

TABLE OF CONTENTS

Foreword 9

Introduction 11

PART I. TEXTUAL MASS

Chapter 1. What is out there? 19

1.1. The Digital Libraries Federation 20

1.2. The Shadow Library 24

1.3. The Incorrupta library 29

Chapter 2. Metadata 33

2.1. Common issues 34

2.2. Normalisation of temporal metadata 36

2.3. Temporal metadata in numbers 41

2.4. Creation or publication date? 42

2.5. Still looking for time travellers (and metadata errors) 44

2.6. Normalisation of publication types 57

2.7. Title normalisation 60

Chapter 3. Texts 63

3.1. Optical Character Recognition 65

3.2. Text preprocessing 81

3.3. Text modernisation 84

PART II. (RE)SEARCHING**Chapter 4. Searching for words 103**

- 4.1. Overview 104
- 4.2. Apache Solr and its configuration 105
- 4.3. Documents 107
- 4.4. Text fields 108
- 4.5. Other fields 118

Chapter 5. From search into research 137

- 5.1. Researching strategies 139
- 5.2. Research dossiers in detail 141
- 5.3. Word sheets 144
- 5.4. Continuous research 146

PART III. MODELLING**Chapter 6. Temporal language models 151**

- 6.1. Related work 151
- 6.2. Language models 152
- 6.3. Evaluating (temporal) language models 156
- 6.4. Language model evaluation as a machine learning challenge 158
- 6.5. RetroGap challenge 164

Chapter 7. Temporal text classification 171

- 7.1. Previous work 175
- 7.2. The RetroC corpus 177
- 7.3. RetroC as a Machine Learning Challenge 180
- 7.4. Baseline solutions 180
- 7.5. Error analysis 182

Chapter 8. Word embeddings for diachrony 185

- 8.1. Basic ideas behind Word2vec models 188
- 8.2. Applications of Word2vec models 193

8.3. Word analogy tasks 195

8.4. Detecting semantic change with word embeddings 198

PART IV. APPLICATIONS

Chapter 9. Lexical ephemera 225

9.1. Motivation 225

9.2. Phanero- and pseudoephemera 230

9.3. Automatic excerption of cryptoephemera 232

9.4. Textual material and results 234

Chapter 10. Traps of culturomics 241

10.1. Handling anomalies 251

Chapter 11. Folkloristics 2.0 257

11.1. Midwife to Murderers 260

11.2. *Surprise* 271

11.3. The vanishing hitchhiker 284

List of excerpts 287

List of figures 291

List of tables 297

Index 299

Bibliography 307