Contents

1 Introduction 1

1.1 Lexical association 1

1.1.1 Collocational association 2

1.1.2 Semantic association 2

1.1.3 Cross-language association 3

1.2 Motivation and applications 4

1.3 Goals and objectives 6

2 Theory and Principles 11

2.1 Notion of collocation 11

2.1.1 Lexical combinatorics 11

2.1.2 Historical perspective 12

2.1.3 Diversity of definitions 14

2.1.4 Typology and classification 18

2.1.5 Conclusion 22

2.2 Collocation extraction 23

2.2.1 Extraction principles 23

2.2.2 Extraction pipeline 26

2.2.3 Linguistic preprocessing 26

2.2.4 Collocation candidates 28

2.2.5 Occurrence statistics 31

2.2.6 Filtering candidate data 33

3 Association Measures 39

3.1 Statistical association 39

3.2 Context analysis 43

4 Reference Data 49

4.1 Requirements . . . 49

4.1.1 Candidate data extraction . . . 49

4.1.2 Annotation process . . . 50

4.2 Prague Dependency Treebank . . . 51

4.2.1 Treebank details . . . 51

4.2.2 Candidate data sets . . . 53

4.2.3 Manual annotation . . . 56

4.3 Czech National Corpus . . . 58

4.3.1 Corpus details . . . 59

4.3.2 Automatic preprocessing . . . 59

4.3.3 Candidate data set . . . 61

4.4 Swedish PAROLE corpus . . . 61

4.4.1 Corpus details . . . 62

4.4.2 Support-verb constructions . . . 62

4.4.3 Manual extraction . . . 63

5 Empirical Evaluation **65**

5.1 Evaluation methods . . . 65

5.1.1 Precision-recall curves . . . 66

5.1.2 Mean average precision . . . 68

5.1.3 Significance testing . . . 70

5.2 Experiments . . . 70

5.2.1 Prague Dependency Treebank . . . 71

5.2.2 Czech National Corpus . . . 73

5.2.3 Swedish PAROLE Corpus . . . 74

5.3 Comparison . . . 76

6 Combining Association Measures **79**

6.1 Motivation . . . 79

6.2 Methods . . . 79

6.2.1 Linear logistic regression . . . 80

6.2.2 Linear discriminant analysis . . . 81

6.2.3 Support vector machines . . . 81

6.2.4 Neural networks . . . 81

6.3 Experiments . . . 82
6.3.1 Prague Dependency Treebank . . . 83
6.3.2 Czech National Corpus . . . 84
6.3.3 Swedish PAROLE Corpus . . . 85
6.4 Linguistic features . . . 86
6.5 Model reduction . . . 87
6.5.1 Algorithm . . . 88
6.5.2 Experiments . . . 89

7 Conclusions **93**

A MWE 2008 Shared Task Results **97**
A.1 Introduction . . . 97
A.2 System overview . . . 97
A.3 German Adj-Noun collocations . . . 99
A.3.1 Data description . . . 99
A.3.2 Experiments and results . . . 99
A.4 German PP-Verb collocations . . . 100
A.4.1 Data description . . . 100
A.4.2 Experiments and results . . . 100
A.5 Czech PDT-Dep collocations . . . 102
A.5.1 Data description . . . 102
A.5.2 Experiments and results . . . 103
A.6 Conclusion . . . 103

B Complete Evaluation Results **105**
B.1 *PDT-Dep* . . . 106
B.2 *PDT-Surf* . . . 107
B.3 *CNC-Surf* . . . 108
B.4 *PAR-Dist* . . . 109

Summary **111**

Bibliography **113**

Index **127**