systems thinking all types of phenomena are seen in terms of 'systems', where by *system* they mean a

> "collection of interrelated elements ('structured set') where a change in one aspect would affect some or all aspects of the system. […] System analysis is the holistic approach to the ultimate understanding, design, and optimization of systems" (Gordesch, 1998, p. 39).

Systems include several parts or subsystems; they are not the sum of their constituent parts, but what arises from the *interaction* of these. Indeed, one of the most important features of any systems theory is emphasis on relationships between objects, patterns of distribution, and the overall context in which these objects are produced (the *environment*). Hence, a system is more than the sum of its parts, and system analysis entails identifying the constituent parts of the system, and "analyzing their interrelations and *functions* as part of the whole"[9] (Seppänen, 1998, p. 183).

Two major types of systems can be hypothesised: open, and closed systems.[10] *Open systems* are characterised by continuous interaction with the *environment*, i.e. the context within which the system exists (extrasystemic elements) and tend towards continuous growth and evolution. Consequently, open systems show constant tension between *homeostasis* (maintenance of the *status quo*) and *development*. On the other hand, the term *closed system* refers to a state of isolation from the environment. No system can ever be completely closed, and closure affects all systems in varying degrees.

Finally, a fundamental part of systems thinking is the use of formal mathematical techniques to model systems behaviour (Lowe & Barth, 1980, p. 570).

It is easy to see how Lotman's theories on culture could be integrated into systems thinking. This was done by Fleischer (1998) who reframed and slightly expanded Lotman's semiosphere. Fleischer's however is not the only existing systemic theory of culture. Here we shall consider two systemic approaches to culture:[11] Fleischer's systems theory, which is a rather comprehensive systemic interpretation of culture in which language, discourse and semantics are foregrounded; and Nobis's theory of behavioural patterns which – though indirectly – provides an interesting view on the level of development of a culture with respect to selected symbols.

### *2.3.1 Michael Fleisher: radical constructivism and semiotics*

Fleischer's theory, variously presented and discussed in his writings in German, is summarised in his essay "Concept of the 'Second Reality' from the perspective of an empirical systems theory on the basis of radical constructivism" (Fleischer, 1998). As we shall see, Fleischer, taking inspiration from Lotman's semiotic ideas,[12] reinterprets the cultural paradigm with a radical constructivist perspective.

The following paragraphs provide a synthesis of his theoretical ideas. Systems theories such as his one can only be clearly understood considering all of their constituent parts; for this reason, all main elements of Fleischer's theory will be taken into consideration, regardless of their level of relevance to the current research.

Fleischer's general framework of reference is *radical constructivism*,[13] a theoretical approach to the concepts of world, understanding, and knowledge

[9] Emphasis in the original.

[10] Other classifications are also possible. See Miller (1978, in Seppänen, 1998, p. 199).

[11] To a linguist, the adjective 'systemic' would most probably bring Halliday's systemic functional theory to mind (described, for example, in Halliday 1978, 1985; Halliday & Hasan, 1989). Indeed, Seppänen (1998), in his historical review of systems thinking, mentions Halliday among systems thinkers in linguistics. Though Halliday mentions culture as the outer circle within which the context of situation and language operate, his interest seems to be primarily on the interaction between the latter two elements, and Halliday's systemic functional linguistics developed into systemic functional grammar, rather than into a theory of culture.

[12] This is evident in the several similarities between their two systems, as well as in the direct attention Fleischer dedicated to Lotman in his own publications (see for example Fleischer, 1989, 2001).

[13] Constructivism grew out of the interpretation of some basic notions by French psychologist Jean Piaget regarding cognitive development. According to Piaget, learning is based on the assimilation of

according to which what we see, perceive or understand is a construction of our own brain, an interpretation based on the possibilities, limits, and logics (schemata) of our mind, which in turn depend on experience (von Glasersfeld, 1984).[14]

Within this radical constructivist framework, Fleisher distinguishes between the First Reality, Culture, and the Second Reality. The First Reality is "the physical, objectively given reality and its laws" (Fleischer, 1998, p. 423), such as the world, society, etc.. *Culture* is "a sign-generating subsystem of the social system" (*ibid*., p. 433) and as such is part of the First Reality. It is described as an open, self-organising *semiotic* system containing "all those phenomena and aspects based upon semiotic processes" (*ibid.*, p. 433). Its configuration is relational and functional. Like all open systems, it is dynamic, increases in order with the passing of time, is partially open and partially closed in order to maintain the steady state, and depends on the environment, i.e. the social system. The system as a whole organizes itself on the basis of chance and necessity, yet its constituent parts are autonomous in their evolution.

The *Second Reality* is defined by Fleischer (*ibid*., p. 430) as "a given and functioning shaping of a system, i.e. a concrete realisation of general laws of the system". In other words, it is a specific cultural realisation organised as a functional, semiotic, relational system. Noticeably, however, the Second Reality is not a concrete entity but rather an image of the First Reality it relates to. Fleischer distinguishes between two types of images: world-images, which represent the view of those belonging to the culture in question (an 'emic' view), and appearance-images, which are the way world-images are received and understood by an external observer (an 'etic' view). The Second Reality is "based on utterances, fixed and manifest opinions and world-images." (*ibid*., p. 429). So, for example, if we considered the First Reality of Great Britain, we could say that it includes culture (the British culture) as a system of semiotic elements. The corresponding Second Reality would then be the way culture is realised through utterances, opinions and world images; in other words, the way the world in general, and the British culture in particular, are seen and described by British inhabitants.

The Second Reality develops at two levels: a general (linguistic) semantic level, and a specific (cultural) semantic level. The "linguistic [semantic formings] provide

new experience and new pieces of information into self-made schemata based on previous experience. If new pieces of information do not fit our schemata, accommodation takes place, i.e. the schemata are changed so as to adequately include the new pieces of information (Piaget, 1937). Piaget's widely quoted statement by constructivists is that intelligence organizes the world by organising itself (Piaget, 1937, p. 311). Consequently, "constructivism proposes that learner conceptions of knowledge are derived from a meaning-making search in which learners engage in a process of constructing individual interpretations of their experiences. The constructions that result from the examination, questioning and analysis of tasks and experiences yields knowledge whose correspondence to external reality may have little verisimilitude. However, to the degree that most of our learning is filtered through a process of social negotiation or distributed cognition [...], generally shared meanings, tend to be constructed." (Applefield, Huber, & Moallem, 2001, p. 2). These concepts inspired the American psychologist Ernst von Glasersfeld (1984), who applied and developed them further, creating what he called 'radical constructivism', focussing on the idea that experience as well as all objects of experience are the result of our ways and means of experiencing.

[14] This view is clearly in opposition to the strong version of the Sapir-Whorf hypothesis (see Section 2.1), according to which our view of the world depends on the possibilities and limits of language, which shapes the logics (schemata) of our mind. It seems compatible, however, with the weak version (see Section 2.1), in so far as language is part of our experience and as such may influence, though not determine, our view of the world.

the basis for the cultural ones, and are under their influence" (*ibid.*, p. 431). In other words, the 'laws' of language are the bricks and mortar for the creation of culture-specific semantic realisations aimed at differentiating one social (sub-)system from another and creating interconnections among its members.

Culture manifests itself at the level of the discourse, which is considered by Fleischer a systemic semiotic repertoire that mediates between a culture and its symbols, the place where symbols are applied and become evident. Within culture, several different subsystems exist at different levels; each of them works as feedback system for the other ones. In Western cultures, these subsystems are typically cultural groups, subcultures, single cultures, and intercultures. Each of them differentiates itself from the other elements at the same level of stratification and from the higher element and manifests itself in the Second Reality through a different type of discourse. The following paragraphs discuss the different levels of stratification of culture and of discourse. Figures 2_2 and 2_3 provide graphical representations of culture, from a synchronic and diachronic perspective, respectively.

*Cultural groups* (the smallest circles in Figures 2_2 and 2_3) are the smallest systems, the elements at the bottom of the stratification hierarchy. A cultural group could be, for example, a circle of friends, a study group, a gang,[15] but also a family, or the employees of a business company. Several cultural groups make up or belong to a *subculture*. Subcultural groups could be, for example, scientists vs. politicians vs. economists vs. lawyers, or different age-groups, or males vs. females. The subculture is the level where *discourse* takes place. All 'subcultures of a geopolitical area as well as neighbouring subcultures which are considered to be subculturally or discoursively close' (*ibid.*, p. 437) make up a *single culture* (for example, the British or the Italian culture). At this cultural level, *interdiscourse* takes place. Finally, the *interculture* is composed of neighbouring or similar single cultures. Examples of interculture could be Mediterranean countries, German-speaking countries, or the European Union. At this level, *intercultural discourse* takes place. Discourse, interdiscourse and intercultural discourse are the linguistic counterpart of each cultural level.

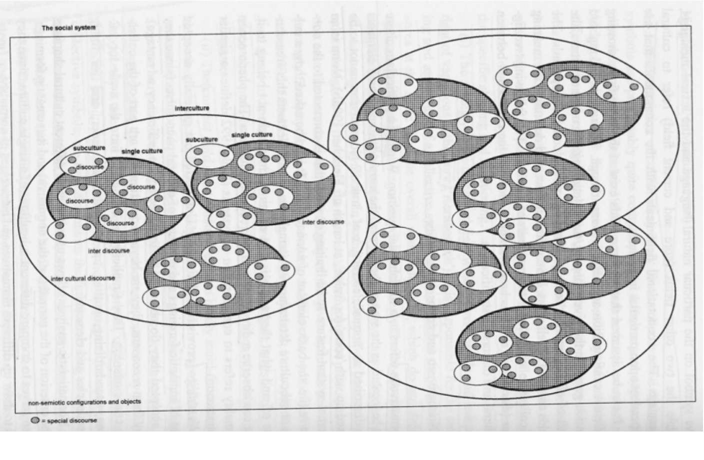


Figure 2_2. Synchronic section of cultural and discourse stratification (from Fleischer, 1998, p. 444)

[15] These are the only three examples provided by Fleisher in his 1998 paper.

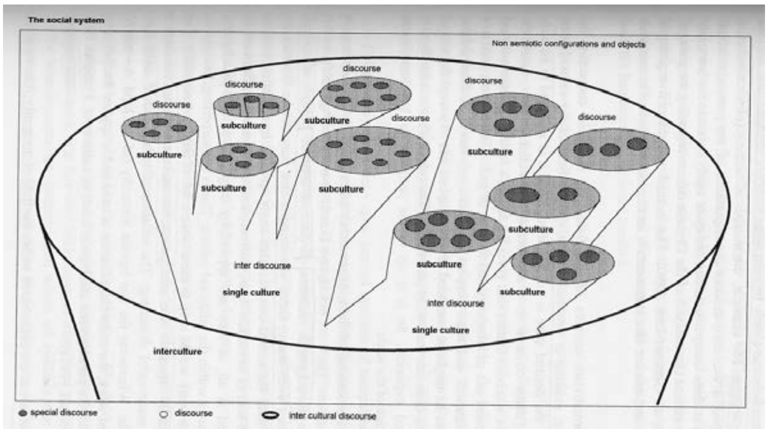


Figure 2_3. Diachronic section of cultural and discourse stratification
(from Fleischer, 1998, p. 445)

In Fleischer's radical constructivist view, discourse is functional to creating, maintaining and developing cultural identity, distinguishing each (sub)culture from the other ones at the same level of classification, and creating the basis for higher level discourse. Discourse, being a semiotic element, is made of symbols. Two types of symbols are particularly relevant: *collective symbols* and *discoursive symbols*. Collective symbols are the most important elements that make up interdiscourse. A symbol can be considered a collective symbol when it has reached a high level of conventionalisation, and agreement exists about its meaning among the members of that single culture. As Wilson & Mudraya (2006, p. 3) nicely explain, "the degree to which a symbol is anchored […] within interdiscourse will determine the conventionalisation of its semantic profile and the extent to which it is open to manipulation". Collective symbols show a very distinctive meaning and a very distinctive rating (positive or negative) that is valid for a whole single culture. Part of the meaning of a collective symbol is culturally dependent, and differs among cultures, because it represents the state of the system. In order to identify collective symbols one should take into consideration frequency and spreading, but also existence of functions and effect of the symbols under investigation (Fleischer, 1998, p. 449). Discoursive symbols are the counterpart of collective symbols at subcultural level. They are connected to a particular subculture, do not occur in other subcultures and, if they do, they show a different semantic content.

Both collective and discoursive symbols are made up of three elements, which Fleischer calls *core*, *current field*, and *connotational field*. The core is a stable element. Collective symbols with long standing have a strong, dominant core. The current field is a rather generalised, but not yet stabilised element. Both core and current field are expressions of cultural meanings. Finally, the connotational field is an expression of individual meaning and as such has no stabilisation at all; it is connected to the particular natural language and to lexical meaning.[16]

[16] Fleischer's terminology might be confusing here. Indeed, he uses the term 'connotational' to refer to *individual, non standardised* meaning components, connected to the particular natural language and lexical meaning, but *not connected to cultural meaning*. This word, however, reminds of 'connotation' and 'connotative meaning', terms used in linguistics, but also in semiotics, to refer to additional meanings evoked by word associations of other words or concepts (Wales, 2001). These additional,

These elements allow interpreting the level of rooting of a word in the given culture: if the connotational field (individual meanings) predominates in the understanding of a word, then that word is not a collective or discoursive symbol. When current field predominates, the given word is about to become a collective or discoursive symbol. Finally, when core meanings predominate, we are dealing with a very strong collective type of symbol. Concrete examples of how to identify core, current and connotational fields are provided in Section 2.4.1.3.

Collective symbols have several functions and are used to a variety of aims, including the following:[17] to ensure or change the discursive status of an utterance; to ensure a system's coherence; to manipulate or polarize opinions or viewpoints; to cancel or prevent arguments; to achieve positive feedback and cultural success; to pose responsibility of what is said on the interdiscourse or on the recipient; to avoid manipulation. A text that is rich in collective symbols is not greatly open to interpretation and, if symbols are used homogeneously, will receive generalised positive feedback.

Another basic element in Fleischer's theory are the interconnected concepts of normativism and normality. Normativism refers to the fact that any particular culture considers some elements or phenomena 'normal', that is to say acceptable, correct and/or real. Normativism includes two states: the desired state (what is normal), which is frequently non specified; and the refused state (what is not normal).

The types of phenomena that fall in the category of normality are

> "unreflected, but generally accepted areas of semantic formings and elements of cultural phenomena (since they are accepted they do not need to be reflected, i.e. efficacy advantage). They form a consens, they are beyond question and are part of the collective consciousness. They (implicitly or explicitly) form the scale for valuation, standardization, putting up a hierarchical order, and for fixing dependencies and elements" (*ibid.*, p. 443).

Normality is therefore the reference point when we want to compare different subcultures or single cultures. Normativistic analysis, which is fundamentally a quantitative type of analysis based on frequency and spreading of specific individual phenomena, helps us understand what is considered normal in a given (sub)culture and allows us to make comparisons between (sub)cultures. Fleischer calls 'normatives' the collective symbols in the group of normality. These have greater force of affiliation and of delimitation than any other type of collective symbol.

It seems appropriate to say that Fleischer expands and systematizes Lotman's core ideas, with particular reference to the 'Second Reality', i.e. a concrete realisation of general laws of culture, as it appears through language; a semiotic, relational system "based on utterances, fixed and manifest opinions and world-images" (*ibid.*, p. 430). The culture layer of the Second Reality is composed of several interacting and at times (partially or totally) overlapping social systems and sub-systems – interculture, single cultures, subcultures, and cultural groups – each expressing itself in a different

frequently unexpressed evaluative meanings, are shared by the speakers of a language. Indeed, connotative meaning, compared by some corpus linguists (Hunston, 2002) to semantic prosody, could be frequently considered as culture-bound (see for example Taylor, 1998).

[17] As Fleischer points out (1998, p. 447) "collective symbols are not the only linguistic or semiotic objects having these functions".

discourse system. At the language level, some of the most relevant elements to identify the subtending cultural layer are collective symbols – symbols with a wide core element, shared by all or most of the members of a cultural system and related in particular to the interculture –, and normatives – collective symbols in the range of normality according to the scales of normalisation of that cultural system. Finally, Fleischer emphasizes quantitative aspects in the identification of collective symbols and normatives, and in the analysis of all phenomena of the Second Reality, regardless of whether we look at them from the point of view of the sender (world images) or the receiver (appearance images).

Cultural analysis can be performed at any of the levels of culture mentioned by Fleischer (interculture, single cultures, subcultures or cultural groups), depending on specific research interests and goals. In the current work I will focus on the English and Italian single cultures, i.e. the cultural elements that are shared by all English natives on the one hand, and Italian natives on the other.

### *2.3.2 Adam Nobis: behavioural patterns and the self-organisation of culture*

Another interesting contribution is offered by Adam Nobis, who analyses self-organising open systems and introduces the concept of ‘enhancement of complexity organisation of evolving objects’ due to interaction.

His theoretical framework for defining culture is inspired by anthropologists Kroeber (1952) and Morin (1973) and is summarised as follows: “culture is a complex of behavioural patterns transmitted non-genetically” (Nobis, 1998, p. 464), where the phrase ‘behavioural patterns’ refers to a stable structure of interaction that emerges among partially spontaneous behaviours. The key concept here is stability. As we have seen, open systems are in constant tension between stability and evolution. Any change in the system determines further changes. Therefore, transmission of behaviour – which is at the basis of evolution – may only take place when that behaviour has a long established network of relations with other behaviours, i.e. a stable behavioural pattern. Behaviours may be of several different types, including mental behaviours, such as for example thinking about a concept. Interestingly, the notion of ‘stable behavioural pattern’ can easily be equated with Fleischer’s notion of conventionalisation, i.e. a mental behaviour, such as thinking of a concept, has features that, at a given point in time, are widely shared by all members in a community/culture.

The evolution of mental behaviours was tested by Nobis in a three-stage diachronic experiment on Polish students aimed at assessing the changing ways of thinking about Europe in Poland (Nobis, 1992a, 1992b, 1998). In a first phase, a group of Polish students were asked what they thought about Europe, and their interviews were recorded. Each of their statements, grouped according to theme (or belief), was analysed in terms of the reason given to support it; and the reasons provided were then collected into what we might call sub-themes. In particular, one of the common themes in the student’s answers referred to European superiority on other continents. The arguments used to support this claim were either cultural, or historical, or economic, or a mixture of more than one of these. A few months later, a different group of Polish students was asked to agree or disagree with a series of statements/opinions comparing Europe to other continents. These statements had been created using the arguments of the students in the previous experiment; some of the statements included homogeneous concepts, while others were combinations of arguments from two, three or four different sub-themes. The most accepted statements were two-element opinions, immediately followed by three-element opinions, one-element opinions, and four-element opinions, in this order. These results led Nobis (1998, p. 470) to formulate the hypothesis that:

> “during social interaction more-element configurations will supersede the few-element configurations, which also means that more-element configurations will be accepted by the increasing number of individuals taking part in the interaction”.

This hypothesis was tested and confirmed two years later on the second group of students, who were interviewed in a repetition of the second experiment. This time, in fact, three- and four-element opinions were as popular as two-element opinions. In other words, the richer the behavioural pattern, the more established (widely accepted, or conventionalised) the pattern is.

While Nobis’ hypothesis was developed to describe the dynamic evolution of culture, it could probably be applied synchronically for cross-cultural comparison. In fact, it could be hypothesised that relevant differences in the number and complexity of the concepts connected to the same key concept in different cultures are indications of different stages of knowledge/acceptance of the key concept.

### 2.4 Analysing culture through language

All the disciplines that are, directly or indirectly, involved in the study of culture, have taken language and linguistic production as their starting point for analysis. Several linguistic aspects can be (and have been) taken into consideration for the analysis and comparison of cultures, including content,[18] genre,[19] syntax,[20] and semantics, the latter being the focus of interest in the current research.

Most work, especially that carried out within the framework of critical linguistics, discourse analysis, translation studies, and semiotics, has traditionally been based on in-depth analysis of a limited number of texts. Recently, however, voices have raised advocating the use of corpus analysis techniques – frequently supported by statistical calculation –[21] and extensive analysis of large quantities of data performed by means of computerized tools have gradually been taken into consideration for these types of study, in these disciplines. Other disciplines, including anthropology, psychology, and consumer research have longer experience with statistical methods, frequently applied to collections of elicited data.

The aim of this section is to accomplish an overview of semantic approaches to the study of culture, in a search for powerful quantitative methods to apply to corpus

[18] By ‘content’ here we analysis of several different textual, intertextual, contextual, and/or semiotic elements, such as relation between text and pictures, anaphora, elision, use of metaphors. See for example Bassnett (1991, pp. 28-29), and Katan (2006).

[19] In different cultures, the same text type may have different ‘rhetorical requirements’ as regards, for example, moves and steps, register (e.g. more or less formal), explicitness. See for example Aston (1988), Wierzbicka (1991, Chapters 4 and 5), and Tosi (2001).

[20] See for example Gerbig (1993), Stubbs (1994), and Galasinski and Marley (1998).

[21] See, for example, Stubbs (1997) and Coffin & O’Halloran (2004).

data.[22] Particular attention will be dedicated to those studies that employ corpora and quantitative research methods, however a brief outline of Wierzbicka's qualitative research work in cultural studies will also be provided, given its relevance in the linguistics panorama.

The studies here presented have been selected according to one or more of the following criteria: relevance of the study within its own discipline; frequent quotations of the given study; novelty of its approach; relevance to the current work. In this review, methodological issues will be foregrounded, along with results, in order to show the types of analysis that have been conducted on corpus data and the connections that can be seen between this type of semantic data and culture. This overview is by no means to be considered exhaustive of all the studies carried out in the different disciplines, or of all possible methodological approaches, especially as far as non-linguistic disciplines are concerned.

For the sake of systematisation, the studies will be grouped according to discipline and research tradition. This type of organisation – though not necessarily the best possible one – has been chosen because it clearly shows how different disciplines are still characterised, to a large extent, by the use of different preferred research methods. Only a few seem to be the studies that experimented with analytical methods or materials from outside the specific research tradition, but their results show the possible benefit of greater interdisciplinary integration.

### *2.4.1 Culture studies in linguistics*

#### *2.4.1.1 (Cross)-cultural semantics: Williams and Wierzbicka*

One of the first linguists to look for words that might be meaningful for the understanding of reality was Raymond Williams. In his 1959 book *Culture and Society*, he noticed that some words change meaning or acquire particular importance in given periods. These words, he believes, mirror changes in the way people think about reality. In particular, he highlighted the following five words: *industry*, *democracy*, *class*, *art*, and *culture*. Subsequently, banking on the idea that comparison between the meaning of selected, relevant words – which he calls 'keywords' – and people's experience of everyday life may shed light on language and the way people use it, he focused on *culture* and published his most famous work *Keywords. A Vocabulary of Culture and Society* (Williams 1976), where he illustrates the words that he believes to be connected to the idea of culture (and society) and that help explain the complexity of its meaning.

Thus, Williams thought that culture could be analyzed through several other words that are in some way related to it. Although he did not provide specific indications about how to select or analyze these words, his work has inspired several linguists. In particular, Wierzbicka seems to have elaborated Williams' definition of 'keywords' when she describes 'cultural key words' as "words which are particularly important and revealing in a given culture" (Wierzbicka 1997: 15-16). While

[22] Following McEnery and Wilson (2001, pp. 76-77), I define as quantitative all those approaches where features are classified and counted and statistical models are constructed in order to explain what is being observed; quantitative approaches differ from qualitative approaches, where "no attempt is made to assign frequencies to the linguistic features which are identified in the data" (*ibid.*, p. 76).

William's attention was not contrastive, Wierzbicka's is cross-linguistic and cross-cultural.

Wierzbicka, like Williams, denies the existence of an objective discovery procedure for the identification of words of this type, but has developed a method for the analysis of their meaning, based on semantic primitives (or universal semantic concepts). Her interest lies in cross-cultural communication, and her theoretical arguments are based on Geertz's and Sapir's assumptions about culture and language. Indeed, she takes up Geertz's (1979; 1998) idea of culture as a system in which patterns of meanings, embodied in symbols, are transmitted between people and across time. These patterns allow people to communicate, perpetuate, and develop knowledge. Furthermore, in line with Sapir (1929) and the so-called Sapir-Whorf hypothesis, she assumes that, within each given culture, language both shapes and reflects reality, and the key to the understanding of this link lies in semantics. Therefore, 'cultural key words' – as she calls them – mirror the values and experience of the people in the given culture and are a major tool of culture perpetration; their meanings, however, are culture-specific and consequently impossible to understand for outsiders. Thus, in order to achieve cross-cultural communication, the meaning of these words should first be made explicit and illustrated by means of "a 'natural semantic metalanguage', based on a hypothetical system of universal semantic primitives" (Wierzbicka 1991: 7). Hence, Wierzbicka (1972; 1980; 1987; 1988; 1989a; 1989b) and her colleagues (see in particular Goddard & Wierzbicka 1994) have developed a list of universal semantic primitives to use in the translation of cultural keywords. In 1991, her list included more than two dozen hypothetical semantic primitives divided by category (Wierzbicka 1991: 8): Pronouns (*I; you; someone; something*); Determiners (*this; the same; two; all*); Classifiers (*kind of; type of*); Adjectives (*good; bad*); Verbs (*want; don't want; say; think; know; do; happen*); Modals (*can; if/imagine*); Place/Time (*place; time; after/before; above/under*); and Linkers (*like; because*). In 1997, the list already included 'nearly sixty candidates' (Wierzbicka 1997: 26), organized in

> "a network of categories, which can be compared (some-what metaphorically) with the parts of speech of traditional grammar. The main point is that the categories […] are, so to speak, both semantic and structural" (ibid.).

In addition to these primitives, she sometimes employs

> "a limited number of other concepts, which are regarded as neither indefinable nor universal or near-universal, but which are still relatively very simple and which recur widely in the languages of the world as separate lexical items. This larger set, whose items can be defined in terms o the basic set of primitives, includes concepts such as 'feel', 'small', 'much', 'a little', 'more', 'less', 'different', and so on" (Wierzbicka 1991: 8).

According to Wierzbicka's method of analysis, for example, the English word 'freedom' and its Latin, and Russian counterparts (Wierzbicka 1997: 125-154)[23] can be described as follows:

[23] Her chapter also includes the Polish word for 'freedom'.

*[English] freedom*

(a) someone (X) can think something like this:

(b) if I want to do something I can do it

(c) no one else can say to me: "you can't do it because I don't want this"

(d) if I don't want to do something I don't have to do it

(e) no one else can say to me: "you have to do it because I want this"

(f) this is good for X

(g) it is bad if someone cannot think this.

*[Latin] libertas* (e.g., *X habet libertatem* 'X has freedom')

(a) someone (X) can think something like this:

(b) when I do something I do it because I want to do it

(c) not because someone else says to me:
"you have to do it because I want you to do it"

(d) this is good for X

*[Russian] svoboda*

(a) someone (X) can think something like this:

(b) if I want to do something, I can do it

(c) when I do something, I don't have to think:
"I can't do it as I want to do it because some (other) people do/say something"

(d) X feels something good because of this

On the basis of these semantic reformulations, along with analogous descriptions of related concepts such as English 'liberty', and Russian '*volja*' (will), and very brief comments about the collocations of these words, Wierzbicka concludes declaring that freedom is a culture specific concept, as it differs in different cultures.

Despite the fact that she insists on her method being "a verifiable, non-speculative way" (ibid.: 30) to study cultural patterns, a few concerns might arise. First of all, Wierzbicka's 'cultural keywords' are chosen subjectively;[24] furthermore, no explanation is provided for the choice of the corresponding words in the various languages: the reader is only left to assume that they are 'official' translations provided by dictionaries. Finally, her considerations depart from a relatively short list of examples of use, which include phrases or sentences by famous writers, set phrases or proverbs. These, along with comments and definitions by previous researchers or writers, are used to explicate the meaning of the concept in terms of semantic (and syntactic) primitives, which still leaves us with doubts about the completeness and objectivity of her non-speculative analysis. Indeed, it really seems that – as Bigi nicely comments – "[…] keywords in Wierzbicka appear to be as a domain in which to apply and verify her theory of semantic universals" (Bigi 2006: 165-166).

What is interesting in Wierzbicka's approach, on the other hand, is her attention to individual words and direct comparison between cross-language translations of the same concept. Also interesting is her attempt to look at collocates, though her analysis of collocates is based on an exiguous number of examples, comments are very brief,

[24] See Chapter 3 for greater details on this issue.

and collocations are used mostly to highlight contrast between near synonyms. In the following sections we shall see how other, more quantitative-oriented, research traditions have used these same elements for (cross)-cultural analysis, but with a different focus. The studies reviewed in the following subsections are all characterised by the use of corpora, but they have been grouped according to what seems to be their major theoretical framework.

### *2.4.1.2 Within the framework of corpus linguistics*

Cultural studies accomplished primarily within the framework of corpus linguistics seem to have all focused on general comparisons between cultures. One of the first studies using corpora to assess cultural specificities is that by Leech and Fallon (1992). The aim of their research was to highlight cultural differences between the American and British cultures by comparing American and British English in the 1960s. On the basis of word frequency tables from the Brown and LOB corpora,[25] created applying the chi-square test and published by Hofland and Johansson (1982), Leech and Fallon grouped words[26] into semantic categories and identified 15 categories where noticeable frequency differences could be seen, namely: sport; transport and travel; administration and politics; social hierarchy; military; law and crime; business; mass media; science and technology; education; arts; religion; personal reference; abstract concepts; and ifs, buts and modality. The authors used frequency differences to draw generalised conclusions about the two cultures: the American culture emerged as masculine, militaristic, dynamic, driven by high ideals, technology, activity and enterprise; the British culture as

> “given to temporizing and talking, to benefiting from wealth rather than creating it, and to family and emotional life, less actuated by matters of substance than by considerations of outward status” (Leech & Fallon, 1992, pp. 44-45)

Leech and Fallon’s cultural analysis and conclusions are entirely based on corpus data and word frequency, and are not contrasted to or commented within any type of qualitative study of the cultures considered.[27]

Leech and Fallon’s results were replicated by Oakes (2003) in a study concerned with the American and British languages of the 1990s. Using the FROWN and FLOB

[25] The Brown corpus, created in 1961, at Brown University (Rhode Island) includes approximately 1 million words of American English from a wide variety of prose texts (Francis & Kucera, 1979). The LOB (Lancaster-Oslo-Bergen) corpus was developed to match the Brown one, and includes about 1 million words from British English texts published in 1961. These two corpora were followed, in the 1990s, by the FROWN and FLOB, created at Freiburg University to match the Brown and LOB and to be used for diachronic studies. The FROWN and FLOB include about 1 million words of American and British English respectively, with sample texts that were published in 1991 (McEnery & Wilson, 2001).

[26] These authors disregarded what they called ‘linguistic contrasts’, i.e. spelling differences (e.g. *color* and *colour*), and lexical choice (e.g. *gasoline* and *petrol*), and did not consider proper nouns.

[27] This strongly clashes with Wierzbicka’s views when she claims that, though important and revealing, “frequency is not everything […]. Frequency dictionaries are only broadly indicative of cultural salience, and they can only be used as one among many sources of information about a society’s cultural preoccupations” (Wierzbicka, 1997, p. 15). Wierzbicka’s words were not openly directed to Leech and Fallon’s study; yet, given the content and date of her considerations, it seems probable that she had this paper in mind. In fact, though in the Introduction to her work she mentions frequency as an indicator of salience, she completely ignores it in her analyses.

corpora.[28] Oakes' results entirely confirmed the findings of the previous study, which can only be taken as evidence that frequency is a good indicator of cultural salience. Indeed, Leech and Fallon's (1992) paper paved the way to a series of other studies which, likewise, focused on general comparisons between cultures and resorted to frequency, grouping of keywords[29] and/or collocates into semantic domains, and keyness as a primary means for highlighting cultural features. Two of these studies are briefly outlined below.

Muntz's (2001) analysis of cultural differences between British English and Australian English is – as the author openly declares – an application of Leech and Fallon's methodology, though a few differences should be highlighted. Muntz, in fact, considered also proper nouns of place names, as "it was felt they could provide information on a country's place within the world, which, in turn, contributes to its identity" (Muntz 2001: 394), and spelling differences, "as they provide evidence of cultural choices and language variation in progress" (ibid.: 394). Furthermore, she developed 11 specific domains,[30] different from Leech and Fallon's ones, and dedicated increased time to verifying concordances of results, eliminating unsuitable ones from the frequency list, and recalculating keyness of each word. Muntz's aim was two-fold: identifying cultural differences in the British and Australian varieties of English, and using these to assess Australian cultural identity. Her study highlighted the above mentioned domains in which Australian identity comes to the fore and provided evidence of Australia's independent identity. Muntz's conclusion (ibid.: 399) was that

> 'the results closely reflect some of the most commonly observed facts and stereotypes about Australia and its culture: the size of its coastline, barren interior, native animals and hot weather. Whilst this assures us of the validity of the corpora, it also tells us that Australians have inherited a way of thinking about such things from Britain. […] in Whorfian terms, then, perhaps the fact that Australians speak a type of English dictates the way Australians think about their physical environment, if not the frequency with which they use those words to describe it.'

Muntz's study clearly shows that relevant semantic domains are culture-dependent and not generally suitable for all cultures or cultural comparisons.

Schmid (2003), instead, applied Leech and Fallon's analytical method to the spoken part of the BNC,[31] in search for confirmation to Deborah Tannen's (1990)

[28] See note 23.

[29] In corpus and computational linguistics, the term *keyword* refers to words which appear in a given corpus with statistically significant higher (or lower) frequency than in a reference corpus (see Chapter 3, Section 3.6.2).

[30] Muntz's domains are: place names referring to other countries; origin/ethnicity adjectives; multiculturalism; prestige forms and borrowings from English; American spelling conventions; colour terms; geography; housing and communities; fauna and flora; weather and clothing; personal reference; abstract concepts (challenge and diversity).

[31] The British National Corpus (BNC) is a monolingual, synchronic, general corpus of British English. It includes samples of written and spoken language (about 90% and 10% respectively) from a wide range of sources, for a total of a 100 million words. The samples were carefully chosen in order to be representative of late 20th century English. The corpus is pos-tagged and encoded for headings, paragraphs, lists, and other features. Sample collection took place between 1991 and 1994.

ideas about gender differences.[32] Words from 18 domains inspired by Deborah Tannen's book[33] were investigated in terms of frequency and collocations. This study confirmed that differences indeed exist in the way British men and women speak, but not all the data were in line with what is suggested in the literature. Our interest in this paper, however, lies in the fact that it extended frequency counts and semantic grouping to collocates, and that the comparison was carried out in order to test a specific theory.

Finally, an interesting corpus study of cultural features is that of Manca (2008), who compared Italian and British farmhouse holiday websites, in an attempt to describe the promotional strategies employed and assess whether they are determined by cultural features. Manca started from an analysis of the collocates of all the adjectives in the wordlist of her corpus of British farmhouse holiday websites. The collocates were grouped according to three main different themes, which she labelled *description of rooms*, *description of surroundings*, and *description of food*. Attention then moved to her comparable corpus of Italian farmhouse holiday websites. Starting from words naming concrete instances of rooms, surroundings, and food (*prima facie* translations of words such as *room*, *kitchen*, *lounge*, etc.), she identified and analysed – dividing them according to semantic domains – the adjectives and phrases that are used in the Italian corpus for the three given themes. Constantly moving from collocate to collocate, and back and forth between the two languages, Manca highlighted that, despite apparent semantic similarities between the adjectives in the two comparable corpora, collocational profiles indicate that different promotional strategies are at work. The discovered differences were then commented and explained within Hall's theory of high- vs. low-context cultures.[34] Manca's corpus-driven study is qualitative in nature more than quantitative, except for basic frequency counts. She took advantage of typical corpus linguistic events: frequency counts, and above all collocational profiles, the latter being a feature that – as we shall see in the following section – has been commonly used by CDA researchers. However, her cross-linguistic approach, along with the explanation of results within a very specific theory of culture makes her contribution extremely relevant to general corpus linguistics, and to the current work.

[32] Gender differences are here equated to cultural differences, insofar as a clear distinction is now commonly accepted in sociolinguistics between males and females. In our reference framework, this distinction can be posited at the level of Fleisher's sub-cultures.

[33] The following domains were selected: personal reference; family; personal relationships; home; food and drink; clothing; car and traffic; computing; sport; public affairs; abstract notions; alleged "women's" and "man's" words; swearwords; hesitators, fillers, backchannel behaviour; linguistic politeness markers; linguistic markers of uncertainty and tentativeness; linguistic markers of conversational cooperation and support.

[34] Hall identified three cultural orientations: time, space, and contexting. By context he means 'the amount of information the other person can be expected to possess on a given subject' (Hall, 1983, p. 61). In high context cultures (HCC), writers expect readers to make use of their contextual knowledge to understand the text. Conversely, in low context cultures (LCC), readers expect contextual elements to be made explicit in the text itself (Hall, 1989).

*2.4.1.3 Within the framework of cultural systems theory*

A well-defined theoretical cultural framework of reference is what characterizes this series of papers. The aim of these papers, in fact, is as much cross-cultural comparison, as verification of specific theoretical elements.

Fleischer's (2002) study aimed to identify the semantic profiles of drinks in three different nations – Poland, France and Germany – and assess the level of conventionalisation of their images in the corresponding cultures. By conventionalisation he means the degree to which the semantic profile of a particular type of drink is common to, or shared by, all the members of a given culture.

To this purpose he presented groups of volunteers from the three cultures with a list of names of drinks and beverages (in alphabetical order), and asked them a rather broad question of the following type: what comes to your mind when you see each of the names of drinks in the list? Therefore, differently from the other linguistics studies, but – as we shall see later – in line with other disciplines, Fleischer's 'corpus' is based on elicited data: freely produced linguistic descriptions of drinks. The respondent's answers were semantically analysed. Three themes were distinguished, namely characteristics and connotations (*Konnotationen/Images*), trademarks and proper names (*Umschreibungen/Marken*), and evaluations (*Wertungen*). Words and phrases in each theme were then grouped into unlabelled semantic categories. Hapax legomena were ignored, as they were considered connected to individual feelings and preferences, rather than to collective (cultural) orientations. Then, Type-Token Ratio (TTR) of answers was calculated,[35] considering the described features as Types and individual instances of each feature as Tokens; and three levels of conventionalisation – high, medium, and low – were established using confidence intervals based on means and standard deviations. TTR was consequently used as an indicator of intra-cultural conventionalisation of the image of drinks.[36] As expected, results showed that different drinks had different levels of conventionalisation in the three cultures considered. So, for example, cocoa and milk show a high level of conventionalisation in all the three cultures. This means that, within each country, a large number of the semantic associations of a given drink are common to all the people belonging to that culture, or, in other words, that the images of cocoa and milk are largely shared by all members of that culture. On the other hand, beer, coke, and coffee show a low level of conventionalisation in Poland, a high level in Germany, and a medium level in France. Inter-cultural differences in conventionalisation were also highlighted, by means of correlation tests. Finally, the semantic profiles of each drink are described and discussed with reference to historical, sociological and more generally cultural events. Fleischer's work is particularly interesting for the direct link it creates between words (names of drinks), verbal associations (descriptions of drinks) and cultural symbols, as well as for the attempt to assess conventionalisation by means of quantitative analysis (Type-Token Ratio).

In line with Fleischer's study is a paper by Wilson and Mudraya (2006). Inspired by Fleischer's (2001, 2003) idea that, though judgements include both individual and cultural components, if the judgements made by different people tend to coincide, the

[35] See Chapter 3.

[36] For a discussion of issues in the use of TTR as a measure of conventionalisation, see Wilson and Mudraya (2006, pp. 680-681)

concept or thing being judged can be considered to have become conventionalised as a collective symbol, these two linguists used quantitative methods to “examine one possible link between the onomasiological and cultural levels” (Wilson & Mudraya, 2006, p. 679), i.e. to analyse the relationship that exists between naming of shoes and the establishing of different types of shoes as cultural symbols in Russia. Their analysis was based on elicited verbal answers collected through two separate tasks in Russian. In the first task the participants (all Russian native speakers) were shown 12 pictures, each presenting a different type of shoe, and were asked to name each type. In the second task, the participants were asked to provide 10 completions of one and the same sentence: “I think that the woman who wears these shoes ….”. The second task was a reviewed version of a widely used activity to study self-identity and aimed to identify specific traits attributed to those who wear a particular style of shoe. Names given to shoe styles and traits were counted and analysed by applying an evenness index, which – similarly to TTR – considers the relation existing between types and tokens, but it also takes into account the evenness of distribution of tokens among types.[37] No significant correlation was found between naming evenness and the evenness of associations and, consequently, the authors’ working hypothesis that ‘where respondents agreed more readily about an appropriate name […] would also […] elicit more conventionalised associations about their potential wearers’ (*ibid.*, p. 687) was disconfirmed for the Russian context. This experiment by Wilson & Mudraya is interesting not only for the quantitative methods applied, but also because it uses data to verify theoretical assumptions, while the other studies simply interpreted data within the chosen theoretical framework. Furthermore, though carried out by linguists, this study, and Fleisher’s before it, is closer in terms of method to studies in other disciplines. In fact, its starting point are not corpora as ‘traditionally’ intended in corpus linguistics but rather an electronic collection (and as such a corpus) of elicited data. Verbal associations are treated and analysed as collocates.

Finally, it is worth noticing that in both the studies here considered focus was on individual words, rather than on the culture itself.

### 2.4.2 Culture studies in other disciplines

Anthropologists Szalay and Maday (1973) were among the first to study subjective culture, or implicit culture, i.e. “psychological variables, images, attitudes, and value orientations” (*ibid.*, p. 33), starting from verbal associations and semantic grouping of verbal responses in one’s native language. The definition of culture these authors adopt is that of “group-specific cognitive organisation or world image” (*ibid.*, p. 33) made up of units of psychological meaning.[38] Psychological meaning is encoded in the words that are elicited by other particular words (hence the need to elicit answers through free verbal associations). Subjective meaning reactions elicited by a particular word are called Elementary Meaning Units or EMUs (Greenberg, 1960, quoted in *ibid.*, p. 34). Psychological meaning includes three major dimensions which are all equally important in the study of culture, namely composition, dominance, and organisation. Indeed, psychological meaning depends on the

[37] Further details on evenness indexes are provided in Chapter 6.

[38] This definition seems to me nicely in line with Lotman’s, and Fleisher’s theories.

composition of several distinct elements: visual imagery, context of use, brand, affective reactions, and function. Furthermore, some EMUs are dominant in a given cultural group, as they are more frequent than any others and their higher frequency influences cognitive processes. Dominance, therefore, is measured in terms of frequency. Finally, the psychological lexicon is organised according to networks based on affinity between elements. In this respect, EMUs cluster into semantic domains, or “large units of cognitive organization” (Szalay & Maday, 1973, p. 34). The method they suggested for and applied in the study and comparison of subjective cultures includes four steps: identification of cultural priorities; assessment of culture-specific composition of dominant EMUs; re-creation of the psychological network (affinities between semantic domains); and inferencing as regards culture-specific hidden attitudes, values, and beliefs. This method, also known as AGA (Associative Group Analysis), was illustrated with respect to the semantic domains of *education*, *manners* and *family*, comparing American and Korean students living in the United States. In order to identify cultural priorities of the two groups, the researchers asked the participants to make a list of 25 important domains of life and to provide for each of them as many associative responses as possible. Fusion of the lists of the Korean students and of the American students led to the definition of two culture-specific lists, which were subsequently merged to create a joined list of common priority topics which could be used as stimuli for the subsequent phases of the study. *Education*, *manners* and *family* were among the common domains, and were used as stimulus words for a free verbal association task. Content analysis was performed on both highly frequent and less frequent domains in each culture. In the content analysis phase, responses were grouped into fewer categories. The identification and categorisation of EMUs was done manually by judges from both cultures. For each stimulus word, affinity indexes – which are necessary to understand the cognitive organisation of each domain and compare domains of different groups – were calculated between the EMUs of each group of participants. Once the affinity structure and cognitive organisation had been established for each domain word, affinity relations between domains (e.g. *manners* vs. *family* and *education*) were assessed. As already noticed in previous studies, the authors observed that “words used in the representation of a particular domain generally show the same consistent cultural patterns” (*ibid*., p. 37), i.e. consistently low or consistently high affinity with the other domains for each group of participants. In the given experiment, *polite, greeting*, *manners*, and *to bow* (EMUs in the *manner* domain) consistently showed low affinity with the *family* and *education* domains for American students, but high affinity for Korean students. These results led the authors to suggest that a relatively low number of properly selected EMUs could be enough for obtaining generalisable cultural trends in broad semantic domains. Finally, the authors concluded that free verbal associations obtained in the native language and solicited with the method described “provide empirical data on the denotative as well as the connotative components of meaning” and “allow to reconstruct culture-specific cognitive organization by its main dimensions” (*ibid*., p. 41). This contribution seems important to the current research for several reasons. First of all, words are considered representations of psychological meaning units and hence of cultural trends. A direct relation is postulated between domain word (stimulus), words associated to domain word (EMUs), and subjective culture. Furthermore, in establishing dominant EMUs, frequency criteria are applied.

Psychologists Potash, de Fileo Crespo, Patel, and Ceravolo (1990) used the Miale-Holsopple Sentence Completion Test (Holsopple & Miale, 1950, cited in *ibid.*) to assess cultural attitudes in American and Brazilian college students. A sentence completion test was chosen by the authors because they considered it “the ideal projective instrument to measure cultural differences because this type of test provides high structure with a diverse variety of topics, permitting comparison among different cultural groups” (Potash, de Fileo Crespo, Patel, and Ceravolo, 1990, p. 657). The participants in the experiment were 39 American and 70 Brazilian college students. The authors’ aim was to compare the two cultures on the following topics: orientations about future; achievement motives and work ethics; interracial tolerance; and sexuality. For this reason, only the subject’s answers to selected sentence stems which the authors considered indicative of the desired variables were analysed. The subjects’ answers were scored as positive, indeterminate or neutral, or negative, and results were compared using the chi-square test. The authors report that American students gave more positive responses to the work ethic stems and more negative ones to interracial marriages. Furthermore, the Brazilian females gave more positive responses to the sexuality stem than did American women. This experiment is interesting for the current research because of its use of a sentence completion task – considered as the ideal projective technique for cross cultural comparisons – and for its analysing responses as positive, neutral or negative.

### 2.4.3 Discussion

Some considerations emerge from this review of empirical studies of culture.[39]

First of all, the limited number of cultural studies in corpus linguistics – which is low if compared to those in the qualitative area[40]– seems to indicate that culture has so far been a rather neglected area of study in this particular branch of linguistics, despite the fact that corpora and quantitative analytical methods are easily applicable to this purpose.

Second, although different research traditions have always been and still tend to be trenched within the boundaries of the paths paved by their respective predecessors, in more recent times attempts have been made to overcome such boundaries and employ methods from other traditions or disciplines. In particular disciplines such as anthropology and psychology seem to have much to offer to linguistics and cross-cultural analysis through language and corpora, both at theoretical and methodological levels.

Third, two broad approaches to the study of culture seem to be distinguishable: overall description of a culture or comparison between cultures through keywords – a sort of *top-down or picture-to-detail* approach in which attention is on the general picture and individual words are used to provide a description of it; and comparison of cultural perspectives based on focus on single concepts – a kind of *bottom-up or*

[39] Adjective ‘empirical’ is used here to emphasize the fact that the core element in these studies is observation of data (deriving from corpus analysis or specifically designed experiments), rather than theory.

[40] Wierzbicka alone can boast dozens of publications in the field.

*detail-to-picture approach* in which attention is constantly on individual words and their meanings, and the general picture is used as a reference framework for their explanation. In either case, keyword and collocate frequencies, as well as grouping of items to create superordinate domains (either entirely semantic or thematic) seem to be established techniques of cross-cultural comparison.

Fourth, as in all types of scientific research, the selection of a clear cultural theoretical framework is of paramount importance for a precise and convincing analysis of results.

The current research uses system's theories on culture, and Fleischer's theory in particular, as its theoretical framework, and corpus linguistics as its methodological framework. Consequently, the next chapter provides an introduction to corpora and corpus linguistics, and an overview of some major issues which are relevant in the current work.

Chapter 3

# Corpora and corpus linguistics

## 3.1 Introduction

The aim of the present chapter is to sketch an organic framework within which to understand the materials and methods used in the current research and which are described in Chapter 5. Consequently, the chapter provides an introduction to corpora and corpus linguistics, with an overview of some major issues. This is not intended as a complete list of all possible topics connected to corpora and corpus analysis; in fact, a selection of topics has been made, on the basis of their relevance to the current work.

## 3.2 What is a corpus and what is corpus linguistics

As any introductory book to corpus linguistics explains, the word *corpus* has always been used by linguists to indicate 'a collection of naturally occurring examples of language, consisting of anything from a few sentences to a set of written texts or tape recordings, which have been collected for linguistic study' (Hunston, 2002, p. 2). With the advent of computers and the development of modern corpus linguistics, the word *corpus* has come to acquire the more specialised meaning of a collection of *electronic* texts, selected and collected for a *specific purpose* according to *specific criteria*. Furthermore, as McEnery and Wilson (2001, p. 32) declare, in linguistics "there is also often a tacit understanding that a corpus constitutes a standard reference for the language variety which it represents". And this thought is most certainly what has long driven linguists in the creation of very large corpora, and in the development of coding standards so that their corpora could be shared within the academic community. However, as the review of studies in Chapters 2 and 4 may be taken to show, when the focus is on features that fall outside the realm of purely linguistic events, such as culture, attitude or behaviour, analyses are performed on collections of data (i.e. corpora) which are intended for a one-off use. For this reason, and being my work centred around culture conveyed by language, rather than language *per se*, I would support here a fairly broad definition of corpus, encompassing only two major basic features: an electronic format, connected to the use of computerised analysis tools; and the idea that a corpus should be designed, i.e. planned for some (general or specific) purpose.

Corpus design entails the application of selection and sampling criteria according to the purpose of the analysis, as well as issues of size, balance, and representativeness. These topics will be dealt with in Section 3.3.1. Furthermore, careful design is an essential prerequisite for the applicability of quantitative methods of analysis and for the generalisability of the results. Key features and methods in the qualitative and quantitative analysis of corpora – such as word lists and frequency; keyword lists and keyness; collocation; semantic preference; and semantic prosody – will be described in Section 3.6.

Finally, the electronic format and the use of computerised tools allow non-linear access to the information in a corpus. This represents a completely different approach to and provides a new perspective on both the language and the content of a corpus. Indeed, while some linguists consider corpus linguistics simply a method of research, others regard it as a new discipline.

In 2004, at the ICAME[1] conference in Verona, during a general meeting on the topic “Corpus linguistics 25 years on”, one particular linguist sounded offended by the fact that corpus linguistics was considered a method, and advocated that it was a real discipline. Her reaction might have been due to the fact that to some extent methodological interests are, by some, still considered Cinderellas with respect to theoretical interests (Leech, 1992, p. 105; Aarts, 2000, p. 7). However, it may be explained in more concrete terms by remembering that the advent of corpora and concordancing tools has changed the way we look at language and has deprived the native speaker of his exclusive status of judge and descriptor of the language. Historically speaking, corpus linguistics and the phraseological view of language it carried along was a radical turn from traditional prescriptive grammars, but also from Chomsky’s generative grammar and distinction between competence and performance.[2]

A look at some definitions of corpus linguistics provided in books and articles shows that the corpus linguistic community is indeed divided between considering corpus linguistics a method (see, for example, Kennedy, 1998; and Svartvik, 2007) or a discipline (e.g. Mahlberg, 2007).[3]

A fact is that, thanks to corpus studies, new and powerful theoretical views about language have emerged, such as the notions of local grammar (Barnbrook & Sinclair, 1995; Hunston & Sinclair, 2000) and pattern grammar (Hunston & Francis, 2000), or Hoey’s (2005) theory of lexical priming. Another fact is that corpus data and computerised analytical methods have been more and more used not only in linguistics, but also in other disciplines, such as the social sciences, psychology,[4] and marketing.

[1] The International Computer Archive of Modern English.

[2] Stubbs (2007a, p. 133) explains this opposition by describing Saussure’s and Chomsky’s approaches as rationalist deductive views situated within the French tradition of dualism, while the perspective adopted by most text and corpus linguists including Firth, Halliday and Sinclair is an empiricist inductive view which rejects dualism and is situated within the British empiricist tradition.

[3] Interestingly, Mahlberg (2007) equals this debate with Tognini Bonelli’s (2001) distinction between ‘corpus-based’ and ‘corpus-driven’, whereby a corpus-based approach avails itself of corpus data to exemplify, clarify and illustrate existing linguistic theories, while a corpus-driven approach analyses corpus data and lets the data drive the description of language or of a specific linguistic event.

[4] See for example the high number of publications by psychologists using data from the CHILDES database (http://talkbank.org/usage/childesbib.pdf), or studies such as Hogenraad (2004), or Hogenraad (2005).

## 3.3 Creating a corpus

The creation of a corpus entails two phases: planning the design of the corpus; and collecting the necessary texts. These two phases, which are equally important and to several extents problematic, are interrelated, and eventually depend on the ways (both technical and theoretical) in which the corpus will be analysed. The following paragraphs will discuss some major theoretical issues, with an eye to the needs of the current research.

### *3.3.1 Corpus design*

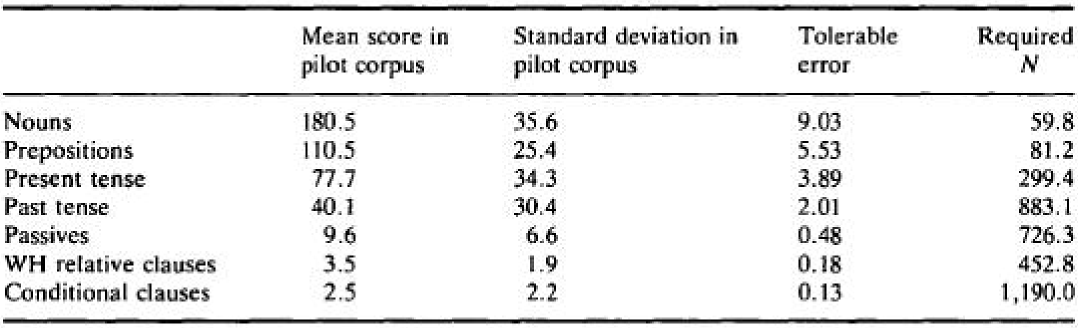
The design of a corpus depends on the purpose for which the corpus is created. If a corpus is created with the only purpose of showing students how to use corpus analysis tools, or "to encourage learners to investigate language data for themselves, the precise contents of that corpus may be relatively unimportant" (Hunston, 2002, p. 27). In most other cases, however, and in particular, when a corpus is created for the purpose of investigating a particular 'type of language' or linguistic event (e.g. British English vs. American English; the language of 19th century popular papers; the phraseology used in English civil law; or East London teenage jargon), the contents of the corpus are important and issues such as size, representativeness, balance, and sampling are usually called into play.

#### *3.3.1.1 Size*

In the last decades, several 'general purpose' corpora, such as the Brown and the LOB corpora (first generation corpora), the British National Corpus and the Bank of English (second generation corpora), have been assembled. Aiming to be representative of language in general, these corpora were created so as to include a wide variety of texts and text types, both written and spoken, and tended to be as large as possible. First generation corpora reached the important target of 1 million words – a great achievement for the time, given the then limited technological resources. But second generation general purpose corpora aim to be several hundred million words.

Indeed, the larger the corpus, the easier it is to retrieve a reasonable number of hits for infrequent or rare linguistic events (McEnery & Wilson, 2001). Furthermore, a very large corpus may also be required to understand the rationale behind grammatical or lexical forms even when they are highly frequent (see for example Mair, 2006 on the size needed for a full study of *get* in English passive constructions; or Granath, 2007 on the size needed to explain the four possible sentence structures after initial *thus*). On the other hand, extremely frequent events, such as function words and auxiliaries, can be easily retrieved in a statistically significant number of hits even in smaller corpora (see for example Biber & Finegan, 1991; Carter & McCarthy, 1995).

Biber (1993, pp. 253-254), estimated the minimal number of texts necessary for representing specific linguistic features in a corpus (Table 3_1).



| | Mean score in pilot corpus | Standard deviation in pilot corpus | Tolerable error | Required *N* |
|---|---|---|---|---|
| Nouns | 180.5 | 35.6 | 9.03 | 59.8 |
| Prepositions | 110.5 | 25.4 | 5.53 | 81.2 |
| Present tense | 77.7 | 34.3 | 3.89 | 299.4 |
| Past tense | 40.1 | 30.4 | 2.01 | 883.1 |
| Passives | 9.6 | 6.6 | 0.48 | 726.3 |
| WH relative clauses | 3.5 | 1.9 | 0.18 | 452.8 |
| Conditional clauses | 2.5 | 2.2 | 0.13 | 1,190.0 |

Table 3_1. Biber's (1993, p. 254) estimates
of required sample sizes (number of texts) for a general corpus

His estimates are based on a sample corpus of "481 texts taken from twenty-three spoken and written registers" (Biber, 1993, p. 253), and calculations are made considering mean score, standard deviation and tolerable error of each individual feature

Specialised corpora can usually afford to be smaller than general corpora. On the basis of their experience, Bowker & Pearson (2002, p. 48) declare that "*well-designed* corpora that are anywhere from about ten thousand to several hundred of thousands of words in size have proved to be exceptionally useful in LSP studies".[5] Indeed several recent corpus studies in LSP are based on small-medium sized corpora (see for e.g. Warren, 2007; Gledhill, 2000; Luzon Marco, 2000; Heyland & Tse, 2005; Banks, 2005, just to mention a few). Scientific support to these empirical habits could come from studies on closure measurements. A corpus can be said to reach closure as regards a particular type of linguistic feature when an increase in the size of the corpus does not bring in new instances of the given feature. In a comparative study of closure in a sublanguage corpus – namely the IBM corpus – and two unconstrained language corpora – the APHB (American Printing House for the Blind) corpus and the Canadian Hansard corpus –, McEnery and Wilson (2001, p. 176) found that the IBM sublanguage corpus showed a very high degree of lexical closure; in fact "the lexicon used by the language of the IBM manuals nearly enumerates itself within the first 110,000 words of the corpus".[6] A corpus that reached closure in a specific feature could be considered representative of the given feature.

Other aspects that are frequently taken into consideration when talking about size are: speed and efficiency of the access software, and the human ability to deal with great amounts of data. Not only the computer might be unable to process great amount of data, but also the human brain. This consideration may lead to the creation of smaller corpora, to the use of sub-corpora, or to the selection (either manual or automatic; randomised or reasoned) of the concordance lines on which analysis is carried out. The last two solutions are suggested by Sinclair (1991, 1992), among others.

[5] Emphasis added.

[6] The IBM corpus showed a greater tendency towards closure also at morphosyntactic and sentence-type levels.

#### *3.3.1.2 Representativeness, sampling and balancedness*

Although size matters, to quote the title of Granath's (2007) paper, this is not the only important issue in corpus design. Another feature to consider is 'representativeness' (or 'representatitivity', as some seem to prefer),[7] generally seen as necessary feature for any attempt at drawing generalisations from corpus data. Unfortunately, after decades of corpus building, representativeness is still a highly controversial and debated issue, at least when talking of general corpora aiming to be representative of the language in question.

Biber (1993) suggested language-internal criteria – such as situational (register) and linguistic (lexical and morphosyntactic features) variability – as essential elements for a corpus to be representative.[8] Váradi (2001), in strong critical opposition to Biber, advocated the use of language-external (i.e. sociolinguistic) criteria and proportional sampling based on objective demographic data.[9] Leech (2007), suggested that "the representation of texts [in a corpus] should be proportional not only to their initiators [i.e. speakers and writers], but also to their receivers" (*ibid.*: 138), as the importance of a text depends on the number of receivers it has. Furthermore, Leech (*ibid.*) sees representativeness and balancedness are scalar values; consequently some degree of representativeness and balancedness should be pursued and aimed at by corpus compilers, though attaining these desiderata to the full might be impossible. Finally, other linguists such as Kilgarriff and Grefenstette (2003) argue that every corpus is representative of nothing else but itself. Interestingly, even a strong supporter of representativeness in corpus design such as Leech (2007, p. 145) accepts that "even without such qualities as representativeness, a corpus retains the merit [...] in showing up 'language as it is actually attested in real life'".

This debate leads me to believe that different possible views of representativity can be considered and applied depending on the purpose for which a corpus is created. Most of the considerations by Biber, Váradi, and Leech reported above, and certainly nearly all of Biber's calculations, are based on the assumption that a corpus is built for the purpose of linguistic analysis. But corpora can be created also for other purposes and in these cases corpus creation may require the application of other internal or external criteria. In particular, as the cultural theories reviewed in Chapter 2 suggest, the parameters that have a major impact on the cultural core that is common to all members of a single culture[10] are neither register, nor text type, nor specific linguistic phenomena, but rather time (the year or decade when the texts were written), and authorship (intended as knowing that the author belongs to the single culture under investigation). Furthermore, a preliminary experiment by Bianchi (2007) – briefly summarised in Chapter 4 – seems to suggest that a relatively large size and

[7] While Renouf (2007, pp. 33-34) argues that the term 'representativity' has replaced in popularity the older form 'representativeness' when talking about this issue in corpus linguistics, Leech seems to make a subtle distinction between the two terms, defining 'representativity' as "the degree to which a corpus is representative" (Leech, 2007, p. 133).

[8] Hence his calculations of sample and corpus size based on calculations of distribution of morpho-syntactic and lexical phenomena such as prepositions, sentence types, and hapax legomena. See Section 3.3.1.1 and Table 3_1.

[9] Such criticism mines the ground upon which corpus linguistics studies have been carried out so far, as no corpus exists that matches the level of statistical representativeness advocated by Váradi.

[10] Here and elsewhere in this work I will use Fleischer's terminology.

heterogeneity may be sufficient when the corpus is created around a specific word/concept.

### *3.3.2 Text collection*

Text collection is usually the most time-consuming part in corpus creation. In fact, depending on the planned design, texts are to be searched for, selected or sampled and last but not least acquired in a suitable electronic format. Despite the recent advances in OCR (Optical Character Recognition) technologies, scanning printed texts still requires careful revision and correction of the acquired text, which can only be carried out manually. This, along with the fact that an increasing number of publications are now available on-line (Meyer, 2002), has gradually led researchers to look at the Web as a potential source of corpus data. The Web lends itself to the creation of different types of corpora, using different types of 'collection methods' that range from manual download of individual web pages to full automatic download of automatically selected sets of pages. An overview of major issues in the use of the Web in corpus studies is provided in Section 3.4.[11]

## 3.4 Corpora and the Web

The last decade has seen a rise in interest towards the Web and its potential in corpus studies, as testified, for example, by the growth of several study and research groups on the topic,[12] and dedicated conferences.[13]

The most commonly used expression to refer to this area of interest is 'Web as corpus'. However, as Bernardini, Baroni, and Evert (2006, p. 10) interestingly point out, this expression subsumes at least four separate senses: 1. querying the Web via commercial search engines and using the retrieved data as concordance lines (i.e. using the web as corpus surrogate); 2. creating corpora from the Web (i.e. using the Web as a corpus shop); 3. considering the Web as a corpus proper; 4. creating a new object, a sort of mini-Web (or mega-corpus) adapted to language research.

The first two scenarios have been fostered by the enormous growth of a multilingual Web, the development of search engines offering rather easy-to-use and flexible text search features, the linguists' need for ever larger text corpora, and an expanding use of corpora in teaching as well as research environments. These two scenarios, though seemingly rather well accepted in the scientific community at large, still require much explicatory and descriptive efforts. The third scenario is still a much debated issue. The last scenario – a development of the second one – is very recent and still under investigation.

The current research project will take advantage of large corpora composed of text retrieved from the Web using spidering tools,[14] and subsequently POS-tagged and

[11] For a wider discussion of corpora and the Web see Gatto (2009).

[12] See for example: the Web as Corpus Special Interest Group of the Association for Computational Linguistics (ACL SIGWAC) (at http://www.sigwac.org.uk/); and the WACKY project (http://wacky.sslmit.unibo.it/doku.php).

[13] The Web As Corpus workshop is now at its seventh edition.

[14] Spidering tools are Web crawling scripts, i.e. programs which browse the Web in a methodical way and retrieve text.

lemmatised and made available to the general public. As such, it falls within the Web as a Corpus Shop and the mini-Web/mega-corpus scenarios.

However, the four scenarios, though different one from the other, are not completely apart, since any use we make of the Web strongly depends on the type and quantity of text available on the Internet, not to mention the methods to access it. For this reason, the following paragraphs will introduce issues related to the Web as corpus proper, before discussing the automated processes for querying the Web and creating corpora from it.

### *3.4.1 The Web as Corpus proper*

As Kilgarriff (2001a, sec. 1) vividly puts it, the Web is an anarchic object showing several features that seem to row against its scientific use as a corpus:

> “First, not all documents [on the Web] contain text, and many of those that do are not only text. Second, [the Web] changes all the time. Third, like Borges's Library of Babel, it contains duplicates, near duplicates, documents pointing to duplicates that may not be there, and documents that claim to be duplicates but are not. Next, the language has to be identified (and documents may contain mixes of language). Then comes the question of text type: to gain any perspective on the language we have at our disposal in the web, we must classify some of the millions of web pages, and we shall never do so manually.”

Nevertheless, Kilgarriff and Grefenstette (2003) argue that the Web can still be considered a corpus in the broad sense of the word, i.e. a collection of (electronic) texts for language or literary – and, I add, also cultural – study. And the Web is just as representative as any other corpus. In fact, as these authors argue, currently available general corpora such as the British National Corpus[15] – created according to shared and accepted criteria –, though 'balanced', are always arbitrary selections of text, and their concept of ‘balance’ is an internal, rather than external one. For example, in the general world, speech events exceed writing events, while the reverse is true in the currently available general-purpose corpora, due to the fact that transcribing speech events is still a problematic and time-consuming task. Similarly, due to both technical as well as theoretical issues, such as size and time limits, fuzziness of text type classifications, and continuous emergence of new genres, the current general corpora only include a selection of text types. On the other hand, the Web contains all traditional text types as well as some emerging ones (Yates, 1996; Leech, 2007). Furthermore, online texts are an excellent resource for the study of emerging usage and current issues, as the Web is “a self-renewing linguistic resource [that] offers a freshness and topicality unmatched by fixed corpora” (Fletcher, 2004, p. 1).[16]

This last point is particularly relevant when analysing culture, since, as we have seen in Chapter 2, culture and cultural associations may change very quickly over time (Nobis, 1998), and considerations about when the corpus and the texts it includes were created are of paramount importance (Bianchi, 2007). Commercially

[15] See Chapter 2, Note 31.

[16] Here and in the following quotes of Fletcher (2004), page numbering refers to the paper retrieved from the Web.

available corpora, be they of a closed or monitor type,[17] become quickly obsolete and tend to include texts from a wide time-span. Therefore, for a synchronic study of current cultural features, the Internet can be seen as an essential resource for the creation of an up-to-date general or specialised corpus.

Besides the issue of representativeness, another concern which usually arises when suggesting the use of the Web as corpus is size. Size is an important issue in corpus linguistics for three main reasons, namely comparing corpora, performing quantitative analyses and statistical calculations, and establishing representativeness. Calculating the size of the Web or of the published web pages in any given language is an insidious task, as the Web is constantly updated and grows almost by the second. Any calculation, therefore, would be almost immediately out of date.[18]

However, if we use the Web simply as a source of data to download for the creation of a corpus, the size issue is at least partially downgraded. Our corpus will have a finite number of words that depends on the purposes for which the corpus is created.[19] Such a corpus could be easily compared to other corpora, and quantitative analyses will be possible on the data of the corpus. Furthermore, as we will see in the following section, Web corpora tend to be more varied in content than traditional corpora, which may have an important impact on the size needed for a corpus.

### *3.4.2 The Web as Corpus Shop: Creating corpora from the Web*

A major issue in the Web as Corpus Shop scenario is representativeness. This explains the effort that, at least in these initial phases of studies in the field, those developing and using automated procedures for creating corpora from the Web put in assessing the representativeness of the retrieved corpora.

Fletcher (2004) compared a small corpus of online documents in English – including only about 11,000 running words – to the written texts in the BNC. The comparison showed differences between the two at the level of spelling (US vs. British), register (interactive vs. narrative style), and type of language (prominence of the language of news and politics vs. prominence of academic language). Furthermore, the Web corpus was more varied as far as frequent lexis is concerned. In fact, the most common 5000 words in the BNC were all present in the Web corpus, while the reverse was not true, and this despite the much smaller size of the web corpus (1/16 of the BNC). Thus, the Web corpus could be considered more representative in terms of the most frequent words.

Studies on several-million-word Web corpora for general purposes created using spidering tools showed that Web corpora assembled following a few reasoned basic criteria concerning preliminary choice of query words and size could be considered comparable to standard balanced hand-collected corpora, in terms of coverage of various text types and topics (see Sharoff, 2006; Ueyama, 2006), though not of register (Baroni & Ueyama, 2006).

[17] Monitor corpora, also called open corpora, are corpora that are constantly being expanded through the addition of new texts. On the other hand, closed corpora, once compiled, are no longer expanded.

[18] It must be said, however, that estimates of the size of the Web at given points in time are possible and have been computed. See for example Lawrence and Giles (1999), Lyman, Varian *et al.* (2003), Kilgarriff and Grefenstette (2003), Grefenstette and Nioche (2000), Gulli and Signorini (2005).

[19] See Section 3.3.1.1 for issues relating to corpus size.

Finally, in a preliminary experiment to the current study which focused on the semantic associations of key word *cioccolato* in Italian, Bianchi (2007) compared a specialised corpus manually created around the key word using the Web as source for text retrieval to 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). The two corpora were compared in terms of semantic and conceptual categories. The comparison showed a limited number of differences, and – by applying the Mann-Whitney test – it was verified that those differences were not statistically significant, as if the two corpora, though constructed with different criteria and purposes in mind, included samples from the same population. Furthermore, the differences could be explained by the time gap between the two corpora.

This preliminary experiment suggested that suitable data for cultural analysis can equally be retrieved from a very large general corpus, or a small-to-medium-sized specialised corpus, provided the latter has been created including a wide variety of texts by different authors. Furthermore, it confirmed that, for this type of cultural analysis, the major concern in corpus creation, along with text variety, seems to be time-coverage, and this is precisely where the Web comes to an aid.

### *3.4.3 Further issues and comments*

An issue that is certainly relevant when dealing with Web corpora retrieved using spidering tools, is that of authorship. In fact, a large quantity of Web text does not bear the author’s name, and once a page has been automatically retrieved and included in a Web corpus any possibility of recovering information about the author is lost. The most common solution to work around this problem is limiting Web searches to a specific language and Internet domain. Almost all the Web corpora created so far, and certainly the ones which will be used in the current research (and which are described in detail in Chapter 5), were created following this procedure. However, for some languages, such as English, which includes several different international varieties and which has been gradually establishing itself as a lingua franca and as 'the' language of the Internet (Crystal, 2003), the sole fact that a page is written in a specific language or appears in a geographically located web site (e.g.: .uk) does not guarantee that the author is native to that language. For other languages, including Italian, whose use is still limited to Italy and a very small area in Switzerland, the chances that a piece of text in that language has been written by a non-native are few. Some attempts have been made to sieve out text by automatically detecting spelling and grammar mistakes (see for example Fletcher, 2004; and Ringlstetter, Schulz, & Mihov, 2006). These methods, however, seem to be still in their infancy, and have not been applied to the Web corpora used in the current study. Nevertheless, no spelling or grammar mistakes which might suggest that the texts were not written by native speakers were noticed while performing manual coding of the Web data. We will come back on this issue later on in the work, after the analyses on the Web corpora have been accomplished and the results have been compared to those of the elicited data.

A second issue is that of readership. Web pages are a form of public communication and, when they are written in an ‘international’ language such as English, the (perspective) audience is international. However, every culture has specific values and beliefs, and the average speaker is absolutely unaware of that. In fact, as already noticed in Chapter 2, values, value orientations, beliefs, and judgments belong to the informal level of culture. This informal level of culture is where people normally react in everyday life and communication (Hall, 1982). Indeed, adaptation of discourse to target readers is only performed by experts in cross-cultural communication, such as professional translators and marketing experts. Consequently, only a specific part of Web communication can be expected to have been adapted to the values of a perspective audience belonging to a different culture from the author’s one. As regards the current research, the authorship and readership issues are of no relevance, given that use of Italian is generally limited to Italy and its native residents. The two issues, however, might bear relevance in the discussion of the English Web data, and in comparing them to the results of the elicited data which were clearly written by native speakers with a native audience in mind.

Finally, although the semantic associations that are common to a whole single culture emerge in language regardless of register (i.e. formal vs. informal language) and text type (poem vs. letter vs. blog vs. news article, etc.), the communicative purpose for which a specific text has been written and the audience to which the text is targeted may influence the semantic content of the text. The Web as a whole is an immense box containing varied but unspecified material which cover all aspects of society and range from scientific papers to gossip news, from marketing advertisements to personal narratives (e.g. blogs), from official legal documents to transcripts of songs and films, from religious text to every day news. But every single document in the Web mirrors only one of those aspects. Unfortunately,

> “automated methods of corpus construction allow for limited control over the contents that end up in the final corpus [and] the actual corpus composition needs therefore to be investigated through post-hoc evaluation methods” (Baroni, Bernardini, Ferraresi, & Zanchetta, 2008, Sec. 3).

The Web corpora chosen for the current experiments – described in Chapter 5, Section 5.2.2.1 – were compared to general reference corpora by their authors. The Web data extracted from those corpora in the current research will be compared to elicited data in Chapter 10.

### 3.5 Annotating a corpus

Annotation (or markup) is the act of adding explicit (meta-)information to a corpus. Different types of information can be added: textual, such as part of speech information (POS tagging), syntactic annotation (parsing), semantic annotation; and meta-textual, such as sociolinguistic information. Depending on the type of information, annotation takes place at word, sentence, paragraph, or file level. Furthermore, annotation can be done manually, automatically by means of specific software tools, or semi-automatically.

An annotated corpus can be queried and analysed starting from the annotated information, as well as from words in the corpus; this is what Hunston (2002, p. 79) calls 'category-based' methodology. Though annotation is not a compulsory step for carrying out corpus investigation involving categories,[20] it certainly makes category-based investigation easier, and is generally considered added value to a corpus (*ibid.*, pp. 79-80). However for annotation to be usable, it has to be systematic, precise, and intelligible to the end-user.

The following paragraphs provide an introductory overview of the annotation processes which will be used in the current work, namely POS tagging, lemmatisation, and semantic annotation. Details of the annotation systems of the corpora used in the current work will be provided in Chapter 5.

### *3.5.1 Part-of-speech tagging*

Part-of-speech tagging – usually called POS tagging, or simply tagging, but also known as grammatical tagging or morphosyntactic annotation (McEnery & Wilson, 2001, p. 46) – takes place at word level and adds morphosyntactic information next to each word in the corpus. The information added makes the grammatical category to which each word belongs explicit, by adding codes such as: adjective, comparative; noun, countable, singular; verb, simple present, 3rd person, etc. Punctuation is also tagged. Different tagsets may distinguish a different number of categories, and consequently include a different number of tags, and they may use very different codes for the same categories.

Deciding the number and types of tags to use is not the only issue in POS tagging. Other issues include how to deal with multi-word units which function as a single grammatical unit (e.g.: *so that*, or *such as*) and contracted forms (e.g.: *don't*, or *it's*).[21]

As Hunston (2002, p. 82) points out, "tagging needs to be done automatically […] otherwise the labour of adding tags by hand would outweigh the advantages of having them". POS tagging was the first type of tagging to be accomplished automatically, and with relatively good results; in fact, in 1971 the TAGGIT program (developed at Brown University) already achieved an accuracy of 77% (Green & Rubin, 1971). Currently, POS tagging techniques have reached excellent levels of accuracy. CLAWS, the tagger developed at UCREL – Lancaster University and which will be used in the current research, can boast an error rate as small as 4%-2% (Rayson, Archer, Piao, & McEnery, 2004). Furthermore, taggers have been created for languages other than English. The most famous and popular language independent tagger is certainly Tree Tagger (Schmid 1997), developed at the University of Stuttgart. The accuracy of this tagger in its English version is over 96% (*ibid.*), in its Italian version it seems to be around 91% for known words and 86% for unknown words according to Sogaard (2009) and about 96% according to Schmid, Baroni, Zanchetta, and Stein (2007). The general Web corpora which will be used in the

[20] A short list of category-based studies carried out on unannotated corpora is offered by Hunston (2002, p. 80).

[21] For a detailed description of how these issues were solved in CLAWS, the POS tagger used in the current research, see Garside and Smith (1997).

current research to create specialised corpora about given key words have been POS tagged by their authors using Tree Tagger.[22]

Finally, the POS tagging process could be finalised with post-editing, i.e. detection and correction of tagging mistakes in the tagged corpus. Post-editing has traditionally been a manual, time consuming, and costly task. Recently, computational linguists have been experimenting with methods for automatic post-editing of POS tagged corpora (see for example Loftsson, 2009). However, neither manual nor automatic methods seem to guarantee an error-free tagged corpus, especially when the corpus is rather large.

POS-tagged corpora allow corpus linguists to perform advanced searches in the corpus, based on POS tags, and are used by computational linguists to train and develop POS taggers. Furthermore, part-of-speech tagging is the first necessary step for other types of annotation, such as lemmatisation, semantic annotation and parsing.

The following sections introduce some basic issues in lemmatisation, and semantic annotation. Parsing, i.e. syntactic annotation, will not be used in the current research; consequently it will not be discussed here.

### *3.5.2 Lemmatisation*

Lemmatisation, i.e. "the reduction of the words in a corpus to their respective lexemes" (McEnery & Wilson, 2001, p. 53), is an important process in corpus linguistic tagging. It differs from stemming as the latter is a semantic process, while lemmatisation is a grammatical one. In a lemmatised corpus, next to each word, its lemma is provided. This entails the automatic recognition of all the inflected forms. In English, inflected forms are found in verbs (e.g.: *plays, played*, and *playing* belong to lemma PLAY, and *goes*, *went*, *gone*, *going* to lemma GO), nouns (e.g. *children* belong to lemma CHILD; *flowers* to lemma FLOWER), and adjectives (e.g. *greater* and *greatest* belong to lemma GREAT). In Italian, inflected forms characterize verbs, nouns, adjectives and articles and the variety of forms belonging to a lemma is much greater than in English. In fact, Italian includes 3 different verb conjugations, about 10 simple verb tenses, and different endings for each person in almost all verb tenses; nouns can be modified by suffixes indicating dimension, affection, etc. (e.g. *casina*, *casetta*, *casettina*, *casuccia*, *casona* are different forms of lemma CASA); adjectives are inflected to distinguish masculine/feminine, singular/plural, and degree (e.g. *bella*, *belli*, *belle*, *bei*, *bellissimo* are forms of lemma BELLO);[23] while articles are inflected to indicate masculine/feminine and singular/plural (*il*, *lo*, *gli*, *i*, *l'* are all different forms of the definite article). In both languages, however, there are cases when a decision has to be made about whether two words belong to the same lemma or to different ones. A controversial case is that of the Italian definite article: *il*, *lo*, *gli*, *i*, *l'* are all forms of the masculine definite article, while *la*, *le*, *l'* are forms of the feminine definite article. Should they be considered as two separate groups/lemmas (as the Tree-Tagger does) or should they be considered as forms of one lemma (the definite

[22] A detailed description of these corpora is provided in Chapter 5.

[23] In theory also diminisher *bellino/a*, and comparative forms *più bello/a/i/e* belong to lemma BELLO, but they do not seem to be treated as such by some taggers, such as the Tree-Tagger.

article, as dictionaries do)? As usual, the answer depends on the aim for which lemmatisation is carried out, i.e. on the granularity needed in the research.[24]

A typical problem is represented by use of apostrophes, as in the case of Italian definite article *l’*, or English Saxon genitive *‘s*. The Tree-Tagger, for example, does not seem to recognize *l’* as an article and treats it as part of the word that follows it. Analogously, *‘s* does not seem to be considered as a genitive, while *child’s* and *children’s* could legitimately and reasonably be classified under lemma CHILD.

Automatic lemmatisation is usually performed by POS taggers, but this process takes place after POS tagging has been completed. During POS tagging, disambiguation of words like *plays* – verb *play* vs. noun *play* – takes place. Next, the lemmatiser adds lemma information to each word/grammatical_category pair. Usually, lemmatisers are based on lemma dictionaries, but they may also include rules for desuffixation after automatic recognition of suffixes; these apply when a word is not included in the dictionary (Baroni, 2004).

### *3.5.3 Semantic annotation*

By semantic annotation, here, we mean “the marking of semantic features of words in a text, essentially the annotation of word senses in one form or another” (McEnery & Wilson, 2001, p. 61). In other words, with semantic annotation every word in the corpus is attached a label which indicates the semantic field to which the word belongs. Semantic fields[25] are conceptual abstractions which include not only synonyms, but also other words that are in some way logically associated to the given concept, including hypernyms and hyponyms. Indeed, these mental abstractions are determined by the way the world is, the way the human mind works, and the operational context within which the semantic classification is needed. Consequently, the phrase ‘Virgin Mary’, for example, could be rightfully classified as ‘religion’, but also as ‘woman’ or ‘mother’. Furthermore, like in hypernymic/hyponymic relations, different ‘levels’ of abstraction are possible: ‘cat’ could be tagged as ‘feline’, ‘mammal’, ‘animal’, or even ‘living being’ if necessary. This issue is sometimes called ‘granularity’ or ‘delicacy of detail’, and choice of one level of granularity over another one is a pragmatic rather than theoretical issue (Wilson, 2003).

Following Schmidt (1988), Wilson and Thomas (1997, p. 55) declare that although “there is no such thing as an ‘ideal’ semantic annotation system”, some general criteria can be listed for the creation or selection of a suitable semantic annotation system. Hence they offer the following criteria (Wilson & Thomas, 1997, p. 55-57):

> 1. It should make sense in linguistic or psycholinguistic terms;
> 2. It should be able to account exhaustively for the vocabulary in the corpus, not just for part of it;
> 3. It should be sufficiently flexible to allow for those emendations which are necessary for treating a different period, language, register or textbase;
> 4. It should operate at an appropriate level of granularity (or delicacy of detail);

[24] See Wilson (2003) for an example of a case when limited granularity could be desirable.

[25] Semantic fields are also called semantic domains, conceptual fields, lexical domains, or lexical fields (Wilson & Thomas, 1997).

5. It should, when appropriate, possess a hierarchical structure;
6. It should conform to a standard, if one exists.

Semantic analysis is a complex task with several issues to be considered, including homography, polysemy, sense ambiguity, units of meaning, and figurative language, as a consequence of the complex network of relationships that subtends words in a language. Indeed, in our mind concepts do not appear to form discrete categories, but rather "fuzzy sets", as prototype theories have shown, and it is not infrequent to find words that fall into more than one semantic field (Wilson, 2003).

The following paragraphs summarize how these problems are dealt with in the UCREL semantic analysis system (USAS), a tool for semantic annotation developed at the University of Lancaster and which will be used in the analytical part of the current work. 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-patient interaction, and from translation to cross-cultural comparisons (see http://ucrel.lancs.ac.uk/usas). In particular, in a cross-cultural study on attitude to shoe fashion by Wilson and Moudraia (2006), the results of automatic tagging with the USAS tagset were compared to those of manual semantic coding and the two coding methods highlighted similar between-group differences. Further details about this tagging system and its semantic categories are provided in Chapter 5.

As described in Rayson, Archer, Piao, and McEnery (2004), semantic annotation employs two main lexical resources: a single word lexicon of 42,000 entries and an idiom lexicon of 18,400 entries, plus an extra single lexicon of about 50 words preceded by wildcard characters to match things like weights and measures. The idiom lexicon – aimed to resolve the tagging of units of meaning – includes phrasal verbs, noun phrases (e.g. *riding boots*), proper names, and true idioms. Tagging of idioms takes priority over tagging of individual words, in order to prevent tagging overlap. Disambiguation of homographs and polysemy is resolved resorting to a combination of seven techniques including POS tagging, which is a pre-requisite in automatic semantic annotation processes, as well as frequency and other types of statistic information and context-sensitive rules.

Finally, USAS's solution to the problem of a word falling into more than one semantic field is attaching several separate labels to the same word (when applicable), and then choosing the most suitable one on the basis of frequency or domain considerations. However, there might be cases when selection of one semantic category only is not applicable. Indeed, this is not the only possible solution to this problem: Wilson (2003), for example opted for assigning more than one category to the same occurrence of a word. The multiple-assignment solution will be adopted also in the current research when tagging data manually (see Chapter 5).

## 3.6 Analysing a corpus: major analytical features and methods

Corpus analysis is accomplished taking advantage of specific software tools, or concordancers. Concordancers may differ in terms of number of features offered, user interface, supported file format, output format, and query language; however, some basic analytical features are common to all of them, namely the frequency word list and concordance features. More advanced concordancers (among which Wordsmith Tools, used in the current research) include other features such as automatic extraction of clusters, collocates, keyword lists, as well as the computation of various types of statistics.

The following paragraphs illustrate the analytical features and methods which have been used or mentioned in the current research. The degree of detail in each paragraph reflects the degree of relevance each feature had in the research. Indeed, my experiment, which focused on semantics but aimed to establish cultural associations of given key words, made ample use of frequency lists and, at least in a preliminary experiment, keyword lists; concordancing was necessary to understand the context of the key words; a look at collocations and semantic preference helped semantic tagging, while colligation was ignored; finally semantic prosody was systematically analysed.

### *3.6.1 Wordlists and frequency*

Wordlists, i.e. lists showing the number of occurrences (raw frequency) of each word in the corpus, provide an overview of the corpus; for this reason they are the first thing that corpus linguists tend to examine, in both quantitative and qualitative studies. As we have seen in Chapter 2, wordlists have also largely been used as a starting point for cross-cultural comparisons.

Wordlists, which can be ordered alphabetically, or by frequency, are always accompanied by information on the total number of running words (tokens), and the total number of word forms (types),[26] in order to allow conversion of raw counts into percentages (normalisation) and comparisons between corpora of different size, as well as calculation of Type-Token Ratio, a measure of lexical variation within the corpus.[27]

If necessary, ‘abridged’ wordlists can be created by applying a specific ‘stop list’[28] which excludes undesired word forms – for example function words – from the wordlist. In the current research, stop lists will be used to filter out function words, as well as other non-desired words such as the various forms of the key word itself, in a series of experiments aimed to explore 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.

[26] *Word forms* or *types* are not to be confused with *lemmas*. In fact, the lemma EAT – to quote an example from Hunston (2002, pp. 17-18) – would include word forms such as *eat*, *eats*, *eating*, *ate*.

[27] More sophisticated concordance packages may also provide other types of statistical information, including standardized type frequency, Type/Token Ratio (TTR), average word length, number of sentences, and average sentence length.

[28] A stop list is a list of words that the researcher wants to exclude from the analysis. The list is created by the researcher – usually in the form of a txt file.

If the corpus is not POS-tagged or lemmatised, the information provided by the wordlist is rather rough, since it will not take into account issues such as polysemy, homography and different word-classes: all occurrences of word *bank*, for example, would be listed under the same entry, regardless of their meaning 'bank of the river', or 'financial institution', and of their being noun or verb ('to bank'). Consequently, entries in an untagged wordlist need to be checked against concordances, to see the contexts in which the given tokens appear. The wordlist of a POS-tagged and/or lemmatised corpus, on the other hand, provides the frequency of lemmas and/or words according to their POS category.[29]

Quantitative comparisons between wordlists is only possible when the frequency counts in the two corpora are normalised to the same figure; it also requires that frequency counts have been conducted in the same way as regards stop lists, numbers, hyphenation, apostrophes, and the like.[30] Comparison of normalised figures, however, only tells us where similarities and differences appear, but not whether they are significant, or due to chance (Meyer, 2002, p. 126). To this purpose, statistical procedures should be applied, and several types of statistics have been proposed, including the *chi-square test*, the *chi by degrees of freedom*, the *log-likelihood test* and the *Mann-Whitney test*. None of these tests is exempt from drawbacks and debate over their application seems to be still open. Most concordancers, however, offer only the *chi-square* and *log-likelihood* options.[31]

The current research will take advantage of wordlists, as a starting point for the identification of semantic categories. Consequently, wordlists from different corpora will not be compared as such, but only after applying semantic analysis. The statistics used to perform quantitative comparisons will be described and discussed in the relevant chapters.

A few more interesting comments could be made about frequency in a corpus list. First of all, an almost linear inverse relationship between word frequency and word rank has been noticed, which is described by Zipf's law. In other words:

> "a word list [and this appears to be true of any word list based on at least a few hundreds of words] contains a very small number of very highly used items, and a long declining tail of items which occur infrequently, with roughly half occurring only once as hapax legomena" (Scott & Tribble, 2006, pp. 27-29).

As a result, the most frequent 150 words in a wordlist typically account for about half of the words in the corpus, though this number may vary depending on factors such as corpus size, genre and register (Powers, 1998). A consequence of the Zipfian distribution of the words in a corpus is the fact that, as the size of a corpus increases new vocabulary enters the corpus following a distribution that is marked, after an initial sharp increase, by a gradual reduction in the number of new words; this is known as Heaps' Law (Heaps, 1978). Although this distribution is not really upper-bounded, due to the presence of proper names and typos, if collecting data from the

[29] For details about lemmatisation, see Section 3.5.2.

[30] For a detailed discussion of issues and possibilities in creating a word list, see Scott and Tribble (2006, pp. 13-20).

[31] For a survey of the various statistics used for comparing corpora see Kilgarriff (1996a, 1996b), and Rayson (2003).

same genre and time period, enlarging a corpus over a certain limit will yield diminishing returns in terms of giving new vocabulary.

Furthermore, some words appear consistently in a high number of texts, while others appear frequently only in a limited number of texts or text types (Scott & Tribble, 2006, p. 29). This suggests the importance of an analysis of the distribution of the words across texts, as well as of their frequency.

### *3.6.2 Keywords and keyness*

The term *keyword* (or *key word*) is widely and constantly used in linguistics and other disciplines; however different meanings are given to this term in different contexts and research traditions. Williams, who paved the way to a rich research tradition in the field of cultural analysis, describes keywords as “significant, binding words in certain activities and their interpretation; they are significant, indicative words in certain forms of thought” (Williams, 1976, p. 13). This is a general definition that can easily be understood and shared; but no indication is provided about how to choose keywords in the analysis of specific contexts, such as culture. Indeed, most linguists working in Williams’ research tradition have not felt the need to investigate possible scientific methods for the selection of cultural keywords.[32]

In corpus and computational linguistics, on the other hand, the notion of *keyword* includes the idea of statistical significance deriving from frequency comparisons. Corpus linguistics *keywords* are usually obtained by comparing the wordlist of the corpus under investigation with the wordlist of a suitable reference corpus; any word of the given corpus whose frequency is found to be outstanding with respect to the reference corpus is considered a keyword. As Baker (2006, p. 123) states, a keyword list “gives a measure of *saliency*, whereas a simple word list only provides *frequency*”.

As was the case with word lists, several statistical methods can be applied for comparing two corpora by (key)word frequency. The chi-square test and the log-likelihood test are frequently used for determining *keyness*, i.e. the degree of outstandingness, or salience, of the specific word in the target corpus. The Wmatrix interface, used in a pilot experiment to the current research, adopts the log-likelihood measure.[33] Keyness can be positive or negative: positive keywords are words that are unusually frequent in the target corpus, while negative keywords are unusually **in**frequent in comparison to the reference corpus.

The reference corpus is usually, but not necessarily, larger and more general than the other one (Hunston, 2002, p. 68).[34] As regards the composition of the reference corpus, Scott and Tribble (2006, p. 65) declare that

> “further research is needed before we can confidently offer a rule of thumb, if one exists. In any case the research purpose is fundamental: in our experience, even the use of a

[32] An exception is perhaps represented by Rigotti and Rocci (2002), who have developed a method for verifying whether selected words are cultural keywords.

[33] A detailed description of Wmatrix is provided in Chapter 5.

[34] Gledhill (1995, 1996) and Bianchi and Pazzaglia (2007), for example, compared different folders of the same corpus, corresponding to the different sections of research articles. Culpeper (2002) extracted the keywords characteristic of six characters of Shakespeare’s *Romeo and Juliet* by comparing the lines spoken by each character to the lines of the remaining five characters (taken together).

> clearly inappropriate reference corpus as in the case of the BNC for studying a Shakespeare play may well suggest useful items to chase up using the concordancer."

To avoid possible terminological confusion, in the current work, the term *keyword* is used when the computational methods described above are applied, while the terms *key word* and *node word* are preferred when a word is chosen according to other, non computational criteria and used as starting point for analysis or for the generation of concordances, respectively. Finally, the term *search word* is used when talking about information retrieval with search engines.

### *3.6.3 Concordancing*

Any word or keyword can be used as starting point (node word) for concordancing. Concordance lines are chunks of text that show the node word in context – hence the term KWIC (Key Word In Context) format. The length of concordance lines depends on the parameters set by the user. In a KWIC concordance, all the occurrences of the node word are displayed one under the other, with the key words vertically aligned and highlighted.[35]

If a corpus is lemmatised, a lemma can be made node word, and the concordancer will search for strings of text containing any of the words belonging to the given lemma. If the corpus is POS tagged, and the software offers specific query options, concordancing can take grammatical category into consideration or even start from a POS tag.

In the current research concordancing will be used at several stages and for different purposes: in the preparatory phases, for extracting sentences containing selected words from general Web corpora (see Chapter 5); and when manually coding wordlists, for seeing the context of each word (see Chapters 7 and 8).

Most software programs allow users to decide the way they want the concordance lines to be shown, in terms of number of words to be displayed, sorting criteria (e.g.: sort alphabetically by node word, by 1 left, and/or by 1 right), and even the presence of specific words in the co-text. KWIC display and a correct use of sorting options facilitate the qualitative analysis of concordance lines and the observation of repeated patterns.

Concordance lines are the typical starting point for the analysis of collocation, colligation, semantic preference and semantic prosody which are usually considered in corpus linguistics the four descriptive components of units of meaning (Sinclair, 2004). As Mahlberg (2007, p. 195)[36] puts it

> "From the level of collocation to semantic prosody the descriptive components of a lexical item become increasingly abstract and move from the fixed core of the item towards its boundaries. Collocation is a very concrete category and accounts for the actual repetition of words on the textual surface around the core. The component colligation introduces a level of abstraction with reference to grammatical categories. Semantic preferences interpret the context of the core in terms of shared semantic

[35] An interesting and precise description of KWIC concordance lines can be found in Tognini Bonelli (2001, 2004) and in Stubbs (2007b, p. 177).

[36] Description of the four levels of analysis as different levels of abstraction is not specific to Mahlberg; in fact, she is following Sinclair and Stubbs.

> features, and finally the semantic prosody accounts for attitudinal or pragmatic meanings."

KWIC displays have greatly changed our way of looking at texts: linguists have passed from linear reading of one text after the other, to non linear and focused access to several texts at once. Also, by looking at chunks of sentences, our attention is necessarily concentrated on the node word and its immediate surroundings, without distractions. On the other hand, 20- or even 50-word chunks can at times be too short to understand all of the semantic components of the given word. A typical case is when a word takes part in an anaphoric chain and its referent can only be understood by going back to the first element of the anaphoric chain; or, as we shall see later, when it comes to analysing semantic prosody. For this reason almost all concordancers allow the user to expand concordance lines to display full sentences, paragraphs or even texts.

### *3.6.4 Collocation, semantic preference and semantic prosody*

As Evert (2007) points out, the term collocation is used in linguistics to refer to various different textual features. In an attempt to make distinctions clearer, he distinguishes between 'lexical collocations' and 'empirical collocations'. Lexical collocations are a series of more or less transparently fixed expressions, ranging from well-known idiomatic expressions and set phrases (e.g. *a school of fish*), to multiword expressions (e.g. *credit card*), to multiword units with mobile elements (e.g. *as far as X is concerned*). The term 'empirical collocations', on the other hand, refers more generally to the fact that some words (collocates) tend to appear more frequently than others in the same linguistic environment, and the study of empirical collocations requires the use of statistical association measures (such as T-score or MI score) to quantify the attraction between co-occurring words.

Although Evert (2007) suggests that this mathematical meaning of 'association' should not be confused with psychological association, a psychological component seems to be present in collocations, alongside a textual and a statistical components (Partington, 1996, pp. 15-16). From a textual point of view, "collocation is the occurrence of two or more words within a short space of each other in a text" (Sinclair, 1991, p. 170). From a statistical point of view, it is "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey, 1991, pp. 6-7). Finally, from a psychological or associative perspective, collocation is the expectations (or 'expectancies', in Firthian terms) that native speakers have of encountering a given word in the same environment as another one (Leech, 1974). In a study on priming, Durrant and Doherty (2010) provide an interesting review of major issues in assessing the psychological reality of collocations, discuss a few studies which suggest that high frequency collocations are psychologically real, and describe two experiments whose results seem to confirm that high-frequency collocations are likely to have psychological reality, though the models currently used to represent priming may need further elaboration.

Semantic preference and semantic prosody are two separate phenomena, but the boundary between the two is not always clear-cut: they frequently appear together (Bednarek, 2008) and they are frequently discussed together. Indeed, they can both be considered as an extension of collocation (see for example Baker & McEnery, 2005; Bednarek, 2008).

Stubbs (2001, p. 65) defines semantic preference as "the relation, not between individual words, but between a lemma or word-form and a set of semantically related words", i.e. the tendency of a word to co-occur with words belonging to one or more specific semantic domains. It has been noticed that a word may have different semantic preferences depending on features such as context, genre, domain, but also literal or metaphorical use (Bednarek, 2008). Furthermore, like collocation, semantic preference varies when syntactic patterning (colligation) varies (see for example Partington, 2004 and his discussion of *sheer*). Finally, different word classes tend to have different semantic preferences (O'Halloran, 2007).

As already mentioned, semantic preference entails a greater level of abstraction than collocation, and the semantic categories are decided by the researcher after looking at the concordance data available, on the basis of his/her intuition of what is most suitable in the project at hand. For example, a series of concordance lines where the node word 'sports car' co-occurred with names of famous American actors could lead to identifying as suitable semantic preference ACTORS, MEN, or even AMERICANS. None of these is preferable to the others *a priori*; only the whole context and aim of the research project may lead to a suitable solution.

When the semantic categories adopted fall into evaluative categories (e.g. 'good/positive/healthy/legal' and 'bad/negative/unhealthy/illegal'), then we enter the realm of semantic prosody. Identifying evaluation in text is a problematic issue, since evaluation can be expressed in several ways. Some lexical items, such as words 'wonderful' or 'good' or 'bad', have an evident evaluative component. However, as Hunston (2004, p. 157) notices, "the group of lexical items that indicate evaluative meaning is large and open and does not lend itself to quantification". Despite this, semantic taggers, such as the USAS tagset, which will be used in the current work, show attempts to list evaluative words, and also phrases (e.g. 'a cut above', or 'below standard', 'hand on heart'). The USAS tagset (Archer, Wilson, & Rayson, 2002) includes a specific category for evaluation (A5), subdivided into 4 subcategories: 'A5.1 Evaluation: Good/bad', 'A5.2 Evaluation: True/False', 'A5.3 Evaluation: Accuracy', and 'A5.4 Evaluation: Authenticity'. Within each category, plus (+) or minus (-) signs indicate positive or negative polarity, respectively. Alongside lexis, lexical-grammatical sequences may be indicators of evaluation (e.g. 'there is something x about y'), as suggested by Hunston and Sinclair (2000). Furthermore, frequently words inherit the positive or negative aura of the collocates they co-occur with (see for example Sinclair, 1991 and his analysis of *set in*). Finally, words and phrases may acquire different evaluative meanings depending on context and "the reader assumptions about value" (Hunston, 2004, p. 158) – to make an easy example, word 'low' indicates positive assessment when collocating with inflation, and negative when next to salaries – but also genre and domain (Bednarek, 2008) – corpus analysis has shown, for example, that phrase 'responsibility for' acquires negative connotation in the news, since it always collocates with negative events such as bombings, explosions, or acts of terrorism, but neutral in business texts where it collocates with budgets, outcomes or decisions (Bednarek, 2008, p. 123). Examples of corpus methodologies which may be used to identify and analyse evaluative language in large collections of texts can be found in Hunston (2004; 2011). A specific area of research concerned with identifying and quantifying expressions of opinion in text is sentiment analysis. Sentiment analysis, a particular type of automatic content analysis focussing on semantic prosody, will be described in Chapter 4, since it frequently used in marketing research.

### 3.7 Some thorny issues

Not all linguists are in favour of corpus linguistics, and its detractors include very famous names such as Chomsky[37] and Widdowson. Chomsky's criticism to corpus linguistics traditionally revolved around the following two points: the use of texts as the primary source of linguistic information, and the finite nature of a corpus.[38] Indeed the corpus perspective, where the data and their frequency of use are key elements in linguistic description, strongly clashes with Chomsky's distinction between competence (I-language) and performance (E-language) and the former's prioritisation over the latter. Furthermore, Chomsky argued that the finite nature of any corpus, even the largest ones, cannot account for the infinite possibilities of language (Chomsky, 1962). Hence, in his view, introspection and not corpus data is the primary key to linguistic research.

Less radical, but nonetheless critical is Widdowson, who considers corpus linguistics as a 'development in E-language description' (Widdowson, 2000, p. 6). Though agreeing that corpus analysis reveals facts about the way language is used that are not directly accessible by intuition or surveys among speakers, Widdowson (2000) sees serious limitations in corpus linguistics connected to its inability to describe member categories (in ethonomethodological terms), to provide insight into the encoded possible and the contextually appropriate and to its showing decontextualised language.

Criticisms such as the ones above have been taken into serious consideration in corpus linguistics and, rather than defeating it, they have aided the development of this field of enquiry. As McEnery and Wilson (2001, p. 5) observe, "[c]oncepts […] such as balance and representativeness […] are a direct response to some of the criticisms Chomsky made." Similarly, awareness of the need to 'recreate' the socio-pragmatic context of corpus data has led to the development and use of tagging schemes which encode sociolinguistic information.[39]

Modern (as opposed to early) corpus linguists are aware of the limitations of corpora and of caveats in their use. The limitations of corpora are summarised by Hunston (2002, pp. 22-23) and are shortly listed and commented below.

First of all, corpora present language out of its context. The word context is to be interpreted here in many senses that range from social and pragmatic context, to visual and audio context. Despite several possibilities exists to include information about

[37] Chomsky's consideration of corpora, however, seem to have slightly changed in recent times (see Aarts, 2000).

[38] A clear review of Chomsky's criticisms to corpus linguistics can be found in McEnery and Wilson (2001, pp. 5-12); mention of the debate is also present in many papers and books about corpora, such as Leech (1992), and Tognini Bonelli (2001).

[39] See for example the following corpora: ICE-GB; The Wellington Corpus of Spoken New Zealand English; The Limerick corpus of Irish English; The Scottish Corpus of Texts and Speech (Xiao, 2008).

textual and contextual data into the corpus, this kind of annotation is time consuming and consequently relatively little used. Similarly, although some multimodal corpus analysis tools have recently been developed (see for example Baldry & Beltrami, 2005) their use is still extremely limited. Finally, several projects have addressed the issue of 'recontextualisation' by annotating important pieces of contextual information, but, to my knowledge, none of them has ever been able to fully provide all of the contextual elements (from the socio-pragmatic to the audio-video ones).

Second, any corpus is a limited sample of language and can only show its own contents. Therefore the linguist must be very careful at making generalisations from a single corpus, as "conclusions about language drawn from a corpus have to be treated as deductions, not as facts" (Hunston, 2002, p. 23). This is particularly true when

> "evidence from a corpus is used to make statements about 'the way the world is' […]. For example, there are roughly twice as many instances of *left-handed* as *right-handed* in the Bank of English corpus. What is the reason for this? One possible explanation is that there are more left-handed people in the world than right-handed people, but we know that this is not so. Another explanation is that left-handed people are considered to have a higher status than right-handed people, and therefore to be more worth talking about. Most left-handers would argue that this does not accord with their daily experience. A third possibility is that right-handedness is considered to be 'the norm' and left-handedness is 'deviant', and that deviance is more often mentioned than normality. Looking at the lines themselves suggests that this is the most likely interpretation, but it is important to recognise that this is an interpretation of evidence, not 'fact'" (Hunston, 2002, p. 66).

Third, a corpus can only provide information about whether something is used or frequent, but not whether something is correct (from the point of view of 'standard grammar') or impossible. As both Chomsky (indirectly) and Widdowson (directly) noticed, we cannot say that something is not possible simply because it does not appear in a corpus.

Fourth, a corpus can offer linguistic evidence but not linguistic information. The corpus only lists several examples of language in use, or frequency counts, but making sense of them is left to the researcher. Indeed, a corpus does not automatically provide answers to linguistic questions. Analysis and intuition are always necessary to make sense of the data.

Awareness of these limitations is probably one of the reasons (though not the only one) that has led corpus linguists to working more and more on specialised corpora, and use general ones as term of comparison. Highly specialised corpora reduce the problem of decontextualisation. Furthermore, as we have see in Section 3.3.1.1, the more a corpus is specialised the smaller it can be, and a small corpus is easier and quicker to annotate. Finally, it has become frequent practice for corpus linguists to carry out the same type of analysis on several different corpora, or to compare corpus results to other types of empirical data or to a specific theory, before drawing generalised conclusions. In fact, when comparing different corpora or different types of empirical data, conclusions are drawn from the analysis of several different samples of the same population, rather than from one single sample. Furthermore, as we have seen in Chapter 2, interpreting corpus data within a clear theoretical framework may help formulate sound hypotheses or draw convincing conclusions, and at the same time prove (or disconfirm) a specific theory.

I agree with Baker in believing that any method of research has “associated problems which need to be addressed and [is] also limited in terms of what [it] can and can not achieve” (Baker, 2006, p. 7). Moreover, as suggested by Fillmore (1992) and others, there is no reason why theoretical linguistics could not go hand in hand with corpora, and various ‘types’ of linguists, including theoreticians, could not make use of corpus data.

The current chapter has outlined some features and key elements of corpus linguistics and has introduced the Web as a source for corpus data. In the next chapter will see how some of the concepts and methods of corpus linguistics recur in or compare to analytical methods in marketing research.

# Marketing research

## 4.1 Introduction

The current chapter provides an overview of the materials and methods most frequently used in marketing research, with particular reference to those connected with textual data, and reviews selected marketing and consumer studies where content analysis of data is performed. The studies have been selected because of their similarities with the materials and methods used in my preliminary experiments and/or in the final design of the work. Finally, Section 4.3 describes two preliminary experiments to the current work, outlines some theoretical and procedural features common to cultural studies, corpus linguistics and marketing research, and explains how these conflate into the current project.

Quoting from Hair, Bush, and Ortinau (2009, p. 4), marketing research is "the function that links an organisation to its market through the gathering of information".[1] This is a broad definition that encompasses several types of data gathering and analytical activities aimed at providing decision makers with information that might help them plan future action and interaction with the desired audience.[2]

As regards data gathering, data collection is carried out on very many different types of sources. A major important distinction is between primary information, i.e. "information specifically collected for a current research problem or opportunity" (*ibid.*, p. 37), and secondary information, i.e. "information previously collected for some other problem or issue" (*ibid.*, p. 37).

As regards analytical activities, a distinction can be made between qualitative and quantitative approaches and also between exploratory research, descriptive research, and causal research. Qualitative approaches, which involve a limited number of subjects (as few as 8-10 subjects), are fast and inexpensive, but their results can hardly be generalised. For this reason, they are typically adopted in exploratory studies. Quantitative studies, on the other hand, involve a large number of people and are usually performed by means of specifically-made multiple-choice questionnaires.

[1] *Marketing research* is not to be confused with *market research*, as the latter focuses specifically on the size and trends of a market and is one of the many faces of marketing research.

[2] As such, marketing research is neither good nor bad in itself. The use that decision makers make of the information gathered, though, may be targeted to gaining personal advantage (as in private business advertisements) or to higher and 'friendlier' goals, as is the case with ethical and social advertising campaigns.

Quantitative studies provide results which can be generalised and are generally performed in descriptive and causal research. Unfortunately, collecting primary data of this type requires careful planning and is highly expensive and time-consuming. In between the two extremes stand rare large scale qualitative studies, and a vast number of quantitative studies performed on a limited number of subjects (as few as 30) for exploratory purposes.

### *4.1.1 Data gathering in marketing research*

Primary information is based on elicited data gathered through a range of direct or indirect questioning techniques (Hair, Bush, & Ortinau, 2009). Direct techniques, such as in-depth interviews and focus groups, involve questioning a small number of subjects on a specific topic and provide the researcher with textual data which is typically analysed using qualitative techniques. The responses collected using direct questioning frequently portray rational and conscious thoughts, as well as socially desirable attitudes (*ibid.*). Indirect techniques, also called projective techniques, include free word association, picture tests, sentence completion tests, and role-playing. Projective techniques – originally developed in the field of psychology – offer a view of the respondent's true opinions and beliefs more neatly than direct ones, and are usually adopted in qualitative studies (Donoghue, 2000). Among the projective techniques used in marketing research, two seem to be of particular relevance in the current research: free word association, and sentence completion tests. Free word association – i.e. "a projective technique in which the subject is presented with a list of words or short phrases, one at a time, and asked to respond with the first thoughts or word that comes to mind" (Hair, Bush, and Ortinau, 2009, p. 185) is among the 10 most common methods used to investigate consumers' needs (van Kleef, Trijp, & Luning, 2005) and has been used, for example, to assess consumers' perception of products (see Guerrero, Claret, Verbeke *et al.*, 2010; Roininen, Arvola, & Lähteenmäki, 2006; Ares, Giménez, & Gámbaro, 2008; Ares & Deliza, 2010) and to assess the cognitive structure of bilingual consumers (Luna & Peracchio, 2002). Interestingly, this is also the technique that anthropologists Szalay and Maday (1973) adopted to study subjective culture, and that Fleischer (2002) used to assess the level of conventionalisation of the image of drinks in Poland, France and Germany (see Chapter 2).

Sentence completion tests – tasks in which "the subjects are given a set of incomplete sentences and asked to complete them in their own words" (Hair, Bush, & Ortinau, 2009, p. 186) have been used, for example, by Belk (1985) to explore the role of materialism in purchase and consumption experiences. In the field of linguistic and cultural studies, this technique was applied by Wilson and Mudraya (2006) to analyse the relationship that exists between naming of shoes and the establishing of different types of shoes as cultural symbols in Russia, and by Potash, de Fileo Crespo, Patel, and Ceravolo (1990) to compare American and Brazilian college students as regards their orientations about future, achievement motives and work ethics, interracial tolerance, and sexuality (see Chapter 2).

Secondary information, i.e. information not specifically collected for the study at hand, includes: customer-volunteered information from electronic customer councils, customer usability labs, e-mail comments, chat sessions, and the like; "data collected by the individual company for accounting purposes or marketing activity reports" (Hair, Bush & Ortinau, 2009, pp. 114-115); or data collected by outside agencies, associations or periodicals. Growing emphasis has recently been put on secondary data, partly as a consequence of the development of the Internet (*ibid.*, 2009, p. 37), and Internet work seems to be gradually replacing field work. Furthermore, secondary information is not as costly as primary information.

### *4.1.2 Research design in marketing research*

Marketing research can be divided into three types: exploratory research; descriptive research; and causal research (Hair, Bush & Ortinau, 2009). Exploratory research aims to outline problems, clarify concepts, collect information, eliminate impractical ideas, and formulate hypotheses. At this stage of research, the researcher can use flexible research designs and methods, and it is customary to resort to convenience sampling, given that the researcher's interest is getting an inexpensive approximation to a specific topic (Guerrero, Claret, Verbeke *et al.*, 2010).[3] Descriptive research is more rigid than exploratory research and requires careful data collection and study design. Descriptive studies can be longitudinal (diachronic) or cross-sectional (synchronic) and aim to describe specific elements of interest, such as the users of a product or service, its demand, the ways it is used, and to make predictions. Finally, causal research performs laboratory and field experiments in order to assess cause-effect relationships between variables (Hair, Bush & Ortinau, 2009).

As regards the analytical methods used in marketing research, these vary depending on the type of data and study. As Aggarwal, Vaidyanathan and Venkatesh (2009) point out, one of the first analytical methods applied to textual data in the marketing field was content analysis; since its first use in the late '70s, it has been adopted, for example, to analyse advertisements, to determine the knowledge structure of salespeople, and to understand communication on home shopping networks. In marketing research, content analysis seems to be preferably applied in exploratory rather than descriptive studies, possibly due to the problems associated with sampling and measurement, or the reliability and validity of content categories, as well as the prohibitive cost of manually coding large amount of data.

Content analysis has been variously defined. Neuendorf (2002, p. 10) lists some of the definitions offered by 'main players in the development of quantitative message analysis'; the elements common to all those definitions suggest to describe content analysis as a quantitative analysis of textual messages (of any type) by means of systematic and replicable measures. As Weber (1990, p. 12) clarifies, "a central idea in content analysis is that the many words of the text are classified into much fewer content categories". Content analysis categories can be decided *a priori*, or while analysing the data. In either case, finalising the coding scheme requires several review steps that go hand in hand with application of the coding scheme to different sets of data by different coders. As we shall see in Section 4.2, although definition of the coding scheme before looking at the data is strongly advocated by content analysis

[3] According to Graveter and Forzano (2008, cited in Guerrero, Claret, Verbeke *et al.*, 2010), convenience sampling is probably used more often than any other kind of sampling in behavioural science research.

guidebooks, such as Weber (1990) and Neuendorf (2002), establishing categories while looking at the data seems to be the preferred option by most researchers in the marketing field. Finally, as Weber (1990) and Neuendorf (2002) clarify, central issues in content analysis are reliability and validity of the classification procedure. Reliability is guaranteed by a consistent application of the coding scheme. When coding is performed manually, different coders should be able to code the same text in the same way: this can be attained by creating and using of a specific codebook which describes the coding categories and explicates how the codes should be interpreted. Automatic coding, on the other hand, requires the use of specific software based on dictionaries and can lead to a higher degrees of consistency. Validity refers to the extent to which the categories adopted in the analysis represent or measure the concept that the researcher is interested in. It is interesting to notice at this point that automatic content analysis is largely applied in other disciplines that are linked to culture, such as the social sciences (see for e.g. McTavish & Pirro, 1990), or linguistics and cultural studies (e.g. Wilson & Moudraia, 2006).

A type of automatic content analysis which has recently been undergoing significant development and is finding application in marketing research is sentiment analysis, or opinion mining,[4] in the form of assessment of positive, negative or neutral sentiment in text. Sentiment analysis has gained momentum with the development of the Web 2.0 – rich in opinionated text types, such as blogs, review portals and other user-generated contents – and by the application of computerized text mining, information retrieval and natural language processing procedures to secondary data available on the Web (Tsytsarau & Palpanas, 2012). Sentiment analysis, which is ultimately performed electronically, often starts with manual analysis of small text samples which are then used to train the specific software (*ibid.*, p. 484). Indeed, determining sentiment polarity is a highly context-sensitive task (Choi, Kim, & Myaeng, 2009). Sentiment analysis may be applied at document, sentence, clause, or even word/phrase level depending on the type of text and the research goals (Thet, Na, Khoo, & Shakthikumar, 2009). Since consumers are largely relying on on-line opinions when making their purchasing decisions (Kaiser, Schlick, & Bodendorf, 2011; Archak, Ghose, & Ipeirotis, 2011), user-generated product reviews (and sometimes also blogs, weblogs and message boards) are analysed in order to understand the standing of a given product on the Web (see for example the articles listed in the Literature Survey section in Jebaseeli & Kirubakaran, 2012). This, however, is not the only possible application of sentiment analysis in marketing. Other applications include, for example, deriving the pricing power of a product feature (Archak, Ghose, & Ipeirotis, 2011) and warning marketing managers about the rising of critical situations (Kaiser, Schlick, & Bodendorf, 2011). Sentiment analysis is worth mentioning in the current work because it analyses positive/negative polarity of text with a logic which is similar to that of semantic prosody (see Chapter 3, Section 3.6.4), and also because of its use of Web text. However, the tools and methods

[4] The two terms are generally used interchangeably, although they originated in different communities and consequently have slightly different notions. As Tsytsarau & Palpanas (2012) explain: "Opinion Mining originates from the IR [Information Retrieval] community, and aims at extracting and further processing users' opinions about products, movies, or other entities. Sentiment Analysis, on the other hand, was initially formulated as the NLP [Natural Language Processing] task of retrieval of sentiments expressed in texts. Nevertheless, these two problems are similar in their essence [...]."

adopted in sentiment analysis and the applications that have been made of it go beyond the scope of the current work. For this reason, no specific sentiment analysis paper will be reviewed in the next section.

Finally, in marketing studies, as in many other scientific fields, it is common practice to validate the results of qualitative studies using quantitative data, and *vice versa*, or different analytical techniques. Yu, Shen, Kelly and Hunter (2006) validated the results obtained from a questionnaire-based quantitative study on 51 subjects by comparing them to the findings of a focus-group meeting. Guerrero, Claret, Verbeke *et al.* (2010) checked the robustness of results obtained with the semantic analysis of free word association tasks by comparing them to the findings of focus-group discussion.

The following section reviews a few marketing and consumer studies where content analysis of data is performed. The studies have been selected because of their similarities with the materials and methods used in my preliminary experiments and/or in the final design of the work.

## 4.2 Review of selected marketing studies

Content analysis techniques are typically applied to elicited data and used in small-scale studies.

Ares, Giménez and Gámbaro (2008) analysed free word associations of the images of five types of natural yogurt. Fifty Uruguayan subjects were asked to evaluate the images and write down the first thoughts that came to their minds. The associations thus elicited were semantically analysed: for each type of yogurt, terms with similar meaning were manually grouped into categories and the categories shared by less than 10% of the participants were discarded; next, the categories observed for the different yogurts were further classified into 19 final categories. For each category, word frequencies were counted and used to compare the different types of yogurt to each other. The results showed that regular yogurt was considered a healthy product having pleasant texture and flavour; low-calorie yogurts were mainly associated with diet or slimming and with texture or other type of sensory defects; finally, yogurts enriched with fibre and antioxidants were mainly related to health, and the prevention of diseases. As the authors declare (*ibid.*, p. 641), “word association thus provided an interest insight into consumers’ perception of yogurts, which could be useful for product development and marketing”.

Codern, Pla, de Ormijana, and Gonzales (2010) employed content analysis to identify the dimensions that lay people and healthcare professionals use to assess the risk of smoking. To this purpose they carried out focus-group interviews with 11 users and 7 professionals. The focus-group discussions were transcribed and manually coded by the researchers. The coding system was developed in subsequent stages: full reading of the transcripts and identification of the recurring topics; review of the codes; and code categorisation into groups. More concretely,

> “two researchers individually generated codes and categories that were then contrasted in search for differences and commonalities. A third researcher followed the process, reading the transcriptions and verifying the codes and their meanings. The three

> researchers involved in the analytical process met regularly to discuss emerging issues.” (*ibid*. 2010, p. 1565)

This is a perfect description of the steps and processes used in the current work to create the initial version of the coding scheme. In some preliminary experiments, two separate coders went through texts about chocolate in English and in Italian and identified the recurring semantic fields; the categories thus separately established were then compared and contrasted in order to create a single list of codes which was reviewed by a third coder (myself). The three coders met frequently to discuss coding issues.[5]

Guerrero, Claret, Verbeke *et al.* (2010) used free word associations of the node word *traditional* to assess the perception of traditional food products in six European regions (Flanders in Belgium, Burgundy in France, Lazio in Italy, the counties of Akershus and Østfold in Norway, Mazovia in Poland and Catalonia in Spain). About 120 subjects in each region were individually asked to name three words in response to the verbal stimulus word *traditional*, while concentrating on food-related issues. For each region, gender and age group, frequencies of elicitation were obtained at three different levels: first, at the level of the words elicited; second, by classifying the elicited words in 55 semantic classes; third, by grouping the 55 classes in ten principal dimensions. Analysis at the level of the 55 semantic categories showed a general tendency of southern European regions to associate the idea of *traditional* with broad concepts such as Heritage, Culture or History, while central and northern European regions tended to focus more on practical issues such as Convenience, Health and Appropriateness. Analysis at the level of the ten principle dimensions showed fewer differences between geographical regions, but highlighted gender differences: women seemed to prefer the Heritage, Health, Origin or Sensory dimensions, while men the Elaboration, Habit, Marketing and Variety ones. Finally, the authors compared results of the word association study to the results of focus group interviews, which confirmed their robustness. This study by Guerrero *et al.* is similar to my experiments in its using a double level of semantic analysis (a wider number of semantic categories, subsequently grouped into a smaller number of broader semantic domains).

The studies reported above applied content analysis to free word associations. However, content analysis is frequently applied to focus groups transcripts and open-ended questions. For example, Brug, Debie, van Assema and Weijts (1995) carried out an explorative study on people’s motivation in consuming fruit and vegetables. Data were collected in focus group interviews. The focus group transcripts were analysed by dividing the sentences into groups depending on their content, each group representing a specific issue which emerged during the discussions. Finally, the issues thus identified and grouped were used to prepare summaries of the focus group meetings.

More interesting is a study on critical success factors in construction project briefing by Yu, Shen, Kelly and Hunter (2006). The authors submitted a questionnaire to 51 experienced construction practitioners. Alongside background information, the questionnaire included an open-ended question aimed to collect opinions on the

[5] A detailed description of the genesis of the coding scheme, along with it subsequent revisions is included in the Appendix.

success factors of project briefing. The open-ended question responses were analysed by assigning responses to coded categories. Through this procedure, 37 critical success factors were identified and coded; the critical success factors were subsequently grouped into five major categories adapted from a careful study of the scientific literature on the topic. Finally, the results of the open-ended responses were compared to the results of a focus group meeting, which confirmed their validity.

In very recent times, however, some researchers seem to be experimenting the application of manual or automatic content analysis on non-elicited data and on what could be considered secondary information.

Aggarwal, Vaidyanathan and Venkatesh (2009), used lexical analysis of the semantic Web to assess the positioning of different brands relative to that of competitors.[6] By means of Google API, they searched the Web for sentences containing specific brand names (e.g., “Stetson”) and analysed their co-occurrence with selected adjectives and descriptors (e.g., “up-to-date”), considering such co-occurrence an indication of subjectivity, i.e. a subjective evaluation or opinion. Significant co-occurrence was established by means of the mutual information score. Finally, frequency of co-occurrence (over a vast amount of textual data) was used to infer brand’s positioning. This study is highly interesting to the current work primarily because of its use of the Web as source of data. Second, because it considers frequency of occurrence of semantic associations as an indication of shared opinion among several subjects.

Finally, a study by Kleij and Munsters (2003) is worth mentioning here, though it does not apply content analysis procedures. The authors involved 165 subjects in the evaluation of different varieties of mayonnaise. The participants were asked to specify their preferences for each type on a 10-point liking scale. Furthermore, they were given the option to freely comment on their assessments. The words in the freely expressed comments were analysed in terms of word co-occurrences. Finally, word co-occurrences were counted for each different product, and the relationship between products and product characteristics as verbalised by the respondents were visualised be means of correspondence analysis. The results of the analyses were compared to preference mapping, a standard procedure in the analysis of sensory drivers of liking based on the use of objective data from trained panel assessment of product characteristics. The authors reported that “the agreement between the correspondence map and the preference map is striking, with the additional advantage being that the correspondence map is stated in terms of consumer language” (*ibid.*, p. 43). This study is relevant to the current research because of its using analytical methods typical of corpus linguistics: first, words are counted (by producing a frequency word list);[7] second, for each of the most frequent words (e.g. *taste*), co-occurrence (collocation) with other words is considered and discussed (e.g. *taste – sour*). Third, because these corpus linguistics analytical procedures are offered as an innovative methodological approach complementary to more traditional preference mapping.

[6] For a discussion of the Web as corpus, see Chapter 3.

[7] The authors of this study do not use this term, but clearly the words counts they mention correspond to a frequency word list.

## 4.3. Cultural studies, corpus linguistics, marketing research, and the current work: common features

As we have mentioned, the current work aims to assess the possibility of using materials and methods typical of corpus linguistics for an analysis of cultural associations of a given node word which could find theoretical or practical applications not only in the linguistic and cultural fields, but also in the marketing one. Such a type of analysis should bank on some common ground among the three fields. For the current purposes, I have identified the necessary common ground in the following features: word associations; semantic/content analysis; and frequency as a measure of the association's importance.

Word associations appear in their psychological dimension in free word association and sentence completion tasks, and in their linguistic dimension in text in general and collocations in particular. Indeed, some parallelism can be seen between empirical collocations and verbal associations or EMUs, to use Szalay and Maday's 1973 terminology, see Chapter 2).[8] Empirical collocations are words that co-occur in the same textual environment; and frequency of co-occurrence determines collocational strength. The collocates of a node word, once grouped into semantic fields or domains, show its semantic preference (Partington, 2004). Analogously, EMUs co-occur in the same psychological environment as the word that triggers them, and they all show high collocational strength to the node word. Classification of words/sentences into semantic/thematic categories is the basic principle of content analysis.

Finally, a higher frequency of one EMU over another could thus be an indication of a cultural (vs. an individual) origin of the EMU itself. This last observation may be better understood considering Fleischer's (1998) theory of culture, illustrated in Chapter 2, according to which the cut-off line between individual and cultural mental associations is frequency of appearance across different subjects belonging to the same cultural group. Furthermore, frequency of elicitation of words in free word association tasks has been related with the strength or importance of a concept in the consumers' minds (Guerrero, Colomer, Guàrdia, Xicola, & Clotet, 2000).

As regards the textual material to use, elicited data in the form of sentence completion tasks or free sentence writing – widely used source of intelligence in marketing research – seems to be accepted, though is not the preferred type of data, in corpus linguistics at least when it comes to analysing culture (see for e.g. Fleischer, 2002; and Wilson and Mudraya, 2006).[9] Certainly, they are in keeping with the definition of corpus I subscribed to in Chapter 3. On the other hand, the use of large

[8] Interestingly, some recent empirical research has shown "a direct predictive relationship between the statistics of word co-occurrence in text and the neural activation associated with thinking about word meanings" (Mitchell, Shinkareva, Carlson, Chang, Malave, Mason, & Just, 2008, p. 1191; Murphy, Baroni, & Poesio, 2009). These results suggest that a direct relation between co-occurrence of words in text and the mental lexicon may exist, though further research is needed in this field.

[9] The term 'elicited data' has been frequently frowned upon by corpus linguists, because it is connected to introspection, a practice that according to some "does not give evidence about usage. […] Actual usage plays a very minor role in one's consciousness of language and one would be recording largely ideas about language rather than facts of it" (Sinclair, 1991, p. 39). This, however, is not a generalised view (see for example Fillmore, 1992; and Nordquist, 2009).

general textual data is more common in corpus linguistics than in marketing. Finally, in both disciplines, the Web is a relatively recent, but promising source use textual data. Consequently, the current work will use data elicited through sentence completion and free sentence writing tasks as a sort of 'control' situation to which Web data can be safely compared.

### *4.3.1 Preliminary experiments*

A preliminary experiment in the analysis of EMUs using corpora of non-elicited data was attempted by Bianchi (2007). The study aimed to highlight EMUs to chocolate in contemporary Italian society and compare the analytical possibilities offered by general and specialised corpora in a task of this kind. Concordances were generated for the Italian words for chocolate in a specialised corpus about chocolate and in a general-purpose corpus. Each concordance line was manually classified in terms of semantic context of the node word, that is the main topic(s) mentioned in the relevant text segment. Classification was based on the lexical meaning of the co-text and was performed through a data-driven, open-coding system. Semantic contexts were then grouped into higher-order categories, which were called 'conceptual fields'. Comparison between the two corpora highlighted what appear to be long-existing and well-established EMUs for chocolate in Italian society. It also suggested the possibility of evolution in the psychological associations of chocolate from the 1980s to 2005. From a methodological perspective, the findings seemed to show that suitable data for cultural analysis can equally be retrieved from a very large general corpus, or a small-to-medium-sized specialised corpus, provided that they include a wide variety of texts by different authors, and that in cultural analysis, the major concern in corpus creation, along with text variety, seems to be time-coverage. In terms of analytical methods, the two levels of analysis used – conceptual fields (higher-level; less fine-grained) and semantic contexts (lower-level; more fine-grained) – were both highly, but differently useful: conceptual fields helped establishing that, despite their apparent differences, these corpora could be considered samples from the same population, and guided the researcher in making sense of results and in establishing some kind of ranking between groups of psychological associations; semantic contexts, on the other hand, was the level where the most interesting EMUs emerged.

Another preliminary experiment (Bianchi, 2010) investigated the suitability of different methodological approaches to automatic semantic tagging in the analysis of cultural traits as they emerge from subjective meaning reactions to given words (EMUs). A first goal of this study was to compare the potential of manual coding to automatic tagging. To this aim, two sets of data elicited from British native speakers were coded manually as well as with Wmatrix, an automatic semantic tagger (see Chapter 5), and for each set of data the results were compared at the level of conceptual domains (superordinate, broader categories) and of semantic fields (subordinate, more fine-grained categories). In order to compare manual and automatic tagging, a specific conversion scheme was developed and applied. At the level of conceptual domains, the conversion scheme was applied to the top 30 items in the semantic frequency list and in the semantic keyword list of the elicited data as offered by Wmatrix, excluding grammatical items. As an intermediate step between manual tagging (sentence-based) and semantic tagging (word-based), it was decided to consider also the top 30 items of the raw frequency list and of the keyword list, as this allowed manual tagging to be applied on the basis of individual words. Therefore, the top 30 semantic items in the lists (excluding the node word) were manually mapped to one or more of the conceptual domains used in manual coding. Those analyses were then compared to the results of manual coding of the whole elicited datasets, which showed that the semantic frequency list performed generally better than the other lists. In fact, it retrieved the same or a higher number of domains and systematically showed strong correlation values at the Spearman test. At the level of semantic fields, comparison was performed using the most frequent 50 items in the semantic frequency list and in the semantic keyword list. When using the semantic frequency lists, the data consistently showed levels of correlation in the modest range; when using the semantic keyword list, results were less consistent. Another goal of the study was to compare elicited data to Web data. At this level of analysis, comparison was performed using automatic tagging only. Consequently, 10,000 sentences were extracted from a general Web corpus for each of the node words of the elicited data and the Web datasets thus created were tagged with Wmatrix. The semantic word lists of the Web data were compared to the semantic word lists of the elicited data. For the sake of experimentation, correlation was computed in three different ways: (1) using the whole semantic frequency lists, (2) using the top 100 items in the lists; and (3) using the top 50 items. All the six cases (three for each node word) showed interesting positive correlation between the elicited and the Web data, the strength of the correlation decreasing from strong to medium to low-medium as the number of items considered decreased.

### *4.3.2 The current work*

Banking on the preliminary experiments described above and on the theoretical ideas reviewed in the previous chapters, the current work will use elicited data gathered through free sentence-completion and sentence-writing tests. The data elicited will be analysed following a content analysis procedure highly similar to that described in Codern, Pla, de Ormijana, and Gonzales (2010) and results will be discussed within the framework of cultural systems theories. Furthermore, the results obtained with elicited data will be compared to non-elicited data from a Web corpus. Indeed, if (freely available) Web corpora gave the same results as more traditional marketing research techniques, marketing research could benefit from a wider range of fast and inexpensive methods. Finally, an automatic semantic tagger will be tested on the elicited data, in order to assess the extent of its possible application in cultural analysis.

The materials and methods employed in the current research are described in detail in Chapter 5. The various analyses performed on the elicited data and on the Web data are reported in Chapters 6-10.

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 <u>expensive</u>* = PRICE; *Chocolate is good for your <u>health</u>* = 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
B - The Body & the Individual
C - Arts & Crafts
E - Emotional Actions, States & Processes
F - Food & Farming
G - Government & the Public Domain
H - Architecture, Building, Houses & the Home
I - Money & Commerce
K - Entertainment, Sports & Games
L - Life & Living Things
M - Movement, Location, Travel & Transport
N - Numbers & Measurement
O - Substances, Materials, Objects & Equipment
P - Education
Q - Linguistic Actions, States & Processes
S - Social Actions, States & Processes
T - Time
W - The World & Our Environment
X - Psychological Actions, States & Processes
Y - Science & Technology
Z - Names & Grammatical Words

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.

## Chapter 6

# Semantic associations of *chocolate* and *wine* in the Italian and English cultures

### 6.1. Introduction

The present chapter addresses R.Q.s 1 and 2 in my Research Question list (see Chapter 1 or Chapter 5), by highlighting the semantic associations of *chocolate*, and *wine* in the Italian and English cultures and comparing them. Following the widely used habit of analysing elicited data in fields such as the social sciences, marketing, and also linguistics, the present chapter makes use of elicited data, collected and semantically tagged as described in Chapter 5, Section 5.1.

As we have seen in previous chapters, elicited data have long been a primary source of intelligence for the analysis of personal and cultural thoughts and behaviours in the marketing field (see Chapter 4), as well as other social areas, and relatively recently also linguistics (see Chapter 2). Widely used methods for eliciting data are projective techniques, such as free-word association tasks and sentence writing or sentence completion tasks, which have proven to be useful in eliciting the affective element behind the concepts involved. The data are then analysed qualitatively and/or quantitatively. A possible analytical method is content analysis, i.e. classifying the many words or sentences in a text into a finite number of semantic categories.

Different scientific disciplines have shown interest in highlighting cultural mental associations by semantic analysis of elicited data (see Chapter 2). Still, a standard, common procedure does not seem to exist, since each study applies a different type of statistical analysis, even when they share data of a similar nature (see Chapter 2).

As suggested by Fleischer (1998; see Chapter 2), the cut-off line between individual and cultural mental associations is frequency of appearance across different subjects belonging to the same cultural group. In other words, the more a mental association is shared among a wide number of subjects in a given cultural group, the more that mental association is conventionalised in the given culture and can be considered a specific feature of the culture itself. More specifically, Fleischer (1998) postulates that symbols (concepts) are made up of three components: core field, i.e. a stable, highly conventionalised meaning; current field, a rather generalised, but not yet stabilised element, and connotational field, i.e. the expression of individual meaning. Both core and current field are expressions of cultural meanings. Furthermore, Nobis (1998, summarised in Chapter 2) suggests that conventionalisation grows with time, and that the more a concept is established within a culture, the more complex its behavioural patterns are.

Taking inspiration from the existing literature, the present chapter develops a computational method for highlighting cultural associations in a corpus of elicited sentences about a given node word, and systematically applies it to four different datasets, two in English and two in Italian. The node words under investigation are: *chocolate*, and *wine.*

Specific reasons led to choosing the node words. *Chocolate* and *wine* are concrete nouns with clear though varied referents in both Italy and the UK, but presumably having different cultural roles in the two countries. Indeed, Italy can boast one of the longest traditions in wine production in the world, and wine is a traditional national product as well as a major export good. On the other hand, Great Britain has never been a 'wine country' either in terms of production or consumption, although wine is currently largely imported and consumed in the country. As regards chocolate, both Italy and Great Britain can boast a solid tradition in chocolate making, with big national enterprises (e.g. Perugina, Talmone, Novi; Cadbury, Bendicks), as well as small local quality chocolate makers. None of the two countries, however, would probably consider chocolate as a traditional national product. Consequently, we would expect wine to be well-rooted in the Italian culture, but less so in the English one, while chocolate is expected to show similar levels of cultural rooting in both countries.[1]

Finally, according to Fleischer, rooting depends on the extension of the highly conventionalised elements (core field), compared to the less conventionalised one (current and connotational fields). According to Nobis, instead, it goes hand in hand with semantic complexity.

## 6.2 Chocolate

### *6.2.1 Inter-culture analysis*

#### *6.2.1.1 Semantic field analysis*

The Italian and English *chocolate* datasets were manually analysed in terms of semantic fields and conceptual domains. For detailed descriptions and discussions of the collecting and coding procedures, of respondents and of their answers, see Chapter 5.

The current section presents and discusses the results of the coding process.

Tables 6_1 and 6_2 list the semantic associations of *chocolate* as they emerged in the English and Italian datasets, in decreasing order of frequency. In the first column, the name of the semantic field is preceded by initials indicating the conceptual domain (e.g. F = Food; FET = Features; FE = Feelings & Emotions). The second and third columns report the Mean number of occurrences of the given field across respondents, and its Standard Deviation. The fourth column highlights the Rank of each field;

[1] Fleischer (2001) and Wilson and Mudraya (2006) use term 'anchored' (German: 'verankert') to refer to the strong connections that links a concept/event/word to a specific type of culture. My preference for the terms 'rooted'/'rooting' is due to its metaphorical implications: the roots of a tree go deep down into the earth in several directions, not only anchoring the tree into the soil, but creating a sort of underground network that at some point intertwines with the roots of other trees.

ranking is based on mean values. The remaining two columns will be presented and discussed later.

As Tables 6_1 and 6_2 show, the English *chocolate* dataset includes 88 fields out of the 95 in the Codebook, while the Italian *chocolate* dataset (Table 6_2) includes 86 fields out of 95.[2]

| Field | Mean | SD | Rank | G2,1 | Cnv | Field | Mean | SD | Rank | G2,1 | Cnv |
|---|---|---|---|---|---|---|---|---|---|---|---|
| F-product/shape | 2.11 | 2.44 | 1 | 0.57 | H | LD-drugs & addiction | 0.15 | 0.42 | 34 | 0.78 | M |
| FET-quality/type | 2.02 | 1.75 | 2 | 0.69 | H | I-fantasy/magic | 0.13 | 0.45 | 35 | 0.67 | H |
| FET-taste/smell | 1.44 | 1.72 | 3 | 0.56 | H | E-economy | 0.11 | 0.58 | 36 | 0.30 | H |
| FE-happiness | 1.33 | 1.34 | 4 | 0.71 | M | E-fair trade | 0.11 | 0.44 | 36 | 0.67 | H |
| F-food | 1.32 | 1.32 | 5 | 0.63 | H | P-family | 0.11 | 0.44 | 36 | 0.67 | H |
| FE-desire | 1.22 | 1.10 | 6 | 0.70 | H | E-religion | 0.10 | 0.34 | 37 | 0.79 | L |
| H-body | 1.09 | 1.14 | 7 | 0.66 | H | FE-seduction | 0.10 | 0.31 | 37 | 1.00 | L |
| H-health | 0.94 | 1.09 | 8 | 0.66 | H | FE-guilt | 0.10 | 0.43 | 37 | 0.67 | H |
| G-geo locations | 0.92 | 1.42 | 9 | 0.59 | H | FE-memory | 0.09 | 0.36 | 38 | 0.76 | M |
| E-event | 0.90 | 1.08 | 10 | 0.67 | H | FE-peace | 0.09 | 0.42 | 38 | 0.67 | H |
| F-composition | 0.78 | 0.89 | 11 | 0.72 | M | P-friendship | 0.09 | 0.33 | 38 | 0.78 | M |
| FET-quantity | 0.64 | 0.91 | 12 | 0.70 | H | G-spreading | 0.09 | 0.33 | 38 | 0.78 | M |
| F-bakery/cooking | 0.62 | 0.99 | 13 | 0.68 | H | COM-comparison | 0.08 | 0.27 | 39 | 1.00 | L |
| E-transaction | 0.62 | 1.05 | 13 | 0.60 | H | L-existence | 0.08 | 0.38 | 39 | 0.61 | H |
| P-women | 0.59 | 1.01 | 14 | 0.65 | H | E-work | 0.06 | 0.28 | 40 | 0.75 | M |
| F-manufacturing | 0.56 | 0.90 | 15 | 0.68 | H | LD-theft | 0.06 | 0.28 | 40 | 0.75 | M |
| FE-passion | 0.55 | 0.89 | 16 | 0.66 | H | E-law | 0.05 | 0.26 | 41 | 0.73 | M |
| F-drink | 0.54 | 0.70 | 17 | 0.75 | M | FE-no reaction | 0.05 | 0.21 | 41 | 1.00 | L |
| F-recipe | 0.53 | 0.86 | 18 | 0.67 | H | FE-bribing | 0.05 | 0.21 | 41 | 1.00 | L |
| CUL-artistic production | 0.53 | 0.85 | 18 | 0.67 | H | I-dream | 0.05 | 0.26 | 41 | 0.73 | M |
| P-children | 0.48 | 0.97 | 19 | 0.61 | H | EN-house | 0.05 | 0.21 | 41 | 1.00 | L |
| E-time | 0.40 | 0.72 | 20 | 0.72 | M | EN-tech | 0.05 | 0.26 | 41 | 0.73 | M |
| C-gift | 0.40 | 0.64 | 20 | 0.78 | M | E-language | 0.03 | 0.24 | 42 | 0.71 | M |
| H-medicine | 0.39 | 0.62 | 21 | 0.78 | M | P-royalty | 0.03 | 0.18 | 42 | 1.00 | L |
| P-men | 0.31 | 1.21 | 22 | 0.54 | H | F-storage | 0.02 | 0.15 | 43 | 1.00 | L |
| FET-price | 0.31 | 0.65 | 22 | 0.65 | H | H-dieting | 0.02 | 0.15 | 43 | 1.00 | L |
| FET-colour | 0.30 | 0.70 | 23 | 0.80 | L | E-war | 0.02 | 0.15 | 43 | 1.00 | L |
| FE-nice/pleasant/pleasure | 0.26 | 0.58 | 24 | 0.67 | H | E-history | 0.02 | 0.21 | 43 | NC | |
| FE-sex | 0.26 | 0.64 | 24 | 0.75 | M | E-holidays | 0.02 | 0.15 | 43 | 1.00 | L |
| FE-unpleasant | 0.25 | 0.65 | 25 | 0.61 | H | E-driving | 0.02 | 0.21 | 43 | NC | |
| FET-sweet | 0.25 | 0.53 | 25 | 0.72 | M | P-gay | 0.02 | 0.15 | 43 | 1.00 | L |
| FET-energy | 0.24 | 0.57 | 26 | 0.71 | M | LD-hiding | 0.02 | 0.15 | 43 | 1.00 | L |
| FE-comfort | 0.23 | 0.54 | 27 | 0.71 | H | EN-nature | 0.02 | 0.15 | 43 | 1.00 | L |
| H-beauty | 0.22 | 0.72 | 28 | 0.63 | H | CUL-culture | 0.02 | 0.15 | 43 | 1.00 | L |
| FE-mood | 0.22 | 0.52 | 28 | 0.77 | M | L-future | 0.02 | 0.15 | 43 | 1.00 | L |
| P-sharing/society | 0.21 | 0.51 | 29 | 0.77 | M | E-playing | 0.01 | 0.11 | 44 | NC | |
| EN-dirt | 0.21 | 0.59 | 29 | 0.62 | H | FE-surprise | 0.01 | 0.11 | 44 | NC | |
| EN-animals | 0.20 | 0.48 | 30 | 0.78 | M | FE-loneliness | 0.01 | 0.11 | 44 | NC | |
| P-people | 0.18 | 0.47 | 31 | 0.77 | M | FE-freedom | 0.01 | 0.11 | 44 | NC | |
| FE-relax | 0.17 | 0.44 | 32 | 0.79 | M | FE-persuasion | 0.01 | 0.11 | 44 | NC | |
| FET-packaging | 0.17 | 0.41 | 32 | 0.83 | L | P-posh | 0.01 | 0.11 | 44 | NC | |
| FE-senses | 0.16 | 0.50 | 33 | 0.70 | H | C-ceremonies | 0.01 | 0.11 | 44 | NC | |
| FE-love | 0.16 | 0.45 | 33 | 0.68 | H | C-party | 0.01 | 0.11 | 44 | NC | |
| FET-physical properties | 0.16 | 0.50 | 33 | 0.70 | H | S-sports | 0.01 | 0.11 | 44 | NC | |

Table 6_1: Semantic associations of *chocolate* for the English

[2] Semantic field ASSESSMENT is not included in this count, given its peculiarities. In the current chapter, as well as in the following ones, this semantic field will be discussed in the dedicated sections.

| Field | Mean | SD | Rank | G2,1 | Cnv | Field | Mean | SD | Rank | G2,1 | Cnv |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FET-quality/type | 2.78 | 1.68 | 1 | 0.68 | H | L-existence | 0.17 | 0.42 | 31 | 0.81 | M |
| F-bakery/cooking | 2.03 | 1.41 | 2 | 0.71 | H | P-age | 0.16 | 0.37 | 32 | 1.00 | L |
| F-product/shape | 1.68 | 1.59 | 3 | 0.65 | H | FET-genuine | 0.16 | 0.45 | 32 | 0.77 | M |
| F-recipe | 1.68 | 1.88 | 3 | 0.62 | H | F-storage | 0.14 | 0.40 | 33 | 0.79 | M |
| F-food | 1.51 | 1.45 | 4 | 0.64 | H | FE-senses | 0.14 | 0.40 | 33 | 0.79 | M |
| FET-taste/smell | 1.51 | 1.23 | 4 | 0.71 | H | P-women | 0.14 | 0.40 | 33 | 0.79 | M |
| G-geo locations | 1.32 | 1.58 | 5 | 0.66 | H | EN-house | 0.14 | 0.40 | 33 | 0.79 | M |
| COM-comparison | 0.94 | 1.19 | 6 | 0.60 | H | P-friendship | 0.13 | 0.38 | 34 | 0.78 | M |
| H-health | 0.94 | 0.93 | 6 | 0.74 | M | S-sports | 0.13 | 0.34 | 34 | 1.00 | L |
| FE-nice/pleasant/pleasure | 0.92 | 1.10 | 7 | 0.63 | H | E-playing | 0.11 | 0.36 | 35 | 0.77 | M |
| FET-quantity | 0.92 | 0.99 | 7 | 0.68 | H | FE-love | 0.11 | 0.32 | 35 | 1.00 | L |
| P-children | 0.90 | 0.87 | 8 | 0.73 | M | FE-guilt | 0.11 | 0.44 | 35 | 0.61 | H |
| H-medicine | 0.89 | 0.97 | 9 | 0.70 | H | FE-peace | 0.11 | 0.32 | 35 | 1.00 | L |
| FE-passion | 0.87 | 1.01 | 10 | 0.64 | H | FE-loneliness | 0.11 | 0.36 | 35 | 0.77 | M |
| CUL-artistic production | 0.86 | 0.96 | 11 | 0.70 | H | I-dream | 0.11 | 0.48 | 35 | 0.68 | H |
| E-event | 0.83 | 0.81 | 12 | 0.71 | H | FE-unpleasant | 0.10 | 0.30 | 36 | 1.00 | L |
| FE-desire | 0.79 | 0.90 | 13 | 0.71 | H | EN-nature | 0.10 | 0.35 | 36 | 0.76 | M |
| F-composition | 0.71 | 0.99 | 14 | 0.67 | H | CUL-culture | 0.08 | 0.27 | 37 | 1.00 | L |
| H-body | 0.71 | 0.79 | 14 | 0.72 | H | E-religion | 0.06 | 0.25 | 38 | 1.00 | L |
| H-beauty | 0.63 | 0.77 | 15 | 0.78 | M | P-men | 0.06 | 0.30 | 38 | 0.73 | M |
| FE-mood | 0.63 | 0.83 | 15 | 0.73 | M | I-fantasy/magic | 0.06 | 0.25 | 38 | 1.00 | L |
| FE-happiness | 0.57 | 0.76 | 16 | 0.71 | H | E-language | 0.05 | 0.21 | 39 | 1.00 | L |
| F-manufacturing | 0.54 | 0.71 | 17 | 0.74 | M | FE-surprise | 0.05 | 0.21 | 39 | 1.00 | L |
| C-gift | 0.48 | 0.86 | 18 | 0.65 | H | P-sharing/society | 0.05 | 0.21 | 39 | 1.00 | L |
| E-transaction | 0.46 | 0.86 | 19 | 0.65 | H | C-party | 0.05 | 0.21 | 39 | 1.00 | L |
| P-family | 0.46 | 0.76 | 19 | 0.75 | M | EN-tech | 0.05 | 0.21 | 39 | 1.00 | L |
| FET-energy | 0.43 | 0.69 | 20 | 0.73 | H | E-war | 0.03 | 0.25 | 40 | NC | |
| F-drink | 0.41 | 0.64 | 21 | 0.78 | M | FE-memory | 0.03 | 0.18 | 40 | 1.00 | L |
| H-dieting | 0.40 | 0.73 | 22 | 0.69 | H | FE-bribing | 0.03 | 0.18 | 40 | 1.00 | L |
| FET-physical properties | 0.40 | 0.68 | 22 | 0.63 | H | G-spreading | 0.03 | 0.18 | 40 | 1.00 | L |
| FET-colour | 0.38 | 0.61 | 23 | 0.78 | M | LD-theft | 0.03 | 0.25 | 40 | NC | |
| FE-comfort | 0.35 | 0.63 | 24 | 0.78 | M | EN-animals | 0.03 | 0.18 | 40 | 1.00 | L |
| E-history | 0.33 | 0.54 | 25 | 0.81 | M | E-economy | 0.02 | 0.13 | 41 | NC | |
| E-time | 0.33 | 0.60 | 25 | 0.78 | M | E-fair trade | 0.02 | 0.13 | 41 | NC | |
| LD-drugs & addiction | 0.33 | 0.67 | 25 | 0.72 | H | FE-competitiveness | 0.02 | 0.13 | 41 | NC | |
| FE-seduction | 0.30 | 0.59 | 26 | 0.70 | H | FE-freedom | 0.02 | 0.13 | 41 | NC | |
| FE-sex | 0.27 | 1.31 | 27 | 0.20 | H | FE-persuasion | 0.02 | 0.13 | 41 | NC | |
| FET-sweet | 0.27 | 0.60 | 27 | 0.70 | H | P-gay | 0.02 | 0.13 | 41 | NC | |
| FE-no reaction | 0.25 | 0.69 | 28 | 0.64 | H | P-royalty | 0.02 | 0.13 | 41 | NC | |
| CUL-studying/intellect | 0.22 | 0.55 | 29 | 0.68 | H | LD-hiding | 0.02 | 0.13 | 41 | NC | |
| FE-relax | 0.21 | 0.51 | 30 | 0.67 | H | C-ceremonies | 0.02 | 0.13 | 41 | NC | |
| P-people | 0.21 | 0.45 | 30 | 0.82 | L | L-future | 0.02 | 0.13 | 41 | NC | |
| EN-dirt | 0.17 | 0.46 | 31 | 0.77 | M | FET-packaging | 0.02 | 0.13 | 41 | NC | |

Table 6_2. Semantic associations of *chocolate* for the Italians

In both cases, the missing fields include the fields CONFIDENCE; SERVING; and EXCESSIVE DRINKING, which is no surprise given that these are fields that were added to the code list while analysing the *wine* datasets, i.e. after analysing the *chocolate* datasets. The remaining fields which are not attested are: STUDYING/INTELLECT; AGE; GENUINE; COMPETITIVENESS; CONFIDENCE; SERVING; and EXCESSIVE DRINKING for English; and PRICE; WORK; LAW; HOLIDAYS; DRIVING; and POSH for Italian. These fields had all entered the coding scheme in the preliminary phases to the current work, after coding two specialized corpora about Chocolate (one in Italian and one in English), and two general corpora in the same languages (see the Appendix).

As regards conceptual domains, both datasets present all the domains considered in the Codebook. Furthermore, in the case of English *chocolate*, no evident clotting of the same domain is visible in any part of the list (i.e. top ranks, as well as middle and bottom ranks are occupied by semantic fields from various domains), while an evident clotting of Food fields in the top 5 positions can be noticed for Italian.

The mean values considered so far, though interesting in so far as they provide a general picture of the semantic fields in each dataset, do not consider distribution of answers across subjects. However, distribution across subjects seems highly relevant, in order to gain a more precise picture of the cultural vs. individual mental semantic associations of the given node words.

Inspired by Wilson and Mudraya (2006), Molinari's evenness measure (G2;1) was applied to the data to assess the level of conventionalisation within each semantic field.[3] Evenness indexes are widely used in biology. Smith and Wilson (1996) introduce the concept of evenness to biologists as follows: “A basic feature of biological communities is the distribution of abundance among species. There are many aspects of this distribution that can be measured, but the simplest feature is evenness. A community in which each species present is equally abundant has high evenness; a community in which the species differ widely in abundance has low evenness”. In the current paper, each semantic field is considered as a ‘community’ or ‘area’ and each subject as a ‘species’ or ‘taxon’ which occurs (in that semantic field) a certain number of times, thus contributing to the composition of that community with a certain number of occurrences. In a comparative experiment on 15 different evenness indexes (Beisel, Usseglio-Polater, Bacmann & Moreteau, 2003), Molinari’s G2,1 resulted among the most sensitive to minor changes in abundance in dominant and median taxa and averagely sensitive to changes in rare taxa, which – within the context of our experiments – translates into highly sensitive to even minor differences in the number of occurrences of the given semantic field in each respondent’s answers. Molinari’s index is computed from raw counts.

As in Wilson and Mudraya (2006) and in Fleisher (2002), evenness values were then divided into three groups, corresponding to high (H), medium (M) and low (L) levels of conventionalisation. Level of conventionalisation is shown by the position of the evenness index with reference to the confidence interval: values that fall below confidence interval indicate a high level of conventionalisation; those falling above confidence interval show a low levels of conventionalisation. The 99% confidence intervals for the *chocolate* data were respectively 0.71-0.79 for English, and 0.73-0.82 for Italian. Please notice that, in this work, due to the presence of tables which are limited in space, all the values are reported rounded to the second decimal, but the analyses were performed using rounding to the fourth decimal, for greater precision. In a few cases, rounding to the second decimal may lead to apparent incongruity, as is the case with FE-MOOD and FET-ENERGY in the Italian *chocolate* dataset (Table 6_2) which seem to have the same evenness value (0.73), but different levels of conventionalization (M and H, respectively). However, such cases of apparent incoherency are very rare and are always explained by having had to round figures to the second decimal because of space limitations.

For an easier reading of results, the evenness values are reported next to mean and SD values in Tables 6_1 and 6_2, in column G2,1, accompanied by indication of their corresponding level of conventionalisation (column Cnv). In the evenness column, NC indicates that the evenness tool was not able to calculate a value for that field, because it contained less than 2 occurrences.

[3] Calculations were performed using an evenness calculator written by Ben Smith of Lund University, Sweden, and available at http://www.nateko.lu.se/personal/benjamin.smith/software. This highly user-friendly program computes 14 different evenness indexes, including Molinari’s.

In both *chocolate* tables, fields with a high level of conventionalisation tend to concentrate in the highest ranks, i.e. the fields with higher mean distributions across subjects seem to be the ones with higher levels of conventionalisation. Low levels of conventionalisation start appearing almost half way through the list and concentrate at the end of it. This seems to suggest that fields with NC values might be considered as having a low level of conventionalisation. However, as we shall see in Section 6.3, *wine* does not show such a neat relation between mean values and level of conventionalisation. Consequently, I shall here ignore all fields marked with NC, not knowing exactly how to assess them.

The distribution of fields across conventionalisation levels is summarised percentage-wise in Table 6_3.

| | **English** | **Italian** |
|---|---|---|
| High | 45.45 | 43.84 |
| Medium | 31.17 | 31.51 |
| Low | 23.38 | 24.66 |

Table 6_3. *Chocolate* – Percentage distribution of fields across conventionalisation levels

The two cultures show a rather similar distribution of fields across conventionalisation levels. The percentage of fields marked by a high level of conventionalisation is around 45% for English and 44% for Italian. Next comes a good 31% of fields with medium level of conventionalisation, while fields with a low conventionalisation are about 23-24%. In both cases, highly conventionalised fields cover slightly less than 50% of the total. If added to medium conventionalisation fields, however, the percentage of fields which – according to Fleischer – could be considered expressions of cultural meanings reaches about 75-76%.

#### *6.2.1.2 Conceptual domain analysis*

The following paragraphs apply the analytical steps described above to conceptual domains. Table 6_4 summarises Mean, SD, Rank, G2,1 and Conventionalisation values for the Italian and English *chocolate* datasets, at this higher level. All values have been computed from raw data, ignoring the existence of subdivisions (semantic fields) within each conceptual domain. However, in the table, the domain name is accompanied by the number of its conceptual fields, in parenthesis.

Both datasets show only three highly conventionalised domains: EVENTS, PEOPLE, and GEOGRAPHY, for the English culture; CULTURE, COMPARISON, and CEREMONY, for the Italian one. In neither case they appear among the most frequent ones in the dataset. A few domains have low levels of conventionalisation: CEREMONY, LOSS & DAMAGE and COMPARISON, in English; LIFE and SPORTS, in Italian. The remaining ones – 8 domains for English and 10 for Italian – have a medium level of conventionalisation. This is schematically summarised in Table 6_5 using percentage values.

| ENGLISH | | | | | | ITALIAN | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Domain** | **Mean** | **SD** | **Rank** | **G2,1** | **Cnv** | **Domain** | **Mean** | **SD** | **Rank** | **G2,1** | **Cnv** |
| Food (9) | 6.49 | 4.27 | 1 | 0.68 | M | Food (9) | 8.71 | 3.59 | 1 | 0.77 | M |
| Features (10) | 5.54 | 3.23 | 2 | 0.74 | M | Features (10) | 6.86 | 2.78 | 2 | 0.77 | M |
| Feelings & emotions (23) | 5.36 | 3.33 | 3 | 0.67 | M | Feelings & emotions (23) | 6.02 | 3.32 | 3 | 0.71 | M |
| Health & Body (5) | 2.67 | 1.94 | 4 | 0.69 | M | Health & Body (5) | 3.57 | 1.97 | 4 | 0.71 | M |
| Events (15) | 2.49 | 2.20 | 5 | 0.61 | H | Events (15) | 2.24 | 1.55 | 5 | 0.68 | M |
| People (11) | 2.05 | 2.62 | 6 | 0.56 | H | People (11) | 2.14 | 1.56 | 6 | 0.73 | M |
| Geo (2) | 1.01 | 1.46 | 7 | 0.58 | H | Geo (2) | 1.35 | 1.63 | 7 | 0.65 | M |
| Culture (3) | 0.55 | 0.86 | 8 | 0.67 | M | Culture (3) | 1.16 | 1.26 | 8 | 0.63 | H |
| Environment (5) | 0.52 | 0.94 | 9 | 0.67 | M | Comparison (1) | 0.94 | 1.19 | 9 | 0.60 | H |
| Ceremony (3) | 0.43 | 0.64 | 10 | 0.78 | L | Ceremony (3) | 0.54 | 0.96 | 10 | 0.60 | H |
| Loss & damage (3) | 0.23 | 0.52 | 11 | 0.78 | L | Environment (5) | 0.49 | 0.82 | 11 | 0.66 | M |
| Imagination (2) | 0.17 | 0.57 | 12 | 0.70 | M | Loss & damage (3) | 0.38 | 0.79 | 12 | 0.65 | M |
| Life (2) | 0.10 | 0.40 | 13 | 0.63 | M | Life (2) | 0.19 | 0.47 | 13 | 0.78 | L |
| Comparison (1) | 0.08 | 0.27 | 14 | 1.00 | L | Imagination (2) | 0.17 | 0.55 | 14 | 0.70 | M |
| Sports (1) | 0.01 | 0.11 | 15 | NC | | Sports (1) | 0.13 | 0.34 | 15 | 1.00 | L |

Table 6_4. *Chocolate* – Conceptual domains in the English and Italian datasets

| | **English** | **Italian** |
|---|---|---|
| High | 21.50 | 20.00 |
| Medium | 57.00 | 67.00 |
| Low | 21.50 | 13.00 |

Table 6_5. *Chocolate* – Percentage distribution of domains across conventionalisation levels

So, as was the case with semantic fields, the two cultures show a similar picture in terms of levels of conventionalisation, with 21.5% vs. 20% of highly conventionalised domains, 57% vs. 67% of domains with medium conventionalisation, and 21.5% vs. 13% of domains with low conventionalisation. It is interesting to notice, however, that, passing from semantic fields to conceptual domains, the picture within each culture has changed. A comparison between Table 6_3 and Table 6_5 shows a clear shift from dominance of highly conventionalised fields to dominance of domains with medium level of conventionalisation, in each culture. According to Fleischer's theory, however, this does not alter the already noted predominance of cultural associations of personal associations, as the former are indicated by high plus medium conventionalisation fields, which is these cases amount to about 78.5% for English and 87% for Italian.

From a methodological point of view, it is interesting to notice that no direct relationship exists between mean frequency and number of semantic fields composing the domain, or mean frequency and conventionalisation, or conventionalisation and number of fields in the domain.

### *6.2.2 Cross-cultural comparison*

So far we have established that *chocolate* is a concept with a reasonably high number of relatively well-established semantic associations in each of the two cultures, and similar percentage distributions of fields across conventionalisation levels. Furthermore we have seen that, although the two datasets share most of the given semantic fields, the latter do not seem to occupy the same ranks.

Consequently, comparison between the English and Italian datasets at the level of semantic fields could possibly tell us whether differences exist between the two cultures when thinking about *chocolate*, and where these differences lie.

First of all, cross-cultural comparison at the level of semantic fields was performed by applying Spearman's Rank Correlation Coefficient.[4] This is a non-parametric (i.e. distribution-free) test, appropriate to ordinal scales, which uses ranks of the x and y variables, rather than data (Fowler, Cohen, & Jarvis, 1998, pp. 138-141). Spearman's *r* "describes the overlap of the variance of ranks" (Arndt, Turvey, & Andreasen, 1999, p. 104). Spearman's test showed strong positive correlation,[5] with Spearman's Rho equal to 0.719 ($p < 0.01$), which suggests that differences between the two datasets do exist, but are rather limited.

In order to try and understand where the cultural differences lie, the datasets were compared using the Welch *t* Test for Independent Samples. T-Tests compare the mean scores of two groups on a given variable; Welch *t* Test for Independent Samples is a "modification of the T-Test for Independent Samples so that it does not assume equal population variances" and has been proven to outperform the ordinary T-Test in almost all circumstances (Fagerland, Sandvik, & Mowinckel, 2011). When comparing two samples with a T-Test, the Null Hypothesis (H0) is that the two samples have the same mean. The Null Hypothesis cannot be rejected when, for the given degrees of freedom (i.e. the total number of subjects -2) and the desired significant value (p), T is lower than the reference value provided in T-Test tables. If T is higher, the alternative hypothesis H1 has to be accepted.

T-Test results were significant ($p < 0.01$) for the semantic fields listed in Table 6_6. In the table, column one lists the name of the semantic field, preceded by initials indicating the corresponding conceptual domain; columns two-five provide the relevant values in the T-test; columns six-nine indicate the field's mean values and conventionalisation levels in the English and Italian datasets. For each field, bold highlights which of the two mean values is the highest.

Consequently, Table 6_6 lists the semantic fields for which the Italian and English population sample taking part in the survey quantitatively differed in their making reference to *chocolate*. But which of these differences are really due to culture and which depend on population sampling?

A look at the conventionalisation level of each semantic field – listed in the table in columns seven and nine, for English and Italian respectively – may help us find an answer to this question. It seems rather safe to state that when a field with a significant T value shows a high level of conventionalisation in one of the two cultures and also a mean value for that culture which is higher than that in the other culture, that difference in means can be taken to be of cultural origins. Consequently, for example, the WOMEN semantic field, which has High conventionalisation in English and Medium conventionalisation in Italian, along with a mean value which is higher in English that in Italian (0.59 vs. 0.14), can be considered a semantic association of

[4] Correlation was performed using SPSS.

[5] According to Fowler *et al.* (1998) the strength of a correlation is to be considered very weak when *r* ranges from 0.00 to 0.19, weak when *r* ranges from 0.20 to 0.39, modest for *r* between 0.40 and 0.69, strong for *r* ranging from 0.70 to 0.89, and very strong for *r* between 0.90 and 1.00.

chocolate typical of the English culture. Similarly, the QUALITY/TYPE semantic field, which shows high conventionalization in both cultures, and higher mean value in Italian (2.78 vs. 2.02), will be considered specific to the Italian culture with respect to chocolate. On the other hand, when a field with a significant higher mean in one culture shows a low level of conventionalisation in that culture, the result may depend on the sample, rather than the culture. An example is the AGE semantic field: despite its having higher mean value in Italian than English (0.16 vs. 0.00), it cannot be considered a semantic association of wine specific to the Italian culture, because its conventionalisation level in Italian is Low. Finally, semantic fields with a higher mean and medium conventionalisation level could possibly be considered culturally more frequent when in the other culture they have a high levels of conventionalisation or when they are virtually absent (NC). Nice examples are the BAKERY/COOKING and HISTORY semantic fields: they show higher mean values in Italian than in English, alongside Medium conventionalisation level in Italian vs. High and NC conventionalisation, respectively, in English. All other cases are uncertain, and need confirmation from other population samples.

| **Field** | **P** | **T** | **Df** | **st.error of df** | **Mean values English** | **Cnv** | **Mean values Italian** | **Cnv** |
|---|---|---|---|---|---|---|---|---|
| Comparison | 0.0000 | 5.6052 | 66 | 0.153 | 0.08 | L | **0.94** | M |
| F-bakery/cooking | 0.0000 | 6.8030 | 104 | 0.207 | 0.63 | H | **2.03** | **M** |
| F- storage | 0.0250 | 2.2866 | 75 | 0.052 | 0.02 | L | **0.33** | M |
| F-recipe | 0.0000 | 4.5354 | 80 | 0.254 | 0.53 | H | **1.68** | **M** |
| H- dieting | 0.0001 | 4.0011 | 65 | 0.093 | 0.02 | L | **0.40** | **H** |
| H- medicine | 0.0005 | 3.5873 | 97 | 1.139 | 0.39 | M | **0.89** | **H** |
| H- beauty | 0.0010 | 3.3603 | 128 | 0.124 | 0.22 | H | **0.63** | **M** |
| E-history | 0.0000 | 4.3301 | 76 | 0.072 | 0.02 | NC | **0.33** | **M** |
| FE-nice/pleasant/pleasure | 0.0000 | 4.3306 | 87 | 0.152 | 0.26 | H | **0.92** | **M** |
| FE-happiness | 0.0000 | 4.4113 | 140 | 0.173 | **1.33** | M | 0.57 | M |
| FE-seduction | 0.0162 | 2.4531 | 86 | 0.081 | 0.10 | L | **0.30** | M |
| FE- mood | 0.0006 | 3.5252 | 96 | 0.118 | 0.22 | M | **0.63** | M |
| P- women | 0.0002 | 3.7299 | 119 | 0.119 | **0.59** | **H** | 0.14 | M |
| P- age | 0.0011 | 3.4203 | 64 | 0.046 | 0.00 | NC | **0.16** | L |
| P-children | 0.0062 | 2.7787 | 141 | 0.152 | 0.48 | H | **0.90** | **M** |
| P- sharing/society | 0.0100 | 2.6154 | 123 | 0.061 | **0.21** | M | 0.05 | L |
| P- family | 0.0016 | 3.2377 | 92 | 0.107 | 0.11 | H | **0.46** | **M** |
| EN- animals | 0.0041 | 2.9272 | 115 | 0.056 | **0.20** | M | 0.03 | L |
| CUL- studying/intellect | 0.0021 | 3.1956 | 62 | 0.070 | 0.00 | NC | **0.22** | **H** |
| FET- quality/type | 0.0084 | 2.6722 | 136 | 0.282 | 2.02 | H | **2.78** | **H** |
| FET- genuine | 0.0065 | 2.8157 | 62 | 0.056 | 0.00 | NC | **0.16** | M |
| FET-price | 0.0000 | 4.4360 | 86 | 0.070 | **0.31** | **H** | 0 | NC |
| FET- packaging | 0.0011 | 3.3540 | 107 | 0.047 | **0.17** | L | 0.02 | NC |
| S- sports | 0.0103 | 2.6356 | 71 | 0.044 | 0.01 | NC | **0.13** | L |

Table 6_6. *Chocolate* – Fields showing significant difference in the T-Test

Consequently, the following semantic fields would appear as distinctively more prominent for Italians than for the English, when talking about *chocolate*: BAKERY/COOKING; RECIPE; DIETING; MEDICINE; BEAUTY; HISTORY; NICE/PLEASANT/PLEASURE; CHILDREN; FAMILY; STUDYING/INTELLECT; QUALITY/TYPE; GENUINE. On the other hand, more prominent for the English than for Italians seems to be: WOMEN; and PRICE.

The two datasets were compared also at the level of conceptual domains. Spearman’s test showed very strong positive correlation, with Spearman’s Rho equal to 0.939 ($p < 0.01$). As regards the T-test, the significant results ($p < 0.01$) are summarised in Table 6_7. The domains which are not listed in the table did not show a statistically significant difference between the two cultures.

| **Domain** | **P** | **T** | **Df** | **St.error of df** | **Mean values English** | **Cnv** | **Mean values Italian** | **Cnv** |
|---|---|---|---|---|---|---|---|---|
| Food | 0.0007 | 34.511 | 144 | 0.643 | 6.49 | M | **8.71** | M |
| Health & Body | 0.0060 | 27.918 | 132 | 0.324 | 2.67 | M | **3.57** | M |
| Culture | 0.0013 | 33.073 | 102 | 0.184 | 0.55 | M | **1.16** | **H** |
| Feature | 0.0084 | 26.740 | 143 | 0.492 | 5.54 | M | **6.86** | M |

Table 6_7. *Chocolate* – Conceptual domains with statistically significant differences between English and Italian

In all the cases, mean values are higher for Italian. However, only CULTURE shows also a high level of conventionalisation. The other fields show medium conventionalisation. Consequently, following the reasoning applied in discussing semantic field cross-cultural differences, it seems safe to state that CULTURE is the only conceptual domain that clearly distinguishes the Italians from the English in thinking about *chocolate*. The other domains in the list might be distinctive of the Italian culture, but this should be confirmed by further data.

## 6.3 Wine

### *6.3.1 Inter-culture analysis*

#### *6.3.1.1 Semantic field analysis*

The analytical procedure adopted for *chocolate* was applied to the analysis of the *wine* datasets. Tables 6_8 and 6_9 show the semantic associations of *wine* as they emerged in the English and Italian datasets, in decreasing order of frequency.

As in the *chocolate* experiment, Molinari's evenness index was computed and three levels of conventionalisation were distinguished using confidence intervals. The 99% confidence intervals for the *wine* data were respectively 0.73-0.82 for English, and 0.75-0.83 for Italian. For an easier reading of the results, the evenness values are reported in Tables 6_8 and 6_9, in column G2,1, accompanied by indication of their corresponding levels of conventionalisation (column Cnv).

Similarly to what happened with *chocolate*, the two datasets share most of the semantic fields in the Codebook, though with different ranks.

In terms of conventionalisation, the Italian *wine* dataset shows a much higher percentage of highly conventionalised fields than the English dataset (61.64 vs. 47.3), and a much lower percentage of fields in the medium range (12.33 vs. 22.97), as summarised in Table 6_10.

| Field | Mean | SD | Rank | G2,1 | Cnv | Field | Mean | SD | Rank | G2,1 | Cnv |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FET-quality/type | 2.54 | 2.06 | 1 | 0.66 | H | C-ceremonies | 0.12 | 0.36 | 32 | 0.81 | M |
| G-geo locations | 1.55 | 2.28 | 2 | 0.52 | H | H-body | 0.11 | 0.35 | 33 | 0.80 | M |
| H-health | 1.42 | 1.37 | 3 | 0.68 | H | P-people | 0.11 | 0.35 | 33 | 0.80 | M |
| FET-taste/smell | 1.29 | 1.20 | 4 | 0.69 | H | G-spreading | 0.11 | 0.31 | 33 | 1.00 | L |
| FET-price | 1.11 | 1.11 | 5 | 0.68 | H | CUL-artistic production | 0.10 | 0.34 | 34 | 0.79 | M |
| F-food | 1.10 | 1.23 | 6 | 0.62 | H | L-existence | 0.10 | 0.47 | 34 | 0.66 | H |
| F-drink | 1.02 | 1.11 | 7 | 0.69 | H | E-holidays | 0.08 | 0.27 | 35 | 1.00 | L |
| E-excessive drinking | 0.91 | 1.21 | 8 | 0.65 | H | FE-no reaction | 0.08 | 0.27 | 35 | 1.00 | L |
| F-composition | 0.73 | 0.82 | 9 | 0.71 | H | FET-sweet | 0.08 | 0.27 | 35 | 1.00 | L |
| FET-quantity | 0.70 | 0.98 | 10 | 0.66 | H | FE-nice/pleasant/pleasure | 0.07 | 0.25 | 36 | 1.00 | L |
| F-recipe | 0.67 | 1.08 | 11 | 0.58 | H | FE-comfort | 0.07 | 0.25 | 36 | 1.00 | L |
| FE-happiness | 0.63 | 0.88 | 12 | 0.71 | H | FE-mood | 0.05 | 0.27 | 37 | 0.75 | M |
| COM-comparison | 0.62 | 1.06 | 13 | 0.65 | H | FE-memory | 0.05 | 0.27 | 37 | 0.75 | M |
| P-women | 0.57 | 1.11 | 14 | 0.50 | H | P-children | 0.05 | 0.23 | 37 | 1.00 | L |
| FE-desire | 0.53 | 0.97 | 15 | 0.61 | H | LD-drugs & addiction | 0.05 | 0.23 | 37 | 1.00 | L |
| E-time | 0.48 | 0.79 | 16 | 0.75 | M | CUL-studying/intellect | 0.05 | 0.23 | 37 | 1.00 | L |
| H-medicine | 0.46 | 0.83 | 17 | 0.67 | H | E-driving | 0.04 | 0.21 | 38 | 1.00 | L |
| P-sharing/society | 0.43 | 0.72 | 18 | 0.72 | H | CUL-culture | 0.04 | 0.25 | 38 | 0.73 | M |
| P-men | 0.42 | 1.05 | 19 | 0.30 | H | E-economy | 0.03 | 0.18 | 39 | 1.00 | L |
| P-posh | 0.42 | 0.78 | 19 | 0.72 | H | E-history | 0.03 | 0.18 | 39 | 1.00 | L |
| FET-physical properties | 0.42 | 0.79 | 19 | 0.70 | H | FE-senses | 0.03 | 0.23 | 39 | 0.71 | H |
| F-manufacturing | 0.40 | 0.74 | 20 | 0.64 | H | FE-peace | 0.03 | 0.23 | 39 | 0.71 | H |
| FE-relax | 0.40 | 0.79 | 20 | 0.74 | M | LD-hiding | 0.03 | 0.18 | 39 | 1.00 | L |
| FET-packaging | 0.40 | 0.68 | 20 | 0.74 | M | EN-nature | 0.03 | 0.23 | 39 | 0.71 | H |
| F-product/shape | 0.38 | 0.70 | 21 | 0.75 | M | EN-house | 0.03 | 0.18 | 39 | 1.00 | L |
| F-bakery/cooking | 0.38 | 0.70 | 21 | 0.71 | H | FE-confidence | 0.02 | 0.15 | 40 | 1.00 | L |
| P-family | 0.38 | 0.88 | 21 | 0.62 | H | FE-seduction | 0.02 | 0.15 | 40 | 1.00 | L |
| FE-unpleasant | 0.36 | 0.82 | 22 | 0.56 | H | FE-freedom | 0.02 | 0.15 | 40 | 1.00 | L |
| F-storage | 0.35 | 0.72 | 23 | 0.66 | H | P-gay | 0.02 | 0.15 | 40 | 1.00 | L |
| E-transaction | 0.33 | 0.60 | 24 | 0.73 | M | LD-theft | 0.02 | 0.15 | 40 | 1.00 | L |
| E-religion | 0.33 | 0.73 | 24 | 0.64 | H | EN-animals | 0.02 | 0.15 | 40 | 1.00 | L |
| P-friendship | 0.33 | 0.73 | 24 | 0.71 | H | FET-genuine | 0.02 | 0.21 | 40 | NC | |
| E-event | 0.25 | 0.49 | 25 | 0.82 | M | H-beauty | 0.01 | 0.10 | 41 | NC | |
| FET-colour | 0.25 | 0.64 | 25 | 0.71 | H | E-playing | 0.01 | 0.10 | 41 | NC | |
| C-gift | 0.23 | 0.54 | 26 | 0.78 | M | E-war | 0.01 | 0.10 | 41 | NC | |
| P-age | 0.22 | 0.53 | 27 | 0.71 | H | E-law | 0.01 | 0.10 | 41 | NC | |
| C-party | 0.19 | 0.47 | 28 | 0.78 | M | FE-sex | 0.01 | 0.10 | 41 | NC | |
| EN-dirt | 0.19 | 0.42 | 28 | 0.84 | L | FE-surprise | 0.01 | 0.10 | 41 | NC | |
| FE-passion | 0.18 | 0.46 | 29 | 0.77 | M | FE-guilt | 0.01 | 0.10 | 41 | NC | |
| F-serving | 0.16 | 0.40 | 30 | 0.83 | L | FE-bribing | 0.01 | 0.10 | 41 | NC | |
| E-work | 0.16 | 0.60 | 30 | 0.62 | H | I-fantasy/magic | 0.01 | 0.10 | 41 | NC | |
| E-language | 0.15 | 0.49 | 31 | 0.70 | H | L-future | 0.01 | 0.10 | 41 | NC | |
| FE-love | 0.12 | 0.39 | 32 | 0.77 | M | | | | | | |

Table 6_8. Semantic associations of *wine* for the English

| Field | Mean | SD | Rank | G2.1 | Cnv | Field | Mean | SD | Rank | G2.1 | Cnv |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FET-quality/type | 3.61 | 2.26 | 1 | 0.72 | H | FE-relax | 0.19 | 0.54 | 32 | 0.67 | H |
| G-geo locations | 1.81 | 1.47 | 2 | 0.71 | H | EN-dirt | 0.18 | 0.46 | 33 | 0.77 | M |
| F-manufacturing | 1.58 | 1.51 | 3 | 0.67 | H | CUL-culture | 0.18 | 0.53 | 33 | 0.80 | M |
| F-recipe | 1.55 | 1.40 | 4 | 0.65 | H | LD-drugs & addiction | 0.16 | 0.45 | 34 | 0.77 | M |
| H-health | 1.50 | 1.17 | 5 | 0.70 | H | C-party | 0.16 | 0.49 | 34 | 0.64 | H |
| F-food | 1.48 | 1.11 | 6 | 0.70 | H | H-body | 0.15 | 0.40 | 35 | 0.79 | M |
| FET-quantity | 1.31 | 1.14 | 7 | 0.74 | H | FE-mood | 0.13 | 0.34 | 36 | 1.00 | L |
| P-friendship | 0.94 | 0.99 | 8 | 0.70 | H | FE-passion | 0.13 | 0.34 | 36 | 1.00 | L |
| FET-taste/smell | 0.85 | 0.88 | 9 | 0.73 | H | FE-peace | 0.13 | 0.50 | 36 | 0.67 | H |
| F-bakery/cooking | 0.82 | 0.88 | 10 | 0.73 | H | EN-house | 0.13 | 0.46 | 36 | 0.80 | M |
| E-language | 0.77 | 0.82 | 11 | 0.71 | H | F-product/shape | 0.11 | 0.45 | 37 | 0.61 | H |
| H-medicine | 0.76 | 0.88 | 12 | 0.67 | H | FE-desire | 0.11 | 0.37 | 37 | 0.77 | M |
| FE-happiness | 0.74 | 0.96 | 13 | 0.63 | H | FE-comfort | 0.11 | 0.37 | 37 | 0.77 | M |
| F-storage | 0.66 | 0.89 | 14 | 0.72 | H | P-posh | 0.11 | 0.48 | 37 | 0.68 | H |
| E-event | 0.66 | 0.94 | 14 | 0.67 | H | L-existence | 0.11 | 0.45 | 37 | 0.61 | H |
| FE-unpleasant | 0.66 | 0.85 | 14 | 0.74 | H | E-holidays | 0.10 | 0.30 | 38 | 1.00 | L |
| E-excessive drinking | 0.65 | 0.83 | 15 | 0.73 | H | P-sharing/society | 0.10 | 0.30 | 38 | 1.00 | L |
| FET-physical properties | 0.65 | 0.91 | 15 | 0.70 | H | FE-memory | 0.08 | 0.27 | 39 | 1.00 | L |
| E-transaction | 0.56 | 0.88 | 16 | 0.66 | H | G-spreading | 0.08 | 0.27 | 39 | 1.00 | L |
| FE-confidence | 0.55 | 0.74 | 17 | 0.70 | H | EN-tech | 0.08 | 0.33 | 39 | 0.75 | H |
| FE-nice/pleasant/pleasure | 0.52 | 0.76 | 18 | 0.74 | H | FE-senses | 0.06 | 0.25 | 40 | 1.00 | L |
| F-composition | 0.48 | 0.70 | 19 | 0.73 | H | FE-love | 0.06 | 0.25 | 40 | 1.00 | L |
| CUL-artistic production | 0.47 | 0.74 | 20 | 0.70 | H | FE-seduction | 0.06 | 0.25 | 40 | 1.00 | L |
| CUL-studying/intellect | 0.47 | 0.76 | 20 | 0.70 | H | FE-loneliness | 0.06 | 0.25 | 40 | NC | |
| F-drink | 0.45 | 0.80 | 21 | 0.69 | H | P-age | 0.06 | 0.25 | 40 | 1.00 | L |
| P-family | 0.45 | 0.99 | 21 | 0.68 | H | E-economy | 0.05 | 0.22 | 41 | 1.00 | L |
| F-serving | 0.40 | 0.71 | 22 | 0.64 | H | FET-energy | 0.05 | 0.22 | 41 | 1.00 | L |
| COM-comparison | 0.39 | 0.64 | 23 | 0.72 | H | H-dieting | 0.03 | 0.18 | 42 | 1.00 | L |
| E-religion | 0.39 | 0.71 | 23 | 0.73 | H | E-law | 0.03 | 0.18 | 42 | 1.00 | L |
| C-gift | 0.37 | 0.66 | 24 | 0.72 | H | P-women | 0.03 | 0.18 | 42 | 1.00 | L |
| FET-packaging | 0.37 | 0.68 | 24 | 0.72 | H | P-men | 0.03 | 0.25 | 42 | NC | |
| EN-nature | 0.35 | 0.68 | 25 | 0.72 | H | P-royalty | 0.03 | 0.18 | 42 | 1.00 | L |
| E-driving | 0.34 | 0.54 | 26 | 0.81 | M | C-ceremonies | 0.03 | 0.18 | 42 | 1.00 | L |
| FET-price | 0.34 | 0.63 | 26 | 0.71 | H | H-beauty | 0.02 | 0.13 | 43 | NC | |
| E-time | 0.32 | 0.78 | 27 | 0.66 | H | E-playing | 0.02 | 0.13 | 43 | NC | |
| FET-colour | 0.29 | 0.64 | 28 | 0.71 | H | FE-sex | 0.02 | 0.13 | 43 | NC | |
| E-work | 0.27 | 0.45 | 29 | 1.00 | L | FE-competitiveness | 0.02 | 0.13 | 43 | NC | |
| FET-genuine | 0.27 | 0.52 | 29 | 0.80 | M | FE-freedom | 0.02 | 0.13 | 43 | NC | |
| E-history | 0.24 | 0.64 | 30 | 0.59 | H | I-fantasy/magic | 0.02 | 0.13 | 43 | NC | |
| P-children | 0.24 | 0.43 | 30 | 1.00 | L | I-dream | 0.02 | 0.13 | 43 | NC | |
| FE-no reaction | 0.23 | 0.53 | 31 | 0.68 | H | LD-hiding | 0.02 | 0.13 | 43 | NC | |
| FET-sweet | 0.23 | 0.56 | 31 | 0.68 | H | S-sports | 0.02 | 0.13 | 43 | NC | |

Table 6_9. Semantic associations of *wine* for the Italians

| | English | Italian |
|---|---|---|
| High | 47.3 | 61.64 |
| Medium | 22.97 | 12.33 |
| Low | 29.73 | 26.03 |

Table 6_10. *Wine* – Percentage distribution of fields across conventionalisation levels

This result is in line with expectations, *wine* being a major and long-standing traditional national product for Italy, but a relatively recent import product for England. However, in both cultures, the sum of high and medium conventionalisation fields – i.e. the fields which highlight cultural meanings – amounts to about 71% and 74% for English and Italian respectively.

### *6.3.1.2 Conceptual domain analysis*

The following paragraphs apply the analytical steps described above to conceptual domains. Table 6_11 summarises Mean, SD, Rank, G2,1 and Conventionalisation values for the Italian and English *wine* datasets, at this higher level. All values have been computed from raw data, ignoring the existence of subdivisions (semantic fields) within each conceptual domain. However, in the table, the domain name is accompanied by the number of its conceptual fields, in parenthesis.

| **ENGLISH** | | | | | | **ITALIAN** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Domain** | **Mean** | **SD** | **Rank** | **G2,1** | **Cnv** | **Domain** | **Mean** | **SD** | **Rank** | **G2,1** | **Cnv** |
| Features (10) | 6.80 | 4.12 | 1 | 0.67 | M | Features (10) | 7.97 | 3.33 | 1 | 0.78 | M |
| Food (9) | 5.20 | 3.37 | 2 | 0.67 | M | Food (9) | 7.55 | 3.08 | 2 | 0.77 | M |
| People (11) | 2.96 | 3.07 | 3 | 0.56 | H | Events (15) | 4.40 | 2.04 | 3 | 0.76 | M |
| Events (15) | 2.85 | 2.11 | 4 | 0.68 | M | Feelings & emotions (23) | 3.89 | 2.62 | 4 | 0.66 | H |
| Feelings & emotions (23) | 2.70 | 2.13 | 5 | 0.65 | M | Health (5) | 2.45 | 1.61 | 5 | 0.70 | M |
| Health (5) | 2.00 | 1.89 | 6 | 0.65 | M | People (11) | 2.00 | 1.68 | 6 | 0.63 | H |
| Geo (2) | 1.66 | 2.29 | 7 | 0.53 | H | Geo (2) | 1.89 | 1.52 | 7 | 0.72 | M |
| Comparison (1) | 0.62 | 1.06 | 8 | 0.65 | M | Culture (3) | 1.11 | 1.29 | 8 | 0.69 | H |
| Ceremony (3) | 0.54 | 0.81 | 9 | 0.72 | L | Environment (5) | 0.74 | 0.97 | 9 | 0.75 | M |
| Environment (5) | 0.27 | 0.58 | 10 | 0.72 | M | Ceremony (3) | 0.56 | 0.78 | 10 | 0.71 | M |
| Culture (3) | 0.20 | 0.45 | 11 | 0.80 | L | Comparison (1) | 0.39 | 0.64 | 11 | 0.72 | M |
| Loss & damage (3) | 0.11 | 0.38 | 12 | 0.77 | L | Loss & damage (3) | 0.18 | 0.46 | 12 | 0.77 | M |
| Life (2) | 0.11 | 0.48 | 12 | 0.65 | M | Life (2) | 0.11 | 0.45 | 13 | 0.61 | H |
| Imagination (2) | 0.01 | 0.10 | 13 | NC | | Imagination (2) | 0.03 | 0.18 | 14 | 1.00 | L |
| Sports (1) | 0.00 | 0.00 | 14 | NC | | Sports (1) | 0.02 | 0.13 | 15 | NC | |

Table 6_11. *Wine* – Conceptual domains in the English and Italian datasets

Disregarding the domains for which the evenness index could not be computed, the English dataset shows two highly conventionalised domains: PEOPLE, and GEOGRAPHY; three domains with low levels of conventionalisation: CEREMONY, CULTURE, LOSS & DAMAGE,; and eight domains with medium conventionalisation: FEATURES, FOOD, EVENTS, FEELINGS & EMOTIONS, HEALTH, COMPARISON, ENVIRONMENT and LIFE. The Italian dataset shows four highly conventionalised domains: FEELINGS & EMOTIONS, PEOPLE, CULTURE, and LIFE; one domain with low levels of conventionalisation: IMAGINATION; and nine domains with medium conventionalisation: FEATURES, FOOD, EVENTS, HEALTH, GEO LOCATIONS, ENVIRONMENT, CEREMONY, COMPARISON and LOSS & DAMAGE.

In percentage terms, the conventionalisation picture is as summarised in Table 6_12.

| | **English** | **Italian** |
|---|---|---|
| High | 15.00 | 19.00 |
| Medium | 62.00 | 64.00 |
| Low | 23.00 | 7.00 |

Table 6_12. *Wine* – Percentage distribution of domains across conventionalisation levels

In parallel with what happened with semantic fields, the Italian culture, compared to the English one, shows a greater number of highly conventionalised domains and a lower number of domains with low conventionalisation. Furthermore, the sum of high and medium conventionalisation domains amount to 75% and 83% respectively.

Furthermore, as was the case with *chocolate*, passing from semantic fields to conceptual domains, the picture within each culture has changed in two ways: 1. while in the case of semantic fields both cultures showed dominance of highly conventionalised fields, in the case of domains the leading level of conventionalisation is the medium one; 2. the total amount of semantic meanings has increased.

Finally, no direct relationship exists between mean frequency and number of semantic fields composing the domain, or mean frequency and conventionalisation, or conventionalisation and number of fields in the domain.

### *6.3.2 Cross-cultural comparison*

Let us now compare the two cultures at the level of semantic fields, to see whether significant differences exist. As with *chocolate*, semantic field comparison between the English and Italian datasets performed by applying Spearman's Rank Correlation Coefficient showed medium-high positive correlation, with Spearman's Rho equal to 0.735 ($p < 0.01$). Furthermore, the *wine* datasets were compared using Welch *t* Test for Independent Samples. T-Test results were significant ($p < 0.01$) for the semantic fields listed in Table 6_13.

Following the logic applied with *chocolate,* the fields with a significant T value and which show a high level of conventionalisation in one of the two cultures along with mean values for that culture which are higher than for the other culture will be considered indicative of cultural differences. The fields with a significant higher mean in one culture but a low level of conventionalisation in that culture, will be ignored as the result may depend on the sample, rather than the culture. Finally, the semantic fields with a higher mean and medium conventionalisation level will be considered culturally more prominent only if in the other culture they show a high level of conventionalisation or are virtually absent. All other cases are uncertain, and need confirmation from other population samples.

| **Field** | **P** | **T** | **Df** | **st.error of df** | **Mean values English** | **Cnv** | **Mean values Italian** | **Cnv** |
|---|---|---|---|---|---|---|---|---|
| F- product/shape | 0.0038 | 2.9395 | 150 | 0.092 | 0.38 | **M** | 0.11 | H |
| F- bakery/cooking | 0.0013 | 3.2867 | 110 | 0.133 | 0.38 | H | **0.82** | **H** |
| F- drink | 0.0003 | 3.6947 | 150 | 0.154 | 1.02 | **H** | 0.45 | H |
| F-manufacturing | 0.0000 | 5.7245 | 81 | 0.207 | 1.58 | H | 0.40 | H |
| F-recipe | 0.0001 | 4.3831 | 151 | 0.200 | 1.55 | **H** | 0.67 | H |
| E-language | 0.0000 | 5.8467 | 151 | 0.106 | 0.77 | **H** | 0.33 | H |
| E- event | 0.0022 | 3.1487 | 83 | 0.130 | 0.25 | M | **0.66** | **H** |
| E- driving | 0.0001 | 4.0889 | 73 | 0.072 | 0.04 | L | **0.34** | M |
| FE-confidence | 0.0000 | 6.6104 | 151 | 0.080 | 0.55 | **H** | 0.02 | L |
| FE- desire | 0.0003 | 3.7065 | 123 | 0.112 | 0.53 | **H** | 0.11 | M |
| FE-nice/pleasant/pleasure | 0.0000 | 5.2408 | 151 | 0.086 | 0.52 | **H** | 0.07 | L |
| P-women | 0.0000 | 3.7979 | 151 | 0.142 | 0.03 | L | **0.57** | **H** |
| P- men | 0.0011 | 3.3454 | 104 | 0.115 | 0.42 | **H** | 0.03 | NC |
| P- children | 0.0024 | 3.1236 | 84 | 0.060 | 0.05 | L | **0.24** | L |
| P-friendship | 0.0001 | 4.3530 | 151 | 0.139 | 0.94 | **H** | 0.33 | H |
| P- posh | 0.0032 | 2.9924 | 149 | 0.102 | 0.42 | **H** | 0.11 | H |
| P- sharing/society | 0.0001 | 3.942 | 129 | 0.084 | 0.43 | **H** | 0.10 | L |
| P- people | 0.0033 | 3.0121 | 90 | 0.036 | 0.11 | M | 0.00 | NC |
| EN- nature | 0.0006 | 3.5868 | 70 | 0.090 | 0.03 | H | **0.35** | **H** |
| CUL- artistic production | 0.0004 | 3.6746 | 78 | 0.100 | 0.10 | M | **0.47** | **H** |
| CUL-studying/intellect | 0.0001 | 4.8607 | 151 | 0.085 | 0.47 | **H** | 0.05 | **L** |
| FET- quality/type | 0.0033 | 2.993 | 123 | 0.359 | 2.54 | H | **3.61** | **H** |
| FET- quantity | 0.0009 | 3.3953 | 118 | 0.178 | 0.70 | H | **1.31** | **H** |
| FET- genuine | 0.0005 | 3.6392 | 74 | 0.069 | 0.02 | NC | **0.27** | M |
| FET-price | 0.0000 | 4.9567 | 151 | 0.156 | 0.34 | **H** | **1.11** | **H** |

Table 6_13. *Wine* – T-Test results

Consequently, I would consider the following semantic fields as distinctively more prominent for Italians than for the English, when talking about *wine*: BAKERY/COOKING; EVENT; WOMEN; NATURE; ARTISTIC PRODUCTION; QUALITY/TYPE; QUANTITY; GENUINE; PRICE. On the other hand, more prominent for the English than for the Italians are: PRODUCT/SHAPE; DRINK; MANUFACTURING; RECIPE; LANGUAGE; CONFIDENCE; DESIRE; NICE/PLEASANT/PLEASURE; MEN; FRIENDSHIP; POSH; SHARING/SOCIETY; PEOPLE; and STUDYING/INTELLECT.

As regards conceptual domains, Spearman's test showed very strong positive correlation, with Spearman's Rho equal to 0.942 ($p < 0.01$). Furthermore, the domains with statistically significant differences at the Welch *t* Test ($p < 0.01$) are listed in Table 6_14.

| **Domain** | **P** | **T** | **Df** | **st.error of df** | **Mean values English** | **Cnv** | **Mean values Italian** | **Cnv** |
|---|---|---|---|---|---|---|---|---|
| Food | 0.0000 | 4.4579 | 138 | 0.572 | 5.20 | M | **7.55** | M |
| Event | 0.0000 | 4.5729 | 134 | 0.340 | 2.85 | M | **4.40** | M |
| Feeling | 0.0038 | 2.9547 | 112 | 0.401 | 2.70 | M | **3.89** | **H** |
| Environment | 0.0010 | 33.912 | 90 | 0.138 | 0.27 | M | **0.74** | M |
| Culture | 0.0000 | 5.3482 | 71 | 0.171 | 0.20 | L | **1.11** | **H** |
| Features | 0.0001 | 4.1143 | 138 | 0.836 | **11.41** | M | 7.97 | M |

Table 6_14. *Wine* – Conceptual domains with statistically significant differences between English and Italian

Only domains FEELINGS and CULTURE emerge as clearly distinctive for the Italian culture. The other domains in the list, having medium level of conventionalisation in both cultures are more ambiguous, and further data are needed.

## 6.4 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 an even smaller number of negative sentences, and a few undecided sentences, as summarised in Table 6_15. In the table, the numerical values are percentages of the total number of sentences in each dataset.

| | **Positive** | **Negative** | **Neutral** | **Undecided** |
|---|---|---|---|---|
| English *chocolate* | 53.92 | 19.03 | 26.35 | 0.69 |
| Italian *chocolate* | 54.21 | 11.85 | 32.75 | 1.19 |
| English *wine* | 46.00 | 19.60 | 27.69 | 6.70 |
| Italian *wine* | 53.59 | 14.49 | 29.62 | 2.29 |

Table 6_15. ASSESSMENT field results in the elicited datasets

Such an analysis, although clearly limited in scope, is sufficient for the purposes of the current work and is a suitable reference term for the methodological comparisons which will be performed in the following chapters.

From a cultural perspective, however, the current analysis of semantic prosody would benefit from extension. In particular, two possible analytical procedures have already been identified: 1. looking at the distribution of the Positive and Negative categories across the various fields/domains;[6] 2. analysing the evaluative adjectives that collocate with the two selected key words.[7]

## 6.5 Conclusions

The present chapter has outlined the semantic associations of the key words *chocolate*, and *wine* in Italian and English minds, as the emerge from the four datasets of elicited data collected. These results will be used in the next chapters as terms of comparison for methodological investigations in the use of different types of data samples and tagging systems.

The following paragraphs summarise the procedures applied and the results obtained, along with some further methodological considerations and a brief discussion of how these data confirm the cultural systems theories used as reference framework. Finally, the last few paragraphs discuss some limitations of the analyses performed and suggest directions for further research.

Observation of the ranking of semantic fields and conceptual domains based on mean values provided a general picture of the most frequent semantic associations in each data set. This picture, however, is approximate, as it disregards distribution of answers across subjects.

A more precise picture was obtained by applying Molinari's evenness index, and by assessing the level of conventionalisation expressed by each value, classified into three groups: High, Medium and Low. Consequently, in each culture and for each node word, it was possible to establish which fields and domains showed high, medium or low level of conventionalisation, respectively corresponding to Fleischer's core, current and connotational fields.

The results are in keeping with expectations. Although *wine* is well established in both countries – with the sum of core (H Cnv elements) and current field (M Cnv elements) predominating over connotational field (L Cnv elements) in both datasets –, among Italian respondents, for whom it is a long-standing traditional national product, the percentage of highly conventionalised semantic fields is remarkably higher than among the British ones, and the percentage of low conventionalisation fields is remarkably lower. A similar picture appeared also from the analysis of conceptual domains.

*Chocolate*, too, appears as a well-established symbol in both cultures, both at the level of semantic fields and conceptual domains. Differently from *wine*, however, the distribution of high, medium and low conventionalisation fields and domains is almost identical in both datasets, which confirms our initial hypothesis of *chocolate* having similar rooting in the two countries.

Finally, for each node word, the English and Italian semantic associations 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

[6] A quick look at the data suggests that, when performing this type of analysis, it will be important to consider only the semantic fields/domains which show a minimum number of hits, alongside a significant a difference between Positive/Negative Assessment.

[7] Methodological inspiration could be taken from the works by Baker (2006), and Aggarwal, Vaidyanathan and Venkatesh (2009), reviewed in Chapter 2.

conventionalisation results, in order to better understand which differences can be safely attributed to culture and which to circumstantial elements, such as population sampling.

Cross-cultural comparisons in terms of conceptual domains highlighted very few, and sometimes ambiguous, differences. Indeed, conceptual domains – though highly useful in the construction of the coding scheme and in its application – proved less useful than semantic fields in cultural and, even more so, cross-cultural analyses. This is most probably due to the fact that they are very wide as categories, and consequently less sensitive as indicators of difference.[8]

At the level of semantic fields, the Italians seem to distinguish themselves from the British for their more frequent matching of *chocolate* to the following concepts: BAKERY/COOKING; RECIPE; DIETING; MEDICINE; BEAUTY; HISTORY; NICE/PLEASANT/PLEASURE; CHILDREN; FAMILY; STUDYING/INTELLECT; QUALITY/TYPE; GENUINE. On the other hand, more prominent for the English than for Italians seem to be: WOMEN, and PRICE. As regards *wine*, the following semantic fields emerged as distinctively more prominent for the Italians than for the English: BAKERY/COOKING; EVENT; WOMEN; NATURE; ARTISTIC PRODUCTION; QUALITY/TYPE; QUANTITY; GENUINE; PRICE. On the other hand, more prominent for the English than for Italians were: PRODUCT/SHAPE; DRINK; MANUFACTURING; RECIPE; LANGUAGE; CONFIDENCE; DESIRE, NICE/PLEASANT/PLEASURE; MEN, FRIENDSHIP; POSH; SHARING/SOCIETY; PEOPLE; and STUDYING/INTELLECT.

Finally, The results of the present study suggest some further general considerations.

A look at the semantic fields which are absent with reference to both key words in the same culture suggests that field presence/absence depends on the key word, rather than the culture. In fact, only one field is systematically absent in the English datasets (COMPETITIVENESS), and none in the Italian ones. This supports the use of dedicated coding systems for different node words.

Furthermore, these experiments suggest that, although the relation between the frequency of occurrence of a semantic field and its conventionalisation is evident, its exact nature might not be a simple and direct one. The quantitative nature of this relation is worth further investigation.

Finally, the current results are in keeping not only with Fleischer's theory but also Nobis's one. The *wine* experiment clearly confirmed that longer standing of a concept (*wine*) in a given culture (Italy) corresponds to stronger cultural rooting, here expressed in terms on higher percentage of highly conventionalised semantic fields. The second of Nobis' hypotheses, postulating greater semantic complexity of longer standing concepts, is supported in the *wine* experiment not by the overall number of semantic fields associated to the given concept, but by the greater number of semantic elements which are shared by several respondents, i.e. those semantic fields or conceptual domains with high levels of conventionalisation.

[8] This is in keeping with results in Guerrero, Claret, Verbeke *et al.* (2010), reviewed in Chapter 4.

The inter-culture analyses performed in this chapter only provide a list of the concepts to which the key words are associated in each culture, but they cannot in any way explain the type of (or reason for) the association. Further steps, such as analysis of individual concordance lines, are needed to understand the exact link between key word and semantic field. Such analyses are beyond the scope of the current investigation, but will be considered in future extensions of this work. Nevertheless, I believe that analyses of this type may be adopted in the exploratory phases of marketing (or cultural) research, where research aims to outline problems, collect information, eliminate impractical ideas, and formulate hypotheses (see Chapter 4, Section 4.1.2).

The current results, however, have limitations deriving from not having controlled the composition of the two population samples (described in Chapter 5). Although the English and Italian groups of respondents show some overall similarities (a majority of university students in the 18-25 age range; data collected in both Northern and Southern areas of the two countries), no precise data is available as regards variables such as the respondents' age, gender or occupation. Consequently, some doubts remain as to the impact of population sampling on the (cross)cultural results. A first confirmation will be provided by comparing the current results to Web data (Chapter 10). Further confirmation could be found by applying, for example, one or more of the following:[9]

- Replication of the study, possibly also with a larger sample size and/or more stratified random sampling.
- Other elicitation methods (e.g. story writing).
- Depth interviews and focus groups, possibly with deliberate attempts to elicit and probe the concepts that showed cultural differences (e.g. ask Italian and English respondents deliberately about women and chocolate and see if there is a difference in how they talk about the subject).
- Content analysis (visual as well as verbal) of representative samples of chocolate/wine advertising from UK and Italian companies addressing the local audience.

For the time being, we will have to accept these results as they are. Should further research disconfirm this cultural comparison and cast doubts on the frequency-plus-T-test method adopted here, nevertheless, the methodological investigations of the chapters that follow will still be valuable. In fact, from now on the focus of the research will shift from finding a suitable way to highlight cultural differences to comparing different types of data and/or coding schemes.

[9] To my best knowledge, no research on chocolate and wine that matches mine or that is in any way useful to explain, confirm or disconfirm my results seems to be currently available.

# 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

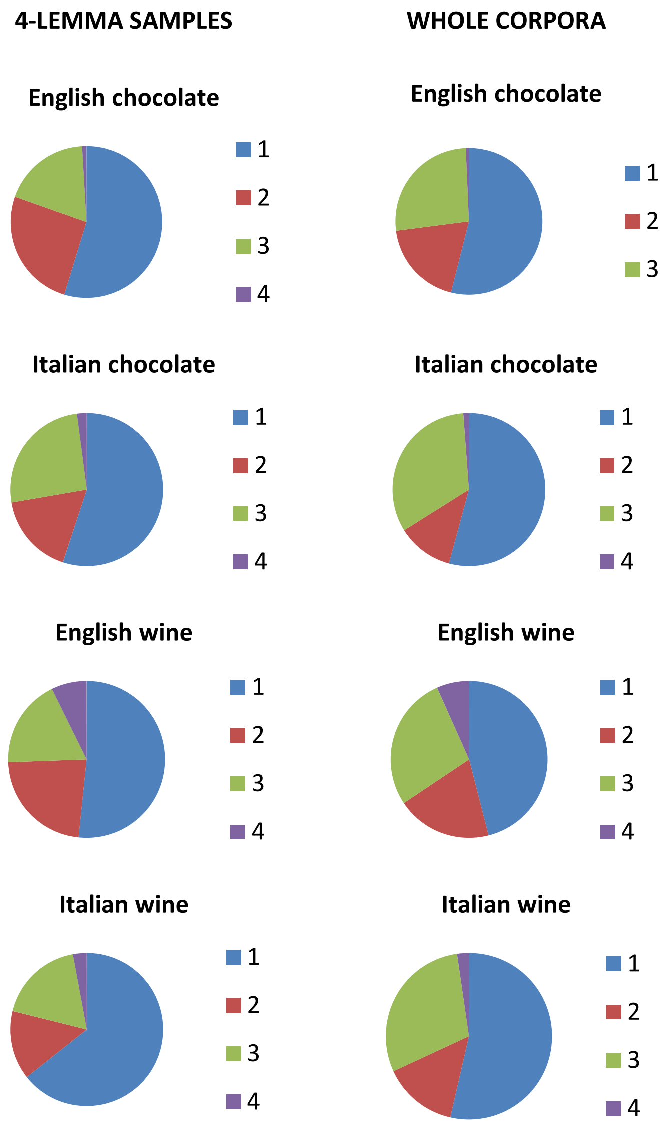


Figure 7_1. ASSESSMENT field results:
4-lemma sampled datasets vs. whole elicited datasets

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** | **T** | **ff** | **st.error** | **mean values** | **Cnv** | **mean values** | **Cnv** |
|---|---|---|---|---|---|---|---|---|
| | **(< 0.05)** | | | **of df** | **English** | | **Italian** | |
| 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** | **T** | **ff** | **st.error** | **mean values** | **Cnv** | **mean values** | **Cnv** |
|---|---|---|---|---|---|---|---|---|
| | **(< 0.05)** | | | **of df** | **English** | | **Italian** | |
| 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** | **T** | **ff** | **st.error** | **mean values** | **Cnv** | **mean values** | **Cnv** |
|---|---|---|---|---|---|---|---|---|
| | **(< 0.05)** | | | **of df** | **English** | | **Italian** | |
| 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** | **T** | **ff** | **st.error** | **mean values** | **Cnv** | **mean values** | **Cnv** |
|---|---|---|---|---|---|---|---|---|
| | **(< 0.05)** | | | **of df** | **English** | | **Italian** | |
| 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

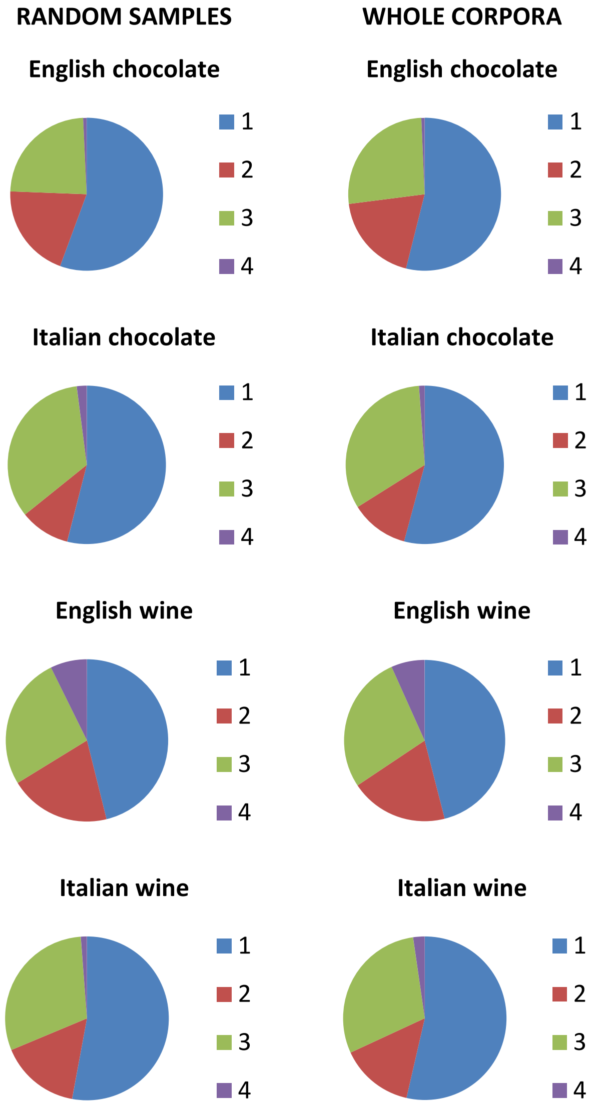


Figure 7_2. ASSESSMENT field results:randomly sampled datasets vs. whole elicited datasets

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

| | | | | | | | | Cn |
|---|---|---|---|---|---|---|---|---|
| **Field** | **G2,1** | **Cnv** | **Field** | **G2,1** | **Cnv** | **Field** | **G2,1** | **v** |
| 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

| | | | | | | | | Cn |
|---|---|---|---|---|---|---|---|---|
| **Field** | **G2,1** | **Cnv** | **Field** | **G2,1** | **Cnv** | **Field** | **G2,1** | **v** |
| 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**<br>**(< 0.05)** | **T** | **ff** | **st.error**<br>**of df** | **mean values**<br>**English** | **Cnv** | **mean values**<br>**Italian** | **Cnv** |
|---|---|---|---|---|---|---|---|---|
| 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**<br>**(< 0.05)** | **T** | **ff** | **st.error**<br>**of df** | **mean values**<br>**English** | **Cnv** | **mean values**<br>**Italian** | **Cnv** |
|---|---|---|---|---|---|---|---|---|
| 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** | **T** | **ff** | **st.error** | **mean values** | **Cnv** | **mean values** | **Cnv** |
|---|---|---|---|---|---|---|---|---|
| | **(< 0.05)** | | | **of df** | **English** | | **Italian** | |
| 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** | **T** | **ff** | **st.error** | **mean values** | **Cnv** | **mean values** | **Cnv** |
|---|---|---|---|---|---|---|---|---|
| | **(< 0.05)** | | | **of df** | **English** | | **Italian** | |
| 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.

Chapter 8

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

## 8.1 Introduction

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

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

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

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

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

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

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

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

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

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

### *8.2.1 Elicited data*

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

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

[3] See Chapter 5.

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

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

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

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

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

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

### *8.2.2 Web data*

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

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

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

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

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

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

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

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

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

## 8.3 Randomly sampled sub-corpora

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

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

### *8.3.1 Elicited data*

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

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

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

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

### *8.3.2 Web data*

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

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

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

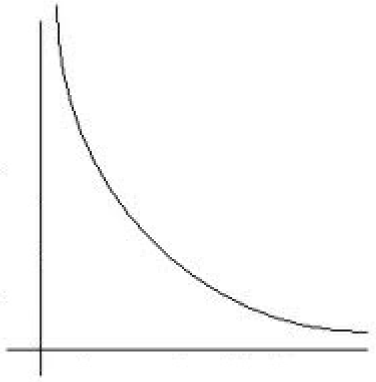
Table 8_6. Randomly sampled Web sub-corpora, tagged with Wmatrix

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

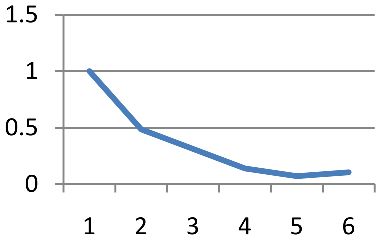
## 8.4 Mathematical progression of field increases: a Zipf-like curve?

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

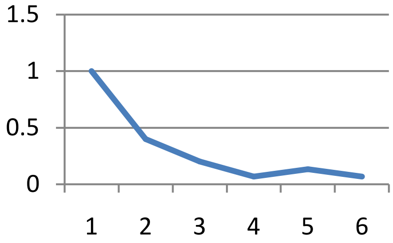
At this point in the research, after having experimented with a wider number of corpora and coding schemes, I have collected a reasonable number of examples of 'category increase progressions' which can be plotted and compared to each other, in order to decide whether Zipf's law can be called into play.



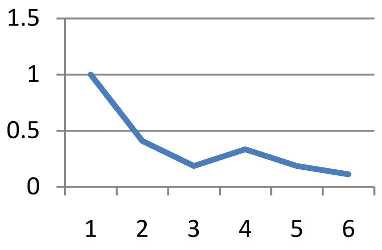
Graph 8_1. Graphic example of Zipf's distribution



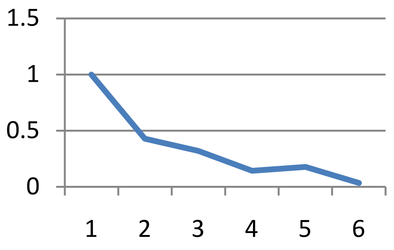
Graph 8_2. Data from Table 7_1



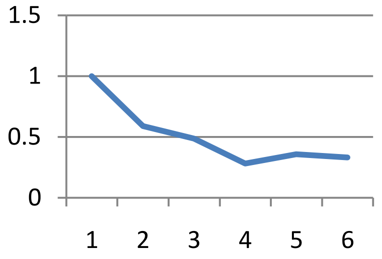
Graph 8_3. Data from Table 7_2



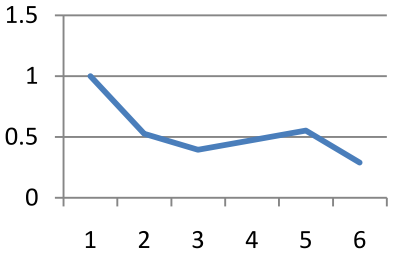
Graph 8_4. Data from Table 7_3



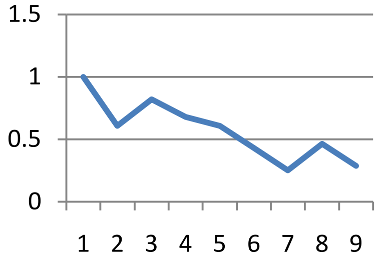
Graph 8_5. Data from Table 7_4



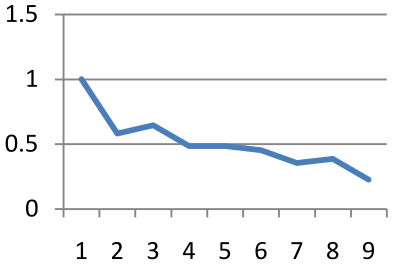
Graph 8_6. Data from Table 8_1



Graph 8_7. Data from Table 8_2



Graph 8_8. Data from Table 8_3



Graph 8_9. Data from Table 8_4

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

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

## 8.5 Conclusions

The current chapter repeated the experiments described in Chapter 7 – i.e. comparing a dataset to the most frequent content words in its wordlist, and to a sub-corpus randomly sampled from the same dataset, the latter being seen as possible shortcuts to the semantic analysis of the whole dataset – but used a different tagging scheme. When passing from manual coding to USAS tagging, differences were observed in the percentages of semantic categories retrieved. Indeed, with USAS tagging, both routes retrieved a smaller percentage of semantic categories. This is most certainly explained by the fact that the USAS dataset – “arranged in a hierarchy with 21 major discourse fields expanding into 232 category labels” (Archer, Wilson, & Rayson, 2002, pp. 1-2), some of which are further subdivided into finer categories marked by a decimal point followed by a further digit, or “one or more ‘pluses’ or ‘minuses’ to indicate a positive or negative position on a semantic scale” (*ibid.*) – includes a higher number of categories and is much finer-grained than the tagset used for manual coding. Indeed, a similar phenomenon was noticed in Chapter 7 when comparing semantic field and conceptual domain results, i.e. a finer-grained scheme and a broader one.

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

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

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

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

Chapter 9

# Automatic tagging

## 9.1 Introduction

In Chapter 8, I used automatic semantic tagging to verify the validity of different types of sampling procedures. On that occasion, automatic tagging of the whole datasets (the elicited, as well as the Web ones) was compared to automatic tagging of sampled subsets. In the current chapter, the same automatic semantic tagger – Wmatrix, the automatic semantic tagger developed at the University of Lancaster – will be applied in order to assess whether it could fruitfully replace manual coding in establishing cultural associations of the given node words (R.Q. 4). More concretely, this chapter compares the results obtained by manual tagging (see 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 current chapter will analyse only the English elicited datasets.

As we have seen in the previous chapters, manual semantic tagging is not only time-consuming, but also highly demanding: it requires the work of at least two well trained coders, as well as an intense effort from each of them in terms of coherent and cohesive application of the given coding scheme. On the other hand, an automated coding procedure would reduce the number of coders to a single researcher, take only a few minutes, and guarantee effortless systematic application of the coding scheme.

In a preliminary experiment (Bianchi, 2010; see also Chapter 4), the *chocolate* and *wine* elicited datasets underwent automatic tagging using Wmatrix and the results of the automatic tagging were compared to manual coding at the level of conceptual domains (superordinate, broader categories) and of semantic fields, by applying the USAS-Codebook conversion scheme described further on in the current chapter. At the level of conceptual domains, the conversion scheme was applied to the top 30 items in the semantic frequency list and in the semantic keyword list of the elicited data as offered by Wmatrix, excluding grammatical items. As an intermediate step between manual tagging (sentence-based) and semantic tagging (word-based), it was decided to consider also the top 30 items of the raw frequency list and of the keyword list, as this allowed manual tagging to be applied on the basis of individual words. Therefore, the top 30 semantic items in the lists (excluding the node word) were manually mapped to one or more of the conceptual domains described in the Codebook. Those analyses were then compared to the results of manual coding of the whole elicited datasets, which showed that the semantic frequency list performed generally better than the other lists. In fact, it retrieved the same or a higher number of domains and systematically showed strong correlation values at the Spearman test. At the level of semantic fields, comparison was performed using the most frequent 50 items in the semantic frequency list and in the semantic keyword list. When using the semantic frequency lists, the data consistently showed levels of correlation in the modest range, with results for *chocolate* being $r = 0.505$ (at $p < 0.01$), and for *wine* $r = 0.558$ (at $p < 0.01$); when using the semantic keyword list, results were less consistent, with strong correlation results for *chocolate* ($r = 0.703$ at $p < 0.01$) and modest correlation results for *wine* ($r = 0.486$ at $p < 0.01$). Finally, the preliminary experiment compared the semantic word lists of the elicited data to the semantic word lists of the Web data. For the sake of experimentation, correlation was computed in three different ways: (1) using the whole semantic frequency lists, (2) using the top 100 items in the lists; and (3) using the top 50 items. All the six cases (three for *chocolate* and three for *wine*) showed interesting positive correlation between the elicited and the Web data, the strength of the correlation decreasing from strong to medium to low-medium as the number of items considered decreased.

The current chapter banks on results of the preliminary experiment described above and expands it in the following directions: 1. expanding the number of items considered in the semantic frequency list; 2. considering highly conventionalised fields/domains and cultural associations; 3. analysing prosody; 4. comparing the results to our ‘control situation’ – i.e. to the results obtained with manual coding of the whole elicited datasets. Furthermore, in Section 9.4, the results of automatic coding will be compared also to manual coding of the most frequent 150 words in the wordlist.

## 9.2 Matching automatic tagging categories to manual coding ones

For the purpose of comparing automatic tagging to manual tagging, automatic semantic tagging was applied to the English elicited data using Wmatrix and the USAS tagset (see Chapter 5, Section 5.3.2). The semantic structure adopted in the USAS tagset is rather different from the one developed and used in the manual tagging process. However, as we shall see in the following paragraphs, comparisons are still possible, by applying a conversion process similar to that used for matching the UCREL semantic taxonomy to that of the Collins English Dictionary (CED) and described by Archer, Rayson, Piao and McEnery (2004).

To allow comparison, the USAS tags were matched to the semantic fields used in the manual coding of the elicited data. For each tag, matching was accomplished by looking at the prototypical examples provided in Archer, Wilson and Rayson (2002), imagining them in the given context (i.e. next to the words *chocolate* and *wine*, but also in the wider context of general speech), and finding a suitable semantic field in the manual tagging list. Examples of matching are provided in Table 9_1.

In the table, the words or expressions specified in the manual coding columns refer to the Codebook semantic field; double slashes (//) indicate that matching is ‘one-to-many’. The word ‘Other’ indicates no matching. For the matching between Codebook semantic fields and conceptual domains, please see Table 2 in the Appendix.

| USAS tag | USAS semantic category | *Chocolate* manual coding | *Wine* manual coding |
|---|---|---|---|
| O4.6+ | Temperature: Hot/on fire | // Drink // Other | // Storage // Other |
| O1.1 | Substances and materials: solid | // Food // Other | // Food // Other |
| I2.2 | Business: Selling | Transaction | Transaction |
| X3.1 | Sensory: Taste | Taste | Taste |
| E2- | Dislike | Passion | Passion |
| L1+ | Alive | Existence | Existence |
| S3.1 | Personal relationship: General | Friendship | Friendship |
| A2.1+ | Change | Other | Other |
| A1.5.1 | Using | Other | Other |

Table 9_1. Conversion schemes: some examples

Different conversion schemes were necessary in order to account for the different fields of the two key words. For example, the elicited corpus showed that USAS tag O4.6+ (Temperature: Hot/on fire), which corresponds primarily to the word 'hot', tends to refer to different semantic fields when next to the word 'chocolate' or 'wine': if chocolate is hot, it is a drink; if wine is hot, we are talking about a storage issue. However, given that both chocolate and wine belong to the same general category of food and drinks, the two conversion schemes show a limited number of differences. A given USAS tag could match one or more categories of the manual codes, or even none of them. Matching was not sought for categories indicating logical or grammatical relations (Table 9_2). Indeed these categories were disregarded in all the analyses.

| Code | Description | Code | Description | Code | Description |
|---|---|---|---|---|---|
| Z4 | Discourse Bin | Z99 | Unmatched | A13.3 | Degree: Boosters |
| Z5 | Grammatical bin | A7 | Probability | A13.4 | Degree: Approximators |
| Z6 | Negative | A7+ | Likely | A13.5 | Degree: Compromisers |
| Z7 | If | A7- | Unlikely | A13.6 | Degree: Diminishers |
| Z7- | Unconditional | A13 | Degree | A13.7 | Degree: Minimisers |
| Z8 | Pronouns | A13.1 | Degree: Non-specific | A14 | Exclusivisers/particularisers |
| Z9 | Trash can | A13.2 | Degree: Maximisers | N1 | Numbers |

Table 9_2. Categories excluded from analysis

One of the major issues in matching two different schemes of this type is how to distribute frequency in the case of 'one-to-many' matching. In this study, when the matching scheme presented 'one-to-many' mapping (about 34% of cases for semantic fields and 30% of cases for conceptual domains, in both datasets), the frequency of the USAS tag was equally distributed among all of the possible matching domains/fields. So, for example the USAS conceptual domain SUBSTANCES AND MATERIALS: SOLID (78%) was equally distributed between Codebook domain FOOD (39%), and in category OTHER (39%). Though this clearly leads to an approximation, it seemed the only possible solution, since manual tags refer to the relationship that exists between the key word (*chocolate* or *wine*) and the rest of the sentence, while automatic tags describe individual words, regardless of the key word. Manually looking at individual concordances in order to recreate the relationship to the key word was discarded in this case, as the aim of the study is precisely to investigate and assess automated procedures.

## 9.3 Analyses

The top 50/100/150 items in the semantic frequency list of the English elicited datasets were compared to the results of manual tagging of the same datasets (see Chapter 6). The 150 limit was arbitrarily chosen considering that a semantic category conflates one or more words in the dataset. Consequently, the top 150 items in the semantic frequency list represent a percentage of the whole dataset which is certainly higher than that of the most frequent 150 words in the wordlist. Consequently, considering that in Chapter 7 over 90% of the highly conventional semantic fields appeared with as few as about 300 words, it seemed reasonable to hypothesise that an even smaller number of the most frequent semantic categories could be enough to highlight all or most of the cultural associations of the node words.

Comparison was performed both qualitatively, and quantitatively, at the level of semantic fields and conceptual domains. In other words, the most frequent 150 USAS tags, once converted into Codebook semantic fields, were compared to semantic fields Tables 6_1 and 6_8 and to conceptual domains Tables 6_4 and 6_11 in Chapter 6. The following paragraphs summarise the results of this comparison.

The results of these qualitative and quantitative comparisons between the most frequent 150 USAS categories in the English elicited datasets and manual semantic analysis of the datasets, at the level of semantic fields, are summarised in Tables 9_3 and 9_4. Column one shows the number of most frequent (Top) semantic tags considered; columns two reports the overall percentage of fields covered (with reference to tables 6_1 and 6_8). Columns three and four show the percentage of highly conventionalised fields (H Cnv) and cultural associations (H+M Cnv) covered. Column five summarizes field increases in passing from one threshold to the next. Finally, the last column reports the results of Spearman’s Rank Correlation test (for $p < 0.01$). Percentages are rounded to the second decimal.

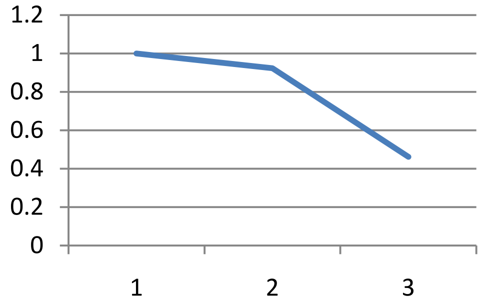
| Matched USAS fields | Codebook fields (%) | H Cnv (%) | H+M Cnv (%) | Field increase | Spearman’s rho |
|---|---|---|---|---|---|
| TOP 50 | 28.41 | 34.29 | 35.59 | + 26 fields | 0.505 |
| TOP 100 | 54.55 | 57.14 | 66.10 | + 24 fields | 0.503 |
| TOP 150 | 67.05 | 74.29 | 79.66 | + 12 fields | 0.492 |

Table 9_3. English *chocolate* elicited dataset: semantic field comparison

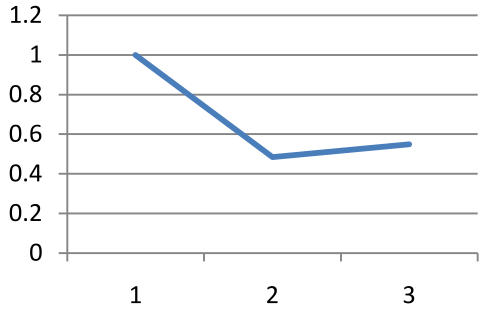
| Matched USAS fields | Codebook fields (%) | H Cnv (%) | H+M Cnv (%) | Field increase | Spearman’s rho |
|---|---|---|---|---|---|
| TOP 50 | 36.47 | 60.00 | 55.77 | + 31 fields | 0.558 |
| TOP 100 | 52.94 | 80.00 | 73.08 | + 15 fields | 0.584 |
| TOP 150 | 68.24 | 80.00 | 80.77 | + 17 fields | 0.525 |

Table 9_4. English *wine* elicited dataset: semantic field comparison

The most frequent 150 items in the USAS frequency list – which represent 56% of each list – showed about 67-68% of the Codebook fields highlighted with manual tagging. This corresponds to 74-80% of the highly conventionalised fields and about 80% of the cultural associations (fields with high or medium conventionalisation). Furthermore, as already noticed in Chapter 8, Zipf’s law does not seem to apply to field increases at different thresholds (see Graphs 9_1 and 9_2), below.



Graph 9_1. Data from Table 9_3



Graph 9_2. Data from Table 9_4

Finally, Spearman's test results are all in the modest range, a result which is similar to the one obtained in the preliminary experiment (Bianchi, 2010). Furthermore, differently from what noticed in Chapter 8, no increasing tendency can be seen when moving from one threshold to the next.

At the level of conceptual domains, the situation is summarised in Tables 9_5 and 9_6, below.

| USAS fields | Overall Codebook domains (%) | H Cnv (%) | H+M Cnv (%) | Domain increase | Spearman's rho |
|---|---|---|---|---|---|
| TOP 50 | 66.67 | 100 | 81.82 | + 10 fields | 0.810 |
| TOP 100 | 86.67 | 100 | 90.91 | + 3 fields | 0.881 |
| TOP 150 | 93.33 | 100 | 100 | + 1 fields | 0.904 |

Table 9_5. English *chocolate* elicited dataset: conceptual domain comparison

| Matched USAS fields | Overall Codebook domains (%) | H Cnv (%) | H+M Cnv (%) | Domain increase | Spearman's rho |
|---|---|---|---|---|---|
| TOP 50 | 66.67 | 100 | 90.00 | + 10 fields | 0.545 |
| TOP 100 | 80.00 | 100 | 100 | + 2 fields | 0.763 |
| TOP 150 | 93.33 | 100 | 100 | + 2 fields | 0.429 |

Table 9_6. English *wine* elicited dataset: conceptual domain comparison

The most frequent 150 items in the USAS frequency list – which represent 56% of each list – showed about 93% of the Codebook domains highlighted with manual tagging, and 100% of the highly conventionalised fields and of the cultural associations. The majority of domains entered the picture already in the top 50 items. Finally, Spearman's test results are in the strong range for *chocolate*, but in the modest range for *wine*. Furthermore, at least in the case of *wine*, Spearman's rho does not increase as the number of USAS fields considered increases.

As regards semantic prosody, i.e. when the semantic categories adopted for analysis fall into evaluative categories (see Chapter 3, Section 3.6.4), the USAS tagset includes a specific category (A5) subdivided into 4 subcategories: 'A5.1 Evaluation: Good/bad', 'A5.2 Evaluation: True/False', 'A5.3 Evaluation: Accuracy', and 'A5.4 Evaluation: Authenticity'. Within each category, plus (+) or minus (-) signs indicate positive or negative evaluation, respectively. In the most frequent 150 semantic items, this category appeared with a clear predominance of positive evaluation. In quantitative terms, *chocolate* showed 70 positive words vs. 30 negative ones, i.e. a positive evaluation which is about 2.3 times bigger than the negative one. *Wine* showed 184 positive words vs. 30 negative ones, with positive evaluation being about 6 times bigger than the negative one. These results are comparable to manual tagging of the two elicited datasets (our control situation) in qualitative terms, but not in quantitative ones (Chapter 6, Table 6_15). In fact, in the whole manually coded datasets, positive assessment was 2.8 times bigger than negative assessment for *chocolate*, and 2.4 times bigger for *wine*.

## 9.4 Concluding remarks

In the current chapter, the English elicited datasets were automatically tagged with Wmatrix, and the most frequent 150 items in the resulting semantic frequency lists were compared to the results of manual coding of the entire datasets, at the level of semantic fields, conceptual domains, and semantic prosody. Since the semantic structure adopted in the USAS tagset is rather different from the one developed and used in the manual tagging process, a conversion scheme was applied which matched the USAS tags to the semantic fields used in the manual coding of the elicited data.

At a qualitative level, the results are encouraging. In fact, comparison showed that the most frequent 150 items in the USAS frequency list – which represent 56% of each list – showed about 67-68% of the Codebook fields highlighted with manual tagging, and about 93% of the conceptual domains, including 74-80% of the highly conventionalised fields and about 80% of the cultural associations, and 100% of the highly conventionalised and cultural domains. Furthermore, the most frequent 150 USAS categories in the semantic frequency list showed marked preference for positive, rather than negative assessment, as was the case in the control situation.

From a quantitative perspective, correlation results assessed using Spearman’s test showed modest correlation for semantic fields and modest/strong correlation for conceptual domains. We must not forget, however, that the conversion procedure adopted introduced quantitative approximations. In fact, in about 34% and 30% of the cases, for semantic fields and conceptual domains, respectively, the frequency of the USAS tags considered was equally (and not proportionally) distributed among two or more Codebook semantic fields, which obviously influenced Spearman’s results.

Finally, the most frequent 150 USAS items in the semantic frequency list were compared to manual coding of the most frequent 300 words in the wordlist (see Chapter 7). At the level of semantic fields, manual tagging of the top 300 words in the wordlist provided better results than the procedure experimented in the current chapter, at both qualitative and quantitative levels. At the level of conceptual domains and semantic prosody, the two procedures seem comparable in terms of results at the qualitative level, but not at the quantitative one. At the level of conceptual domains, manual coding of the top items in the wordlist showed not only about 100% of highly conventionalised fields and of the cultural associations, but also strong/very strong correlations with the whole datasets. On the other hand, the top 150 items in the semantic frequency list recovered 100% of the highly conventionalised and cultural domains, but showed inconsistent correlation results (modest correlation for *wine* and strong for *chocolate*). Finally, the ASSESSMENT field is characterised in all the cases under analysis by prevalence of positive vs. negative assessment, but the proportion between the two types of assessment is markedly different (4.2 times bigger in the *chocolate* manually coded top 300 words; 2.4 times bigger in the *wine* manually coded top 300 words; 2.3 times bigger in the *chocolate* top 150 USAS tags; and 6 times bigger in the *wine* top 150 USAS tags).

It seems clear from the current results that the approximations involved in the application of the conversion scheme have variably influenced the quantitative comparisons. It is noticeable, however, that, despite approximations, the most frequent 150 semantic categories were able to retrieve over 70% of the high conventionalisation fields/domains and cultural associations, with the already noticed 'improvement' in the number of semantic categories when passing from a more detailed coding scheme to a less detailed one.

Finally, comparison of the most frequent 150 USAS items in the semantic frequency list to manual coding of the most frequent 300 words in the wordlist suggests that, at least for small corpora, such as the elicited ones used in the current work, using an automatic semantic tagging tool is worth only if the tagging semantic categories can be used without further conversion. The case is likely to be different when using larger corpora. In fact, if we consider that both procedures are sensitive to corpus size, when working with very large corpora, the top N items in the semantic frequency list would be more representative of the overall corpus that the top N words in the frequency list.

Chapter 10

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

## 10.1 Introduction

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

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

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

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

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

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

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

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

## 10.2 Sampled Web sub-corpora: creation and coding

As already noted in Chapter 8, with small sized datasets such as the elicited ones, random samples in the 25-35% size range of the original dataset provided very good results. Consequently, I decided that 25% could be a more than suitable sampling limit for the sampling of the much larger Web data.

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

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

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

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

### 10.3 Inter-culture analysis

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

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

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

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

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

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

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

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

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

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

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

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

### *10.3.1 Semantic field ASSESSMENT*

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

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

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

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

Table 10_3. *Chocolate*: Semantic field ASSESSMENT

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

Table 10_4. *Wine*: Semantic field ASSESSMENT

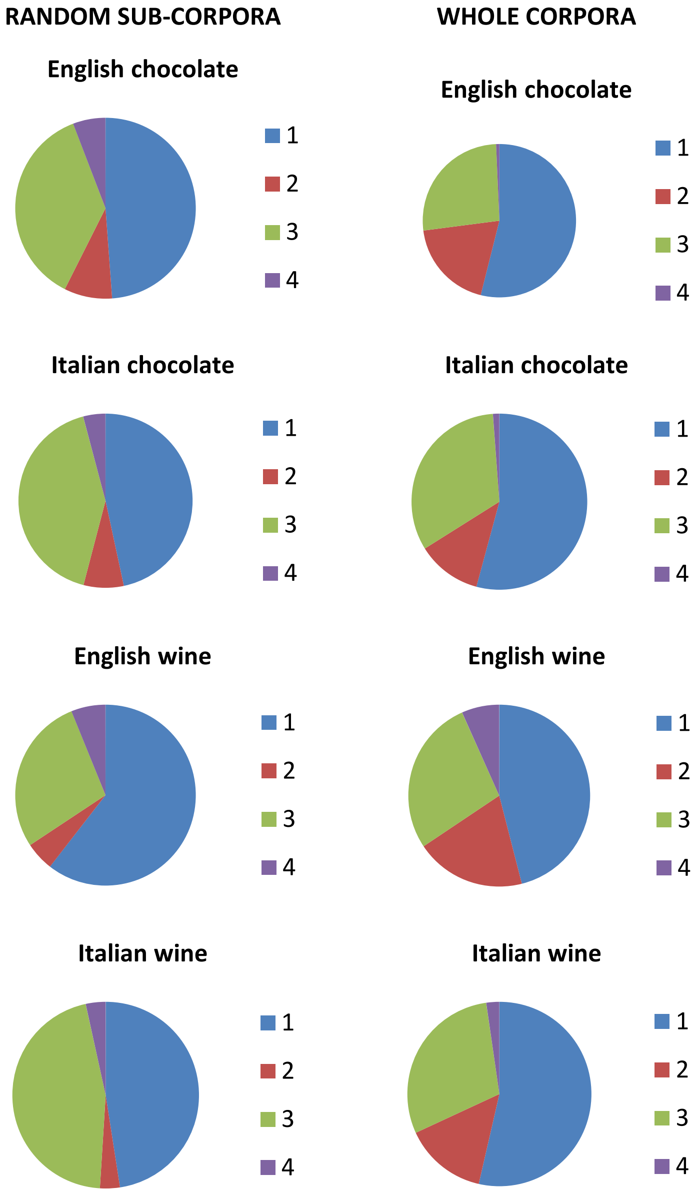


Figure 10_1. ASSESSMENT field results:
random sub-corpora vs. whole elicited datasets

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

## 10.4 Cross-cultural comparison

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

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

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

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

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

## 10.5 Concluding remarks

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

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

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

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

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

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

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

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

Chapter 11

# Conclusion

## 11.1 Introduction

The current chapter reviews the various phases of this work, and summarises the results of the analyses. Furthermore, it discusses the limitations of the current approach, and suggests possible extensions and ideas for future work. Despite its several limitations, the work may be considered to contribute to the current state of knowledge and research in corpus linguistics and cultural studies in several ways which will be illustrated in a concluding section.

The following paragraphs review the aims, theoretical background, and research questions that guided the investigation.

In a general attempt to contribute to our understanding of cultural systems, and of the relationship between text, semantics, and culture, I selected and outlined two models of culture which lend themselves to semantic as well as quantitative analyses; these are the systemic models by Fleischer (1998) and Nobis (1998), described in Chapter 2. In particular, according to Fleischer, discourse, i.e. the linguistic level at which culture shows itself and develops, is characterised by symbols. In turn, symbols are composed of three elements: the core, which is a stable semantic element that is shared by all members of the cultural community; the current field, a semantic element which is shared by several, though not all, members of the community but which is spreading; and the connotational field, a semantic element that is specific to single individuals. Both core and current fields are expressions of cultural meanings and can be identified by analysing the frequency and distribution of the semantic associations of a given concept/word across the members of the cultural community. Consequently, Fleisher's theory will help us establish the level at which a concept (expressed by a word) is rooted (or anchored in Fleisher's terminology) in a given culture at a given point in time. Nobis' theory on the other hand will help us compare two cultures in terms of their relative development with reference to the same symbol. In fact, Nobis banks on the generalized systemic idea that systems are in constant tension between stability and evolution – the latter being achieved by transmission of behaviours (including mental behaviours) – and suggests that transmission of behaviour may only take place when that behaviour has a long established network of relations with other behaviours, i.e. a stable behavioural pattern. Nobis' notion of 'stable behavioural pattern' can easily be equated with Fleischer's notion of conventionalisation: a mental behaviour, such as thinking of a concept, has features that, at a given point in time, are widely shared by all members in a community/culture. Consequently, if two cultures show relevant differences in the number of the concepts connected to the same key concept we can hypothesise different stages of knowledge/acceptance of the key concept. These theories were tested on two concepts, *chocolate* and *wine*, in the English and Italian cultures. These concepts were selected, among other considerations, because one of them (*wine*) could be expected to have different rooting in the two cultures, while the other (*chocolate*) to have similar rooting, given the two countries' climatic conditions and food production histories.

Furthermore, a review of semantic approaches to the study of culture in different disciplinary areas, including linguistics, anthropology, and psychology (Chapter 2), suggested elicited data and non-elicited data as equally possible materials for cultural analysis. The same review also suggested that corpora and quantitative analytical methods are easily applicable to this purpose. In particular, the most frequently used analytical methods seem to revolve around the use of frequency lists, keyword and collocate frequencies, as well as grouping of items to create superordinate domains (either entirely semantic or thematic). These topics, along with other major issues in corpus linguistics such as the use of the Web as a source for corpus data, were detailed and discussed Chapter 3.

Finally, given my desire to perform an analysis of cultural associations of a given concept which could find theoretical or practical applications not only in the linguistic and cultural fields, but also in the marketing one, Chapter 4 overviewed the materials and methods most frequently used in marketing research, reviewed selected marketing and consumer studies where analysis of linguistic data is performed, and established some methodological common ground among linguistics, cultural studies, and marketing. Such common ground can be summarised in the following features: use of elicited data, but also of Web data; analysis of word associations; semantic/content analysis; and frequency as a measure of the association's importance.

The theoretical and methodological elements outlined above provided the framework for the experimental part of the work. This part of the work – which focussed on the development of a suitable analytical method to establish and compare the cultural mental associations of *chocolate* and *wine* in Great Britain and Italy, and on testing different types of datasets for cultural analysis – was operationalised in five Research Questions. Research Questions 1 and 2 – What are the semantic associations of chocolate, and wine in the Italian and English cultures?; and What are the differences between the Italian and English cultures with reference to chocolate, and wine?, respectively – were addressed in Chapter 6, and the results of that Chapter are summarised in Section 11.2.1 below. Research Question 3 – Could we identify the cultural associations of the two words without coding the entire dataset? – was addressed in Chapter 7 (on elicited data) and Chapter 8 (on Web data); the results of the analyses performed are summarized and discussed in Section 11.2.3, below. Research Question 4 – Could we identify the cultural associations of the two words using an automatic semantic tagger? – was addressed in Chapter 9 and is summarised in Section 11.2.4, below. Finally, Research Question 5 – Could we identify the cultural associations of the two words using a general (Web) corpus? – was addressed in Chapter 10, and the results are discussed in Section 11.2.2, below. Please notice that in the current chapter the sections have been organized in a way that is slightly different from the progression of the research questions. This was done because, at this stage, it seemed useful to highlight the cultural part of the work separately from the purely methodological one.

## 11.2 Summary of the experimental work and results

The current experimental work can be divided into two logical parts. The first part is concerned with finding the most suitable materials and methods for the analysis of linguistic data to highlight cultural features. This has been subdivided here into two separate sections, summarising the results of elicited data analysis (Section 11.2.1) and of Web data analysis (Section 11.2.2), respectively. The second logical part is concerned with methodological issues, such as testing different sampling procedures in order to avoid having to code large datasets (Section 11.2.3), and using an automatic semantic tagger in place of manual coding of the data (Section 11.2.4).

### *11.2.1 Retrieving cultural associations in elicited data (control situation)*

The current work used elicited data on *chocolate* and *wine*, gathered through free sentence-completion and sentence-writing tests in English and Italian, to highlight the cultural associations that each key word has in the cultures considered. The elicited data were manually analysed using content analysis procedures (i.e. semantic coding), and the semantic categories that emerged from the content analysis were quantitatively measured in terms of overall frequency, as well as frequency distribution across subjects – the latter being calculated by applying Molinari's evenness index. Furthermore, by looking at the position of the evenness index with reference to the confidence interval, it was possible to establish the level of conventionalisation of the various fields and domains, in each culture and for each node word. Three conventionalisation levels were considered: high, medium, and low, respectively corresponding to Fleischer's core, current and connotational fields. Finally, the results of the Italian and English datasets were compared by the Welch *t* Test for Independent Samples.

In keeping with expectations, *chocolate* appeared as an equally long- and well-established symbol in the two cultures; on the other hand, *wine* – though well-established in both countries – showed longer rooting in the Italian culture, as the Italian respondents' answers showed a remarkably higher percentage of highly conventionalised semantic fields and domains, and a remarkably lower percentage of low conventionalisation fields and domains than the British ones.

In the light of the current experiments on elicited data, the Italians seem to distinguish themselves from the British for their more frequent matching of *chocolate* to the following concepts: BAKERY/COOKING; RECIPE; DIETING; MEDICINE; BEAUTY; HISTORY; NICE/PLEASANT/PLEASURE; CHILDREN; FAMILY; STUDYING/INTELLECT; QUALITY/TYPE; GENUINE. On the other hand, more prominent for the English than for Italians seem to be: WOMEN; and PRICE.

As regards *wine*, the following semantic fields emerged as distinctively more prominent for the Italians than for the English: BAKERY/COOKING; EVENT; WOMEN;

NATURE; ARTISTIC PRODUCTION; QUALITY/TYPE; QUANTITY; GENUINE; PRICE. On the other hand, more prominent for the English than for the Italians were: PRODUCT/SHAPE; DRINK; MANUFACTURING; RECIPE; LANGUAGE; CONFIDENCE; DESIRE; NICE/PLEASANT/PLEASURE; MEN; FRIENDSHIP; POSH; SHARING/SOCIETY; PEOPLE; and STUDYING/INTELLECT.

As we have argued in Chapter 6, this is only a list of the mental associations in which the English culture seems to differ from the Italian one. Neither the qualitative nor the quantitative analyses performed in this work can in any way explain the type of the association or the reasons for the differences. Further steps, such as analysis of individual concordance lines, are needed to understand the exact link between key word and semantic field in each culture. Such analyses are beyond the scope of the current investigation, but will be considered in future extensions of this work. Nevertheless, I believe that analyses of this type may be adopted in the exploratory phases of marketing (or cultural) research, where research aims to outline problems, collect information, eliminate impractical ideas, and formulate hypotheses.

### *11.2.2 Comparing Web data to the control situation*

The elicited data – considered as the control situation – were compared to non-elicited sentences on *chocolate* and *wine* from general Web corpora in English and Italian.

The Web corpora – analysed through randomly sampled subsets of about 1800-2000 sentences – retrieved over 90% of the semantic fields with high conventionalisation and of the cultural associations attested in the corresponding elicited datasets, and 100% of the domains. However, the corpora also retrieved most of the low conventionalisation fields, along with a few extra fields whose conventionalisation level is not known (although one could speculate that – being those fields totally absent in the elicited corpora – they could be classified as having low conventionalisation). The Web results were quantitatively compared to the elicited ones by means of Spearman's Rank Correlation Coefficient and showed modest correlation at the level of semantic fields, and strong correlation, or higher, at the level of conceptual domains. Furthermore, no marked and systematic differences can be seen between the results of the English data vs. those of the Italian data.

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

Consequently, despite our initial fears that issues such as uncontrolled authorship and readership (see Chapter 3) could represent a bias in the use of English Web data, comparisons between the Web corpora and the elicited data suggest that large general Web corpora can be considered representative of the cultural associations of a node word. In fact, randomly sampled Web subsets of only 1800-2000 sentences, included all the relevant cultural associations of the node word.

Furthermore, when the coding scheme adopted was broad and included few categories, the general Web corpora appeared to be representative not only at a qualitative level, but also at a quantitative one.

Unfortunately, as argued in Chapter 6, we cannot rely on frequency alone to establish conventionalisation. Only the very highest ranks in the frequency list are systematically occupied by low conventionalisation fields, and only the very lowest ranks are systematically occupied by high conventionalisation ones. Any other position in the list can hardly tell us something about conventionalisation level. Consequently, if we had only Web data, and no control elicited data, we would have to assess the conventionalisation level of each field/domain by applying an evenness index, in order to establish which of the retrieved semantic fields/conceptual domains can be safely considered cultural associations. Fundamental pre-requisite for applying the evenness computation is the possibility to group the Web sentences according to subject/author or website. This – along with T-test analyses for cross-cultural comparisons – could not be done in the current work, because at the time when the Web data were retrieved, the Sketch Engine did not provide information about the website each text was taken from. The updated version of Sketch Engine, however, does provide this type of information, and its users can now benefit from the possibility to assess the distribution of concordance lines across Web sites (i.e. authors).

### *11.2.3 Testing different procedural approaches*

The current work experimented different procedural approaches. In particular, focus was on finding an alternative route to manual coding of the whole dataset, as this is a costly and complex procedure when the number of sentences in the dataset is very high. The various procedures were all tested on the elicited datasets, while only the procedures that had showed better results were applied to the Web datasets .

One of the procedures adopted was random sampling of the sentences in the dataset, a rather standard procedure to create smaller, but representative sub-sets. 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.

The other two procedures were based on analysis of a limited number of the most frequent words in the datasets. These less standard procedures were inspired by previous linguistic studies of culture and by Fleischer's theories which suggest the existence of a relationship between cultural associations, their level of conventionalisation and frequency of occurrence of the given associations. This led me to testing the following two possibilities: 1. performing manual semantic analysis of 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; and 2. using the four most frequent content words to extract sentences from the manually coded dataset and create a sampled sub-corpus.

The random sampling technique proved to be the most representative route, as it systematically showed higher results that the others at all levels of analysis, including separate analysis of semantic field ASSESSMENT. In fact, 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, with correlation results falling in the very strong range. Furthermore, this procedure showed qualitative and quantitative results that were perfectly comparable to those of the whole dataset as regards analysis of the ASSESSMENT category.

The other two techniques provided interesting results at both qualitative and quantitative level, but not when it came to analysing semantic field ASSESSMENT. In fact, 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, with correlation results in the strong range. 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, with correlation results in the strong-very strong range. Separate analysis of the ASSESSMENT category, however, showed qualitative and quantitative results that were not comparable to those of the whole dataset.

### *11.2.4 Testing the use of an automatic semantic tagger*

Finally, an automatic semantic tagger (Wmatrix/USAS tagset) was tested on the elicited data, in order to assess the extent of its possible application in cultural analysis. The automatic semantic tagger was used in two different scenarios: 1. an ‘autonomous’ scenario, where the USAS tagset was automatically applied to the elicited and Web datasets and the results of the tagging process were compared to the most frequent content words in their wordlists, and to sub-corpora randomly sampled from the same datasets; and 2. a ‘comparative’ scenario, where the USAS tags retrieved in the datasets were converted into Codebook tags and results were compared to those of manual coding.

In the ‘autonomous’ scenario, it was noticed that by applying USAS tagging the most frequent content words in the wordlists and the sub-corpora randomly sampled from the datasets both retrieved a smaller percentage of semantic fields than with manual coding. This is a consequence of the very high granularity of the USAS dataset which – with all its categories and subcategories, as well as ‘pluses’ or ‘minuses’ to indicate a positive or negative position on a semantic scale – includes almost 400 different labels. Interestingly, however, although the percentage of semantic fields retrieved was lower when USAS tagging was applied (about 44.5% vs. 68-70% when analysing the most frequent words in the wordlist; 66.79% vs. 84.09% for *chocolate*, and 80.36% vs. 86.9% for *wine*, when analysing the random sub-corpus), in both cases Spearman’s test results were very similar to those obtained with manual tagging, the correlation index being always in the strong range. Finally, when the USAS tagset was applied to the Web corpora, it became evident that the number of most frequent semantic words in the wordlist necessary to highlight the most frequent semantic associations of the node word depends on corpus size. In fact, the top 300 word retrieved only about 25% of the semantic fields. Even extending to 450 the number of words considered, the percentage of fields retrieved was still very low (about 31%). However, correlation values were in the medium range and showed constant linear increase, thus strengthening the hypothesis that the most frequent words in corpus are highly representative of the contents of the corpus.

In the 'comparative' scenario, the English elicited datasets were automatically tagged with Wmatrix, and the most frequent 150 items in the resulting semantic frequency lists were compared to: A. the results of manual coding of the entire datasets, and B. manual coding of the most frequent 300 words in the wordlist. In both cases, a conversion scheme which matched the USAS tags to the semantic fields used in the manual coding of the elicited data was applied. Comparison showed that the most frequent 150 items in the USAS frequency list – which represent 56% of each list – showed about 67-68% of the Codebook fields highlighted with manual tagging, and about 93% of the conceptual domains, including 74-80% of the highly conventionalised fields and about 80% of the cultural associations, and 100% of the highly conventionalised and cultural domains. Furthermore, the most frequent 150 USAS categories in the semantic frequency list showed marked preference for positive, rather than negative assessment, as was the case in the control situation. From a quantitative perspective, correlation results assessed using Spearman's test showed modest correlation for semantic fields and modest/strong correlation for conceptual domains. This is most certainly due to the quantitative approximations adopted in the conversion procedure. In fact, in about 34% and 30% of the cases, for semantic fields and conceptual domains, respectively, the frequency of the USAS tags considered was equally (and not proportionally) distributed among two or more Codebook semantic fields, which obviously influenced Spearman's results. Finally, the most frequent 150 USAS items in the semantic frequency list proved to be less representative of the whole dataset than manual coding of the most frequent 300 words in the wordlist.

## 11.3 Some methodological considerations

Methodological issues were a major concern in the current work from the very begging. A summary of the methodological considerations emerging from the investigation is provided in the following paragraphs.

The current research confirms that, alongside elicited data, which are a typical source of linguistic material in marketing research, the Web can be a useful source of data for analysing cultural associations of a given word or concept. The current work tested freely available large general Web corpora, from which sentences containing the word under analysis were extracted. In the current research, it was not possible to compute evenness measures in the Web corpus, and the results could only be interpreted by comparing them to the elicited data. The comparison, however, showed that a small random sample of the Web data included all the relevant cultural associations of the node word. This leads to believe that if the Web data are collected in a way which allows the researcher to group the Web sentences according to subject/author or website, the Web data could be interpreted regardless of the presence of 'control' data.

Furthermore, the research suggests that a relatively small number of sentences including the given key word is sufficient to understand its cultural associations. In fact, the current research tested eight random sub-sets sized 20-30% of the original datasets, and including from about 400 to about 2000 sentences. Each of them retrieved about or over 95% of the cultural associations. This procedural finding is particularly relevant when dealing with a very large dataset including several thousand sentences. Indeed, manual coding is not only time-consuming, but also highly demanding: it requires the work of at least two well-trained coders, as well as an intense effort from each of them in terms of coherent and cohesive application of the given coding scheme. And the larger the dataset, the greater the effort in applying the scheme consistently and coherently. Also an analysis of the most frequent words in the wordlist could, if necessary, be employed as an alternative route to tagging the whole corpus, bearing in mind – however – that the number of words to consider depends on the size of the original corpus and that this procedure introduces approximations. The effectiveness of these two procedures are easily explained by the fact that cultural associations emerge from a combination of frequency and spreading across a large number of subjects.

Finally, if a suitable automatic semantic tagger is available, quick and consistent semantic analysis of the whole corpus can be easily obtained, and the cultural associations can be identified by looking at the most frequent semantic categories in the corpus. However, if the corpus under analysis is small, such as the elicited ones used in the current work, the use of an automatic semantic tagging tool is recommended only if the semantic categories of the automatic tagging can be used without further conversion, since conversions introduce approximations.

As regards approximations, however, it must be said that, as Hubbard (2010, p. 23) clarifies, measurement is “a quantitatively expressed reduction of uncertainty based on one or more observations” and, in many circumstances, having even an approximate idea of the variables and their values represents a big leap ahead from our original level of knowledge about the given object. This is indeed the case of exploratory market research, where – as we have seen – the researcher’s aim is to acquire an inexpensive approximation and uses it to outline problems, eliminate impractical ideas, and formulate hypotheses.

A further consideration regards tagset granularity. In the current work, three different tagsets were applied to the same data: the Codebook semantic field tagset, including 96 semantic fields; the Codebook conceptual domain tagset, with its 16 conceptual domains; and the USAS tagset which includes almost 400 different tags. The Web data, as well as the elicited wordlists and sampled sub-corpora were compared to the ‘control’ data after applying each of the three tagsets. Throughout the work it consistently appeared that when passing from a more detailed to a less detailed tagset (e.g. semantic fields vs. conceptual domains; USAS tagging vs. manual tagging), semantic category coverage increased, and also correlation values increased. This is in keeping with observations by Guerrero, Claret, Verbeke *et al.* (2010, reviewed in Chapter 4), who applied to their data a double grouping process with categories which are comparable to our semantic fields and conceptual domains and noticed that greater differences between cultures appeared at the level of semantic fields. Furthermore, I also share their considerations about the advantages and limits of grouping semantic categories into conceptual domains, when they say that its main advantage

> "is its simplicity, although the double grouping process increases the subjectivity of the results obtained. In addition, some difficulties may be observed when trying to obtain a reduced number of classes because it was not always easy to group the different classes together under a common dimension or concept. It is also important to notice that using this approach the more subtle differences between regions may disappear" (*ibid.*, p: 230).

I would add, however, that the grouping of semantic fields into conceptual domains facilitated choosing one semantic field over another when performing manual tagging. So I would suggest creating and using a two level tagging scheme when coding the data, but limiting analyses to the more fine-grained level in the tagset.

Finally, a look at the semantic fields which are absent with reference to both key words in the same culture suggests that field presence/absence may depend on the key word, rather than the culture. In fact, only one field is systematically absent in the English datasets (COMPETITIVENESS), and none in the Italian ones. It must be remembered that, in the current work, the overall numer and range of fields and domains emerged from the data themselves. Consequently, absence is relative to the coding scheme; any semantic field which does not appear in the Codebook could be considered 'absent' in both cultures and for both node words. Nevertheless, I would tentatively declare that this finding supports the use of dedicated coding systems for different node words. However, the relationship that links presence/absence of a semantic field and culture requires further investigation, on a much wider number of node words.

## 11.4 Limitations of current work and future directions

The current work has some limitations and possible directions for development have already been identified.

A major limitation derives from not having controlled the composition of the two population samples when collecting the elicited data. Although the English and Italian groups of respondents show some overall similarities (a majority of university students in the 18-25 age range; data collected in both Northern and Southern areas of the two countries), no precise data was available in the current research as regards variables such as the respondents' age, gender or occupation. The fact that the elicited data and the Web data analysed in Chapter 10 showed similar results, to some extent confirms similarity between the two population samples. Further confirmation could be found by applying, one or more of the following:

1. Replication of the study, possibly also with a larger sample size and/or more stratified random sampling.
2. Other elicitation methods (e.g. story writing).
3. Depth interviews and focus groups, possibly with deliberate attempts to elicit and probe the concepts that showed cultural differences (e.g. ask Italian and English respondents deliberately about women and chocolate and see if there is a difference in how they talk about the subject).

4. Content analysis (visual as well as verbal) of representative samples of chocolate/wine advertising from UK and Italian companies addressing the local audience.

For the time being, we will have to accept these results as they are. Should further research disconfirm this cultural comparison and cast doubts on the frequency-plus-T-test method adopted here, nevertheless, the methodological investigations performed in comparing different types of data and/or coding schemes will still be valuable.

Next, the analyses performed in this work highlight the semantic categories which are culturally connected to the given key words, but do not allow the researcher to understand the kind of relation that exists between the category and the key word. For example, when Italians talk about *children* and *wine*, what exactly do they refer to: that wine can or cannot be given to children? A possible way to answer this question would be to look at concordance lines. An analysis of the concordance lines of each semantic category could represent an interesting extension to the current work, and might provide greater insight into cultural specificities.

Third, the analysis of the ASSESSMENT field performed in this work, although clearly limited in scope, was sufficient for the purposes of the current work and was a suitable reference term for the methodological comparisons which were performed in the various chapters. However, from the perspective of cultural and even more so consumer research, the current level of analysis of semantic prosody appears excessively limited. We may expect different cultures to focus on different features when positively or negatively assessing a concept. Furthemore, we noticed in Chapter 4 how sentiment analysis plays a fundamental role in marketing research. Opinionated text may, for example, orient consumer behaviour when purchasing products, warn marketing managers about the rising of critical situations, or help establish the pricing power of a product feature. Consequently, in order for assessment analysis to find any application in marketing, even in exploratory phases, it needs to investigate further factors such as the reasons behind positive/negative evaluation, and the features of the concept/product which triggered the evaluation. With reference to the procedures and tools adopted in the current work, an extension of the analysis of the ASSESSMENT field could see the application of the following analytical methods: 1. looking at the distribution of the Positive and Negative categories across the various fields/domains;[1] 2. analysing the evaluative adjectives that collocate with the two selected key words;[2] 3. retrieving key relations between words, such as "the attributes assigned to various persons or things, and the various modifying and negating words and phrases associated with these" (Wilson, 1993, p. 6).[3]

Fourth, in the current research, it was not possible to compute evenness measures in the Web corpus, and the results could only be interpreted by comparing them to the elicited data. The comparison, however, showed that a small random

[1] A quick look at the data suggests that, when performing this type of analysis, it will be important to consider only the semantic fields/domains which show a minimum number of hits, alongside a significant a difference between Positive/Negative Assessment.

[2] Methodological inspiration could be taken from the works by Baker (2006), reviewed in Chapter 2, and Aggarwal, Vaidyanathan and Venkatesh (2009), reviewed in Chapter 4.

[3] In previous versions of the Wmatrix system, features for the automatic retrieval of key relations were available (Wilson 1993). Unfortunately, in the on-line version these features are no longer available.

sample of the Web data included all the relevant cultural associations of the node word. This leads to believe that if the Web data are collected in a way which allows the researcher to group the Web sentences according to subject/author or website, the Web data could be interpreted regardless of the presence of ‘control’ data.

Furthermore, the current analyses concerned two consumables. It might be interesting to test the procedures and automatic semantic tools described in this work on other topics of cultural and/or marketing interest. A relevant candidate is certainly node word *traditional* in food talk, since the results could be compared to those of Guerrero, Claret, Verbeke *et al.* (2010) in their study on the perception of traditional food products (see Chapter 4). Another node word that comes to mind is *flexibility*, as this word seems to have different semantic prosodies in the Italian and Anglo-American cultures: Italians seem to praise ‘flexible procedures’ (almost an oxymoron for the English or Americans), and strongly oppose ‘flexibility’ when it refers to the need to adapt themselves to a constantly changing job market.

Finally, in the current work experimentation with automatic tagging was possible only for English, since no semantic tagger based on a coding scheme similar to that of Wmatrix exists for Italian. Currently, a Finnish and a Russian version of the USAS tagset exist, alongside the English one.[4] It would be interesting to develop an Italian USAS tagset and test it on the *chocolate* and *wine* data.

## 11.5 Contribution to knowledge and concluding remarks

Despite its clear limitations, the current work can be considered to contribute to knowledge in several ways.

First of all, the current work is characterized by an interdisciplinary perspective which links linguistics, marketing research and cultural studies. The combination of the three fields seems innovative and certainly provides interesting methodological as well as theoretical ideas from which all the three disciplines could benefit.

Second, the quantitative comparisons between the entire datasets (elicited as well as Web) and smaller samples of the data accomplished in this work add useful pieces of information to our general knowledge in corpus linguistics.

Third, the procedure adopted to establish the cultural associations of the key words was specifically developed after careful analysis of similar experiments described in the scientific literature of different disciplines (linguistics, cultural studies, and consumer research) and in the light of the cultural systems theories by Fleischer (1998) and Nobis (1998). In particular, Fleischer (1998) suggests a quantitative type of analysis based on frequency and spreading of specific individual phenomena. Consequently, I believe that the procedure adopted in the current work represents an improvement to previous frequency-based measurements of cultural semantic associations. In fact, mere observation of raw or mean frequency of fields and conceptual domains provides an approximate picture of the semantic associations, as it disregards distribution of answers across subjects. On the other hand, the use Molinari’s evenness index – inspired by Wilson and Mudraya (2006), but here applied in a different way – introduces a quantification of spreading. Confirmation of the

[4] See http://ucrel.lancs.ac.uk/usas/

validity of the procedure applied comes from the fact that the results are in keeping with expectations.

Finally, this work is among the rare applications of Fleisher's theory of culture, and provides results which seem to support the theory. Furthermore, Nobis' system theory is also confirmed. In fact, the *wine* experiment clearly confirmed that longer standing of a concept (*wine*) in a given culture (Italy) corresponds to stronger cultural rooting (Nobis, 1998), here expressed in terms on higher percentage of highly conventionalised semantic fields. The second of Nobis' hypotheses, postulating greater semantic complexity of longer standing concepts, is supported in the *wine* experiment not by the overall number of semantic fields associated to the given concept, but by the greater number of semantic elements which are shared by several respondents, i.e. those semantic fields or conceptual domains with high level of conventionalisation. These two system theories, though still little known among linguists and consumer researchers, have much to offer to cultural analysis. Furthermore, they lend themselves to quantitative research and, thus, to corpus linguistics.

To conclude, I believe that the current work has been rather successful in its aim to contribute to our understanding of cultural systems, and of the relationship between text, semantics, and culture. Furthermore, it provides theoretical as well as practical ideas for improving cultural analysis through language.

All the three main areas of studies considered in this interdisciplinary research may benefit from its theoretical reviews and discussions and the results of its analyses. In particular, I believe that analyses of the types performed in the current work could be adopted in the exploratory phases of marketing or cultural research, where research aims to outline problems, collect information, eliminate impractical ideas, and formulate hypotheses.

# References

Aarts, B. (2000). Corpus linguistics: Chomsky and fuzzy tree fragments. In C. Mair, & M. Hundt (Eds), *Corpus linguistics and linguistic theory* (pp. 5-13). Amsterdam/Atlanta: Rodopi.

Aggarwal, P., Vaidyanathan, R., Venkatesh, A. (2009). Using Lexical Semantic Analysis to Derive Online Brand Positions: An Application to Retail Marketing Research. *Journal of Retailing*, 85, 2, 145–158.

Applefield, J.M., Huber, R., & Moallem M. (2001). Constructivism in Theory and Practice: Toward a Better Understanding. *The High School Journal*, 84, 2, 35-53. Retrieved October, 5, 2012, from www.davidvl.org/250CourseSpr04/b67.pdf.

Archak, N., Ghose, A., & Ipeirotis, P.G. (2011). Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science*, 57, 8, 1485–1509.

Archer, D., Rayson, P., Piao, S., & McEnery, T. (2004). *Comparing the UCREL semantic annotation scheme with lexicographical taxonomies. Proceedings of the EURALEX-2004 Conference* (pp. 817-827). Lorient, France.

Archer, D., Wilson, A., & Rayson, P. (2002). *Introduction to the USAS category system*. *Benedict project report*, October 2002. Retrieved May, 20, 2009, from http://ucrel.lancs.ac.uk/usas/usas%20guide.pdf.

Ares, G., Deliza, R. (2010). Identifying important package features of milk desserts using free listing and word association. *Food Quality and Preference*, 21, 621-628.

Ares, G., Giménez, A., Gámbaro, A. (2008). Understanding consumers’ perception of conventional and functional yogurts using word association and hard laddering. *Food Quality and Preference*, 19, 636-643.

Arndt, S., Turvey, C., Andreasen, N.C. (1999). Correlating and predicting psychiatric symptom ratings: Spearman's r vs. Kendall's τ correlation. *Journal of Psychiatric Research*, 33, 97-104.

Aston, G. (1988). *Negotiating Service: Studies in the Discourse of Bookshop Encounters*. Bologna: CLUEB.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London-New York: Continuum.

Baldry, A.P., Beltrami, M. (2005). The MCA Project: concepts and tools in multimodal corpus linguistics. In M. Asplund Carlsson, A. Løvland & G. Malmgren (eds.), *Multimodality: Text, culture and use. Proceedings of the Second International Conference on Multimodality* (pp. 79-108). Kristiansand: Agder University College/Norwegian Academic Press.

Banks, D. (2005). The case of Perrin and Thomson: An example of the use of a mini-corpus. *English for Specific Purposes,* 24, 2, 201-211.

Barnbrook, G. & Sinclai,r J. (1995). Parsing Cobuild entries. In J. Sinclair, M. Hoelter, & C. Peters (Eds), *The language of definitions: The formalization of dictionary definitions of natural language processing* (pp. 13-58). Luxemburg: Office for Official Publications of the European Communities.

Baroni, M. (2004). Part-of-speech Tagging (e lemmatizzazione). [Part-of-speech Tagging (and lemmatization)]. Retrieved January, 20, 2010, from http://sslmit.unibo.it/~baroni/compling04f/pos.pdf.

Baroni, M., Bernardini, S. (2003). The BootCaT Toolkit. Simple Utilities for Bootstrapping Corpora and Terms from the Web. Readme file of the Bootcat suite. Retrieved November, 25, 2007, from http://sslmit.unibo.it/~baroni/Readme.BootCaT-0.1.2.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2008). The WaCky Wide Web: A collection of very large, linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation*, 43, 3, 209-226.

Baroni, M., Kilgarriff, A. (2006). Large linguistically processed web corpora for multiple languages. *Proceedings of EACL*. Trento. Retrieved September, 12, 2007, from www.aclweb.org/anthology-new/E/E06/E06-2001.pdf.

Baroni, M., Kilgarriff, A.. Pomikálek, J., Rychlý, P. (2006). WebBootCat: a Web Tool for Instant Corpora. In *Proceeding of the EuraLex Conference 2006*. Vol. 1 (pp. 123-132). Trento, Italy: Edizioni dell'Orso.

Baroni, M., Ueyama, M. (2006). Building general- and special-purpose corpora by Web crawling. In *Proceedings of 13th NIJL International Symposium* (pp. 31-40). Retrieved December, 15, 2009, from http://clic.cimec.unitn.it/marco/publications/bu_wac_kokken_formatted.pdf.

Bassnett, S. (1991). *Translation studies*. London: Routledge.

Beisel, J.N., Usseglio-Polater, P., Bacmann, V., & Moreteau, J.C. (2003). A Comparative Analysis of Evenness Index Sensitivity. *International Review of Hydrobiology*, 88, 1, 3-15.

Bednarek, M. (2008). Semantic preference and semantic prosody re-examined. *Corpus Linguistics and LinguisticTheory*, 4, 2, 119-139.

Belk, R.W. (1985). Materialism: Trait Aspects of Living in the Material World. *Journal of Consumer Research*, 12, 265-280.

Bernardini, S., Baroni, M., Evert, S. (2006). A WaCky Introduction. In M. Baroni, & S. Bernardini (eds.). *WaCky! Working papers on the Web as Corpus* (pp. 9-40). Bologna: Gedit.

Bianchi, F. (2007). The cultural profile of chocolate in current Italian society: A corpus-based pilot study. *ETC – Empirical Text and Culture Research*, 3, 106-120.

Bianchi, F. (2010), Understanding culture. Automatic semantic analysis of a general Web corpus and a corpus of elicited data. *ETC – Empirical Text and Culture Research*, 4, 1-29.

Bianchi, F., & Pazzaglia, R. (2007). Student writing of research articles in a foreign language: metacognition and corpora. In R. Facchinetti (ed.), *Corpus Linguistics 25 Years on* (pp. 261-289). Amsterdam: Rodopi.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8, 4, 243-257.

Biber, D., & Finegan, E. (1991). On the exploitation of corpora in variation studies. In K. Aijmer, & B. Altenberg (eds.), *English Corpus Linguistics: Studies in Honour of Jan* Svartvik (pp. 204-220). London: Longman.

Bigi, S. (2006). Focus on Cultural Keywords. *Studies in Communication Sciences*, 6,1, 157-174.

Bowker, L., Pearson, J. (2002). *Working with Specialized Language. A practical guide to using corpora*. London & New York: Routledge.

Broder, A., Glassman, S., Manasse, M., Zweig, G. (1997). Syntactic clustering of the Web. In *Proceedings of the Sixth International World-Wide Web Conference*. Retrieved December, 15, 2009, from http://www.std.org/~msm/common/clustering.html.

Brug, J., Debie, S., van Assema, P., Weijts, W. (1995). Psychosocial Determinants of Fruit and Vegetable Consumption among Adults: Results of Focus Group Interviews. *Food Qualily and Preference*, 6, 99-107

Carter, R., McCarthy, M. (1995). Grammar and the spoken language. *Applied Linguistics*, 16, 141-158.

Cavalli-Sforza, L.L. (1996). *Geni, popoli e lingue* [Genes, Peoples, and Languages]. Milano: Adelphi.

Choi Y., Kim Y. Myaeg S. (2009). Domain-specific Sentiment Analysis using Contextual Feature Generation. *TSA '09. Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion* (pp. 37-44). Hong Kong, China.

Chomsky, N. (1962). Transformational Approach to Syntax. *Third Texas Conference on Problems of Linguistic Analysis in English May 9-12, 1958, Studies in American English* (pp. 124–158). Austin, Texas: The University of Texas.

Codern, N., Pla, M., de Ormijana, A.S., Gonzales, F.J. *et al.* (2010). Risk Perception Among Smokers: A Qualitative Study. *Risk Analysis*, 30, 10, 1563-1571.

Coffin, C., O'Halloran, K. (2004). Teaching Critical Discourse Analysis: the role of Corpus and the Concordancer. Paper presented at TALC 2004, 6-9 July 2004, Granada.

Cogozzo, V. (2005). *Corpora e cultura: uno studio comparato del contesto semantico del lemma 'cioccolato'* [Corpora and culture: a comparative study of the semantic context of lemma 'chocolate']. Unpublished graduation thesis. University of Genova, Italy.

Crystal, D. (2003). *English as a Global Language*. Cambridge: Cambridge University Press.

Culpeper, J. (2002). Computers, language and characterization: An analysis of six characters in Romeo and Juliet. In U. Melander-Marttala, C. Östman, & M. Kytö (Eds), *Conversation in Life and in Literature* (pp. 11-30). Uppsala: Universitetstryckeriet.

Donoghue, S. (2000). Projective techniques in consumer research. *Journal of Family Ecology and Consumer Sciences*, 28, 47-53. Retrieved May, 15, 2012 from www.up.ac.za/saafecs/vol28/donoghue.pdf.

Durrant, P. & Doherty, A. (2010). Are high frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, 6, 2, 125-155.

Evert, S. (2004). The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714.

Evert, S. (2007). *Corpora and collocations*. Retrieved December, 15, 2011, from http://purl.org/stefan.evert/PUB/Evert2007HSK_extended_manuscript.pdf. Shorter version of the text in A. Lüdeling & M. Kytö (eds.) (2008), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.

Fagerland, W., Sandvik, L., & Mowinckel, P. (2011). Parametric methods outperformed non-parametric methods in comparisons of discrete numerical variables. *BMC Medical Research Methodology*, 11-44.

Fairclough, N. (2003). *Analysing Discorse*. London & New York: Routledge.

Fillmore, C.J. (1992). ‘Corpus linguistics’ or ‘computer-aided armchair linguistics’?. In J. Svartvik (Ed.), *Directions in Corpus Linguistics* (pp. 35-60). Berlin: Mouton-de-Gruyter.

Fleischer, M. (1989). *Die sowjetische Semiotik. Theoretische Grundlagen der Moskauer und Tartuer Schule* [Sovietic semiotics. Theoretical bases of the Moscow-Tartu school]. Tübingen: Stauffenburg-Verl.

Fleischer, M. (1998). Concept of the ‘Second Reality’ from the perspective of an empirical systems theory on the basis of radical constructivism. In G. Altman, W.A. & Koch (Eds), *Systems. New Paradigms for the Human Sciences* (pp. 423-460). Berlin-NewYork: De Gruyter.

Fleischer, M. (2001). *Kulturtheorie: Systemtheoretische und evolutionäre Grundlagen*. Oberhausen: Athena.

Fleischer, M. (2002). Das Image von Getränken in der polnischen, deutschen und französischen Kultur. *ETC - Empirical Text and Culture Research*, 2, 8-47.

Fleischer, M. (2003). *Wirklichkeitskonstruktion: Beiträge zur systemtheoretischen Konstruktivismusforschung* [The construction of reality: contributions to constructivism and systems theory research]. Dresden: Thelem.

Fletcher, W. (2004). Making the Web more useful as a source for linguistic corpora. In U. Cornor, & T. Upton (eds.), *Corpus Linguistics in North America 2002.* Amsterdam: Rodopi. Retrieved September, 12, 2007 from http://kwicfinder.com/AAACL2002whf.pdf.

Fowler, J., Cohen, L, & Jarvis, P. (1998). *Practical Statistics for Field Biology*. Chichester (UK): Wiley.

Francis, W.N,. & Kucera, H. (1979). *Brown Corpus Manual*. Retrieved June, 14, 2007, from http://icame.uib.no/brown/bcm.html.

Galasinski, D., & Marley, C. (1998). Agency in foreign news: A linguistic complement of a content analytical study. *Journal of Pragmatics*, 30, 565–587.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech, & A. McEnery (Eds), *Corpus annotation: Linguistic information from computer text corpora* (pp. 102-121). London: Longman.

Gatto, M. (2009). *From Body to Web. An Introduction to the Web as Corpus*. Bari: La Terza.

Geertz, C. (1979). *Meaning and order in Moroccan society: Three essays in cultural analysis*. Cambridge: Cambridge University Press.

Geertz, C. (1998). *Interpretazione di culture*. Bologna: Il Mulino. [Italian translation of Geertz C. 1973. *The Interpretation of Cultures*.]

Gerbig, A. (1993). The representation of agency and control in texts on the environment. In R. Alexander, J. Bangs, J. Door (Eds), *Language and Ecology*. *Proceedings from Symposium on ecolinguistics* (pp. 61-73). Amsterdam.

Gledhill, C. (1995). Collocation and genre analysis: the phraseology of grammatical items in cancer research abstracts and articles. *Zeitschrift für Anglistik und Amerikanistik*, 43, 11-29.

Gledhill, C.J. (1996). Collocation and the rhetoric of scientific ideas. Corpus linguistics as a methodology for genre analysis. Retrieved March, 21, 2003 form http://helmer.aksis.uib.no/allc/gledhill.pdf.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes*, 19, 2, 115-135.

Goddard, C., Wierzbicka, A. (Eds.) (1994). *Semantic and lexical universals: Theory and empirical findings*. Amsterdam: John Benjamins.

Gordesch, J. (1998). Evolutionary modelling. In G. Altman, & W.A. Koch (Eds.), *Systems. New Paradigms for the Human* Sciences (pp. 39-57). Berlin-NewYork: De Gruyter.

Grace, S., Cramer, K. (2003).The Elusive Nature of Self-Measurement: the Self-Construal Scale versus the Twenty Statements Test. *Journal of Social Psychology*, 143, 5, 649-668.

Granath, S. (2007). Size matters – or thus can meaningful structures be revealed in large corpora. In R. Facchinetti (Ed.), *Corpus Linguistics 25 Years* on (pp. 169-186). Amsterdam & New-York: Rodopi.

Green, B., & Rubin, G. (1971). *Automatic Grammatical Tagging of English*. Technical Report, Department of Linguistics, Brown University, RI.

Greenberg, J.H. (1960). Concerning inferences form linguistic to nonlinguistic data. In H. Hoijer (Ed.), *Language in culture* (pp. 3-20). Chicago: University of Chicago.

Grefenstette, G., & Nioche, J. (2000). Estimation of English and non-English language use on the www. In *Proc. RIAO (Recherche d'Informations Assist_ee par Ordinateur)* (pp. 237-246). Paris. Retrieved September, 12, 2007, from http://arxiv.org/ftp/cs/papers/0006/0006032.pdf.

Graveter, F.J., Forzano, L.A.B. (2008). *Research methods for the behavioural sciences. Cengage Learning EMEA*, International Editions 3e. UK: Gardners Books.

Guerrero, L., Claret, A., Verbeke, W. *et al.* (2010). Perception of traditional food products in six European regions using free word association. *Food Quality and Preference*, 21, 225-233.

Guerrero, L., Colomer, J., Guàrdia, M.D., Xicola, J., & Clotet, R. (2000). Consumer attitude towards store brands. *Food Quality and Preference*, 11, 387-395.

Gulli, A., Signorini, A. (2005). The Indexable Web is more than 11.5 billion pages. *Proceedings of WWW 2005* (pp. 902-903). Japan: Chiba. Retrieved September, 12, 2007 from http://www.divms.uiowa.edu/~asignori/papers/the-indexable-web-is-more-than-11.5-billion-pages/size-indexable-web.pdf.

Hair, J.F., Bush, R.P., & Ortinau, D.J. (2009). *Marketing Research. In a Digital Information Environment*. Boston: McGraw-Hill.

Hall, E.T. (1982). *The Hidden Dimension*. New York: Doubleday.

Hall, E.T. (1989). *Beyond Culture*. New York: Doubleday.

Halliday, M.A.K. (1978). *Language as Social Semiotics. The social Interpretation of Language and Meaning*. London: Arnold Ed.

Halliday, M.A.K. (1985). *An Introduction to Functional Grammar*. London: Arnold Ed.

Halliday, M.A.K., & Hasan, R. (1989). *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*. Oxford: Oxford University Press.

Heaps, H.S. (1978). *Information Retrieval: Computational and Theoretical Aspects*. Orlando: Academic Press.

Heyland, K., Tse, P. (2005). Hooking the reader: a corpus study of evaluative *that* in abstracts. *English for Specific Purposes*, 24, 2, 123-139

Hoey, M. (1991). *Patterns of lexis in text. Oxford*: Oxford University Press.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Hofland, K., & Johansson, S. (1982). *Word Frequencies in British and American English*. Bergen: The Norwegian Computing Centre for the Humanities.

Hogenraad, R. (2004). What happens when a psychologist takes on a literary topic? Literary theorists don't like it. A reply to Gerard Steen and Edmund Nierlich. *IGEL News*, 14. Retrieved May, 23, 2005, from http://www.arts.ualberta.ca/igel/Newsletter14.htm#Hogenraad.

Hogenraad, R.(2005). What the Words of War Can Tell Us About the Risk of War. *Peace and Conflict: Journal of Peace Psychology*, 11, 2, 137-151.

Holsopple, J.Q., & Miale, F.R. (1950). *Sentence completion: a projective method for the study of personality*. Springfield, IL: Thomas.

Hubbard, D.W. (2010). *How to Measure Anything. Finding the Value of "Intangibles" in Business*. Hoboken, New Jersey: John Wiley & Sons Inc..

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2004). Counting the uncountable: Problems of identifying evaluation in text and in a corpus. In A. Partington, J. Morley, L. Haarman (Eds), *Corpora and Discourse* (pp.157-188). Bern: Peter Lang.

Hunston, S., & Francis, G. (2000). *Pattern Grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hunston, S., & Sinclair, J. (2000). A local grammar of evaluation. In S. Hunston, & G. Thompson (Eds.), *Evaluation in Text* (pp. 74-101). Oxford: Oxford University Press.

Jabaseeli, A.N., & Kirubakaran, E. (2012). A Survey on Sentiment Analysis of (Product) Reviews. *International Journal of Computer Applications*, 47, 11, 36-39.

Kaiser, C., Schlick, S., & Bodendorf, F. (2011). Warning system for online market research – Identifying critical situations in online opinion formation. *Knowledge-Based Systems*, 24, 824–836.

Katan, D. (2006). It's a Question of Life or Death: Cultural differences in advertising private pensions. In N. Vasta (Ed.), *Forms of Promotion Texts, contexts and cultures* (pp. 55-80). Bologna: Pàtron Editore. Retrieved April, 14, 2007 from: www.sslmit.univ.trieste.it/katan/Papers/Vasta%20Paper.doc.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London & New York: Longman.

Kilgarriff, A. (1996a). Comparing word frequencies across corpora: Why chi-square doesn't work, and an improved LOB-Brown comparison. In *Proceedings from ALLC-ACH'96* (pp. 169-172). Retrieved December, 20, 2008, from http://torvald.aksis.uib.no/allc/kilgarny.pdf.

Kilgarriff, A. (1996b). Which words are particularly characteristic of a text? A survey of statistical approaches. In L.J. Evett, & T.G. Rose (Eds.), *Language Engineering for Document Analysis and Recognition (LEDAR), AISB96 Workshop proceedings* (pp. 33-40). Brighton, England. Faculty of Engineering and Computing, Nottingham Trent University, UK. Retrieved May, 28, 2006 from http://www.kilgarriff.co.uk/Publications/1996-K-AISB.pdf.

Kilgarriff, A. (2001a). The web as corpus. In R. Rayson, A. Wilson, T. McEnery, A. Hardie, & K. Shereen (Eds), *Proceedings of Corpus Linguistics 2001* (pp. 342-344). Lancaster. Retrieved May, 17, 2007, from http://www.kilgarriff.co.uk/Publications/2001-K-CorpLingWAC.txt.

Kilgarriff, A. (2001b). Comparing Corpora. *International Journal of Corpus Linguistics*, 6, 1, 97-133.

Kilgarriff, A., Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics,* 29, 3, 333-347.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. *Euralex XI Proceedings* (pp. 105-11), Lorient, France. Retrieved May, 28, 2007, from ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-08.pdf.

Kleij, F., Musters, P.A.D. (2003). Text analysis of open-ended survey responses: a complementary method to preference mapping. *Food Quality and Preference*, 14, 43-52.

Kluckhohn, F.R., & Strodtbeck, F.L. (1961). *Variations in value orientations*. Evanston, Illinois: Row, Peterson.

Kroeber, A.L. (1952). *The nature of culture*. Chicago: The University of Chicago Press.

Kroeber, A.L., & Kluckhohn, C. (1952). *Cultures. A Critical Review of Concepts and Definitions*. Cambridge, Mass.: Harvard University.

Lawrence, S., & Giles, C.L. (1999). Accessibility of information on the web. *Nature*, 400, 107-109.

Leech, G. (1974). *Semantics*. Harmondsworth: Penguin.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (ed.), *Directions in corpus linguistics: proceedings of Nobel symposium 82* (pp. 125-148). Berlin & New York: Mouton de Gruyter.

Leech, G. (2007). New Resources, or Just Better Old Ones? The Holy Grail of Representativeness. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds), *Corpus Linguistics and the Web* (pp. 133-149). Amsterdam: Rodopi.

Leech, J., & Fallon, R. (1992). Computer Corpora: What do they tell us about Culture?. *ICAME Journal*, 16, 29-50.

Loftsson, H. (2009). Correcting a PoS-tagged corpus using three complementary methods. *Proceedings of the 12th Conference of the European Chapter of the ACL* (pp. 523-531). Retrieved January, 18, 2010 from http://www.aclweb.org/anthology/E/E09/E09-1060.pdf.

Lotman, J.M. (1980a). Prefazione [Preface]. In S. Salvestroni (a cura di), *Testo e contesto. Semiotica dell'arte e della cultura* [Text and context. The semiotics of art and culture] (pp. 3-5). Roma-Bari: Laterza.

Lotman, J.M. (1980b). Un modello dinamico del sistema semiotico [A dynamic model of the semiotic system]. In S. Salvestroni (a cura di), *Testo e contesto. Semiotica dell'arte e della cultura* [Text and context. The semiotics of art and culture] (pp. 9-27). Roma-Bari: Laterza.

Lotman, J.M. (1980c). La cultura come intelletto collettivo e i problemi dell'intelligenza artificiale [Culture as collective memory and the issues of artificial intelligence]. In S. Salvestroni (a cura di), *Testo e contesto. Semiotica dell'arte e della cultura* [Text and context. The semiotics of art and culture] (pp. 29-44). Roma-Bari: Laterza.

Lotman, J.M. (1980d). Il fenomeno della cultura [The phenomenon of culture]. In S. Salvestroni (a cura di), *Testo e contesto. Semiotica dell'arte e della cultura* [Text and context. The semiotics of art and culture] (pp. 45-60). Roma-Bari: Laterza.

Lotman, J.M. (1985a). Introduzione [Introduction]. In S. Salvestroni (a cura di). *La semiosfera: l'asimmetria e il dialogo nelle strutture pensanti* [The Semiosphere: asymmetry and dialogue in thinking structures] (pp. 49-51). Venezia: Marsilio.

Lotman, J.M. (1985b). La Semiosfera [The Semiosphere]. In S. Salvestroni (a cura di). *La semiosfera: l'asimmetria e il dialogo nelle strutture pensanti* [The Semiosphere: asymmetry and dialogue in thinking structures] (pp. 55-76). Venezia: Marsilio.

Lotman, J.M. (1985c). La cultura e l'organismo [The culture and the organism]. In S. Salvestroni (a cura di). *La semiosfera: l'asimmetria e il dialogo nelle strutture pensanti* [The Semiosphere: asymmetry and dialogue in thinking structures] (pp. 77-82). Venezia: Marsilio.

Lotman, J.M. (1985d). La metasemiotica e la struttura della cultura [metasemiotics and the stucture of culture]. In S. Salvestroni (a cura di). *La semiosfera: l'asimmetria e il dialogo nelle strutture pensanti* [The Semiosphere: asymmetry and dialogue in thinking structures] (pp. 83-90). Venezia: Marsilio.

Lotman, J.M. (1993). *La cultura e l'esplosione: prevedibilità e imprevedibilità* [Culture and explosion: foreseeability and unforeseeability]. Milano: Feltrinelli.

Lotman, J.M. (1994). *Cercare la strada. Modelli della cultura* [Looking for the way. Models of culture]. Venezia: Marsilio.

Lotman, J.M. (1997). Il simbolo nel sistema della cultura [Symbol in the system of culture]. In R. Galassi, & R. De Michiel (a cura di). *Il simbolo e lo specchio. Scritti della scuola semiotica di Mosca-Tartu* [The symbol and the mirror.

Papers of the Moscow-Tartu semiotic school] (pp. 53-66). Napoli: Edizioni Scientifiche Italiane.

Lotman, J.M., & Uspenskij, B. (1975). Sul meccanismo semiotico della cultura [The sign mechanism of culture]. In J.M. Lotman, & B. Uspenskij (Eds), *Semiotica e cultura* [Semiotics and culture] (pp. 59-96). Milano-Napoli: Ricciardi.

Lowe, J.W.G., & Barth, R.J. (1980). Systems in archaeology: a comment on Salmon. *American Antiquity*, 45, 3, 568-574.

Luzon Marco, M.J. (2000). Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes*, 19, 1, 63-86.

Luna, D., & Peracchio, L.A. (2002). Uncovering the cognitive duality of bilinguals through word association. *Psychology & Marketing*, 19, 6, 457-475.

Lyman, P., Varian, H.R. *et al.* (2003). *How much information 2003*. Technical report, School of Information Management and Systems, University of California at Berkeley. Retrieved September, 18, 2006 from http://academic.research.microsoft.com/Publication/3195543/how-much-information-2003-school-of-information-management-and-system.

Mahlberg, M. (2007). Lexical items in discourse: identifying local textual functions of *sustainable development*. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, Discourse and Corpora. Theory and analysis* (pp. 191-218). London-New York: Continuum.

Mair, C. (2006). Tracking ongoing grammatical change and recent diversification in present-day standard English: the complementary role of small and large corpora. In A. Renouf, & A. Kehoe (Eds.), *Language and Computers 55, The Changing Face of Corpus Linguistics* (pp. 355-376). Amsterdam & New York: Rodopi.

Manca, E. (2008). From phraseology to culture: the case of qualifying adjectives in the language of tourism. In U. Römer, & R. Schulze (Eds.), *Patterns, meaningful units and specialized discourses*. [*International Journal of Corpus Linguistics 13(3), Special Issue*] (pp. 368-385). Amsterdam: John Benjamins.

McArthur, T. (1981). *Longman lexicon of contemporary English*. London: Longman.

McEnery, T., & Wilson A. (2001). *Corpus Linguistics, 2nd edition*. Edinburgh: Edinburgh University Press.

McTavish, D.G., & Pirro, E.B. (1990). Contextual content analysis. *Quality and Quantity*, 24, 245-265.

Meyer, C.F. (2002), *English Corpus Linguistics. An Introduction*. Cambridge: Cambridge University Press.

Mitchell, T., Shinkareva, S., Carlson, A., Chang, K., Malave, V., Mason, R., & Just, M. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320, 1191-1195.

Morin, E. (1973). *Le paradigme perdu: la nature humaine* [The lost paradigm: The human nature]. Paris: Editions du Seuil.

Muntz, R. (2001). Evidence of Australian cultural identity through the analysis of Australian and British corpora. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, S. Khoja (Eds), *Proceedings of the Corpus Linguistics 2001 Conference* (pp. 393-399). Lancaster University.

Murphy, B., Baroni, M., & Poesio, M. (2009). EEG responds to conceptual stimuli and corpus semantics. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 619–627). Retrieved December, 15, 2011 from http://aclweb.org/anthology-new/D/D09/D09-1065.pdf.

Neuendorf, K.A. (2002). *The content analysis guidebook*. Thousand Oaks-London-New Delhi: Sage Publications.

Nisha Jesebaseeli A., Kirubakaran E. (2012). A Survey on Sentiment Analysis of (Product) Reviews. *International Journal of Computer Applications*, 47, 11, 36-39.

Nobis, A. (1992a). Das Selbstorganisationssystem Europa [The self-organizing system of Europe]. *Snakalog. An International Yearbook of Slavic Semiotics*, 3, 141-160.

Nobis, A. (1992b). Das besser Europa [The better Europe]. In E. Skotnicka-Illasiewicz (Ed.). *Dilemmata der europäinischen Identität* [Dilemmas of the European Identity] (pp. 29-33). Warszawa: Stiftung "Polen in Europe".

Nobis, A. (1998). Self-organization of culture. In G. Altmann, & W.A. Koch (eds.), *Systems. New Paradigms for the Human Sciences* (pp. 461-478). Berlin-NewYork: De Gruyter.

Nordquist, D. (2009). Investigating elicited data from a usage-based perspective. *Corpus Linguistics and Linguistic Theory*, 5, 1, 105-130.

O'Halloran, K.A. (2007). Critical discourse analysis and the corpus-informed interpretation of metaphor at the register level. *Applied Linguistics*, 28, 1-24.

Oakes, M. P. (2003). Contrasts Between US and British English of the 1990s. In E.H. Oleksy, & B. Lewandowska-Tomaszczyk (Eds), *Research and Scholarship in Integration Processes* (pp. 213–22). Lo´dz´: University of Lo´dz´ Press. Retrieved March, 30, 2007 from www.cet.sunderland.ac.uk/IR/oakesRSIP2003.pdf.

Partington, A. (1996), *Patterns and Meanings. Using Corpora for English Language Research and Teaching. Studies in Corpus Linguistics*. Amsterdam-Philadelphia: John Benjamin Publishing Company.

Partington, A. (2004), 'Utterly content in each other's company'. Semantic prosody and semantic preference'. *International Journal of Corpus Linguistics*, 9, 1, 131-156.

Piaget, J. (1937). *La construction du réel chez l'enfant* [The Construction of Reality in the Child]. Neuchâtel: Delachaux et Niestlé.

Potash, H.M., de Fileo Crespo, A., Patel, S., & Ceravolo, A (1990). Cross-Cultural Attitude Assessment With the Miale-Holsopple Sentence Completion Test. *Journal of Personality Assessment*, 55, 3&4, 657-662.

Powers, D. M. W. (1998). Applications and Explanations of Zipf's Law. In D.M.W. Powers (Ed.), *New Methods in Language Processing and Computational Natural Language Learning* (pp. 152-160). Somerset, NJ: ACL. Retrieved January, 5, 2012 from http://acl.ldc.upenn.edu/W/W98/W98-1218.pdf.

Pullman, M., McGuire, K, & Cleveland. C. (2005). Let Me Count the Words: Quantifying Open-Ended Interactions with Guests. *Cornell Hotel and Restaurant Administration Quarterly*, 46, 3, 323-343.

Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. thesis, Lancaster University.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13, 4, 519-549.

Rayson, P., Archer, D., Piao, S., & McEnery, A.M. (2004). The UCREL semantic analysis system. *Proceedings of the Beyond Named Entity Recognition Semantic Labeling for NLP Tasks Workshop, Lisbon, Portugal* (pp. 7-12). Retrieved May, 10, 2009, from http://ucrel.lancs.ac.uk/usas/usas20guide.pdf.

Renouf. A. (2007). Corpus development 25 years on: from super-corpus to cyber-corpus. In R. Facchinetti (Ed.), *Corpus Linguistics 25 Years on* (pp. 27-50). Amsterdam & New-York: Rodopi.

Rigotti, E. & Rocci, A. (2002). From argument analysis to cultural keywords (and back again). In F. Van Eemeren *et al.* (Eds.), *Proceedings of the Fifth Conference of the International Society for the Study of Argumentation*. Amsterdam: Sic Sat: 903-908. Retrieved May, 3, 2008, from http://www.ils.com.unisi.ch/rigotti_rocci_keywords.pdf.

Ringlstetter, C., Schulz, K., Mihov, S. (2006). Orthographic Errors in Web Pages: Toward CleanerWeb Corpora. *Computational Linguistics*, 32, 3, 295-340.

Roininen, K., Arvola, A., Lähteenmäki, L. (2006). Exploring consumers perceptions of local food with two different qualitative techniques: Laddering and word association. *Food Quality and Preference*, 17, 20–30.

Rossini Favretti, R., Tamburini, F., & De Santis, C. (2002). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In A. Wilson, P. Rayson, & T. McEnery (Eds.), *A rainbow of corpora: Corpus linguistics and the languages of the world*. Munich: Lincom-Europa. Retrieved April, 22, 2007 from http://citeseer.ist.psu.edu/660888.html.

Sapir, E. (1929). The Status of Linguistics as a Science. *Language*, 5, 207-214.

Sapir, E. (edited by Mandelbaum D.G.) (1949). *Selected writings in language, culture and personality*. Berkeley: University of California Press.

Schmid, H. (1997). Probabilistic part-of-speech tagging using decision trees. In D.B. Jones, & H.L. Somers (Eds), *New Methods in Language Processing* (pp. 154-1649. London: Routledge. Retrieved January, 22, 2010 from http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf.

Schmidt, K.M. (1988). Der Beitrag der begriffsonrientierten Lexikographie zur systematischen Erfassung von Sprachwandel und das Begriffswörterbuch zur mhd. Epik [The contribution of lexicography for the systematic recording of language change and the lexical dictionary of Middle High German epic]. In W. Bachofer (Ed.), *Mittelhochdeutsches Wörterbuch in der Diskussion* [The Middle High German Dictionary under discussion] (pp. 35-49). Tübingen: Max Niemeyer.

Schmid, H.J. (2003). Do women and men really live in different cultures? Evidence from the BNC. In A. Wilson, P. Rayson, & T. McEnery (eds.), *Corpora by the Lune. A festschrift for Geoffrey Leech* (pp. 185-221). Frankfurt: Peter Lang.

Schmid, H., Baroni, M., Zanchetta, E., & Stein, A. (2007). Il sistema 'tree-tagger arricchito' – The enriched TreeTagger system. *IA Contributi Scientifici*, 4, 2, 22-23. Retrieved January, 26, 2010, from http://evalita.fbk.eu/2007/proceedings/09-IA-IV-2-UniTn-pos.pdf.

Scott, M. (2008). *Wordsmith Tools version 5*. Liverpool: Lexical Analysis Software Ltd.

Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Philadelphia: John Benjamins.

Seppänen, J. (1998). Systems ideology in human and social sciences. In G. Altman, & W.A. Koch (Eds), *Systems. New Paradigms for the Human Sciences* (pp. 180-302). Berlin-NewYork: De Gruyter.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni, & S. Bernardini (Eds.). *WaCky! Working papers on the Web as Corpus*. Bologna: Gedit.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (1992). The automatic analysis of corpora. In J. Svartvik (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82* (pp. 379-397). Stockholm, 4-8 August 1991. Berlin: Mouton de Gruyter.

Sinclair, J. (2004). *Trust the text. Language, Corpus and Discourse*. London: Routledge.

Smith, B., & Wilson, J.B. (1996). A consumer's guide to evenness indices. *Oikos*, 76, 70–82.

Sogaard, A. (2009). Ensemble-based POS tagging of Italian. Proceeding of EVALITA 2009. Retrieved January, 26, 2010 from http://tcc.itc.it/events/EVALITA/reports/PoSTagging/POS_UNI_COPENHAGEN.pdf.

Stubbs, M. (1994). Grammar, Text, and Ideology: Computer-assisted Methods in the Linguistics of Representation. *Applied Linguistics*, 15, 2, 201- 223.

Stubbs, M. (1997). Whorf 's Children: Critical comments on Critical Discourse Analysis (CDA). In A. Ryan, & A. Wray (Eds.), *Evolving models of* language (pp. 110–116). Clevedon: BAAL in association with Multilingual Matters.

Stubbs, M. (2001). *Words and Phrases. Corpus Study in Lexical Semantics*. Oxford: Blackwell.

Stubbs, M. (2007a). On texts, corpora and models of language. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, Discourse and Corpora. Theory and* analysis (pp. 127-162). London & New York: Continuum.

Stubbs, M. (2007b). Multi-word sequences in English. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, Discourse and Corpora. Theory and analysis* (pp. 163-189). London & New York: Continuum.

Svartvik, J. (2007). Corpus linguistics 25+ years on. In R. Facchinetti (Ed.), *Corpus Linguistics 25 Years on* (pp. 11-25). Amsterdam & New York: Rodopi.

Szalay, L.B., & Maday, B.C. (1973). Verbal Associations in the Analysis of Subjective Culture. *Current Anthropology*, 14, 1-2, 33-42.

Taylor, C. (1998). *Language to Language. A practical and theoretical guide for Italian/English translators*. Cambridge: Cambridge University Press.

Thet, T.T., Na, J., Khoo, C.S.G., Shakthikumar, S. (2009). Sentiment analysis of movie reviews on discussion boards using a linguistic approach. *TSA '09. Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion* (pp. 81-84). Hong Kong, China.

Tognini Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: Benjamins.

Tognini Bonelli, E. (2004). Working with corpora. In C. Coffin, A. Hewings, & K. O'Halloran (Eds), *Applying English functional grammar and Corpus Approaches* (pp. 11-24). London: Arnold.

Tosi, A. (2001). *Language and Society in Changing Italy*. Clevedon, UK: Multilingual Matters.

Tsytsarau M., Palpanas T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24, 3, 478-514.

Ueyama, M. (2006). Creation of general-purpose Japanese Web corpora with different search engine query strategies. In M. Baroni, & S. Bernardini (Eds.), *WaCky! Working papers on the Web as Corpus* (pp. 99-126). Bologna: Gedit.

van Kleef, E., van Trijp, H.C.M., & Luning, P. (2005). Consumer research in the early stages of new product development: a critical review of methods and techniques. *Food Quality and Preference*, 16, 181-201.

Váradi, T. (2001). The linguistic relevance of corpus linguistics. In P. Rayson, A. Wilson, T. McEnery, A. Har Hardie & S. Khoja (Eds.), *Proceedings of the Corpus Linguistics 2001 Conference* (pp. 587-593). Lancaster University: UCREL Technical Papers 13. Retrieved January, 11, 2008, from http://ucrel.lancs.ac.uk/publications/CL2003/CL2001%20conference/papers/varadi.pdf.

von Glasersfeld, E. (1984). An Introduction to Radical Constructivism. In P. Watzlawick (Ed.), *The Invented Reality*. New York: Norton. Retrieved January, 27, 2007 from http://srri.nsm.umass.edu/vonGlasersfeld/onlinePapers/html/082.html.

Wales, K. (2001). *A Dictionary of Stylistics*. London: Longman/Pearson Education.

Warren, M. (2007). An initial corpus-driven analysis of the language of call center operators and customers. *ESP Across Cultures*, 4, 80-97.

Weber, R.P. (1990). *Basic Content Analysis*. Newbury Park, CA: Sage.

Wentan, Li (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE T Inform Theory*, 38, 1842–1845.

Whorf, B.L. (1956). *Language, Thought, and Reality*. New York: John Wiley & Sons, and The Technology Press of M.I.T.

Widdowson, H.G. (2000). On the limitations of linguistics applied. *Applied Linguistics*, 21, 1, 3-25.

Wierzbicka, A. (1972). *Semantic Primitives*. Frankfurt: Athenäum.

Wierzbicka, A. (1980). *Lingua mentalis: The semantics of natural language*. Sydney/New York: Academic Press.

Wierzbicka, A. (1987). *English speech act verbs: A semantic dictionary*. Sydney: Academic Press.

Wierzbicka, A. (1988). *The semantics of grammar*. Amsterdam: John Benjamins.

Wierzbicka, A. (1989a). Semantic primitives and lexical universals. *Quaderni di Semantica*, 10, 1, 103-121.

Wierzbicka, A. (1989b). Semantic primitives: the expanding set. *Quaderni di semantica*, 10, 2, 309-332.

Wierzbicka, A. (1991). *Cross-Cultural Pragmatics. The Semantics of Human Interaction*. Trends in Linguistics. Studies and Monographs, 53, Berlin/New York: Mouton de Gruyter.

Wierzbicka, A. (1997). *Understanding Cultures Through Their Key Words*. Oxford: Oxford University Press.

Williams, R. (1959). *Culture and Society*. London: Chatto & Windus.

Williams, R. (1976). *Keywords. A Vocabulary of Culture and Society*. London: Fontana.

Wilson, A. (1993). Towards an Integration of Content Analysis and Discourse Analysis: The Automatic Linkage of Key Relations in Text. *UCREL Technical Paper 3*, Linguistics Department, Lancaster University.

Wilson, A. (2003). Developing Conceptual Glossaries for the Latin Vulgate Bible. *Literary and Linguistic Computing*, 17, 4, 413-426.

Wilson, A., & Mudraya, O. (2006). Applying an Evenness Index in Quantitative Studies of Language and Culture: A Case Study of Women's Shoe Styles in Contemporary Russia. In P. Grzybek, & R. Köhler (Eds.), *Exact Methods in the Study of Language and Text* (pp. 709-722). Köhler: de Gruyter.

Wilson, A., & Moudraia, O. (2006). Quantitative or Qualitative Content Analysis? Experiences from a cross-cultural comparison of female students' attitudes to shoe fashions in Germany, Poland and Russia. In A. Wilson, P. Rayson, D. Archer (Eds.), *Corpus Linguistics around the World* (pp. 203-217). Amsterdam: Rodopi.

Wilson, A., & Thomas, J. (1997). Semantic Annotation. In R. Garside, G. Leech, & A. McEnery (Eds), *Corpus Annotation. Linguistic information from computer text corpora*. London & New York: Corpora.

Xiao, R.Z. (2008). Well-known and influential corpora. In A. Lüdeling, & M. Kyto (Eds.), *Corpus Linguistics: An International Handbook. Volume 1*. Berlin: Mouton de Gruyter.

Yates, S. (1996). English in Cyberspace. In S. Goodman, & D. Graddol (Eds.), *Redesigning English. New texts, new identitities* (pp. 106-133). London & New York: Routledge.

Yu, A.T.W., Shen, Q., Kelly, J., & Hunter K. (2006). Investigation of Critical Success Factors in Construction Project Briefing by Way of Content Analysis. *Journal of Construction Engineering and Management*, 132, 11, 1178-1186.

# Codebook
# for tagging semantic associations of the words Chocolate and Wine

By Francesca Bianchi
Revision n.3: 25.03.2011

The current codebook describes the coding scheme and the coding procedure to adopt in the manual coding of semantic associations of the words Chocolate and Wine.

## 1. The coding scheme

### *1.1. Step 1. Setting the coding scheme*

A set of semantic categories is given for the coding task. These semantic categories – listed in Table 1 – were set in a preliminary cross-cultural corpus study based on four data sources: two specialized corpora about Chocolate (one in Italian and one in English), and two general corpora in the same languages. From each corpus, concordances for the word Chocolate were extracted and each concordance line containing the node word was classified in terms of semantic field of the node word, that is the main topic(s) mentioned in the relevant text segment. Classification was based on the lexical meaning of the co-text and was performed through a data-driven, open-coding system. Once the first few categories were established by looking at the first concordance line, a category list was created and used for the classification of subsequent concordance lines. When none of the categories in the list fitted a given concordance line, one or more new categories were created and added to the list. For the classification of semantic fields, full sentences were usually considered, and sometimes also wider contexts.

This process was carried out individually by two different coders, one working on the English data and the other on the Italian ones. Halfway through the coding task, the two coders met and merged their category lists, thus creating a wider category list (which was called *List of Semantic fields*, or *first level list*) to be used by both for the completion of the coding tasks. The final *List of Semantic fields* and all of the coded data were eventually reviewed by a third coder. In the rare cases of disagreement between the reviewer and the coder (0.1%), the suggestions of both were accepted, and the concordance line under consideration was classified as a case of multiple attributions.

For the purpose of comparison, semantic fields were then grouped into higher-order categories which were called ‘conceptual domains’ (*second level list*). These superordinate groupings identify domains relevant to all and only the semantic fields contained in them. The grouping of semantic fields into conceptual domains, which was also established inductively by the same coders who had performed semantic field attributions, was carried out considering the labels identifying semantic fields and also concordance line content, as possible polysemy of the semantic field labels might have been misleading for classification purposes.

### *1.2 Step 2. Refinement of the coding scheme and creation of Codebook*

The coding scheme set in Step 1 was then applied to a set of elicited data (in English) around node word ‘chocolate’. The coder was the same person who did the code review in Step 1.

During this phase the coding scheme was enriched with a few new semantic fields: Drink (under conceptual domain Food); Transaction, Fair Trade, and Time (under Events); Guilt, Comfort, Relax, Piece, and Bribing (under Feelings & Emotions); Men, Gay, and Posh (under People); Hiding (under Loss & Damage); Quantities, Price, and Packaging (under Features). Furthermore, semantic field Taste (under Features) was renamed Taste/Smell; semantic field Arts (under Culture) was renamed Artistic production; semantic field Films (under Culture) was eliminated as it is subsumed in Artistic production; semantic field Eating (under Health & Body) was eliminated as it is subsumed either in Food (e.g. I have chocolate when I feel hungry), Health, or Medicine (e.g. Binge eating with chocolate was my favourite activity at the time); conceptual domain Psychology, which included semantic fields Psychology and Morals was eliminated, as it could be subsumed in Medicine (e.g. Chocolate contains substances called Phenylethylamine and Seratonin, both of which are mood lifting agents found naturally in the human brain) or Religion (e.g. The ingestion of chocolate by so many of the women in church inevitably caused great interruption). Finally, an extra semantic category was added, that of Comparison (under Comparison).

The resulting coding scheme was described in Codebook (rev.1).

### *1.3 Step 3. Further refinement of the coding scheme (Codebook_rev2 and Codebook_rev3)*

The coding scheme described in 1.2 was applied to two separate sets of elicited data (in English) around node words ‘chocolate’ and ‘wine’, respectively. Each set was coded by two separate coders: the researcher, and another coder who had receive specific training on the use of the coding scheme. During the coding procedure the two coders met a couple of times to discuss the need for further domains and/or fields. When a new semantic field was agreed upon and added to the list, each coder reviewed the sentences s/he had already tagged.

During this phase, and mostly due to the needs that emerged in connection to the key word ‘wine’, the following semantic fields were added:

- Storage; and Serving (in conceptual domain Food);
- Work; Driving; Excessive drinking; Holidays (in Events)
- Confidence (in Feelings & Emotions)
- Age (in People)
- Physical properties (in Features)

The resulting coding scheme (Codebook_rev2) is summarised in Table 1, for a total of 16 Conceptual domains and 92 semantic fields.

Subsequently the coding scheme in Table 1 was applied to an Italian set of elicited data on chocolate. As before, the data were coded by two separate coders: the researcher, and another coder who had receive specific training on the use of the coding scheme. During the coding procedure the two coders met a couple of times to discuss the need for further domains and/or fields. When a new semantic field was agreed upon and added to the list, each coder reviewed the sentences s/he had already tagged.

**Table 1: Codebook_rev2: Coding scheme**

| **Conceptual domain (second level list)** | **Semantic fields (first level list)** |
|---|---|
| Food | Product/shape; Bakery/cooking; Manufacturing; Food; Composition; Recipe; Drink; Storage; Serving |
| Health & Body | Health; Medicine; Body; Beauty |
| Events | Language/etymology; Economy; Religion/mythology; War; History; Law; Event; Transaction; Fair Trade; Time; Work; Driving; Excessive drinking; Holidays |
| Feelings & Emotions | Senses; Love; Desire; Pleasure; Sex; Happiness; Seduction; Mood; Passion; Competitiveness; Memory; Surprise; Loneliness; Freedom; Persuasion; Guilt; Comfort; Relax; Peace; Bribing; Confidence |
| People | Women; Men; Gay; Children; Posh; Friendship; Royalty; Sharing/society; People; Family; Age |
| Geography | Geographical locations; Spreading |
| Imagination | Fantasy/magic; Dream |
| Loss & Damage | Theft; Drugs and addiction; Hiding |
| Ceremonies | Ceremonies; Party; Gift |
| Environment & Reality | Nature; Animals; House; Dirt; Technology |
| Culture | Artistic production; Culture; Studying/intellect |
| Life | Future; Existence |
| Features | Quality/type; Colour; Sweet; Genuineness; Energy; Taste/Smell; Quantities; Price; Packaging; Physical properties |
| Sports | Sports |
| Comparison | Comparison |
| Assessment | Assessment |

During this phase, the following semantic fields were added:

- Dieting (in conceptual domain Health and Body);
- Playing (in Events)
- No reaction; Unpleasant (in Feelings & Emotions)

Furthermore, semantic field Pleasure (in Feelings & Emotions) was renamed Nice/Pleasant/Pleasure

The resulting coding scheme (Codebook_rev3) is summarised in Table 2, for a total of 16 conceptual domains and 96 semantic fields.

**Table 2: Codebook_rev3: Coding scheme**

| **Conceptual domain (second level list)** | **Semantic fields (first level list)** |
|---|---|
| Food | Product/shape; Bakery/cooking; Manufacturing; Food; Composition; Recipe; Drink; Storage; Serving |
| Health & Body | Dieting; Health; Medicine; Body; Beauty |
| Events | Playing; Language/etymology; Economy; Religion/mythology; War; History; Law; Event; |

| | |
|---|---|
| | Transaction; Fair Trade; Time; Work; Driving; Excessive drinking; Holidays |
| Feelings & Emotions | 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 | Women; Men; Gay; Children; Posh; Friendship; Royalty; Sharing/society; People; Family; Age |
| Geography | Geographical locations; Spreading |
| Imagination | Fantasy/magic; Dream |
| Loss & Damage | Theft; Drugs and addiction; Hiding |
| Ceremonies | Ceremonies; Party; Gift |
| Environment & Reality | Nature; Animals; House; Dirt; Technology |
| Culture | Artistic production; Culture; Studying/intellect |
| Life | Future; Existence |
| Features | Quality/type; Colour; Sweet; Genuineness; Energy; Taste/Smell; Quantities; Price; Packaging; Physical properties |
| Sports | Sports |
| Comparison | Comparison |
| Assessment | Assessment |

The semantic fields are briefly described below (Table 3). For each of them examples are also provided. In order to avoid biasing the current tagging, examples are taken from the Web data analysed in Step 1. When this is not possible an invented example is provided.

**Table 3: Semantic fields, descriptions and examples**

| **Sem. Field** | **Description** | **Examples** |
|---|---|---|
| Assessment | refers to the way the connotation the keyword acquired within the whole sentence, and may be Positive (P), Negative (N), Neutral (O) or Undecided (U). | **P:** Chocolate is good for your health.<br>**N:** I don't eat chocolate. It's too sweet.<br>**O:** Chocolate is made from cocoa beans.<br>**U:** I adore chocolate but try to eat little of it because it makes me fat. |
| Comparison | Comparison is expressed between two concepts, one of them being the key word, or between actions, one of them involving the key word | White chocolate is tastier than dark chocolate.<br>I don't dislike chocolate, but I prefer fruit. |
| Product/shape | A specific Brand or shape of the product is mentioned | Chocolate salami: made of extra dark chocolate with roasted hazels. |
| Bakery/cooking | The product is used in cooking or backing, or a type of backed or cooked product is mentioned | The cake is brick-shaped, with layers of soft pastry alternating with almond and chocolate cream, iced with chocolate and marzipan, and decorated with milk chocolate circles. |
| Storage | Mention of method or place of storage, or of feature that derives from storage. | White wine is usually served chilled.<br>Can I freeze wine? |
| Serving | Mention of how to serve the keyword, or description of serving process or object | Wine is best drunk from a glass.<br>Wine glasses come in various sizes. |
| Drink | The product is used as a drink | Drinking chocolate was not condemned by the Church which, in fact, in 1669, thanks to Cardinal Brancaccio – who dedicated an ode to it –, declared that the Solomonic saying "Liquidum non frangit jejunum" could equally apply to chocolate, i.e. that chocolate could be drunk at times of fasting without committing sin. |
| Manufacturing | The manufacturing process is described, or mentioned is made of a particular manufacturer | A couple of years ago, however, the European Parliament issued a directive that allowed the substitution of cocoa butter, traditionally employed in the production of chocolate, with other types of vegetable fats.<br>Nestlé is a Swiss brand of chocolate. |
| Food | The key word is considered as food or as accompaniment to food | These are modern, active women, who care about their physical fitness, have sound food knowledge, and see chocolate as an excellent natural and tasty |

| | | |
|---|---|---|
| | | energy integrator, unlike other integrators. They do sport, but are not slaves to it.<br>I would use chocolate to eat.<br>I have chocolate when I feel hungry.<br>My mum and brother also like to drink wine with a nice meal. |
| Composition | Some or all of the ingredients that make the product are listed | A couple of years ago, however, the European Parliament issued a directive that allowed the substitution of cocoa butter, traditionally employed in the production of chocolate, with other types of vegetable fats.<br>Amedei Chocolate: Cocoa min. 70% Dark chocolate 66%. |
| Recipe | Suggestions are offered of how to eat the key word, or a recipe is provided | Chocolate with strawberries is the ideal match.<br>Some people add chocolate to the wine and herbs sauce in which wild pork is stewed. |
| Dieting | Reference to dieting, being on a diet, or food having too many calories | Essendo a dieta, devo evitare alcuni alimenti, tra cui il cioccolato. [Being on a diet, I should avoid eating certain types of food, chocolate included.] |
| Health | Reference to general health or lack of it | People who make use of chocolate enjoy more constant health and are less prone to little illnesses that undermine the joy of life. |
| Medicine | Reference to specific medical topics, illnesses, or substances | Chocolate becomes desirable and even indispensable at difficult times of the day, due to its association with serotonine.<br>Binge eating with chocolate was my favourite activity at the time. |
| Body | Reference to parts of the body or to body shape, in connection with the key word | Chocolate is bad for your teeth.<br>Rose’s so thin because she eats little chocolate.<br>Chocolate is fattening. |
| Beauty | Mention of beauty (or lack of) in connection to the key word | In Turin I had a nice chocolate massage.<br>Chocolate makes you spotty. |
| Playing | Reference to playful activities | Con la cioccolata si possono fare molti giochi. [Chocolate can be used for playing different types of games.]<br>Vorrei tuffarmi in una piscina di cioccolato. [I’d like to dive in a swimming-pool full of chocolate.] |
| Language | The key word is referred to as a word | The word chocolate comes from the old Maya word Coxatal. |
| Economy | Reference to economy in general or to a specific economic situation | Fair trade chocolate stimulates the economy of third world countries.<br>Americans eat about 5 billion dollar’s worth of chocolate every year making us the world’s eighth largest consumer. |
| Transaction | Reference to buying and selling, or importing and exporting | You can buy good chocolate in any supermarket. |
| Fair trade | Reference to fair trade | Fair trade chocolate stimulates the economy of third world countries. |
| Driving | Reference to driving | Don’t drive after drinking wine. |
| Excessive drinking | Reference to excessive drinking | Whenever I think of wine I think of getting drunk.<br>She drank far too much wine at the party. |
| Work | Mention of profession directly connected to keyword, or use of keyword in the working environment | Some people have jobs as chocolate tasters.<br>Wine tasters are often pretentious.<br>I don’t drink wine when I’m working. |
| Religion | Reference to religious events, ceremonies, rites or people in their religious significance or capacity | Drinking chocolate was not condemned by the Church which, in fact, in 1669, thanks to Cardinal Brancaccio – who dedicated an ode to it –, declared that the Solomonic saying "Liquidum non frangit jejunum” could equally apply to chocolate, i.e. that chocolate could be drunk at times of fasting without committing sin. |
| War | Reference to real or metaphorical | At the end of the Second World War, the Americans |

| | | |
|---|---|---|
| | war or conflict | distributed chocolate bars, along with tinned food and drinks.<br>The two countries were fighting for the monopoly of the chocolate market. |
| History | Mention of historical facts about or the history of the key word | Drinking chocolate was not condemned by the Church which, in fact, in 1669, thanks to Cardinal Brancaccio – who dedicated an ode to it –, declared that the Solomonic saying "Liquidum non frangit jejunum" could equally apply to chocolate, i.e. that chocolate could be drunk at times of fasting without committing sin. |
| Law | Mention of the legal status of the product, or of legal directives connected to it | A couple of years ago, however, the European Parliament issued a directive that allowed the substitution of cocoa butter, traditionally employed in the production of chocolate, with other types of vegetable fats. |
| Event | An event is mentioned, deprived of its religious significance | I want a chocolate cake at my birthday party.<br>I get chocolate at Easter. |
| Holiday | Reference to holidays in connection to keyword | Wine tasting cruises and holidays are very popular. |
| Time | Reference to a specific time or time span in connection with the key word | I like eating chocolate in the long winter nights in front of my fireplace.<br>I could eat chocolate all day. |
| No reaction | Reference to absence of feelings | L'immagine riportata sopra non mi dice nulla. [The picture on top means nothing to me.] |
| Unpleasant | Reference to unpleasant feelings | Quando penso al cioccolato mi sento male. [When I think about chocolate, I feel sick.] |
| Senses | General reference to the five senses | Chocolate is a joy to all the five senses. |
| Love | Mention of the key word as an expression of love for a person or in connection to love to for a person | Chocolate makes a good Valentine gift. |
| Desire | Explicit or implicit mention of a desire for the key word or the key word used to satisfy a desire | Chocolate becomes desirable and even indispensable at difficult times of the day, due to its association with serotonine. |
| Nice/Pleasant/Pleasure | Direct or indirect mention of the key word being or producing pleasure | Chocolate, a pleasure for the palate and the eyes. |
| Sex | Direct or indirect mention of the key word in connection to sex | Chocolate is better than sex. |
| Happiness | Direct or indirect mention of the key word producing or leading to happiness | Chocolate, milky or darky, or even almondy, you make happy every chappy. |
| Seduction | Using the key word to seduce, or description of the key word as sensuous | Chocolate is sensual.<br>On our first date, I used chocolate to make an impression. |
| Mood | Direct or indirect mention of the key word generally acting on mood | Chocolate can change your mood.<br>Chocolate helps me get over bad times. |
| Passion | Strong feelings (love, hate, obsession, craving, strong desire, etc.) for the key word | In the new millennium there is certainly a craving for chocolate: it is given as a present, it is talked about, and it has its fun clubs and web sites. |
| Confidence | Mention of keyword in direct or indirect reference to confidence | I'll ask him out when I've got enough wine inside me.<br>I would use wine to gain confidence. |
| Competitiveness | Mention of the key word in direct reference to competitions | The amount of chocolate involved in this competition has relighted the imagination to incite candy eaters and all citizens all around the world. |
| Memory | The key word triggers memories or is remembered | Chocolate reminds me of my childhood. |
| Surprise | The key word as cause of surprise or similar feeling | Chocolate is awesome. |
| Guilt | Direct or indirect mention of a feeling of guilt | When I eat chocolate I feel guilty. |

| | | |
|---|---|---|
| Comfort | Direct or indirect mention of a feeling of comfort | Chocolate reminds me of luxury.<br>Chocolate is comforting. |
| Relax | Direct or indirect mention of a feeling of relaxation | Chocolate is good at the end of a long day. |
| Peace | Direct or indirect mention of a feeling or state of peace | If more people ate Chocolate there would be more peace in the world. |
| Loneliness | Direct or indirect mention of a feeling or state of loneliness | You'll have a chair to rest on, and hours as empty as chocolate eggs. |
| Freedom | Direct or indirect mention of a feeling or state of freedom | Chocolate makes you free. |
| Persuasion | Use of the key word to persuade | Its taste of chocolate convinced me that we were doing the right thing. |
| Bribing | Key word used for bribing someone | I would use chocolate to bribe my daughter to be obedient. |
| Women | Mention of women in connection to the key word (either by using a generic word for females or a female name) | These are modern, active women, who care about their physical fitness, have sound food knowledge, and see chocolate as an excellent natural and tasty energy integrator, unlike other integrators. They do sport, but are not slaves to it.<br>Sweet Agnese of chocolate hue, come to think of it, I've never kissed you. |
| Men | Mention of men in connection to the key word (either by using a generic word for males or a male name) | Men eat less chocolate than women.<br>Tony prefers dark chocolate. |
| Gay | Mention of homosexuals in connection to the key word | I've got a gay friend who always gives me a box of chocolates for my birthday. |
| Children | Mention of children and babies in connection to the key word | One day, while I was going to school, I saw a boy with many sweets and chocolates in his rucksack. |
| Age | Reference to age or age group (except children) | Wine is for adults.<br>Teens drink cheap wine. |
| Friendship | Mention of the key word in direct reference to friendship and friends | There is nothing better than a good friend – except a good friend with chocolate. |
| Royalty | Mention of the key word in direct reference to nobility or nobles | From the court of Spain chocolate spread like a collective cult among the noble élites of Europe. |
| Posh | Mention of the key word in direct reference to someone or something being posh, including the key word itself | Belgian chocolate boxes look posh. |
| Sharing/society | Reference to the social use of the key word | The whole world like chocolate.<br>Chocolate makes the world go round. |
| People | Mention of the key word in connection to people in general | Most people love chocolate. |
| Family | Mention of the key word in connection to family as an institution, or to members of the family | I'm not a chocolate brother, but I don't mourn, 'cause I swear this vanilla kid got its going. |
| Geo locations | Mention of the key word in connection to specific geographical locations or brand names that are identified with a specific nation (e.g. Coca-Cola = USA) | Travelling around the towns of chocolate.<br>Lattenero is top quality milk chocolate. This particular taste is achieved using a high percentage of cocoa from selected plantations in Venezuela. |
| Spreading | Mention of the key word in connection to multiple geographical locations | Chocolate is made all over the world. |
| Fantasy/magic | Mention of the key word in connection to a magical or fantastic world | Every drop of that river is hot melted chocolate of the finest quality! |
| Dream | Mention of the key word in connection to dreaming | Chocolate is often found in dreams.<br>My daughter dreamt of chocolate last night. |
| Theft | Mention of the key word in connection to thieving | Where did you steal that chocolate bar from the candy store? |
| Drugs and addiction | Mention of the key word in | People may become addicted to chocolate as much as |

| | | |
|---|---|---|
| | connection to illegal drugs, or direct or indirect comparison to an illegal drug | cocaine consumers (are addicted to cocaine). |
| Hiding | The key word or someone directly connected to it hides or is hidden | To save chocolate from my sister's eagerness I have to hide it under my bed. |
| Ceremonies | Mention of the key word in connection to a specific ceremony (e.g. marriage; baptism, etc.) | Chocolate is little used at weddings. |
| Party | Mention of the key word in connection to a party | Experience this transformation by hosting a chocolate tasting party for friends. |
| Gift | Using the key word as a gift to others or yourself | In the new millennium there is certainly a craving for chocolate: it is given as a present, it is talked about, and it has its fun clubs and web sites.<br>I would treat myself with chocolate. |
| Nature | Mention of the key word in connection to natural elements | Chocolate begins by luring visitors into a tropical rain forest where they can examine a replica of a Theobroma cacao tree, which produces the seeds that are used to make the sublime substance. |
| Animals | Mention of the key word in connection to animals, either living or fantastic or made of the key word | In 1575, Benzoni said that "chocolate is more like a drink for pigs".<br>At Easter I was given a small chocolate bunny. |
| House | Mention of the key word in connection to parts of a house | The windows were chocolate, and all the walls and ceilings were made of chocolate, so were the carpets and the pictures and the furniture and the be beds. |
| Dirt | Mention of the key word in connection to dirt | Mind not to dirt the sofa with chocolate.<br>Chocolate may grease your fingers, if you're not careful. |
| Tech | Mention of the key word in connection to technology or technical objects (also made of the key word) | At Easter I was given a small chocolate Ferrari. |
| Artistic production | Mention of the key word in connection to books, films, paintings and the like | I've seen the film Willy Wonka and the Chocolate Factory five times. |
| Culture | Mention of the key word in connection to culture in general or cultural events/places | Have you been to the Chocolate museum in Brussel? |
| Studying/intellect | Mention of the key word in connection to study and intellect | Chocolate makes you brighter. |
| Future | Talking about the future | In the future, cars will be powered by chocolate. |
| Existence | Mention is made to one's living, or to life in general | Life without chocolate is not worth living.<br>The woman lived in a chocolate house. |
| Quality/type | Mention of different types or qualities of the product | Chocolate, milky or darky, or even almondy, you make happy every chappy.<br>White chocolate and milk chocolate are sweeter than dark chocolate, and I like them better. |
| Quantity | Mention of quantity of product | Too much chocolate is sickening. |
| Physical properties | Reference to physical properties of the keyword | Chocolate melts in the sun.<br>Wine is my favourite alcoholic drink. |
| Colour | Direct or in direct mention of the product's colour | Sweet Agnese of chocolate hue, come to think of it, I've never kissed you. |
| Sweet | Direct or in direct mention of the product being sweet | White chocolate and milk chocolate are sweeter than dark chocolate, and I like them better. |
| Genuine | Direct or in direct mention of the product being genuine | These are modern, active women, who care about their physical fitness, have sound food knowledge, and see chocolate as an excellent natural and tasty energy integrator, unlike other integrators. They do sport, but are not slaves to it. |
| Energy | Direct or in direct mention of the product being energetic. | I eat chocolate immediately before setting off for my daily 30 km bike ride. |
| Taste/smell | Taste or smell is either directly or indirectly mentioned or involved | These are modern, active women, who care about their physical fitness, have sound food knowledge, |

| | | |
|---|---|---|
| | in the statement | and see chocolate as an excellent natural and tasty energy integrator, unlike other integrators. They do sport, but are not slaves to it.<br>Chocolate, a pleasure for the palate and the eyes. |
| Price | Reference to specific price or general mention of the product being cheap or expensive | Very good chocolate may be expensive. |
| Packaging | Mention or description of product’s packaging | Chocolate comes in lovely carton boxes. |
| Sports | Mention of the key word being used in connection to sports | These are modern, active women, who care about their physical fitness, have sound food knowledge, and see chocolate as an excellent natural and tasty energy integrator, unlike other integrators. They do sport, but are not slaves to it. |

## 2. The current coding task

### *2.1 The data*

The data were elicited by means of questionnaires with sentence completion and sentence writing tasks. In fact, the questionnaires began with the following completion sentences:

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

These were followed by a request to write 20 sentences that include the word given. Some respondents wrote less that 20 sentences, or even no sentence at all.

### *2.2 Coding procedure*

The unit of data collection is the questionnaire, while the unit of analysis is the sentence.

Coding is done manually and requires the coders to assign **one ore more semantic domains (chosen among the ones given)** to whole sentences on the basis of their assessment of the semantic areas/domains that are explicitly or implicitly mentioned in the given sentence. Decisions might be triggered by specific words in the sentence [e.g. *Very good chocolate may be expensive* = PRICE; *Chocolate is good for your health* = POSITIVE + HEALTH], but also by considerations regarding thematization [e.g. *Chocolate is tasty but makes you fat* = NEGATIVE; *Chocolate makes you fat but is very tasty* = POSITIVE], 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. *Chocolate is smooth and creamy* = POSITIVE, because usually smooth and creamy have a positive connotation; *I eat chocolate before sitting an exam* = POSITIVE + ENERGY, because it's common knowledge that an exam is a hard task that drains your energies).

An Excel table is provided for the coding task. The first column lists all sentences collected. The answers appear in the order they were given, one questionnaire after another. Change of respondent usually takes place with the following sentence: "Whenever I think of chocolate I…", or "Whenever I think of wine I….". Columns from B onwards list the Semantic fields to choose from.

For semantic field Assessment, please assign a value of Positive (P), Negative (N), Neutral (O) or Undecided (U) by typing the corresponding letter in the cell. For the other semantic fields, please enter X when the field is present, nothing when not present. Since multiple attributions are possible, a concept like Hate or Loathing will be marked as PASSION + NEGATIVE.

At the end of the coding process, we suggest you check your coding in the following way:

- activate the filter feature in the excel table by selecting the row listing semantic fields (usually the second raw)
- filter the sentences, semantic field after semantic field

If a coder feels that the descriptions of the semantic fields need extending or fine tuning, they should take note of the sentences which fit the category but not the description. These will be discussed with the other coders a the end of the coding process.

Coordinamento SIBA
UNIVERSITÀ DEL SALENTO

http://siba2.unisalento.it

eISBN 978-88-8305-093-0 (electronic version)

http://siba-ese.unisalento.it