NARROWING THE GAP BETWEEN TERMBASES AND CORPORA IN COMMERCIAL ENVIRONMENTS

KARA CORDELIA WARBURTON

DOCTOR OF PHILOSOPHY

CITY UNIVERSITY OF HONG KONG

JULY 2014

CITY UNIVERSITY OF HONG KONG
香港城市大學

Narrowing the Gap Between Termbases and Corpora in Commercial Environments
商用術語庫和語料庫之間的協調問題研究

Submitted to
Department of Linguistics and Translation
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

Kara Cordelia Warburton

July 2014
二零一四年七月

ABSTRACT

This research investigates the terminological data in terminology databases (termbases) and in corresponding corpora from commercial sources. Four companies in the information technology (IT) sector are used as case studies. Our broad objective is to increase awareness about some of the issues and challenges faced by terminologists in commercial settings. We demonstrate that there are significant gaps between the termbases and the corresponding corpora, that such gaps reduce the effectiveness of the termbases, and that they can be minimised by adopting a corpus-based approach to term identification.

We begin by establishing that the language used in a company contains terminology. After reviewing the conventional theories and methodologies of the field of terminology, we challenge the suitability of some of their precepts for companies that require terminological resources that are both repurposable and production-oriented. We then reveal features in the termbases that depart from established norms. Using a batch concordance technique, we quantify the gap between the termbase terms and the corpora. We then attempt to explain this gap by examining termbase terms that occur in various frequency ranges within the corpora. Using empirical observations, we formulate some guiding principles for selecting terms for termbases with respect to various features including term length, part of speech, term variation, and the use of certain types of modifiers.

We discover that keywords hold potential for discovering multi-word terms that, if documented in termbases, would significantly increase the correspondence between termbases and corpora. We conclude that termbases developed in companies would increase in value if corpus-based approaches to term identification were adopted. This challenges the conventional understanding of what is a term; to open the field of terminology to commercial applications and environments, termhood needs to be established based on communicative purpose and end-use of terminological resources in addition to purely semantic criteria.

Keywords: terminology, terminography, termbases, corpora, keywords, LSP

CONTENTS

ABSTRACT.....i
CITY UNIVERSITY OF HONG KONG.....iii
ACKNOWLEDGEMENTS.....iii
TYPOGRAPHICAL CONVENTIONS.....iv
LIST OF TABLES.....v
LIST OF FIGURES.....viii
CHAPTER 1 INTRODUCTION AND MOTIVATION.....1
1.1 About this thesis.....1
1.2 Key terms and definitions.....2
1.3 Terminology as a discipline and a vocation.....3
1.4 Terminology management and terminography.....6
1.5 Applications of terminological resources.....8
1.6 Implications for commercial applications.....8
1.7 Motivation of the current research.....12
1.7.1 The close ties to translation.....13
1.7.2 The restricted focus of termbases.....15
1.7.3 The need for terminological resources to serve multiple purposes.....17
1.7.4 Improving term identification.....21
CHAPTER 2 LITERATURE REVIEW.....23
2.1 Terminology and LSP.....23
2.1.1 The role of subject field.....23
2.1.2 A closed set of linguistic properties.....26
2.1.3 Communicative context and communicative function.....26
2.1.4 Conscious acquisition.....27
2.2 Terminology and genre.....28
2.3 What is a term?.....30
2.3.1 Disambiguating term.....31
2.3.2 Theoretical interpretations.....35
2.3.2.1 General Theory of Terminology.....35
2.3.2.2 Socio-cognitive Theory.....39

2.3.2.3 Lexico-semantic Theory and Textual Terminology....40
2.3.2.4 Communicative Theory....42
2.3.3 Views on variation....42
2.3.4 Predominance of nominal forms....47
2.3.5 Predominance of multi-word terms....51
2.4 Methodologies....52
2.4.1 Onomasiological vs semasiological approaches....52
2.4.2 Thematic vs ad-hoc methodologies....55
2.5 The contributions of corpus linguistics....56
2.5.1 Corpus linguistics and lexicography....56
2.5.2 Corpus linguistics and terminology....57
2.6 Summary....61
CHAPTER 3 CRITICAL DISCUSSION OF THE LITERATURE....65
3.1 Company-specific language as an LSP....65
3.2 The notion of term, in commercial environments....66
3.2.1 Purpose or application of terms....67
3.2.2 Importance of non-nouns....68
3.2.3 Prevalence of variants....70
3.2.4 Semi-technical vocabulary....73
3.3 Views on theory and methodology....74
3.4 The role of corpora....75
3.5 Genre as a deterministic factor for terminology....76
3.6 Summary....78
CHAPTER 4 RESEARCH OBJECTIVES....80
4.1 Research questions....80
4.2 Research methodology....81
4.3 Expected outcomes....85
CHAPTER 5 DESCRIPTION AND PREPARATION OF THE DATA....87
5.1 Description of the data....87
5.1.1 Minitab....87
5.1.2 SAS....91

5.1.3 Symantec....92
5.1.4 Hewlett Packard....95
5.1.5 Summary....98
5.2 Preparation of the data....99
5.2.1 Preparing the corpora....99
5.2.1.1 Problems and issues....99
5.2.1.2 Minitab....101
5.2.1.3 SAS....103
5.2.1.4 Symantec....103
5.2.1.5 Hewlett Packard....105
5.2.1.6 Summary of changes....105
5.2.2 Preparing the termbases....106
5.2.2.1 Minitab....106
5.2.2.2 SAS....117
5.2.2.3 Symantec....118
5.2.2.4 Hewlett Packard....119
5.2.2.5 Corpus-valid terms....119
CHAPTER 6 ANALYSIS OF THE DATA....121
6.1 Analysing the termbases....121
6.1.1 Review of key standards....121
6.1.2 Entry model and data categories....123
6.1.3 Size of the corpus in relation to the termbase....127
6.1.4 Observations....129
6.2 Analysing the termbase terms....131
6.2.1 Frequency....131
6.2.1.1 Normalising the frequency counts....132
6.2.1.2 Average frequency of termbase terms....133
6.2.1.3 Establishing comparable frequency ranges....134
6.2.1.4 Number of termbase terms that occur at frequency ranges....135
6.2.2 Case....137
6.2.3 Length....137

6.2.4 Word class....140
6.2.5 Variants....141
6.2.5.1 Minitab....143
6.2.5.2 SAS....146
6.2.5.3 Symantec....147
6.2.5.4 Hewlett-Packard....147
6.2.6 Observations....149
6.3 Termbase terms that do not occur in the corpus....151
6.3.1 Distribution....152
6.3.2 Differences in case....152
6.3.3 Differences in number....155
6.3.4 Term length....159
6.3.4.1 Resetting the boundaries of MWTs....161
6.3.5 Observations....168
6.4 Termbase terms that occur infrequently in the corpus....169
6.4.1 Distribution in the termbase....169
6.4.2 Term length....171
6.4.3 Other properties....173
6.4.4 Observations....176
6.5 Termbase terms that occur frequently in the corpus....177
6.5.1 Distribution in the termbase....177
6.5.2 Term length....178
6.5.3 Other properties....181
6.5.3.1 Validation of verbs....185
6.5.4 Observations....188
6.6 Verbs in the corpus....189
6.7 Variants in the corpus....192
6.8 Observations....194
CHAPTER 7 EXPLORING KEYWORDS....196
7.1 Potential significance and related research....196
7.2 Keyword identification....199

7.3 Keyword categorisation....203
7.4 Frequency of keywords versus frequency of termbase terms....203
7.5 Keywords that are under-represented in the termbases....205
7.5.1 Top-ranking keywords....206
7.5.1.1 Minitab....206
7.5.1.2 SAS....207
7.5.1.3 Symantec....210
7.5.1.4 Summary....212
7.5.2 Mid- and low-ranking keywords....213
7.5.2.1 Minitab....213
7.5.2.2 SAS....214
7.5.2.3 Symantec....215
7.5.2.4 Summary....215
7.5.3 Keywords that are non-existent or rare in the reference corpus....216
7.5.3.1 Minitab....217
7.5.3.2 SAS....218
7.5.3.3 Symantec....219
7.5.3.4 Summary....220
7.6 Collocate relationship measures....220
7.6.1 Log likelihood....221
7.6.2 Z-score....223
7.6.3 Specific Mutual Information....224
7.6.4 Dice Coefficient....225
7.6.5 MI3....226
7.6.6 T-Score....227
7.6.7 Comparison and selection....228
7.7 Concordances and collocations....230
7.7.1 Top-ranking keywords....231
7.7.1.1 Minitab....231
7.7.1.2 SAS....236
7.7.1.3 Symantec....239

7.7.1.4 Summary....246
7.7.2 Mid- and low-ranking keywords....247
7.7.2.1 Minitab....247
7.7.2.2 SAS....250
7.7.2.3 Symantec....252
7.7.2.4 Summary....256
7.7.3 Keywords that are non-existent or rare in the reference corpus....257
7.7.3.1 Minitab....257
7.7.3.2 SAS....258
7.7.3.3 Symantec....260
7.7.3.4 Summary....262
CHAPTER 8 CONCLUSIONS AND IMPLICATIONS....263
8.1 The gap between termbases and corpora....263
8.2 Economic impacts....267
8.3 A purpose-driven notion of terminography....270
8.4 Implications for theory and practise....272
8.5 Further reflections....276
8.6 Limitations and further research....277
BIBLIOGRAPHY....280
APPENDIX A – Words and expressions removed from the Minitab termbase....295
APPENDIX B – Words and expressions removed from the Symantec termbase....300
APPENDIX C – Data category description for the Minitab termbase....301
APPENDIX D – Software used in the research....303
APPENDIX E – List of abbreviations....304
APPENDIX F – Calculation formulae for the collocate relationship measures....305
APPENDIX G – Sample legal agreement....307