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.