A CORPUS-BASED STUDY OF ADJECTIVES IN CONTEMPORARY ENGLISH

CAO JING

DOCTOR OF PHILOSOPHY
CITY UNIVERSITY OF HONG KONG
AUGUST 2011
CITY UNIVERSITY OF HONG KONG
香港城市大學

A Corpus-Based Study of Adjectives in Contemporary English
基於語料庫的當代英語形容詞研究

Submitted to
Department of Chinese, Translation and Linguistics
中文、翻譯及語言學系
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
哲學博士學位

by

Cao Jing
曹競

August 2011
二零一一年八月
ABSTRACT

This thesis describes a systematic study of adjectives in contemporary English. In particular, the study employs large collections of naturally occurring texts to investigate the use of adjectives against different text categories as embodied in the corpora. A major objective is to chart the correlation between different types of texts and the variational characteristics of adjectival constructions. The study draws on the most recent development in corpus linguistics as well as the state of the art in artificial intelligence in general and machine learning techniques in particular for the extraction and analysis of differentiating adjectival features in relation to a taxonomy of text categories. Three major research issues are identified, including 1) the correlation between adjectives and text categories, 2) the correlation between adjectives and subject domains, and 3) the possible application of adjectives in the field of automatic text classification.

Two corpora are employed in the investigation. The British National Corpus (BNC) is used for lexical investigation and the British component of International Corpus of English (ICE-GB) is used for the exploration of syntactic features. While the empirical data obtained from the BNC confirms the findings from previous studies that adjectives are preferred in writing than in speech, and that adjectives tend to occur more often in more formal texts than in less formal texts, an analysis of the textual dispersion of adjectives across different text categories has further suggested that adjectives are more likely to be a stylistic factor rather than a content factor. Linear regression analyses also
show a strong positive correlation between the distributional probability of adjective use and text formality. Cluster analyses reveal that, using the value of *Chi by degrees of freedom* computed on the combination of chi-square and adjective frequency, text categories that are stylistically related to each other can be meaningfully grouped together. Machine learning techniques available in Weka are used to evaluate the performance of adjectival probability in automatic text classification in terms of precision, recall and *F*-measure. The results suggest that adjectives can be used as a strong characteristic for automatic text classification.

The syntactic features of adjective phrases (AJPs) in the ICE-GB are examined in terms of their internal structures and external, clausal functions. In addition to the survey of distributions of syntactic features, multiple regression analyses suggest a significant correlation between the distributional probability of syntactic properties and text formality. Such a correlation is then tested by using machine learning techniques. Results show that the AJP syntactic probability can contribute to automatic text classification, and that the performance of AJP syntactic probability differs in different classification tasks.

In addition, adjectives in terminology are investigated to provide additional information about adjectives in specialized domains. A sub-corpus is constructed specially for the study out of the ICE-GB, and is annotated at the terminological level. The distribution of adjectives in term expressions, defined as term-ADJs, is examined across eight subject domains. Linear regression analyses of the observed distributional features
suggest a strong positive correlation between the use of term-ADJs and text formality, and results obtained from machine learning techniques show that term-ADJs in general contribute more to text classification than nouns, verbs and adverbs.

To sum up, the research described in this thesis represents the following unique contributions to the existing body of knowledge. First, it is the first large-scale corpus-based empirical study of adjectives in present-day British English. Secondly, it is the first corpus-based linguistic study that proposes the use of machine learning techniques to help provide linguistic insight into language use. Significant results have been achieved. Firstly, the investigations have answered some of the fundamental questions concerning the distributional behavior of adjectives across a broad range of text categories and subject domains. Secondly, the study has also lent some profound insight into the relationship between the distributional probability of adjectives and text formality, hence addressing at the same time the issue of applying adjectives to the automatic measurement of text similarity for automatic text classification. Thus, this research has made several contributions through 1) in-depth experiments on the correlation between adjectives and formality degree of texts, 2) use of adjectives to measure text similarity, 3) in-depth investigation of adjective use in terminology, and 4) in-depth experiments on the use of adjectival characteristics in the field of automatic text classification. The findings in this research will serve as the basis for the future work to build a regression model so as to express the relationship between adjectives and text categories.
CONTENTS

ABSTRACT ................................................................................................................................................. i
ACKNOWLEDGMENTS .............................................................................................................................. iv
CONTENTS .................................................................................................................................................. vi
LIST OF TABLES ....................................................................................................................................... xii
LIST OF FIGURES ..................................................................................................................................... xv

PART I INTRODUCTION, LITERATURE REVIEW, AND METHODOLOGY .............................................. 1

CHAPTER 1 INTRODUCTION .................................................................................................................. 2
  1.1 Adjectives ......................................................................................................................................... 2
  1.2 Motivations ....................................................................................................................................... 4
  1.3 Objectives ....................................................................................................................................... 6
  1.4 Organization of the Dissertation ................................................................................................... 8

CHAPTER 2 LITERATURE REVIEW ....................................................................................................... 11
  2.1 What are Adjectives ....................................................................................................................... 11
    2.1.1 The Syntactic Characteristics ................................................................................................. 12
    2.1.2 The Semantic Characteristics ............................................................................................... 14
    2.1.3 Summary .................................................................................................................................. 16
  2.2 Theoretical Studies ......................................................................................................................... 17
    2.2.1 Ferris (1993) .......................................................................................................................... 18
    2.2.2 Tucker (1998) ........................................................................................................................ 21
  2.3 Corpus-based Studies ..................................................................................................................... 24
    2.3.1 Syntactic Characteristics of Adjectives .................................................................................. 25
    2.3.2 Distributional Characteristics of Adjectives ......................................................................... 28
      2.3.2.1 Biber (1988) .................................................................................................................... 29
      2.3.2.2 Biber et al. (1999) ........................................................................................................... 30
      2.3.2.3 Rayson et al. (2000) ......................................................................................................... 32
2.3.2.4 Yamazaki (2002) ................................................................. 33
2.3.3 Adjectives in Natural Language Processing ........................... 34
2.4 Discussion and Summary ....................................................... 35

CHAPTER 3 METHODOLOGY ................................................................. 38
3.1 Genres, Registers, and Text Typology .................................... 38
3.2 Research Questions ............................................................... 39
3.3 Statistical Analysis and Machine Learning ............................ 42

PART II THE FORM TYPES OF ADJECTIVES ..................................... 46

CHAPTER 4 THE BNC AS A CORPUS RESOURCE ............................ 48
4.1 Overview ............................................................................... 48
4.2 Corpus Composition ............................................................. 49
4.3 Encoding Scheme .................................................................. 51
4.3.1 BNC Documents ............................................................ 51
4.3.2 BNC Text Categories ...................................................... 51
4.3.3 BNC Word Units and Parts of Speech .............................. 52
4.4 C5 Tagset for Adjectives in the BNC ..................................... 54

CHAPTER 5 THE DISTRIBUTION OF THE FORM TYPES OF ADJECTIVES ....... 57
5.1 Overview ............................................................................... 57
5.2 Creation of Sub-corpus and Extraction of Adjectives .............. 58
5.3 Proportion of Adjective Tokens .......................................... 60
5.4 Textual Dispersion of Adjective Types ................................. 62
5.5 Chapter Summary ............................................................... 65

CHAPTER 6 ADJECTIVE OCCURRENCES AND TEXT CATEGORIES .................. 67
6.1 Overview ............................................................................... 67
6.2 Creation of Training and Test Datasets ................................. 68
6.3 Adjectives and Text Formality ............................................. 70
PART III THE SYNTACTIC CHARACTERISTICS OF ADJECTIVES

CHAPTER 7 THE ICE-GB AS A CORPUS RESOURCE

7.1 Overview
7.2 Corpus Structure
7.3 Annotation Scheme
7.3.1 General Annotation Scheme
7.3.2 Part-of-Speech Annotation for Adjectives
7.3.3 Syntactic Annotation for Adjectives Phrases
7.4 Comparison with the BNC

CHAPTER 8 THE DISTRIBUTION OF ADJECTIVE PHRASES

8.1 Overview
8.2 Distribution of Adjective Phrases across Text Categories
10.4 Corpus Construction ................................................................. 148
  10.4.1 Corpus Resource for Term Annotation .......................... 149
  10.4.2 Term Annotation ............................................................ 150
    10.4.2.1 Annotation Guidelines ............................... 151
    10.4.2.2 Inter-annotator Agreement .......................... 153
    10.4.2.3 Basic Statistics .............................................. 157
  10.5 Distribution of Term-ADJs across Domains ..................... 158
    10.5.1 Adjectives in Term Expressions ............................. 159
    10.5.2 Term-ADJs across Domains ................................. 164
  10.6 Distributional Probability of Term-ADJs and Domains ....... 168
    10.6.1 Linear Regression Analysis ................................. 168
    10.6.2 Application in Text Classification ....................... 169
      10.6.2.1 Pre-processing ........................................ 170
      10.6.2.2 Experimental Results ............................... 172
  10.7 Chapter Summary ............................................................. 176

PART V CONCLUSION ......................................................................................... 177

CHAPTER 11 CONCLUSIONS AND FUTURE WORK ............................................ 178
  11.1 Conclusions ................................................................. 178
  11.2 Future Work ................................................................. 182

BIBLIOGRAPHY ................................................................................................. 184

APPENDIX A: Two Sets of Word Class Tags in the BNC ................. 191
APPENDIX B: A Frequency List of Common Adjectives across Text Categories .... 192
APPENDIX C.1: A Frequency List of Single-Category Adjectives – CONV ........... 205
APPENDIX C.2: A Frequency List of Single-Category Adjectives – OSP ............ 206
APPENDIX C.3: A Frequency List of Single-Category Adjectives – AC ............. 207
APPENDIX C.4: A Frequency List of Single-Category Adjectives – FIC ............ 208
APPENDIX C.5: A Frequency List of Single-Category Adjectives – NAC ............ 209
APPENDIX C.6: A Frequency List of Single-Category Adjectives – NEWS ........... 210
LIST OF TABLES

TABLE 2.1 Semantic features of adjectives .......................................................... 16
TABLE 2.2 Adjectival constructions of English .................................................. 18
TABLE 2.3 Relations of attributive adjectives .................................................... 26
TABLE 2.4 Construction of postmodifying adjectives ......................................... 27
TABLE 2.5 Descriptive statistics of attributive and predicative adjectives .......... 30
TABLE 2.6 Corpus composition (Biber et al., 1999) ......................................... 31
TABLE 2.7 Corpus composition (Rayson et al., 2000) ....................................... 32
TABLE 4.1 The composition of the BNC ............................................................ 50
TABLE 4.2 Specification of c5 tags for adjectives ............................................. 54
TABLE 4.3 Specification of ambiguous tags concerning adjectives ................. 55
TABLE 5.1 Basic statistics of the sub-corpus ...................................................... 58
TABLE 5.2 Basic statistics of adjectives across text categories ......................... 60
TABLE 5.3 The textual dispersion of adjective types in BNC ............................ 63
TABLE 6.1 Basic statistics of the training set ...................................................... 69
TABLE 6.2 Basic statistics of the test set ............................................................ 63
TABLE 6.3 Manual ranking of the eight text categories .................................... 73
TABLE 6.4 Adjective density of the training set ................................................ 74
TABLE 6.5 Manual vs. automatic rankings in the training set ............................ 75
TABLE 6.6 Results from linear regression test ................................................... 77
TABLE 6.7 Basic statistics of adjectives in the test set ...................................... 79
TABLE 6.8 Expected automatic ranking of the test set ..................................... 80
TABLE 6.9 Calculation of the chi-square in two text categories ....................... 83
TABLE 6.10 CBDF values of 8 text categories .................................................. 84
TABLE 6.11 CBDF values of 5 text categories .................................................. 86
TABLE 6.12 Precision, recall and F-measure for two category classification ....... 88
TABLE 6.13 Precision, recall and F-measure for eight category classification .... 89
TABLE 6.14 Precision, recall and F-measure for five category classification ...... 90
TABLE 7.1 The composition of the ICE-GB ...................................................... 96
TABLE 7.2 Three levels of text category classification in the ICE-GB ........................................97
TABLE 7.3 POS Annotations of adjectives .................................................................99
TABLE 7.4 General comparison between the BNC and ICE-GB .................................102
TABLE 8.1 Basic statistics of adjective tokens in the ICE-GB ....................................104
TABLE 8.2 Basic statistics of AJPs across super categories ......................................108
TABLE 8.3 Basic statistics of AJPs across macro categories ......................................108
TABLE 8.4 Basic statistics of AJPs across micro categories ......................................109
TABLE 8.5 Automatic vs. manual rankings across micro categories ..........................111
TABLE 8.6 Results from Spearman rank correlation .................................................112
TABLE 8.7 Distribution of AJP internal structures across super categories .........113
TABLE 8.8 Distribution of AJP internal structures across macro categories ............114
TABLE 8.9 Distribution of AJP internal structures across micro categories ............115
TABLE 8.10 Syntactic functions of AJPs across super categories ..........................120
TABLE 8.11 Syntactic functions of AJPs across macro categories ..........................121
TABLE 8.12 Syntactic functions of AJPs across micro categories ..........................123
TABLE 9.1 Results from ANOVA tests .................................................................130
TABLE 9.2 Results from multiple regression tests ...............................................131
TABLE 9.3 Weighted average accuracy for super categories .....................................133
TABLE 9.4 Weighted average accuracy for macro categories ................................134
TABLE 9.5 Weighted average accuracy for micro categories ................................134
TABLE 9.6 Results from ANOVA tests .................................................................135
TABLE 9.7 Results from multiple regression tests ...............................................136
TABLE 9.8 Weighted average accuracy for super categories .....................................138
TABLE 9.9 Weighted average accuracy for macro categories ................................138
TABLE 9.10 Weighted average accuracy for micro categories ................................138
TABLE 9.11 Results from multiple regression tests ...............................................140
TABLE 9.12 Weighted average accuracy for the best subsets ................................141
TABLE 10.1 The structure of the sub-corpus .........................................................150
TABLE 10.2 Results from inter-annotator experiment ..........................................154
TABLE 10.3 Interpretation of Kappa .................................................................155
TABLE 10.4 A summary of the inter-annotator agreement ..................................156
TABLE 10.5 Basic statistics of the newly built corpus..................................................157
TABLE 10.6 Basic statistics of term-ADJs in terms.......................................................159
TABLE 10.7 Basic statistics of single-domain term-ADJ types.......................................160
TABLE 10.8 Basic statistics of term-ADJs in termNP constructions.................................162
TABLE 10.9 Proportion of term-ADJs in for main termNP constructions............................162
TABLE 10.10 Proportion of adjectives across domains ...................................................164
TABLE 10.11 Term-ADJ vs. Non-term-ADJs.....................................................................165
TABLE 10.12 Proportion of term-ADJs across domains....................................................167
TABLE 10.13 Automatic vs. manual rankings .................................................................168
TABLE 10.14 Results from linear regression test............................................................169
TABLE 10.15 The structure of the dataset for textual classification.................................170
TABLE 10.16 The structure of the dataset for stylistic classification...............................171
TABLE 10.17 The structure of the dataset for topical classification..................................171
TABLE 10.18 Weighted average accuracy for textual classification...............................173
TABLE 10.19 Weighted average accuracy for stylistic classification..............................173
TABLE 10.20 Weighted average accuracy for topical classification...............................174
LIST OF FIGURES

Fig. 2.1 The potential structure of $qlgp$ (adjectival apex) ..............................................22
Fig. 2.2 Distribution of adjectives across texts..................................................................33
Fig. 3.1 A screenshot of Weka explorer .........................................................................45
Fig. 4.1 Structure of $A00.xml$ in the BNC ...................................................................51
Fig. 4.2 An example of an annotated sentence in the BNC ..............................................53
Fig. 4.3 An example of $AJO-AVO$ case from $AOA.xml$ ................................................55
Fig. 5.1 Proportion of adjectives across text categories .................................................61
Fig. 5.2 Textual dispersion of adjective types across text categories ..............................63
Fig. 6.1 A hypothetical plot relating text categories to formality degrees ......................67
Fig. 6.2 Graph of adjective density by manual ranking ..................................................76
Fig. 6.3 Graph of $Y=-2.652+1.054X$ ........................................................................78
Fig. 6.4 A dendrogram of the eight BNC text categories ...............................................85
Fig. 6.5 A dendrogram of the five BNC text categories ..................................................86
Fig. 6.6 The matrix of adjective performance in eight category classification .......... 86
Fig. 7.1 An example of multi-layer annotations in the ICE-GB ....................................98
Fig. 7.2 Example of an adjective phrase ......................................................................100
Fig. 8.1 AJP density across super categories ................................................................108
Fig. 8.2 AJP density across macro categories ..............................................................109
Fig. 8.3 AJP density across micro categories ...............................................................110
Fig. 8.4 AJP internal structure across super categories ................................................113
Fig. 8.5 AJP internal structure across macro categories ...............................................114
Fig. 8.6 Internal structure of $AJHD$ across micro categories .......................................116
Fig. 8.7 Internal structure of $AJPR+AJHD$ across micro categories .........................117
Fig. 8.8 Internal structure of $AJHD+AJPO$ across micro categories .........................117
Fig. 8.9 Internal structure of $AJPR+AJHD+AJPO$ across micro categories ............ 118
Fig. 8.10 Distribution of the four main functions across macro categories .................122
Fig. 8.11 Distribution of $NPPR$ across micro categories ...............................................124
Fig. 8.12 Distribution of $CS$ across micro categories ....................................................124
Fig. 8.13 Distribution of NPPO across micro categories ......................................................... 125
Fig. 8.14 Distribution of CO across micro categories ............................................................. 125
Fig. 9.1 Comparison of $F$-measures .................................................................................... 142
Fig. 10.1 An example of single-word terms .............................................................................. 152
Fig. 10.2 An example of multi-word terms .............................................................................. 152
Fig. 10.3 Proportion of terms across ICE-GB domains .......................................................... 158
Fig. 10.4 Proportion of term-ADJs against terms ................................................................. 160
Fig. 10.5 An example of a termNP with two adjectives ........................................................ 161
Fig. 10.6 Proportion of term-ADJs in four main termNP constructions ............................... 163
Fig. 10.7 Adjective density across ICE-GB domains ............................................................. 165
Fig. 10.8 Proportion of term-ADJs vs. non-term-ADJs across ICE-GB domains ................. 166