

CONTENTS

1	Introducing Language Technology and Humanities	1
1.1	<i>Why Language Technology for the Humanities?</i>	1
1.2	<i>Structure of the Book</i>	3
	<i>References</i>	6
2	Design of Text Resources and Tools	7
2.1	<i>Text Resources in the Humanities</i>	7
2.1.1	<i>Text Resources and Corpora</i>	9
2.1.2	<i>Data and Metadata</i>	10
2.2	<i>Corpus Design and Creation</i>	12
2.2.1	<i>Designing a Text Resource</i>	12
2.2.2	<i>Humanities Corpora</i>	14
2.3	<i>Use Case: The Diorisis Ancient Greek Corpus</i>	16
2.4	<i>Corpus and Natural Language Processing Tools</i>	20
2.4.1	<i>Text-Processing Pipeline</i>	20
2.4.2	<i>Pre-processing and Tokenization</i>	21
2.4.3	<i>Stemming, Lemmatization, and Morphological Annotation</i>	23
2.4.4	<i>Part-of-Speech Tagging</i>	25
2.4.5	<i>Chunking and Syntactic Parsing</i>	27
2.4.6	<i>Named Entities</i>	28
2.4.7	<i>Other Annotation</i>	29
2.5	<i>Conclusion</i>	31
	<i>References</i>	32

3	Frequency	35
3.1	<i>Concept of Frequency</i>	36
3.2	<i>Application: The “Characteristic Vocabulary” of the Moonstone by Wilkie Collins</i>	39
3.3	<i>Application: Terms with ‘Turbulent History’ in the Early English Books Online</i>	43
3.4	<i>Conclusion</i>	46
	<i>References</i>	46
4	Collocation	47
4.1	<i>The Concept of Collocation</i>	48
4.2	<i>Probability of a Bigram</i>	49
4.3	<i>Observed and Expected Probability of a Bigram</i>	50
4.4	<i>Strength of Association: Pointwise Mutual Information (PMI)</i>	52
4.5	<i>Strength of Association: Log Likelihood Ratio</i>	54
4.6	<i>Application: What Residents of Modern London Complained About</i>	54
4.7	<i>Conclusion</i>	58
	<i>References</i>	59
5	Word Meaning in Texts	61
5.1	<i>The Study of Word Meaning</i>	61
5.2	<i>Distributional Approaches to Word Meaning</i>	62
5.3	<i>Word Space Models</i>	64
5.3.1	<i>Words in Space</i>	64
5.3.2	<i>Word Embeddings</i>	68
5.4	<i>Use Case: Exploring Smell in Historical Health Reports</i>	69
5.4.1	<i>Visualizing Words in the Semantic Space</i>	71
5.4.2	<i>Measuring Distances in the Semantic Space</i>	72
5.5	<i>Use Case: Finding Semantic Change in a Web Archive</i>	75
5.6	<i>Conclusion</i>	78
	<i>References</i>	78
6	Mining Textual Collections	81
6.1	<i>Textual Similarity, an Old Problem</i>	82
6.2	<i>How to Construct a Feature Space</i>	83
6.2.1	<i>Feature Selection</i>	84
6.2.2	<i>Feature Scoring</i>	88

6.2.3	<i>Representation as a Geometric Space</i>	89
6.2.4	<i>The Document–Term Matrix</i>	90
6.2.5	<i>Representation as a Vector Space</i>	90
6.2.6	<i>Summary</i>	93
6.3	<i>Application: Discovery of Similarity in the Anglo-Saxon Chronicle</i>	93
6.3.1	<i>Transformation of the Anglo-Saxon Chronicle into a Document Collection</i>	94
6.3.2	<i>Feature Extraction and Feature Selection</i>	95
6.3.3	<i>Construction of the Document–Term Matrix</i>	96
6.3.4	<i>Feature Scoring</i>	97
6.3.5	<i>Rendering a Feature Space Through Projection to a Lower-Dimensional Space</i>	99
6.3.6	<i>Measuring the Cosine Similarity Between Annals</i>	102
6.3.7	<i>Clustering</i>	104
6.3.8	<i>Topic Modelling</i>	107
6.3.9	<i>Topic as a Hidden Layer</i>	110
6.3.10	<i>Hierarchical Topic Modelling</i>	111
6.3.11	<i>Summary of Topic Modelling</i>	113
6.4	<i>Conclusion</i>	113
	<i>References</i>	114
7	The Innovative Potential of Language Technology for the Humanities	117
7.1	<i>Bridging Concepts Between Humanities and Language Technology</i>	117
7.2	<i>A Critical View of Language Technology</i>	121
	Index	123