

# Contents

<b>Foreword by Willard McCarty</b> .....	v
References .....	xii
<b>Language Technology for Cultural Heritage, Social Sciences and Humanities: Chances and Challenges</b> .....	xxi
Caroline Sporleder, Antal van den Bosch and Kalliopi Zervanou	
1    From Quill and Paper to Digital Knowledge Access and Discovery .....	xxi
2    Mutual Benefits .....	xxii
3    Challenges .....	xxv
4    This Volume .....	xxvii
References .....	xxxii
<b>Part I Pre-Processing</b>	
<b>Strategies for Reducing and Correcting OCR Errors</b> .....	3
Martin Volk, Lenz Furrer and Rico Sennrich	
1    Introduction .....	4
2    The Text+Berg Project .....	5
2.1    Language Identification .....	7
2.2    Further Annotation .....	8
2.3    Aims and Current Status .....	8
3    Scanning and OCR .....	9
3.1    Enlarging the OCR Lexicon .....	9
3.2    Post-correcting OCR Errors .....	10
4    Evaluation .....	15
4.1    Evaluation Setup .....	15
4.2    Evaluation Results .....	16
5    Related Work .....	19
6    Conclusion .....	20
References .....	21

<b>Alignment between Text Images and their Transcripts for Handwritten Documents</b> .....	23
Alejandro H. Toselli, Verónica Romero and Enrique Vidal	
1 Introduction .....	24
2 HMM-based HTR and Viterbi Alignment .....	26
2.1 HMM HTR Basics .....	26
2.2 Viterbi Alignment .....	28
2.3 Word and Line Alignments .....	29
3 Overview of the Alignment Prototype .....	29
4 Alignment Evaluation Metrics .....	30
5 Experiments .....	32
5.1 Corpus Description .....	32
5.2 Experiments and Results .....	33
6 Remarks, Conclusions and Future Work .....	35
References .....	36
<b>Part II Adapting NLP Tools to Older Language Varieties</b>	
<b>A Diachronic Computational Lexical Resource for 800 Years of Swedish</b> .....	41
Lars Borin and Markus Forsberg	
1 Introduction .....	42
2 Lexical Resources for Present-Day Swedish .....	44
2.1 SALDO .....	44
2.2 Swedish FrameNet++ .....	46
3 A Lexical Resource for 19th Century Swedish .....	47
4 A Lexical Resource for Old Swedish .....	48
4.1 Developing a Computational Morphology for Old Swedish .....	51
4.2 The Computational Treatment of Variation in Old Swedish .....	56
4.3 Linking the Old Swedish Lexical Resource to SALDO .....	58
5 Summary and Conclusions .....	58
References .....	59
<b>Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change</b> .....	63
Eiríkur Rögnvaldsson and Sigrún Helgadóttir	
1 Introduction .....	63
2 Tagging Modern Icelandic .....	64
2.1 The Tagset .....	64
2.2 Training the Tagger .....	65
3 Tagging Old Icelandic Texts .....	66
3.1 Old vs. Modern Icelandic .....	67
3.2 The Old Icelandic Corpus .....	67
3.3 Training the Tagger on the Old Icelandic Corpus .....	68

- 4 Tagged Texts in Syntactic Research ..... 70
  - 4.1 Object Shift ..... 71
  - 4.2 Passive ..... 73
- 5 Conclusion ..... 74
- References ..... 75

**Part III Linguistic Resources for CH/SSH**

**The Ancient Greek and Latin Dependency Treebanks ..... 79**

David Bamman and Gregory Crane

- 1 Introduction ..... 79
- 2 Treebanks ..... 80
- 3 Building the Ancient Greek and Latin Dependency Treebanks .... 81
- 4 Ancient Greek Dependency Treebank ..... 83
- 5 Latin Dependency Treebank ..... 84
- 6 The Influence of a Digital Library ..... 84
  - 6.1 Structure ..... 86
  - 6.2 Reading Support ..... 88
- 7 The Impact of Historical Treebanks ..... 90
  - 7.1 Lemmatized Searching ..... 91
  - 7.2 Morphosyntactic Searching ..... 91
  - 7.3 Lexicography ..... 92
  - 7.4 Discovering Textual Similarity ..... 94
- 8 Conclusion ..... 95
- References ..... 96

**A Parallel Greek-Bulgarian Corpus: A Digital Resource of the Shared Cultural Heritage ..... 99**

Voula Giouli, Kiril Simov and Petya Osenova

- 1 Introduction ..... 100
- 2 Background ..... 100
- 3 The Bilingual Greek–Bulgarian Literary and Folklore Corpus: Selection and Description ..... 101
  - 3.1 Corpus Specifications ..... 101
  - 3.2 Collection Description ..... 102
  - 3.3 Metadata Descriptions ..... 103
- 4 Text Annotation and Processing ..... 104
  - 4.1 The Greek Pipeline ..... 105
  - 4.2 NLP Suite for Bulgarian ..... 106
  - 4.3 Sentence Alignment ..... 108
- 5 Tools Customization and Metadata Harmonization ..... 108
- 6 Bilingual Glossaries ..... 109
- 7 Content Management ..... 110
- 8 Conclusions ..... 111
- References ..... 111

**Part IV Personalisation**

**Authoring Semantic and Linguistic Knowledge for the Dynamic Generation of Personalized Descriptions . . . . . 115**

Stasinou Konstantopoulos, Vangelis Karkaletsis, Dimitrios Vogiatzis and Dimitris Bilidas

- 1 Introduction . . . . . 115
- 2 Authoring Domain Ontologies . . . . . 117
- 3 Description Adaptation . . . . . 119
  - 3.1 Personalization and Personality . . . . . 119
  - 3.2 Representation and Interoperability . . . . . 121
- 4 Adaptive Natural Language Generation . . . . . 122
  - 4.1 Document Planning . . . . . 122
  - 4.2 Micro-Planning . . . . . 123
  - 4.3 Surface Realization . . . . . 125
- 5 Intelligent Authoring Support . . . . . 126
  - 5.1 Profile Completion . . . . . 126
  - 5.2 Interaction Log Mining . . . . . 128
- 6 Related Work . . . . . 129
- 7 Conclusion . . . . . 129
- References . . . . . 131

**Part V Structural and Narrative Analysis**

**Automatic Pragmatic Text Segmentation of Historical Letters . . . . . 135**

Iris Hendrickx, Michel Génèreux and Rita Marquilhas

- 1 Introduction . . . . . 135
- 2 Corpus of Historical Letters . . . . . 137
  - 2.1 Annotated Data Set . . . . . 139
- 3 Experimental Setup . . . . . 141
- 4 Text Segmentation . . . . . 143
  - 4.1 Classifying Each Word . . . . . 145
  - 4.2 Segment Production (Smoothing) . . . . . 146
- 5 Semantic Tagging . . . . . 148
- 6 Conclusions . . . . . 150
- References . . . . . 152

**Proppian Content Descriptors in an Integrated Annotation Schema for Fairy Tales . . . . . 155**

Thierry Declerck, Antonia Scheidel and Piroska Lendvai

- 1 Introduction . . . . . 156
- 2 Summary of Propp’s Analysis . . . . . 156
- 3 Preprocessing Propp . . . . . 159
  - 3.1 Relaxing the “Fairy Tale Grammar” . . . . . 159
  - 3.2 Functions and Moves . . . . . 160

- 4 Functions and Frames . . . . . 160
  - 4.1 Proppian “Frames” and FrameNet . . . . . 160
  - 4.2 APftML Frame Elements . . . . . 161
  - 4.3 Functional Annotation . . . . . 163
- 5 Fairy Tale Characters . . . . . 165
  - 5.1 Characters vs. Dramatis Personae . . . . . 166
- 6 Temporal and Spatial Structure . . . . . 167
- 7 Dialogue and Narration . . . . . 168
- 8 Conclusion . . . . . 169
- References . . . . . 169

**Adapting NLP Tools and Frame-Semantic Resources for the Semantic Analysis of Ritual Descriptions . . . . . 171**

Nils Reiter, Oliver Hellwig, Anette Frank, Irina Gossmann, Borayin Maitreya Larios, Julio Rodrigues and Britta Zeller

- 1 Introduction . . . . . 171
- 2 Computational Linguistics for Ritual Structure Research . . . . . 173
  - 2.1 Project Research Plan . . . . . 173
  - 2.2 Related Work . . . . . 174
- 3 Ritual Descriptions . . . . . 174
  - 3.1 Textual Sources . . . . . 175
  - 3.2 Text Characteristics . . . . . 175
- 4 Automatic Linguistic Processing . . . . . 177
  - 4.1 Tokenizing . . . . . 177
  - 4.2 Part of Speech Tagging and Chunking . . . . . 177
  - 4.3 Anaphora and Coreference Resolution . . . . . 180
- 5 Semantic Annotation of Ritual Descriptions . . . . . 184
  - 5.1 Adaptation of Existing Resources . . . . . 185
- 6 Detecting Ritual Structure . . . . . 188
- 7 Future Work and Conclusions . . . . . 190
  - 7.1 Future Work . . . . . 190
  - 7.2 Conclusions . . . . . 190
- References . . . . . 191

**Part VI Data Management, Visualisation and Retrieval**

**Information Retrieval and Visualization for the Historical Domain . . . . . 197**

Yevgeni Berzak, Michal Richter, Carsten Ehler and Todd Shore

- 1 Introduction . . . . . 197
- 2 Background . . . . . 198
- 3 Information Extraction from a Historical Collection . . . . . 199
  - 3.1 Dataset . . . . . 199
  - 3.2 Extraction of Named Entities . . . . . 200
  - 3.3 Aliasing . . . . . 200

- 4 Visualization of Document Similarities ..... 202
  - 4.1 Similarity measurement ..... 202
  - 4.2 Visualization of similarities ..... 203
- 5 Graphical User Interface ..... 204
- 6 The Benefit for Historical Research ..... 207
- 7 Conclusion and Outlook ..... 209
  - 7.1 Topic Models ..... 209
  - 7.2 Clustering and Layouting ..... 210
  - 7.3 Evaluation ..... 210
  - 7.4 Adaptation to Other Domains ..... 211
- References ..... 211

**Integrating Wiki Systems, Natural Language Processing, and Semantic Technologies for Cultural Heritage Data Management ..... 213**

René Witte, Thomas Kappler, Ralf Krestel, and Peter C. Lockemann

- 1 Introduction ..... 213
- 2 User Groups and Requirements ..... 214
  - 2.1 User Groups ..... 214
  - 2.2 Detected Requirements ..... 215
- 3 Related Work ..... 216
- 4 Semantic Heritage Data Management ..... 217
  - 4.1 Architectural Overview ..... 217
  - 4.2 Source Material ..... 219
  - 4.3 Digitization and Error Correction ..... 219
  - 4.4 Format Transformation and Wiki Upload ..... 220
  - 4.5 Integrating Natural Language Processing ..... 223
  - 4.6 Semantic Extensions ..... 225
- 5 Summary and Conclusions ..... 229
- References ..... 229



<http://www.springer.com/978-3-642-20226-1>

Language Technology for Cultural Heritage  
Selected Papers from the LaTeCH Workshop Series  
Sporleder, C.; van den Bosch, A.; Zervanou, K. (Eds.)  
2011, XXXII, 232 p., Hardcover  
ISBN: 978-3-642-20226-1