Contents

Preface 1

1 Introduction 3

General Background 11

2 Discourse Relations 13

2.1 Discourse Relations . . . 13

2.2 The Penn Discourse Treebank . . . 14

2.3 Discourse Connectives . . . 15

2.4 Discourse Arguments . . . 17

2.4.1 Notation of the arguments . . . 18

2.5 Semantic Types of Discourse Relations . . . 18

2.6 Annotation Process . . . 22

2.6.1 Theoretical starting points . . . 22

2.6.2 Inspiration from the PDTB approach . . . 23

2.6.3 Annotating on top of syntactic trees . . . 23

2.6.4 Two-phase annotation . . . 24

2.7 Other Annotated Phenomena . . . 26

2.7.1 List structures . . . 26

2.7.2 Discourse special: headings, captions, metatexts . . . 27

2.7.3 Genre annotation . . . 27

2.8 Summary . . . 27

3 Coreference 29

3.1 Basic Terms . . . 29

3.2 Related Work . . . 32

3.3 Grammatical and Textual Coreference 33

3.3.1 Grammatical coreference 34

3.3.2 Textual coreference 36

3.4 Coreference of Nominal Groups with Different Referential Potential 39

3.5 Coreference Annotation in the PDT 43

3.5.1 Scope of annotated expressions 44

3.5.2 Embedded nominal groups 44

3.5.3 Syntactic zeros 45

3.5.4 Non-referring expressions 48

3.5.5 Coordinative constructions 48

3.5.6 Coreference with specific and generic nominal groups 49

3.5.7 Discourse deixis 50

3.5.8 Prepositional phrases 50

3.6 Summary 50

4 Bridging Relations **51**

4.1 Typology of Bridging Relations 51

4.2 Annotation of Bridging Relations in Corpora 54

4.3 Annotation of Bridging Relations in the PDT 54

4.3.1 Meronymical relation between a part and a whole 55

4.3.2 The relation between a set and its subsets/elements 56

4.3.3 The relation between an entity and its singular function 58

4.3.4 The relation between coherence-relevant discourse opposites 58

4.3.5 Non-coreferential explicit anaphoric relation 59

4.3.6 Further underspecified bridging relations 60

4.4 Discussion and Summary 61

5 Topic–Focus Articulation **63**

5.1 What Is Topic–Focus Articulation 63

5.2 The Importance of Topic–Focus Articulation 64

5.3 The Theoretical Basis 64

5.4 Basic Terms of Topic–Focus Articulation 67

5.4.1 Context and contextual boundness 67

5.4.2 Communicative dynamism 69

5.4.3 Topic and focus 71

5.5 Detection of Topic and Focus 73
5.5.1 Question test 74
5.5.2 Test with negation 75
5.6 Representation of TFA in the Prague Dependency Treebank 76
5.6.1 Annotation of contextual boundness in the PDT 76
5.6.2 Annotation of communicative dynamism in the PDT 78
5.7 Summary 80

Data **81**

6 Prague Dependency Treebank **83**
6.1 Layers of Annotation 84
6.2 Discourse Coherence Phenomena 86

7 Inter-Annotator Agreement **89**
7.1 Within a Single Sentence 90
7.2 Crossing the Sentence Boundary 92
7.3 At the Document Level 94
7.4 Summary 94

8 Searching in the PDT **97**
8.1 Basics of the PML-TQ Language 98
8.1.1 Node selection 98
8.1.2 Relations between nodes 99
8.1.3 Negative query 101
8.1.4 Crossing the layers of annotation 102
8.2 Discourse Coherence Phenomena and the PML-TQ 103
8.2.1 Non-dependency relations 103
8.2.2 Topic–focus articulation and anaphora 105
8.2.3 Output filters 107
8.2.4 Output filters in discourse 108
8.3 Hands on the Data 111
8.3.1 Data to download 111
8.3.2 Data for searching 114

Case Studies 115

9 Relation of Discourse Analysis to Syntax 117

9.1 Features of Syntactic Analysis Used for Discourse-Level Analysis 118

9.1.1 Syntactico-semantic labels for relations between clauses 121

9.1.2 Scope of discourse arguments 132

9.1.3 Connectives 136

9.2 Discourse Structure from the Syntactic Point of View 139

9.2.1 Discourse relations realized intra-sententially and inter-sententially . 140

9.2.2 Discourse relations in subordinate versus coordinate structures . . . 143

9.3 Summary 146

10 Morphosyntactic Characteristics of Czech Connectives 149

10.1 General Characteristics 149

10.1.1 Part-of-speech classification 150

10.1.2 Form and inflection 151

10.1.3 Origin 153

10.1.4 Placement in the sentence and in the argument 155

10.1.5 Subordinate, coordinate and inter-sentential connectives 156

10.2 Characteristics of Most Frequent Connectives 157

10.2.1 Frequency 157

10.2.2 Part-of-speech characteristics 159

10.2.3 Intra- and inter-sentential use of connectives 160

10.2.4 Degree of connectivity 161

10.3 Summary 163

11 Multiword Discourse Phrases 165

11.1 A Scale of Explicitness and Implicitness of Discourse Relations 165

11.2 Terminology 166

11.3 Current Annotation of Secondary Connectives in the PDT 167

11.4 Lexico-Syntactic Characteristics of English AltLexes in the PDTB 167

11.5 Syntactic Characteristics of Czech Secondary Connectives 168

11.5.1 Secondary connectives realized by verbal phrases 171

11.5.2 Prepositional phrases 171

11.5.3 Secondary connectives realized by (semi-)clauses 172

11.6 Lexical Characteristics of Secondary Connectives in Czech 173

11.7 Semantic Characteristics of Secondary Connectives in Czech 174

11.8 Summary 176

12 Exploration of Weak Coherence and Coherence Disruptions 179

12.1 Terminology, Data and Workflow 180

12.1.1 Unsignaled relations in the RST Discourse Treebank 180

12.1.2 Treatment of no relation in the PDTB 181

12.1.3 Treatment of no relation in the PDT 182

12.2 Results and Discussion 185

12.2.1 New types of relations 187

12.2.2 Reader's expectation as coherence factor 190

12.3 Summary 195

13 Contextually Bound Expressions without a Coreference Link 197

13.1 Data 199

13.2 Reasons of Contextual Boundness without Anaphoric Links 200

13.2.1 Deduction of contextual boundness from previous context 202

13.2.2 Extralinguistic reasons for contextual boundness 204

13.2.3 "Scene setting" circumstances 206

13.2.4 Contextually bound expressions representing measures 207

13.2.5 Technical reasons 209

13.3 Discussion 211

13.4 Summary 213

14 Tracing Salience 215

14.1 Motivation 215

14.2 Related Work 217

14.3 Related Linguistic Phenomena annotated in the PDT 219

14.4 The Salience Algorithm 223

14.5 Learning Salience 229

15 Summary 233

List of Abbreviations 239

Bibliography 241

Sources 257

Subject Index 259

Name Index 263