

Contents

<i>List of Contributors</i>	page xi
<i>Foreword</i>	xiii
<i>Introduction</i>	xvi
Part I Fundamentals	
1 Language Variation	3
JAMES A. WALKER	
1.1 Introduction: Defining Language Variation	3
1.2 Types of Linguistic Variables	5
1.3 Dimensions of Variation	7
1.4 Conclusion	15
2 Phonetic Variation in Dialects	17
RACHAEL TATMAN	
2.1 Introduction	17
2.2 Vowels	22
2.3 Consonants	23
2.4 Suprasegmentals	24
2.5 Conclusion	24
3 Similar Languages, Varieties, and Dialects: Status and Variation	27
MIRIAM MEYERHOFF AND STEFFEN KLAERE	
3.1 Introduction	27
3.2 Language: More Than Communication	28
3.3 Language and Dialect	28
3.4 Creoles as a Class of Natural Languages	32
3.5 Introducing the Corpus of Spoken Bequia Creole English	34
3.6 Transforming Spoken Word into Categorical Data	36
3.7 Feature Interrelationship	38
3.8 Graphical Models to Visualise Interactions	40
3.9 Feature Relations at a Marginalised Level	42
3.10 Feature Interrelationship within Communities	43
3.11 Community Distinction	44
	vii

viii	<i>Contents</i>	
	3.12 Speaker-Specific Creole Frequency	45
	3.13 Conclusions: Principled Methods for Exploring Systematic Dialects within Languages	48
4	Mutual Intelligibility	51
	CHARLOTTE GOOSKENS AND VINCENT J. VAN HEUVEN	
	4.1 Introduction	51
	4.2 How to Measure Intelligibility	58
	4.3 Extra-linguistic and Para-linguistic Factors Influencing Intelligibility	67
	4.4 Linguistic Determinants of Intelligibility	70
	4.5 Relationship between Intelligibility and Language Trees	83
	4.6 Conclusions, Discussion, and Desiderata for Future Research	87
5	Dialectology for Computational Linguists	96
	JOHN NERBONNE, WILBERT HEERINGA, JELENA PROKIĆ AND MARTIJN WIELING	
	5.1 Introduction	96
	5.2 Dialectology	96
	5.3 Dialectometry	98
	5.4 Edit Distance on Phonetic Transcriptions	100
	5.5 Geography of Distributions	104
	5.6 Validation	108
	5.7 Emerging Opportunities and Issues	111
	5.8 Conclusions	112
Part II Methods and Resources		
6	Data Collection and Representation for Similar Languages, Varieties and Dialects	121
	TANJA SAMARDŽIĆ AND NIKOLA LJUBEŠIĆ	
	6.1 Representing Language Variability in Corpora	122
	6.2 Types of Micro-Variation and the Corresponding Data Collection Procedures	124
	6.3 Privacy and Linguistic Micro-Variation	133
	6.4 Conclusion	134
7	Adaptation of Morphosyntactic Taggers	138
	YVES SCHERRER	
	7.1 Introduction	138
	7.2 Model Transfer Methods	145
	7.3 Normalization and Other Data Transfer Methods	149
	7.4 Tagger Adaptation and Multilingual Models	156
	7.5 Conclusions	158

Contents	ix
8 Sharing Dependency Parsers between Similar Languages	167
ŽELJKO AGIĆ	
8.1 Introduction	167
8.2 A New Hope? Notable Exceptions	172
8.3 To Conclude: A Glimpse of the Future	178
 Part III Applications and Language Specific Issues	
9 Dialect and Similar Language Identification	187
MARCOS ZAMPIERI	
9.1 Introduction	187
9.2 A Supervised Text Classification Problem	189
9.3 Collecting Data	190
9.4 Competitions	192
9.5 DSL Shared Task 2015	195
9.6 Conclusion and Future Perspectives	198
10 Dialect Variation on Social Media	204
DONG NGUYEN	
10.1 Introduction	204
10.2 Social Media for Dialect Research	206
10.3 Processing Data	209
10.4 Patterns in Social Media	210
10.5 Future Outlook	213
11 Machine Translation between Similar Languages	219
PRESLAV NAKOV AND JÖRG TIEDEMANN	
11.1 Introduction	219
11.2 Models and Approaches	219
11.3 Character-Level Machine Translation	223
11.4 Closely Related Languages as MT Pivots	228
11.5 Bitext Combination	234
11.6 Language Adaptation	238
11.7 Other Approaches	243
11.8 Applications and Future Directions	244
11.9 Conclusions	246
12 Automatic Spoken Dialect Identification	254
PEDRO A. TORRES-CARRASQUILLO AND BENGT J. BORGSTRÖM	
12.1 Introduction	254
12.2 Background	255
12.3 Resources for Dialect Identification	260
12.4 State of the Art	265
12.5 Standardized Evaluations and Recent Performance	269
12.6 Challenges and Future Outlook	272

x	<i>Contents</i>	
13	Arabic Dialect Processing	279
	NIZAR HABASH	
	13.1 Introduction	279
	13.2 Arabic and Its Variants	279
	13.3 Challenges of Arabic Dialect Processing	282
	13.4 Arabic Dialect Resources	285
	13.5 Arabic Dialect Processing Tools and Applications	289
	13.6 Conclusion and Outlook	293
14	Computational Processing of Varieties of Chinese: Comparable Corpus-Driven Approaches to Light Verb Variation	304
	MENGHAN JIANG, HONGZHI XU, JINGXIA LIN, DINGXU SHI, AND CHU-REN HUANG	
	14.1 Introduction	304
	14.2 Computational Approaches to Language Variations in Chinese and Other Languages	310
	14.3 Classification of Varieties of Mandarin Chinese	316
	14.4 Conclusion	319