Contents

Preface

Acknowledgments

Chapter 1 - Introduction 1
1.1 An interdisciplinary perspective 1
1.2 Aims of the project and Research Questions 2
1.3 Outline of the current work 2

Chapter 2 - Culture 5
2.1 Introduction 5
2.2 Lotman and the semiotic perspective 8
2.3 Systems theory 11
2.3.1 Michael Fleisher: radical constructivism and semiotics 12
2.3.2 Adam Nobis: behavioural patterns and the self-organisation of culture 17
2.4 Analysing culture through language 18
2.4.1 Culture studies in linguistics 19
2.4.1.1 (Cross)-cultural semantics: Williams and Wierzbicka 19
2.4.1.2 Within the framework of corpus linguistics 22
2.4.1.3 Within the framework of cultural systems theory 25
2.4.2 Culture studies in other disciplines 26
2.4.3 Discussion 28

Chapter 3 - Corpora and corpus linguistics 31
3.1 Introduction 31
3.2 What is a corpus and what is corpus linguistic 31
3.3 Creating a corpus 33
3.3.1 Corpus design 33
3.3.1.1 Size 33
3.3.1.2 Representativeness, sampling and balancedness 35
3.3.2 Text collection 36
3.4 Corpora and the Web 36
3.4.1 The Web as Corpus proper 37
3.4.2 The Web as Corpus Shop: Creating corpora from the Web 38
3.4.3 Further issues and comments 39
3.5 Annotating a corpus 40
3.5.1 Part-of-speech tagging 41

3.5.2 Lemmatisation 42
3.5.3 Semantic annotation 43
3.6 Analysing a corpus: major analytical features and methods 45
3.6.1 Wordlists and frequency 45
3.6.2 Keywords and keyness 47
3.6.3 Concordancing 48
3.6.4 Collocation, semantic preference and semantic prosody 49
3.7 Some thorny issues 51

Chapter 4 - Marketing research 55
4.1 Introduction 55
4.1.1 Data gathering in marketing research 56
4.1.2 Research design in marketing research 57
4.2 Review of selected marketing studies 59
4.3. Cultural studies, corpus linguistics, marketing research, and the current work: common features 62
4.3.1 Preliminary experiments 63
4.3.2 The current work 64

Chapter 5 - The current study: materials and method 65
5.1 Introduction 65
5.2 Materials 67
5.2.1 Elicited data 67
5.2.1.1 The questionnaires 67
5.2.1.2 The respondents 68
5.2.1.3. The elicited datasets 68
5.2.2 The Web datasets 70
5.2.2.1 UKWAC and ITWAC 70
5.2.2.2 The Sketch Engine 72
5.2.2.3 Creating the project datasets 72
5.3 Procedure 73
5.3.1 Manual Coding 73
5.3.1.1 Origin and development of the Codebook 73
5.3.1.2 Manually coding whole datasets 74
5.3.1.3 Manually coding wordlists 75
5.3.1.4 The coding scheme 75
5.3.2 Automatic semantic tagging 76
5.4 Research design 77

Chapter 6 - Semantic associations of *chocolate* and *wine* in the Italian and English cultures 79
6.1. Introduction 79
6.2 Chocolate 80
6.2.1 Inter-culture analysis 80
6.2.1.1 Semantic field analysis 80
6.2.1.2 Conceptual domain analysis 84

6.2.2 Cross-cultural comparison 85
6.3 Wine 88
6.3.1 Inter-culture analysis 88
6.3.1.1 Semantic field analysis 88
6.3.1.2 Conceptual domain analysis 91
6.3.2 Cross-cultural comparison 92
6.4 Semantic field ASSESSMENT 93
6.5 Conclusions 94

Chapter 7 - Alternative routes to highlight cultural semantic associations of a given key word 97
7.1 Introduction 97
7.2 Route one: using the most frequent words in the dataset 98
7.2.1 Creating the wordlists 98
7.2.2 Coding the most frequent content words in the wordlist 99
7.2.3 Semantic fields analysis at different thresholds 100
7.2.4 Conceptual domains analysis 104
7.2.5 Semantic field ASSESSMENT 106
7.3 Route two: creating a sub-corpus by sampling using the most frequent lemmas in the dataset 106
7.3.1 Semantic fields analysis 107
7.3.2 Conceptual domains analysis 111
7.3.3 Semantic field ASSESSMENT 113
7.3.4 Conventionalisation level analysis and cross-cultural comparison 114
7.4 Route three: random sampling 119
7.4.1 Semantic fields analysis 119
7.4.2 Conceptual domains analysis 122
7.4.3 Semantic field ASSESSMENT 123
7.4.4 Conventionalisation level analysis and cross-cultural comparison 124
7.5 Conclusions 127

Chapter 8 - Alternative routes to highlight cultural semantic associations of a given key word: further experiments 131
8.1 Introduction 131
8.2 Most frequent 50/100/150/200/250/300 content words 132
8.2.1 Elicited data 132
8.2.2 Web data 133
8.3 Randomly sampled sub-corpora 134
8.3.1 Elicited data 135
8.3.2 Web data 135
8.4 Mathematical progression of field increases: a Zipf-like curve? 136
8.5 Conclusions 137

Chapter 9 - Automatic tagging 139
9.1 Introduction 139
9.2 Matching automatic tagging categories to manual coding ones 140
9.3 Analyses 142
9.4 Concluding remarks 144

Chapter 10 - Semantic associations of *chocolate* and *wine* in general Web corpora 147
10.1 Introduction 147
10.2 Sampled Web sub-corpora: creation and coding 148
10.3 Inter-culture analysis 149
10.3.1 Semantic field ASSESSMENT 151
10.4 Cross-cultural comparison 153
10.5 Concluding remarks 154

Chapter 11 - Conclusion 157
11.1 Introduction 157
11.2 Summary of experimental work and results 159
11.2.1 Retrieving cultural associations in elicited data (control situation) 159
11.2.2 Comparing Web data to the control situation 160
11.2.3 Testing different procedural approaches 161
11.2.4 Testing the use of an automatic semantic tagger 162
11.3 Some methodological considerations 163
11.4 Limitations of current work and future directions 165
11.5 Contribution to knowledge and concluding remarks 167

References 169

Appendix 183
Codebook for Semantic associations of the words Chocolate and Wine 183

List of Tables

Table 3_1 Biber's (1993:254) estimates of required sample sizes (number of texts) for a general corpus 34
Table 5_1 Sentence distribution across *chocolate* and *wine* respondents 68
Table 5_2 Elicited data summary – English 69
Table 5_3 Elicited data summary – Italian 70
Table 5_4 The Web datasets from the WACKY corpora 73
Table 5_5 *Chocolate* and *wine*: Summary of semantic fields and conceptual domains 75
Table 5_6 Semantic fields in the UCREL Semantic Analysis System tagset 77
Table 6_1 Semantic associations of *chocolate* for the English 81

Table 6_2 Semantic associations of *chocolate* for the Italians 82

Table 6_3 *Chocolate* – Percentage distribution of fields across conventionalization levels 84

Table 6_4 *Chocolate* – Conceptual domains in the English and Italian datasets 85

Table 6_5 *Chocolate* – Percentage distribution of domains across conventionalization levels 85

Table 6_6 *Chocolate* – Fields showing significant difference in the T-Test 87

Table 6_7 *Chocolate* – Conceptual domains with statistically significant differences between English and Italian 88

Table 6_8 Semantic associations of *wine* for the English 89

Table 6_9 Semantic associations of *wine* for the Italians 90

Table 6_10 *Wine* – Percentage distribution of fields across conventionalization levels 90

Table 6_11 *Wine* – Conceptual domains in the English and Italian datasets 91

Table 6_12 *Wine* – Percentage distribution of domains across conventionalization levels 91

Table 6_13 *Wine* – T-Test results 92

Table 6_14 *Wine* – Conceptual domains with statistically significant differences between English and Italian 93

Table 6_15 ASSESSMENT field results in the elicited datasets 93

Table 7_1 *Chocolate* English wordlist: Semantic fields analysis at different thresholds 101

Table 7_2 *Chocolate* Italian wordlist: Semantic fields analysis at different thresholds 101

Table 7_3 *Wine* English wordlist: Semantic fields analysis at different thresholds 101

Table 7_4 *Wine* Italian wordlist: Semantic fields analysis at different thresholds 101

Table 7_5 English *chocolate* wordlist: Semantic fields in top 300 content words 102

Table 7_6 Italian *chocolate* wordlist: Semantic fields in top 300 content words 103

Table 7_7 English *wine* wordlist: Semantic fields in top 300 content words 103

Table 7_8 Italian *wine* wordlist: Semantic fields in top 300 content words 103

Table 7_9 Conceptual domains in the *chocolate*, and *wine* datasets' most frequent 300 content words 104

Table 7_10 *Chocolate* English wordlist: Conceptual domain analysis at different thresholds 104

Table 7_11 *Chocolate* Italian wordlist: Conceptual domain analysis at different thresholds 105

Table 7_12 *Wine* English wordlist: Conceptual domain analysis at different thresholds 105

Table 7_13 *Wine* Italian wordlist: Conceptual domain analysis at different thresholds 105

Table 7_14 ASSESSMENT field results in the top 300 words 106
Table 7_15 *Chocolate* English elicited sub-corpus: Semantic fields analysis at different thresholds 108
Table 7_16 *Chocolate* Italian elicited sub-corpus: Semantic fields analysis at different thresholds 108
Table 7_17 *Wine* English elicited sub-corpus: Semantic fields analysis at different thresholds 108
Table 7_18 *Wine* Italian elicited sub-corpus: Semantic fields analysis at different thresholds 109
Table 7_19 English *chocolate*: Semantic fields in the 4-lemma sampled sub-corpus 109
Table 7_20 Italian *chocolate*: Semantic fields in 4-lemma sampled sub-corpus 110
Table 7_21 English *wine*: Semantic fields in 4-lemma sampled sub-corpus 110
Table 7_22 Italian *wine*: Semantic fields in 4-lemma sampled sub-corpus 111
Table 7_23 Conceptual domains in the *chocolate*, and *wine* sampled 4-lemma sub-corpora 112
Table 7_24 *Chocolate* English elicited sub-corpus: conceptual domains 112
Table 7_25 *Chocolate* Italian elicited sub-corpus: conceptual domains 112
Table 7_26 *Wine* English elicited sub-corpus: conceptual domains 112
Table 7_27 *Wine* Italian elicited sub-corpus: conceptual domains 112
Table 7_28 ASSESSMENT field results in the 4-lemma sampled sub-corpora 113
Table 7_29 English *chocolate* 4-lemma sub-corpus: Conventionalisation results 115
Table 7_30 Italian *chocolate* 4-lemma sub-corpus: Conventionalisation results 115
Table 7_31 English *wine* 4-lemma sub-corpus: Conventionalisation results 116
Table 7_32 Italian *wine* 4-lemma sub-corpus: Conventionalisation results 116
Table 7_33 *Chocolate* sub-corpora: T-Test results for semantic fields and conceptual domains 117
Table 7_34 *Wine* sub-corpora: T-Test results for semantic fields and conceptual domains 118
Table 7_35 English *chocolate*: Semantic fields in the randomly sampled sub-corpus 120
Table 7_36 Italian *chocolate*: Semantic fields in the randomly sampled sub-corpus 120
Table 7_37 English *wine*: Semantic fields in the randomly sampled sub-corpus 121
Table 7_38 Italian *wine*: Semantic fields in the randomly sampled sub-corpus 121
Table 7_39 Randomly sampled sub-corpora: semantic field results 121
Table 7_40 Conceptual domains in the *chocolate*, and *wine* randomly sampled sub-corpora 122
Table 7_41 Randomly sampled sub-corpora: conceptual domains results 122
Table 7_42 ASSESSMENT field results in the randomly sampled sub-corpora 123

Table 7_43 English *chocolate* random sub-corpus: Conventionalisation results 124
Table 7_44 Italian *chocolate* random sub-corpus: Conventionalisation results 124
Table 7_45 English *wine* random sub-corpus: Conventionalisation results 125
Table 7_46 Italian *wine* random sub-corpus: Conventionalisation results 125
Table 7_47 *Chocolate* random sub-corpora: T-Test results for semantic fields and conceptual domains 126
Table 7_48 *Wine* random sub-corpora: T-Test results for semantic fields and conceptual domains 127
Table 7_49 Semantic fields: Summary of results 128
Table 7_50 Conceptual domains: Summary of results 128
Table 8_1 English Elicited *chocolate*: most frequent 300 words, tagged with Wmatrix 133
Table 8_2 English Elicited *wine*: most frequent 300 words, tagged with Wmatrix 133
Table 8_3 English Web *chocolate*: most frequent 300 words, tagged with Wmatrix 134
Table 8_4 English Web *wine*: most frequent 300 words, tagged with Wmatrix 134
Table 8_5 Randomly sampled elicited sub-corpora, tagged with Wmatrix 135
Table 8_6 Randomly sampled Web sub-corpora, tagged with Wmatrix 136
Table 9_1 Conversion schemes: some examples 141
Table 9_2 Categories excluded from analysis 141
Table 9_3 English *chocolate* elicited dataset: semantic field comparison 142
Table 9_4 English *wine* elicited dataset: semantic field comparison 142
Table 9_5 English *chocolate* elicited dataset: conceptual domain comparison 143
Table 9_6 English *wine* elicited dataset: conceptual domain comparison 143
Table 10_1 Random Web sub-corpora: Semantic fields 149
Table 10_2 Summary of extra fields and domains retrieved by the random Web sub-corpora 150
Table 10_3 *Chocolate*: Semantic field ASSESSMENT 152
Table 10_4 *Wine*: Semantic field ASSESSMENT 152

List of Figures

Figure 2_1 Lotman's semiosphere 9
Figure 2_2 Synchronic section of cultural and discourse stratification (from Fleischer 1998: 444) 14
Figure 2_3 Diachronic section of cultural and discourse stratification (from Fleischer 1998: 445) 15
Figure 7_1 ASSESSMENT field results: 4-lemma sampled datasets vs. whole elicited datasets 114

Figure 7_2 ASSESSMENT field results: randomly sampled datasets vs. whole elicited datasets 123

Figure 10_1 ASSESSMENT field results: random sub-corpora vs. whole elicited datasets 152

List of Graphs

Graph 8_1 Graphic example of Zipf’s distribution 136

Graph 8_2 Data from Table 7_1 136

Graph 8_3 Data from Table 7_2 136

Graph 8_4 Data from Table 7_3 137

Graph 8_5 Data from Table 7_4 137

Graph 8_6 Data from Table 8_1 137

Graph 8_7 Data from Table 8_2 137

Graph 8_8 Data from Table 8_3 137

Graph 8_9 Data from Table 8_4 137

Graph 9_1 Data from Table 9_3 143

Graph 9_2 Data from Table 9_4 143