Contents

List of Contributors xi
Foreword xiii
Introduction 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
# Table of Contents

## Part I Methods and Resources

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.12</td>
<td>Speaker-Specific Creole Frequency</td>
<td>45</td>
</tr>
<tr>
<td>3.13</td>
<td>Conclusions: Principled Methods for Exploring Systematic Dialects within Languages</td>
<td>48</td>
</tr>
<tr>
<td>4</td>
<td>Mutual Intelligibility</td>
<td>51</td>
</tr>
<tr>
<td>5</td>
<td>Dialectology for Computational Linguists</td>
<td>96</td>
</tr>
<tr>
<td>6</td>
<td>Data Collection and Representation for Similar Languages, Varieties and Dialects</td>
<td>121</td>
</tr>
<tr>
<td>7</td>
<td>Adaptation of Morphosyntactic Taggers</td>
<td>138</td>
</tr>
</tbody>
</table>

---

## Part II Methods and Resources

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>51</td>
</tr>
<tr>
<td>4.2</td>
<td>How to Measure Intelligibility</td>
<td>58</td>
</tr>
<tr>
<td>4.3</td>
<td>Extra-linguistic and Para-linguistic Factors Influencing Intelligibility</td>
<td>67</td>
</tr>
<tr>
<td>4.4</td>
<td>Linguistic Determinants of Intelligibility</td>
<td>70</td>
</tr>
<tr>
<td>4.5</td>
<td>Relationship between Intelligibility and Language Trees</td>
<td>83</td>
</tr>
<tr>
<td>4.6</td>
<td>Conclusions, Discussion, and Desiderata for Future Research</td>
<td>87</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>96</td>
</tr>
<tr>
<td>5.2</td>
<td>Dialectology</td>
<td>96</td>
</tr>
<tr>
<td>5.3</td>
<td>Dialectometry</td>
<td>98</td>
</tr>
<tr>
<td>5.4</td>
<td>Edit Distance on Phonetic Transcriptions</td>
<td>100</td>
</tr>
<tr>
<td>5.5</td>
<td>Geography of Distributions</td>
<td>104</td>
</tr>
<tr>
<td>5.6</td>
<td>Validation</td>
<td>108</td>
</tr>
<tr>
<td>5.7</td>
<td>Emerging Opportunities and Issues</td>
<td>111</td>
</tr>
<tr>
<td>5.8</td>
<td>Conclusions</td>
<td>112</td>
</tr>
<tr>
<td>6.1</td>
<td>Representing Language Variability in Corpora</td>
<td>122</td>
</tr>
<tr>
<td>6.2</td>
<td>Types of Micro-Variation and the Corresponding Data Collection Procedures</td>
<td>124</td>
</tr>
<tr>
<td>6.3</td>
<td>Privacy and Linguistic Micro-Variation</td>
<td>133</td>
</tr>
<tr>
<td>6.4</td>
<td>Conclusion</td>
<td>134</td>
</tr>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>138</td>
</tr>
<tr>
<td>7.2</td>
<td>Model Transfer Methods</td>
<td>145</td>
</tr>
<tr>
<td>7.3</td>
<td>Normalization and Other Data Transfer Methods</td>
<td>149</td>
</tr>
<tr>
<td>7.4</td>
<td>Tagger Adaptation and Multilingual Models</td>
<td>156</td>
</tr>
<tr>
<td>7.5</td>
<td>Conclusions</td>
<td>158</td>
</tr>
</tbody>
</table>
# Table of Contents

## Part I Similar Languages, Varieties, and Dialects

8 Sharing Dependency Parsers between Similar Languages  
**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 II Dialect and Similar Language Identification

9 Dialect and Similar Language Identification  
**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

## Part III Applications and Language Specific Issues

10 Dialect Variation on Social Media  
**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  
**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  
**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 Standarized Evaluations and Recent Performance 269  
12.6 Challenges and Future Outlook 272
### 13 Arabic Dialect Processing

**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

**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