Computer-Aided

Categorisation and Quantification of Connectives

in English and Arabic

(Based on Newspaper Text Corpora)

Volume (1)

Adnan Jabbar Radhi Al-Jubouri

Doctor of Philosophy

The University of Aston in Birmingham

July 1987

The University of Aston in Birmingham

Computer-Aided Categorisation and Quantification of Connectives in English and Arabic (Based on Newspaper Text Corpora)

Adnan Jabbar Radhi Al-Jubouri,
Ph.D., 1987

Summary

This study presents a detailed contrastive description of the textual functioning of connectives in English and Arabic. Particular emphasis is placed on the organisational force of connectives and their role in sustaining cohesion. The description is intended as a contribution to a better understanding of the variations in the dominant tendencies for text organisation in each language. The findings are expected to be utilised for pedagogical purposes, particularly in improving EFL teaching of writing at the undergraduate level.

The study is based on an empirical investigation of the phenomenon of connectivity and, for optimal efficiency, employs computer-aided procedures, particularly those adopted in corpus linguistics, for investigatory purposes. One important methodological requirement is the establishment of two comparable and statistically adequate corpora and the design of software and the use of existing packages to achieve the basic analysis. Each corpus comprises ca. 250,000 words of newspaper material sampled in accordance with a specific set of criteria and assembled in machine readable form prior to the computer-assisted analysis. A suite of programmes have been written in SPITBOL to accomplish a variety of analytical tasks, and in particular to perform a battery of measurements intended to quantify the textual functioning of connectives in each corpus. Concordances and some word lists are produced by using OCP.

The results of this research confirm the existence of fundamental differences in text organisation in Arabic in comparison to English. This manifests itself in the way textual operations of grouping and sequencing are performed and in the intensity of the textual role of connectives in imposing linearity and continuity and in maintaining overall stability. Furthermore, computation of connective functionality and range of operationality has identified fundamental differences in the way favourable choices for text organisation are made and implemented.

Key Terms

connectivity
cohesion
text organisation
computerised text analysis
English and Arabic languages

Acknowledgement

I would like first to thank Professor F. E. Knowles, for his able guidance and continuous encouragement and support during the various stages of this project, and for his valuable assistance in supervising the design of the various SPITBOL programmes that have been used in the experimental stage of the work.

I would also like to thank Mr Eric C. Richards, system manager, Aston University Computer Centre, for his sound advice and professional suggestions for dealing with most of the computing problems, particularly those related to hardware.

I am grateful to the Department of Scholarships, Iraqi Ministry of Higher Education and Scientific Research, for granting me a scholarship to conduct this project.

Finally, I am greatly indebted to my family for their patience, love and support.

CONTENTS

PAGE

Volume (1)

Abstract 2

Acknowledgements 3

Contents 4

List of Tables 29

List of Figures 35

Chapter 1 Introduction 42

1.0 Perspective 42

1.1 Motivations for the Investigation 43

1.1.1 Preliminaries 43

1.1.2 Some Problems in EFL Written Text Composition 44

1.1.3 Pedagogical Consequences of Incompetent EFL Writing 46

1.1.4 A Brief Assessment of Current Writing Manuals 48

1.1.4.1 EFL Writing Materials 48

1.1.4.2 General Writing Materials 52

1.1.4.3 Conclusion 56

1.1.5 Improving Instruction in EFL Written Text Composition 59

1.2 Text, Cohesion and Written Composition 63

1.2.1 Text and Composition 63

1.2.2 Cohesion and the Text 65

1.2.3 Connectives and Text Composition 67

1.3 Nature of the Investigation 68

1.4 The Investigatory Apparatus 71

1.4.1 Requirements of Efficiency 71

PAGE

1.4.2 Main Components 74

1.4.3 Operational Mode 75

1.5 The Calculus of Connectives: A Preliminary Note 78

1.5.1 Fields for the Quantitative Study of Language 78

1.5.2 Features of the Calculus 79

1.6 Aims of the Study 81

1.6.1 Ultimate Aim 81

1.6.2 Primary Aims 83

1.6.3 Secondary Aims 86

1.7 Scope and Limits of the Study 87

1.8 Plan of the Thesis 89

Footnotes to Chapter 1 92

Chapter 2 The Investigatory Apparatus: Validation of Approach 93

2.0 Perspective 93

2.1 Text Linguistics: Scope and Tasks 94

2.1.1 Preliminaries 94

2.1.2 Some Arguments for a Science of Text 96

2.1.3 Categories of Text Models 99

2.1.4 Concluding Remarks 100

2.2 Contrastive Analysis 101

2.2.1 Preliminaries 101

2.2.2 CA in Linguistics 102

2.2.2.1 Comparing Languages 102

2.2.2.2 Contrastive Linguistics 103

2.2.3 Types of CA Studies 104

PAGE

2.2.4 CA and FL Pedagogy 105

2.2.5 The Predictive Power of CA 107

2.2.6 Some Proposals for CA 111

2.3 Textual Contrastive Analysis 113

2.3.1 Preliminaries 113

2.3.2 Some Early Attempts at TCA 114

2.3.3 Gleason's Contrastive Discourse Structure 116

2.3.4 Kaplan's Contrastive Rhetoric 120

2.3.5 Hartmann's Model of Contrastive Textology 125

2.4 TCA in the Project 129

2.4.1 Preliminaries 129

2.4.2 Nature of Comparison 130

2.4.3 Dimensions of Comparison 131

2.4.4 Procedures of Comparison 132

2.5 Conclusion 140

Footnotes to Chapter 2 144

Chapter 3 Text, Textuality and Cohesion 145

3.0 Perspective 145

3.1 On the Entity of Text 146

3.1.1 Text as a Linguistic Unit 146

3.1.2 Attempts at Defining Text 147

3.1.3 Toward a Definition of Text 150

3.1.4 Angles for Viewing Text 154

3.1.5 Definition of Text in this Project 156

3.2 Textuality 158

3.2.1 Text Actualisation 158

PAGE

3.2.2 Text as a Manifestation of a Cybernetic System 160

3.2.3 Constitutive Principles of Textuality 162

3.2.4 Regulative Principles of Textuality 164

3.3 Cohesion 165

3.3.1 Preliminaries 165

3.3.2 Models of Textual Cohesion 167

3.3.2.1 Enkvist's Stylistic Model 167

3.3.2.2 The Functional-Systemic Model (Halliday; Hasan) 173

3.3.2.2.1 Halliday's Early Views 176

3.3.2.2.2 Hasan's Treatment of Cohesion 177

3.3.2.2.3 Halliday and Hasan's Categorisation of Cohesion 179

3.3.2.3 The Stratificational Model (Gutwinski) 186

3.3.2.4 The Procedural/Relational Model (Beaugrande and Dressler) 192

3.4 Cohesion: a Synthesis of Views 203

3.4.1 Preliminaries 203

3.4.2 Characterising Features 203

3.5 Cohesion: Types of Cohesive Relations 206

3.5.1. Reference 207

3.5.2 Substitution 209

3.5.3 Ellipsis 210

3.5.4 Lexical Cohesion 213

3.6 Cohesion as a Process of Textual Linearisation 215

3.6.1 Preliminaries 215

3.6.2 Type of Linear Orientations 216

PAGE

3.6.3 Linearity and Text Actualisation 217

3.6.4 Regulative Principles of Linearisation 218

3.6.4.1 Grouping 219

3.6.4.2 Sequencing 220

3.6.4.3 Salience 224

3.7 Conclusion 227

Footnotes to Chapter 3 231

Chapter 4 Connectives 234

4.0 Perspective 234

4.1 The Connective in Logic 235

4.1.1 Introduction 235

4.1.2 Propositions, Statements and Sentences 236

4.1.3 Truth Values 238

4.1.4 Connectives and Truth Relations 239

4.1.5 Types of Connection 240

4.1.5.1 Conjunction 241

4.1.5.2 Disjunction 241

4.1.5.3 Implication 242

4.1.5.4 Equivalence 244

4.1.5.5 Negation 244

4.1.6 Logical Connectives vs. Natural Language Connectives 245

4.1.7 Concluding Remarks 252

4.2 The Connective in Linguistic Studies 254

4.2.1 The Connective in Traditional Grammar 254

4.2.1.1 English Traditional Grammar 254

PAGE

4.2.1.2 Arabic Traditional Grammar 257

4.2.1.3 Concluding Remarks 260

4.2.2 The Connective in Some Theories of Modern Linguistics 261

4.2.2.1 Bloomfield's Distributional Model 261

4.2.2.2 Transformational Generative Theory 263

4.2.2.3 Concluding Remarks 265

4.3 Connectives in Some Models of Textual Cohesion 266

4.3.1 Preliminaries 266

4.3.2 The Connective in Enkvist's Stylistic Model 267

4.3.3 The Connective in the Functional-Systemic Model 268

4.3.4 The Connective in Gutwinski's Stratificational Model 272

4.3.5 The Connective in the Procedural/Relational Model 274

4.4 Some Other Textual Descriptions 277

4.4.1 Van Dijk's Semantic Description 277

4.4.2 Quirk et. al.'s Description 281

4.5 The Concept of the Connective 291

4.5.1 Introduction 291

4.5.2 Nature of the Connective 292

4.5.2.1 The Connected Objects 293

4.5.2.2 Some Characteristic Features of the Connective 295

4.6 Conclusion 309

Footnotes to Chapter 4 313

Chapter 5 <u>Automation and Processing of the Corpora</u> 314

5.0 Perspective 314

PAGE

5.1 Intuition vs. Corpus Linguistic Analysis 315

5.2 Pilot Study 318

5.2.1 Aims and Set-up 318

5.2.2 Results 320

5.3 Criteria for Assembling the Corpora 321

5.3.1 Introduction 322

5.3.2 Type of Text 322

5.3.2.1 Type of Arabic vs. English 323

5.3.2.2 Source of Selection 326

5.3.2.3 Texts Excluded 328

5.3.3 Synchronicity 328

5.3.4 Size 328

5.3.5 Organisation 329

5.3.6 Automatic Accessibility 330

5.4 Sampling 331

5.5 Hardware Facilities 336

5.6 Data Input 338

5.6.1 Preliminaries 338

5.6.2 Automatic Input 338

5.6.2.1 Use of OCR 338

5.6.2.2 Limitation of KDEM 340

5.6.3 Manual Input 342

5.7 Editing and Proofreading 343

5.8 Text Processing: Word List and Concordance Generation 345

5.8.1 Uses of Word Lists and Concordances 345

5.8.2 Some Decisions 346

5.9 Software Tools for Linguistic Analysis 350

PAGE

5.9.1 Programming Languages 350

5.9.2 Packages 353

5.10 Using OCP: Problems, Procedures and Evaluation 356

5.10.1 Preliminaries 356

5.10.2 Accessing OCP 356

5.10.3 OCP Output 357

5.10.4 OCP Performance: Some Comments 357

5.11 Tagging Procedures 361

5.11.1 Preliminaries 361

5.11.2 Taxonomy of Tagging Systems 361

5.11.3 The Tagging System in this Project 365

5.11.3.1 Features 365

5.11.3.2 Tag Assignment 367

5.12 Automation and Processing of Connectives 377

5.12.1 Preliminaries 378

5.12.2 Identification 378

5.12.2.1 Assembling the Dictionary 378

5.12.2.2 Editing the Dictionary 380

5.12.2.3 Dictionary Lookup 381

5.12.2.4 Problems of Multi-word Connectives 382

5.12.2.5 Disambiguation 383

5.12.3 Patterning 384

5.12.4 Categorisation 385

5.12.4.1 Preliminaries 385

5.12.4.2 Levels of Functionality 385

5.12.4.3 Range of Operationality 386

PAGE

5.12.4.4 Connective Tags: An Example 387

5.12.4.5 Some Reserved Tags 388

5.12.4.6 Editing and Verification of Tags 389

5.12.4.7 Word Lists and Concordances of Connectives 390

5.12.5 Quantification 390

5.13 Conclusion 392

Footnotes to Chapter 5 395

Volume (2)

Preface 2

Contents 3

List of Tables 13

List of Figures 18

Chapter 6 Categorisation and Description of Connectives 21

6.0 Perspective 21

6.1 Problem of Nomenclature 24

6.2 Structural Types of Connective Expressions 27

6.3 Categorisation of Connectives 32

6.3.1 Features 32

6.3.2 Composition of the Scheme 35

6.4 Additive 36

6.4.1 General Comments of Textual Role 36

6.4.2 Additive Categories 37

6.4.3 Appending 38

6.4.3.1 Repertoire 38

6.4.3.2 Textual Functioning 39

6.4.3.3 Some Textual Patterns 41

PAGE

6.4.4 Enumeration 44

6.4.4.1 Repertoire 44

6.4.4.2 Textual Functioning 45

6.4.4.3 Some Textual Patterns 48

6.4.5 Amplification 52

6.4.5.1 Repertoire 52

6.4.5.2 Textual Functioning 53

6.4.5.3 Some Textual Patterns 54

6.4.6 Comment 55

6.4.6.1 Repertoire , 55

6.4.6.2 Textual Functioning 56

6.4.7 Continuity 60

6.4.7.1 Repertoire 60

6.4.7.2 Textual Functioning 60

6.4.7.3 Some Textual Patterns 62

6.4.8 Coupling 63

6.4.8.1 Repertoire 63

6.4.8.2 Textual Functioning 63

6.5 Comparison 65

6.5.1 General Comments on Textual Functioning 65

6.5.2 Similarity 67

6.5.2.1 Repertoire 67

6.5.2.2 Textual Functioning 67

6.5.3 Degree 70

6.5.3.1 Repertoire 70

6.5.3.2 Textual Functioning 70

PAGE

6.6 Alternative 73

6.6.1 Repertoire 73

6.6.2 Textual Functioning 75

6.6.3 Some Textual Patterns 75

6.7 Reformulation 80

6.7.1 General Comment on Textual Functioning 80

6.7.2 Restatement 81

6.7.2.1 Repertoire 81

6.7.2.2 Textual Functioning 81

6.7.2.3 Some Textual Patterns 82

6.7.3 Exemplification 84

6.7.3.1 Repertoire 84

6.7.3.2 Textual Functioning 84

6.7.3.3 Some Textual Patterns 84

6.7.4 Summary 86

6.7.4.1 Repertoire 86

6.7.4.2 Textual Functioning 86

6.8 Orientative 87

6.8.1 General Comments on Textual Functioning 87

6.8.2 Adjustment 89

6.8.2.1 Repertoire 89

6.8.2.2 Textual Functioning 92

6.8.3 Confirmation 101

6.8.3.1 Repertoire 101

6.8.3.1 Textual Functioning 102

6.9 Temporal 105

6.9.1 General Comments on Textual Functioning 105

PAGE

6.9.2 Sequence 106

6.9.2.1 Repertoire 106

6.9.2.2 Textual Functioning 107

6.9.2.3 Some Textual Patterns 110

6.9.3 Simultaneity 110

6.9.3.1 Repertoire 110

6.9.3.2 Textual Functioning 111

6.9.3.3 Some Textual Patterns 111

6.9.4 Temporal Span 112

6.9.4.1 Repertoire 112

6.9.4.2 Textual Functioning 113

6.9.4.3 Some Textual Patterns 114

6.9.5 Temporal Positioning 115

6.9.5.1 Repertoire 115

6.9.5.2 Textual Functioning 116

6.9.5.3 Some Textual Patterns 117

6.9.6 Temporal Circumstance 120

6.9.6.1 Repertoire 120

6.9.6.2 Textual Functioning 120

6.9.7 Time Frequency 123

6.9.7.1 Repertoire 123

6.9.7.2 Textual Functioning 123

6.10 Spatial 124

6.10.1 Repertoire 124

6.10.2 Textual Functioning 124

6.11 Causal 126

PAGE

6.11.1 General Comments on Textual Functioning 126

6.11.2 Cause/Reason 129

6.11.2.1 Repertoire 129

6.11.2.2 Textual Functioning 130

6.11.3 Result/Inference 131

6.11.3.1 Repertoire 131

6.11.3.2 Textual Functioning 132

6.11.4 Magnitude/Degree 134

6.11.4.1 Repertoire 134

6.11.4.2 Textual Functioning 135

6.11.5 Purpose 136

6.11.5.1 Repertoire 136

6.11.5.2 Textual Functioning 137

6.11.6 Conditional 137

6.11.6.1 Repertoire 137

6.11.6.2 Textual Functioning 138

6.12 Adversative 143

6.12.1 General Comments on Textual Functioning 143

6.12.2 Antitheticity 143

6.12.2.1 Repertoire 143

6.12.2.2 Textual Functioning 144

6.12.2.3 Some Textual Patterns 145

6.12.3 Contrast 148

6.12.3.1 Repertoire 148

6.12.3.2 Textual Functioning 149

6.13 Conclusion 152

PAGE

Chapter 7 The Quantification of Connectives:
A General Calculus of Observations 154

7.0 Perspective 154

7.1 Some Quantitative Descriptions of Connectives 156

7.1.1 Conjunctions in Czech Newspaper Texts 156

7.1.2 Properties of Conjunctions across Genres: Kramsky 157

7.1.3 Properties of Conjunctions across Genres:
Smith and Frawley 158

7.1.4 Density of "Rhetorical" Connectives 160

7.1.5 Concluding Remarks 162

7.2 Measures Used in the Calculus of Connectives 163

7.2.1 Preliminaries 163

7.2.2 Frequency Distribution and Related Measurements 164

7.2.3 Measurements Based on Type-Token Mathematics 165

7.2.4 Entropy and Redundancy 171

7.2.5 Measure of Growth and Tuldava's Index 173

7.2.6 Repeat Rate 175

7.2.7 Measures of Interval 176

7.3 A Global Statistical Profile of the Corpora 180

7.3.1 Statistical Profile of the English Corpus 180

7.3.1.1 Frequency and Rank Distribution 180

7.3.1.2 Type-Token Measurements 181

7.3.1.3 Sentence Length 183

7.3.1.4 Paragraph Length 185

7.3.1.5 Measurement and Extrapolation of Growth 186

7.3.2 Statistical Profile of the Arabic Corpus 189

7.3.2.1 Frequency and Rank Distributions 189

PAGE

7.3.2.2 Type-Token Measurements 191

7.3.2.3 Sentence Length 194

7.3.2.4 Paragraph Length 196

7.3.2.5 Measurement and Extrapolation of Growth 198

7.4 Quantitative Characteristics of Connectives: A General Profile 201

7.4.1 Preliminaries 201

7.4.2 General Profile of English Connectives 202

7.4.2.1 Frequency and Rank Distributions 202

7.4.2.2 Connective-Sentence Distribution 207

7.4.2.3 Type-Token Measurements 209

7.4.2.4 Entropy and Redundancy 211

7.4.2.5 Repeat Rate Indices 212

7.4.2.6 Measures of Intervals 212

7.4.2.6.1 Gap Distribution 212

7.4.2.6.2 Levin's Measure of Interval 214

7.4.2.7 Measure of Growth 214

7.4.3 General Profile of Arabic Connectives 221

7.4.3.1 Frequency and Rank Distributions 221

7.4.3.2 Connective-sentence distribution 226

7.4.3.3 Type-Token Measurement 228

7.4.3.4 Entropy and Redundancy 230

7.4.3.5 Repeat Rate Indices 231

7.4.3.6 Measures of Intervals 231

7.4.3.6.1 Gap Distribution 231

7.4.3.6.2 Levin's Measure of Interval 232

PAGE

7.4.3.7 Measure of Growth 232

7.5 Conclusion 238

Footnotes to Chapter 7 241

Chapter 8 The Quantification of Connectives: Calculus of Functional Categories 243

8.0 Perspective 243

8.1 A Statistical Preview of Functionality 244

8.1.1 Profile of Functional Categories in English 245

8.1.2 Profile of Functional Categories in Arabic 247

8.2 Additive Connectives 250

8.2.1 Additive Connectives in English 250

8.2.1.1 Global Profile 250

8.2.1.2 Categories of Additive Connectives in English 256

8.2.2 Additive Connectives in Arabic 258

8.2.2.1 Global Profile 258

8.2.2.2 Categories of Additive Connectives in Arabic 264

8.3 Alternative Connectives 265

8.3.1 Alternative Connectives in English 265

8.3.2 Alternative Connectives in Arabic 270

8.4 Comparative Connectives 275

8.4.1 Comparative Connectives in English 275

8.4.1.1 Global Profile 275

8.4.1.2 Categories of Comparative Connectives in English 281

8.4.2 Comparative Connectives in Arabic 282

8.4.2.1 Global Profile 282

PAGE

8.4.2.2 Categories of Comparative Connectives in Arabic 288

8.5 Reformulatory Connectives 289

8.5.1 Reformulatory Connectives in English 289

8.5.1.1 Global Profile 289

8.5.1.2 Categories of Reformulatory Connectives in English 295

8.5.2 Reformulatory Connectives in Arabic 296

8.5.2.1 Global Profile 296

8.5.2.2 Categories of Reformulatory Connectives in Arabic 301

8.6 Orientative Connectives 303

8.6.1 Orientative Connectives in English 303

8.6.1.1 Global Profile 303

8.6.1.2 Categories of Orientative Connectives in English 309

8.6.2 Orientative Connectives in Arabic 310

8.6.2.1 Global Profile 310

8.6.2.2 Categories of Orientative Connectives in Arabic 316

8.7 Temporal Connectives 318

8.7.1 Temporal Connectives in English 318

8.7.1.1 Global Profile 318

8.7.1.2 Categories of Temporal Connectives in English 323

8.7.2 Temporal Connectives in Arabic 326

8.7.2.1 Global Profile 326

8.7.2.2 Categories of Temporal Connectives in Arabic 331

PAGE

8.8 Spatial Connectives 334

8.8.1 Spatial Connectives in English 334

8.8.2 Spatial Connectives in Arabic 339

8.9 Causal Connectives 344

8.9.1 Causal Connectives in English 344

8.9.1.1 Global Profile 344

8.9.1.2 Categories of Causal Connectives in English 349

8.9.2 Causal Connectives in Arabic 351

8.9.2.1 Global Profile 351

8.9.2.2 Categories of Causal Connectives in Arabic 357

8.10 Adversative Connectives 359

8.10.1 Adversative Connectives in English 359

8.10.1.1 Global Profile 359

8.10.1.2 Categories of Adversative Connectives in English 364

8.10.2 Adversative Connectives in Arabic 366

8.10.2.1 Global Profile 366

8.10.2.2 Categories of Adversative Connectives in Arabic 372

8.11 Conclusion 374

Footnotes to Chapter 8 376

Volume (3)

Preface 2

Contents 3

List of Tables 10

List of Figures 12

PAGE

Chapter 9 Connectives in English and Arabic: A Contrastive Account 16

9.0 Perspective 16

9.1 The Quantitative Characterisation of Connectives 18

9.1.1 Introduction 18

9.1.2 Variation in Sentence-Connective Distribution 19

9.1.3 Type-token Indices 22

9.1.4 Frequency Distribution 25

9.1.5 Measure of Repetitiveness 25

9.1.6 Variation in Entropy and Redundancy 30

9.1.7 Variation in Growth 32

9.1.8 Preview of Functional Categories 37

9.2 Interlingual Variations in the Functional Categories 39

9.2.1 Introduction 39

9.2.2 Additive Connectives 40

9.2.2.1 Quantitative Variations of Additive Connectives 40

9.2.2.2 Textual Variations of Additive Connectives 49

9.2.2.2.1 Appending 49

9.2.2.2.2 Enumeration 53

9.2.2.2.3 Amplification 55

9.2.2.2.4 Comment 56

9.2.2.2.5 Continuity 59

9.2.2.2.6 Coupling 62

9.2.3 Comparative Connectives 63

9.2.3.1 Quantitative Variations of Comparative Connectives 63

PAGE

9.2.3.2 Textual Variations of Comparative Connectives 70

9.2.3.2.1 Similarity 70

9.2.3.2.2 Degree 70

9.2.4 Alternative Connectives 74

9.2.4.1 Quantitative Variations of Alternative Connectives 74

9.2.4.2 Textual Variations of Alternative Connectives 79

9.2.5 Reformulatory Connectives 82

9.2.5.1 Quantitative Variations of Reformulatory Connectives 82

9.2.5.2 Textual Variations of Reformulatory Connectives 91

9.2.5.2.1 Restatement 91

9.2.5.2.2 Exemplification 92

9.2.6 Orientative Connective 94

9.2.6.1 Quantitative Variations of Orientative Connectives 94

9.2.6.2 Textual Variations of Orientative Connectives 102

9.2.6.2.1 Adjustment 102

9.2.6.2.2 Confirmation 106

9.2.7 Temporal Connectives 107

9.2.7.1 Quantitative Variations of Temporal Connectives 107

9.2.7.2 Textual Variations of Temporal Connectives 116

9.2.7.2.1 Temporal Sequence 116

9.2.7.2.2 Simultaneity 125

9.2.7.2.3 Temporal Positioning 130

PAGE

9.2.7.2.4 Span 134

9.2.7.2.5 Circumstance 136

9.2.8 Spatial Connectives 141

9.2.8.1 Quantitative Variations of Spatial Connectives 141

9.2.8.2 Textual Variations of Spatial Connectives 145

9.2.9 Causal Connectives 147

9.2.9.1 Quantitative Variations of Causal Connectives 147

9.2.9.2 Textual Variations of Causal Connectives 156

9.2.9.2.1 Cause-Reason 156

9.2.9.2.2 Result-Inference 158

9.2.9.2.3 Condition 162

9.2.9.2.4 Degree-Magnitude 164

9.2.9.2.5 Purpose 166

9.2.10 Adversative Connectives 167

9.2.10.1 Quantitative Variations of Adversative Connectives 167

9.2.10.2 Textual Variations of Adversative Connectives 175

9.2.10.2.1 Antitheticity 175

9.2.10.2.2 Contrast 178

9.3 Range of Operationality 180

9.3.1 Preliminaries 180

9.3.2 Types of Range 181

9.3.3 Range of Connectives: A General Profile 183

9.3.4 Range of Additive Connectives 187

9.3.5 Range of Alternative Connectives 189

PAGE

9.3.6 Range of Comparative Connectives 191

9.3.7 Range of Reformulatory Connectives 192

9.3.8 Range of Orientative Connectives 194

9.3.9 Range of Temporal Connectives 196

9.3.10 Range of Spatial Connectives 197

9.3.11 Range of Causal Connectives 198

9.3.12 Range of Adversative Connectives 200

9.4 Conclusion 202

Footnotes to Chapter 9 205

Chapter 10 Conclusions and Pedagogical Implications 206

10.0 Perspective 206

10.1 Text Organisation 207

10.2 Text Organisation in English and Arabic 209

10.2.1 Preliminaries 209

10.2.2 Textual Grouping in English and Arabic 210

10.2.3 Textual Sequencing in English and Arabic 211

10.3 Textual Organisation and Connectivity 219

10.3.1 Preliminaries 219

10.3.2 Types of Connectivity 219

10.3.3 Extension 220

10.3.3.1 Operational Scope 220

10.3.3.2 Types of Extension 221

10.3.3.3 Extension via Additivity 221

10.3.3.3.1 Textual Role of Additivity 221

10.3.3.3.2 Additivity in English and Arabic 222

10.3.3.4 Extension via Comparison 232

PAGE

10.3.3.4.1 Textual Role of Comparison 232

10.3.3.4.2 Comparison in English and Arabic 233

10.3.4 Variation 235

10.3.4.1 Operational Scope 235

10.3.4.2 Types of Variation 236

10.3.4.3 Variation via Adversativity 236

10.3.4.3.1 Textual Role of Adversativity 236

10.3.4.3.2 Adversativity in English and Arabic 237

10.3.4.4 Variation via Alternation 239

10.3.4.5.1 Textual Role of Alternation 239

10.3.4.5.2 Alternation in English and Arabic 240

10.3.5 Elaboration 241

10.3.5.1 Operational Scope 241

10.3.5.2 Types of Elaboration 242

10.3.5.3 Elaboration via Reformulation 242

10.3.5.3.1 Textual Role of Reformulation 242

10.3.5.3.2 Reformulation in English and Arabic 243

10.3.5.4 Elaboration via Orientation 245

10.3.5.4.1 Textual Role of Orientation 245

10.3.5.4.2 Orientation in English and Arabic 246

10.3.6 Enhancement 248

10.3.6.1 Operational Scope 248

10.3.6.2 Types of Enhancement 248

10.3.6.3 Enhancement via Temporality 249

10.3.6.3.1 Textual Role of Temporality 249

10.3.6.3.2 Temporality in English and Arabic 250

10.3.6.4 Enhancement via Spatial Connectives 254

PAGE

10.3.6.4.1 Textual Role of Spatial Connectives 254

10.3.6.4.2 Spatial Connectives in English and Arabic 254

10.3.6.5 Enhancement via Causality 255

10.3.6.5.1 Textual Role of Causality 255

10.3.6.5.2 Causality in English and Arabic 256

10.3.7 Concluding Remarks 259

10.4 Implications for Teaching EFL Writing Skills at the Undergraduate Level 261

10.4.1 Preliminaries 261

10.4.2 The Writing Programme: An Introduction 262

10.4.3 Some Guidelines for Designing a Writing Programme 264

10.4.3.1 Requirements 264

10.4.3.2 Some Pedagogical Suggestions 265

10.4.4 Typology of Exercises 272

10.4.5 Concluding Remarks 277

10.5 Suggestions for Further Research 278

10.5.1 Text Analysis 278

10.5.2 Computerised Corpus Analysis 281

10.5.3 Lexicography 282

10.5.4 EFL Writing Pedagogy 283

10.6 A Final Note 284

Footnotes to Chapter 10 286

Appendices

Part 1 Non-Statistical Appendices 288

Appendix 1 Reflections on the Word as a Unit of Linguistic Measurement 288

Appendix 2 Approaches to Text-Based Studies 340

Appendix 3 Toward a Definition of Text: Criteria and Requirements 357

Appendix 4 Manual Input: Problems of Transliteration 359

Appendix 5A OCP Command File for Concordance Generation of English 376

Appendix 5B OCP Command File for Concordance Generation of Arabic 377

Appendix 6 Steps for Accessing OCP 378

Appendix 7A A Sample Page from the English Corpus Containing Tagged Connectives 383

Appendix 7B A Sample Page with Deleted Connectives 384

Appendix 7C A Sample Page with Text Deleted and Connectives Retained 385

Volume (4)

Preface 2

Contents 3

Appendices - Part II Statistical Appendices 18

Appendices - Part III Appendices on Microfiche 285

References 286

LIST OF TABLES

Volume (2)

PAGE

6.1 Categorisation scheme of connectives 36

7.1 Indices based on type-token measurement in the English corpus 182

7.2 Sentence length in the English corpus 184

7.3 Paragraph length in the English corpus 185

7.4 Calculation of growth of vocabulary in the English corpus 188

7.5 Indices based on type-token measurement in the Arabic corpus 192

7.6 Sentence length in the Arabic corpus 195

7.7 Paragraph length in the Arabic corpus 197

7.8 Calculation of growth of vocabulary in the Arabic corpus 201

7.9 Connectives achieving 75% of token coverage in the English corpus 204

7.10 Rank distribution of connectives in the English corpus 206

7.11 Connective-sentence distribution in the English corpus 208

7.12 Indices based on connective type-token measurement in the English corpus 210

7.13 Calculation of global growth of connectives in the English corpus 217

7.14 Growth of connective types within connective tokens in the English corpus 220

7.15 Connectives achieving 75% of token coverage in the Arabic corpus 223

7.16 Rank distribution of connectives in the Arabic corpus 225

7.17 Connective-sentence distribution in the Arabic corpus 227

7.18 Indices based on connective type-token measurement in the Arabic corpus 229

7.19 Calculation of growth of connectives in the Arabic corpus 235

PAGE

7.20 Growth of connective types within connective tokens in the Arabic corpus 237

8.1 A summary account of the size of connective categories in terms of tokens in the English corpus 245

8.2 A summary of the distribution of types within the functional categories of connectives in the English corpus 247

8.3 A summary of the size of connective categories in terms of tokens in the Arabic corpus 248

8.4 A summary of the distribution of types within the functional categories of connectives in the Arabic corpus 248

8.5 Rank distribution of additive connectives in the English corpus 251

8.6 Global growth of additive connectives in the English corpus 254

8.7 Local growth of additive connectives in the English corpus 255

8.8 Distribution of categories of additive connectives in the English corpus 257

8.9 Rank distribution of additive connectives in Arabic corpus 258

8.10 Global growth of additive connectives in the Arabic corpus 262

8.11 Local growth of additive connectives in the Arabic corpus 263

8.12 Distribution of categories of additive connectives in the Arabic corpus 264

8.13 Rank distribution of alternative connectives in the English corpus 265

8.14 Global growth of alternative connectives in the English corpus 268

8.15 Local growth of alternative connectives in the English corpus 269

8.16 Rank distribution of alternative connectives in the Arabic corpus 270

PAGE

8.17 Global growth of alternative connectives in Arabic corpus 273

8.18 Local growth of alternative connectives in the Arabic corpus 274

8.19 Rank distribution of comparative connectives in the English corpus 276

8.20 Global growth of comparative connectives in the English corpus 278

8.21 Local growth of comparative connectives in the English corpus 280

8.22 Distribution of categories of comparative connectives in the English corpus 281

8.23 Rank distribution of comparative connectives in the Arabic corpus 283

8.24 Global growth of comparative connectives in the Arabic corpus 286

8.25 Local growth of comparative connectives in the Arabic corpus 287

8.26 Distribution of categories of comparative connectives in the Arabic corpus 288

8.27 Rank distribution of reformulatory connectives in the English corpus 290

8.28 Global growth of reformulatory connectives in the English corpus 293

8.29 Local growth of reformulatory connectives in the English corpus 294

8.30 Distribution of categories of reformulatory connectives in the English corpus 295

8.31 Rank distribution of reformulatory connectives in the Arabic corpus 297

8.32 Global growth of reformulatory connectives in the Arabic corpus 299

8.33 Local growth of reformulatory connectives in the Arabic corpus 300

8.34 Distribution of categories of reformulatory connectives in the Arabic corpus 302

PAGE

8.35 Rank distribution of orientative connectives in the English corpus 303

8.36 Global growth of orientative connectives in the English corpus 306

8.37 Local growth of orientative connectives in the English corpus 308

8.38 Distribution of orientative categories in the English corpus 309

8.39 Rank distribution of orientative connectives in the Arabic corpus 311

8.40 Global growth of orientative connectives in the Arabic corpus 314

8.41 Local growth of orientative connectives in the Arabic corpus 315

8.42 Distribution of categories of orientative connectives in the Arabic corpus 316

8.43 Rank distribution of temporal connectives in the English corpus 318

8.44 Global growth of temporal connectives in the English corpus 321

8.45 Local growth of temporal connectives in the English corpus 322

8.46 Distribution of categories of temporal connectives in the English corpus 323

8.47 Rank distribution of temporal connectives in the Arabic corpus 326

8.48 Global growth of temporal connectives in the Arabic corpus 329

8.49 Local growth of temporal connectives in the Arabic corpus 330

8.50 Distribution of categories of temporal connectives in the Arabic corpus 332

8.51 Rank distribution of spatial connectives in the English corpus 335

PAGE

8.52 Global growth of spatial connectives in the English corpus 337

8.53 Local growth of spatial connectives in the English corpus 338

8.54 Rank distribution of spatial connectives in the Arabic corpus 339

8.55 Global growth of spatial connectives in the Arabic corpus 342

8.56 Local growth of spatial connectives in the Arabic corpus 343

8.57 Rank distribution of causal connectives in the English corpus 344

8.58 Global growth of causal connectives in the English corpus 347

8.59 Local growth of causal connectives in the English corpus 348

8.60 Distribution of categories of causal connectives in the English corpus 350

8.61 Rank distribution of causal connectives in the Arabic corpus 352

8.62 Global growth of causal connectives in the Arabic corpus 355

8.63 Local growth of causal connectives in the Arabic corpus 356

8.64 Distribution of categories of causal connectives in the Arabic corpus 357

8.65 Rank distribution of adversative connectives in the English corpus 359

8.66 Global growth of adversative connectives in the English corpus 362

8.67 Local growth of adversative connectives in the English corpus 363

8.68 Distribution of categories of adversative connectives in the English corpus 364

8.69 Rank distribution of adversative connectives in the Arabic corpus 367

PAGE

8.70 Global growth of adversative connectives in the Arabic corpus 370

8.71 Local growth of adversative connectives in the Arabic corpus 371

8.72 Distribution of categories of adversative connectives in the Arabic corpus 372

Volume (3)

9.1 Levin's measures of compactness of the distribution of connectives in the English and Arabic corpora 30

9.2 Entropy and redundancy of connectives in the English and Arabic corpora 31

9.3 Frequency distribution of connectives of temporal sequencing occurring within a distance of 20 words in the Arabic corpus 124

9.4 Frequency distribution of connectives of temporal sequencing occurring within a distance of 20 words in the English corpus 125

9.5 Connective cores used for constructing time relators with referential elements (as observed in the English corpus) 132

9.6 Connective cores used for constructing time relators with referential elements (as observed in the Arabic corpus) 132

9.7 Distribution of operational range of connectives 184

9.8 Distribution of range of additive connectives 188

9.9 Distribution of range of comparative connectives 190

9.10 Distribution of range of alternative connectives 191

9.11 Distribution of range of reformulatory connectives 193

9.12 Distribution of range of orientative connectives 195

9.13 Distribution of range of temporal connectives 196

9.14 Distribution of range of spatial connectives 197

9.15 Distribution of range of causal connectives 198

9.16 Distribution of range of adversative connectives 200

LIST OF FIGURES

Volume (1)

PAGE

2.1 Classification of approaches to language comparison 103

2.2 A bi-directional procedure of CA 104

2.3 A unidirectional procedure for CA 105

2.4 Graphic representation of the movement of paragraphs in Kaplan's data (Kaplan, 1966, p.15, 1972, p.64) 123

2.5 Direction of TCA in the project 139

3.1 Categories of cohesion (Halliday 1964) 176

3.2 A summary of Halliday and Hasan's categorisation of cohesive ties 182

4.1 Logical relations in Quirk et. al. (1972, p.661) 284

Volume (2)

7.1 Distribution of types and tokens in the English corpus 182

7.2 Distribution of sentence lengths in the English corpus 184

7.3 Distribution of paragraph lengths in the English corpus 186

7.4 Growth of vocabulary in the English corpus 188

7.5 Distribution of types and tokens in the Arabic corpus 192

7.6 Comparison of the frequency distribution of types and tokens in the English and Arabic corpora 193

7.7 Distribution of sentence lengths in the Arabic corpus 195

7.8 Distribution of paragraph lengths in the Arabic corpus 197

7.9 Growth of vocabulary in the Arabic corpus 198

7.10 Comparison of growth of vocabulary in the English and Arabic corpora 199

7.11 Distribution of connectives in the English corpus 203

PAGE

7.12 Distribution of the 10 most frequent connectives in in the English corpus 205

7.13 Distribution of Connectives per sentence in the English corpus 209

7.14 Distribution of connective types and tokens in the English corpus 211

7.15 Distribution of distance between occurrences of connectives in English 213

7.16 Growth of connective tokens in the English corpus 217

7.17 Growth of connective types within tokens in the English corpus 221

7.18 Distribution of connectives in the Arabic corpus 222

7.19 Distribution of the 10 most frequent connectives in the Arabic corpus 224

7.20 Distribution of connectives per sentence in the Arabic corpus 226

7.21 Distribution of connective types and tokens in the Arabic corpus 229

7.22 Distribution of Distance between occurrences of connectives in Arabic 232

7.23 Growth of connective tokens in the Arabic corpus 235

7.24 Growth of connective types within tokens in the Arabic corpus 238

8.1 Distribution of categories of connectives in the English corpus 246

8.2 Distribution of connective types within the categories of connectives in the English corpus 246

8.3 Distribution of categories of connectives in the Arabic corpus 249

8.4 Distribution of connective types within categories of connectives in the Arabic corpus 249

8.5 Distribution of categories of additive connectives in the English corpus 257

8.6 Distribution of categories of additive connectives in the Arabic corpus 264

8.7 Distribution of categories of comparative connectives in the English corpus 282

8.8 Distribution of categories of comparative connectives in the Arabic corpus 289

8.9 Distribution of categories of reformulatory connectives in the English corpus 296

8.10 Distribution of categories of reformulatory connectives in the Arabic corpus 302

8.11 Distribution of categories of orientative connectives in the English corpus 310

8.12 Distribution of categories of orientative connectives in the Arabic corpus 317

8.13 Distribution of categories of temporal connectives in the English corpus 324

8.14 Distribution of categories of temporal connectives in the Arabic corpus 333

8.15 Distribution of categories of causal connectives in the English corpus 351

8.16 Distribution of categories of causal connectives in the Arabic corpus 358

8.17 Distribution of categories of adversative connectives in the English corpus 365

8.18 Distribution of categories of adversative connectives in the Arabic corpus 373

Volume (3)

9.1 Comparison of the distributions of connectives per sentence in the two corpora 20

9.2 Tokens and Types of Connectives in the English and Arabic corpora 20

PAGE

9.3 Comparison of the distribution of connectives in the English and Arabic corpora 26

9.4 Comparison of the distribution of connective types in the English and Arabic corpora 26

9.5 Comparison of the distribution of the 10 most frequent connectives 27

9.6 Comparison of growth of connective tokens 33

9.7 Comparison of growth of connective types 33

9.8 Comparison of growth of connective types within tokens in the two corpora 36

9.9 Comparison of the distribution of categories of connectives 36

9.10 Comparison of shares of connective categories in connective mass in the English and Arabic corpora 38

9.11 Comparison of the distribution of connective types within categories of connectives 38

9.12 Comparison of the distribution of additive connectives 41

9.13 Comparison of the distribution of additive connectives within connective tokens 41

9.14 Comparison of "global" growth of additive connectives 44

9.15 Comparison of "local" growth of additive connectives 44

9.16 Comparison of the distribution of categories of additive connectives in the English and Arabic corpora 46

9.17 Comparison of distribution of shares of additive categories 46

9.18 Comparison of the distribution of comparative connectives 64

9.19 Comparison of the distribution of comparative connectives within connective tokens 64

9.20 Comparison of the global growth of comparative connectives 67

9.21 Comparison of the local growth of comparative connectives 67

PAGE

9.22 Comparison of the distribution of categories of comparative connectives in the two corpora 69

9.23 Comparison of the distribution of shares of comparative categories 69

9.24 Comparison of the distribution of alternative connectives 76

9.25 Comparison of the distribution of alternative connectives within connective tokens 76

9.26 Comparison of the global growth of alternative connectives 78

9.27 Comparison of the local growth of alternative connectives 78

9.28 The semantic content of propositions entering in an alternative relation in (1) English (2) Arabic 81

9.29 Comparison of the distribution of reformulatory connectives 83

9.30 Comparison of the distribution of reformulatory connectives within connective tokens 83

9.31 Comparison of the global growth of reformulatory connectives 86

9.32 Comparison of local growth of reformulatory connectives 86

9.33 Comparison of the distribution of the categories of reformulatory connectives in the two corpora 89

9.34 Comparison of the distribution of shares of reformulatory connectives 89

9.35 Comparison of the distribution of orientative connectives 95

9.36 Comparison of the distribution of orientative connectives within connective tokens 95

9.37 Comparison of the global growth of orientative connectives 98

9.38 Comparison of the local growth of orientative connectives 98

9.39 Comparison of the distribution of categories of orientative connectives in the two corpora 100

9.40 Comparison of the distribution of shares of orientative categories 100

PAGE

9.41 Comparison of the distribution of temporal connectives 108

9.42 Comparison of the distribution of temporal connectives within connective tokens 108

9.43 Comparison of the global growth of temporal connectives 110

9.44 Comparison of the local growth of temporal connectives 110

9.45 Comparison of the distribution of the categories of temporal connectives in the two corpora 113

9.46 Comparison of the distribution of shares of temporal categories 113

9.47 Comparison of the functional scope of "until" vs. "’ilā ’an" and "ḥattā" 135

9.48 Comparison of the distribution of spatial connectives 142

9.49 Comparison of the distribution of spatial connectives within connective tokens 142

9.50 Comparison of the global growth of spatial connectives 144

9.51 Comparison of the local growth of spatial connectives 144

9.52 Comparison of the distribution of causal connectives 148

9.53 Comparison of the distribution of causal connectives within connective tokens 148

9.54 Comparison of global growth of causal connectives 150

9.55 Comparison of local growth of causal connectives 150

9.56 Comparison of the distribution of categories of causal connectives in the two corpora 153

9.57 Comparison of the distribution of shares of causal categories 153

9.58 Clustering of connectives introducing result/consequence on the relation continuum 159

9.59 Comparison of the distribution of adversative connectives 168

9.60 Comparison of the distribution of adversatives within connective tokens 168

PAGE

9.61 Comparison of global growth of adversative connectives 171

9.62 Comparison of local growth of adversative connectives 171

9.63 Comparison of the distribution of the categories of adversative connectives in the two corpora 174

9.64 Comparison of the distribution of shares of adversative categories 174

10.1 Text development via appending in Arabic 224

CHAPTER ONE

Introduction

1.0 Perspective

The structure of connected discourse has received growing attention during the last decade and a half, and has been examined from different angles and for different purposes. A number of models and theoretical frameworks, some significantly more adequate than others, have been proposed to describe and interpret the various patterns of connectivity. The final word has not been said, nor will, perhaps, ever be, which gives this area of linguistic investigation inexhaustible richness and prosperity in both its theoretical and metascientific perspectives.

This study presents a contrastive linguistic description of connectives in English and Arabic based on machine-readable corpora and assisted by computer techniques where applicable. The study is an exposition of the textual as well as the quantitative properties of the mechanism of connectives as a fundamental part of the machinery of textual cohesion. The findings are intended to be manipulated for pedagogical purposes, and hence a practical dimension is added to the work.

The principal purpose of this introductory chapter is to set the scene for the study. This is accomplished by deliberately separating and then examining the various issues that constitute the background of the investigation. This will help pinpoint the direction that the study, in its entirety, is to take and delimit its aims, scope and essence. Accordingly, this chapter aims to achieve the following tasks: a) a discussion of the motivations behind the initiation of this work, b) an elaboration of (a) through i) exposing the type of the problems associated with EFL written text composition, ii) discussing the shortcomings of existing writing manuals; c) a brief discussion of the concepts of cohesion and connectives and an outline of their relevance to text composition; d) an identification of the type of investigatory apparatus to be developed for the study; e) a specification of the aims of the investigation and an outline of its scope and limits; f) and finally, a brief outline of the plan of this work.

1.1 Motivations for the Investigation

1.1.1 Preliminaries

Scientific inquiry (including research in linguistics) is undertaken because a certain phenomenon, event or state of affairs is found problematic. This means that within the framework of existing store of knowledge, scientists (including linguists) cannot understand its occurrence, general nature or particular properties. Consequently, an investigation is initiated with a general standard aim of gaining knowledge within the framework of which whatever is problematic will lose its problematic character.

The contrastive study of some properties of connectives in English and Arabic is such a problematic phenomenon. While there exists a handful of studies on connectives in English, each trying to view them from a different angle, there is an absence of a study that is devoted to Arabic connectives and their textual role. Furthermore, no coherent and sophisticated attempt exists in the present state of the art for an observationally adequate comparison of the properties of connectives in English and Arabic. There should, therefore, be very little to say in a way of justifying the initiation of an investigation into this phenomenon: the motives are transparent enough.

However, in addition to these, we have other more compelling motives for undertaking the investigation. These other motives, taken together, reflect our ultimate concern and preoccupation: improving EFL instruction in advanced writing skills for Arabic-speaking learners. Written text composition is identified with its search for a textually integrated and convincingly demonstrated pattern of discourse. Connectives are devices that, among others, establish cohesion and thus reflect textual integration. The implications, therefore, that a study of connectives in English and Arabic can offer are valuable in delimiting the components that need to be incorporated in EFL instruction, particularly at the stage of pedagogical material design and production. These issues will be elaborated within the next sections.

1.1.2 Some Problems in EFL Written Text Composition

In EFL situations, the problems that Arabic speaking learners of English face when producing written texts are not merely restricted to undeveloped knowledge of grammar and lexicon. Written text composition by those who are sufficiently sophisticated in the use of grammar and choice of lexical items has been found to suffer from an inconsistent and in many cases deplorably deficient approximation to a native-like composition (cf., for instance, the investigations carried out by Kaplan 1966, Dudley-Evans and Swales 1980, Koch 1981, Williams 1982, 1984a, Al-Jubouri 1984, 1987 and Holes 1984). Such a situation is attributable to a number of factors, the most prominent of which is the manner in which written text organisation diverges in the two languages, and the comparative diversity in the type and use of the necessary conventions that signal relations between parts of discourse. This divergence is sharply brought about when an Arab writing in English attempts to bring his knowledge of the textual organisation of Arabic to bear on the written mode of English.

At present there is no adequate theory of how Arabic written text conventions are transferred to situated communicative strategies in English text composition, and perhaps there will not be one for a considerable time to come. This is due to the limited number of genuine attempts at investigating this phenomenon and the fact that these attempts often suffer from a lack of consistent orientation. Moreover, some of these studies are in the form of unpublished dissertations and theses and are, in most cases, left to slumber in the stacks of university libraries.

However, a number of recent as well as current interlingual investigations of the various aspects of textuality in English and Arabic can, taken in their entirety, point to a specific direction of research (cf., for instance, Shamaa 1977, Al-Shabab 1986, Williams (forthcoming, Al-Jabr (forthcoming)). One major aim of our project is an informative comparison of one fundamental principle of textuality, namely cohesion, as realised through the surface-expressed ties of connectives. The value of the project lies in its explicit exposition of the textual resources that English and Arabic make use of, each independently, for realising connectivity within text.

But before we start discussing the aims of the project and the methodological mechanism used to attain them, we would like to reflect briefly on some pedagogical problems that incompetent written composition can create, in order that sufficient motivation for carrying out the project can be substantiated. The problems that are discussed are essentially peculiar to a specific situation that, professionally speaking, I was closely associated with for a considerable length of time. The situation is that of undergraduate university students whose major subject is English and who are training to be secondary school teachers of English. The situation also includes Arabic-speaking teachers of English who attend in-service teaching training courses which involve a strong advanced English component. Despite this specificity of situations, the problems that will be outlined are of some general applicability.

1.1.3 Pedagogical Consequences of Incompetent EFL Writing

It has been observed through personal experience with the various facets of EFL pedagogy (particularly in the situation mentioned above) that as the learner progresses in mastering new linguistic skills, a discrepancy simultaneously starts to emerge between these skills and the learner's written performance. Difficulty in achieving a native-like approximation in writing results in seriously impeding the learner's functioning in English, particularly as the focus in learning shifts from passive skills to more advanced skills such as letter, essay, report or article writing and translation. Written texts produced by incompetent students, as a survey conducted at my university once showed, are often stigmatised, to the learners' disappointment, as "incorrect, poor, awkward, loose, repetitive, incoherent, etc.". Such a situation leads to two consequences.

First, the learners develop "writing anxiety" because they feel their written composition is judged by complicated, or at least unexplained, standards. Anxiety can, within very limited range, sustain a positive effect by heightening attention and effort and, hence, can result in improvement of productivity. But beyond this range, writing anxiety leads to negative results: it drains the learner's resources that should otherwise be devoted to the task itself (cf. Murray 1971). Anxiety leads to the learner's conviction that writing in a foreign language is difficult, challenging and, within an educational curriculum that requires testing and grading, even threatening (cf. Sarason's 1980 discussion of anxiety created by testing and intensified by incompetent writing).

The second consequence (cf. Beaugrande 1984) is the tendency of both teachers as well as learners, in their effort to upgrade and monitor the skill of writing in English, to resort to ready-tailored pedagogical manuals and textbooks specifically designed to cater for the demands of advanced writing and 'free' composition. Such textbooks are, in the current state of the art, not scarce and new books appear every year. They all form sources of reference and authority for training and use.

1.1.4 A Brief Assessment of Current Writing Manuals

A close examination of the types of these text books would indicate two categories. The first constitutes text books designed for EFL written practice. The other subsumes those books that are designed for general users including English native students. We now consider critically the efficiency of these text books for the ELT situation we are concerned with.

1.1.4.1 EFL Writing Manuals

An examination of a number of such textbooks will immediately classify them into two groups. The first group comprises textbooks that are wholly devoted to the development of the writing skill, and are therefore used in conjunction with other instructional material in the classroom. In the second group of textbooks, writing is a component in more general course work. Textbooks examined in the first group include Chaplan (1970), Swales (1974), Lawrence (1974), Imhoof and Hudson (1975), Arnold and Harmer (1978), Jordan (1980), and Johnson (1981). Within the second group, we examined Chaplan (1977), Mackin and Carver (1971) and Al-Hamash and Al-Jubouri (1981). (Those textbooks have been prescribed in courses of various lengths for advanced learners in the institutions I have been connected with).

The following general comments can be stated:

1. There is a strong tendency towards variation in classroom techniques and in the typology of exercises offered. Although this

is a positive feature and one that is essential for minimising boredom, the question remains how those various techniques and exercises within a particular course book could contribute to optimising written textual development.

2. Part of the instructional material is constructed with little thought for the type of advanced learner concerned or his immediate needs. It constitutes no more than a bank of instructional material, and the burden of selecting, grading and adapting is left to the discretion of the practising teacher. Our experience has shown that efficient and effective accomplishment of these tasks is often beyond the capability of the average language teacher (for a variety of reasons, such as lack of awareness of effective techniques, limitation of time, poor planning, enormous classes of mixed ability, resistance to custom-made experimental material in favour of the more "authoritative" textbook, and sometimes the textbook is imposed for the purpose of unifying standards at institutes and has therefore to be followed progressively).

3. Related to comment 2 is the incapability of the instructional material to cater for the differences between the cultural patterns of thought that are exhibited in the written mode of the learner's native language and those peculiar to English (cf. on this subject Kaplan 1966 and Dudley-Evans and Swales 1980). It has been pointed out (cf. for instance Yorkey 1977) that in teaching English writing to Arabic-speaking learners, the use of subordination "especially the use of adverbial clause of time, place, result, concession, cause, purpose or condition is a matter

which requires considerable instruction and practice" (p.68). In addition, Yorkey advocates the specific teaching of a tightly organised, logical presentation of ideas. Such considerations are often ignored in writing course books. The writers, in their appeal to a general audience, show conscious unawareness of these specific requirements. Individualisation is simply not part of their set of aims.

4. The recently emerging concern of textbook writers (including myself at certain stages) with the functional/notional syllabus has led to incorporating functions and notions in the development of the skill of writing (cf. particularly Arnold and Harmer 1978 and Johnson 1981, cf. also, to a lesser degree, Lawrence 1974). Although such conceptions are, in theory and, to some extent, in practice, are both justified and well-defended, doubts still remain over the effectiveness of techniques in achieving the terminal goals. For instance, it is not clear how, particularly in the case of self-study, the learning of the various ways of expressing specific functions can directly contribute to the production of a piece of writing that manifests 'proper' texture. By 'proper' we mean conforming to the principles of textuality and the parameters of design (see Chapter 3 for an explication of these principles and parameters).

5. The materials in certain cases (cf. Chaplan 1977, Jordan 1980 and Johnson 1981) hardly display a consistent scheme of gradation. This, we believe, reflects two deficiencies: an incoherent set of goals that the textbooks attempt to attain, and a

lack of awareness of the specific or even possible place of the textbook within a total EFL syllabus. Both deficiencies are derived from the textbook writer's incompetent prediction of the peculiarities of a particular ELT situation.

6. Coupled with comment 5 is the lack of integration of the units of instructional material. This is evident on two levels. The first is a horizontal level: the units do not integrate among themselves in a coherently planned progression that encourages the development of the skill of writing. The second level is vertical: the composition component in a particular unit often bears very little relevance, especially in topic content or in complexity (cf. Chaplan 1977 in particular) to the main topic area and difficulty level of the main unit. This lack of relevance is neither theoretically nor practically warranted, and hence an unfavourable discrepancy is created in the general planning, which demands the teacher's constant intervention and modification.

7. There is an absence of the deliberate teaching of the efficient and appropriate use of the mechanisms of connectivity which lend text its surface texture. It is not clear whether this is due to ignorance of the textual role of cohesion in text or whether the instruction material is based on the implicit assumption that learners have already mastered the textual signalling of surface structure even before the course work starts.(1) Whatever the assumption is, there is a demand for a careful development of the learner's rendering of tight texture. This is achieved by assisting the learner in the efficient, effective and appropriate handling of the mechanism of cohesion of English (cf. comments to

this effect by Keen 1978, Montgomery 1982, Pincas 1982 and Coupland 1984).

All the textbooks examined offer rules of usage. These are to be commented on within our examination of the second type of reference manuals and textbooks, those designed for the general user including native speakers of English.

1.1.4.2 General Writing Manuals

Such textbooks have been prolific during the last two decades. The motivation they have received comes from two main sources. The first is the advent of theories of New Rhetoric, particularly in the context of teaching writing in higher schools and colleges in the United States (cf. for instance within this area, the contributions made by Christensen 1963a, 1963b, 1965, Rodgers 1966, Young, Becker and Pike 1970). The second source is the literacy crisis in the American educational system which has been diagnosed by researchers and specialists in written text composition (cf. for instance, Black 1982, Beaugrande 1984).

Writing textbooks and reference manuals have come to be regarded as "authorities on usage" (cf. Beaugrande 1984, p.12) and have therefore a wide impact not only on English-speaking learners of writing, but also on advanced EFL learners either directly or through influencing the mode of practice of their non-native teachers. The aim of the manual of instruction is supposedly to speed up the teaching and learning of a practical craft. The teacher offers assignments and corrects what his students produce on the assumption that they will improve their next production. If the student understands the principles behind the corrections in his written text, he will be in a better position to alter and improve his productive schemata (cf. Hirsch 1977). It is assumed therefore that consulting a good manual can reinforce the teacher's explanation of the corrective principles.

The influence that the writing instruction manuals exercise is by no means always positive. Negative impact can be traced in the way correctness of usage is determined and dictated. On this point, Beaugrande (1984) argues that correct usage is "determined by decree and assertiveness" which eventually renders the learner "anxious and insecure by authoritarian attitudes" (p.13). Such a negative influence can be displayed through surveying this collection: Cooper (1964), Sanders et al. (1966), Arena (1975), Corder (1979), Strunk and White (1979), Corbett (1980), McMahen and Day (1980), McCrimmon (1980), Eisenberg (1982) and Clanchy and Ballard (1983). The survey is concerned with observing the utility of these books as reference manuals of writing instruction in the classroom. More specifically, we would like to examine what bearing the formulation and dictation of rules can have and whether any consideration is made for the operations performed by the different parts of the mechanism of cohesion. Within this collection we include two books of the previous group: Swales (1974) and Zinkin (1980).

The following observations are the outcome of this brief survey:

1. One immediately discernible shortcoming is the unwarranted

interface between general strategies of writing, such as removing ambiguities and needless wordage or enforcing intellectual and sound reasoning, and the author's personal biases concerning correctness and specificity. For instance McCrimmon's (1980) account of paragraph development, tone, style and persuasion is followed by a description of usage of such points as "dangling modifiers" (p.416), "faulty complement" (p.458), "confusion of adjective and adverb" (p.461). Similarly, Strunk and White (1979) mix their advocacy of precision, clarity and brevity with such distinctions as "compared to" vs. "compared with", "shall" vs. "will", "while" vs. "although". Eisenberg (1982) in her search for good style, includes distinctions between expressions such as "owing to the fact that" vs. "because", and advises the students to avoid such "nonwords" as "implement" and "utilization" (p.150). Similar rules are also advocated by McMahen and Day (1980 pp.388-418).

2. Implicit, though often explicit, in the rules and pieces of advice offered in these instruction manuals is the right/wrong dichotomy. This takes up an authoritarian attitude that can be detrimental in a classroom situation, particularly within EFL pedagogy. Stubborn adherence to what an instruction manual calls 'right' can create an unfavourable prescriptive instruction where language use is fragmented, reduced or distorted. We are fully aware that the imposition of such a dichotomy is motivated by such a mixture of good intent and convenience, but the authoritarian manner in which the right/wrong rule set is constructed lacks the basis of observable usage and, therefore, "many rules merely embody personal biases" (Beaugrande 1984 p.15). In reality, such an apparently

relentless dichotomy is governed and monitored by a graded continuum between two poles: acceptable and unacceptable. The validity of the grades are explainable in terms of the writers' own options. In the classroom practice, danger lurks when untrained non-native teachers as well as students misunderstand the advice as a strict rule. Often they are puzzled or even worried by seemingly conflicting rules or contradictory pieces of advice on a particular usage: the manual imposes one thing and discoursal reality shows another (cf. Harris 1979 and Rose 1981). In most situations, great effort, time, ingenuity and intelligence are invested in the struggle to follow the rules, which can, in effect, impede rather than speed up effective teaching and learning.

3. Rules and pieces of advice on usage are sometimes unworkable. This is due to an ambiguity within the rule itself or to the vagueness of the terms used by the author of the instruction manual. For instance, Zinkin's (1980) rules of "present it well" (pp.8-11), "write simply" (pp.12-21), "be coherent and consistent" (pp.22-23) are neither clear nor workable in the variety of classroom situations. Her definition of "ambiguity" is itself ambiguous: "Ambiguity is not lack of coherence, but obscurity" (p.23). It is difficult to imagine how these rules, stated as they are, can assist students to produce a coherent, unambiguous written text.

4. The manuals surveyed do not offer a direct exposition of the functioning of the mechanisms of textuality, particularly that of cohesion. Most aspects discussed are rhetorical (eg. rhetorical organisation, forms of writing, audience, etc.) or stylistic (e.g.

distance, effectiveness, etc.). Such aspects of cohesion as treated under the subtitle of sentence combining are either embedded within a grammatical exposition of the clause (cf. Arena 1975, Corder 1979) or are given a flimsy and unrealistic weight (cf. McCrimmon 1980) or they are completely overlooked (cf. Clanchy and Ballard 1983).

5. Classroom pedagogical problems associated with writing are hardly predicted or treated. For instance, there is an absence of systematic step-by-step classroom guidance and teaching techniques. Additionally, such problems as the discrepancies in a mixed-ability writing classroom or remedial work for textual idiosyncrasies and inconsistencies are left undiscussed. Lacking also is a set of specifications for evaluating and measuring achievement after a particular stage of instruction. All these issues are left to the teacher to work out for himself and his students, a mammoth task, which, in our EFL situation, can prove difficult to achieve satisfactorily.

1.1.4.3 Conclusion

By way of concluding our brief assessment of writing manuals, we would like to point out that mastering efficient written text composition is accepted as a goal of EFL practice. The planners of an EFL syllabus and the designers of the teaching material are, however, not always confident about when the skill of writing should be taught. Recent trends in EFL pedagogy, with their undisputed emphasis on the supremacy of the spoken language, have tended to push the teaching of the writing skill to the background. EFL textbooks that are based on these trends and that aim at a language proficiency programme delay the written work component to the end of the unit of instruction. Often this component includes exercises that are disintegrated, both in theme and language complexity, from the main unit. Additionally, lack of integration is evident in the manner in which the written work components throughout the textbook are related to each other.

EFL textbooks that are dedicated to teaching writing tend to be more concerned with methods and classroom techniques than with ensuring the effective attainment of their aims. On the other hand general written composition manuals devoted to native speakers of English are characterised by an over-emphasis on mechanics and usage and, more often than not, by an ambiguity in the treatment of rhetorical and stylistic aspects of writing. In an EFL situation, such manuals can be used as a general reference rather than a course book.

More fundamental to our purpose is the fact that both types of instruction manuals surveyed offer very little guidance to such specific teachers and·learners as those participating in an Arabic EFL situation, guidance that can monitor and evaluate their progress towards attaining the desired goals. Needless to say, since these manuals address a wide range of learning situations, including the native users of the language, they are hardly based on any valid foundation of objectively observable, empirically describable contrastive textual analysis. Bald assertiveness is implemented instead of a careful and demonstrably consistent work-out of the variations in the textual system of the target language (English) and that of the source language (Arabic). One essential aspect of this system is the ways and means of concatenating clauses and sentences throughout the development of the text.

When objectively evaluated, these surveyed manuals will fail to adequately satisfy the criteria of pedagogical sufficiency, efficiency, and appropriateness. This immediately questions the necessity of a course book in the writing classroom, particularly in the EFL situation we are concerned with. Indeed there is an advocacy for abandoning manuals of writing instruction (cf. the views in Hirsch 1977) and relying on the expertise and discretion of the practising teacher. In support of this view is the argument advanced by some teachers that "the student's actual learning takes place in the process of producing-and-correcting, which is an individual process for each student, rather than uniform subject matter to be gleaned from a book" (p.165). But without going into the details of weighing this argument, we would like to assert that a well grounded and appropriately written textbook can provide ample assistance for the teacher in planning and shaping up his classroom practice. Additionally, there is hardly any valid evidence at the moment that can prove the existence of a significant correlation between efficient EFL written text production and the non-use of writing manuals.

The recourse of EFL teachers and students, in their anxiety to achieve a better standard of written composition, to the existing textbooks has resulted in disappointed anticipation. This is reflected in the students' written products.

1.1.5 Improving Instruction in EFL Written Text Composition

The brief assessment offered in the last section recognises one of the dilemmas pervading EFL practice, namely the degree to which instruction in written text composition and the accompanying materials used should be structured. The discrepancies observed in the writing manuals reflect controversies that have spanned full gamut: from promotion of intensive practice of isolated skills, to programmes totally without form or structure that leave development of the writing skills to incidental learning.

The position we take is that a higher or lower degree of structuring in and of itself will not necessarily upgrade instructional efficiency, though it can contribute to that end in conjunction with other factors. The thrust of instructional planning should come from careful attention to specific requirements which provide a focus for writing-development activities.

An essential requirement, and one that in relation to broader goals of EFL writing development forms the backbone for the structural framework of syllabus design and material production, is the provision of an adequate and appropriately prepared instruction component that aims at enabling Arabic-speaking advanced learners of writing to produce written text composition that complies to a satisfactory degree with the standards of textuality, design and linearity peculiar to English texts. The constitution and structuring of this component should be monitored by the findings of two types of research. The first is concerned with describing and interpreting the processes of EFL written text composition. This includes such tasks as the investigation of how items are selected and used in real contexts, the exploration of criteria for the construction of a comprehensive theoretical model to represent the detailed aspects and distinctions entailed in text production, the description of the cognitive predispositions and development that make human actions meaningful and relevant. The general aim is the provision of a larger and clearer picture of the mental operations that take place during EFL written text production. This type of research falls within the realm of psychology and psycholinguistics.

The second type of research investigates the variations in the textual patterning of English as compared to Arabic. It purports to identify and describe the resemblances and divergences in the manner in which the constituent elements of English and Arabic text group themselves together to form units of information, which in turn group themselves into larger units, larger wholes. The general aim is to offer an insight in the typology and constitution of texture and structure of text in English as opposed to Arabic. This type of research is the main concern of the rapidly-expanding field of text-linguistics.

It is our contention that research of the second type should precede and function as a prerequisite to research of the first type. The operations involved in text processing and production (the concern of the first type of research) are sensitive to context and knowledge of text organisation, including connectivity and linearity (the concern of text linguistics), as well as to such factors as ideologies, belief systems, values, and prior experience.

EFL text organisation is subject to obstacles such as contamination during the recovery and use of internalised textual knowledge. The nature and extent of the contamination cannot be determined unless the different modes of text organisation in English and Arabic are empirically identified and characterised.

Accordingly, real advance in the systematic description of the processes involved in EFL text production as well as the genuine intent of producing highly serviceable material and techniques for promoting EFL writing skills, require research that aims at describing variations in the manner textuality is realised in English and Arabic. Our study purports to carry out one aspect of such a task, namely the description of variation in the ways cohesion is realised as sequential connectivity through the use of connectives.

Within this context, we are aware of two objections facing the validity of what has been said. The first one is general and hinges on a suspicion of theoreticians on the part of practitioners in the field of EFL pedagogy. This suspicion, while not always unwarranted, is here dismissed on the assumption that practices depend on and derive from theory. The theory may be an unofficial, unarticulated one, held by one or several practitioners, or it may be an official theory, widely held and supported. The quality of the practices, in the majority of EFL situations, are traced back to the theories which gave rise to them. Conversely, changes in theory are bound to affect practices. If linguistic theories have, on the whole, not been conducive to enlightened and effective practice in EFL writing, this may have been due to the indifference of theoreticians of language to the needs of practitioners, or to inherent limitations in linguistic theories. Such a situation need not last long. Theories and studies within the framework of text linguistics offer systematic treatments of aspects of textuality and organisation that can prove essential for the teaching of such EFL skills as reading and writing. These aspects include internal cohesion of texts, the connectedness of parts of texts, the development of thematic material, paragraphing, paraphrase and restatement.

The second objection is more specific. It is posed again by practitioners as well as textbook writers and syllabus designers, and is expressed in this way. If we wait for good research to point out to us the best way to organise an EFL composition course for a particular group of learners, and the format of a pedagogically efficient composition textbook, we would wait a long time, and meanwhile we would still need to teach the skills of writing. The objection is a valid one and can be regarded as part of a larger problem where EFL policies have traditionally been diffuse, fragmented or uninformed. However, remedy seems to lie in the cooperation of theoreticians and practitioners on a large scale and without the customary delays. Progress demands that theory and practice maintain a constant constructive evaluation of accepted priorities, attitudes, diagnostics, and so on. The findings of theories or models of language and discourse and the available investigations of textual problems within one language or across languages should be exploited towards furnishing insights into more efficient methodology that can ensure success in the writing classroom. Through cumulative learning positive results can be reflected in the learner's written product.

It is with such conditions as background that we approach the main task of this study, namely the description of similarities and diversities in the manner connectives operate within English and Arabic. Such a description requires an efficient investigatory apparatus that can effectively be used to empirically discover and verify the means through which connectives, as an essential marker of the cohesion of text, vary in their categories and distribution in the two languages. But before we describe the components of our investigatory apparatus, we would like to offer brief introductory comments (more detailed explication is provided in chapter 4) on text, cohesion, and connectives and their place in written text composition.

1.2 Text, Cohesion and Written Composition

1.2.1 Text and Composition

First, a few introductory comments are offered on the concept 'text'. The concept, as is the case with all theoretical entities in linguistics, has been viewed from different angles and therefore a number of definitions have been attempted. (See 3.1 for a detailed account). A structural definition would consider text as a sequence of lower-level constituents, such as sentences, and a formula can be given in which text is regarded as $S(+S)^n$ where n>1. According to this definition a text is made up of a minimum of two sentences. Unfortunately the concept 'sentence' itself is controversial (cf. Williams 1984b) and different schools of linguistics offer different conceptions. One classical example is the concept offered by the Arab grammarians centuries ago defining the sentence as an entity after which a brief silence seems best. However, one can argue that this definition can equally be true of 'text', and yet there exists a substantial amount of research, particularly within text linguistics, that would place these two concepts on two different levels of abstraction, each with its own inherent properties.

A functional view of text (cf. Halliday 1978, 1985, Hasan 1978, 1979, Halliday and Hasan 1976) would describe text as a "unit of language in use". Accordingly, a text can be a mere sentence or a whole novel, the emphasis being on its role as a communicative unit in language.

But whatever one's views on the nature and place of text in linguistic theory, one must eventually accommodate to the fact that much of adult language behaviour displays itself in text creation. Hence, text takes up a particular value within language functioning, to the extent that text receivers (listeners or readers) sometimes go to great lengths to interpret as text anything that is said or written, and are ordinarily ready to assume any kind of displacement, for instance some error in production or in their own understanding, rather than admit that they are being faced with 'non-text'. But despite such an attitude, text should be regarded as an entity that is structured (or 'textured' to use Halliday's term) in such a way that it can readily be differentiated from non-text. Text structuring (or texture) can be accounted for in terms of principles of textuality and linearity, a task that is considered central in text linguistic literature (cf. discussions in Wikberg 1978, Enkvist 1978b, 1985, Beaugrande 1980, Pugh 1981, Al-Jubouri and Knowles 1986, Al-Jubouri 1987).(2)

1.2.2. Cohesion and the Text

To approximate such a task, text must be conceived of as the product of a process of composition and concatenation (Scinto 1983). This implies that text is made up of constituents of lower value than the text itself that constitute the input to the process. Concatenation is by no means a haphazard process; it is constrained to a considerable extent by the requirements of textuality (see 4.3 below) and is achieved through meeting the organisational demands of linearity (see 4.3 below). The operations involved in this process are the core of cohesion, and the organisational unity of the final product characterises it as a text.

Concatenation of the lower-level text constituents is realised through connectivities that take up two dimensions. The first dimension represents interconnection of surface constituents indicated, implicitly or explicitly, by specific connection signals that reflect the working of the different mechanisms of cohesion and that ensure propositional development within the text. The role of the signals is manifested by the fact that the constituents cannot by themselves contribute to textual development. Their contribution starts when they exhibit connectivity whereby each constituent is interrelated with another, thus building up the text.

The second dimension, discussed by Beaugrande (1980) under conceptual connectivity (cf. also the discussion of various aspects of coherence in van Dijk 1977a, Hobbs 1978, 1979, Vasiliu 1979, Sørenson 1981, Witte and Faigley 1981, Robinson 1984 and the papers in Neubauer 1983 and Tannen 1984), refers to the ways in which the components of the textual world, i.e. the configuration of concepts and relations which underlie the surface text, are mutually accessible and relevant. Beaugrande and Dressler define a 'concept' as "a configuration of knowledge (cognitive content) which can be recovered or activated with more or less unity and consistency in the mind", while to them "relations" are links "between concepts which appear together in a textual world: each link would bear a designation of the concept it connects to" (p.4). Coherence in this sense is envisioned as the product of connecting concepts and relations into a network composed of what Beaugrande and Dressler call "knowledge spaces" (p.94) (i.e. internally organised configurations of context in the mind) centred around main "topics". It follows that in employing a text, i.e. in speaking or writing, the language user constructs chains of concepts and relations to organise the textual world around a particular topic. Maintaining conceptual connectivity creates an interaction of text-presented knowledge with users' stored knowledge of the world.

The first dimension is more relevant to our aims and will be explored with more depth in the relevant sections of this study. Text concatenation is accomplished on an explicit surface level through the manipulation of cohesive devices, one of which is connectives. A preliminary consideration of connectives is now attempted, with particular reference to the question of composition. A more extensive elaboration is offered throughout the study.

1.2.3 Connectives and Text Composition

The textual role of connectives can be observed in sustaining cohesion among text constituents and in characterising overall texture. If we assume (somewhat loosely) that a constituent embodies a single idea, then the sequential connection of constituents is a necessary prerequisite and a corollary of the development of complex thought and expression.

Improper rendering of connection, particularly through inefficient use of connectives, has been found as an indicator of immature writing of native speakers of English. Black (1982) refers to a study carried out by Cooper et al. (1979) of the Fall 1979 entering class at the State University of New York at Buffalo. This study indicates that students have major writing problems, and, in identifying the locus of these problems, specifically states that the students "have great difficulty creating written text which has adequate connections and relationships from sentence to sentence" (Black 1982 p.200). The recommendation advocated is "not for more drill in the mechanics of writing but for better teaching of the written composition process especially at the intersentence level" (op. cit.).

In EFL written text composition, failure to handle connection properly is further aggravated by the student's use of his native language (in our case, Arabic) resources for sentence combining.

Yorkey (1977) refers to the Arabic-speaking learners' use of "wa-wa method" of sentence combining in their effort to produce a connected text. This failure explains why students who perform admirably in standard grammar exercises and appear to have a good command of English nevertheless fail to produce acceptable texts (cf. Pincas 1982 pp.55-6, cf. also Leavelle 1984).

Connectives are then a perennial problem in EFL text production and should be given some attention. Classroom practice has shown that it is relatively easy to persuade students of the importance of such items as "on the one hand", "subsequently", "paradoxically", "therefore", but when it comes to produce a text of a fair length, the global view required to produce these appropriately is often missing. Students either fail to see the need to provide the information the reader needs to "bridge the gaps" between sentences, or insert the wrong connective or impose an unnecessary one. In most cases, the selection of the wrong connectives is based on the students' internalised knowledge of the form, function and distribution of connectives in Arabic. The product is a text that manifests incomplete propositional development of an argument through absence or misuse of what Nash (1980 p.21) calls 'overt marking of transition' or 'directive clues' (cf. Coupland 1984).

Having outlined the relevance of cohesion and connectives to text composition, we now consider the type of research we intend to carry out, and the nature of the methodology involved.

1.3 Nature of the Investigation

Sufficient motivation has now been evidenced for initiating an investigation into the nature of connectives as a cohesive phenomenon. More specifically, we are interested in the description and measurement of the variations in the patterns of connectives as manifested in English texts opposed to Arabic texts. We propose to carry out this investigation by performing experimental work whereby the behaviour of connectives is observed, described and quantified as they occur within statistically adequate corpora of running text.

In conducting such a piece of inquiry we are aware of a number of problems that require effective decision-making. Two types of problems can be identified and are then outlined, due to their immediate impact on shaping the process of the investigation. We refer to them as substantive problems and metascientific ones.

Substantive problems concern such issues as the nature and efficiency of the theoretical framework within which connectives are to be analysed and described. Once this is identified, a synthesis of views has to be worked out concerning such controversial issues as textuality, cohesion and connection. Another problem hinges on the diversity of views concerning some basic concepts that are considered fundamental for the investigation itself. Among these are the concepts of "text", "connective" and "word".

The second type of problems are metascientific. They concern the establishment of an efficient investigatory apparatus that can be used for achieving the primary and secondary aims of the project. Another problem that demands specific attention is the formulation of criteria and requirements for assembling a corpus of text where the phenomenon of connectivity is to be studied. The need for using a corpus in the conduct of linguistic inquiry is itself a source of controversy and therefore calls for reconciliation.

To overcome these problems, a series of decisions have to be made, explicitly or otherwise. In the course of our experimental work decisions are sometimes made with full recognition of the implications. At other times, however, decisions, due to a variety of factors, are made simply as a matter of convenience or by rules of thumb. In general, decisions to handle our substantive problems involve selection and definition. The selection of a theoretical approach to text-based analysis entails the use of theoretical terms most of which have been defined in a number of different ways. Our use of these terms is sometimes based on operational definitions conceived of in the hope of facilitating the process of investigation.

Metascientific problems require a variety of classes of decisions. The first involves decisions about sampling: manner of selection, size of sample and methods of assembling the corpora. The second class involves decisions about the type of experimental work and its potential effect on a) the possibility of making generalisation, b) the applicable statistical techniques, and c) the appropriate design method. Another class of decisions are concerned with the experimental design: the manner of observing connectives, the nature of the categories to be identified, the sequence of implementation, whether computer techniques are to be employed and the size of computer involvement. A fourth type of decision involves measurement. There has been increasing recognition in recent years that the conclusions we draw from experimentation depend on what and how we measure. Methods and procedures of analysis are obtainable from quantitative linguistics. However, some methods and procedures are much more specific, efficient or appropriate than others. Hence, it is necessary to display an adequate understanding of the specific assumptions, uses and limitations of the various statistical techniques in order to manipulate them effectively.

All these decisions make up an essential part of the investigatory apparatus and provide a framework for setting the steps involved in the research. The next section considers the nature, function and components of this apparatus.

1.4 The Investigatory Apparatus

1.4.1 Requirements of Efficiency

Each linguistic project that proposes to investigate a complex phenomenon should provide for a plausible, coherent and well-specified investigatory apparatus. Such an apparatus can be presumed to offer a progressively definitive identification, analysis and explanation, and should therefore include all arguments, research procedures, methodological assumptions and decisions that can have a direct bearing on the investigation.

The question of the validity, necessity or sufficiency of the use of an apparatus is one that is related to the more general debate on the methodological status of linguistics. The debate is by no means a concluded one (cf. for instance, the papers in Perry

1980) and since it is connected with the fundamentals of model-building in language theorisation, it will perhaps continue at all levels of linguistic inquiry. We shall therefore attempt to avoid considering these issues. Rather, we are more interested in specifying the characteristics expected in the apparatus we intend to use and its components.

An apparatus should possess, as to its eventual aim, some ultimate degree of efficiency with which the investigation can achieve its goals. One fundamental requirement of efficiency is the correct identification of aims that the apparatus seeks to achieve. Potentially successful inquiry can be badly frustrated by a faulty design of aims. Wrong aims will either produce fallacious inferences or, at least, hinder the work of the apparatus. Correct aims that are set too high or too broad can result in some methodological indeterminacy within the function of the apparatus that can affect the scientific status of the evidence or conclusions. Aims that are out of reach can invalidate any investigatory apparatus. On the other hand, too narrowly-defined aims can direct the apparatus in such a way that the evidence produced will be of little theoretical or empirical importance.

The efficiency of the investigatory apparatus hinges on its capability of effectively responding to these two metascientific questions: a) what is the project attempting to find out, and b) how much will the resulting evidence advance our knowledge of the phenomenon under focus. Consideration of these two questions can determine to a large extent the value as well as the degree of sophistication that the design of an apparatus exhibits.

These two questions are then prior to the question that constitutes the whole function of the investigatory apparatus: how do we, as effectively as possible, set out to achieve our aims? Indeed, the degree of sophistication of any apparatus develops relative to the particular aims of the project, and hence its overall value can be judged on whether it has effectively or ineffectively achieved the aims.

Another requirement for an efficient apparatus is the appropriate theoretical framework within which the apparatus can function. This entails that different theoretical approaches will influence and in many cases determine the manner in which the apparatus operates. Using an apparatus for an approach other than the one it is designed to function within can lead to a number of theoretically or practically questionable results.

A third requirement is the appropriate selection of data on which the apparatus as a whole can operate. Observed data are in fact the basis, in a sense, for postulating inferences or rules, since we would not want to acknowledge something as a regularity if it were not verified in at least a sizeable proportion of natural language text.

The fourth requirement of the efficiency of the apparatus is an appropriate mechanism for measurement. This mechanism does not only operate on the data, but it also functions as an evaluatory measure for the conclusiveness of the findings. Among the widely used measures of conclusiveness in empirical studies are experimental controls, logical consistency as well as statistical procedures. In our apparatus this mechanism resides with the statistical component. It is responsive to the adequate delivery of answers (description, explanation, synthesis) demanded by the aims. It is also responsive to the manner in which the arguments are constructed and evidence is used or deduced.

These requirements, we believe, promote the efficiency of the apparatus. An efficient apparatus is capable of rendering the linguistic phenomenon under investigation, connectives in the case of our project, more amenable to systematic investigation and informative analysis. In the next section we consider very briefly the main components of the apparatus.

1.4.2 Main Components

The investigatory apparatus used in our project is composed of a number of components. Each component has its contribution in the analysis of the corpora while at the same time it interacts with other components, thus reflecting unified analytic instrumentation in its overall examination of the corpora. Identifying these components and their interactive role helps elucidate the workability of the whole apparatus. What follows is a brief description of these components. A fuller discussion is reserved for the next chapter.

The theoretical component represents the base of the apparatus. It includes a number of substantive decisions (see 1.3 above) that refer first to the type of text-based approach selected for the analysis, as well as to its contrastive nature. It includes a number of theoretical statements concerning text cohesion and connectives, some are of axiomatic nature while others are belief statements representing our conjectures and interests.

The procedural component includes decisions and steps taken to ensure an efficient and systematic selection and manipulation of the data for the purpose of operatively adequate attainment of the aims. A number of these decisions concern the application of computer techniques, which, because of the size of the corpora, are an essential part of the research. The potentials of computer use in linguistic studies are obviously extensive and this has been proved through numerous studies and projects in the sixties and seventies. This component of the apparatus organises the selection of the corpora and their coding in machine-readable form. Furthermore, it arranges and regulates use of relevant packages and programs for analysis.

The third component concerns the statistical methods selected from a repertoire that is available in 'quantitative linguistics'. These procedures make up in their totality what is here called a calculus of connectives (see 1.5 below).

Given the above brief characterisation of the apparatus, we can now consider its operational mode within our study. It is our hope that the apparatus can function in a systematic manner that can render a principled achievement of the aims of the investigation.

1.4.3 Operational Mode

The investigation we intend to carry out should include three characteristics which qualify and direct the functions of the apparatus. First, the investigation is systematic and controlled, basing its operations on a bottom-up model of analysis. Second, the investigation is empirical; we turn to experience and observation for validation and our findings are drawn from and checked against objective reality. The set of observations made are ordered and analysed to answer the crucial questions posed by the investigation and provide a more dependable as well as a deeper and fuller understanding of the nature of the phenomenon of textual connectives in English as compared to Arabic. And third, our investigation is self-corrective; not only does it have built-in checks to prevent any distortion of the data, but, in addition, the procedures adopted are open for scrutiny and are therefore an open target for other researchers to either challenge and refute or corroborate and extend.

The apparatus we envisage functions in the following operational stages:

a. Background issues: The investigation begins with a consideration of a set of theoretical issues concerning text, textuality and linearity and the manner in which connectives operate to achieve textual cohesion. In this way the phenomenon to be investigated is isolated, with all its potential variables, and an attempt is made for the formation of some general notions concerning the nature, form and function of connectives.

b. Observation: Two text corpora are set up in machine readable form to provide data bases for empirical observation.

Methodological options available, decisions taken and choices made, whether on theoretical, implementational or pragmatic grounds, are discussed.

c. Categorisation: This is a basic procedure for reducing isolated data to a functional basis. The aim is the systematisation of otherwise incomprehensible masses of data. Connectives that are isolated through observation are investigated and categorised according to the type of interpropositional relationships they express.

d. Quantification: This is a more sophisticated stage where precision of measurement allows more adequate analysis of the phenomenon of textual connectives by mathematical means. Statistical procedures, making up in their entirety a calculus of textual observations of connectives, or, for short, a calculus of connectives, are developed to measure the properties of connectives.

e. Discovery of Relationships: Through the previous stages, variations in the patterning of connectives in English and Arabic are identified and explained. Findings are drawn from the comparison which can be manipulated in the next stage.

f. Practical Implications: The findings are examined and suggestions are made for pedagogical purposes.

Having discussed the nature, efficiency, components and operationality of the investigatory apparatus, we now pause to consider very briefly what we have meant by the "calculus of connectives" mentioned above. This arrangement is deliberately made in order to clarify our use of this term and avoid possible misinterpretation. A more detailed account of the calculus is offered in relevant chapters (cf. Chapters 7, 8 and 9).

1.5 The Calculus of Connectives: A Preliminary Note

1.5.1 Fields for the Quantitative Study of Language

In assessing quantitative linguistics, Herdan (1962) suggests that there are two fields for the use of statistical procedures in the study of language. In the first, statistics is used as an auxiliary tool, mainly for the purpose of testing hypotheses. The statistical procedures can, on any level of language study, evaluate, whether by themselves or in conjunction with other methodological procedures, the evidence in favour of one or the other hypothesis. The particular problems under investigation are themselves not statistical in nature, and, therefore, statistical procedures that are employed in other branches of knowledge are, as a rule, adequate, probably with some modification imposed by the very nature of the linguistic problems. Such are the procedures suggested by Anshen (1978), Hatch and Farhady (1982), Butler (1985a) and, to some extent, Williams (1970) and the recent introductory work by Kenny (1982).

In the second field, the problems themselves are of a quantitative nature and therefore require statistical procedures. Put differently, the linguistic problems and concepts that are operated on within the description or interpretation require for their precise use or analysis certain statistical procedures which are not superimposed but, rather, constitute a part of linguistic thought and method. Such statistical procedures are, among others, discussed at length by, for instance, Yule (1944), Herdan (1956, 1960, 1962, 1964, 1966) and Brainerd (1974). (See a brief review in Johnson 1976).[3]

We believe that there are a number of linguistic phenomena, the description of which will fall within a grey area that results from an overlap of the two fields. In other words certain linguistic problems are essentially of a qualitative, non-mathematical nature that, despite this, manifest certain properties that are quantitatively determined. Within this area falls the statistical description of the phenomenon of connectives.

Connectives, we believe, manifest specific qualitative properties peculiar to their textual role and the semantic relationships that they secure within or across sentences. In addition, they display quantitative properties that can characterise text connectivity within a particular language (and even within a particular genre, cf. for instance Smith and Frawley 1983, but this divergence is overlooked in this research). Hence, a qualitative as well as quantitative comparison with properties of connectives in other languages is, at least at the theoretical level, feasible. The set of quantitative procedures used within this project for the description of the statistical properties of connectives is here labelled the calculus of connectives.

1.5.2 Features of the Calculus

The need for the calculus arises from two sources. The first one is related to the requirement of objectivity in describing a phenomenon. This requirement is dictated by the empirical nature of the study and the fact that within linguistic analysis it is necessary, particularly when interlingual comparison is attempted, to eliminate human bias, or at least minimise it and reduce it to insignificance. In achieving this, the linguist, however, must be aware of another type of bias that can adversely affect the value of the results, namely statistical bias. Our analysis of the computations of the calculus should therefore aim at nullifying the effect of this factor.

The other source for demanding a calculus is related to the question of corpus versus language. Such a question hinges on whether we adopt the view of language as a "virtual" system or an "actual" system (cf. the treatment of these concepts in Chapter 3, see also Beaugrande 1980, 1984). So long as the analysis is based on the view that language is a virtual system, there is no need for the calculus. Conversely, if the opposition of corpus versus language is brought in, then the need for the calculus emerges. This is simply because this opposition is nothing but that between sample versus statistical population (Herdan 1960, 1962). Since, within this study, a need is explicitly made for the use of parallel corpora (cf. 2.5 and 5.1 below) in achieving an interlingual contrastive textual analysis of connectives, a consideration and development of the calculus are immediately called for. The typology of the procedures that make up the calculus is not discussed in this preliminary note (See Chapter 3).

1.6 Aims of the Study

In specifying the aims of our research, we would like to distinguish two sets of aims: primary and secondary. The primary aims are concerned with the fundamental problem of the investigation and constitute its core and specific point of focus. The secondary aims are not the immediate concern of the investigation, but their achievement strengthens and supports the perspective within which the primary aims are to be realised and optimises the efficiency of their attainment. Both sets of aims contribute to the pursuit of one ultimate aim for this study. Specifications of all these aims now follow.

1.6.1 Ultimate Aim

We have argued in the preceding sections for the existence of a demand for an interlingual study of connectives, which, we believe, constitute a problematic aspect in textuality in English as compared to Arabic. The word "problematic" is used here in a metascientific context to describe an aspect that represents a state of affairs which linguists do not fully understand. A linguistic aspect is problematic because there is a gap in the existing knowledge. In other words, the fragment of knowledge in which this state of affairs could have been understood (or at least better understood) is missing.

The ultimate aim of our investigation is then the search for insight into and understanding of the system of textual cohesion as realised through connectives. However, the question as to the nature of insight represents a complex philosophical problem, which, in order to be adequately defined, requires a highly technical and philosophical discussion. Such a discussion falls outside the scope of our study. We shall therefore approach the problem from another angle (cf. Botha 1981) and reformulate it as follows. Under what circumstances can we have the feeling that we have (gained) insight into, or understanding of, a problematic state of affairs such as the similarities and variations in the behaviour of connectives in English as opposed to Arabic?

The question is obviously of crucial importance. The "circumstances" referred to are described by scientists (including linguists; cf. Botha 1981) in terms of expressions such as "regularity", "pattern", "structure", "mechanism" and "cause". One can then claim that to have insight into the system of connectives, as indeed with any other problematic state of affairs, involves being able to see it as a manifestation of an underlying regularity, fit it into an underlying pattern, identify it as part of a structure, indicate its mechanism and point out the cause of whatever is problematic. These claims, overlapping and informal as they are, are, as far as our central problematic phenomenon is concerned, insufficient to create an adequate insight. Furthermore, the use of the word "feel" in the question posited above has the apparent indication that our ultimate aim of the study is subjective when scientific knowledge is required to exhibit objectivity or intersubjectivity.

We shall therefore characterise more precisely the circumstances in which we have insight into a problematic state of affairs such as the one under focus. This characterisation is intended as a framework for the primary aims of the study, within which lies the answer to the question about the nature of insight.

To have insight into the similarities and variations in the system of English and Arabic connectives involves:

a. being able to give a description of the regularity, pattern, structure, mechanism and, where applicable, causes underlying this system;

b. being able to give an explanation of a contrastive nature to the working of the system;

c. being able to infer pedagogical predictions and suggestions that constitute guidelines for a more effective teaching of advanced writing to Arabic-speaking learners.

The ultimate aim is then realiseable in three more immediate aims that are here labelled primary aims and to a lesser extent in a number of secondary aims. These are specified next in more detail.

1.6.2 Primary Aims

1. The first immediate aim and the main thrust of the investigation is the description of connectives in English and Arabic. This description is characterised as follows:

a. The description constitutes an image of the properties of connectives as they occur in a statistically adequate corpus of running text.

b. It approaches (a) by:

i. first discussing the broader area of textuality, linearity and cohesion.

ii. specifying the role of connectives as cohesive ties and categorising the type of interpropositional relationships obtainable through their use.

iii. working out a calculus of textual observations of connectives that pinpoint their statistical behaviour in both the English and Arabic corpora.

c. More specifically, the following are to be considered:

i. What is text and how is textual cohesion related to the broader questions of textuality and linearity?

ii. What models of cohesion are posited and how does this research approach them?

iii. What kind of entity is the connective? What relationships do connectives express?

iv. In categorising connectives how are the different subcategories interrelated?

v. What is the nature of the multifunctional connectives?

d. In working out the calculus of connectives, the following statistical features are considered:

i. The frequency distribution of connectives and their categories.

ii. Quantitative properties related to type-token statistics.

iii. The measurement of growth rate of connectives in texts.

iv. The measurement of interval rate in the occurrence of connectives.

v. The calculation of probabilities of repeatedness.

2. The second immediate aim of our investigation is a contrastive textual evaluation and explanation of similarities and variations in the patterning of relationships expressed by English and Arabic connectives. Evidence is drawn from observation of connectives in the corpora and is verified against intuitive judgements. Hence a consideration of the dichotomy of corpus versus intuition is essential in delimiting the nature of these judgements.

Explanation is here based on the description of connectives mentioned in the first aim. It does its work, not by invoking something beyond what might be described, but by putting one fact into relation with others. Each element of what is being described is better clarified and understood through its relation with other elements, and it is because they come to a common focus that together they shed light on the problematic aspects of connectives.

3. The third aim is that of making implications for EFL pedagogy based on prediction of the typology of textual problems involved. More specifically we would like to find out how a statement of the properties of connectives can be manipulated in formulating suggestions that can be incorporated in designing or supplementing teaching materials in advanced EFL writing skills for Arabic-speaking students.

These aims give rise to a variety of metascientific questions. The answers intended to some of them constitute the secondary aims of this study.

1.6.3 Secondary Aims

The secondary aims are directed towards the formation of heuristic strategies that guide the use of the investigatory apparatus and consolidate the empirical nature of the investigation. One fundamental secondary aim is the conduct of an experimental corpus-based text analysis for search of the patterns and categories of connectives mentioned in the primary aims. This aim can be pursued through the attainment of a number of constitutive aims, which results in the optimisation of the perspective within which the study is to be accomplished. These constitutive aims are:

1. The experiment is computer-aided. Computer techniques are to be employed as part of the function of the investigatory apparatus. This requires a) an examination of the software available, and b) an identification of the related problems, such as the method of automation of the data-bases, the computational status of "word", the status of multi-word connectives in computer use, the appropriate tagging procedures, etc. Suggestions for practical solutions are to be made and incorporated.

2. Two corpora of running text, comparable in size and method of selection, are to be prepared in machine readable form, one in English and the other in modern standard Arabic (refer to Chapter 5 for a definition of this term). Means of encoding the texts, their possibilities and limitations are to be considered.

3. Word lists and concordances are to be produced in order to assist in achieving the task of identifying connectives, providing sufficient context for their use and giving a statistical image of

their distribution. Problems of a computational nature related to the attainment of this aim are to be considered and solutions are to be justified and implemented.

4. A pilot experiment is to be conducted to assess the feasibility of carrying out the experimental component of the project and to determine the significance of the findings.

1.7 Scope and Limits of the Study

The aims as formulated in 1.6 above make it explicit that our investigation seeks to accomplish an observationally adequate description and explanation of language-pair-specific textual connectivity. The aims are achieved through an empirical-inductive approach that attempts to have scientific substance and relevance and is founded on basic theoretical assumptions that will be discussed in later chapters. As with any study the scope is restricted by a number of factors and we feel that this should be clarified right from the outset. One should not only make claims of what will be accomplished, but also categorically specify what is not to be included or at least expected from the investigation.

One of the distinctive attributes of the description and explanation offered for the patterns and functions of connectives in English and Arabic - and one which, we believe, is shared by the majority of studies, particularly empirical ones - is "openness". In no sense can we claim that our characterisation of connectives is final. We fully endorse the view (cf. Kaplan 1973, Cohen and Manion 1980) that the functions that descriptions and explanations perform cannot be appreciated without a full awareness of how far from finality they are in the actual conduct of inquiry.

The ways in which the findings of the description, explanation and prediction are characterised as open can be identified as follows:

1. The findings are partial: Given the restrictions of time, fund and equipment and the limitations of the corpora, we could only explore some of the properties that qualify the phenomenon of textual connectives.

2. The findings are conditional: They are based on systematic observation of a limited, though statistically adequate, corpus of text, and, therefore, hold true only for a certain range of connectives.

3. The findings are approximate: The magnitudes they yield bear only an approximation to truth, this is again due to the factors mentioned in (1) above.

4. The findings are limited: This is due to two factors: a) the research is experimental in nature and hence it is, in general terms, only relatively conclusive, and perhaps will remain so even if more elaborate extension is attempted, and b) the findings are appropriate to particular contexts in which they can be practically manipulated; for other contexts they may demand a different angle of view, or even a different theoretical and experimental orientation.

In short, we would like to state that the findings of this investigation are produced by a corpus-oriented investigatory apparatus and are indeed of statistical nature. They do not claim to completely cover the reality of the phenomenon of textual connectives from all perspectives (such as those of text linguistics, psycholinguistics and/or the science of communication), something which would be beyond the reach of empirical investigation anyway. However, the more successful our findings are in demonstrating regularities in the mechanism of connectives, involving even connectives that occur less frequently, and thus conducting an optimal investigation, the greater will be its textual and theoretical information value.

1.8 Plan of the Thesis

A last note concerns the plan of this work. The thesis falls into four volumes. Volume (1) includes five chapters that discuss a number of relevant theoretical and methodological issues. Chapter One is an introductory chapter that sets the scene for the study. It includes arguments for initiating the investigation, points out the directions that it will take, outlines the type of investigatory apparatus to be developed and states the aims, essence and scope of the study. Chapters Two and Three are concerned with a more detailed outline of the operational mode of the investigatory apparatus. A discussion is offered in Chapter Two for the selection of text linguistics as a framework of analysis and considers with some details the nature of the interlingual analysis to be performed. Chapter Three and Four start a background theoretical consideration of a number of relevant questions. Chapter Three discusses text, textuality, linearity and cohesion, while Chapter Five discusses connectives. The arguments in both chapters are supported with a sufficient review of literature. Chapter Five offers a detailed description of the experimental work. The Chapter starts by a brief account of the set-up and results of the pilot experiment. It, then, discusses the question of corpus versus intuition and summarises decisions taken in this respect. Later, it gives a full account on assembling the English and Arabic corpora and examines problems related to computerising the input. The Chapter then reports on the steps of computer processing of the corpora, tagging and the production of word lists and concordances.

Volume (2) falls into three chapters (6-8) that are mainly concerned with analysing connectives and discussing the results of their quantification and, thus, form the gist of the research. Chapter Six offers a description of the functional categories of connectives and describes their behavioural patterns in English and Arabic. Chapter Seven and Eight are devoted to a description of the calculus of connectives. This is preceded by an account of the various quantitative features of each corpus.

Volume (3) of the thesis is comprised of two chapters and Part I of the appendices. Chapter Nine is contrastive in nature and aims at describing quantitative and textual variations in the behaviour of connectives. The final chapter aims at reconsidering the problem of connectivity across the two languages in the light of the evidence that the project has provided. It, then, formulates some pedagogical implications for the teaching of writing, concludes the study and offers suggestions for further work.

The Appendices are made up of three distinct parts. Part I is included within Volume (3) and comprises non-statistical appendices. Appendix (1) gives a detailed account on the status of the "word" as a unit of linguistic measurement. Appendix (2) offers a broad review of text-based approaches. The rest of the appendices offer some explanatory details that clarify or strengthen certain arguments in the thesis.

Volume (4) contains Parts II, III, and IV of the appendices. Parts II and III comprise the statistical appendices that are referred to within the thesis. Part IV contains concordances and word lists and is produced on 55 microfiches.

Footnotes to Chapter 1

(1) Jordan (1980 pp.93-96) offers a brief list of some common connectives as an appendix. But no practice is offered, and for information about their use, he refers the learner to a "good dictionary"!

(2) For an account of written composition within a cognitive perspective see Bartlett (1982), Beaugrande (1982), Cooper (1982), Scardamalia (1982), Freedle and Fine (1983) and the papers in Whiteman (1981) Vol. 2. See also the papers in Winterowd (1975) for a treatment of composition, theory and pedagogy, within a rhetorical perspective. For an account of composition within a textual (especially, systemic-functional) perspective, see the papers in Couture (1986).

(3) For a short account of mathematical linguistics see Plath (1961). See also Rosengren's (1971) discussion of the quantitative concept of language. For a comprehensive annotated bibliography of statistical stylistic/linguistic studies see Bailey and Dolezel (1968).

CHAPTER TWO

The Investigatory Apparatus: Validation of Approach

2.0 Perspective

One of the primary tasks of the linguist as an experimenter is to provide an adequate explication of his apparatus of investigation and to ensure that the components of such an apparatus can jointly provide an optimally informative and systematic account of the properties, known or discovered, of the phenomenon investigated. It is therefore the aim of this and the next chapters to attempt a detailed characterisation of the investigatory apparatus as used in this project and outline its theoretical as well as methodological orientation. More specifically, we would like to examine the framework, both theoretical and methodological, within which the project was carried out.

This aim will be achieved in two major steps. First, we shall concern ourselves with a short critical account of the theoretical orientation and tasks of text linguistics as the model within which this project is to be conducted. A review of the various approaches to text-based analysis that are available to the researcher is offered in Appendix (2).

The next step is a detailed consideration of the nature and tasks of contrastive analysis and the demands of a more textual type of contrastive analysis. The chapter ends with some guidelines for both theoretical and methodological implementations, which will be elaborated in the next few chapters.

A note on the organisation of this chapter is in order. To substantiate its aims, this chapter is divided into five main sections. Section 2.1 concentrates on the aims, tasks and orientation of text linguistics as the basic framework for this study. Section 2.2 tackles the problem of contrastive analysis, and its theoretical as well as its applied perspectives. Later, in Section 2.3, we examine the arguments for a more textual type of contrastive analysis, reviewing - as we do so - certain relevant approaches. We draw upon these approaches in validating the type of contrastive textual analysis we intend to follow and for outlining the main procedures of analysis. This is included in Section 2.4. The last Section summarises the chapter and concludes the main arguments.

2.1 Text Linguistics: Scope and Tasks

2.1.1 Preliminaries

Research in text linguistics can be dated back about twenty years. It is not possible, nor is it indeed necessary, within the scope of this limited section, to offer a detailed chronological survey of the arguments and views that contributed to or accompanied the emergence of this trend in linguistics. Such a survey can be found elsewhere (cf., for instance, van Dijk 1972, Rieser 1978, Beaugrande 1980, Beaugrande and Dressler 1981, and for a more detailed documentary survey see Hatim 1981). Here we are primarily interested in the relevance of the main framework of the theorisation of text linguistics to the issues of cohesion and texture, and in particular to the task of understanding the textual role of connectives. The quest for this relevance is approached through an examination of the scope and tasks of text linguistics and the arguments that constitute the backbone of the rationale behind its development. Needless to say that, as with any other theoretical trend in linguistics, there has been, since the start of the evolution of textual formulations, much talk against and in favour of text theory and text grammar. Moreover, the last few years have seen a great amount of highly interesting descriptive, theoretical and applied work, published in many papers and books and extending to such disciplines as poetics, cognitive psychology, and social sciences; yet, at the same time there have been certain misunderstandings as well as problems that were not adequately handled. However, the exposition of all this does not concern us here, nor do we find the clarification of problematic issues of any direct relevance to the phenomenon we intend to investigate. Accordingly, our outline will avoid these thorny areas, thus concentrating on prominent questions of immediate concern.

Interest in studying textual structures started in the sixties and was extended and intensified during the seventies especially within the framework of European linguistics. Early textual formulations are associated, among others, with Dressler, Schmidt, Harweg, Halliday, Hasan, Isenberg, Bellert, and van Dijk. Fundamental contributions have also been made by such scholars as Petöfi, Rieser, Raible, Pike, Grimes, Longacre, Enkvist, Ihwe, Kummer, Hartmann and others. The scope of text linguistics has been delimited by the very arguments that were put forward to justify its legitimacy. Some are examined below with some detail.

2.1.2 Some Arguments for a Science of Text

One of the earliest arguments hinges upon linguists' concern over the limitation of sentence grammar. This is enunciated by a number of linguists belonging to varying schools of linguistic inquiry. Harris, as early as 1960, introduces in his preface to the Phoenix edition of his book two new additions that are regarded as marking the end of classical structuralism (Rieser 1978). In the first, he proposes to describe language "as consisting of specified sets of kernel sentences and a set of transformations." (Harris 1960, p vi). In the second, he specifies that the current linguistic analysis does not go beyond the sentence; the stringent demands of its procedures are not satisfied by the relations across sentence boundaries. He admits, however, that "there are...structural features which extend over longer stretches of each connected piece of writing or talking" (p. vii). The tools necessary for describing connected pieces of language are provided by his "discourse analysis", which he proposed in a number of earlier works (Harris 1952a, 1952b, reprinted in Harris 1963 and in Harris 1970).

Van Dijk (1972) argues that many relevant and systematic phenomena of natural language are properties of "discourse", and that these properties cannot be adequately described in the existing types of sentence grammar. The formulation of text grammars is accordingly expected to provide a more adequate framework for the description of many problematic phenomena dealt with in modern linguistics (cf. also Morgan 1981).

One such phenomenon, for instance, is the native speakers' ability to disambiguate ambiguous sentences. In a text, disambiguation is "automatically" performed by the semantic and textual representation of preceding and/or following sentences. In other words, "the semantic description of a sentence S_i has to be coherent with that of the sequences $<S_1, S_2,..., S_{i-1}>$ and $<S_{i+1}, S_{i+2},...,S_n>$" (van Dijk 1972, p. 4).

Another phenomenon that is explicable in terms of a text grammar is the means by which a stretch of language is describable as a text, mainly cohesion among the various constituents that make up the text. Sentence grammars are not capable of offering satisfactory, i.e. sufficiently general and consistent, explication of these means and their interrelatedness in a text. Related to this is the identification of two types of textual structures: global referred to as "macro-structure" (cf., van Dijk 1972, 1980) and surface, more "local" structure referred to as "micro-structure".

One essentially relevant phenomenon is the native speakers' ability to concatenate clauses and sentences into pairs, triples,...n-tuples, as a strategy for creating texture and building up the text. The concatenation is not haphazard; and the native users are capable of recognising textually faulty connection. The type and pattern of connective used and the semantic relationships that enter in the connection can operate as a control that ensures meaningful concatenation.

Another related problem is the native speaker's ability to assign well-formed semantic representation (i.e. grammatical interpretation) to sentences and constituents that are "often very different from the corresponding grammatical units as they are described in terms of the .... grammatical system" (Quirk et al. 1985, p. 1423). Elliptical constructions fall within this category of units. Here the semantic representation of preceding and following sentences usually provide the elements necessary for a possible interpretation. Moreover, interpretation is further determined by the systematic (pragmatic, referential) and unsystematic (ad hoc) or extralinguistic (psychological, social, historical) factors of the context or of the communication process.

One other argument advanced for the insufficiency of sentence grammars is the provision of criteria that determine the properties, and lead to a formal account, of the typology of texts: newspaper leaders and commentaries, news texts, literary texts, advertisements, daily conversation, public addresses, technical reports, business letters, etc. Nor do sentence grammars provide guidelines for distinguishing narrative from argumentative or descriptive texts.

These arguments and the need to investigate phenomena that stretch across the sentence boundary have led scholars to set up textual models of linguistic analysis. This trend gained wide recognition in the early seventies, particularly in Europe, where linguists, unlike their American counterparts, Who were preoccupied with isolated, invented sentences, advocated the study of language use, and encouraged the utilisation of textual discourse as material for conducting their projects.

2.1.3 Categories of Text Models

The accumulation of research that followed and the rich palette of enquiry that has resulted have identified a diversity of trends and models in text linguistics (cf, for instance, the papers in van Dijk and Petöfi 1977, Dressler 1978, Petöfi 1979, Petöfi 1982, van Dijk 1985 and the discussions in Enkvist 1984 and Beaugrande 1985). Such are the intensity and diversity of contribution that text linguistics is becoming recently an overall designation for any linguistic exploration of the text.

In his paper (1984), Enkvist identifies a number of text models in the the current state of the art. If we exclude interactional text models which constitute a major thrust within "discourse analysis" rather than "text linguistics", we can then recognise three broad categories of text models. This, admittedly, is not an exhaustive categorisation, as it does not pretend to include the full bibliographical data. The demarcation of categories is fuzzy as overlap is inevitable.

The first category that Enkvist recognises includes models that view text as strings of sentences which are given as input for analysis and description. The primary aim is a description and explanation of such textual phenomena as co-reference, connection (via conjunctive-type elements), thematic progression and overall theme-dynamic patterns. These models are capable of describing cohesion in text and are thus of direct interest to our work. A survey of some of these models are offered in the next two chapters.

The second category of models depart from the assumption that a text is composed of a set of predications and interpredicational semantic relations. The aim is to exhibit how these predications can be textualised through a process of grouping which involves conjunction and embedding. Such operations are monitored by a text strategy and accordingly different strategies will result in different textualisations of the same input predications. A representative model in this category is van Dijk's work on coherence (1977a).

The third category of models, which Enkvist labels "cognitive text models", start out with a body of experience and knowledge from which predications can be drawn. Modelling this predication-producing process is based on associative networks (cf. Kintsch and van Dijk 1978, Findler 1979). The concepts are themselves placed in the nodes of the network, their relations appearing as paths between the nodes. A text strategy constitutes a set of points of entry into, and paths chosen through, the network.

2.1.4 Concluding Remarks

The above discussion is intended to justify the primacy of text linguistic theorisation for accounting for the phenomenon of text connectivity which we are interested in. Other text-based approaches have exhibited limitations that can affect the angle through which we perceive textuality in general. The tools that text linguistic studies have utilised, we would like to maintain, are capable of efficiently probing into the question of sequential connectivity, identifying its textual patterns and progression. This issue will picked up in the next Chapter.

We would also like to maintain that the selection of a theoretical model for describing and explaining a phenomenon exercises some influence on the means and procedures through which interlingual analysis is carried out. For instance, the choice of structural or operational modelling of language as communicative vehicle can direct the analytic apparatus in two different paths, each leading to a description of different aspects of the same phenomenon. In a situation where the analysis is expected to assist in formulating guidelines for practical applications, the interrelation between a model and contrastive analysis becomes a crucial issue. We would like, therefore, within the next sections to examine the nature, scope and tasks of contrastive analysis in the hope of arriving at a·clearer delimitation of the typology of procedures we intend to employ in this project.

2.2 Contrastive Analysis

2.2.1 Preliminaries

The literature on contrastive analysis (CA) is enormously extensive and covers a wide range of activities. The motivation for the surge of linguistic research on CA is the need to discover and systematise the differences as well as the correspondences that a pair or more of languages exhibit when compared with one another, and the conditions under which such differences are made obvious. For instance, a study may expose the manner in which English and Arabic, or two variants of English and Arabic, or two stages in the development of either, differ with respect to passivisation, word order, morphological structure, etc.

However, the attitude of linguists toward CA has during the last two decades ranged from hostility (for instance in the review of Halliday et al. 1964 in English Language Teaching Journal 1966 p. 76) via passive lip-service (cf. Lee 1968, Richards 1971, and also Di Pietro 1971), to the re-creation of interest in the field (Hartmann 1980, James 1980, and the papers in Hartmann 1977, and Fisiak 1980, 1981, 1984).

Since our project involves fundamentally a contrastive analysis component in which properties of connectives in English and Arabic texts are investigated and compared, an introductory consideration of CA methodology is deemed sufficiently appropriate to be made within the context of this Chapter. The aims of the next few sections are: a) the provision of a general introductory account of CA as a discipline within linguistics, b) a consideration of approaches to Textual Contrastive Analysis (TCA) and their relevance to this work, c) the establishment of the requirements of TCA for the present study and the inclusion of a set of propositions on the manner in which such requirements are to be methodologically satisfied.

2.2.2. CA in Linguistics

2.2.2.1 Comparing Languages

CA can be considered one of the oldest preoccupations of modern linguistics. Its roots can be traced in the diachronic comparative linguistic studies carried out, especially in Germany, in the first half of the nineteenth century, most notably by the Germans F. Bopp, the brothers A. W. and F. Schlegel, J. L. K. Grimm, A.

Schleicher and the Dane R. Rask (Ducrot and Todorov 1979 p. 9). Those studies are characterised by the interest in establishing correspondences between Indo-European languages and are devoted to the discovery of not only resemblances but also kinships among them.

Within the modern synchronic perspective of linguistic studies, approaches to the comparison of languages can be classified in two groupings (Figure 2.1). The first group represent studies that have been termed "areal and typological linguistics", which are "concerned with establishing common patterns due to geographical proximity of the respective speech communities and the classification of language groups according to their structural characteristics" (Hartmann 1980 p. 24).

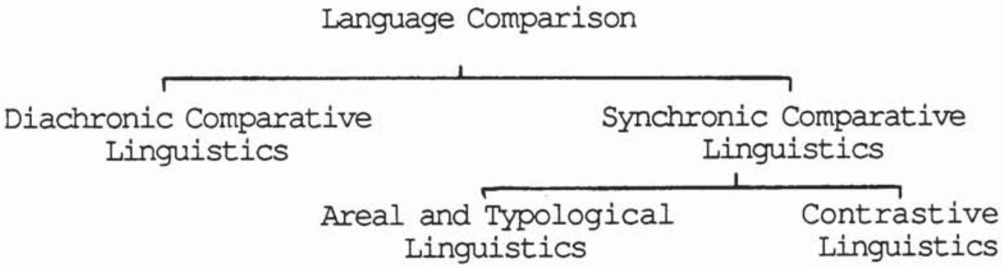
Language Comparison
Diachronic Comparative Linguistics
Synchronic Comparative Linguistics
Areal and Typological Linguistics
Contrastive Linguistics

Fig. 2.1 Classification of Approaches to Language Comparison

The second group is one that has received most attention since the forties and is usually labelled Contrastive Linguistics. This will be the focus of treatment in the next section.

2.2.2.2 Contrastive Linguistics

This term covers most of the activities where two or more languages are compared for similarities or differences. It was adopted by Whorf (1941) (quoted in Fisiak 1981) and the approach to these studies is labelled Contrastive Analysis (Hartmann 1980).

Although the two terms are well familiar in linguistic research, a number of linguists prefer less universal terms, such as "confrontative analysis" (Mrazovic 1974), "linguistic confrontation" (Akhmanova and Melencuk 1977) and "comparative descriptive linguistics" (Ellis 1960).

2.2.3 Types of CA Studies

A close examination of contrastive studies available can distinguish two types: theoretical (or "analytical", "descriptive" or "confrontative") and practical (or "applied", "didactic" or "differential") (cf. Sharwood Smith 1974, Hartmann 1977, Fisiak 1981). These two types are closely related; however, they differ in the formulation of their aims and in the general methodological perspective through which the procedures of comparison are to be carried out.

The theoretical type of contrastive studies aims at providing an exhaustive account of the correspondences and diversities between two or more languages, determining how and which elements are comparable and working out a model for their comparison. The methodological process of the comparison is bi-directional (Figure 2.2, cf. Fisiak 1981). A category X is examined for means and manner of realization in both languages A and B. The theoretical conclusions that a systematic description brings together can assist

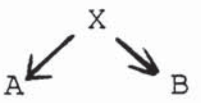

Fig. 2.2 A Bi-directional Procedure for CA

in forming a better understanding and achieving a deeper insight into the workings of the two languages, thus directly contributing to linguistics in general.

Practical contrastive studies draw upon the findings of theoretical contrastive studies. Their most immediate aim is the provision of a framework for the comparison of languages, selecting and organising the conclusions to suit and serve a specific purpose, e.g. teaching, translation, etc. The process of investigation is unidirectional, starting with the first member of the pair and moving towards the other. Hence exists the use of such terms as L1 vs. L2 (in bilingual studies), source vs. target languages (in translation), native vs. foreign languages (as well as the previous two pairs of terms) for foreign language padagogy.

$$\begin{tikzcd}
 & X \arrow[dl] & \\
A \arrow[rr, dashed] & & B
\end{tikzcd}$$

Fig. 2.3 A Unidirectional Procedure for CA

2.2.4 CA and FL Pedagogy

Contrastive studies have been recognised as a vital part of the foreign language teaching operation. Their effectiveness lies in the efficiency of the design of syllabi and teaching materials and in the appropriateness of selection and adoption of methodology and classroom techniques. On this point Fries (1945 p.9) argues that "the most efficient materials are those that are based upon a scientific description of the language to be learned, carefully compared with a parallel description of the native language of the learner". Similar arguments aroused great interest in the area of

CA and numerous studies and projects started to appear.

The basic assumption underlying these studies as pronounced by Lado (1957 p.2) is "that the student who comes in contact with a foreign language will find some features of it quite easy and others extremely difficult. Those elements that are similar to his native language will be simple for him, and those elements that are different will be difficult".

This view is supported by a number of linguists and has been widely accepted in CA investigations. For instance, Ferguson in his introduction to Moulton (1962) specifies that "a careful contrastive analysis of two languages offers an excellent basis for the preparation of instructional materials". This is also echoed in a number of FL textbooks and teachers' guides(1).

The advocacy of the relevance of CA to FL pedagogy has encouraged the emergence of practical orientations in CA. The underlying aim of the hosts of studies is the discovery, prediction and systematisation of the learning difficulty and the identification of the learning burden faced by FL learners. This is achieved by comparing the various grammatical aspects of the two languages (at the phonological, morphological, syntactic and semantic levels of description). These practical orientations and the surge of interest towards application particularly in the wake of Weinreich's (1953) and Lado's (1957) work, not only obscured the theoretical objectives of CA, but also included within its domain a number of studies that were not essentially pedagogical (cf. Fisiak 1981). As a result, criticisms were raised and some linguists and

FL pedagogy practitioners expressed their doubts about the utility of practical CA orientation.

CA was then subjected to critical scrutiny. One of the most debated issues revolves around the relevance of CA to FL pedagogy and the doubts that have been cast have brought forward mixed reactions. In one extreme there are opponents of CA who question the role of CA as a valid foundation for predicting errors and for the design of FL teaching material (cf. Wardaugh 1970). In the other extreme there are the proponents of CA who strongly back the principle of implementing the findings of contrastive investigation in practical areas (cf. James 1969, 1971, 1979, 1980; Fisiak 1980, 1981, 1984 and the papers therein).

In between the two polar extremes of pros and cons there stands a moderate group who is willing to accommodate the basic ideas of CA and incorporate them in resarch design and application (cf. Di Pietro 1971, Nickel 1971 Sanders 1981).

The problem of the predictive power of CA is outlined next.

2.2.5 The Predictive Power of CA

As stated earlier, the tasks of CA is the synchronic contrast of two language aspects, carried out in such a way that similarities and dissimilarities can be revealed. On the basis of the findings, a prediction of probable errors produced by learners can be made. Accordingly, teachers, textbook writers, and FL material designers as well as examiners are able in advance to prepare themselves as to the kinds of errors a student of a particular language background is likely to produce and the linguistic difficulties he is prone to experience. A set of possible solutions can then be proposed to avoid or minimise the errors and to facilitate as efficiently as possible FL acquisition. These suggestions are then verified by subjecting them to classroom trials. This is essential since there is evidence that while one student substitutes one form for a particular element, another of the same language background will substitute a different form for the same element (Dardjowidjojo 1974 p. 47). The argument that classroom verification is an integral part of practical CA is supported by Lado (1957) who explicitly states that a list of problems produced by CA remains hypothetical until "final validation is achieved by checking it against the actual speech of the students" (p. 72).

These arguments came under fire by the generativists and others. The following is a summary of their criticisms.

1. A serious charge against CA and one which is backed by Wardaugh (1970) is the nature of CA. Wardaugh describes many contrastive studies as "snippets of information" about two languages. The obvious response (cf. Di Pietro 1974) to such a criticism is that, if it were to be accepted at all, it could just as well be levelled at linguistics in general. Very few formal descriptions of single languages that linguists have so far provided can be characterised as somewhat complete or even extensive. "In fact there can be no such thing as a 'complete' account of the grammar of a language, because language is inexhaustible" (Halliday 1985a, p. xiii).

2. Another serious criticism concerns the function of CA. Wardaugh (1970) claims that the most that CA can do is the explanation of errors produced by a learner but not their prediction. Related to this is the argument that the making of an error in FL learning can depend on factors extraneous to linguistic structure. Hence, it is dubious whether a practically satisfactory extent of predictability is to be attained until these factors are also investigated and their adversity to FL learning process is properly uncovered and accurately assessed. This claim seems to revolve around two points: a) the procedures through which CA is undertaken (ie a methodological consideration), and b) the practical and pedagogical goal that CA aims to attain.

These claims, as well as a number of others, has been strongly refuted (cf. James 1972, Sharwood Smith 1981). First, advocates of the validity of CA never claimed a one hundred percent predictability of errors, nor did they maintain total accuracy of prediction. In addition the adverse effects of factors extraneous to linguistics are already recognised. On this point Carroll (1971 p. 113) states "the teacher's ability to manage learning behaviour remains one of the most unexplored, unstudied variables in educational research". Nickel (1971) adds some psychological parameters which have relevance to the learner. One is the learner's greater difficulty to "encode" than "decode" in a foreign language. Another is the possibility of interference resulting from other languages being studied or already learned. In addition, CA predictions, as stated earlier, require classroom validation in some empirical manner, otherwise they will remain totally conjectural.

Furthermore, the need for CA based FL pedagogy is justified on the grounds that more efficient methodology and teaching and testing material can be devised. Most FL teachers and learners lack the ability of creating a linguistic environment similar to the one in which a child acquires his native language. There are numerous restrictions in the time available, limitation of aims, variation in learner's motivations, unrealistic classroom situations and the fact that the target language is not the the language of the community outside the classroom. To be able to explain errors, a teacher or a learner is required to possess a special linguistic training which enables him to locate errors related to native language interference and distinguish them from those that are peculiar to such factors as overgeneralisation (cf. Richards 1971) or insufficient classroom practice. Indeed in many parts of the world, even in developed countries, not all language teachers have the privilege of knowing linguistics to make their own explanations. A reference to CA findings is therefore a valid procedure.

It follows that if practical CA can assist in reducing the time and effort of learning and teaching a language, it is already a major contribution which warrants the practice of practical CA investigations. This is feasible through systematic descriptions, at various levels, of both the source and target languages. Such descriptions, when combined, can offer a total image of the workings of the respective linguistic systems. In other words a theoretical account of the correspondences and diversities of the languages precede applicational considerations and empirical validation.

What emerges from these arguments is the view that despite claims to diminish the significance of both theoretical and practical CA, and despite the fact that CA studies may have been too limited in scope and in practical implications, the field of contrastive linguistics is nevertheless still theoretically justified. Most claims of CA opponents centre on studies that are possibly misguided in specifics or whose practical implications lack validation. The value of such claims, however, is that they are indicative of possible procedural flaws that have to be avoided in order to produce a sound piece of contrastive investigation.

2.2.6 Some Proposals for CA

In this section a number of general proposals are made concerning the tasks of CA. It is felt that the incorporation of these tasks among those already advocated can promote the theoretical and applicational value of CA studies.

One significant development that CA can undergo which extends its theoretical activities and add a further dimension, and a vital one, to its practical application is its inclusion, among the different levels of analysis and description, of a textual (including discoursal, stylistic and rhetorical) level. Most contrastive studies have handled linguistic aspects at the sentence level, and therefore the whole discipline of contrastive linguistics has suffered some of the shortcomings attributed to sentence grammar. Among those is its inability to account for such problems as inter-sentential variations exhibited by different languages including means of connectivity, paragraph organisation, and means available in any language for realising text typologies.

A relevant expansion that CA can manipulate effectively, both theoretically and practically, and one that most explicitly operates on a textual level, involves the addition of sociocultural elements. Since language is a means of communication which occurs in specific contexts, which, in turn, are socially constrained, it follows that a systematic study that reveals diversities in the social contexts in the native versus the foreign cultures is a significant indicator that offers proper guidance for those engaged in the FL teaching operation, at the various levels of its hierarchy. Unfortunately, neither anthropologists nor ethnographers have been able to bring together a body of contrastive cultural information derived from systematic analyses of the native and the target cultures, which can be regarded sufficiently informative and coherent to provide guidelines in FL instruction. One only hopes that in order to remedy this deficiency studies on cognitive anthropology can systematise cultural data, thus offering explanations of cultural biases that can be of benefit not only to pedagogy but also to translation and bilingualism theories, and in bilingual or multilingual lexicography.

Further elaboration of textual CA, its scope and tasks, is offered in the next sections. It only remains here to conclude by stating that CA, despite the criticisms it received, is still both theoretically and practically valid, and that a demand for expansion to include higher levels of analysis is mandatory for a better understanding of language and a more efficient manipulation of findings.

2.3 Textual Contrastive Analysis (TCA)

2.3.1 Preliminaries

A major criticism launched against CA (cf. James 1971), and one, we believe, that is bound to have considerable consequence on contrastive linguistics, whether in aim, essence, scope or methodology, is concerned with the ways that our view of language may or may not be adequate for descriptive typological comparison. Pre-textual linguistic methods imposed a total reliance on contrastive analysis of sentence-level linguistic elements, which in effect characterised the analyses with the same deficiencies that sentence grammar suffered from (cf. 2.1 above).

Accordingly, CA studies, both theoretical and practical, can be revitalised by reshaping their framework to include a textual-level of analysis as its major component. It was Gleason (1968, see also 2.3.3 below) who had tentatively suggested that CA at the textual level, rather than that of the word or sentence, might be a better framework for focusing "on what may well prove to be the most interesting of all contrastive problems, the differences in the way connected discourse is organized and the way that organization is signalled to the hearer or reader" (1968 p. 58).

However, Gleason's tenets were not substantiated at that time, and seemed to receive little attention. The reasons that we can offer for this neglect can be summarised as follows:

a. Text linguistics and discourse analysis studies were then just emerging (cf. Rieser 1978, Beaugrande 1980) and, therefore, there did not exist an appropriately sufficient theoretical development on which textual contrastive models could be erected. Dressler's famous "Einführung" appeared in 1972 and before this publication the textual attempts were characterised by drastically insufficient maturity.

b. The general preoccupation of linguistics at that time lay in the generative theories, and in particular transformational-generative grammar. A number of contrastive linguists attempted to establish a generative approach to CA, alleging that the insufficiency of CA tenets is attributed to the structural framework within which CA had operated. This preoccupation, as a result, minimised the possibilities of serious consideration of TCA, and, accordingly, most of Gleason's stimulating arguments received hardly any notice.

The aim of the next few sections is twofold:

a) an outline of the most prominent studies in TCA and their theoretical contribution to a TCA theory, and,

b) a specification, based on this and previous section (ie 2.2), of the most immediate requirements of TCA as related to the investigatory apparatus of our project. Having established these requirements, a note is included on the applicational nature of the study.

2.3.2 Some Early Attempts at TCA

TCA is not an entirely new phenomenon in text-based research. Most studies that can be considered forerunners operated within stylistic and rhetorical framework. Vinay and Darbelnet's "comparative stylistics" (reported in Kachru and Stahlke 1972, Hartmann 1977 and 1980, and Wilss 1982) can be considered as the first viable brand of TCA and, as Hartmann 1980 p. 27 describes it, the most original attempt then to give discourse its proper place in language comparison.

Vinay and Darbelnet's point of departure was not the global comparison of language structure, but rather "the situationally equivalent text". To them a target language message may be considered equivalent to the source language text on two accounts: a) it has the same "meaning", and b) the situations to which the message relates are identical. Thus, by taking situational appropriateness as a common denominator of contrasting source and target language texts, Vinay and Darbalet attempt giving "the practising translator and foreign language learner a method for producing target-language versions which would be stylistically appropriate in corresponding context of situations" (Hartmann 1980 p. 27).

The basic tenet assumes the association of conventionalised styles with different communicative situations. Accordingly, the translation of a text requires the creation of a "situationally equivalent" counterpart in the target language, a view that was considered revolutionary when it was introduced, though it came to be deprecated through the engagement of linguists of that period of the arguments against and in favour of behaviourist structuralism (initiated by Chomsky's (1959) now classical review of Skinner's views).

These early proposals, with their textual contrastive implications, make contribution not only to stylistics but also to translation theory. "Translation theory" is a blanket term that covers the body of knowledge available about the process of translating (cf. Newmark 1981 p. 19). Translation itself is "an operation performed on languages: a process of substituting a text in one language for a text in another" (Catford 1965 p.1). It therefore follows that most of the activities performed in this area reflect, directly or indirectly, aspects of TCA.

Another area where proposals encouraged a TCA attitude is content analysis. An instance of research in this area is Edelman's investigation of political discourse (1964), in which he distinguishes at least four distinct styles(2). Edelman proposes that "a similar analysis of other cultures would no doubt bring to light different typologies of language forms, with different persisting meaning" (p. 151).

Taken in their totality, these early contributions to TCA, despite their value in their peculiar areas of investigation, make no coherent formulation of an approach or a framework for reference. Such formulation was put forward within Gleason's proposals for a contrastive discourse analysis and Kaplan's studies in contrastive rhetoric. Their arguments and conclusions are discussed next.

2.3.3 Gleason's Contrastive Discourse Structure

Gleason's (1968) argument departs from a number of basic assumptions:

a. An "acceptable" piece of discourse differs fundamentally from a randomly selected series of sentences.

b. The phenomena to be accounted for in contrastive linguistics arise most noticeably in the course of careful translation, and some of the problematic issues concern the attainment of connectivity between successive sentences while conveying the intended message, that is, the achievement of "proper" discourse structure.

Gleason's premises are restricted to a single form of discourse, narrative, justified on the account that more comprehensive coverage would complicate the statement. However, this restriction, he maintains, scarcely alters the basic principles.

According to Gleason, language provides at the minimum some guidance in mapping the typical stream of narrative activity into an articulated sequence. Some features are common to all or many languages, while others are peculiar to one or two. The main task of a full language description is to cover all the aspects which are linguistically controlled, while the main task of contrastive analysis (at least at the theoretical level) is to indicate which features are unique and which shared.

Gleason proposes a model within the framework of stratificational grammar in which he postulates a class of linguistic units which he labels "Action". Another class is called "Connections". The "tactics" provide for the arrangement of these units in long chains, which Gleason calls "Event-line", generally with these two classes alternating.

Within this conception, languages differ in the way the following three kinds of linguistic apparatus figure in the skeletal structure of the narratives.

a) Differences in the organisation of the Event-line, i.e. in the inventory of semologic units or in the tactics controlling their arrangements.

b) Differences in the grammatical organisation of the sentences, i.e. in the lexical units employed, or in the tactic patterns.

c) Differences in the way semologic Event-lines are realised in grammatical sentences, i.e. in the complex mapping relations between the two strata.

A second sector of discourse structure that Gleason describes is the "Participants". These are semologic constituents of narratives related to some or all of the Actions by semologic Roles. Instances of such Roles are: agent, goal, beneficiary, affected, causer, that have to be distinguished from grammatical functions within clauses, such as subject, direct object or indirect object.

A single participant may be related to several Actions. Furthermore, Actions differ as to how many and what Roles they permit or require. All these must be specified for the use of the semologic tactics, taking particular notice of the complexity of realisational relations of these Roles[3].

The identification of Participants, Gleason notes, cannot be performed by sentence grammar. Such a task requires analysis operating over stretches much longer than a single sentence. Furthermore, patterns of participant identification across languages can exhibit profound interlingual differences. One difficulty in translation which is an outcome of these differences is the requirement of a total restructuring of the system of Participant identification.

Gleason's proposals, despite the intricacy and richness of their implications, suffer from some shortcomings. First, the model as posited is not complete and needs verification through sufficient textual analysis. What is presented is based on superficial examination and random observation of text portions; one of them is even contrived for illustrative purposes. Additionally, the interaction between the organisation of the Event-line and the identification of the Participants demands clarification. Furthermore, the model neglects the role and nature of the Connection posited between Actions. We are left in doubt as to how Connections are realised and what contribution they can make to the stock of known characteristics of discourse in general and of the Actions they operate on in particular.

But despite these comments, the model posited and the arguments advanced to support it have the merit of being a significant contribution to textual comparison and an attempt to restore the viable operationality that CA so badly required, at a time when CA was beginning to lose some of its credibility both in theory and in practice.

2.3.4 Kaplan's Contrastive Rhetoric

Most of the views reflected in Kaplan's works (for instance, Kaplan 1966, 1967, 1968) are included, with some necessary modification, in Kaplan (1972). This modification is necessitated by the accumulation of information in theoretical linguistics during that period. Our review of Kaplan's views will therefore be based on examination of this work, though we will still refer to some of his earlier studies, particularly his 1966 treatise.

Kaplan's premise departs from the assumption that "the organisation of a paragraph, written in any language by any individual who is not a native speaker of that language, will carry the dominant imprint of that individual's culturally-coded orientation to the phenomenological world in which he lives and which he is bound to interpret largely through the avenues available to him in his native language" (1972, p. 1, author's underlining). This view, is to some extent, a resonance of the old "Whorf-Sapir hypothesis" that language predetermines certain modes of observation and interpretation for its speakers.

Another assumption that Kaplan posits concerns the status of the basic unit of analysis. According to his theory, Aristotle's contention that discourse is a stream of words and that, therefore, the word is its basic unit is a fallacy. So is Bloomfield's assumption that the sentence is the basic unit of syntax. What has to be recognised is that within a discourse a unit may exist, and operate, containing within itself a group of "independent" but subordinated units. Christensen (1967) as well as other rhetoricians of the "New" persuasion (see Appendix 2) have demonstrated that the paragraph, for instance, contains a number of quite varied units not essentially related to the concept of the sentence.

A third assumption concerns the failure of rhetoricians in their engagement with taxonomy to grasp the essential view that communication is not static. Accordingly, what is important is not the taxonomical labelling of its parts, but their functional interrelationship.

A fourth assumption is that logic (not in the strictest philosophical sense) is the basis of rhetoric and that it is evolved out of a culture (cf. Dufrenne 1963 pp.35-37), i.e. it is not universal. Consequently, rhetoric is not universal either, but varies from one culture to another. This is supported by the view that every language offers its speakers a ready-made interpretation of the world (Spitzer 1953 pp. 83-84). It follows that the expected sequence of thought in the English language (which is essentially Platonic-Aristotelian, shaped by the Roman, Medieval European and later Western thinkers) is different, say, from that of Arabic, a Semitic language with a completely different cultural background.

A final assumption is the recognition that a paragraph is employed in writing to suggest a cohesion which may usually not exist in speaking. It is an artificial unit of thought that lends itself to patterning quite readily. An English expository paragraph reflects the thought patterns that the English readers appear to expect as an integral part of their communication. The development of thought patterning can be inductive, i.e. the paragraph begins with a topic statement and then by a series of subdivisions of that topic statement, each supported by illustrative exemplification, proceeds to develop the central idea and relate it to all other ideas in the essay. Alternatively, the patterning can represent deductive reasoning, when a series of examples and illustrations are first stated to be related later into a single statement at the end of the paragraph.

To verify these assumptions, Kaplan carried out some experimental work involving a large number of non-English speakers of various language background groupings, which he later divides into Semitic, Oriental, Slavonic. For the purpose of the experiment, Kaplan asked the participants to write down passages in English. He later set out to examine these passages to find out where rhetorical patterns diverge from those which are peculiar to the English language.

Through analysis of the data Kaplan comes out with a number of specifications of the rhetorical patterns, especially within paragraphs. The "movement" of the paragraph, he argues, differs in different languages, and this is schematised in the representation reproduced in Figure (2.4) below.

Kaplan concludes his arguments with two cautious remarks: a) the rhetorical patterns and categories discussed are in no sense meant to be mutually exclusive, and b) much more detailed and more accurate descriptions are required before any meaningful contrastive system can be elaborated.

Fig. 2.4 Graphic representation of the movement of paragraphs in Kaplan's data (Kaplan, 1966 p. 15, 1972 p. 64)

The research design of Kaplan's empirical validation of his hypothesis is not without flaws. It is felt that the identification of these can serve the provision of a better mechanism of investigation, particularly in a project of a contrastive nature such as ours.

One comment concerns the type of the data gathered. Kaplan's method of having non-English speakers of various languages write expository passages in English hinges upon the expectation that, say, native speakers of Japanese would produce English passages that are different in organisation from those produced by Arabic speakers. Each reflects a rhetoric organisation peculiar to his culture and therefore have a different shaping of reality. This type of data suffers from some weaknesses that are projected upon the findings.

a. The norm that is used for measuring differences is English where the paragraph development is represented with a straight line, a view that is not shared by researchers engaged in contrastive rhetoric. Hinds (1983), for instance, reports a study conducted by Cheng (1982) where a different representation is effectively argued for the English expository paragraph.

b. Kaplan's data, all passages in English, may or may not exhibit the rhetoric organisation peculiar to a native language. One can argue that the organisation detected in the passages is an idiosyncrasy of a stage of development of an "intralanguage" rhetorical organisation and that, when sufficient knowledge is attained of the participants' previous foreign language training, a sketch can be made of this developmental stage in the acquisition of the FL textual grammar and rhetorical organisation.

c. Kaplan's reasoning, at least as far as Arabic is concerned, is notoriously deductive. He constructs his arguments on two assumptions: i) parallelism is the general rhetorical pattern of expository Arabic, ii) this pattern is culturally acquired since Arabic is deeply influenced by the Koran. He then moves to locate parallelistic forms in his sample of passages written English. When he finds any, he refers them to the influence of the mother tongue rhetorical organisation; but he does not attempt to explain the existence of patterns that are not parallelistic in form.

d. Kaplan's categorisation of languages is statistically as well as linguistically most dubious. The term "Semitic" is used to indicate Arabic and Hebrew. The size of the group of Semitic speaking participants is 129, of whom only 3 are Hebrew speaking. We feel that any findings derived from the data of this group, if applicable of Arabic at all, cannot adequately apply to Hebrew, unless, of course, there already exists a preconception of the rhetorical organisation of Hebrew derived from sources other than the data. Since the data is not representative of Hebrew, we think that Kaplan's category of "Semitic" languages refers to Arabic, and that his characterisation of the rhetorical patterns and schematisation of the movement of the Semitic paragraph is applicable to Arabic. Related to this point is the erroneous use of the category of "Oriental" languages (See the discussion in Hinds 1983 pp. 186-187).

These flaws, however, should not demerit Kaplan's work. As a contribution to TCA it stands as one of the early and most distinguished attempts. As a contribution to contrastive rhetoric it has stood out as the only source in the literature for almost two decades. It is only recently that some rhetorical studies with contrastive orientation, partly stimulated by Kaplan's work, have started to appear (for instance, Koch 1981, Williams 1982, Hinds 1983, Al-Jubouri 1984).

2.3.5 Hartmann's Model of Contrastive Textology

Hartmann's proposals are expounded in his 1980 volume, although hints are suggested in the introductory chapter of his work (1977). In setting the scene for the model, Hartmann departs from the basic premise that contrastive analysis without a text base is ineffective and incomplete. His proposal for "contrastive textology" is motivated by the desirability for "an adequate theoretical framework" (1980 p.34). The scheme, he admits, are extensions of existing rather than untried ones, which is legitimately warranted since, as mentioned earlier, very few precedents are available for what a contrastive textological model should look like.

Contrastive textology aims at combining both the contrastive dimension and a textological dimension in a unified approach. In order to set up a suitable model, and by analogy with the familiar levels of contrastive phonology, lexicology and grammar, Hartmann posits a supra-hierarchical level, subdivided by the three semiotic dimensions, which result in the components: a) text pragmatics (or communicative textology), b) text syntax (or combinatorial textology), and c) text semantics (or referential textology). These components are fused into one eclectic whole, directed towards the goal of accounting for the communicative potential of texts within and across languages.

The pragmatic component is concerned with the different ways in which the correlation between functional variety and discourse manifestations are handled. The aim is a "situational discourse typology", the kind pioneered in the genre classifications of rhetoric, dialectology, stylistics and register analysis.

The syntagmatic component, Hartmann maintains, handles the "texture" of text, how successive portions of discourse are strung together to form complete texts. The aim is a description of inter-sentence connectivity, the kind of "combinatorial textology" attempted in a number of theoretical and descriptive studies of cohesion and textual composition (such as Halliday and Hasan 1976, Gutwinsky 1976, and Werlich 1976, refer to the review in Chapter 3). Hartmann observes that "none of these [i.e. studies] are methodologically uniform, which makes their evaluation and adaptation to contrastive analysis difficult" (1980 p.36), which is true in so far as eclecticism in the set-up of the model is a fundamental requirement.

Finally, the semantic component in the model handles the different ways in which referential information is distributed among the consistent elements of a text. The aim is an explanatory account of the means of "information structure", the kind of "referential textology" suggested by the Prague School notion of functional sentence perspective.

In order to promote the effectiveness and validity of the results of contrastive textology, Hartmann develops the notion of "parallel texts". The rationale behind it is his view that "all interlinguistic contrasts are manifest in texts" (ibid p.37), a basic fact that practical activities such as FL pedagogy, translation and bi-lingual lexicography have always been aware of, though it took some time for contrastive linguists to realise and appreciate that, for instance, "textual equivalence is itself one way of establishing comparability" (Halliday et al. 1964 p.123).

The notion of "parallel texts" as a procedure for comparing translationally equivalent texts has been used by a number of linguists and institutions either as a technique in translator training or as a procedure to arrive at contrastive description (for

English-Arabic contrastive analysis using the procedure to arrive at some problematic aspects of translation, see El-Sheik 1978 and Shamaa 1977).

The procedure of parallel texts, as Hartmann develops it, dictates "that we should a) incorporate what we know about phonology, lexicology, grammar and textology within a discourse framework and b) combine the conceptual-logical, critical-exegetical, correlational-sociological, and experimental-scientific approaches into an eclectic whole" (p.37). Accordingly, parallel texts can be used to attain interlingual comparisons at all levels and with any method.

In discussing the procedure, Hartmann, posits three major groups of parallel texts, though he cautiously observes that "refinements in the classification of parallel texts depend on progress in intra-linguistic discourse typology and inter-lingual equivalence criteria" (p.37). Class A of parallel texts include texts that are the result of a full-scale professional translation with the source text becoming a situationally appropriate target language text. Class B include texts that are typically the result of a deliberate adaptation of a message for the purpose of conveying an identical one in the target language. Class C of parallel texts are typically unrelated except by the investigator's recognition that they share a similar context, though created independently, as, for instance, when instances of a specific text-type (e.g. marriage columns in newspapers) are compared across pairs of languages.

The advantage of parallel texts, in Hartmann's words, (cf. also

2.4 below) "lies in the fact that they document contrasts between discourse types within and across languages and thus confirm the evidence we possess from studies in contrastive lexicology ... and bilingual interference ... for the existence of relatively separate language varieties" (p.39). This view, we believe, remains legitimate as long as the size of the sample examined is adequately representative, a consideration that is left out by Hartmann. Furthermore, it is essential to provide a more detailed characterisation of the proposed classes of texts, a task that he obviously leaves to further research in the area.

2.4 TCA in the Project

2.4.1 Preliminaries

It is generally recognised (cf. for instance, Hamp 1968, Marton 1974) that contrastive analysis is a legitimate branch of theoretical linguistics, irrespective of its pedagogical uses and implications. It follows that the notions and features that TCA employs vary greatly depending on what linguistic theory one espouses. The studies that have been reviewed (2.2 above) confirm the existence of a level of analysis for investigating and contrasting interlingual features that are higher than the sentence. While Gleason examines these features from a discoursal perspective, suggesting a macro-structural design for narrativity, Kaplan performs his comparison from a rhetorical perspective, examining such features as organisation and unity in text. In this section we would like to outline the nature and dimension of the TCA component of this study, drawing largely on insights from these reviewed studies. The discussion will examine the dimensions, requirements and procedures that are to be adopted. These issues pinpoint the operational direction of the investigatory apparatus of this study.

2.4.2 Nature of Comparison

For the purposes of this project the textual contrastive analysis is defined as systematic comparison of selected aspects of connectives in English and Arabic texts, the intent of which is to reveal the typology and extent of variations in the patterning of connectives. The ultimate goal of the comparison is , as stated earlier (cf. Chapter 1), applicational: to provide teachers and textbook writers with a body of information and some general as well as specific guidelines which can be of service in the preparation of instructional material, the planning of courses and the development of classroom techniques in teaching EFL written text production.

The analysis is here labelled "textual" because it operates on a level beyond the sentence and can therefore not be resolved by resorting to sentence grammar. Further justification for the adoption of the tools of text theory in the analysis has already been discussed. It suffices here to mention that comparison of the phenomenon of connectives is better formulated if described in terms of textual relations that obtain in a coherent text. It is interesting to notice that even if one does not intend from the beginning to postulate a textual level of analysis describable in terms of text theory, the contrastive investigation proper, if conducted consistently, makes it necessary to approach the phenomenon from a textual vantage point. Such a point presupposes taking into consideration means of expressing connectivity, a concept that cannot adequately be described at a sentential level by use of the current theories of sentence grammar. For connectivity, whether sequential or conceptual, cannot be contrasted across languages by resorting to theories of the complex or embedded sentence. If these theories are capable of providing tools for confronting certain aspects of intrasentential connectivity, it will fail to account for intersentential relations.

2.4.3 Dimensions of Comparison

In attempting to set the dimensions for the comparison in this work we depart from the premise that textuality in a particular language imposes at the minimum certain requirements whereby the sequences in a text are identified and related and whereby texture is constructed and maintained. Some requirements built on the existence of certain textual phenomena (such as reference or connection) are common to many, if not all languages, while others are peculiar to one or two. The main task of a full description of a phenomenon is to cover all the aspects which are linguistically controlled, while the main task of textual contrastive analysis (at least on the theoretical level) is to indicate which aspects are unique and which shared.

The dimensions that characterise the TCA we propose to carry out in this project reflect a number of theoretical assumptions. These concern certain issues that will be handled in more detail in the next chapters. We discuss them here in so much as they concern the tasks aimed at in the analysis. In general we propose the following as theoretical dimensions of the variations to be explicated through the analysis:

1. Languages differ in the way text is created from constituents. When such constituents are concatenated sequentially, some languages require more specifications in certain positions than others do. The differences exist in the minima allowed. A clear-cut and well-defined upper limit is extremely difficult to envisage.

2. Two languages may offer different repertoires of possibilities for locating text out of constituents. This entails an exercise of choice and preference. For instance, when two languages allow the same alternative realisations, their users may display different preferences. The briefest permissible possibility is not always the preferred one. The question of preference is governed by the extent that a strategem can achieve rhetorical effect and textual unity.

3. In text creation, languages differ in the manner and size of distribution of their cohesive ties, particularly connectives. One language may manifest a bigger or smaller density or diversity of ties than do others. The distribution is here assumed to constitute the dominant imprints of the textual orientation of a particular language. Whether this is culturally determined (cf. Whorf-Sapir's hypothesis, Kaplan's claims in his 1966 treatise and the views of Koch 1981) is left for further research that is based on sociocultural comparison, and is therefore ignored in this study.

2.4.4 Procedures of Comparison

The accomplishment of a CA exercise is made by exploiting a number of methods, each with certain procedures. Halliday et al.

(1964) identify two methods. The first is 'describe then compare', which has obvious procedures: one cannot achieve a comparison of how two languages function unless one describes first how each of them works. One favoured procedure (cf. Brooks 1962 p.155) is the identification and description of the full inventory of patterns discoverable in the code of a given language. This should result in an awareness not only of their totality, but also of the order of their frequency and of their internal relationships. This description is followed by a comparison of the two inventories of patterns, whereby similarities and differences are outlined.

The second method is 'compare patterns, not whole languages'. This is justified on the assumption that one can no more compare such two genetically different languages as, for instance, English and Arabic than compare cheese with chalk. Each language consists of a complex of a large number of patterns at various levels and at different degree of delicacy.(4) The procedures involved here depend to a large extent on the nature and level of the pattern being investigated. Further, they are modified by whether any application is sought, and if so, to what end (e.g. FL pedagogy, translation, etc.).

If our TCA project is to adopt the first method, it will then have to initiate a comprehensive comparison of all features of textuality and design within the total typology of text across English and Arabic, a project that demands the combined efforts of a team of researchers, sufficient funds and adequate time. Such a project is feasible if sponsored by an institution or academy of language studies. Given the specific aims and limited scope of our study, we then have to adopt the second method but try to draw upon the validity of the first. In other words, we shall concentrate our efforts in a technical description of connectives as one type of cohesive tie in both English and Arabic. Once the description is completed, a contrastive account is initiated whereby interlingual variations and similarities are confronted and described. In general, since the study is applicational, intended to offer pedagogical implications, the procedures that are to be manipulated in the contrastive analysis are an amalgam drawn from a number of contrastive approaches.

These procedures are outlined below. As will be shown, partly here but mainly throughout the mainstream of the study, these procedures are under two influences that shape their direction. The first concerns the constraint imposed by the theoretical limitation of text grammar at the present state of the art. The second is the intent of this project to provide pedagogical implications. The procedures are as follows:

1. Relevant Definition: If there existed a general typology of textual constituency and relevant conceptualisation that are consistent and well delimited and therefore can be employed without necessary justification or contrived amendment, TCA would proceed simply by taking individual textual units or their combinations and contrasting their surface realisation in one language with that in another. In that case, the question of the nature of the units of analysis and of the all-important "tertium comparationis" would not arise. But, unfortunately, this is not the case, and in the absence

of clearly delimited textual units for its starting point, we find ourselves in our TCA compelled to attempt a definition of the basic units and conceptions relevant to the analysis and make that our point of departure (cf. Ch. 3 and 4).

2. Categorisation of Connectives: The establishment of categories of connectives, while considered here as a tool for their description and therefore directly contribute towards the realisation of the primary aims of the study (cf. Chapter 1), can itself be a procedure for comparison. This is, we believe, because comparison depends on description, hence a successful comparison is constrained by the quality of the underlying description. On this particular point, Halliday et al. (1964 p.118) states that "comparison resting on sound linguistic and phonetic theory is more powerful, for whatever purpose is envisaged for it, than ad hoc impressionistic comparisons lacking a descriptive foundation". Accordingly, comparison is preceded by an independent description of connectives first in English, then Arabic (see Chapters 9-10).

3. Criteria for Comparability: A mandatory prerequisite for the initiation of a comparison is the establishment of comparability, i.e. to ascertain that the phenomenon that is to be contrasted interlingually is in fact comparable. This calls for the setting up of a number of assumptions that can serve as categories that are applicable for both languages.

In so much as our project is concerned there are three possible categories for making comparison:

a.Equivalence of Nomenclature: This refers to the equivalence of terminology used for describing the phenomenon under investigation in each language. A word of caution is in order here. It is often the case that what a thing is called is not a dimension or attribute of that thing. This is particularly applicable to a linguistic term employed in two languages as genetically wide as English and Arabic. Hence, care should be exercised when comparing categories in different languages that have the same name. The term "connectives", to start with, is itself a source of problem, since the nearest Arabic equivalent "Hurufu Al-<atf" indicates a category of "particles" that is too limited in size and function, and hence a large number of items that can legitimately be termed "connectives" are left out (cf. Chapter 5 for more detail). Similarly, categorisation of "connectives" where such terms as "additive" or "causal" are used have to be constructed and defined in such a way that these categories are comparable in two languages. The question is then that of redefinition of nomenclature so that we avoid making the mistake of assuming that because the nomenclature is the same, the categories referred to must be the same.

b. Formal Equivalence: This refers to the ways of identifying the categories we need in order to undertake the comparison. The question rests on how we arrive at the category of "connectives" across English and Arabic. In this project we tackle this question by setting criteria for the textual function of the connective and the textual environment where it can be operational. If a particular item achieves the specified textual function of connectivity within the environment, then it is regarded as a connective in either language. The mode of functioning and the textual environment can then be subjected to comparison. It is clear that unless we employ a unified scheme that is descriptively adequate and theoretically systematic, we shall end up with incompatibility of categories and relations used in the description of each language. This has the consequence that each category will remain language-specific and hence, is not comparable.

c. 'Meaning' Equivalence: This refers to the capability of the phenomenon of connectives to demonstrate contextual equivalence. One way of ascertaining this is by reference to translation (cf. Halliday et al. 1964, also Vinay and Darbelnet 1958). If the categories to be compared, or its members, are not at least sometimes equivalent in translation, they cannot adequately serve as a basis for comparison. We agree that translation is a controversial topic, and, moreover, this is not the right place to discuss it; but unless we believe that by-and-large elements that create sequential connectivity, particularly the category called connectives in this project, can have translational and contextual equivalence in another language, we should scarcely be able to translate a text and, furthermore, we should hardly be interested in teaching a foreign language.

4. Comparison: Comparison is the core of the effort expended in any contrastive analysis. The goal is to make statements that bring out the type, nature and extent of similarities and variations in the realisation of the phenomena investigated across the two languages, which later can be manipulated for practical purposes. In this project comparison is achieved through the following:

a. Inventories: The categories and subcategories of connectives in English and Arabic are juxtaposed and variations in typology of functional relations are outlined.

b. Observation of textual environment: The textual environment for a category of connectives are examined interlingually. Range of functioning is then projected and described.

c. Statistical comparison: Results of the computation of the distribution of connectives, measurement of gap and growth rate as well as other statistical measurements, incorporated in what we have called the calculus of connectives, are compared and a comparative picture is developed.

Comparison starts with descriptive statements of the patterning of a particular category or subcategory of connectives in English and proceeds in the direction of Arabic, and back again towards English. This will ensure that we do indeed get a contrastive description rather than just two parallel descriptions of connectives in the two languages. Since English connectives are the ones the use of which is the focus of the pedagogical application, they will also be the ones in terms of which the contrastive statements are made.

To clarify this further, we would like to state that steps have been taken to systematise the comparisons and minimise arbitrariness in the contrastive treatment. First, as mentioned earlier, efforts are made to secure that the comparison is based on a unified conception of entities and on a unified scheme for analysis. Secondly, the comparison is related to a prior description of connectives in both languages, since, as argued before, unless one has a clear image of each, it is difficult to adapt adequately the description of one to fit the categories of the other. Thirdly, the comparison is not planned to face one way; rather it is intended to be bi-directional (Fig. 2.5). The relevant issue here is that, after categories of connectives in both languages are fully understood, the profile of connectives of one language is deliberately established by being viewed through the angle set up to view the profile of the other. The procedure can at any point be reversed by peeping through an opposite angle at the second profile to identify and inspect the first. But the earlier profile remains the basis for comparative statements unless it is drastically modified.

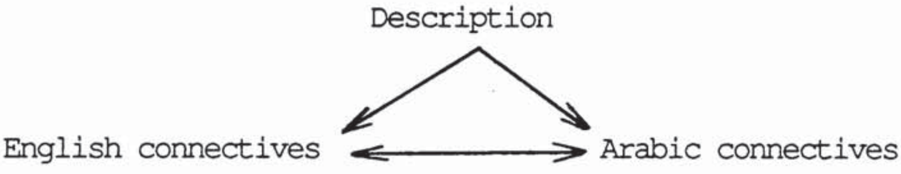

Fig. 2.5 Direction of TCA in the Project

The rationale behind the adoption of such a procedure is that there exist cases of mutual exoticism in the manner textuality is established through the use of connectives. The rhetorical effect of connectivity can be so different (cf., for instance, the use of the Arabic additive connective 'wa' with its seemingly equivalent English counterpart 'and') that a straight comparison is always open-ended. Besides, as manifested through observational analysis, of the two parallel corpora, the degree of explicitness can be so variable that the contrastive statements made will be distorted if a "neutral" comparison (one that brings together two languages that have been separately or independently described) is carried out. Furthermore, the type of confrontation represented in Figure 2.5 is expected within this study to yield statements viable enough to be of particular use in pedagogical applications.

2.5 Conclusion

The task of accounting for the behaviour of connectives in English and Arabic requires an assessment and delimitation of the general framework in which the various theoretical as well as experimental methods and procedures (that, in their totality, constitute the investigatory apparatus of the project) can most effectively be exploited. Since evidence that supports the formulation of generalisations concerning the behaviour of connectives can best be produced through examining natural texts as objects of enquiry, an early step was to examine the approach to text-based studies that is most relevant. (See also Appendix 2).

Since connectives are here regarded as means for creating and sustaining cohesion in text (cf. Ch. 4; see also Ch. 1 above), and since cohesion has been recognised as an essential standard in textuality (see Ch. 3 below), it became obvious that the employment of methods and procedures of the fast-expanding field of text linguistics can assist in proffering a valid description. An overview of the models currently operative in text linguistics has exhibited the diversity of issues that are pursued and that have, collectively, helped establish its credentials as a scientific approach.

This decision has an obvious consequence with respect to the contrastive component of the project. It is by now recognised that at a theoretical level, contrastive linguistic studies are influenced to a considerable extent by the theory and model of linguistic description that is adopted. Although it is a truism that in applied contrastive work, the approach may well be eclectic, i.e. picking and choosing among different theories and models, even using different theories for different areas of work, it is nevertheless the main framework of the theory that determines the depth and systematicity of the interlingual comparison. This is particularly evident in theoretical contrastive work.

This line of reasoning has led us to a detailed consideration of the various aspects of CA, its predictive power, its use and abuse. The general basic assumption of CA is that while languages are different, there is always a certain degree of similarity between them. This is indicated by the fact that most of what is written or said in one language is translatable into another language, and that one language can be taught to speakers of another totally different language. However, similarity is only partial, even with cognate languages or with dialects of the same language, let alone two widely different languages as English and Arabic. One of the main theses of CA rests on the assumption that languages consist of some isolatable elements that enter into certain arrangements. These elements are assigned to various hierarchical ranks of structural units and to levels according to specific criteria, mostly of paradigmatic, distributional or extralinguistic nature. If this categorisation of language elements is implemented on two (or more) languages (or varieties of a language) by a consistent application of a language theory, the results will be capable of manifesting greater or lesser similarity or diversity in relation to the isolated elements and their properties.

One serious drawback we have encountered in CA formulations is their inability to operate adequately across the sentence boundary. A more textual formulation is better equipped to secure the full contrastive significance of a) the patterns of interplay between various stretches of text, and b) the role of cohesive devices in sustaining and elaborating connectivity. Prominent among these patterns is the cohesive role of connectives, their diversity of patterns, syntactic realisations, intensity of use, range of connection and the type of semantic relationships they express.

Accordingly, a number of textual formulations and procedures in contrastive textual analysis were examined. Gleason's (1968) comparative discourse analysis examines the differences in the way connected discourse is organised and the way that organisation is signalled to the text recipient (hearer or reader). Kaplan's work on the rhetorical patterns of the paragraph, particularly his 1966 treatise, views the organisational role of the paragraph as a reflection of thought patterns that can culturally vary from one language to another. Hartmann's model of contrastive textology, particularly as expounded in his 1980 work, posits a textual analytic apparatus that has three dimensions: communicative, combinatorial and referential. One very significant notion that is elaborated in the model is the requirement for parallel texts as a procedure for contrastive purposes.

The theoretical and procedural tools for TCA in the project are predominantly eclectic (though still under text linguistic blanket). The main method for comparison focuses on connective patterning within a statistically adequate corpus of texts. The procedures include specification of the typology of constituents, categorisation of textual relationships that are realised on the surface text through the use of connectives, and achieving a profile of the quantitative behaviour of connectives in the two languages.

These procedures are monitored by a set of comparability criteria that ascertain that the phenomenon under investigation is in fact comparable. Comparison, then, starts with descriptive statements of the patterning of connectives first in English, then in Arabic. Contrastive statements are effected through viewing patterns and categories of connectives in one language in the light of the other. It is a reciprocal procedure which aims to ensure optimal viability.

To sum up, we have argued in this Chapter that a comparison of connectives in English and Arabic is more efficiently carried out within a text linguistic framework and conducted through procedures of a textual type of interlingual confrontation. Since connectives aim to achieve cohesion within a text, we need now to examine the concept of text, textuality and cohesion before we discuss the specific role of connectives. This will be the task of the next Chapter.

Footnotes to Chapter Two

(1) In the Institute for the Development of English Language Teaching in Iraq (IDELTI) a number of projects, which I actively participated in during the period 1975-1982, were carried out, under the direction of Professor K I Hamash, for the preparation of language teaching textbooks, teaching guides and teacher training manuals (which are currently in use on a nationwide basis). One of the underlying principles advocated is the implementation of the findings of CA, a task that practically proved difficult owing to the relatively small number of linguistic aspects investigated in the English-Arabic CA studies available.

(2) These are a) hortatory language (e.g. in election speeches of party candidates), b) legal language (e.g. in parliamentary bills and statutes), c) administrative language (e.g. in reports and civil service memoranda), d) bargaining language (e.g. in committees and lobbies) (see Hartmann 1980).

(3) For example, the Role realised by the object of "please" is realised by the subject of "like", while the subject of "please" and the object of "like" realise the same Role.

(4) It is unwise trying to envisage a single, general statement accounting for all the patterns that exist in a language and, therefore, it is not possible to produce an overall contrastive statement accounting for the difference between the two languages. The alternative accordingly is to examine a particular pattern, say typology and function of adverbials, in the two languages.

CHAPTER THREE

Text, Textuality and Cohesion

3.0 Perspective

One of the main issues in text studies is the analysis of the mechanics that contribute to textual well-formedness, particularly the formal relationships which must obtain among text components in order that a text functions in a meaningful way. Investigation into these formal relationships provides empirical inroads to the more general issue of how text constituents (of various size, type, and complexity) can accumulate to produce a higher order textual construct that manifests functional unity. Such an issue sees text in two perspectives: text as a product, i.e. an output having certain construction that can be represented in systematic terms, and text as a process, i.e. a continuous process of semantic choice, a movement through the network of the virtual system of language (cf. 3.2 below), with each set of choices constituting the environment for a further set (cf. Halliday and Hasan 1985).

The aims of this Chapter is to explore some of these issues. The first main issue is an examination of the properties of textuality. This includes observations of text as a product and reveals something of its dynamic unfolding as a process. Later, a closer and more rigorous examination is made of one essential principle of textuality, that of cohesion. The main focus will be the mechanics that are required in a text in order to relate its components and influence its organisation. The discussion of these issues serves as a perspective for the understanding of the textual function of connectives and hence is essential to this project. In carrying out this examination, we do not claim thoroughness; some issues need further elaboration that can best be left for further work. Nor do we claim conclusiveness; in text linguistics, a number of issues are examined and re-examined, each time in a different light and within a different framework of analysis. However, the examination is adequate in the sense that it lays the basis for the qualitative and quantitative statements made on connectives in the next few chapters.

3.1 On the Entity of Text

3.1.1 Text as a Linguistic Unit

The introduction of the concept "text" as a unit of linguistic analysis has proved operational in overcoming some of the shortcomings in linguistic theorisation outlined in the previous chapter. Its theoretical value has been manifested in a number of ways.

1. The concept "text" has helped to extend the system of linguistic levels put forward by modern linguistic theories that are based on the sentence. This extension has facilitated the understanding and explication of a number of textual issues such as cohesion and coherence (Dressler 1972, Beaugrande 1980) and their relevance to such problems as text typology (cf. Werlich 1976), macro-structures and macro-propositions (cf. Dijk 1972, 1979, Wirrer 1979). It has also proved convenient for linguists whose object of research is language as a system and who are conspicuously engaged

in its definition (Segre 1979).

2. It breaks the usual scheme of linguistic units seen as a mere additive progression and based on the linear character of any stretch of language, spoken or written. This scheme considers the morpheme as a succession of phonemes, and the sentence, in turn, as a syntagmatic cluster of morphemes (cf. Gracio-Berrio 1979).

3. The concept "text" has made it possible to shed a better light on a number of problems that have suffered certain shortcomings in treatment when based on analyses at the sentence level. These problems include issues related to translation theory and practice (cf. Wilss 1982), foreign language teaching and learning (cf. Werlich 1975, Keen 1979, Beaugrande 1980, and the papers in Kohonen and Enkvist 1978), and a host of psycholinguistic issues, such as those related to listening and reading comprehension (cf., for instance, Frankel 1977, Kintsch and Dijk 1978, the papers in Freedle 1977, Levelt and Flores d'Arcais 1978, Spiro et al. 1980, Fine and Freedle 1983), and text production (cf., for instance, Beaugrande 1984 and the papers in Whiteman 1981, Nystrand 1982, Martlew 1983).

3.1.2 Attempts at Defining Text

The surge in research on text linguistics has resulted in the formulation of widely differing definitions of text. Broadly speaking, four views to the concept can be identified in the current state of the art (cf. Ballmer 1982, where three views are discussed).

a. Text as a sentence: The first view identifies text with one single sentence. This view is adopted by Dascal and Margalit (1974) and employed as their justification for taking a negative attitude towards text linguistics. Dascal and Margalit (1974) criticise the concept of text as toò vague. Their focal view centres on the argument that "no evidence has been produced...supporting the claim that the description of the relations between independent sentences is not equivalent to the description of the conditions for constructing complex sentences by recursively embedding or coordinating other sentences" (p.199). This position has been refuted by a number of linguists, notably Petőfi and Rieser (1973) and Itkonen (1975, 1976, 1979). Itkonen sets out to prove in his detailed arguments that the use of recursivity is, to a grater or lesser extent, unjustified in natural language grammars and that, therefore, the position of sentence grammarians vis-a-vis text grammarians is correspondingly weak.

b. Text as a sequence of sentences: The more popular view among text linguists is to identify text with a sequence of sentences that manifests a number of properties, mainly cohesion and coherence (see, for instance, van Dijk 1972, Petőfi 1973, Garcia-Berrio 1979, Itkonen 1979, Longacre 1979, Wirrer 1979, Albaladejo Mayordomo 1982, Gindin 1982). Sgall (1979) questions this view stating that "it should be clear that a text is a sequence of sentence tokens (utterances, sentence uses or occurrences) rather than of sentences" (p.89) This statement is corroborated by his emphasis on use, claiming that a text "does not exist before being uttered by a speaker of the given language" (op. cit).

A divergent, but essentially relevant, view is expressed by Glinz (1979).

c. Text as a unit higher than a sequence of sentences: A different conception of text proposes an intermediate unit between a sentence and a text, though the actual status of such a unit is not well defined. Instead of a binary opposition (sentence/text) Langleben (1979), for instance, suggests a tripartite hierarchy of text, sentence cluster, sentence, with demarcation (opposition) line separating a text as "free form" (i.e. complete and closed in itself) from a sentence cluster and a sentence as "non-free forms" (i.e. constituting an essentially incomplete part of some other conglomeration). A similar view is expressed by Kukharenko (1979) whose views revolve on the claim that if text is to be considered a sequence of sentences, it is natural to expect the latter to be the minimal unit of the text-level, and the term used to label such a unit is , again, "sentence cluster". According to Kukharenko (1979, p.235) "sentence-clusters (adhered to as superphrasal unities, syntactical complexes, prosaic stanzas) present semantic, topical and lexico-grammatical unities of two or more sentences, often coinciding with paragraphs". A text can, then, make up one or more sentence-clusters and the textual structure presents itself in a hierarchy of linear elements. An exception is made in two cases: a one-sentence-text and a one-sentence-cluster-text.

d. Text as an autonomous entity: According to this view a text is an operational unit of language, a functional-semantic concept and is not definable by size (cf. Halliday 1977). Within this framework, a text is to be defined as a fundamental unit of semantics, and not as a kind of supersentence (Halliday 1978 p.135). A sequence of sentences, according to Halliday (1978), is in fact the realisation of text rather than constituting the text itself.

Furthermore, this view regards text as a unified whole in which there operates simultaneously a multiplicity of integrative devices (Halliday 1977, 1978, Fowler 1977, Hasan 1979). These can be studied under "texture" (Halliday and Hasan 1976, Hasan 1978, 1979), and "progression" and "localisation" (Fowler 1977). Text, according to this view, has a generic structure, is inherently cohesive, and constitutes the relevant environment for selection in the "textual" system of the grammar. Halliday (1977, p.195) argues that by "text", we understand "a continuous process of semantic choice. Text is meaning and meaning is choice, an ongoing current of selections each in its paradigmatic environment".

Having classified the various approaches to the definition and adoption of the term "text", we now consider critically the views that have been expressed in this respect, in an attempt on our part to arrive at a working definition that is to be manipulated within the course of this study.

3.1.3 Toward a Definition of Text

The previous section has made it clear that there are discrepancies and conflicting views in the explication of the term 'text' in the current literature of text linguistics. Additionally, the various explications are not without problems. Some of the general shortcomings that can be identified are summarised below:

1. The wide and general coverage that the term assumes over variable stretches of language results in its inclusion of all kinds of discourse, written and oral. Indeed one is obliged to employ an attribute (e.g. written text, literary text, etc.) when dealing with texts in the traditional sense. Although we subscribe to the theoretical motives which have led to the adoption of a single term (though it need not have been "text") we still feel that this loose and therefore unfortunate use has resulted in some vagueness comparable to that of other linguistic entities such as the "sentence" or the "word".

2. The seemingly loose delimitation of text size has created some misgivings over the accuracy of the concept "text", particularly in empirical terms. A text can refer to a whole novel, to an epic or to a section or a stanza. Portions such as an individual aphorism or entry into a diary, the shortened version of a fragment of a conversation have equally been designated "text" as the entire entity from which they are derived. This situation has caused some sceptics such as the Italian linguist Berruto to question the entity of "text": "What are the explicit, formal criteria which identify the beginning and end of a text? Is the Paolo and Francesca canto itself a text, or is part of the text 'The Divine Comedy' (or perhaps of the text 'Inferno')?" (Berruto 1979, p.500, cf. also Bertineto 1979).

3. The controversy over the nature of text constituency has created conflicting definitions. A text in some conceptions is a "sequence of words forming an actual utterance in a language"

(Hartmann and Stork 1972, p.236), while, as exhibited in the previous section, in others it is a unit consisting of more than one sentence. A slightly different view that attempts to discard with the term "sentence" as a constituent but retains "utterance" states that "stretches of speech of any extension, either spoken or written, are called texts" and that "any short stretch of text constitutes an utterance" (Pilch 1976, p.24).

4. There has been a confusion between the adoption of "text" to refer to a unit of theoretical analysis (a unit of theory language) for the purpose of establishing a virtual system, and the use of the same term to refer to a unit of actual system (a unit of object language). Halliday (1961 p.243), for example, employs the term as an object language construct; later, however, he uses the term as a theoretical construct (1975 p.123, 1977 pp.193-194). A distinction between these two entities is referred to in Dressler (1972), where the term "emic text" is used to refer to text as a theoretical device and "etic text" to refer to text as a unit of actual speech (cf. also Garcia-Berrio 1979 and Mortara Garavelli 1979).

The distinction of "text" in these two uses is sensitively delicate. The explication of "text" as an object language unit suggests the equality of extension between the commonplace term and the explican, while the explication of text as a theoretical unit implies the assumption and application of a determinate epistemological model and relies on the definition of the usage in a theory language. It can include, though not necessarily so, the equality of extension between the object and the theory language term.

It should be noted, however, that problems of explication such as these are not new in linguistic studies. They constitute part and parcel of linguistic inquiry and reflect the intricacy of and variation between theoretical models set up to facilitate our understanding of the nature of language. Contemporary linguistics, perhaps even more than any other discipline that studies Man, is burdened with the difficulty and requirement of defining its object of investigation, identifying its methods and specifying its tasks, and text linguistics does not escape this fate (cf. the comments in Giuliani et al. 1979, p.170).

Problems arise, as indeed in the conception of the entity "text", when a central constituting criterion, a defining element on its own, is in conflict with the postulation of certain limiting criteria (cf. Garcia-Berrio 1979). While it is true that a consensus for delimiting a central criterion for defining the concept of "text" is beyond the current state of the art of text theories, the same can be said of the entities "phrase", "clause" or "sentence" in sentence theories (cf. Kiefer 1977, Tolhurst 1979, Graustein and Thiele 1979a, 1979b, Harrah 1981, Morgan 1982). It is because of this situation that we are trying to define this concept as it is referred to in this study. Our attempt will start by examining the angles from which the concept is viewed, followed by a consideration of the relevant properties that "text" exhibits as compared to "non-text". The latter step will lead to a consideration of the principles of textuality.

3.1.4 Angles for Viewing Text

Two angles can be identified for viewing the concept of text.

1. A more syntactic-semantic angle. Accordingly text is classified as such by means of internal features and is based on the following form: whenever an entity E, regardless of the communicative situation CS in which it is used, exhibits the set of semantic-syntactic features X (x1, ..., xn), then E is a text.

2. A more pragmatic angle. Here text is classified as such according to external features. Schematically: whenever a communicative situation CS, regardless of the manner in which the entity E is made, exhibits the set of features Y (y1, ..., yn), then E is a text.

Viewed from angle (2) only the conception of "text" manifests a specific weakness. Since "text", viewed in this manner, will be made utterly dependent upon classes of external factors, and since these factors are variable and resist total confinement, an explication of the entity of text will be precluded. Most language users also connect a set of X features and not only a set of Y features. Furthermore, we are, in the present state of the art, far from having a complete picture of the communicative situations, acts and processes. For a viable and fully developed description of the set of Y features we need to draw on the cooperation of a number of disciplines: sociology, psychology, communication theory as well as linguistics. At the moment such a description is in its infancy, though we do not deny the existence of some recognisable attempts (among others, see for instance, Leech 1983, Levinson 1983, and the discussion in St. Clair 1980 and the papers in St. Clair and Giles (eds.) 1980).

The two angles specified above for examining the entity of "text" create different conceptions. In addition, different objectives are established, which vary in accordance to whether a priority is given to syntactic-semantic aspects of "text", i.e. internal features (in which case the objectives hinge around the description of the mechanism that unequivocally maps the semantic content onto the phonic or graphic strings), or to psycho-sociological aspects, i.e. external features (in which case the objectives are concerned with the actual linguistic realisations and the analysis of the mental processes and the social situations that determine them).

Our contention regarding this question commences from the admissibility of uniting the two angles and the scrutinisation of "text" in a broad, but still sufficiently relevant, perspective. Within this perspective, syntax, semantics and pragmatics are regarded as dimensions of language conceived of as a semiotic system. The word "dimension" in this respect refers to a total aspect of a participating language system.

These three dimensions are systematised, not independently from each other as early phases of linguistic studies used to assume (cf. Trager 1950), an outlook that seemed to be successful for the description of sounds, though even then, as Pike (1967, pp.362-3) observes, it was not fully upheld even by its own defenders. This assumption of the independence of syntax from semantics was maintained with considerable vigour by Chomsky (1957, 1965) and his followers. And although syntax and semantics were studied for many years, little regard was shown to the ways people use grammar and meaning in communication (Beaugrande and Dressler 1981). The use of language was relegated to the domain of pragmatics, the exploration of which has only recently started, and much concentrated effort is required for pursuing this direction.

We would like to adopt the view (cf. Schank 1975a-b, Beaugrande 1980) that there exists an interaction among these three semiotic dimensions, without which the utilisation of text would simply not be operational. The correlation and co-existence among these dimensions cannot be ignored or even reserved for a subsequent examination and interpretation. Any explication of text (as indeed the case with any model of language, cf. Walker 1978, Beaugrande and Dressler 1981) in which syntactic-semantic features are regarded autonomous from context cannot function as a reliable construct. Accordingly, the inception of a definition that caters for the study of text in communication can prove most effective as a starting point for the analysis of text organisation and for the explanation of real communicative behaviour.

3.1.5 Definition of Text in this Project

The most essential requirement for an entity to be designated text is its textuality (see 3.2 below), a term used to cover a phenomenon that is displayed through its principles or its "standards" (see Beaugrande 1980 and Beaugrande and Dressler 1981).

"A presentation is likely to be rejected as non-text only if the standards of textuality are so strongly defied ... that communicative utilization is no longer feasible" (Beaugrande and Dressler 1981 p.34). These principles, once sufficiently exhibited in a presentation, whether oral or written, characterise it as a coherent unit of mutually relevant elements, an integrated manifestation whose most decisive trait is its occurrence in communication.

The criteria and requirements for a sound conception of text, particularly as related to this study, are summarised in Appendix (3). With these criteria and requirements in mind, we can now formulate our view of text as a linguistic entity. Text is regarded as the naturally occurring organised manifestation of language. Its size is immaterial since language occurrences may have the surface format of a single word, one sentence or a sequence of sentences. Text is a meaningful configuration of language intended to communicate, definable in accordance with the extent of its adherence to the standards of textuality.

Some of the portentous implications of this definition are:

a) The emphasis on the naturalness of occurrence: a contrived piece of language, such as examples found in grammar books, are by themselves only means for illustration and cannot be regarded as text (unless they are incorporated within a text).

b) The communicative nature of the entity of "text": Text forms an essential part in a communicative act. It creates a message by dynamically relating the encoder's knowledge and perspective of reality to the context.

c) The concern with communication requires the expansion of the traditional restrictive confines of linguistics in order to enhance its interaction with other language-related disciplines: psychology, sociology, computer science, statistics, philosophy, cybernetics, education and literary studies. This justifies in our study the requirement to resort to computer application and statistical work in analysis, categorising and quantifying.

d) The role of textuality: The principles of textuality are considered essential in characterising a manifestation as "text". These will be outlined in more detail in the next section.

3.2 Textuality

3.2.1 Text Actualisation

In the previous sections we stated that the definability of an entity as "text" is dependent upon its extent of adherence to the principles of "textuality". We would like in this section to delineate the scope and essence of this term. Our conception is loose modification of Beaugrande (1980, 1984) and Beaugrande and Dressler (1981).

We depart from a distinction that Beaugrande makes (following Hartmann 1963 and Gülich and Raible 1977) between the "virtual" and "actual" systems of language. System as a notion is applicable not only to a language level, but also to the entity of "text" (cf. Fowler 1977). The intersystem of a particular language (English or

Arabic, for instance) consists of "virtual" systems, i.e. "functional unities of elements whose potential is not yet put to use" (Beaugrande 1980 p.16). Examples for such systems include the repertories of sounds, morphological forms, sentence patterns, concept names, etc., which a particular language possesses and are awaiting use.

A system is "actualised" when it is put in use, that is, when some of its options are being activated (taken from a storage and implemented). A text is, therefore, an actual system: a functional unity that evolves through processes of decisions of selection among choices of the virtual system. In other words, text is created through processes describable in terms of preferences, i.e. elements are selected from virtual systems and this selection is viewed in relation to the other choices which were available in the system.

This distinction is not new (cf. McTear 1984): it is reminiscent of Halliday's systemic framework. Halliday assumes that "at every place in the structure of every unit, one or more choices are made" (Halliday 1961/1966 p.14), and that "the grammar is based on the notion of choice" (1969/1976 p.3). Halliday maintains that the user of a language, "like a person engaging on any kind of culturally determined behaviour, can be regarded as carrying out, simultaneously and successively, a number of distinct choices" (ibid p.3). Thus, within the systemic framework "the description of a linguistic item is the set of features selected from the total available" (ibid p.4), that is, from the "virtual" systems. Stated differently, "the description of a sentence, clause or other item may be just a list of the choices that the speaker has made" (op. cit.).

The evolution of text is labelled "actualisation" by Beaugrande (1980, 1984, cf. Beaugrande and Dressler 1981), a useful term that reflects the creation of text as an "actual" system. Text production is one example of actualisation, and the text itself is one example of an actual system. Any further utilisation of a text, such as reading, interpreting, translating, quoting, etc., is also actualisation and activation, though under new conditions and different environment.

3.2.2 Text as a Manifestation of a Cybernetic System

The process of text actualisation is describable in terms of cybernetics. One of the basic features of cybernetics is that it does not only consider control systems in their static state but also during movement and development (Lerner 1972). In a number of cases, such a dynamic approach reveals relationships and facts which otherwise would remain undiscovered. Such a functional property of systems, for instance, as "stability", which is of decisive importance for evaluating the serviceability of many systems, would be impossible to account for without considering the dynamics of their internal organisation.

Stability is used to describe the constancy of any behavioural feature of a system. It refers to the property of a system which enables it to return to its original state after a disturbance (Muller 1964/1968 p.168)[1]. This property is applicable to text as an actual system. The environment if "actualisation" requires the adaptation of the intersystems of language according to the demands imposed on them by the context. In other words, actualisation should manifest a sufficient degree of stability to constantly adapt to the contexts. Consequently, the actualised text system possesses and displays as part of its internal organisation the appropriate modifications and adaptations performed during the process of actualisation.

The systems remain stable so long as they support utilisation and continuity (Beaugrande 1980) (though texts can occasionally experience greater or lesser discontinuities). It follows that the stability of the text as a cybernetic system relies heavily on the continuity of occurrences in participating systems. Continuity is reflected by the relations that hold the textual system together. Stated differently, continuity is reflected by the connectivities that characterise a text: the unbroken access among the occurring elements of the participating language systems. Continuity can be experienced as the fuzziness of the boundaries among the elements that constitute the text. It is textually displayed through sequential and conceptual connectivities created respectively by the mechanisms of "cohesion" and "coherence". Furthermore, continuity is reflected through planning connectivity, so that each component (i.e. text fragment) is relevant to some interactive or communicative plan that aims to satisfy specific goals. The context determines to a large extent the number of actual occurrences needed for connectivity to prevail(2).

The three types of connectivities make up the different domains of textuality and constitute the basis for the actualisation and utilisation of text. Textuality is monitored by "constitutive" as well as "regulative" principles (3). The constitutive principles figure as the criteria upon which the entity of "text" as opposed to "non-text" is specifically dependent. The regulative principles determine the quality of a presentation that has satisfied the requirements demanded by the constitutive principles. Below is an outline of these two sets of principles.

3.2.3 Constitutive Principles of Textuality

Beaugrande (1980, 1984) and Beaugrande and Dressler (1981) suggest seven constitutive principles of textuality. These are: a) cohesion, b) coherence, c) intentionality, d) acceptability, e) situationality, f) intertextuality, and, g) informativity.

For an explication of cohesion and coherence as postulated in this study see 3.5 below. Both these two principles are concerned with the intrinsic features of text.

Intentionality and acceptability relate to the attitudes of the text users: the producer and the recipient respectively, during the process of actualising the text. Intentionality subsumes the text producer's attitude that the presented configuration is to be considered not only as a cohesive and coherent entity but also as manifesting relevance to the plans and goals of the producer. By "relevance" is meant the capability of the text of affecting the chances of the plans and goals. "Plan" is here employed in the sense of a set of steps configured with the intention of leading to a specific goal. "Goal" is definable as a future state of the world whose attainment is envisaged (and, usually, desired) and intended to be brought about by the actualisation of the text.

It should be noted that intentionality possesses a range of "tolerance" where it remains in effect even when the principles of cohesion and coherence are not fully satisfied, and when the plan does not lead to or attain the envisaged goal.

Acceptability, on the other hand, subsumes the text recipient's attitude to regard the presented configuration as a cohesive and coherent entity having some relevance to the recipient, e.g. to acquire knowledge or provide cooperation in a plan. This attitude is affected by such factors as text type, social or cultural setting and the desirability of goals.

Acceptability also possesses a tolerance range where it remains operational even when the context brings disturbances, or where the recipient does not share the producer's envisioned goal.

Situationality is the term used (cf. Beaugrande 1980, 1984, Beaugrande and Dressler 1981) to subsume all the ways in which a text is relevant to (i.e. affects the creation of) a real or recoverable situation. The text is seen as an action capable of both monitoring and managing a situation (where "action" is defined as an event enacted by an agent to change a situation).

Intertextuality refers to the ways in which the text presupposes knowledge of other texts. In Beaugrande's words, intertextuality "subsumes the relationships between a given text and other relevant texts encountered in prior experience" (1980 p.20).

Beaugrande goes on to maintain that intertextuality is the major factor in the establishment of text types, where expectations are formed for whole classes of language occurrences.

The seventh principle of textuality is informativity, which concerns the extent to which text events are uncertain, new, known, or surprising. In cybernetic terms, informativity is the extent to which an event disturbs the stability of a textual system and requires regulation. Considered from an operational perspective, informativity can be subdivided into "familiarity", i.e. the degree to which an event or operation has been encountered by the processor, and "unfamiliarity", i.e. the degree to which any portion of the text is unpredictable in view of the whole.

We have now glanced at the seven constitutive principles of textuality, the existence of which determines the definability of the entity "text" and which, when defied, bring textual communication into a halt. The next section considers the regulative principles that operate as a control over the constitutive principles.

3.2.4 Regulative Principles of Textuality

The regulative principles are instrumental in defining variations in the manner in which texts are actualised. All texts, by definition, are responsive to the constitutive principles, but they differ in the design of their actualisation. The regulative principles (cf. Beaugrande 1980) are:

a) Efficiency: This refers to the utilisation of a text (by producer or recipient) with the least outlay of effort.

b) Effectiveness: This principle subsumes the degree to which the text has an impact on the recipients, thus promoting processing depth, and forwarding the producer's chances to attain his goal. It is concerned with the relevance of text materials to steps in the producer's plan.

c) Appropriateness: This term refers to the extent to which the producer's or recipient's choices fit the current setting. Appropriateness (cf. Beaugrande op. cit.) relies on the proportionality between the demands of a communicative situation and the degree to which the constitutive principles of textuality are upheld.

3.3 Cohesion

3.3.1 Preliminaries

The term cohesion was made current in linguistic research by Halliday (1964) when he made an early version of his functional-systemic account of cohesion. Since then, several models have been proposed which are intended to characterise organisation in text and discourse and to delineate the nature of dependence in the connectedness among sentences in written text or "turns" in spoken text. Common to all these models and to related studies is the focus on the various aspects of the global structuring of discourse and text and the view that cohesion is a defining property of text (and therefore of acceptable discourse). Additionally, these models and studies share the position that cohesion is a tangible property, i.e. it can be identified and, at some level, measured. A further common stand is their emphasis on sequence in text while tracking the various cohesive mechanics and defining their scope of operation.

Differences among these models represent diverse modes of conceptualising organisation beyond the sentence level, and variations in the degree of emphasis and in their scope of analysis. Happily, this area in text theorisation is far from a unified field of enquiry and the diversity generates vitality in current research.

Although more can be said on this issue, it is not practical to attempt a review within this thesis of the variety of models and approaches proposed for cohesion. Such a task can perhaps be carried out in a separate study. However, by way of setting the scene for our own account of cohesion, a review will be made of four well-known models in text research: the stylistic model as represented by Enkvist (1973) (cf. also his 1978a thesis), the functional-systemic model as initiated by Halliday (1964), expanded by Hasan (1968) and developed further by Halliday and Hasan (1976) (cf. also Halliday 1977, Hasan 1978, 1979, and the summary in Halliday and Hasan 1985), the stratificational model as proposed by Gutwinski (1976) and the procedural/relational model suggested by Beaugrande and Dressler (1981) (cf. also Beaugrande 1980). The review of the proposals in these models is sufficient, in our view, to shed light on the diversity of emphasis and scope and will assist in formulating a synthesis that will represent, with some elaboration, our own position.

3.3.2 Models of Textual Cohesion

3.3.2.1 Enkvist's Stylistic Model

The model that Enkvist proposes (Enkvist 1973, see also Enkvist 1978a) views textual cohesion from a linguistic-stylistic perspective, with potential application to the analysis of predominantly literary texts (cf. Gutwinski's model, see below).

Enkvist departs from the assumption that a) styles can be viewed as varieties of language that correlate with specific constellations of contextual features, and b) stylistic analysis should always be based on comparison, tacit or implicit. A body of text defined by contextual criteria can be stylistically investigated by being compared to other texts that are recognised as relevant and sensible norms of comparison. To Enkvist, comparison is the only key to stylistic differentials, that is, to the style markers that characterise one text as different from other texts. It follows that all stylistic descriptions must begin with an inventory of stylistic markers.

These markers, according to Enkvist (1973, p.110), can be located in individual sentences as well as spans larger than the sentence. For "single sentences have style, and stylistic incongruities such as the use of a colloquial word in an otherwise solemn, high style frame may occur within the bounds of one sentence." On the other hand the manner in which sentences are strung together into texts can also function as a style marker, particularly in contexts characterised by the use of textually deviant sentence strings. Patterning of sentence sequences is an essential stylistic aspect. "If certain patterns of sentence sequence are significantly more frequent in a given text than in a norm chosen for its contextual relationship with that text, they qualify as style markers precisely like any other linguistic features." (p.115).

Textual style markers are classified into two major fields: A) theme dynamics, B) cohesion devices between sentences and textual units, including linkage which overtly marks relations between sentences. These are discussed below.

A. Theme Dynamics

Enkvist's development of theme dynamics as an apparatus for the description of patterns of sentence sequence is based on syntax and draws on studies of theme as elaborated by the Prague linguists (Mathesius, Fibras, Daneš, Adamec), by Halliday (1967a, 1967b, 1968a) and others (e.g. Dahl 1969). However, Enkvist maintains that within intersentence grammar and text linguistics, the investigator should not be satisfied with an apparatus capable only of discussing the statics of theme and rheme. There is therefore a need for theme dynamics expressly designed for description of thematic cohesion in strings of sentences. These dynamics chart "the patterns by which themes recur in a text and by which they run through a text, weaving their way from clause to clause and from sentence to sentence". (p.116).

Theme dynamics consist of three parts:

1) Theme statics, that is, a theory of theme in a clause and sentence. Theories of this type are already available.

2) A theory and method of thematic identification, which facilitates the comparison of thematically definable parts of different sentences and the decision whether to regard them the same or different, irrespective of whether they are expressed with the same words or not. At present, lack of sufficiently rigorous semantic theory of synonymy leads to maintaining some very rough-and-ready systems of theme identification. Themes may thus be regarded as the same if they fit into certain patterns of semantic relationship such as a) repetition, b) reference, c) synonymy, d) antonymy, e) comparison, f) contracting hyponymy, g) expanding hyponymy, h) co-membership of the same field, i) sustained metaphor (see Enkvist 1973 pp.117-118 for a more detailed discussion and exemplification).

Two remarks are to be made in regard to these categories. First, sentences can often be linked thematically by the simultaneous use of more than one device of thematic identification. Secondly, the categories listed above can further be subdivided for greater delicacy. For instance, a subclass of the category (h) can be assigned the label "indexical", a semiotic term, to indicate a special word-field relationship as in, for example, "sun" and "shadows".

3) A taxonomy of patterns of theme movement through the successive sentences of a text. This component is required even if the second one has been identified and established. In this

respect, and despite the various difficulties that a theoretical conception of the terms "theme" and "rheme" causes, one can operationally and strictly discuss thematic movement in terms of two positions: I(nitial) and N(on-initial). In this case, (and assuming that I and N are precisely defined) one can work out four possible patterns of thematic movement: I to I, I to N, N to I and N to N.

Enkvist recognises that there are various possible principles of classifying thematic movements. One criterion is syntactic function: a theme may move from the subject of one sentence to the subject of another, from subject to object, from object to subject, and so on. Another is syntactic structure: thematic features may move from a noun phrase to a verb phrase, and so on.(4) One principle suggested for classification concerns the measurement of distance between sentences of related themes. For instance, while in some texts thematic movement progresses from sentence n to sentence n + 1, in other texts the progression may advance from sentence n to n + 2 or n + 3, and so forth.

Methodologically, an investigator, in order to cut time-consuming effort in laboriously processing large bodies of text, has to fall back on simple techniques for surveying and identifying themes and thematic movement. One type of theme-dynamic display is the "cohesion chart", in which "the clauses and sentences are numbered and plotted against the various cohesion devices" (ibid, p.121). Another is the "stylistic profile", in which "relative numbers of cohesion devices are given in staple diagrams or histograms" (op.cit).

B. Cohesion

The second field Enkvist proposes for classifying textual markers comprises various types of cohesion. Enkvist recognises the textual role played by such cohesive devices as anaphoric and cataphoric reference, pronominalisation, the use of referential do or one. In addition to these, there are a number of cohesive features that are less formal yet still amenable to linguistic analysis and description. Enkvist then focuses on four types of such features: contextual cohesion, lexical cohesion, clausal linkage and iconic linkage.

Contextual cohesion "keeps together passages occurring in the same matrix of contextual features" (ibid, p.122). For example, in a novel, a dialogue has a contextual matrix different from a descriptive passage in the same novel. Similarly, in a play, stage directions are under the contextual constraints of a matrix different from that of the dialogue in the play. Each verbal strand displays typical and distinct cohesive patterns.

The second type of cohesion, lexical cohesion, suggests that "coherent texts often have a homogeneous vocabulary, which contributes to their unity" (op.cit). Homogeneity of vocabulary may be affected by a number of factors. One vital factor is the subject matter of the text; for instance, an article on plants is likely to contain a high density of terms related to plants. Other factors comprise various contextual features, including style: a colloquial text is likely to employ a stylistically homogeneous, colloquial vocabulary.

The third type of cohesive features, clausal linkage, is discussed in the next chapter when textual studies of connectives are reviewed. The "fourth type" is "iconic linkage", a term borrowed from semiotics. It subsumes "those situations in which two or more sentences cohere because they are, at some level of abstraction, isomorphic" (ibid, p.123). For example, one line of Pope is highly likely to be metrically isomorphic with another line of Pope. In identifying iconic linkage, one is compelled to determine the level of abstraction at which the isomorphism is significant as an iconic link. As a rule such isomorphisms have to be realised at, or close to, the surface.(5) Instances of iconic linkage include rhythmic and metrical regularities, rhyme, alliteration and assonance. Furthermore, iconic links may also be syntactic, linking, for instance, "The old gentleman elegantly kissed the young lady" with "The striped tiger cruelly bit the innocent lamb".

Other cohesive features that Enkvist proposes is the consistent use of certain tenses (Weinrich 1964) and the consistent use of such aspects of point of view as can be linguistically defined (cf. Sinclair 1968). At this point the model approaches the borderline between text linguistics in the strict sense and poetic and narrative analysis of the kinds developed by the Russian formalists and French neo-structuralists. Enkvist leaves the border open for any translation of literary concepts into linguistic terms.

The relevance of all these textual and intersentential patterns for stylistics is summarised in two points (Enkvist 1973 pp.125-126):

First, they reveal the kinds of conceptual frames employed if agreement is reached that style is not merely a quality of sentences but also of texts. In this case, means for describing style must be devised, "which reckon with textual, intersentential features and not only with terms that refer to phenomena within the confines of single sentences" (ibid p.125).

Secondly, patterns of textual cohesion provide the investigator with "a vast arsenal of additional style markers". Accordingly, stylistic differentials between text and norm can be expressed with the aid of densities of cohesion devices. For instance, one can test a hypothesis such as "X's scientific style is characterised by a comparatively high density of thematic movement, from rheme in sentence _n_ to theme in sentence _n_ + 1". Furthermore, Enkvist suggests that observations of textual cohesion patterns and of devices of theme dynamics "may also yield material for practical tasks such as the teaching of composition and normative stylistics" (ibid p.126), a view that we share with him.

3.3.2.2 The Functional-Systemic Model (Halliday; Hasan)

A major and integral strand in functional-systemic linguistics has been the study of the "discoursal" or "textual" function of language.(6) This strand has motivated and directed the development of the functional-systemic conception of cohesion, a model that has been greatly influential over the last decade.

First we start with a number of basic assumptions fundamental to the understanding of the model. Halliday (cf., for instance,

1977) views language as made up of three levels, (or strata in Lambs's stratificational sense, cf. Lamb 1966): semantic (semology), lexico-grammatical (lexicology: syntax, morphology and lexis), phonological (phonology and phonetics). The second assumption concerns the semantic level, which is viewed as having four components: experiential, logical, interpersonal and textual. The first two are so closely related that they can be combined under the label "ideational".

Each stratum and component is described as a network of options, sets of interrelated choices.(7) The description is viewed as both paradigmatic and open-ended. Each component of the semantic system is assumed to specify its own structures as the "output" of the options in the network (ie, each act of choice contributes to the formation of the structure). These structures are mapped one on to another through the lexico-grammatical stratum so that they form a single integrated structure that represents all components simultaneously.

Another assumption ascribes to the lexico-grammatical system an organisation by rank (as opposed to immediate constituent structure). That is, units identified at a particular rank make up the building blocks for the rank above. Each one of these units is, at the same time, the result of the combination of the units at the lower rank. Each rank is "the locus of the structural configurations, the place where structures from different components are mapped on to each other" (Halliday, 1977, p.177).

To Halliday, the grouping of semantic components differs according to the perspective from which they are viewed. For instance, from a higher level of semiotic vantage point, "it is the textual component that appears as distinct, since the textual component has an enabling function in respect to the other components: language can effectively express ideational and interpersonal meanings only because it can create text" (ibid, p.178). To avoid any ambiguity that can result from the above formulation, Halliday stresses that the entire semantic system is "text-forming", in the sense that "a text is the product of meanings of all four kinds - experiential, logical, and interpersonal, as well as textual" (ibid p.181). However, it is the textual component whose function is specifically that of creating text, of differentiating between "language in the abstract" and "language in use". Stated differently, the semantic options available in the textual component assist in distinguishing between text, as language relevant to its context, from non-text or decontextualised language such as words listed in a dictionary.

The textual component, then, embodies the specifically text-forming resources of the linguistic system. This component as worked out for English is composed of the following: 1) The structure-generating systems, which are of two kinds: a) thematic system, and, b) information systems (see Halliday 1968a, 1968b). 2) The cohesive relations: non-structural resources in the sense that they are not realised through any form of structural configuration. These resources, both structural and non-structural, provide "texture" (in the sense of text constitution) in the language.

Halliday does not imply that these are universal features; but the systems in each network and the way the systems are realised are specific to such languages as English.

These assumptions form the basis of the functional-systemic model of cohesion, particularly as envisaged by Halliday and Hasan (1976). However, before discussing the characterising features of this model, we would like to examine the earlier conception of cohesion within this model. This requires a brief discussion of Halliday's early views on the subject and Hasan's subsequent treatment.

3.3.2.2.1 Halliday's Early Views

In his earliest treatment of cohesion, Halliday (1964) suggests that the categories subsumed under cohesion are two types: grammatical and lexical (Figure 3.1).

Fig. 3.1 Categories of Cohesion (Halliday 1964)

The distinction between "structural" and "non-structural" cohesion was reiterated and reinforced in Halliday et al. (1964), with a focus on the view that cohesion may occur both within and between sentences. In the same work (pp.247-8), Halliday et al. observe that

> "... all languages display certain features which can be regarded as cohesive; such features are of different types and belong to different levels, but all contribute to the internal binding of the text. The fundamental type of cohesion is of course the relation of structure itself: two or more items entering into a structure always cohere. But there are other, non-structural features exerting a similar force. In English these include grammatical anaphora, grammatical substitution and lexical anaphora, grammatical substitution and lexical anaphora ... All such features, including those which extend across sentence boundaries, will figure somewhere in a description of English .."

3.3.2.2.2 Hasan's Treatment of Cohesion

Subsequent treatments of cohesion within the functional-systemic approach started to focus on non-structural cohesion. The first detailed treatment was made by Hasan (1968). In this work she attempts to "identify some of the features which distinguish a text from a disconnected set of sentences, in order to try and establish what it is that determines that a passage of English forms a text". Cohesion is defined in this way (ibid, p.8):

> "It is the internal, linguistic features characterizing a text and distinguishing it from an agglomeration of sentences that we are here referring to under the name of 'cohesion'".

These features are cohesive by virtue of their contribution to the unity of a text.(8) This is specifically performed by linking one sentence to a sentence or group of sentences preceding or following.

Hasan labels any single linguistic feature having this function as a "cohesive tie".

Cohesive ties are then categorised as "lexical", "phonological", and "grammatical". Lexical cohesion includes such features as repetition of items and use of near-synonyms. Phonological cohesion refers to such aspects as intonation (in conversation), meter and other phonological aspects of verse (in poetry).

However, it is grammatical cohesion that receives the main emphasis in Hasan's study. Features that are classified under the type of cohesion include reference, substitution, ellipsis and conjunction. Hasan points out that not all of these categories "are sharply distinct; but the division offers a means of presenting the facts in a reasonably systematic manner" (ibid, p.25).

Hasan's work, particularly her methodology, has been criticised (most strongly, perhaps, by Widdowson 1973) on the grounds that it is atomistic and fragmentary. In Widdowson's view, while Hasan's partial taxonomy of cohesive devices, with its many examples, is a useful reference tool, "it does not show how the devices are differentially used, how they relate with each other and with lexical cohesion to create text" (p.114). Widdowson outlines, through contrasting Hasan's approach with that, for instance, of Hilyer (1970), that Hasan selects the grammatical level of analysis and then proceeds to show which elements from this level fulfil a cohesive function. In his opinion, she should have adduced different categories from a representative text to show how "it is the manner in which they relate that makes the constituent sentences hang together" (Widdowson, 1973, p.115).

This criticism can be rejected on the basis that it does not take account of the analyst's purpose, which, in Hasan's work, is purely descriptive. After identifying a grammatical feature of cohesion, Hasan is simply concerned with describing how it is realised. There are, therefore, self-imposed limitations on what she has set out to do.(9) The claim, therefore, that Hasan's analysis is inadequate because it does not exhibit how the linguistic items that realise cohesive relations are used, differentially or interactively, to create text, ignores to a large extent the aims and purposes of the analysis itself.

3.3.2.2.3 Halliday and Hasan's Categorisation of Cohesion

Some of the criticisms in Widdowson (1973) were rectified in Halliday and Hasan's monumental work (1976). Here, Halliday and Hasan spell out, clearly and elaborately, their conception of cohesion and interpretation of its role in the creation of text. Some of the main characterising elements of "cohesion" in their theoretical statement can be summarised as follows:

a. First of all, cohesion is a component of "texture", a term that refers to the property of "being a text". A text has texture, which distinguishes it from non-text. A text "derives its texture from the fact that it functions as a unity with respect to its environment" (ibid, p.2). Cohesion refers to the resources that a language has for creating texture. These resources are linguistic features that exist in a text and that can be identified as contributing to its total unity and thus giving it texture.

b. In addition, cohesion is a semantic concept (not a structural one) in that "it refers to relations of meaning that exist within a text and that define it as a text" (ibid, p.4). Cohesion "does not concern what a text means; it concerns how the text is constructed as a semantic edifice" (ibid, p.26).

c. Cohesive relations are properties of text, not of any structural unit such as the sentence; they are non-structural semantic relations. Halliday and Hasan accept that structure is a unifying relation: the elements of any structure have an internal unity which satisfies the expression of part of a text. All grammatical units - sentences, clauses, groups, words - are treated as internally "cohesive" because they are structured. Indeed, structure is one means of expressing "texture". If every text consists of one sentence, then structure will suffice to explain its internal cohesiveness. But a text "typically extends beyond the range of structural relations as these are normally conceived of" (ibid p.7). Since texts cohere, cohesion then depends on something other than structure. In other words, text-forming relations are non-structural; they are not describable in terms of constituent structures.

d. Cohesion occurs whenever one element in a text depends for its interpretation on some other element(s) in the text; the one presupposes the other in the sense that it can only be effectively interpreted by reference to it. "When this happens, a relation of cohesion is set up, and the two elements, the presupposing and the presupposed, are thereby at least potentially integrated into a text" (ibid, p.4).

e. The term used to describe a single instance of cohesion, i.e. the cohesive relation between the presupposing and the presupposed items, is a "tie" (cf. Hasan 1968). This concept assists in analysing a text in terms of its cohesive properties, and in giving a systematic account of its patterns of texture. Indeed, a segment of text can be characterised in terms of the number and kinds of ties which it exhibits.

f. The text-forming resources are categorised into grammatical and lexical types. Each of these is then categorised into subclasses. Figure 3.2 displays the categories of cohesion as proposed in the model. A few remarks on these categories are in order.

Reference is a semantic relation characterised by the specific nature of the information that is signalled for retrieval. This information constitutes the referential meaning, the identity of the thing(s) referred to; "and the cohesion lies in the continuity of reference, whereby the same thing enters into the discourse a second time" (ibid p.31). The linguistic elements that are interpretable by reference to something other than themselves include the personals, demonstratives (including <u>the</u>) and comparatives (these elements are perfectly intelligible on their own, but they are interpretable only when we know who or what they refer to). Personal reference depends on the concept of personal roles in the

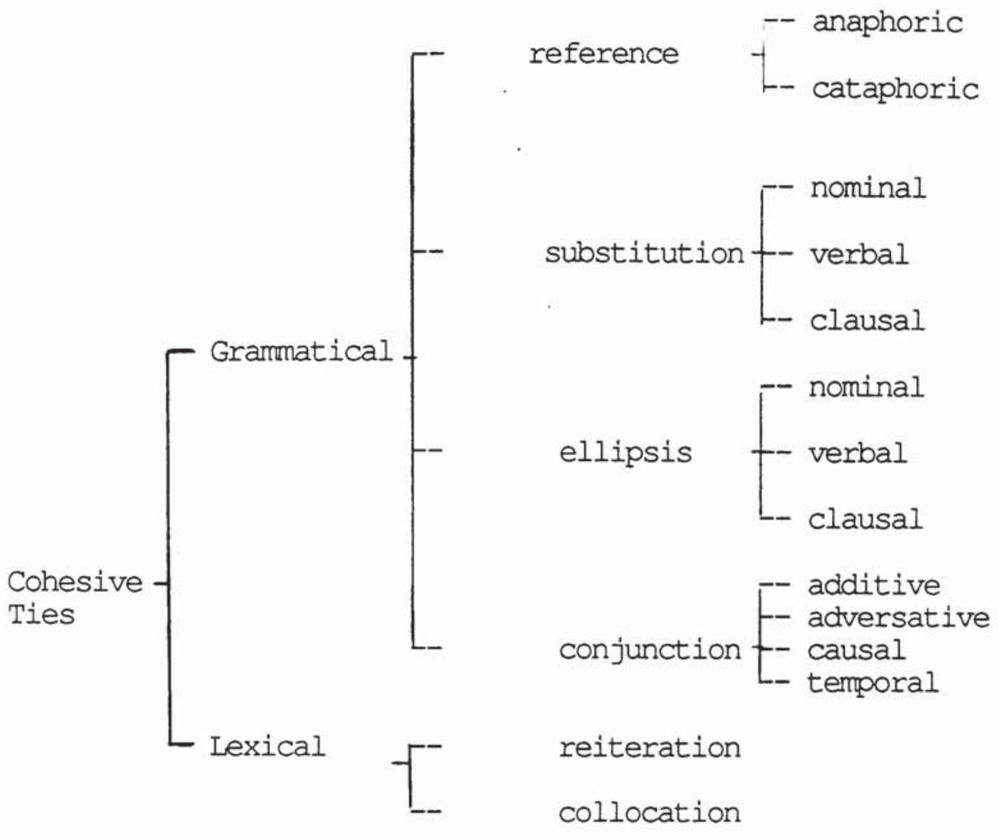

Fig. 3.2 A Summary of Halliday and Hasan's Categorisation of Cohesive Ties

speech situation, i.e. some person or object other than the speaker and addressee(s) e.g. she. Demonstrative reference is based on proximity, i.e. "near" or "not near" e.g. this. Comparative reference involves a conception of likeness and unlikeness between phenomena, e.g. earlier. The article the functions as an unmarked demonstrative; it signals that the referent can be identified but without locating it on any semantic scale.(10)

The reference may be "exophoric" (referring to some phenomenon located outside the text and in the context of situation), or "endophoric" (referring to an element within the text). Endophoric referential elements typically relate to parts of the text that have preceded (i.e. anaphoric reference) or to parts that follow (i.e. cataphoric reference).

Exophoric reference contributes to the creation of text by linking the language to the context of situation. However, it does not contribute to the integration of one part of the text with another. Hence, it does not contribute directly to cohesion as the model defines it. Accordingly Halliday and Hasan take little account of exophoric reference, while their focus lies on endophoric reference, which is treated as the norm.

Reference, in the endophoric sense, is treated as a text-forming agency, since it contributes to the making of a text. It is a signal that the interpretation of a particular item is to be sought somewhere else in the text. Reference is thus independent of the linguistic structure, and so may extend beyond any structural unit.

Two other types of cohesive relation are substitution and ellipsis. These function as alternatives to repetition of a particular item, and hence cohere with the passage in which that item occurs. Substitution and ellipsis are essentially the same process; ellipsis can be interpreted as that form of substitution in which the item is replaced by nothing.

Compared to reference, substitution is a relation between linguistic items, such as words cr phrases, whereas reference is a relation between meanings. In terms of the linguistic system, reference is a relation on the semantic level while substitution is a relation on the lexico-grammatical level. But from the point of view of textual cohesion, substitution resembles reference in being anaphoric, and hence constituting a link between parts of a text.(11)

There are three types of substitution, defined grammatically (rather than semantically, since substitution is a grammatical relation). These types are nominal (with the substitutes one, ones, same), verbal (with the substitute do and its various forms), and clausal (with so, not).

Ellipsis involves the notion of "something left unsaid, but understood nevertheless". There is some presupposition in the structure that assists the supply of the missing elements. Ellipsis does not refer to any instance in which there is some information that the speaker has to supply from his own evidence; that would indeed apply to every sentence that is ever spoken or written and would (as Halliday and Hasan view it) not help in explaining the nature of a text. Rather, ellipsis refers to sentences, clauses, etc. "whose structure is such as to presuppose some preceding item, which then serves as the source of the missing information" (ibid p.143).

Lexical cohesion is achieved by the use of vocabulary, a) by reiteration, and b) by collocation. Reiteration can be displayed in a matrix. There is a certain arbitrariness in both dimensions of the matrix, but each is motivated by general considerations. The vertical dimension represents the organisation of the system while the horizontal one indicates the patterning of text. The vertical dimension is a scale: one end represents repetition of the same lexical item whereas the other end represents the class of "general words" that refer to the lexical item. Between the two ends are such types as synonyms, near-synonyms and superordinates. The horizontal dimension shows the referential relationship between the reiterated item and the base lexical item: co-referential, inclusive, exclusive or unrelated.

Collocation as a type of cohesive relation is achieved through the association of lexical items that regularly co-occur. Their proximity in a discourse is treated as a contribution to texture. Halliday and Hasan extend the basis of the lexical relationship that features as a cohesive force and maintain that "there is cohesion between any pair of lexical items that stand to each other in some recognizable lexico-semantic (word meaning) relation" (ibid p.285). This includes words related by a particular type of oppositeness (called "complementarity" in Lyons' (1968) classification, e.g. boy, girl), words drawn from the same ordered series (e.g. Tuesday.. Thursday), and words drawn from unordered lexical sets (e.g. road .. rail, red .. green). The members of a pair often stand in some identifiable semantic relation: part to whole, part to part, hyponyms of the same superordinate term, and so on. The effect is not limited to a pair of words. It is very common for long cohesive chains to be built up out of lexical relations of this kind.

Having discussed each of their categories of cohesion and the ways in which they are realised with copious exemplification, Halliday and Hasan then propose and illustrate (ibid, Ch.8) a procedure for the analysis of texts which displays how a network of cohesive devices of different categories interact within a text.

First they outline the principles of analysis; then they suggest a coding scheme for the various types of cohesion, and finally they work out an analysis of seven short passages of text.

Halliday and Hasan have made a major contribution to a better understanding of the linguistic resources exploited by language users in the creation of text. Their model has been influential in that it inspired research, of a linguistic, psycholinguistic and sociolinguistic nature, in text and textuality.(12)

3.3.2.3 The Stratificational Model (Gutwinski)

Gutwinski proposes a linguistic framework for the study of cohesion in literary texts based on the stratificational theory of linguistics as described by Lamb (1966).(13) He acknowledges in addition indebtedness to Halliday's systemic grammar and, to his conception of cohesion, particularly in his 1964 and 1966 studies. However, he departs from the Hallidayan model because of what he believes is a lack of explicitness in developing "a semology or even a fully worked-out tactic for its upper stratum (lexical hierarchy or lexis)" (Gutwinski 1976 p.23), a problem that he also associates with tagmemics. In his view a model of semologic structure has to underlie any serious attempt to handle connected discourse. Stratificational theory, he claims, is adopted as the theoretical framework because of its capability of recognising and developing strata, one of which is semology while the others are phonology and grammar. Although cohesion as a linguistic phenomenon belong to the grammatic stratum, a truly comprehensive description can only be made by stating it in terms of the units of, and the relations obtaining on, the semologic stratum.

Gutwinski, however, concedes that the structure of the semologic stratum "is not directly observable since it is not represented directly in the grammar and even less so in the phonology of the language" (ibid p.25). But then he claims that semologic structure "finds its manifestation in the relatively shallower structure of the grammar and is still recoverable from it" (op.cit.).

Accordingly, cohesion as a term is employed for the relations that exist among the sentences and clauses of a text. These relations, which in Gutwinski's view, occur on the grammatic stratum, are signalled by certain grammatical and lexical features reflecting discourse structure on a higher, semologic stratum. These features account for textual connectivity of sentences and clauses. "They do not by themselves constitute cohesion but they mark which clauses and sentences are related and in what manner" (ibid p.26). It is this relatedness of clauses and sentences that constitutes the internal cohesion of a text.

A good understanding of cohesive relations in a text, Gutwinski believes, will help us in reconstructing the text's discourse structure. Since cohesion is established as a manifestation of discourse structure, it follows that a text, which is envisaged as a continuous discourse having structure, will display cohesion. Gutwinski asserts further that this cohesion "may differ in kind and degree depending on how it is structured on the semologic stratum and what options have been chosen while realising the semologic structure on the grammatic structure". Accordingly, he concludes, texts may exhibit strong or weak cohesion, but there will be no text that does not manifest cohesion.

These assumptions form the background of Gutwinski's description of cohesion. His use of Gleason's (1968) model (see 2.4.2 above) forms the backbone of his account of the semologic structure of cohesion. This model provides a reticulum or network of semologic units, generated by the semotactics (the tactics of the semologic stratum) in a way that ensures the generation of an event-line.(14) Sentences and clauses are then generated in a manner that conforms to the requirements of the semologic structure. Gutwinski identifies two types of tactics involved in this generation: semologic and grammatic tactics, which explain why certain grammatical choices that are feasible in formulating an isolated sentence may not be so for the same sentence if it occurs as part of the configuration of a text. The choice in the latter case, is determined by the semologic reticulum. Conversely, one can determine the semologic structure (or part of it), and consequently the discourse structure, by examining closely the clauses of a text and establishing the type of grammatical choices that were made in their generation.

Before Gutwinski proceeds to the discussion of the typology of cohesive features, he makes a note of what he calls "a cohesive factor", that is the order in which sentences follow one another in a text. The importance of this factor is represented by the imposition of an interpretation to a conglomeration of sentences by virtue of their appearing in a certain order together. If no interpretation is feasible, that sequence of sentences is not a text. "Order" is then a cohesive factor that, either by itself or in combination with other factors, indicate the kind of cohesive relations that obtain between sentences and clauses.

The cohesive features that Gutwinski postulates, and later investigates in literary samples from Henry James and Hemingway, are categorised, following Halliday (1964), into two main classes: grammatical and lexical. However, his listing differs from that of Halliday in the manner of classification and presentation, and in some detail. Gutwinski gives two reasons to justify these differences. First, his present listing "will achieve a greater consistency with the theory of cohesion presented". Secondly, it will "provide a workable descriptive framework for the examination of texts for the purpose of establishing their cohesive features" (p.59).

A. Grammatical

1. Anaphora and cataphora

(a) pronouns

(i) personal pronouns, e.g. he, him, she, it, they

(ii) demonstrative pronouns: e.g. this, these, that, those

(iii) relative pronouns: e.g. who, which, that, whom, whose

(b) determiners: e.g. the, this, these, that, those

(c) personal possessives, e.g. his, its, their

(d) substitutes

(i) verbal (do)

(ii) nominal (one)

(iii) partial

(e) adverbs, e.g. there, then

(f) submodifiers, e.g. such, so

2. Coordination and subordination

(a) connectors

3. Enation and agnation

(a) enate sentences

(b) agnate sentences

B. Lexical

1. Repetition of item

2. Occurrence of synonym or item formed on same root

3. Occurrence of item from same lexical set (co-occurrence group).

In his classification, Gutwinski drops a distinction that Halliday (1964) adopts (cf. Hasan 1968 and Halliday and Hasan 1976) between structural and non-structural categories. Relations between clauses are not studied unless they are signalled by connectors. Enation and agnation (see below) are not covered by structural cohesion since they refer to inter-sentence relations.

A few explanatory notes on the main categories are in order (we postpone the discussion of subordination, coordination and connectors to the next chapter when we review some textual treatments of connectives). Anaphora in Gutwinski's classification has been broadened to include not only cataphora but substitution as well, a point of departure from the functional-systemic classification. The inclusion of substitution is justified on the grounds that it represents essentially the same cohesive relation as anaphora. In this, Gutwinski follows Hockett's conception of anaphora as discussed in his 1958 work. Substitution is classified into three parts: nominal (through the use of one, ones), verbal (through the use of do and its inflections) and partial which subsumes the phenomenon of ellipsis and its various manifestations. Thus the category of anaphora has a wide coverage.

The terms enation and agnation were originally introduced by Gleason (1965). Enation obtains when two sentences have identical structures, that is, "if the elements (say, words) at equivalent places in the sentences are of the same classes, and if constructions in which they occur are the same" (Gleason 1965: p.199). Often enation functions cohesively in conjunction with lexical cohesion and may be reinforced by other features of grammatical cohesion. Agnation subsumes relations that are opposite and complementary to enation.(15) The use of an agnate structure is considered as a cohesive factor in a certain stretch of text since it is dictated by the previous structures in that stretch for achieving a particular function: linking, summarising or resumptive.

Lexical cohesion includes repetition across sentence boundaries, which helps relate various sentences in a text. It subsumes occurrence of the same lexical items or of synonyms or other members of the same co-occurrence class (lexical sets) in sentences that are in close proximity. Gutwinski admits that the determination of how distant sentences can be and still display lexical cohesion is an empirical question, related to such considerations as memory span.

Gutwinski, in concluding his discussion of cohesive features, points out that there are other linguistic phenomena which ought to be considered in a full study of cohesion. These include modality, sequence of tenses, use of certain adjectives, comparatives and adverbials, repetition of whole clauses or parts of them and of entire paragraphs (the latter can occur in works of literature). This creates a motivation for further work in this area, both in describing familiar features of cohesion or in discovering new ones.

3.3.2.4 The Procedural/Relational Model (Beaugrande and Dressler)

This exposition of cohesion is explicitly stated in Beaugrande and Dressler (1981), although most of it is introduced in Beaugrande (1980) under different headings. The account starts with a number of assumptions that make up the backbone of the procedural/relational approach.

One main assumption concerns the function of the language system of syntax. The most obvious illustration of this function is the imposition of organisational patterns of various size and complexity upon the surface text (defined as the presented configuration of words). The major units of syntax are patterns of well-marked dependencies: the phrase, the clause, and the sentence, all capable of being utilised in a short as well as long span of time and processing resources. Accordingly, cohesion has to be procedurally postulated within two perspectives. The first views cohesion as sequential connectivity between elements <u>within</u> phrases, clauses and sentence, while the second concerns connectivity within stretches of text of longer range. The two perspectives are closely related to each other since "each occurrence is instrumental in ACCESSING at least some other occurrences" (Beaugrande and Dressler 1981 p.48, their emphasis). This assumption is the core of the concept of cohesion and the two perspectives point out to the mechanisms by which it is elaborated. This is outlined below.

A. Short-range cohesion

The assumptions that underlie cohesion of this type draw support from relational grammar (particularly as discussed in Cole and Sadock (eds) 1977, Perlmutter and Postal 1978, and Johnson and Postal 1980). One basic feature of this kind of grammar is its focus on the connectivity of grammatical occurrences in surface structure. The motivation behind the theorisation hinges on the argument that text perception must evolve in real time, a factor that explains why people, instead of waiting for sentence completion to build a derivational tree (as derivational models of grammar imply), actually start connecting perceived elements as soon as possible.

Within this framework, Beaugrande and Dressler (1981) view the basic phrases and clauses of English as configurations of links between pairs of elements, many of them having further linkage. A related question that imposes itself concerns the manner and order in which these links are created. To answer this and other related questions on cohesion, Beaugrande (1980) and Beaugrande and Dressler adopt a relational kind of syntax that is designated "augmented transition network", a formalism that has been found to perform the best in the simulation of language processing on computers.

This type of syntax relies heavily on the recognition and enumeration of grammatical dependencies that obtain between elements in phrases, clauses or sentences. The network is, in fact, a configuration of "nodes" (in this case "grammatical states") connected by "links" (in this case "grammatical dependencies"). The processor traverses the links to access the nodes, making the data at the nodes active and current. This operation is identical to a process of problem-solving, whereby a hypothesis is tested concerning the typology of dependency between the nodes. The data at the nodes can determine, and therefore should be treated as an "instruction" about, the preferential or probable links that can be tested next. Thus the types of links are limited through avoidance of blocked pathways where the probability for a failure in traversing the next node is higher than that of success. It is a simple form of means-end analysis where the processor focuses on the main differences between the first point (the initial state) and the final point (the goal state).

To exemplify the working of the network, Beaugrande and Dressler use a fragment from a school reader (1981 p.1 for a quotation of this fragment). The first sentence reads:

A great black and yellow rocket stood in the desert.

Beaugrande and Dressler then outline the idealised sequence of operation when the systemic processor advances from one state to another (see, in particular, their illustrative figure on page 51 of their volume). On registering the first micro-state, in this instance the determiner "a", the processor is able to recognise the macro-state of noun phrase. Each macro-state is capable of limiting the number and typology of the full range of probable occurrences. The macro-state has a "control centre" (for instance the head in a noun phrase or the verb in a verb phrase, etc.) which manifests the heaviest linkage to other states. Accordingly, the highest priority of the processor, upon entering the noun phrase macro-state in the example above, is to discover the head. When this hypothesis fails, the processor revises it in favour of the next hypothesis in the priority list, that of modifier. The hypothesis succeeds and the processor then postulates the next state (S3) to be the head, and so on. When "and" is encountered, the processor predicts that a) the next state (S4) is most probably of the same type as the previous one (S3), and b) it is, in addition, probably the last of its type in the sequence. Hence, a simple "recursion" of the micro-state "modifier" is performed. The processor then succeeds in finding the head (S5) "rocket", which is the control centre of this macro-state.

Beaugrande (1980) and Beaugrande and Dressler (1981), in order to understand the procedural ordering of the operations that the processor performs, view processing in another perspective, summarised in terms of "stacking". This implies that each element is picked up and placed on top of a "hold stack" (a concept that Beaugrande and Dressler borrow from Rumelhart 1977). This refers to the active list of working elements to be integrated into a connected structure. In a "pushdown stack" each entry goes to the top of the stack and pushes the rest one notch down. When the control centre, the head in the example above, reaches the top of the stack, the stack is cleared in reverse order. This means that the last link is established first in the network and so on until the first link is set up. As a result, a network is built that shows the grammatical dependencies of the macro-state "noun phrase".

The rest of the sample is processed in the same manner. The processor will construct the verb phrase network. This macro-state is registered upon encountering the verb "stood". Since this is already the head, the processor would then search for a modifier. The search is augmented by anticipating not one class but subclasses of modifiers, e.g. adverb vs. prepositional phrases. If the adverb is hypothesised as the current preference, the processor will advance to test this hypothesis. Failure to establish this link causes the processor to retract and test the hypothesis of a prepositional phrase, a macro-state within the overall verb-phrase macro-state. The sub-goal that is set up is to find the head of the phrase ("desert"), which is identified after the determiner "a".

The cohesion within this sentence is expressed in terms of a labelled transition network where the nodes are the grammatical states and the links are the dependencies. The network, as has been demonstrated, is constructed in real time by making "transitions" from one node to the next, an operation that requires specifying or discovering the relation between the current node and its successor. Beaugrande (1980 pp.47-8) suggests the following list of link types for labelling transitions in actualised networks of grammatical dependencies. In each link type, the control centre is labelled first; a) verb-to-subject, b) verb-to-direct object, c) verb-to-indirect object, d) verb-to-modifier, e) verb-to-auxiliary, f) verb-to-dummy, g) head-to-modifier, h) modifier-to-modifier, i) head-to-determiner, j)component-to-component, k) junction.

Beaugrande admits that his list is not intended to be definitive. One might argue for a more comprehensive list, depending on the extent of thoroughness of syntactic processing postulated. However, this list serves to designate the current use of elements and helps in conjunction with predictions, preferential ordering of hypotheses, and hypothesis-testing to reduce the enormous amount of searching and combining required.

B. Long-range Cohesion

The previous section has outlined that in closely-knit units such as phrases, clauses and sentences, cohesion is sustained by fitting elements into short-range grammatical dependencies. In long-range stretches of text, there are devices for exhibiting "how already used structures and patterns can be re-used, modified, or compacted" (Beaugrande and Dressler 1981 p.49). These devices are: a) recurrence, b) parallelism, c) paraphrase, d) use of pro-forms, e) ellipsis, f) tense and aspect, g) junction, h) functional sentence perspective, i) intonation.

Two points are raised concerning the function of these devices.

1. These devices sustain cohesion by achieving repetition, substitution, omission and signalling relationships.

2. The devices in performing their cohesive role are less

obligatory than those which serve for closely-knit units. Missing elements in the latter case create a more noticeable disturbance within, or in the vicinity of, the units (phrases, clauses or sentences). Thus long-range cohesive devices contribute to efficiency rather than satisfy grammatical obligations.

Recurrence is a direct repetition of elements. The most obvious type of recurrence is that of lexical element, i.e. repetition of the same words or expressions. As a cohesive device, it is usually kept within limits since unduly frequent recurrence of items tends to lower informativity. However, recurrence is prominently used to a) assert or affirm one's viewpoint, b) convey surprise at occurrences that seem to conflict with one's viewpoint, c) express repudiation (cf. Halliday and Hasan 1976), i.e. rejecting some material stated (or implied) in the previous discourse, d) express the need to overcome irrelevant interruptions and get on with a statement (see text in Beaugrande and Dressler 1981 p.55), e) to express instances of iconicity, i.e. an outward resemblance between surface expressions and their content, particularly in poetic texts.

Other forms of recurrence are partial recurrence, parallelism and paraphrase. Partial recurrence refers to using the same basic word-components but in a different word class (cf. the device of polyptocon in classical rhetoric). "In this fashion, an already activated concept can be re-used while its expression is adapted to various settings" (ibid p.56). Parallelism entails re-using surface formats but filling them with different content, while paraphrase, on the other hand, is repetition of content with a change of expression (ibid pp.57-9 for examples).(16)

Further cohesive devices are used to compact (i.e. shorten and simplify) the surface text (even though there is a relative loss of determinacy). One obvious device is the use of pro-forms: "economical, short words empty of their own particular content, which can stand in the surface text in place of more determinate, content-activating expressions" (ibid, p.60). The best-known pro-forms are the pronouns which function as co-referents (ie, they share reference) to nouns or noun phrases. Co-reference is achieved either through the use of anaphora or cataphora (cf. Halliday and Hasan 1976).

There are other elements besides pronouns that Beaugrande and Dressler correlate with pro-forms. These include pro-verbs and pro-modifiers. The function of a pro-verb is performed by the verb "do" to "keep current the content of a more determinate verb or verb phrase" (ibid p.62). In this function, the verb "do" can co-refer with a considerable block of content.

The function of a "pro-modifier" is achieved by "so" and "such". These can stand for whatever modifiers connected to the verb in the original verb phrase. "So" can even stand for a whole clause (achieving "clausal substitution" in Halliday and Hasan 1976), thus signalling that the content of the clauses is to be kept active and current.

Textual compactness is also achieved by ellipsis. In the procedural approach advocated, "ellipsis is present only when text processing involves an apperceptible discontinuity of the surface text" (ibid, p.67 their emphasis). Typically, ellipsis operates through a sharing of structural components among clauses of the surface text. This is usually performed via an anaphoric function: the complete structures occur before the elliptical one. But the distance between the two must be kept within limits, otherwise the elided structure will be hard to recover or determine, and savings are lost on search and matching operations.

Cohesion is further supported by tense and aspect. These categories are realised differently in various languages (cf. Appendix 1). Accordingly, each language has its own means of distinguishing a) past, present and future time, b) continuity vs. single points, c) antecedent vs. subsequent, d) finished vs. unfinished.

The variety of means available in languages for expressing tense and aspect is a strong indication of the complexity and subjectivity involved in the organisation of time in the textual world. Even within the same language, an event can be expressed in different perspectives, for instance whether the event is seen as a closed unit at a single point in time, a multi-part unit extending over an unbounded expanse of time or a multi-part unit with defined time boundaries. Cohesion is sustained through viewing text-world events and situations as related. Where there are gaps, a process of updating is employed to indicate how the text-world is evolving.

A special aspect of cohesion which represents "an interaction between syntax, informativity and communicative settings" is exhibited in functional sentence perspective. Simply stated, it refers to the correlation between priorities of knowledge or informativity and the arrangements of words in clauses and sentences. In other words, the positions in which content materials are placed within stretches of clauses and sentences are suggestive of organisation according to priorities and degrees of informativity. A text producer tends to create a point of orientation before presenting new or more specific content material; a tactic that creates focus on crucial elements. Accordingly, informativity tends to rise towards the end of a clause or sentence. The cohesive effect of this aspect results when the sequencing of surface text gives signals about the shared knowledge to be manipulated during a given stage of the communicative interaction. For example, "due to the strategic usefulness of presenting known material first, the subjects of English sentences are often, though certainly not always, expressions (re)activating established or predictable content... The latter stretch of the predicate is, in turn, especially serviceable for creating focus" (ibid p.76, their emphasis).

A subsidiary cohesive system that is relevant to spoken texts only is that of intonation. Beaugrande and Dressler adopt the account of intonation in discourse proposed by Brazil (1975) (see also Brazil 1983, 1985 and Coulthard and Brazil 1982). Basically Brazil adopts Halliday's (1967) "tones" but re-names them to suggest the kinds of discourse actions involved. The "tone" is the rising or falling tendency of a "tone group" (defined as a stretch of text uttered as a unit).(17) The basic choice is between a "falling" tone and a falling-rising one. The first one is normally suggestive of the discourse action of "informing" (or proclaiming) while the second is usually associated with "invoking" (or referring). "Informing" is accomplished when the speaker presents predominantly new, unexpected, corrective or contrastive material, while "invoking" is done when the speaker presents predominantly known or expected material (see examples in Brazil 1975 p.6).

In addition, Brazil (ibid p.7) identifies two "marked" or intensified options pointing to an extra measure of speaker's involvement. The first is an intensified information action usually associated with a rise-fall, while the second is an intensified invoking action usually having a simple rising tone. Finally, Brazil identifies a low rising tone with neutrality, i.e. avoiding commitment to any one type of discourse action. This basic scheme is combined with a differentiation of "keys" to refer to types of pitch.

In general, intonation as a cohesive system is viewed as "the imposition of characteristic audible contours of tone and key upon texts in discourse, providing major cues about expectations, attitudes, intentions and reactions (Beaugrande and Dressler 1981 p.80).

This exposition of text cohesion and the explication of the typology of cohesive devices are, as Beaugrande and Dressler admit, never complete or exhaustive. The notion of "text cohesion" is regarded as substantially broad. There are two factors that support this view. One concerns the "operationalisation" of syntactic or grammatical structures as configurations utilised in real time, and the other refers to the "interaction" of syntax or grammar with the various other factors of textuality (op.cit). Thus Beaugrande and Dressler leave the door ajar for new conception and further extensions of the issues they have developed within their model.

3.4 Cohesion: A Synthesis of Views

3.4.1 Preliminaries

The notion of cohesion has, as our survey has shown, been defined from a number of perspectives. This is due partly to the intractability of the notion itself, but mainly to the differences in the analysts' persuasion, the analytical objectives and the material subjected to the analysis (see Hatim 1981 p.330 for similar comments). Therefore, in order to arrive at a conception that can serve as a working basis for exploring the cohesive role of connectives, we have to view cohesion from a perspective that is wide enough to accommodate some seemingly conflicting features of cohesion. The purpose of this and the next sections is to impose some uniformity on the diversity of views. This is achieved by first holding to the definition of text we have worked out and to the explication of textuality and its principles.

3.4.2 Characterising Features

In general, cohesion refers to the range of possibilities that exist for linking parts of text (of various size and complexity) together. But in order to gain an insight into this phenomenon, we have to view it from two perspectives: cohesion as a relation and cohesion as a process. Viewed from the second perspective, cohesion refers to the arrangement of text constituents into a working order that secures sequentiality. This involves the organisation of text constituents in a usable format (usually involving a linear modality, that is textual sequentiality) in order to serve the purpose of communication. In text actualisation, the issue of text organisation is not a trivial one and has therefore to be examined from a detailed standpoint. This will be discussed in the next section under "text linearisation".

Viewed from the first perspective, cohesion refers to a configuration of relations that interact cumulatively to enable a passage (spoken or written) to function as a text. On the surface level, these relations reflect dependencies among text constituents of various ranks. Normally two types of such relations are identified: the first concerns short-range dependencies that exist within the clause and that tie up the main constituents into a communicatively meaningful unit. Such relations are predominantly structural and can, perhaps, be best accounted for through sentence grammar (see Halliday 1985a for a brief outline of the clause.) Such relations are not the focus of this work.

The second type of relations concern dependencies of a longer range than the clause, that is dependencies that span the clause, clause-complex or paragraph boundaries. Normally, there exist text devices that explicitly signal the type of textual dependency and the kind of linkage involved. More specifically, these devices display how a textual dependency is to be interpreted and the textual range and direction where the interpretation is to be located.

Before we discuss the typology of relations that are explicitly signalled by cohesive devices, we would like to offer certain comments that characterise the concept of "cohesion" as we view it.

1. Cohesive relations have to be understood as a motivated manifestation of textuality that confirms and consolidates text organisation. These relations have an enabling function in the formation of text (and in relating it to its context) such that sequential connectivity is maintained and made recoverable.

2. Related to the previous point is the nature of the cohesive devices. These function textually not by virtue of their individual meaning or occurrence but through their imposition of a structure to the underlying connectivity of text-knowledge. One way of achieving this is by activating elements of knowledge, usually through a continual interaction of text-presented knowledge with previously stored knowledge.

3. Accordingly, cohesive relations as expressed by the various cohesive devices have to be interpreted before any continuity can be established. If the interpretation is suspended or not accessed at all (cases of stylistic deviance are excepted, cf. Enkvist 1973, 1978), a text will appear as a collection of sentences with holes and discontinuities that will disqualify it as a text or, to mention the least, disturb its stability.

4. A cancellation or disuse of cohesive devices will require a

considerable amount of lengthy restatement, repudiation and expansion to keep text-presented knowledge current and to help provide sufficient cues to update material. This will be achieved on the expense of a compact formatting of text and will burden the text unnecessarily.

5. On the other hand, too much reliance on cohesive devices, particularly ellipsis, reference or substitution, may outweigh or cancel any savings gained through an appropriate use of these devices, thus causing a loss in terms of text intelligibility. This is due to unnecessary expenditure of effort on pattern matching and retrieval that will occur beyond the point of diminishing returns.

6. Thus the distribution of cohesive devices and their manipulation for the achievement of effectiveness is regulated to a considerable extent by the principles of appropriateness and efficiency, which, in this case, are often language-specific (as this work intends to consolidate). In other words, cohesion is regulated differently in different languages since text stability is differently monitored by the regulative principles of textuality. It has been noted (cf. Gutwinski 1976, Halliday and Hasan 1976, Smith and Frawley 1983) that such variations exist in regards to different genres within one particular language. This issue, though we agree with it in principle, goes beyond the scope of this work and therefore cannot be investigated here.

3.5 Cohesion: Types of Cohesive Relations

The means of expressing cohesive relations can be grouped into two main categories:

1. Hard-core cohesive devices, including reference, substitution, ellipsis, use of connectives and lexical cohesion. These will be discussed in some detail within this section.

2. Soft-core or non-framing cohesive devices, including textual phenomena that interact and collaborate with the first type of devices to produce cohesive effects. These include iconic linkage (mainly parallelism and paraphrase), order, focus and emphasis. These will later be discussed as means for achieving text linearisation (see 3.6 below)

Since this work is concerned with connectives as cohesive signals, the rest of the devices will receive only marginal treatment below.

3.5.1 Reference

Reference concerns the use of alternative surface expressions to the same identity in a textual world. Although it is possible to extend this type of cohesion to include most other types (substitution, lexical cohesion, and even ellipsis), it is here restricted to the following:

1. Reference via proforms:
   a. Pronouns

Personal pronouns, eg	English:	he, him, she
	Arabic:	huwa, hu, hiya
Relative pronouns,	English:	who, which
	Arabic:	Allad̲ī, Allatī

b. Deictic English: the, this, these
Arabic: Al, hād̲ā, tilka

2. Reference via comparatives:

a. Identity English: same, other
Arabic: nafs, 'āxar

b. General similarity English: similar
Arabic: šabīh, mit̲l

c. Specific similarity

Proforms normally differ from their co-referring expressions in a number of ways (Beaugrande 1980 referring to Paduceva 1970 and Dressler 1972, cf. also Beaugrande and Dressler 1981):

1. Pro-forms have a wide range of potential application.

2. They are comparatively empty of inherent content. They derive their actualised content from their co-referring expressions.

3. Proforms are usually shorter. Dressler sees this fact is in agreement with Zipf's (1935) law that the more frequently a word is used, the shorter it tends to be or become.

4. Most proforms require a distinctive surface appearance. In both English and Arabic, pronouns maintain different forms for gender, number, person, and case.

The use of reference is distinguished in three axes of orientation:

1. Anaphora: this refers to using a proform to point back to the co-referring expression. Anaphora is the most common directionality in reference.

2. Cataphora: here the proform points forward to the co-referring expression.

3. Exophora: here the proforms do not point to co-referring expressions within the text, but to entities that lie in the situation. There are doubts (cf. Halliday and Hasan 1976) to whether exophora can be appropriately treated as cohesive since the actualised content of the proforms is not recoverable in the text itself.

The use of reference increases textual efficiency in that it creates compactness in surface structure since reference items are shorter than the expressions they replace, or, at least, express the relation in fewer words. This allows substantial savings in processing effort. At the same time reference assists in keeping content current in active storage. These gains, it should be noted, can sometimes suffer reduction, when, in certain cases, indeterminacies appear concerning the identification of the co-referring expressions. In these cases savings are wasted on search and matching operations. However, the co-operative nature of textuality (particularly as indicated by the constituent principles of intentionality and acceptability) is such that indeterminacies of reference are kept to the minimum. (For a discussion of reference and coherence see Garnham et al. 1982, see also Stenning 1978, Webber 1979, Kurson 1985, and the papers in Kreiman and Ojeda 1980).

3.5.2 Substitution

Like reference, substitution refers to the replacement of expressions of various sizes with simpler, shorter items. The difference between the two relations, as specified by Halliday and Hasan, is that "substitution is a relation in the wording rather than in the meaning" (p.88).[18] That is, substitution is a relation between linguistic items: words, phrases or clauses, a relation on the lexico-grammatical level; while reference is a relation between meanings, a relation on the semantic level.

Accordingly types of substitution are defined grammatically: the criterion being the grammatical function of the substitute item. We can distinguish three types of substitution: a) nominal, b) verbal and c) clausal.

The list of items that occur as substitutes is a short one and is predominantly proforms (Beaugrande and Dressler 1981):

Nominal: English: one, ones; same
Arabic: wāḥid; nafs [ḏāt, <ayn] (al-šay')
Verbal: English: do
Arabic: yaf<al
Clausal: English: so, not
Arabic: hākaḏā, kaḏālika

The primary meaning of substitution is anaphoric; it presupposes an element to which it is already structurally related. The role of substitution in textuality is manifested in achieving compactness and increasing efficiency.

3.5.3 Ellipsis

Another cohesive relation that contributes to compactness and efficiency is ellipsis. Normally, ellipsis functions via sharing structure components among clauses. Beaugrande (1980) indicates that ellipsis has been marked with controversy. He summarises the dispute as follows:

> "The surface structures in texts are often not so complete as they might be in the judgement of the investigator. Language theories with clearly drawn boundaries of grammatical or logical well-formedness necessarily proliferate the treatment of utterances as elliptical, according to the explicitness of the well-formed idealizations." (p.155; his emphasis)

An extreme standpoint (cf., for instance, Clark and Clark 1977, p.16) would view most utterances as elliptical. For instance, words linked with "and" or "or" are treated as elliptical coordinated clauses. It is questionable whether this view has any empirical reality. A more plausible method of determining the presence of ellipsis is suggested by Beaugrande and Dressler (1981) in their procedural/relational approach. Ellipsis, in their view, is present only when text processing involves an "appreciable discontinuity of the surface text" (p.67). Accordingly, a sequence is elliptical if it is noticeably discontinuous, and must be given connectivity through transfer from the preceding sequence (ellipsis is mainly anaphoric).

A detailed analysis of ellipsis as a cohesive relation (with particular reference to English) is put forward by Halliday and Hasan (1976). In their model ellipsis is treated under three headings: nominal, verbal and clausal ellipsis, referring respectively to ellipsis within the nominal group, verbal group and ellipsis that affects the verb and other elements in the clause.

Views on ellipsis in Arabic are contradictory. Williams 1984 (using a small set of data) claims that unlike English, Arabic tends to resist ellipsis. In his view, ellipsis is no more than a peripheral element of the grammatical system of Arabic. To illustrate that, Williams refers to the tendency of the Arabic verb to carry its subject (or a subject marker) with it whereas the English verb is able to shed it. Williams concludes that this basic difference renders the use of parallelistic forms more acceptable and more frequent in Arabic compared to English.

However, these claims are refuted by Al-Jabr (forthcoming, personal communication) who believes that Arabic has a similar tendency to using ellipsis to that in English. In his data (a bigger and more varied set than Williams's) there is no statistically significant difference between the number of occurrences of instances of ellipsis in English and Arabic.(19) This area of cohesion in Arabic needs further and more elaborate investigation.

But whatever view we adopt, the textual role of ellipsis should be duly recognised. Generally, ellipsis sustains cohesiveness and increases efficiency in all cases where the complete structure is recoverable, i.e. where there is presupposition in the structure of what is to be supplied. In text production, this necessitates keeping the distances between the presupposing (complete) and the presupposed (elided) components within reasonable limits. Efficiency will be affected, even damaged, if the distance is too big or if ellipsis is too heavy. In such cases all savings (usually made via utilising compact structures) will be cancelled by the demand to carry out intensive search and problem-solving.

3.5.4 Lexical Cohesion

The types of cohesive relations discussed above can be categorised as grammatical. They are expressed (except for ellipsis) through the use of synsemantic expressions (members of a closed system). Lexical cohesion, on the other hand, has no particular forms or items that function cohesively, but every autosemantic lexical item (members of an open set) may set up a cohesive relationship with another item (or other items) in the text. Thus lexical cohesion refers to the relation achieved via the selection of vocabulary; in particular it refers to the cohesive effect produced by the recurrence of surface expressions with the same or related conceptual content and reference.

The extent and variability of lexical cohesiveness is determined by the degree of lexical bonding between items in the text. In other words, it is the closeness or remoteness of the relationships between autosemantic items that specifies their textual cohesive force. There are three factors that regulate lexical cohesiveness (see Halliday and Hasan 1976 and Schneider 1979).

a) The relatedness of items in the linguistic system: This refers to the degree of proximity among items in the lexical system (and textual world). For instance, in the linguistic system there is a closer relationship between "dawn" and "morning" than between "dawn" and "day"; the latter, in turn, are more closely related than "dawn" and "month".

b) The relatedness of items in the text: This refers to proximity among the occurrences of items in text. Proximity here concerns the distance separating one item from another in terms of number of words, clauses or sentences in between.

c) The overall frequency of items in the linguistic system: Items of high frequency (e.g. "good", "take") or items that enter with equal readiness with words of every possible range of lexical meaning (e.g. "man", "way") have little cohesive force.

Accordingly, cohesion is strong if the lexical items are identical or synonymous, occur in close proximity (same sentence, for instance) and are of low relative frequency in the system of the language. Conversely, cohesion is weak if the items are remotely related (i.e. share few common elements of conceptual content), occur far apart in the text and are of very high frequency.

Lexical items that have cohesive force can manifest one of the following type of relationship to one another:

1. Identity (of referent):

    a. repetition: e.g. tourists - tourists

    b. synonym tourists - visitors

    c. hyperonym tourists - holiday-makers

    d. general term tourists - people

2. Non-Identity (of referent):

    a. ordered sequence spring - summer - autumn - winter

b.	antonyms	virtue - vice
c.	hyponym	lakes - ponds
d.	specific term	place - corner

One more type of lexical cohesion with a strong referential force is when a lexical item summarises a bigger chunk of text than a word (i.e. a clause, sentence or paragraph) eg, "The miners opted for a strike. This move cost them months of suffering". The word "move" refers to (summarises and labels) the whole of the previous sentence. Such items are also cohesive in the sense that they reflect the judgement and standpoint of the text producer, as when the word "move" is replaced by "solution" or "madness" (further analysis of the types and range of lexical cohesion goes beyond this work).

3.6 Cohesion as a Process of Textual Linearisation

3.6.1 Preliminaries

In this section a brief treatment is offered for the concept of cohesion as a process of linearisation. A detailed analysis goes beyond the scope of this work. Furthermore, a number of issues related to this topic are, in the present state of the art, still in their experimental stage and therefore the various views concerning the process are not conclusive. Nor shall we try to offer any in this exposition.

By "linearisation" we refer to the imposition of surface linearity upon underlying relational configurations. "Linearity", a key word in text utilisation, refers to the strictly serial formatting of text. We shall discuss below two aspects of the process of linearisation: a) the types of linear orientation, and b) the regulative principles of linearisation.

3.6.2 Types of Linear Orientations

Text as a manifestation of an actualised system is presented in a linear manner. Without attempting to state the obvious, linearity, in its most primitive form, is displayed in the fashion in which a sequence of visual marks or images on paper (in the case of a written text), or a sequence of sounds with intervening pauses (in the case of a spoken text).

Sequential elements of a spoken text, it has often been stated (cf., for instance, Leech and Short 1981), occur linearly in time whereas those of a written text occur linearly in space. Put in a different way, sounds are more temporally oriented and therefore emphasis in spoken texts lies on temporal succession, while images are more spatially oriented and therefore emphasis in written texts lies on spatial succession (cf. Nystrand 1982c). The contrast results partly from the acoustic versus visual modalities (on this topic see O'Connor and Hermelin 1978). The acoustic modality offers a continuous representation, whereas the visual (i.e. graphological) modality offers a discrete one.

The components of the visual configuration are simultaneously available, whereas those of the acoustic one are successively available (cf. LaBerge and Samuels 1974; Dogette and Richards 1975).[20] It follows that the so-called "tyranny of succession"

(Leech and Short 1981 p.211) is predominantly noticeable in the ephemeral medium of spoken texts, since a spoken element, once uttered, cannot be erased or recalled.(21) In a written text, the permanence of the graphic medium permits amendments during the process of actualisation: for instance, re-editing by the producer and re-reading by the recipient. However, in both types of text, utilisation occurs in a fixed order. The text, whether actualised by the producer or recipient, is not a static object, but a dynamic phenomenon.

3.6.3 Linearity and Text Actualisation

The linearity of the occurrence and association of language elements in a text can be regarded as one aspect of the systematic linearity of human action (cf. Beaugrande 1980 referring to Lashley 1951). The human being possesses a faculty for linear activities (cf. Piaget 1976, Jaffee 1977) of which linguistic ordering of elements (syntactic or semantic) would be considered as just one special instance.

A significant requirement of a general theory of linear action is the correlation of the linear modalities of speech and writing with the levels and phases of actualisation. Beaugrande (1980, 1984) recognises four phases:

a. The goal-planning phase: In this phase pathways of actions are set up that might lead to a goal.

b. The ideation phase: In this phase, conceptual configurations are created that act as control centres for working with text content.

c. The conceptual development phase: This is when ideas are specified, enriched and interrelated.

d. The expression phase: In this phase concepts are assigned natural language expression.

An important characteristic with each phase is reflected in the extent of choice available. A set of options exist in each phase and are arranged in non-linear configurations. These options form the central resources obtained for the actualisation of text.

3.6.4 Regulative Principles of Linearisation

Text linearity is regulated by a number of general principles which relate processing capacities and motivations to the surface text. The principles outlined below are relevant to the model developed by Beaugrande (1980, 1984) and (loosely) to the discussion in Leech and Short (1981), Nash (1980), Jordan (1984) and Hoey (1979, 1983). These are:

1. Grouping
2. Sequencing
3. Salience

These principles regulate cohesion by allowing the process of actualisation to navigate freely within the constituents of the text. This point will be elucidated below in our brief summary of the role of each of these principles in sustaining cohesion in the text. Further and more detailed analysis goes beyond the scope of this work. (We shall refer to these principles when discussing the textual role of connectives).

3.6.4.1 Grouping

This principle concerns how constituents enter into an arrangement that is both meaningful and effectively compatible with text producer's goals. Text arrangement requires a set of decisions on the type and number of steps and their serialisation in a plan. Hence, arrangement is relevant to the order in which sequences are combined for conveying maximum informativity, to the relationships that obtain once the order of sequences has been set, and to the demand for creating prominence, emphasis, focus and contrast (Nash 1980, Werth 1984).

One portentous aspect of grouping is the size of sequence. Normally, constituents are grouped into sequences of manageable proportion. A text (particularly an English one) that consists of one long sequence can burden utilisation by requiring more text processing. On the other hand, processing navigates more freely when a text exhibits some degree of segmentation. Sequence size (particularly word and sentence lengths) are considered a possible indication of stylistic variation and therefore can assist in resolving disputed authorship.

Another important aspect of grouping and text arrangement is the lay-out (Nash 1980). Text sequences form units, and these in turn fall within bigger units. A written text is marked off by disjunctives, insets, parentheses, indentations, etc. Lay-out can be a possible indication of stylistic preferences. Moreover, it may reflect some rhetorical pressure imposed by the text. Nash (1980), for instance, states that to carry out a sense of phrasing and intonation, a text producer can arrange his text into rather short sequences and lay them on the page in such a manner that visual spacing reflects oral timing. A different lay-out, for instance serialised paragraphing indicated by numerals, sets the information in an optimum order so that there is no overflow between one set of information and another. This renders the convention useful for the presentation of various kinds of official material (directives, contracts, brochures, public notices, regulations, etc.).

3.6.4.2 Sequencing

This principle concerns activities and procedures, the role of which is to arrange text components (constituents of various sizes) into a working order. We would perhaps understand the textual force of sequencing better if we look at four of its aspects: juxtaposition, regression, progression and pause. The overall effect is to regulate the flow of the text constituents and integrate their parts.

1. Juxtaposition

This refers to the placement of text components next to each other. Juxtaposition is the essence of linearisation. Clauses (and bigger constituents) are usually juxtaposed according to certain ordering strategies that attempt to implement steps in a text plan. Normally, these strategies "reflect the extent to which they make text processing and storage easier: the mind does not have to strain itself by searching for an unconventional organizational mode" (Beaugrande and Dressler p. 122).

One strategy, particularly favoured in narrative, is the adoption of a temporal ordering that exhibits a chronological sequence (Leech and Short 1981 p.236). Another more general strategy with a wider application in various text types is to follow the pattern "situation-problem-solution-evaluation" (Hoey 1979, 1983, Jordan 1984). This strategy is related to the first one in the sense that it, too, follows a natural time-sequence. Other strategies, for instance in describing a scene or a room, involves moving from higher to lower, central to peripheral, mobile to stationary (Beaugrande 1980 pp.116, 204 referring to De Soto, London and Handel 1965, and Linde and Labov 1975).

In general, clear information-structuring involves appropriate selection of high priority information with sensible ordering. However, global patterns such as the "situation-problem-solution-evaluation", though channelling the development of the text-world, do not necessarily determine the format of juxtaposition. These patterns are supportive and are capable of providing an ordered progression of underlying events, actions, states, views, etc. But the writer is free to express those events, actions or views in some other order than their temporal and/or causal sequence, provided sufficient guidance and signalling is provided.

This is combined with other aspects of sequencing which indicate how text components look backward or forward (giving a cyclical impression of organisation) or pause to signal a threshold of termination. These are outlined next.

2. Regression

This refers to the means in which a step or a series of steps in the organisation plan of a text, or a text constituent, is influenced by previous steps or constituents in the same text. As text constituents are grouped and combined into bigger sequences or blocs, current sequences are continually affected by those that had already been stated. In other words, the sequences that have already been introduced in the text have a bearing on the current organisation of text, particularly on the choices available for the text producer. For instance, a topic sentence at a particular paragraph narrows down the set of alternatives available and the manner in which the selected ones are actualised. Even "the selection of a first word has in greater or lesser degree committed the speaker to a particular construction or at least a set of alternative constructions" (Boomer 1965 p.156, cf. also Beaugrande 1984).

Among the various types of textual regression is the one that figures predominantly in the means for maintaining cohesion of the text. This is affected by intentional reactivation of particular constituents, i.e. reusing of surface expressions at a later point (see discussion of cohesion above).

Another type is paraphrase. Here content remains constant but the surface realisation varies in expression, size, level of complexity, attitude, etc. Regression via paraphrase can enrich the introduction of the previous relevant sequence (by supporting, explaining, intensifying or re-directing it in a way that fits the plan of a text).

A third type of regression is parallelism, a term that denotes re-using a surface format with different components. Here new content is packaged into already activated structuring, thus intensifying focus on content and purpose.

3. Progression

This is the converse of regression and refers to all means in which text sequences are linearised in such a way that they can influence subsequent parts of a text.

Progression operates via anticipation. Some text constituents are readily and correctly anticipated. This is reflected, for instance, in the reader's ability to constantly anticipate: make hypotheses, and test them, about what is coming next. (cf. Goodman 1970, Al-Jubouri 1976). Current decisions can then be aligned with anticipated ones. However, too much progression conflicts with informativity (one of the principles of textuality): the higher the degree of correct anticipation of a constituent, the less informative it is.

Some types of progression coincide with those of regression. One is the anticipated recurrence of certain expressions in a subsequent sequence; a well-known case is cataphoric reference. Another type is anticipated paraphrase. The textual function of this type of progression is to help reduce the load that a prior sequence creates for utilisation. This function is achieved by allowing more focus on content and purpose and by activating phrasing of the content using different sets of expression. Parallelism can also be a type of progression. Here a constituent would anticipate the occurrence of a parallelistic form.

4. Pause

This aspect of sequencing refers to the requirement for retarding or suspending the linear sequence from time to time, particularly when such a text constituent as the sentence or paragraph comes to its threshold of termination. In this case it indicates that a step (or a group of steps) in a plan has been carried out and that the next step (or group of steps) is to be initiated. Sometimes a pause is created when a step (or a series of steps) is suspended for the purpose of presenting, for instance, a sub-plan. This technique is often employed for achieving a rhetorical effect.

3.6.4.3 Salience

This principle here refers to the extent to which text components can impose a striking impression on the senses. Salience is thus related to the manner in which the steps in a text plan can successfully achieve the texts producer's goals. Accordingly, different schemes of linearisation of text sequences can create different degrees of salience. The preference for a particular formatting is a rhetorical consideration, and can be taken as an indication of the rhetorical nature of a particular type of text or an imprint of an individual's style.

The operational effect of salience on the linear development of a text is probably better understood if two components of salience are examined, each separately. These are "prominence" and "involvement". Of course, more components can be envisaged and discussed, but that will perhaps be better left for future work. It should be remembered, however, that these components operate together to provide confirmation for the proposed type of linearity.

1. Prominence

This is defined by Halliday (1971 p.340) as "the general name for the phenomenon of linguistic highlighting, whereby some feature of language stands out in some way". To understand the function of this aspect of textual salience we must assume the existence of central components in conjunction with peripheral ones. The central components, spaced within the text sequences, carry the textual core: a central point of emphasis which carries the weight of information. Prominence has the textual role of distributing the flow of control within text sequences (of various types, sizes and complexity). It, therefore, represents the priority that certain components (constituents) have over others and can influence the way they are strung together in a linear order.

One facet of prominence that displays its impact on linearisation of sentences has been designated "functional sentence perspective". In many languages the early component of the sentence, the "theme" or "topic", is normally used to express the point of orientation (known, i.e. given, or predictable content) whereas central materials are mentioned late in the sentence. In some cases, a central component can be placed in the early stretch of the sentence. Such a format is normally reserved for emphasis and hence the structure displays more prominence.

Prominence is bivalent. Werth (1984) distinguishes "accent" and "contrast", two terms that he conflates under one term "focus". Accent marks new or revived information in a text and can be of two broad types: information accent and attention-accent. The first marks freshly-introduced semantic material. The second, i.e. attention-accent, provides for prominence previously-occurring material when, for some reason, it needs to be highlighted (for instance, for the purpose of renewing the present relevance of a piece of information).

Contrast, when associated with an item, indicates that a previous piece of information has a negative relationship with the item. In other words, contrast on an item implicitly makes it deny some other item in the discourse. Werth (ibid pp.131-147) discusses how various grammatical categories and lexical usages can be affected when they are associated with contrast. However, we are still short of research that explicitly shows a) how these factors can affect linear arrangement of text constituents, b) how they interrelate with other principles of linearity in consolidating text cohesion, and c) how various cohesive devices can collaborate with those factors of prominence to produce a total cohesive effect.

2. Involvement

In Beaugrande's view (1984 p.182), involvement designates "the intensity with which people are participating in discourse interaction". In written text, it is a reflection of the writer-reader relationship, particularly the type and extent of the role that each plays in the text. In general the writer tries to establish a role and a tone of voice in an effort to create a successful rapport (Nash 1980). This is usually reflected in details of language and style; one such is the manner in which material is linearised. According to Osgood and Bock (1977 p.93) if the text producer (the writer) is the centre of concern, salient elements migrate to the beginning of a sequence (a sentence). But if the text receiver (the reader) is the centre, salient elements come later. This early occurrence of salient material is indicative of the text producer's immediate reaction, while late occurrence is a better guideline for the text receiver. On this point Beaugrande (1984 p.186) comments (referring to studies conducted by Lunsford 1977, 1981, Odell 1977, and Maimon 1979) that unskilled writers often use non-strategic sentence formats that fail to guide focus, and that this tendency may be related to their strong ego-involvement vs. that of the audience (text receivers).

3.7 Conclusion

This chapter has been concerned mainly with a) outlining the notion of textuality: its nature and principles, and b) discussing the concept of cohesion, particularly as adopted in this study. We departed from a distinction between "virtual" and "actual" systems of language. A "virtual" system refers to unities of elements whose potential is not yet put to use. Such a system is "actualised" when it is put in use and the resulting artifact is itself an actual system. It follows that text is by nature "systemic", i.e. it functions as a system.

Viewed in a cybernetic perspective, a text displays "stability", a functional property of systems that is highly relevant for evaluating the efficiency of many systems. As a text is produced or received, the system is actualised and traverses a series of steps. The system regulates itself according to the demands of the context and to the requirements imposed by the various steps in the text plan, so that each change can be met with a suitable adjustment in organisation. The systems remain stable as long as they support utilisation and continuity. Hence text stability is heavily dependent on text continuity.

In a text, continuity is reflected by connectivities: sequential and conceptual connectivities as well as connectivity of planning. These represent the various domains of textuality and constitute the basis for text actualisation.

The notion of textuality refers to the quality that distinguishes text from non-text. It can be defined according to two sets of principles: constitutive and regulative. The constitutive principles include cohesion, coherence, informativity, intentionality, acceptability, situationality, and intertextuality. These determine, when available, that a presentation is a text. The regulative principles determine variations in the manner in which a text is actualised. They include: efficiency, effectiveness and appropriateness.

An examination of a few theoretically diverse models of cohesion has specified the same type of connecting links and suggested that such links are essential for the integrated and cohesive quality found in text. Anaphoric reference, ellipsis, substitution, logical connectives, for example are mentioned in these models. Accordingly, such mechanisms, with some necessary modification, have been adopted as a basis for exploring cohesion.

As a heuristic device, two facets of cohesion are distinguished and examined: cohesion as a set of relations and cohesion as text linearisation. The two aspects are interrelated. The first aspect represents cohesion as a configuration of relations that exist among text components. These relations can be explicitly expressed through a variety of "hard-core" devices. Reference concerns the use of alternative surface expression to the same identity in a textual world. Substitution, like reference, refers to replacement of expressions of various sizes with simpler, shorter items. Ellipsis is a phenomenon that functions via sharing structure components among clauses. These types of cohesive devices are grammatical in nature. Cohesion can also be marked by lexical items. Here cohesiveness is determined by the degree of lexical bonding between items in the text. The type of lexical relations are either that of identity between a pair (or more) of items (represented in repetition of the item or expressions of its synonym, hyponym, or a related general term), or a non-identity (ordered sequence including the presupposed item, or expression of antonym, some related hyponym or a specific term.

Cohesion as text linearisation concerns a textual phenomenon that interacts and collaborates with the cohesive devices of the first aspect of cohesion in order to produce cohesive effect. It refers to the imposition of surface linear order upon underlying relational configurations, and, hence, it concerns the way text is organised. A number of principles regulate text linearity. The first concerns grouping of text components (or constituents), i.e. how constituents enter into an arrangement that is both meaningful and compatible with the text producer's goals. The second principle refers to text sequencing, a phenomenon that textually functions via juxtaposition of constituents, regression (the degree to which a current text constituent is influenced by a previous one), progression (the extent to which a current constituent or text component affect later ones) and pause (a halt to indicate an end of a step). Salience refers to the extent to which a linear ordering can impose a striking impression on the senses. Two components of salience are examined: prominence and involvement. The first refers to the linguistic phenomenon of highlighting, where a feature stands out. The second refers to the way writer-reader relationships can impose a particular linear organisation of constituents.

To conclude, these proposals are consistent with textual treatments that focus on the dependence of one textual sequence upon another, such as the one intended in this project. Since text is linear in organisation, an association between text sequences must necessarily be manifested sequentially. Thus a clause (or a sentence, or any bigger sequence) relies on earlier or subsequent ones for the total cohesive effect. The mechanisms that explicitly relate sequences or mark their dependence can analytically be specified. The various relationships that obtain produce collectively the textual property - cohesion.

Footnotes to Chapter 3

(1) The stability of a system is based on its structure (defined in cybernetic terms as the interrelationship of elements of a system). With the increased development of its organisation, a system is not influenced by the external disturbances or incidents. A system is ultra-stable if it remains unaltered by the most extreme disturbances or changes of its surroundings. If disturbances exceed a particular level, then the ultra-stable system adopts another pattern of behaviour (see Muller 1964/1968 p.169, Lerner 1972 pp.46-47).)

(2) Illustrations of the regulatory operations required in this respect are provided in Beaugrande 1980 pp.17-18.

(3) These two terms are borrowed from Searle (1969) p.33ff, though not within this particular context. Cf. Beaugrande's (1980, 1984) use of the term "standards").

(4) A different classification of patterns of theme movement are suggested by Daneš (1970a, 1970b). In this classification, theme identification and theme movement are fused into a single taxonomy. The patterns are (Enkvist 1973 p.120-121): a) simple linear progression (in which the rheme of one sentence becomes the theme of the next), b) passages with run-through themes (a sequence of sentences with the same theme but different rhemes), c) progression of derived themes (there is one "Hypertheme" and several hyponymic "Teilthema", and d) the development of a split rheme (the themes of successive sentences are co-members of a concept forming the rheme of the initial sentence). Another treatment of theme in relation to English is given in Jones's 1977 study of expository prose.

(5) In this conviction, it will be meaningless to claim isomorphism for sentences just because they share some deep structure such as NP + VP.

(6) It seems that Halliday uses the terms interchangeably. "Discoursal" component (1968) becomes "textual" component (1970a) without any apparent alteration in the frame of reference.

(7) The network of options, in Halliday's views (cf. Halliday 1977), has the form "if a, then either b or c". Variants of this general form are postulated to include "if a, then either y or z and either m or n; if x, or if m, then either p or q; if both y and n, then either r or s or t" and so on.

(8) It should be mentioned that Hasan (1968) in discussing cohesive ties agrees that structural relations (as opposed to non-structural, which are the main concern of her study) are unifying factors. "Any linguistic item forming a single complete structure - any sentence, or clause, or group - will always be internally 'cohesive' precisely by virtue of its structure. If a text could contain only one sentence the concept of cohesion would not need to

be distinguished from that of structure, since the unity of such a text can be fully defined and stated in structural terms. This is indeed true of sentences which are complete texts.." (p.21). These views are reiterated in Halliday and Hasan (1976).

(9) Cf. arguments to this effect in Frankel (1977).

(10) Cf. a preliminary study of text deixis in Kurzon (1985). Kurzon believes that deixis are different from referential elements and that certain languages have two different sets of elements for deixis and for anaphora. English uses the same set of elements.

(11) However, Halliday and Hasan point to another distinction that results from the different nature of the two types of relation. "Because reference is basically a non-verbal relation, a reference item may point in any direction, and pointing to the preceding text is only one among the set of possibilities. Substitution on the other hand, being a verbal relation, is essentially confined to the text". (1976 p.90).

(12) Examples of such research are the studies conducted by Chapman (1979, 1983), Eiler (1979, 1983), Henderson (1979), Coleman (1980), Hartnett (1980), Johns (1980), Lieber (1980), Pritchard (1980), Stine (1980), Szwedek (1980), Tatilon (1980), Maynard (1982), Williams (1982), Been (1982), Smith and Frawley (1983), Rottweiler (1984), Stoddard (1984), Al-Jabr (Forthcoming) (cf. also Källgren 1978, Kittredge 1981, Gumperz 1984).

(13) The stratificational theory as developed by Gleason, Lamb and others views language as consisting of several systems, called "stratal systems", each of which is said to be associated with a "stratum of linguistic structure" (cf. Lamb 1966 p.1). The number of strata varies with the different postulation at the different stages of the development of the theory. Lamb suggests that all natural languages have at least four and some may have up to six. Gleason (1968) suggests three: phonology, grammar and semology. These systems interact with each other in complicated ways. The two important characteristics of the stratal systems are: a) each stratum has its own units (inventory) and its own syntax (tactics) specifying how these units can be arranged in structures, b) the relationship between strata is one of realisation or manifestation: units and structures of one stratum are not composed by those of lower stratum but only realised by them (see Gutwinski's summary 1976 p.36-53).

(14) As discussed in 2.4.2 above, Gleason's model is based on narrative discourse.

(15) Gleason (1965 p.202) defines agnation in the following way. "Pairs of sentences with the same major vocabulary items, but with different structures (generally shown by differences in arrangement, in accompanying function words, or other structure markers) are agnate if the relation in structure is regular and systematic, that is, if it can be stated in terms of general rules".

(16) See Beeston (1974), Al-Jubouri (1984) and Al-Jabr (Forthcoming) for an account of parallelism and paraphrase as forms of repetition in Arabic.

(17) For a good treatment of "intonation" and "tones" see Kingdon (1954), Halliday (1967c), Crystal (1969), and Lehiste (1970, 1975). For a brief and pedagogically oriented treatment see Al-Hamash and Al-Jubouri (1975) and the series of monographs that were inspired by that work and that the two authors produced for teacher-training courses in the period of 1976-1982.

(18) This indicates that the classification of cohesive relations into different types should not be treated as implying a rigid division. Many instances of cohesive relations occur on a borderline between two types and can therefore be interpreted as one or the other. For instance Beaugrande and Dressler include substitution in their discussion of reference via proforms, while Halliday and Hasan make a distinction between the two.

(19) It should be mentioned that Al-Jabr in his analysis overlooks the ellipsis of the subject pronoun labelled in Arabic grammar "damir mustatir".

(20) There is evidence, however, (cf. Das et al. 1975 p.82) that not all acoustic processing is temporal, nor all visual processing spatial. For instance, there is an argument (cf. Just and Carpenter 1980) that readers move through a text with characteristic timing as they fixate on successive words with their eyes. A further treatment of this topic goes beyond the scope of this study. What concerns us more is the role of connectives in sustaining linearity and maintaining textuality.

(21) An exception must be made where spoken text is recorded; this form of text creates and sustains a situation different from the original one where the text producer formulates his desired goals. We take such a form of text as indicative of data rather than information.

CHAPTER FOUR

Connectives

4.0 Perspective

This chapter is primarily concerned with the conception of connectives and the development of its role in text organisation. The concept of connectives will first be traced in logic and linguistics, thus providing a threshold for later theorisation. It will be displayed that models of logic as well as models of sentence grammar have failed to explain the textual role that connectives exercise in organising the text and the impact they exert over the finished product as a whole.

Theorisation departs from the assumption that the utilisation of text relies on the extent of proper choice made during the text production phases. Such choice is constrained, within a textual context by the typology, size and behavioural patterns of the connectives and the relationships they uphold for clausal sequencing as well as the range of connectivity they maintain for the establishment of textuality.

The treatment, though focusing on connectives in particular, is still rather general in the sense that no contrastive attempt is endeavoured at this stage of the study. We believe that while a contrastive textual treatment is the axis of this work, it is nevertheless essential to offer a foundation of theoretical conception on which later tasks will be established and formulated, including the discovery of categories of connectives, taxonomy of their textual range and the description of their quantitative properties. This treatment, we hope, will not conflict with the experimental nature of the project and the requirements of a bottom-up analysis and synthesis.

The chapter is then divided into three main sections. The first is an attempt at monitoring the conception of connectives as formulated in two types of pre-textual research. By the label "pre-textual" we here include linguistic as well as logical theorisation operating at, and so confined by, the sentence level. The general aim is to trace the operationality of these various formulations and assess their viability in investigating textual organisation. The second component of the Chapter examines the status of connectives within textual framework and reviews a number of textual analyses. The last section reflects on the previous two and offers the necessary foundational elements for our own conception of this phenomenon. The chapter ends with some conclusions that can also serve as preparatory considerations for the later stages of this work.

4.1 The Connective in Logic

4.1.1 Introduction

The discussion of connectives in logic receives its validity within the scope of this study from the established cooperation between logic and linguistic studies. Methods developed within formal logic to study the semantics of artificial languages have been applied to the semantics of natural language; and within text linguistics methods and techniques from logical theory have become increasingly common. One of the attempts to incorporate logical semantics within the discussion of textual issues, such as text, connection and coherence, was made by van Dijk (1977a and 1977b).

However, we do not claim any comprehensiveness. Our account of the connectives in propositional logic is concise and touches only the most relevant points. The aim is to unfold the disparities that exist between the role of connectives in logic and their role in natural language texts. And in order to restrict this account within a manageable limit, and thus keep it clear and succinct, we have intentionally excluded some issues that a more detailed discussion might have otherwise included. This outline is based on Suppes (1957), Mitchell (1962), Alexander (1969), Kreyche (1970), Ballard (1972), Zierer (1972), Beilin and Lust (1975), Allwood et al. (1977), Hodges (1977), Copi (1982).

4.1.2 Propositions, Statements and Sentences

To clarify this brief account of logical connectives, it will help to outline some relevant concepts. In line with the assumptions of a classical bivalent logic, propositions are defined as statements that are either true or false, and hence, they differ from questions, commands and exclamations. For while questions may be asked, commands given and exclamations uttered, only propositions can either be asserted or denied.

The term proposition is usually distinguished from the term sentence. The latter is a physical manifestation, tangible in terms of a beginning and end, and measurable in terms of number of words. A proposition refers to what a sentence may be uttered to assert.

It follows that a one-to-one correspondence between a sentence and a proposition is not necessary. Two different sentences may in the same context have the same meaning. For example, [4.1a] and [4.1b] below:

[4.1a] The director signed the document.

[4.1b] The document was signed by the director.

are two different sentences, since they exhibit a different word order and since the number of words are not the same. Yet, the two sentences have exactly the same meaning. Furthermore, a sentence is language-specific in the sense that it is always a sentence of a particular language, the language in which it is enunciated, whereas a proposition is not peculiar to a language in which it may be formulated. For example, [4.2a] and [4.2b] below:

[4.2a] The delegates have arrived.

[4.2b] Ḥaḍarat al-wufūdu.

are different, for they are in different languages: English and Arabic. Yet they express a single meaning, and they can be used to assert the same proposition.

Sometimes the same sentence can be used to express very different propositions, particularly when the contexts are different. The following sentence:

[4.3] The present British prime-minister is a lady.

could have been uttered in 1985 to make a true statement about Margaret Thatcher, but it would have been uttered in 1975 to make a false statement about Harold Wilson. The terms statement and proposition, though not exact synonyms, are often used in much the same sense in the context of logical investigation. (Note, however, that in 5.3 below we use the term proposition in a different way from its logical sense).

Other useful terms are argument, conclusion and premise. In the logician's sense an argument is any group of propositions of which one (the conclusion) is claimed to follow from the others (the premises), i.e. premises are regarded as providing support or grounds for the truth of the conclusion. It should be noted that no proposition by itself, in isolation, is either a premise or a conclusion. It is a premise where it occurs as an assumption in an argument, whereas it is a conclusion only when it follows from propositions assumed in that argument. Thus premise and conclusion can be regarded as relative terms.

4.1.3 Truth Values

It is customary in propositional logic to divide all statements (propositions) into two general categories, simple and compound. A compound proposition is one which can be analysed into at least two simple ones. A simple proposition cannot be further analysed in terms of other propositions. Thus the statement in [4.4] is simple while the one in [4.5] is compound.

[4.4] The beach is crowded.

[4.5] The beach is crowded and the sun is hot.

Since any proposition is either true or false, it follows that every proposition has a *truth value*, where the truth value of a true proposition is true, whereas the truth value of a false statement is false.

In the sense that simple propositions are not made up of other component propositions, they do not depend on other propositions for their truth value. Compound propositions may be divided into two different categories, according to whether or not the truth value of the compound proposition is determined by anything other than the truth values of its components. A compound proposition is said to be *truth-functional* if, and only if, its truth or falsity is a function of the truth or falsity of its component propositions. A component of a compound statement can be defined as a truth-functional component of it "provided that if the component is replaced in the compound by different statements having the same truth value as each other, the different compound statements produced by those replacements will also have the same truth values as each other". (Copi 1982, p.280, notes that other, rather more complicated, definitions have also been proposed). Compound propositions that are not truth-functional usually entail expressions of desire, belief, and the like. This type of statement is left out from this brief account.

4.1.4 *Connectives and Truth Relations*

In symbolic logic, connectives such as *and* and *or* (generally symbolised in the propositional calculus as the conjunction and disjunction operators [.] and [ˇ] respectively) form compound statements out of simple ones. By convention, a connective is truth-functional when it possesses the property of making the truth values of the simple statements it connects. Given any two component simple statements, e.g. statement p and statement q, linked by a logical connective, there are only four possible combinations to represent their truth or falsity.

1) p is true and q is true;
2) p is true and q is false;
3) p is false and q is true;
4) p is false and q is false.

These possibilities can be detailed in truth tables such as the one below (where T in the propositional calculus stands for "True" and F "False"):

	p	q
1)	T	T
2)	T	F
3)	F	T
4)	F	F

Thus the truth value of any compound statement is determined by i) the truth value of the component statements, ii) the values of the connectives. To understand the relation between these two factors, we have to examine the types of propositional connection, a task left to the next section.

4.1.5 Types of Connection

In logic, interest has traditionally been shown in only a few types of connection: conjunction (and), disjunction (or), implication (if ... then), and equivalence (if and only if). Included also among connectives is negation (not), which actually does not combine statements but operate on one statement at a time. These five types of connection are briefly discussed below.

4.1.5.1 Conjunction

Conjunction is performed via the connective *and* [.] to construct a compound statement which is true only if all its component statements (conjuncts) are true. If any conjunct is false, the compound statement, the conjunction, is also false. That is, the conjunction *p.q* is false when *p* is true and *q* is false, when p is false and *q* is true, and when *p* is false and *q* is false.(1) The truth table below displays the truth values of *p.q* for every possible combination of truth values of *p* and *q*:

p	*q*	*p.q*
T	T	T
T	F	F
F	T	F
F	F	F

4.1.5.2 Disjunction

Disjunction is formed by inserting the connective *or* [∨] between two statements. The two component statements so combined are called "disjuncts". Disjunction is intended to assert that at least one (and perhaps *both*) of the disjuncts is true. That is the disjunction [*p∨q*] is true when *p* is true and *q* is true, *p* is true and *q* is false, *p* is false and *q* is true. Only when *p* and *q* are false is the disjunction made false. These truth values are shown in the truth table below:

p	*q*	$p \vee q$
T	T	T
T	F	T
F	T	T
F	F	F

When the connective *or* is used to make such a compound assertion, *or* is being used in its *inclusion* sense. (We shall discuss later the difference between inclusive and exclusive disjunction).

4.1.5.3 *Implication*

An implication (also called a hypothetical, a conditional or an implicative statement) is a compound statement composed of two simple statements combined by the connective "if ... then" (symbolised $\supset$ or $\rightarrow$). The first part is the antecedent (or the implican), the second is the consequent (or the implicate).

Implication may be of several kinds. Examine examples 4.6-9.

[4.6] If you put any more salt in your lunch, then you are going to be very thirsty.

[4.7] If the two variables of a logical disjunction have the value "false", then the overall value of the proposition is also "false".

[4.8] If John is a bachelor, then John is unmarried.

[4.9] If Steve ever manages to pass the driving test, then I'll eat my hat.

In statement [4.6] the relation between the antecedent clause and the consequent is a causal nexus. In statement [4.7] the consequent is logically derived from its antecedent while in statement [4.8] the consequent follows from its antecedent by the very definition of the term "bachelor". In statement [4.9] the consequent does not follow from its antecedent either by logic or by definition. The speaker starts from the assumption that "Steve will never pass his driving test" and affirms it by the facetious promise to eat his own hat should he be proved wrong.

Although the four conditional statements are different in that each asserts a different type of implication, there is a common partial meaning. This is defined as $\underline{p} \rightarrow \underline{q}$, which is considered as an abbreviation of $\sim(p.\sim q)$. This partial meaning, symbolised by →, is a type of implication called material implication by logicians. The connective that → refers to is truth-functional, just like the symbols for conjunction and disjunction.

It has been stipulated that material implication is true wherever the antecedent is false or its consequent is true. As such, it is defined by the truth table.

$\underline{p}$	$\underline{q}$	$\underline{p \rightarrow q}$
T	T	T
T	F	F
F	T	T
F	F	T

4.1.5.4 Equivalence

Equivalence roughly corresponds to "if and only if". This connective, termed bi-conditional, establishes a double material implication, i.e. an implication in two directions at the same time: one going from antecedent to consequent ($p \rightarrow q$) and another going from consequent to antecedent ($q \rightarrow p$). The antecedent and the consequent are said to be materially equivalent (or equivalent in truth value) when they are either both true or both false. This notion is expressed by the symbol "≡". Material equivalence is a truth function and can be defined by the following truth table:

p	q	p ≡ q
T	T	T
T	F	F
F	T	T
F	T	T

4.1.5.5 Negation

The negation of any proposition is the denial of its truth value. If the truth value of the proposition is true, then the truth value of its negation is false. Conversely, if the truth value of the proposition is false, then the truth value of its negation is true. It is customary to use the curl, or tilde, "~" to symbolise negation. Thus, where p is any statement whatever, its negation is written ~p. It is obvious that "~" is a truth-functional operator (since it reverses the truth value of the original proposition).

It is helpful to notice that if a proposition is negated an even number of times, its truth value will not change. The following table makes this explicit:

p	~p	~~p	~~~p	~~~~p
T	F	T	F	T
F	T	F	T	F

It is clear in the above table, that the values of columns 3 and 5 are the same as the truth value of the original statement, p.

4.1.6 Logical Connectives vs. Natural Language Connectives

The comments so far made are intended to show that logical connectives are interpreted in truth-functional terms. Their role is to yield a truth value of compound propositions, computable from the truth values of the simple component statements, irrespective of the meaning of the connected proposition. This role does not run parallel to that of the connectives in natural language: it is a more restricted role. This will be clarified through a comparison between the use of and, or and if in logic and their use in natural language.

Logical and is a truth-functional connective; its role is therefore restricted to producing a truth-functional compound statement. Natural language and is multi-functional: the same connective may express different types of connection, and one type of connection may be expressed by various connectives (see Chapter 6 below). Examine the examples below:

[4.10a] Sally's neat and Sally's sweet.

[4.11a] The delegates arrived and the conference started.

[4.12a] John took a sleeping pill and fell asleep.

[4.13a] Move and you are a dead man.

In [4.10] and corresponds to the logical conjunctive connective, whereas statements [4.11-13] and can be intuitively paraphrased into (and) then, (and) so, if...then. It follows that natural language and may be used to express not only conjunction, but temporal, causal and conditional relations.

Furthermore, logical and is commutative, i.e. [p.q] is always equivalent to [q.p]. In natural language, while and in statement [4.10] is commutative, as in [4.10b], where the two component propositions change place, in [4.11-13] and is non-commutative, i.e. the order of the component propositions cannot be reversed without inflicting a change in meaning, as in [4.11b], or becoming unacceptable, as in [4.12b-13b]:

[4.10b] Sally's sweet and Sally's neat.

[4.11b] The conference started and the delegates arrived.

? [4.12b] John fell asleep and took a sleeping pill.

? [4.13b] Be a dead man and you move.

It follows that and in natural language may be considered of two meanings:

i) And(1) is equivalent to the logical connective, i.e. there may be symmetrical equivalence between statements; A and(1) B = B and(1) A.

ii) And(2) designates the condition in which symmetry does not exist: A and(2) B $\neq$ B and(2) A.

Another alogical component of conjunctive meaning in language relates to iteration (Beilin and Lust 1975, p.190 referring to Ziff 1977 p.45). Ziff notes that the statement in [4.14a] gives the meaning of [4.14b],

[4.14a] George ate: he ate and he ate and he ate and he ate.

[4.14b] George ate a great deal.

while the formal representation of this statement does not:

$$p \cdot p \cdot p \cdot p \cdot p = p$$

Disjunction by the use of or poses similar discrepancies to those imposed by conjunction. The account of disjunction given earlier (cf. 5.1.5.2) describes inclusive disjunction where one or both of the disjuncts are true. Often, however, a stricter type of disjunction seems to be intended, usually denoted as "exclusive disjunction" and given its own sign ($\bar{\vee}$). In this case, disjunction will be true only if exactly one disjunct is true, and it, therefore, has the following truth table:

p	q	$p \bar{\vee} q$
T	T	F
T	F	T
F	T	T
F	F	F

Here the disjunction will be false when both component statements are true and when both are false. Accordingly, exclusive or is truth-functional, but it is not the same or of inclusive disjunction, the one often defined in logic.

There is no unanimous agreement on the interpretation of the natural language expression of disjunction. On the one hand, some believe it confounds the two meanings, i.e. inclusive and exclusive (Geis and Zwicky 1971), although some consider the inclusive interpretation is the more common reading (Quine 1966, van Dijk 1977b). On the other hand, Dik (1968) and Lakoff (1971) assume that disjunction in natural language is of the exclusive type. This is confirmed by Moravcsik (1970) who has studied the disjunctive connective in forty languages, though he also suggests that at times they are ambiguous as to whether they are inclusive or exclusive. This ambiguity is indeed indicative of the ostensible disparity between the logical and the linguistic interpretations of the connective.

A further disparity occurs in the cases when the meaning of or is conflated with condition or cause. Examples [4.15a] and [4.16b] (borrowed from Quirk et al. 1985 pp.933-934) clarify these uses:

[4.15a] Don't be too long, or you'll miss the bus.

[4.16b]They liked the apartment or they wouldn't have stayed so long.

Statements [4.15a] expresses a conditional and can be interpreted as in [4.15b] below, whereas statement [4.16a] expresses "inferential disjunction", where or may be viewed as the negative counterpart of because, interpreted as in [4.16b].

[4.15b] If you are too long, you'll miss your bus.

[4.16b] They liked the apartment because they stayed there the whole summer.

The connective or in this use is non-commutative and manifests an asymmetry, in the sense that clause ordering, unlike in logical disjunction, is not free. This asymmetric nature of or is evident in the fact that commuted versions of [4.15a] and [4.15b] are not acceptable.

The disparity between the meaning of the material implication connective "if... then", symbolised as →, and natural language "if... (then)", raises difficulties of two types. The first concerns the actual difference in meaning between the two, while the other concerns the truth-functional nature of the logical connective.

Copi (1982 cf. also Mitchell 1964, van Dijk 1977b) stresses that while natural language "if ... then" may express various types of conditional propositions (cf. examples [4.6-9] above), only a partial common meaning is expressed by → . In other words, the connective → cannot express the whole or complete meaning of any of statements [4.6-9]. This is because the meaning of logical → is exhibited in its truth-function: it denies that its antecedent is true while its consequent is false. Copi (1982, p.293) formalises this meaning as "the negation of the conjunction of its antecedent with the negation of its consequent". In the calculus, $p \rightarrow q$ is an abbreviation of ~(p.~q). This is a comparatively restricted meaning. Natural language conditional statements may fuse a further meaning with this: a causal connection (as in statement [4.6]), a logical connection (as in statement [4.7]), a definitional connection (as in statement [4.8]) or a decisional connection (as in [4.9]).

The second type of disparity between logical and natural language "if... then" resides in the truth values of the compound statement. According to the truth tables of material implication the proposition p → q is true if p is true and q is true, if p is false and q is true or false, whereas it is false if p is true and q is false. Accordingly the statements [4.17-19] express true conditionals:

[4.17] If Shakespeare wrote "Hamlet", Birmingham is an industrial city.

[4.18] If Shakespeare wrote "Volpone", Birmingham is an industrial city.

[4.19] If Shakespeare wrote "Volpone", Birmingham is the capital of France.

For in statement [4.17], both antecedent and consequent are true; in statement [4.18], the antecedent is false and the consequent true; and in statement [4.19], both the antecedent and the consequent are not true. However, does the fact that such statements do not occur in a natural text and that, even if they are made, we should not know whether to call them true or false, discredit the truth value of p → q?

In natural language conditional statements, the antecedent gives the condition under which the consequent is realised. For instance, to assert statement [4.20],

[4.20] If Labour comes to power, Britain is going to give a moral lead to the world through unilateral disarmament.
(The Sunday Telegraph 22/5/1983)

is to state a condition under which Britain will give the world a moral lead. Thus, for a conditional statement to be plausible, it is usually necessary that we should see the realisation of both the antecedent and the consequent as relevant to a common subject matter or a topic (cf. Mitchell 1964, Allwood et al. 1977, van Dijk 1977a, 1977b). Such relevance is lacking in statements [4.17-19].

However, "subject matter or meaning is strictly irrelevant to material implication, which is a truth function. Only truth and falsehood are relevant here" (Copi 1982, p.318). In other words, no such relation as that of "implication" need hold between antecedent and consequent. The minimum condition that must be satisfied, if an implicative proposition is to be true, is that it is not the case that the antecedent is true and the consequent is false. It is perhaps a little paradoxical that this minimal relationship should be called "material implication"; but what the logician is interested in is to show that "if a significant proposition is substituted for "$p \supset q$", it will entail a proposition of the form "$\sim q \supset \sim p$" (Mitchell 1964, pp.62-63). Related to statement [4.17], this proposition means "if it is the case that Birmingham is not an industrial city, then Shakespeare did not write 'Hamlet'". Thus the implicit claim in the calculus is that it is logically irrelevant what, if any, relation of relevance holds between the antecedents and the consequents of an implicative proposition.

Related to this is a further difference between logical → and natural language "if ... then". Since an implicative proposition is true regardless of whether its consequent is true or false, it follows that $p \rightarrow q$ and $p \rightarrow \sim q$ represent forms of propositions which can both be true together. In other words, statement [4.21a] and [4.21b] are both true.

[4.21a] If you take these pills regularly, you will feel better in a few days.

[4.21b] If you take these pills regularly, you will not feel better in a few days.

In natural language the inconsistency between statements [4.21a] and [4.21b] is immediately diagnosed since one is the "contrary" of the other (in this case, while the first statement is factual, the second sounds counterfactual). Moreover, a conditional in natural language is falsified if, though the condition is satisfied, the consequent is not realised.

4.1.7 Concluding Remarks

The contrastive account in the last section helps to draw a number of conclusions concerning the relation between connectives in logic and natural language. We admit that interrelations between the two types of connectives do exist to a limited extent, despite the disparities. In this respect we accept Strawson's (1952) view that the parallel between the two systems of connection is not a "simple mistake" (p.8). However, the disparities are so prominent that we have to adopt the view that the three connectives (and, or, if...then) do not represent conjunction, disjunction and implication exactly or unambiguously.

In summary, logical connectives are a small set of operators with a restricted role: to connect two simple propositions and determine the truth value of the compound statement. They perform this role regardless of relevance of meaning of the two component statements and irrespective of their relevance to a common topic or subject matter. This elementary apparatus is too meagre to describe natural language texts, particularly when they reflect an extensive and richly interconnected world. Although more sophisticated extensions of formal logic (intensional logic, temporal logic, Boolean logic, alethetic logic, and what-have-you) are considerably more adequate, they still elude and frustrate the linguist who is not a specialist in formal logic (not to mention the plight of the linguistic student!). Further, these logics, though unobjectionable in themselves, will still cause ample obscurity (not to mention chaos) if used to establish a text world model for linguistic communication. For instance, when accounting for an assertion, logics miss important factors in text actualisation: rhetorical choices, textual organisation, continuities and the text producer's intentions. Additionally, uncertainties, exceptions and unexpected turns cannot satisfactorily be described in a strictly logical framework. Indeed, as Allwood et al. (1977) argue, a lot more than truth-functional properties are involved in human communication.

We therefore intend to abandon the description of connectives in English and Arabic in terms of formal logic. No effort will be made to account for the truth function of the various connectives.

Instead, our concept of the connective is based on its cohesive force within the text and is compatible with our synthesis of cohesion (conceived of both as a set of relations and as linearisation, see Ch.4 above). This will be discussed later, after an outline of the place of the connective in linguistic and text linguistic studies.

4.2 The Connective in Linguistic Studies

This section discusses the treatment of the connective in some sentence linguistic grammars: traditional and modern. The aim is to delineate the limitation of these theories and, in addition, to complete the descriptive profile of the connective that we have started.

4.2.1 The Connective in Traditional Grammar

4.2.1.1 English Traditional Grammar

In English traditional grammar, connectives received comparatively little attention. Both in the specific description of English and in the general theory of language, they have been restricted to a casual treatment of conjunction. Vorlat (1963), in a study of English grammar during the period 1585-1735 (in which she analysed the 14 oldest known grammar works written on English soil) shows that conjunction, studied as a part of speech, was defined by its connective function. Some grammarians also referred to its ordering function, i.e. it establishes a set of main classes and subclasses, although, in general, there was "no agreement on whether the conjunction links only sentences and clauses, or also words"

(1963, Vol.I, p.181). Cobbett in his 1818 grammar (William Cobbett, eds. Nickerson and Osborne 1983) defines conjunctions as follows:

> "Conjunctions are so called because they conjoin or join together words, or parts of a sentence".
> (p.41, his emphasis)

He divides the types of conjunction into two (p.75): copulative, expressing a "union in the actions, or states of being, expressed by the verb", and disjunctive, expressing a "disunion".

A later and more interesting treatment of conjunction is incorporated in Earle's (1871) grammar of English. Earle establishes a group of items which he calls "The Link-Word group"[(2)]. This group comprises prepositions and conjunctions. These "act as the intermediaries of words and sentences" (p.434). The distinction between the two is "based on the definition that prepositions are used to attach nouns to the sentence, and conjunctions are used to attach sentences or to introduce them" (loc. cit.). In characterising conjunction, Earle makes two important statements: a) conjunction is essentially a symbolic word (this is reminiscent of logic), but it nevertheless comprises "within its vocabulary a great deal of half-assimilated presentive matter" (pp.444-445); b) of all the parts of speech "the conjunction comes last in the order of nature" (p.444). In Earle's view, since a conjunction joins "two sentences", it presupposes "the completion of the simple sentence" and therefore implies "the pre-existence of other parts of speech" (loc. cit.). Hence "the conjunctions are as a whole a comparatively modern formation" (p.445). The rest of the work exemplifies the use of a number of conjunctions: for, till, until, because, too-to, as, as..as, as..so, so..as, so, so..that, then, where, whereas, whether, or, nevertheless, just, no more than, yet. Another, and indeed more detailed, treatment of interclausal connection is Märtzner (1885). Under the heading of "Satzfügung", he divides connection into two major types "die Beiordnung der Sätze" and "die Unterordnung der Sätze". The first type is divided into two classes of connection: syndetic and asyndetic. The syndetic connection is further classified according to meaning, into four subclasses: a) copulative, b) disjunctive, c) adversative and d) causal.

A further attempt at describing conjunction is made by Sweet (1892). Conjunctions are divided into "full-conjunctions" and "half-conjunctions". The former can be subdivided into "primary": those that are used as conjunctions and conjunctions only (e.g. and, or) and "secondary": those that are originally adverbs or prepositions, but may act as conjunctions (e.g. for, since). The "half-conjunction" type subsumes adverbs which closely resemble conjunctions, e.g. still, nevertheless. The distinction between full-conjunctions and half-conjunctions is that the latter connect logically only, not formally also, as the former do. Sweet then classifies conjunctions (both types full- and half-conjunctions) "according to meaning" (ibid p.145) into: a) affirmative (copulative), b) alternate, c) negative, d) adversative, e) concessive, f) hypothetical, g) temporal, and h) causal.

Later grammarians showed similar scanty attention to the problem of connectives as those before, regarding it mainly as a minor part or sub-part of speech. Jesperson (1933), for instance, treats conjunction as a subclass of particles which include, in addition, adverbs and prepositions (pp.68-9). Onions (1906) makes only a brief remark on conjunction, in the last pages of the book (pp.145-146) as an afterthought. Partridge (1949, p.60) defines the conjunction simply as a "connecting word"; to clarify this, he quotes the definition of the term in Webster's New International Dictionary!

In summary, traditional grammarians have been preoccupied with the establishment of the parts of speech and other grammatical paradigms, set by the Latin tradition (cf. Padley 1976 for further remarks on this point). In addition, emphasis has been placed on such matters as correctness, linguistic precision and literary excellence.[3] Often grammatical treatments are characterised with obsolete exemplification and the inclusion of some archaic expressions regardless of their frequency of distribution. For instance, Earle (1871) discusses the use of the preposition sithence (p.441) and the conjunctions preadventure (p.446), then to mean than (p.448), howbeit, according as (p.454). The result is the absence of a systematic description of conjunctive structures based on a representative data and characterised with analycity and rigour.

4.2.1.2 Arabic Traditional Grammar

Arabic traditional grammar classifies conjunctions as "ḥurūf al-<aṭf" [particles of connection], a far more limited category than that discussed in English grammar. The types of connection and the

syntactic function of the particles are summarised in Ibn Mālik's "'alfiyya". According to this view, connection, "al-<aṭf", is of two types: "<aṭf bayān" [Explicative Apposition] and "<aṭf nasaq" [connection of sequence]. The latter is more related to our conception of connection than the former.[4]

The connection of sequence "<aṭf nasaq" refers to the connection of two items (words, phrases or clauses) with one of a limited set of "particles of connection". These are "wa [and], fa [and, then], ḥattā [even], t̲umma [then, 'am [or], 'aw [or], 'immā ... wa'immā [either ... or], lākin [but], lā̄ [(but) not], bal [but, instead]".

These connectives are extensively exemplified in traditional grammar, and their behaviour and the various constraints that govern their use are exhaustively discussed, including their influence on the case or mood of connected items. For instance, generally, they all dictate that the second connected item carries the same case or mood of the first, an important factor that gives rise to the creation of parallelistic strings (see Chapters 6 and 9 below, cf. also Al-Jubouri 1984).

However, a close inspection of the meaning and behaviour of these particles reveals that what traditional Arab grammarians refer to as conjunction is in fact "coordination", not only of clauses, but also of lower rank constituents. Subordination is restricted to the study of "condition" where a number of connectives (with a practical common conditional denotation) are grouped together mainly because of similarity in their syntactic and morphological patterning. Under "particles of condition" (which are distinguished from "particles of connection") are included "'in [if, whether], man [whoever], mā [whatever], mahmā [whatever], matā [when, whenever], 'ayna [where, wherever], 'ayyāna [wherever, whenever], 'annā [whenever, wherever], ḥayṯumā [wherever], 'ayya [whatever], 'iḏā [if], law [if], kayfamā [whatever way], 'iḏ mā [if, when], la'in [if], 'ammā [as for], lawlā [had it not been for]".

It should be noted that each of these connectives expresses a variable "degree" of condition and that they possess a varying shade of meaning that can rhetorically distinguish one from another. This intricacy of usage is one factor that contributes to the complexity of condition in Arabic. Another factor is that "condition" has not been well studied. Traditional grammar studies condition insofar as it affects the verb mood of the verbs in the subordinate and main clauses "jumlatu wa jawābu al-šarṭ [literally: conditional clause and its response]". In other words, condition is studied within syntactic and morphological perspectives, rather than a semantic, rhetorical or a textual one. In this respect, Al-Muttalibi (1981, p.13) admits that traditional grammatical research, both old and new, has not concerned itself in standardising the conditional "term": there have been various interpretations of the conditional connectives to the extent that overlap has rendered the same connective having different interpretations with different scholars. Indeed very few scholarly studies have been made to re-examine conditional clauses and their connectives.(5)

More recently, Arabic connectives have been studied under coordinate and various types of subordinate clauses in some scholarly grammar books. The classification has been based on the Western tradition of classical grammar. One extensive study along these lines is Cantarino's (1975, Vol.3) which concentrates entirely on coordination and subordination. A slightly different approach is Beeston's (1968, 1970) where sections are allocated to the study of circumstantial clauses, qualifying clauses, that-clauses, in addition to conditional clauses. Note that Beeston, in line with traditional Arab grammarians, uses the term "connectives" to label "coordinators". A similar approach is followed by Tritton (1943), Haywood and Nahmad (1965) and Wickens (1980).

4.2.1.3 Concluding Remarks

It should be remembered that all these attempts to study the connectives concentrate primarily on the structural patterning of the clause. The treatment of the connectives and their semantic interpretations are therefore a peripheral rather than a central concern. This is of course in line with the aims and perspective of a sentential grammar. No study, within this framework, has focused on the rhetorical choices, the types of organisational contingency, or the semantic "defaults" that each class of subordinators or coordinators can have. The concern on describing structural details of the sentence is connected with providing sufficient motivation for advocating linguistic purism and accuracy. Hence the traditional approach has fallen short of achieving a coherent profile on how connectives operate within a text and of taking account of types of connectives other than coordinators and subordinators.

4.2.2 The Connective in Some Theories of Modern Linguistics

Since it is not appropriate within the scope of this work to review the place of the connective in all theories of modern linguistics, we have selected, in a random-like manner, two theories: Bloomfield's distributional model and the transformational generative grammar. While we apologise for omitting the rest of the theories, we believe that these two will offer an insight into sentence grammars which have studied the connectives.

4.2.2.1 Bloomfield's Distributional Model

The study of connectives within Bloomfield's theoretical model, as indeed with most other sentence linguistic theories, is confined to the study of coordinative and subordinative constructions. Bloomfield (1935, p.194) distinguishes two types of syntactic constructions, where a syntactic construction "shows us two (or sometimes more) free forms combined in a phrase". One is the "exocentric" construction, where the phrase may belong to a form-class other than that of any constituents. For instance, the actor-action construction is exocentric since it does not belong to the nominative expression of the actor, nor to the finite verb of the action. The other type of syntactic construction is "endocentric", belonging "to the same form-class as one (or more) of the constituents" (loc. cit.).

Subordinate clauses belong to the first type of construction. The constituents in a subordinate clause are usually a subordinating expression (a conjunction) and an actor-action phrase (as in "if she leaves"); the clause as a whole ("resultant" phrase in Bloomfield's terminology) does not express the function of any of its constituents, but serves as a "modifier". Normally the subordinating conjunction is peculiar to the construction and serves to characterise the resultant phrase, in this case the subordinate clause.

Coordinative constructions are, in Bloomfield's view, endocentric. For example, the phrase "boys and girls" belong to the same form-class as the constituents boys, girls. These constituents are the "members of the coordination", and the other constituent, and, is the coordinator. Bloomfield points out that there may be minor differences of form-class between the resultant phrase and the members. For instance, "Sally and Susan" is plural, while each member is singular.

This view of coordination (and one can include subordination as well) came under fire when Dik (1968) questioned Bloomfield's basic concepts of "construction", "position", "function" and "form class". Dik points out (p.21) that the notion of "form-class" in fact "comprises quite different types of categorisation", and is further "influenced by the lack of clarity in the status of the notion 'position', which is, again, a consequence of the difference between theory and practice regarding the notion 'construction'". Additionally, Dik introduces some counter-examples to Bloomfield's conception of coordination, and concludes that the differences in form-class between the coordination as a whole and its members are not 'minor', as Bloomfield claims, no matter how this term is interpreted. They are major differences that, together with the general remarks made above, are "sufficient to reject a general definition of 'coordinative construction' in terms of the endocentric-exocentric dichotomy" (ibid, p.22).

One can add that Bloomfield's model places its emphasis on the syntactic properties of coordination and subordination and their relevant endocentric-exocentric constructions. There are definitely other properties, semantic and textual, that are left untouched, probably because their discussion would go beyond the scope of the model. Needless to say that the description of intersentential connection is not the concern of the distributional theory.

4.2.2.2 Transformational Generative Theory

The interest of the transformational generative theory in the connective is restricted to its proposals for coordination. Chomsky (1957, p.35) considers the process of conjunction as one of the most productive processes for forming new sentences. A "simplified" rule that he formulates for generating conjunction looks like this:

> "If S1 and S2 are grammatical sentences, and S1 differs from S2 only in that X appears in S1 where Y appears in S2 (i.e., S1 = .. X .. and S2 = ..Y..), and X and Y are constituents of the same type in S1 and S2, respectively, then S3 is a sentence, where S3 is the result of placing X by X + and + Y in S1 (i.e., S3 = .. X + and + Y..)". (loc. cit.).

Chomsky argues that the phrase structure model (based on constituent analysis) can generate coordination only under very limited possibilities, and that it actually fails if the constituents X and Y disagree with the rule above.

In Chomsky's view, coordination, like negation and questions, is achieved by a "generalised transformation". (This is different from "singulary transformation" in that it acts on two or more structures at once, whereas singulary transformation acts on single structures). In the standard version of his theory (1965), Chomsky makes a theoretical shift: conjunction is an aspect of the base of the syntactic components, and is generated by base rules through "generalised phrase-markers". The general rule for conjunction in the standard theory is formalised as follows:

> "if XZY and XZ'Y are two strings such that for some category A, Z is an A and Z' is an A, then we may form the string X Z and Z' Y, where Z and Z' is an A".
> (Chomsky 1965, p.212, fn.9).

Note that although the two rules (in the 1957 and 1965 versions) is essentially the same, the second is a rather simpler formulation and is claimed to operate in the deep structure. More succinctly, "sentences are described as compound if their deep structures contain two or more conjoined sentences" (Jacobs and Rosenbaum 1968, p.253). This rule of conjunction is repeated with some detail in Chomsky (1975), where it is given as a criterion for grammatical description in terms of phrase structure (cf. in particular pp.223-227) and in transformation analysis (pp.301-302). In the extended standard version of TG (cf. Radford 1981), Chomsky suggests a coordinate structure constraint where coordinate structures are treated as "islands", i.e. no "subpart" of it can be moved, though this does not block the movement of the island as a whole.

One of the weaknesses of the conjunction rule is its limitation. It only states that under certain conditions a coordination may be formed, but it does not state that all possible types of coordination are generated in this way. The examples that Chomsky (and other transformationalists) produce to support these rules admit only simple subject-predicate combinations in the deep structures, and attempt to illustrate cases of coordination reduction. Further problems have been identified (cf. for instance the arguments in Dik 1968, pp.74ff), which concern such aspects as transformationalists' claims of simplicity, the problem of plural nouns that have in many cases to be reduced to singular ones, the problem of referential identity, the solution of which does not seem to lead to an empirically adequate account of the relations between the sentences concerned. In summary, the objections raised "throw grave doubts on the adequacy, the generality, and the simplicity of the transformational description of coordinations" (Dik 1968, p.92).

4.2.2.3 Concluding Remarks

To confine the treatment of the connective in modern sentential grammar to only two models is to be crudely reductionist. However, what interests us is one common factor: connectives have not been duly explicated. One obvious reason is that a characterisation of the textual role of connectives requires a description of the mechanisms of connection between clauses/sentences in a text. Such a description is only partially made within the framework of sentential grammar with all its theoretical restriction. In the first place, one of the main achievements of modern sentence linguistics, particularly during the first half of this century, is the elaboration of the "lower" level of the linguistic system, i.e. of phonology and morphology. The ideas developed for the study of syntax have concentrated on the clause as the primary object of analysis, with connectives receiving only a casual attention. For instance, Bloomfield's theoretical model identifies two constructions: endocentric and exocentric, with coordination (including interclausal coordination) being assigned to the former and subordination to the latter. Within the standard version of TG, the syntactic description of connectives generally starts from base structures which are not made explicit. The extended standard version adds a constraint on element movement in coordinate structures. Some TG scholars (cf., for instance, Burt 1975), consciously ignore the whole area of coordination or subordination.

It was within textual models of linguistics that textual connectivity received particular attention. Indeed one of the main differences between sentence and text grammars, as far as the type of task is concerned, is that the former regulates sentence construction whereas the latter regulates connections between sentences. This leads us to a close consideration of the place of connectives in some textual models, a task for the next section.

4.3 Connectives in Some Models of Textual Cohesion

4.3.1 Preliminaries

The establishment and elaboration of connectives and their textual role is a developmental factor in text linguistic literature. The problem of connectivity by special linkage "devices" has been recognised since the start of the research on textual problems embodied within text linguistics (cf., for instance, Hasan 1968, Dressler 1972). The examination of connectives in textual studies will be covered in this and the next section (4.4). In this section we shall examine the place of connectives in the textual models of cohesion discussed in the previous Chapter. In the next section we shall examine the treatment of connectives in other textual studies.(6)

4.3.2 The Connective in Enkvist's Stylistic Model

In his stylistic model of cohesion (see 3.3.2.1 above), Enkvist describes the role of connectives under "clause linkage", which "provides .. an arsenal of formal means marking the ways in which clauses cohere within sentences and sentences within texts" (1973, p.122). In categorising the relations between clauses/sentences, Enkvist borrows a "logical" scheme of classification suggested by Milic (1969 p.21). The scheme envisages eight types of relations:

1. Additive, where a proposition has no organic relation with its predecessor and is expressible by and.

2. Initial, which refers to the first sentence of a paragraph.

3. Adversative, where a proposition changes the direction of the argument; it is realised by such connectives as but.

4. Alternative, which pertains to a proposition that may be substituted for a previous one; this relation is expressed by or.

5. Explanatory, which involves a restatement, definition or expansion of the previous proposition; an example of a relevant connective is "that is".

6. Illustrative, expressing an instance or illustration as

denoted by "for example".

7. Illative, suggesting a conclusion, particularly through such a connective as "therefore".

8. Causal, pointing out the cause for a preceding conclusion, as expressed by "for".

Enkvist then states that density patterns of types of sentence linkage "may offer us a battery of additional style markers" (ibid p.123). This important statement is unfortunately left without further elaboration nor any exemplification.

4.3.2 The Connective in the Functional-Systemic Model

Perhaps the most extensive as well as influential account of connectives in text linguistic studies is Halliday and Hasan's description of conjunction within their functional-systemic model of cohesion. The description starts with a note that conjunction is rather different in nature from the other cohesive relations (reference, substitution and ellipsis). Conjunctive elements "are not primarily devices for reaching out into the preceding (or following) text", which is the primary function of referential items, "but they express certain meanings which presuppose the presence of other components in the discourse" (Halliday and Hasan 1976, p.226).

Conjunctive cohesion is set up when any pair of adjacent sentences are related by one of a small set of semantic relations. These are of two types: external or internal. Halliday and Hasan summarise them in this way:

> "..[a conjunctive relation] may be located in the phenomena that constitute the content of what is being said (external) or in the interaction itself, the social process that constitutes the speech event (internal)". (ibid, p.321).

The classification is made according to the functional-semantic component from which each type is derived: either the meaning resides in the ideational or interpersonal component. An external relation exists between external phenomena and reflects the "ideational" function of language; "it is a relation between meanings in the sense of representations of 'contents' (our experience of) external reality (ibid p.240). An internal conjunctive relation, on the other hand, is that which is internal to the communication process, reflecting the interpersonal function of language (that is, in the functional-systemic terminology, the speaker/writer's choice of speech role and rhetorical channel, his attitudes and his judgements).

The difference between the two types is illustrated in the following two examples, where the relation is basically a temporal one (from Halliday 1979 p.190):

[4.22a] First all of the machine broke down. Next it started to make alarming noises inside.

[4.22b] First of all the machine has broken down. Next it doesn't belong to me anyway.

In [4.22a] the temporal successivity is a relation between events: first one thing happened, then another. In other words, the temporal successivity is in the "thesis", in the content of what is being said. In [4.22b], on the other hand, the two sentences are two steps in the argument, and the temporal successivity is seen in the speaker's organisation of his discourse, his unfolding of his role in the speech situation.

Halliday and Hasan then categorise conjunction into four broad semantic relations, which, in their view, may be described in most general terms under the four headings of "and", "yet", "so", and "then". The relations are respectively: additive, adversative, causal and temporal. Halliday and Hasan maintain that the reduction of the numerous varied kinds of conjunction to this small number of basic types leaves the door ajar for a considerable amount of subclassifying; the complexity of facts has to be relegated to a later, or more 'delicate', stage of the analysis. Their reason for selecting a simple overall framework is a pragmatic one: "it seems to have the right priorities, making it possible to handle a text without unnecessary complication" (Halliday and Hasan 1976, p.239).

It is worth mentioning that the distinction of external/internal relations is common to all four categories of semantic relations. The language user, upon using any of the different relations, may exploit either type (external or internal) in order to create text. To elaborate this, Halliday and Hasan set up a detailed scheme that subclassifies the four main conjunctive types; additive, adversative, causal and temporal, in relation to the external/internal meaning (ibid pp.242-243). The total number of subcategories is 56: 21 are external/internal (3 additive, 3 adversative, 9 causal and 3 temporal), 27 subcategories can be used in an internal meaning only (7 additive, 5 adversative, 6 causal and

9 temporal) and 11 subcategories are used in an external meaning (1 adversative and 10 temporal). The categorisation is considerably detailed and discrete and can offer insights into this type of cohesive relation.

Two points have to be made about Halliday and Hasan's classification, which, despite its comprehensiveness, make it less compatible with the aims of our study. First it is a semantic classification of the underlying conjunctive relations and not the connectives that are used to signal them. These relations are general and can be associated with different threads of meaning at different places in the text and can take various structural forms. Take as an example the temporal successivity as expressed in [4.22] above. Although the cohesion is achieved via the conjunctive expression "First ... Next", it is the underlying semantic relation of successivity in time that actually has the cohesive power. This relation can take other structural forms (see examples in Halliday and Hasan 1976, p.228).

The second point is related to the first one. Because conjunction is a highly generalised component within the semantic system, it can be signalled by connective devices that are not restricted to conjuncts (whether adverbs or adverbial phrases). This is a valid point and we subscribe to it. However, Halliday and Hasan delete from their list of conjunctive expressions all connectives that introduce subordinate clauses, despite the fact that such connectives may signal similar semantic relations to the ones that conjunctive expressions indicate. The main reason behind this omission is that subordination, as well as coordination, are structural and not cohesive relations, and that cohesion is a relation between sentences (despite the difficulty of defining this concept), and not a relation within a sentence.

These two constraints and the particular way in which the classification of relations is schematised make it necessary, in order to meet the aims of this project, to adopt our own conception of connectives. This will be elucidated subsequently.

4.3.3 Connectives in Gutwinski's Stratificational Model:

In his stratificational model of cohesion, Gutwinski studies connectives under coordination and subordination, which he employs as cover terms for the cohesive relations expressed by connectives between clauses and sentences. In other words, these two terms do not subsume the same range of grammatical phenomena for which they have traditionally been used. Gutwinski justifies this on the assumption that "the connectivity of two or more sentences due to the presence of connectors whose function is to link these sentences into a morphologic construction larger than a single sentence is essentially of the same kind as the grammatical connectivity, marked also by connectors, of clauses within a sentence" (Gutwinski 1976, p.73). Accordingly, Gutwinski sees no reason (except, in his view, for the arbitrarily set limits of grammatical analysis) to restrict the study of the relations of coordination and subordination indicated by connectors, to clauses in the sentence structure.

Thus, connectives are divided into two large groups: coordinating and subordinating. Each group is then subdivided according to function. The classification is influenced by one proposed by Gleason (1961 pp.342-343). The following is Gutwinski's classification with some illustrative examples:

A. Coordinating Connectives:

Cumulative or Additive: and, likewise, moreover, in addition, furthermore

Disjunctive: or, nor, else, otherwise, alternatively

Adversative: but, however, nevertheless, on the contrary, on the other hand.

Illative: therefore, so, for this reason, then.

B. Subordinating Connectives:

1. Causal: because, since, as, for the reason that.
2. Purposive: that, in order that, so that, lest, for the purpose of.
3. Conditional: if, unless, provided that, whether.
4. Concessional: though, although, in spite of the fact that, notwithstanding that.
5. Comparative: as, than.
6. Temporal: as, as soon as, while, before, until, since, when, ere.

Gutwinski admits that the relations covered by coordination and subordination could be grouped and named in several ways, some of which might be more satisfactory than his. But he seems more interested in the existence of such cohesive relations than in the way they are categorised. Gutwinski's main aim is to use such a simplified framework to "a preliminary study" of cohesion in selected literary texts. A more intricate framework, though it can be developed, is never attempted.

4.3.4 The Connective in the Procedural/Relational Model

In this model of cohesion (cf. Beaugrande 1980, Beaugrande and Dressler 1981), one important device that fulfils the functions of long range cohesion is "junction". This is regarded as a "clear device for signalling the relationships among events or situations" (Beaugrande and Dressler 1981, p.71). These various relationships often obtain without the explicit use of junction "simply because people have predictable ways of organizing knowledge" (Beaugrande 1980, p.159). The model, therefore, uses the term "junction" only when there are junctive expressions (i.e. connectives, such as, and, or, but, because, etc.).

Beaugrande and Dressler (ibid) identify four types of junction:

1) Conjunction links "things" that have the same status, e.g. both are true in the textual world. Conjunction signals simple addition of events in a temporal and causal sequence. Beaugrande (1980, p.160) believes that since those relations are "recoverable from content, the junctive expression 'and' is dispensable, or replaceable with subordination". To illustrate that Beaugrande gives this example. (loc. cit.):

[4.23a] Three boys are playing football. One boy kicks the ball. It goes through the window. [etc.]

[4.23b] Three boys are playing football when one boy kicks the ball so that it goes through the window. [etc.]

This non-committal nature of conjunction, in the view of Beaugrande and Dressler (ibid p.72), makes it the "default junction", since, unless specified otherwise, "events and situations are combined additively in a textual world".

2. Disjunction: This category of relations require an explicit connective, normally "or" (sometimes used for forming a phrasal connective "either .. or"). Disjunction connects two items or knowledge configurations that, in regard to their environment, have an alternative status but only one seems to be valid in the textual world. Beaugrande and Dressler (ibid, p.72) believe that the processing of disjunction is perhaps difficult because a processor, in order to maintain the integration of a text world, would seek to select the valid alternative and attach it, discarding others.

3. Contrajunction: This refers to the linkage of two knowledge configurations that, in regard to their environment, appear incompatible, yet they co-exist in a textual world. This relation is signalled most often by "but". Other connectives that indicate contrajunction are "however", "yet", "nevertheless", etc. The textual function of contrajunction is "to ease problematic transitions at points where seemingly improbable combinations of events or situations arise" (ibid, p.73).

4. Subordination: This type of linkage indicates that the determinacy of one knowledge configurations is contingent upon access of the other. The relation is represented by a large

repertoire of connectives: causal (such as, "because", "since", "as", "thus") and temporal ("then", "next", "after", "since", "while"). Hence, subordination signals more diffuse dependencies than do other categories of "junction", some of these dependencies are inferrable via world knowledge.

One other use of subordination is the signalling of modality, i.e. probability or necessity of events and situations. The connective "if" indicates the condition under which a particular state or event is enabled in the textual world. Modality is essential for achieving a "projection" of events and situations, those that will, can, may happen or might have happened in the textual world.

In Beaugrande and Dressler's view (ibid, p.74) the use of connectives as explicit signals "is rarely obligatory, because text users can recover relations such as additivity, incongruity, causality, etc. by applying world knowledge". This is a questionable statement and, coming from two leading figures in text linguistics, is rather disappointing. It is true that the application of world knowledge will help a text receiver to recover the relations that hold between statements. But text is not a mere collection of sentences with inferrable relations holding them together. It is an artifact that reflects a disciplined proportionateness among its various components, and exhibits the fusion of various principles of textuality. The use of connectives is monitored by the demands of cohesion, coherence, informativity and acceptability, and therefore cannot be cancelled as "rarely obligatory". Further, text reflects efficiency, effectiveness and appropriateness of design. There are rhetorical constraints that monitor these factors, constraints that are often text-type and/or language specific. The deliberate absence of connectives can, at least on some occasions, violate these constraints. This can impair text continuity by creating a rhetorical imbalance in text linearity: in the way constituents are grouped, sequenced or made prominent. This leads to textual gaps that distort text fluency and reduce its effectiveness. (We shall return to this point subsequently when we discuss the nature of the concept of connective.)

Beaugrande and Dressler, however, admit that the intricacies of the role of connectives are far greater than their account might imply. They further agree that by using connectives "text producers can exert control over how relations are recovered and set up by receivers" (loc. cit.). In this perspective, they argue, the use of connectives "demonstrates how communicative interaction, not just grammatically obligatory rules, decides what syntactic format participants use." (ibid pp.74-75). To clarify this point, they state that the use of connectives help make reception of a text efficient and help the text producer during the organisation and presentation of a textual world.

4.4 Some Other Textual Descriptions

4.4.1 Van Dijk's Semantic Description

Van Dijk's work on the connectives (1977a, 1977b, also his 1977c modified in 1979 and reprinted as Chapter 6 in 1981) attempts to "provide a partial description of connectives in natural language, especially of their abstract underlying structure" (1977a, p.11). In his view, connection constraints are dependent on the requirement that the respective facts they denote are related. The term "connectives" is used to subsume a set of expressions from various syntactic categories that express relations between propositions, facts or speech acts. The description focuses on connections between clauses and sentences, leaving out what Van Dijk calls "phrasal connectives", i.e. connectives making (noun or verb) phrases from two or more words or phrases.

In classifying connectives, van Dijk distinguishes two types: semantic and pragmatic. Semantic connectives express relations between facts and propositions. Unlike the classical connectives of logical languages, they are not truth-functional, but intensional, i.e. they can be described in terms of meaning relations (e.g. possibility, probability, necessity of their conditional link). A list of types of connectives, drawn from traditional grammar, is made (8 in 1977a p.53, and 11 with examples in 1977b, pp.14-15). However, Van Dijk maintains that these types, owing to syntactic and stylistic determinants, may be specific variants of a restricted number of abstract basic connectives. To identify these, he first sets up a formal semantic framework that takes into account a number of discoursal aspects, such as topic of conversation, point of view, the possible world, etc. Within this framework, abstract basic connectives are reduced to four: conjunction, disjunction, conditionals and contrastives. Van Dijk then sets out to identify each category, and specify semantic rules for their description.

Pragmatic connectives express relations between speech acts. Their use may be accompanied by different constraints, phonological and syntactic. For instance, pragmatic connectives, in Van Dijk's view, are often sentence-initial, followed by a pause and expressed with a specific intonation contour. It should be noted that the distinction between the two types of connectives, semantic and pragmatic, is not clear-cut. Fuzziness arises when the use of pragmatic connectives leaves traces of their semantic meanings. Van Dijk explains this by the assumption that each connective has a certain (minimal) meaning which may be further specified depending on its semantic or pragmatic use. For this reason, a description of pragmatic connectives requires an interpretation in terms of "functions" with respect to pragmatic "contexts".

Van Dijk then outlines the pragmatic use of a few major English connectives: and, but, or, so, if, contrasting them with their semantic uses. For instance, to illustrate the uses of and, Van Dijk (1977c, 1979, 1981) gives these two sentences.

[4.24a] Yesterday we went to the movies and afterwards we went to the pub for a beer.

[4.24b] Why didn't Peter show up? And, where were you that night?

[4.24c] Harry has counted me out. And, I even hadn't had a chance.

In [4.24a], and is used semantically to express a relation between two facts that are ordered in time. In [4.24b], And is used to indicate the fact that the speaker has another question; i.e. it signals an addition to the first speech act. In [4.24c], the speaker uses And to mark a contradiction or protest, perhaps to prevent the hearer from drawing false conclusions from the first speech act.

In general, natural connectives, with the exception of enumerative conjunction and disjunction, are of the "conditional" (relevant) type. This means that the situation, fact or speech act denoted in the consequent is to be interpreted in worlds determined by the antecedent (together with the topic of conversation).

A further characteristic of connectives lies not only in their ability to make sentences out of sentences, but also in their ability to build "sequences of sentences". In other words, connected "propositions" may be expressed either in composite sentences or in sequences. The semantic rules, it is assumed, that hold for the sentential connectives also hold for those occurring in sequences. The type of connectives that connect sequences are normally coordinators and sentence adverbs, not subordination like "because", "although", "if...then".

However, van Dijk emphasises that sentential and especially sequential connection need not be signalled by explicit connectives. The co-occurrence of sentences may express the connections between propositions. On the other hand, disjunctions and <u>if</u>-conditionals are normally not expressed without their explicit connectives because the facts denoted may not hold in the actual world. The general rules in this respect sound as follows:

> "... in sequences which do not use connectives (asyndetic sequences), the sentences are interpreted to have truth values with respect to a given topic of conversation, relating them indirectly. As a second general rule it will be assumed that facts thus connected by one topic of

> conversation are further to be connected in the closest possible way, viz as reason/cause and consequences... In cases where conditional relations are exceptional, i.e. do not hold in most possible situations, the explicit connective must be used". (1977a, pp.87-88).

To conclude, van Dijk's work is a highly interesting contribution to a better understanding of the semantic and pragmatic nature of connection. Most of the points in his characterisation of connectives are valid and worth noting particularly in so far as his formal semantic framework is concerned. One drawback is perhaps resident in the apparatus itself. The concern with the logico-semantic nature of connections between clauses, sentences or sequences has forced him to neglect consideration of rhetorical factors. In addition, the description of connectives is based on contrived exemplifications, rather than on studying samples of natural language texts, a problem that directly affects his conclusions.

4.4.2 Quirk-et al.'s Description

The treatment that Quirk et al. put forward for the connectives falls into two related versions: one made in their (1972) work and the other in their more recent volume (1985). Both versions are summarised below. It should be mentioned that Quirk's interest in connectives goes back to his 1954 paper in which he studied subordinators and conjuncts as the main means of signalling the concessive relation in Old English. However, the main basis for the two versions, particularly the first one, is Greenbaum's work on adverbials in English (1969).

In their first version of the description of sentence connectives, Quirk et al. deal primarily with "syntactic devices that enter into sentence connection" (p.652). There are other "factors" that may be present in addition to syntactic devices: implication in the semantic context and lexical equivalence. The former refers to "relationships implied by the juxtaposition of sentences with their semantic interpretation" (p.653). The latter (i.e. lexical equivalence) refers to the relationship exhibited by successive sentences through their vocabulary (labelled as "lexical cohesion" in models of textual cohesion).

The syntactic devices interact with the other two factors in the connection between sentences. Other factors that may be relevant include features of the situation, including: the visible scene, the medium of communication, the relationship between the participants in the communication, and the specific purposes of the communication.

The syntactic devices used for connecting sentences are grouped under the following headings: a) time and place relaters, b) logical connectors, and, c) substitution. The third one subsumes the cohesive relations that in models of cohesion fall under the heading of "reference" and "substitution". These do not concern us at the moment, and, therefore, only the first two types of devices are outlined below.

1. Time and Place Relators

a. Time Relators

This type of sentence connection can be established "by time-relationships signalled by adjectives or adverbials with temporal significance or by tense, aspect and modality. Three major divisions of time-relationships are suggested.

i. Temporal ordering previous to given time reference: e.g.

Adjectives: earlier, former, preceding, previous.

Adverbials: already, as yet, before, beforehand, previously.

ii. Temporal ordering simultaneous with given time-reference, e.g.

Adjectives: co-existing, concurrent, contemporary, simultaneous.

Adverbials: at present, at this point, concurrently, here, in the interim, meantime, meanwhile, simultaneously, in the meanwhile, and the relative when.

iii. Temporal ordering subsequent to given time reference, e.g.:

Adjective: following, later, subsequent.

Adverbials: after, afterwards, at once, finally, immediately, last, later, next, after that, after this.

b. Place Relators:

These refer to expressions that denote place-relationships and that have a role in connecting sentences. They are of two types:

i. Place-relators where ellipsis is involved. Examples from Quirk et al. 1972, p.660):

[4.24] He examined the car. The front was slightly damaged.

[4.25] The traffic lights soon changed. He walked across quickly.

ii. Place adverbs that do not involve ellipsis, e.g. here, there, elsewhere, the relative where.

2. Logical Connectors:

Logical connectors are grouped under the coordinators and, or, but, "in as much as similar interpretation could obtain if the coordinator alone is used" (ibid p.661). Some are grouped under the conjunction for. Accordingly, we have four major types, each is categorised into more explicit relationships. Figure 4.1 is a summary of these relations (loc. cit.).

Fig. 4.1 Logical Relations in Quirk et al. (1972, p.661).

These relations are mentioned earlier in the work (ibid, p.520) when a semantic classification of conjuncts is given, as defined by their role in clause and sentence connection. The difference in the two schemes is that the relations expressed by or and but are amalgamated under one group: "contrastive", which is then subdivided into: reformulatory, replacive, antithetic and concessive. Another difference is the omission of "cause" as a category (since it is covered by the class "result"). This is replaced by the category "temporal transition".

The second version of Quirk et al.'s account of sentence connection gives the same semantic classification for the role of conjuncts as that given in the earlier version. However, in discussing and categorising sentence connection in general, the second version displays deeper insight and more comprehensive detail. The authors are definitely conscious of the accumulation of knowledge during the thirteen years that separate the two versions, and have therefore incorporated into their account some fundamental findings of research in text linguistics.

This version starts with a general classification of sentence connection. There are five main types:

1. Asyndetic connection: Here the mere juxtaposition of sentences is an icon of connectedness, "even where the juxtaposed parts have no grammatical or lexical features in common" (Quirk et al. 1985, p.1425).

2. Structural parallelism: A strong impression of connectivity

is given when neighbouring sentences share grammatical features of tense, aspect, clause, structure, or word order.

3. Connection by sequence: Here juxtaposition of sentences is iconic of connected sequence. The sentences have enough grammatical features in common, such as the same subject or the same tense, that they often imply temporal and causal connection.

4. Thematic connection: This involves sequential proceeding from the known (given) to the unknown (new), "thus forming a chain in which what was unknown becomes the known as a point of departure towards a further unknown item" (ibid, p.1430). The connection should not be taken as linearly straightforward, and indeed the interpretation of text is more often than not dependent on world knowledge and on common sense. This is particularly true when thematic connections are linked with lexical items that are in turn linked to situational context.

5. Rhematic connection: This involves syntactic and/or lexical connection through the rhemes of adjacent sentences.

6. Syntactic connection: In this type an overt connecting item, usually and is used.

Quirk et al. then concentrate on the relation between parts of a text. Connection between the various parts can be achieved by connective features that are grouped into four categories:

a) Pragmatic and semantic implication: Here interpretation of the text parts relies not only on our knowledge of the language, but also on our world knowledge.

b) Lexical linkage: This is achieved via lexical recurrence, similarity and inclusion. It can also work through lexical opposition and exclusion. Lexical links may also be established by additional information fed in by the text producer, and may, at times, depend on the general knowledge of the text receiver.

c) Prosody and punctuation: Different prosodic realisations reflect different interpretations and different bases of linkage. Prosodic features, such as stress, rhythm and intonation, are relevant to information processing, and "since the linkage between parts of a text reflects the build up of information ... it follows that prosody is a vitally important factor in textual coherence" (ibid, p.1443). Punctuation, looked at as a surrogate and inadequate representation of the widely differing prosodic patterns, can, in a written text, assign specific interpretations not only to sentences but to larger units (such as the paragraph) as well.

d) Grammatical devices: The treatment of this type of connection, when compared to its earlier version, is far more detailed. The types of grammatical devices are discussed under 9 main heads:

i) Place and time relators: Similar treatment to that in the earlier version, with additional notes on "distance indicators" (ibid, p.1450). Moreover, two classes of relators are discussed: a closed-class items, such as afterwards, and an open-class that have a quasi-grammatical status, such as "a long time, in 1985, all year".

ii) Tense, aspect and narrative structure: This refers to the fine distinction in time relationships by the discrete indication of time and aspect that all finite clauses carry. Although the temporal contrasts are limited, compared to the use of time relators, they contribute to textual cohesion and progression. Narrative texts, in particular, comprise great time-reference complexity: "the narrative will weave backwards and forwards, a mixture of tenses and aspects, of finite and non-finite clauses enabling the narrator to depart from the linear sequence of historical order so as both to vary the representation and to achieve different ... effects" (ibid, p.1455).

iii) Determiners, pro-forms and ellipsis: This refers to the role of substitution and ellipsis in sustaining clausal linkage.

iv) Discourse reference: This refers to the type of cohesive linkage performed when a signal marks the identity between what is being said and what has been said before.

v) The textual role of the adverbial: Certain adverbials, particularly conjuncts and some disjuncts, help sustain logical connection between one part of the text and another and assist in interpreting the text to the text receiver. Quirk et al. (1985) still categorise this group into four according to the broad meaning of the coordinators and, or and but, and the conjunction for, though they no longer subclassify the relations in minute detail. Instead, they place an emphasis on the role of the adverbials as indicators of relational structures. These are seen of three types:

a. General to particular: This relationship can be indicated by such connectives as for example, thus, even, indeed, where a text proceeds from a general point to a particular one. Note how each of these adverbials can assist the relationship at the marked point in this example (ibid, p.1470).

> [4.25] Many of the audience became openly hostile. My uncle wrote a letter to the management next day.

The text can proceed in the reverse direction (particular to general). Note the use of in fact in this example (ibid, p.1433):

> [4.26] The ordinary household saw is not easy to use. In fact, any sort of woodwork calls for great manual skills.

b. Progression: Adverbials help to indicate the direction and mark the successive stages of progression, whether it is locational, temporal or logical. For example (ibid, p.1471):

> [4.27] First, boil the rice in well-salted water; drain it immediately. Next warm the lightly buttered base of a small pie dish. You may now put the rice in the dish. Then add the cheese, tomato and onion. The pie is at last ready to be put in the oven.

c. Compatibility: Frequently, the interpretation of a text depends on recognising whether or not two parts are compatible with each other. Adverbials help mark the relation either of matching (e.g. in addition, so, too), or contrast (e.g. on the other hand).

vi) Coordination and Subordination: The use of both devices provides a text with a variety of expression and with a well-ordered presentation of information.

vii) The part played by questions: These are inserted in text in order to hold the text receiver's interest, or to provide a focus and so facilitate information processing. Questions may also serve a role as polite equivalents of requests, particularly in spoken text. Rhetorical questions are often used for communicative effect, for instance to confirm what has been explicitly stated or assumed.

viii) Participant involvement: This is indicated in the way the text producer addresses the receiver. Occasionally the relation of both is made explicit. For instance in legal documents identification and authority are made heavily explicit. Equally, however, both participants are left implicit. In many texts, the addressee is unmentioned and the text producer leaves it implicit to who is the authority for the communication (refer for examples in Quirk et al. 1985, pp.1479-1481).

ix. Information processing: This aspect of grammatical organisation is seen as the motivation behind other grammatical features, such as the use of coordination and subordination, rhetorical questions and (connective) adverbials. It is through this feature of connection that the text producer sets the information focus, prepares his text receivers for the text direction or re-direction, alerts them to the climatic strategies or the present alternatives, points out the evidence or counter-evidence, draws their attention to the causes, consequences or contrast. In fact, the text producer provides the essence of anticipation of the information.

These nine types of sentence connection indicate how intricate text organisation can be. But despite the detailed analysis that Quirk et al. offer in both versions, it suffers from a drawback. There is ample comprehensiveness of analysis, but very little synthesis. The various aspects of connection do not lead to formulating a theoretical basis that serve as a "proper" textual framework. This can be defended on the grounds that the account is mainly descriptive and is therefore in line with the main aims of the task of both 1972 and 1985 contributions. We can still draw a lot of assistance and insight from this detailed description in both our analysis and synthesis.

4.5 The Concept of the Connective

4.5.1 Introduction

In the previous chapter we argued that cohesion refers to a configuration of relations whose cumulative interaction helps a passage to function as text. In text actualisation, these relations provide the semantic structuring whereby text fragments (of various types: clauses, clause-complexes, paragraphs) can make meaningful wholes. This entails the existence of dependencies which can, on the surface level, be either expressed explicitly, via hard-core cohesive devices, or implicitly, via either soft-core cohesive devices or the mere juxtaposition of fragments of various ranks.

If text is viewed form an internal angle as a syntactic-semantic configuration, then the cohesive relations that hold its constituents together must also be syntactic-semantic, a point that is forcefully argued in Halliday and Hasan (1976). They maintain (for instance p.303) that cohesive devices are lexico-grammatical phenomena of one kind or another. And although we have stated that we view text from a broad angle that permits external features, it is the syntactic-semantic relations that figure predominantly in the discussion of cohesion, particularly in describing the relations expressed by connectives. External features are more applicable in describing other principles of textuality, for instance, coherence, informativity, intentionality and acceptability.

Accordingly, our discussion of the concept connective will focus on its cohesive role, its textual characterising features and the typology of its surface realisation. Characterisation of the semantic features will be discussed in Chapter 6. This is because we would like to discuss those features as they figure in the two corpora. In other words, the characterisation of the semantic relations that are expressed by connectives will not be imposed as a ready-made scheme of analysis. Rather, it will be derived from a study and analysis of the behaviour of connectives as they occur in the two corpora.

4.5.2 Nature of the Connective

Although connectives, along with other cohesive devices, are operative in relating one text component with another, they exhibit differences in textual functioning. Relations such as reference, substitution and ellipsis are concerned with establishing a semantic association between items in the text. They obtain when a particular item in a text is interpreted by reference to a previous (or subsequent) item (or object) within (though occasionally outside) the text. In text utilisation, the presupposing items, as they occur earlier in the text, are kept in storage and are later reactivated through the use of the presupposed items. Reference, substitution and ellipsis are, in this perspective, operations of reactivation. In the case of cataphoric reference, presupposed items occur first and are suspended or held, so to speak, on a stack to anticipate the occurrence of the presupposing item. The relation expressed by connectives, however, is not of that type. Referential meaning does not account for the textual force of this relation, though, on occasions, it may constitute part of it. This point will be shelved for the moment and discussed in a later chapter. We shall concentrate in the rest of this chapter two points: a) the nature of the connected units, i.e. the type of units or objects, the concatenation of which is signalled by connectives, and b) some general characterising features of the connectives.

4.5.2.1 The Connected Objects

One of the first, and obvious, requirements for a serious study of connectives, and of textual connectivity in general, is an understanding of the nature of the connected objects. A number of proposals have been suggested, all commensurate with the status and nature of text and other requirements of the various theoretical models. For instance, in linguistic grammar, the terms "sentence" and "clause" have been widely used to describe the connected objects. In Halliday and Hasan's influential model of cohesion, cohesive relations obtain between sentences, rather than between clauses that make up one sentence. In logical and semantic models, "propositions", as bearers of truth-values, are employed to refer to conjuncts, disjuncts, antecedents or consequents. We would like to show below that in a textual study these three terms are not necessarily mutually incompatible. They may in fact account for various aspects of the same unit.

For the purposes of this analysis, we would first like to assume that connectives operate on "propositions". The term "proposition" is vague and requires qualification. It is not employed here as it is used technically in logic and semantic. Rather, it is used in a "linguistic" sense, and somewhat informally, to refer to the conceptual content of a clause. Again, the expression "content" is not iconic with "propositional content" as used in logic and in some semantic models; rather, it is an abstract construct representing the "meaning" of a clause or a sentence.

The difference between sentence/clause meaning and propositional content is discussed in Lyons (1981). Lyons (p.119) stresses that "propositional logic fails to give a full account of sentence-meaning". For instance, one part of the meaning of a clause that is definitely not part of its propositional content is its "thematic meaning". The term "proposition", as we use it, is relevant to the function of the clause as a message, which has specific reference to processes, persons, objects, abstractions, qualities, states and relations of the text world.

In this conception, a proposition is seen as a configuration of knowledge with at least two concepts reflecting the theme-rheme, or given-new dichotomies, and hence corresponds to the informational structure of the clause. The proposition refers to the relation that obtains between the two concepts (i.e. conceptual structure) and determines how the concepts are mutually relevant. Syntactically, we would like to assume, a simple proposition corresponds to a simple finite clause with at least two major functional constituents, the subject and the predicator. Thus we shall regard the simple clause and the simple proposition as two aspects of the same entity: the former is surface/structural and the latter is logico-semantic/informational. Further, we shall reserve the term sentence to the physical manifestation of a simple clause or a clause-complex.

4.5.2.2 Some Characteristic Features of the Connective

The following comments are intended to outline the nature of the connective as a cohesive tool and its main characterising features.

To start with, we would like to assume that the cohesive force of the connectives resides in the relation they signal between one proposition or a sequence of propositions and another (proposition or sequence). Whether connectives have intensional or extensional meaning with reference to the propositions they attempt to relate has been a controversial issue. More linguists believe that the cohesive force lies in the juxtaposition of clauses and bigger sequences, and not in the connective per se. In Halliday and Hasan's description of conjunction, one of the key notions is that it is "the underlying semantic relation ... that actually has the cohesive power" (1976, p.229) and not the linguistic forms which explicitly signal it. Van Dijk (1977b, p.46) argues that "connection is not dependent on the presence of connectives" and that, conversely, "the presence of connectives does not make sentences connected". He concludes that "the connection between propositions is determined by the RELATEDNESS OF THE FACTS denoted by them" (p.47, his emphasis). Winter (1971, p.43), referring to Huddleston et al.'s (1968) large corpus of scientific English, specifies that "for every four possible clause relation pairs, one is accounted for in terms of its outer-clause relation". In other words, such relations as expressed by connectives are largely left implicit.

The essence of this view is that the logical relation between two (or more) clauses or bigger textual constituents is usually implicit and that in text utilisation, these relations can be recovered, even when they are not explicitly signalled, by reference to the textual world, to the linguistic, semantic and situational environment surrounding the propositions. It follows that the logical relation exists independently of any explicit cohesive signal and that, by implication, connectives have no meaning of their own: their meaning is predetermined by the linguistic environment.

This view of the connective is contradicted by Arapoff (1968) who argues that "sentence connectors frequently determine what the logical relationship between two sentences will be" (p.243). This argument is made on the grounds that "while one connector may create one set of suppositions about the truth or nature of the two sentences it connects, another connector will create quite a different set of suppositions about the same two sentences" (loc. cit.) It follows that "two sentences will have different semantic interpretations according to the sentence connector used to connect them" (loc. cit.).

In deciding a stand among these conflicting views, one is tempted (as, for instance, in Frankel 1977) to favour the first attitude where there seems to be a weight of evidence. However, we would like to approach this issue with some caution. There are a number of points we would like to put forward in considering the cohesive force of the connectives, all of which seem, in our view, to mark off the nature of the relation of connection.

a. We agree that juxtaposition of clauses can account for a logical relation between propositions. The relevance of juxtaposition for directing text linearity and for sustaining cohesion has already been pointed out (cf. 4.6 above). The relation remains implicit unless it is specifically signalled out via the use of connectives. But juxtaposition is regulated by a number of constraints in order that a relation can be specified. These constraints reflect types of relatedness that represent an essential aspect of conceptual connectivity. One type concerns the consistency of relatedness of individuals and objects with states, events anu situations. Another is the compatibility of the two propositions in relation to the textual world, definable as "the cognitive correlate of the knowledge conveyed and activated by a text in use" (Beaugrande 1980, p.24). A third type of relatedness is that of presupposition between the propositions, in its various interpretations: logical, lexical, textual and contextual (i.e. pragmatic). Presupposition, in text grammar, is defined in terms of precedence and identity of relations of propositions (cf. van Dijk 1977b). When a relation is explicitly signalled by a connective, it indicates that these conditions are satisfied and hence the nature of the relevance is specified.

b. Often the nature of the juxtaposition is such that a proposition is made dependent upon another, previous or subsequent. Conceptually, this dependency can take the form of relevant implication and entailment. Structurally, dependency of clauses is evident in subordination. The explicit use of a connective will specify the type of dependency, its range and complexity, not only structurally, but conceptually as well. This can be exemplified in the following:

[4.28a] Output must recover. Only then employment will pick up.

[4.28b] After output recovers, employment will pick up.

Although the temporal ordering in both examples is similar, the use of the connective in [4.28a] imposes a temporal dependency without the need of a structural embedding; in [4.28b] the connective "after" signals conceptual as well as structural subordination, where the clause is embedded in the main (superordinate) one, thus making a single unit (a clause complex) (cf. the discussion of types of nexus in Foley & van Valin 1984).

c. The choice of signalling relations between propositions or leaving them implicit is a rhetorical consideration and is governed by the organisational pressure of the text. Rhetorical choices in this respect are not always deliberately made. Rather, they are under subconscious review. Nash (1980 pp.40-44) shows, through a study of an essay, how, for instance, extensional terms of the kind that imply a point of view, an authorial stance (e.g. as far as ... are concerned, for example, beside) occur predominantly in the first and last paragraphs. This distribution may be symptomatic of a rhetorical pressure imposed by the text. In Nash's words (p.44), the writer "makes his entry and takes his leave in a style that suggests a sort of literary socialising; between the greeting and the leavetaking comes the matter of getting down to argumentative business". A different text, with different rhetorical requirements, will display a different distribution of connectives in the text.

Furthermore, we would like to suggest, and this study will set out to verify, that preference for explicit or implicit signalling of relations is language-specific. While a certain language tolerates implicit expression of connection without causing any rhetorical disturbance, and thus text stability is maintained, another language may display a lesser degree of tolerance.

d. Related to the last point is the rhetorical issue of clarity. Nash (1980) argues that there are gaps between sentences, not the intervals required of typography, but "spaces which interpretation must leap" (p.21). The absence of an explicit signal for the relation between two propositions may create ambiguity in interpretation. Such ambiguity stems from the possibility of having several potential relations, despite the fact that the text as a whole can make compensations in this respect. It is important to note that, on such occasions, as Widdowson (1973) demonstrates, the relation can be disambiguated via the language user's choice of connective to signal the relation that he intends. Widdowson (1979, p.130), referring to the arguments and exemplification in Krzeszowski (1975) (although mainly concerned with appropriateness of sentences under certain extra-linguistic circumstances) indicates that extra-linguistic circumstances can easily be imagined in which two utterances would be related if the connective introducing the second one changes, for instance, from therefore to nevertheless or moreover.

Nash (1980, pp.21-22) supports this argument by giving this example:

> [4.29a] Nottingham is a city which has undergone a great deal of 'development'. It retains some of the charm that once earned for it the title of Queen of the Midlands.

There is an ambiguity on how to relate these two sentences. The inverted commas round 'development' may give a clue that the word should be read pejoratively. In this case the relationship between the two sentences might be expressed in this way:

> [4.29b] Nottingham is a city which has undergone a great deal of 'development'. Nevertheless, it retains some of the charm that once earned for it the title of Queen of the Midlands.

However, this relation is not necessarily the one implicit in the two sentences as in [4.29a]. One can argue, despite the inverted commas, that the relation between the two sentences is to be understood in this fashion:

> [4.29c] Nottingham is a city which has undergone a great deal of 'development'. Consequently it retains some of the charm that once earned for it the title of Queen of the Midlands.

The connective consequently signals a different relation. Even the word in inverted commas now has a different meaning, not the implied attitude of disapproval as in [4.29b], but an implication that the word is now a term currently in fashion and that 'development' has positive results for the city of Nottingham.

Nash argues that the two sentences are separated by a gap which can be bridged in several ways, depending on the intended meaning. The example, despite its apparent simplicity, requires a directive clue, an overt signal of transition, to render it meaningful. In written texts, there are numerous instances where rhetorical ambiguity can only be removed by the use of appropriate connectives. In our view, text producers are sensitive to the possibility of such ambiguity and usually take steps to pre-empt it. Where the presence of a connective is crucial for a correct interpretation of the logico-semantic relation between a particular pair of sentences, a text producer will select an appropriate explicit signal. On the occasion when this procedure fails, unless a particular rhetorical effect is to be achieved, text continuity will be impaired, thus leading to a certain degree of discontinuity that can, in turn, disturb text stability.

e. The relationship between the propositional relation and the connective that signals it is very close and complex. The connective, in our view, is the most explicit means whereby relations are signalled. However, we accept that the meaning of a substantial number of connectives (but not all, see below) is also determined by the logico-semantic relations they attempt to signal. This is particularly true in the case of connectives that can signal more than one relation. Take the following two examples and examine the meaning of "and" in each:

> [4.30] [The emergency actions] have allowed the situation to be kept under control and have given policy makers a breathing space for a calmer assessment of the situation.
>
> (The Guardian, 14/6/1983)

> [4.31] Deny them this satisfaction and you really do condemn them to a life without joy or purpose.
>
> (The Sunday Telegraph 5/6/1983)

The propositional relation signalled in [4.30] is additive with some causal meaning imposed by juxtaposition of two propositions referring to two events. Hence, "and" is characterised as an additive connective. But in [4.31], the relation is that of a condition, where the first alternative sets a condition for the second to take place. The same sentence can be reproduced using "if" instead of "and" and the same relation can still be signalled. Hence, "and" is categorised as causal/conditional. It follows that while connectives, as linguistic representations, are important in capturing propositional relations, these in turn are useful in categorising the individual meaning of connectives. Only in this respect one can claim an intensional meaning for such connectives.

f. There is evidence in linguistics, for instance in Halliday and Hasan (1976), and also in the work on clause-relation conducted by Winter (1971, 1977, 1982) and Hoey (1979, 1983, cf. also Dea 1977, Jordan 1978, Edwards 1982, Hoey and Winter 1982 and Jordan 1984), that propositional relations and the connectives that signal them have no one to one relation. While a connective has a close and complex relationship with the propositional relation it signals, as we discussed above, certain relations can be signalled by other means. No one can formulate it better than Halliday and Hasan (ibid p.227) when they stated succinctly,

> "There is a range of different structural guises in which the relations that we are here calling CONJUNCTIVE may appear. These relations constitute a highly generalized component within the semantic system, with reflexes spread throughout the language, taking various forms; and their cohesive potential derives from this source."

Let us exemplify this by selecting the temporal relation of succession. In the following examples, [4.32b], [4.32c] and [4.32d] are possible renderings of the original text [4.32a]:

[4.32a] After the local election results are
announced, Mrs Thatcher will be
under intense pressure to declare
her intentions.
(The Observer, 24/4/1983)

[4.32b] After the announcement of the local election
results, Mrs. Thatcher will be under
intense pressure to declare her intentions.

[4.32c] The local election results being announced,
Mrs Thatcher will be under intense pressure
to declare her intentions.

[4.32d] Mrs Thatcher will be under intense pressure
to declare her intentions. This will follow
the announcement of the local election results.

In each of these examples the temporal relation appears in a different realisation, depending on the other semantic patterns with which it is associated. In the original example, the relation is structurally expressed as a relationship between predications, with one clause being contingent on another via the conjunction "after". The same relation can be expressed as a minor predication; that is, it may be realised prepositionally as in [4.32b] or participially as in [4.32c]. It can, further, be realised in two separate sentences, as in [4.32d], with one of them having a lexical marker of time sequence, in this case the verb "follow" in the second sentence.

There are, of course, other structural as well as lexical ways of signalling the temporal relation. They all represent methods of predicting or signalling time sequence and hence are considered cohesive agents. This has a number of implications.

i) Implicitness of a semantic relation between two or more propositions is, more often than not, only apparent, and refers only to the absence of a logical connective. Often there exist other explicit methods of signalling the same relation. The text receiver is able to recognise that the same phenomenon may be embodied in different structural and lexical constructions, and he is aware of the stylistic and rhetorical variations that accompany the choice of one particular method over another.

ii) If a comprehensive investigation of signalling of propositional relations is intended, then a study of the connectives will be insufficient. However, we are particularly interested in this project with one major method of signalling: the use of connectives. Discussion of other methods goes beyond the scope of this work.

g. Connectives, viewed from a broad textual perspective, display a duality of functioning in a text. They do not only assist in creating cohesion, they also help maintain text coherence. This duality has been observed by a number of linguists, and, therefore, before we take this issue any further, we would like to examine briefly some of the observations made.

Austin (1962, p.75) notes that particles (i.e. sentence connectives) may be used to make explicit the illocutionary force of an utterance (i.e. sentence-like unit). By way of exemplification, Austin states "we use 'therefore' with the force of 'I conclude that'" (loc. cit). In this role, as Mountford (1975) argues, connectives serve as "discourse markers of interactive structure, marking the propositional sequencing of propositional acts" (p.41). In other words, connectives, within Mountford's framework, help give discourse its coherence, a point that echoes Widdowson's (1973) view that the use of a connective not only establishes *cohesion* between two sentences; it also makes the nature of their *coherence* explicit.

Mountford's interpretation of this duality of role is that connectives "signal the procedures that language-users ... use to *make sense* both at a propositional level and in terms of the linear and hierarchical relationships of illocutionary acts" (ibid, p.142, his emphasis). These "interactive" procedures relate propositions to each other in two ways: as "sentence-like objects" by reference to the rules that govern the non-structural, text-forming component of the language system (i.e. cohesion), and as "locutions" by reference to the communicative rules by which "cognitive content and communicative value [are] conveyed" (loc. cit.) (i.e. coherence, in Mountford's description). Hence the textual role of connectives can be described in terms of interactivity.

In further describing the "interactive" acts that "hold between propositions and the illocutionary force" (ibid, p.143), Mountford explains that the signals of "interactivity" are "logical connections of various kinds realized by logical connectors which are, grammatically speaking, conjuncts" (loc. cit.). Although "conjuncts", if used in the same sense that Greenbaum (1969) and Quirk et al. (1972, 1985) use it, are not, in our view, the only structural realisation of connectivity (see Ch.6 below), the message that Mountford is trying to pass is clear, at least in terms of his model of discourse analysis. The various kinds of logical connection are "species of interactive act ... [which] may be signalled by particular connective devices" (loc. cit.).

Although these observations make interesting guidelines, it is not our aim to fuse incompatible approaches to text and discourse analysis with our own. However, two points are worth making in this respect. They concern respectively the function of connectives in text cohesion, which is our main concern in this project, and the role of connectives in manifesting text coherence.

i) The function of connectives is not restricted to signalling propositional relations between the adjacent clauses. That indeed is one aspect of its cohesive force. Another aspect, which has not been pursued in detail, concerns the role that connectives play in text linearisation: connectives respond to the way text components are organised in a linear ordering and can figure in sustaining such principles as grouping, sequencing and salience. Hence, in maintaining sequential connectivity, connectives perform a two-fold function: it relates propositions in terms of logico-semantic relations, and so consolidates the transition between one text sequence (clause or bigger) and another, and at the same time assist in achieving a sense of linear organisation. This function will be elaborated in later chapters particularly as it concerns the variations in English and Arabic.

ii) Connectives can maintain coherence in a variety of ways. We are, however, not going into the details of examining the illocutionary force expressed by each connective and the role it plays in illocutionary development. That can be traced elsewhere (cf., for instance, Mountford 1975, Widdowson 1973, 1978), although a specialised study for this topic is still required. What we would like to do is to point to the case where the use of connectives may help in capturing text conceptual connectivity. We have argued earlier that a proposition, as we employ the term in this study, refers to a relation that obtains between at least two concepts. A concept has fuzzy boundaries. It consists of a "control centre" in a "knowledge space" around which are organised whatever other basic components the concept subsumes (Beaugrande 1980, p.67, referring to Scragg 1976, p.104). To represent text underlying conceptual connectivity, one can, in agreement with Beaugrande (1980) and

Beaugrande and Dressler (1981), postulate a conceptual-relational network. The network consists of nodes and links (similar to Beaugrande and Dressler's grammatical networks we mentioned on reviewing their model of cohesion), and is composed of knowledge states.

Concepts in this network will occupy the nodes and look along all of its relational links in the knowledge space. Link labels announce the type of concept that is attained by traversing links in the most appropriate direction. The processor works from a current state to a following one by trying to identify the type of the node to be attained.

When one conceptual configuration (corresponding to a proposition) has been traversed, another is immediately set up. Often, with the absence of cohesive devices, particularly connectives, the underlying configuration of the next proposition (or set of propositions) is not readily integrated. In such cases, inferencing must be made to bind things together. This operation demands the provision of reasonable concepts and relations to fill in a gap and so stop a discontinuity in a textual world, a task that can efficiently be performed by connectives.

The function of connectives in such a network starts in marking and directing access routes from certain nodes to others. They serve to speed up access by reducing unnecessary processing expended on inferencing, testing hypotheses and setting preferences. Thus a space is bridged where a pathway might fail to reach. Once this takes place, text world knowledge is updated and the number of the next possible links are restricted and identified, or at least made predictable.

This network may sound simplistic, and indeed a number of factors, stylistic and/or rhetorical, have to be incorporated. Moreover, further work is required to find out whether such a network and the role of connectives is psychologically plausible, i.e. whether text users have a uniform threshold for noticing and filling in gaps and discontinuities. To quote Beaugrande and Dressler (1981, pp.102-103), a number of questions can be posed and therefore research is needed to answer them; for instance:

> "... how similar are the textual world of the producer and that of the typical receiver? Do they, for instance, agree on what is or not worth mentioning? Are there major differences in the richness of their mental representations for text world situations and events?"

In the current state of the art, these questions, and many more relevant ones, are far from answered.

4.6 Conclusion

The study of connectives is not a new or recent undertaking. It originated with classical grammar in both the Western and Arab traditions. The classical approach to English grammar has been moulded on the Latin tradition with some modifications that have been introduced gradually to match the peculiarities of the English language. Within this framework, the English connective has been studied under the heading of "conjunction", and is generally used to refer to coordinators and subordinators. In Arabic, connectives are studied under the heading of "<aṭf nasaq" [connection of sequence].

This type of connection is equivalent to coordination in Arabic and is signalled by a limited number of particles "huruf al-<atf", the function of which is to connect two items of "equal status": two words, phrases or clauses, and where the case or mood of the consequent item (when it is inflectable) follows (i.e. is identical with) that of the antecedent. Subordinators in Arabic are studied under the heading of "condition" and hence only a limited number of such connectives have been systematically analysed.

The inadequacy of the traditional approach to describe connectives is attributed to a) its inability to recognise the role of sentence connectives in relating sentences, and bigger sequences, to create text; b) its vague criteria in describing conjunction; c) the lack of interest of Arabic traditional grammar in studying subordination (except condition) and inclusion of a number of subordinators in the study of adverbs or particles, and d) the traditional Arab grammarians' particular interest in one aspect of clause connection: the morphological status of the connected items, i.e. the morphological changes that various particles of connection can inflict upon the connected items, their case or mood. Hence the traditional approach, despite its merits, does not provide a model that can serve as a fully-fledged and generally applicable theory of textual connectivity.

The task of modern sentence linguistics is to provide a general theory of sentences and to produce language-specific grammars which can account for all possible sentences of the particular language. The term sentence is used here both as a theoretical entity and as an empirically given string. Connectives have been studied only within this perspective.

Since connectives function as a cohesive device in relating text components and maintaining text cohesion, it follows that the scope of sentential models is too restricted to account adequately for text connection. Indeed, the textual role of the connectives, their rules and constraints have (to a varying degree of adequacy and delicacy) only to be handled within a text linguistic framework. Models of cohesion (such as the stylistic, functional-systemic, stratificational and procedural-relational) have all discussed the phenomenon of connectivity, the most comprehensive as well as influential of all being perhaps the functional-systemic model.

We proceed, then, within a textual framework, to identify the concept of the connective. In carrying out this task, we are required first to unify the various conceptions of the connected unit. We assume, in this study, that the semantic role of the connectives is to signal the relations that obtain between one proposition or one sequence of propositions and another. The term proposition is used in an informal manner to refer to the content of the clause. In this way, "clause" refers to the formal aspect of the connected unit, "proposition" to its content, while we have reserved the term "sentence" to subsume the physical manifestation of the clause/clause complex, a manifestation with two important markers: capitalisation of the first letter (applicable to English) and termination with an appropriate sentence terminator (applicable both to English and Arabic).

The cohesive force of the connective is seen in the textual role that it exercises in organising text. Connectives specifically signal relations between propositions that would otherwise remain implicit. The preference towards signalling relations is constrained to a great extent by rhetorical as well as informational (and other textual) considerations. We would like further to suggest that the way connectives are distributed is language-specific: while greater degrees of implicitness of intersentential relations are tolerable in one language, in another they may severely disrupt text stability, affecting both text cohesion and coherence. This is one of the main tasks that this project is aimed to verify in relation to English and Arabic. The investigatory apparatus is intentionally designed to identify, categorise and measure connectives in a large body of text. This involves using computer techniques, the description of which will be reserved to the next chapter.

Footnotes to Chapter 4

(1) The truth value of the conjunction holds no matter how complex the compound statement may be. If we had a conjunction made up of twenty conjuncts, then in order for the conjunction to be true, all of its twenty conjuncts would have to be true. If only one of them is false, then the whole conjunction will be false.

(2) Earle (1871, p.434) admits that this term is borrowed from E. Thring's "On the Principles of Grammar" (The Clarendon Press, Oxford), but maintains that he has varied the scope of the term.

(3) Vorlat (1963, Vol.I, pp.87-89) summarises the problems as follows: a) the words which are declinable in Latin are not necessarily so in English; b) English has a part of speech, i.e. the article, for which there is no Latin equivalent; c) some grammarians prefer to follow a different description of nouns and pronouns; some would even like to include "expletives" as a class.

(4) The first type of connection "<aṭf bayān" [Explicative Apposition] is the asyndetic connection of a noun with a preceding noun, which it more nearly defines, e.g.

> Yūqadu min šajaratin mubārakatin zaytūnatin.
> [It is] lighted by (the oil of) a blessed tree, an olive.

Here "zaytūnatin" [an olive] is in explicative apposition to "šajaratin mubarakatin" [a blessed tree].

(5) To the best of my knowledge, only two such studies have been made within the traditional framework: one by Barakat (1977) who studied condition in an old tribal dialect of Arabic, and the more comprehensive study by Al-Muttạlibi (1981) of condition particularly in modern Iraqi poetry.

(6) In order to keep the review within a manageable limit, we shall not consider, though we are fully aware of, treatments such as Quasthoff (1982) (within the cognitive framework), Posner (1982) (within the poetic framework), Warner (1981) and Burtoff (1983) (textual/logical studies), Ortner (1983) (a textual/grammatical account), and Cohen (1984) (a computational approach to the function of "clue words").

CHAPTER FIVE

Automation and Processing of the Corpora

5.0 Perspective

This chapter gives a description of the various stages of the experimental work, particularly the steps taken in the automation of the two corpora and the computerisation of the linguistic analysis. It is, in other words, a description of the investigatory apparatus in action. The chapter is written with two general aims in mind. First it discusses the type of the methodological decisions taken and the rationale behind each. Secondly, it articulates the various problems that have emerged in conducting the project and the procedures and alternative steps that have been adopted to surmount them.

More specifically, we would like to achieve the following main tasks:

1. A discussion of the type of linguistic data used in this project and its relevance to the linguistic analysis. This is a reflection of the controversy that is still raging in linguistics over the place of introspection as opposed to a corpus-based analysis.

2. A brief description of the pilot experiment which was conducted at the early stages of the work for the purpose of verifying overall feasibility and assessing the requirements of the project.

3. An account of the procedures we have followed in assembling the two corpora in machine-readable form. This involves discussion of sampling procedures and methods of data encoding, and consideration of the problems that accompany automating Arabic script.

4. A discussion of the various stages of text processing, the type of tagging procedures used and the production of word lists and concordances.

5. A brief critical survey of the software tools used in linguistic analysis including programming languages and packages, particularly concordance generators. This is followed by an evaluation of the merits and weaknesses of OCP from a user's perspective.

5.1 Intuition vs. Corpus Linguistic Analysis

Although the long-standing debate argument in favour of intuition or corpus as an empirical basis for linguistic investigation has now abated, it must, nevertheless, be noted that it is still central in any meaningful discussion of the choice of the analytical apparatus. In this section we would like to present our position, since this, we believe, will place in a proper perspective the primary descriptional instrumentaria used in the project.

Data gathering and the use of corpus analysis have been employed for years as procedures for linguistic analysis. This position was challenged by Chomsky's now thirty years old formulation of the "fundamental aim" of a grammar, which, Chomsky claims, should account for "all and only the grammatical sentences of a language" (1957 p.13, cf. Bach 1973 p.5) "else it could not be called a description at all" (Lees 1957 p.382). With this formulation came the demand for intuition-based hypotheses or theories that are falsifiable only by producing crucial intuition-based counter-examples.

Chomsky and his exegetes have often linked this argument with their opposition between competence and performance. Competence is regarded as the set of possibilities given to the native speaker-hearer of a particular language owing to the fact - and this fact alone - that he has mastered that language (cf. Ducrot and Todorov 1979). These possibilities include constructing and recognising an infinite number of grammatically correct sentences, judging their acceptability, perceiving their meaningfulness, identifying the ambiguous ones, acknowledging that some sentences that may sound different can have a strong grammatical resemblance while others that have similar realisation can be grammatically dissimilar.

Performance, on the other hand, is seen as a specific set of utterances produced by native speakers. These can suffer from having features that are irrelevant to the abstract rule system of the language, such as hesitations and unfinished structures, arising from a variety of psychological and social constraints acting upon the speaker.

While intuition is linked with competence, corpus is linked with performance. According to Chomsky and his disciples, those "flashes of insight, those perceptions of pattern, which mark off the brilliant scientist from the dull cataloguer of data" (Lees 1957 p.380) were to replace corpus-related elaboration of linguistic phenomena (cf. the discussion in Ringen 1977).

In order to specify our position as regards this controversy, we should like first to distinguish between two kinds of aims in linguistic investigation. The first concerns the establishment of grammaticality and, explicitly or by implication, non-grammaticality (Ulvestad 1979 p.90). In this respect, intuition-based hypotheses are better suited for investigatory purposes. Normally these hypotheses rely on elicitation procedures that require organising "intuitional" tests. The tests can either be achieved through a monointrospectional method, i.e. the linguist uses his own or an informant's presumably competential intuition for arriving at the required answers, or a multi-introspectional method, by which the intuition of several native speakers is resorted to.

The second type of aims revolves around the description of a particular linguistic phenomenon, its manner of occurrence and recurrence, taxonomy of classes, and within-class variations. Such a description can be most efficiently achieved through observation based on empirical data, which usually take the form of a linguistic corpus. Quantitative procedures, while of no concern to the first type of aims, are fundamental to the second.

To this latter type belong the aims of this study. Our main concern is the provision of a description of the functional role of connectives as cohesive means, their occurrence and co-occurrence.

This task demands corpus-analytical routines as fact-producing procedures in order to enable the analysis to be initiated and concluded in a systematically meaningful manner. However, these routines need occasionally to be interfaced with introspectional procedures. This helps establish symmetrical regularities and constructional impossibilities with a fair degree of certainty. In particular, introspectional procedures are employed in problematic instances where functional typology of connectives is fuzzy and therefore demands to be closely and carefully monitored.

Since this study intends to carry out a contrastive analysis of a textual nature, the need is, as has been argued in Chapter 2, intensified for the use of two parallel corpora where the phenomenon of connectives can be observed, measured and compared. It is then that significant differential statements can be formulated with a high degree of confidence. To ascertain the validity of these and other views a pilot experiment is first conducted before any serious effort is expended on sampling the corpora. This is described in the next section.

5.2 Pilot Study

5.2.1 Aims and Set-up

The pilot study represents the first step in the project programme and was carried out with a few aims in mind:

1. It determines the feasibility of the project as a whole. Since the linguistic task of contrasting the system of connection in English and Arabic requires detailed preparatory work, a pilot

experiment helps to specify if this task is feasible before any step is taken in the direction of data collection and data processing.

2. It checks the functioning of the various aspects of the investigatory apparatus and determines what procedures may promote its efficiency in carrying out the rest of the project.

3. It assesses the capabilities and limitations of one of the packages designed for linguistic computing and decides, in the light of the results, the amount of programming needed to supplement the package.

4. It suggests statistical procedures that may be used as an integral part of the calculus of connectives.

For a pilot study, it was decided that two to three texts are to be used, totalling approximately 1,000-1,500 words in English and a similar size of text in Arabic. The selection of text was governed by two parameters: a) only signed newspaper articles were to be selected, and b) the selection was random. Accordingly, two newspaper texts in English and three in Arabic were selected. The total number of orthographic words in the English texts was 1,431 and in the Arabic texts 1,120.

These texts were keyboarded on a micro and then transmitted to the mainframe computer. OCP (the Oxford Concordance Program) was run to produce word lists and concordances, first for the entire data and then for connectives. Calculations were then made to compare the distribution of connectives in both sets of data.

5.2.2 Results

Results of the pilot study, despite its limited size and scope and the simplicity of its procedures, have indicated differences in the general distribution of connectives in English and Arabic, and in the distribution of functional categories. However, the main contribution of the pilot study is methodological, i.e. the provision of guidelines for the main project. Throughout the various stages of the study, further research has suggested itself at almost every turn, and it has been apparent that more work is required in order to determine more specifically the pattern and structure of the English and Arabic connectives and to assess more accurately their magnitude and scatter and their general pattern of behaviour.

Some of the guidelines that concern computer techniques are:

a) A large database is required for the main experimental work. The bigger the database is, the more informative the results are, and the more accurate is the representation, the image or the replica that the theoretical description offers of the underlying reality.

b) Computer techniques can be used to achieve automatic identification and categorisation of connectives and for producing their quantitative profile.

c) More elaborate statistical procedures than the ones used in the pilot experiment are needed for the purpose of specifying regularities and patterns of behaviour of connectives, and for arriving at a more competent contrastive account.

On the whole, the results have been seen as valid enough to justify further research in the field in the hope of arriving at an integrated and axiomatised theoretical construct. In the light of this, phases of the experimental work were planned and their requirements were considered and implemented. This will be discussed in the rest of this chapter.

5.3 Criteria for Assembling the Corpora

5.3.1 Introduction

A general principle that was adhered to throughout this stage of the work is comparability of the corpora, that is, there should be a shared similarity in terms of size, text type and composition. This requires laying a set of criteria that can serve as guidelines for selection and compilation and give the corpora their distinctive characteristics. The following set represents the main criteria followed in this project:

1. Text type: This controls selection according to three specifications: a) type of Arabic vs. English; b) type of genre; c) type of exclusions.

2. Synchronicity: This controls the period in which the texts have first appeared (i.e. published).

3. Size: This controls the length of the corpus as a whole or its subdivisions according to a standard unit of measurement.

4. Presentation: This controls the method of arrangement of the various subdivisions and components of each corpus.

5. Automatic accessibility: This controls the format in which the corpus is to be assembled.

In addition to these there are statistical considerations that have been followed during the process of selection. Each of these criteria will be discussed in turn below.

5.3.2 Type of Text

5.3.2.1 Type of Arabic vs. English

The term "Arabic" is not applied to one unified language, and hence any linguistic investigation of Arabic has to specify at the outset the form or version that is studied. It should be noted that within the Arabic speaking world there is a considerable dialect diversity that distinguishes the spoken mode of the language.(1) These dialects represent localised varieties employed in the speech of everyday life and display many and sometimes substantial mutual differences.(2)

In addition to the regional dialects, there is a superposed "literary" language which may be divided historically into three periods:

a) Classical Arabic: this is the Arabic used during the sixth through the thirteenth, and even later, centuries. It is the language of literary output, including pre-Islamic poetry and orations, the language of the Holy Qur'an and the great bulk of all that has been considered best in Arabic literature.(3)

b) Medieval Arabic: this continued from the thirteenth to nineteenth centuries, an age of decadence and declining social and political importance in the Arab world.

c) Modern Standard Arabic (MSA)(4): this form started to develop in the middle of the nineteenth century when a literary renaissance took place, marked with certain very notable trends toward breaking with the forms and ideals of Medieval Arabic. Modern Standard Arabic is, however, still modelled in its main features on Classical Arabic and therefore, whatever the differences between the written Arabic of these three periods, there are strong bonds of continuity in grammar and in a substantial size of vocabulary, with the Qur'an still considered the acme of perfection.

Today, MSA is the normal vehicle for all written communication, and the variety that is used throughout the whole Arab-speaking world. It is the language of literacy, and is therefore considered of high prestige, the counterpart of the elevated literary language of thirteen centuries duration. Although, strictly speaking, it has no native speakers, it is still the language of formal speeches and religious sermons, and the language often used in radio broadcasts, particularly those aimed at a pan-Arab audience.(5) MSA is the variety that is studied in this project.

5.3.2.2 Source of Selection

The corpora are based on a selection of newspaper texts. We assume that the material is generally representative of present-day written English and Arabic, a position that is similar to the one adopted by Allén (1970) in compiling the frequency dictionary of present-day Swedish and by Knowles (1981) in compiling the Polish frequency dictionary. In general, newspaper prose is of particular interest since it is distributed through an important mass medium, a vehicle for the transmission of information, expression of views and for entertainment. Newspapers form part of man's total environment in mass society and therefore cannot be ignored without conscious and sustained effort. Far more commonly, their use engenders a degree of passivity which makes them efficient for the moulding of tastes and preferences.

However, it must be admitted that authentic texts in general reflect the topic, genre and the context in which they are produced, as well as the typology of writer-reader relation, and are therefore linguistically constrained by these factors. Hence they are incapable of reflecting exactly the language system in its totality, a problem that is standard in corpus linguistics and is manifested specifically in deriving the lexicon of the language from a text corpus. Consequently, it is important to proffer an accurate description of the composition of the corpus before any analysis is initiated.

The English corpus is based on a sample of articles from a number of quality newspapers all published in Britain: the Guardian, the Times, the Sunday Times, the Daily Telegraph, the Sunday Telegraph and the Observer. The Arabic corpus is based on a sample of articles from six morning papers: Al-'anba' (Kuwait), Al-Ahram (Egypt), Al-Thawra, Al-Jumhuriyya (two Iraqi newspapers), Al-Arab and Al-Sharq Al-Awsat (published in London with readership in and outside Britain).

The texts are all signed newspaper articles. This criterion has been adopted on the assumption that such texts have more author individuality and suffer less editorial intervention and revision than unsigned texts which, particularly in Arabic, can be the product of an editorial team work or a translation, acknowledged or unacknowledged, of articles published in other languages.

The genre that characterises the texts is predominantly argumentative. The goal state of such texts is the inducement of shared belief, and this is usually achieved via different rhetorical strategies. Normally, journalistic texts of this type ought to show highly developed techniques for controlling the focus of attention, upholding interest and maintaining effectiveness while arguments are introduced and discussed. There are, for instance, grouping and sequencing techniques that monitor the assignment of control centres in the text world and reflect that on the manner in which propositions are arranged and related, and the way the flow of informativity is upheld or withheld. Superposed on these techniques is a calculated activating and subsequent overturning of reader expectations on the conceptual level, with the result that the text world undergoes revision during the process of constructing or reconstructing the arguments. The operations are all aimed at culminating in achieving persuasion. On the surface level, these operations are signalled by a density of expressions, of which one important type is connectives (see Chapter 4).

The topics of the texts are diverse, covering a wide range of political, economic and social subjects. The political subjects comprise articles on current politics at home and abroad, political organisations, local affairs and international questions. The economic subjects comprise articles on commerce, industry, agriculture, monetary problems, financial policies, world trade and international economic crises. The social subjects comprise articles on the family and home, private persons, women, improvement of environment and leisure time activities. However, owing to the period of publication, two, then current, topics, both political, are more recurrent than the others. One is the June 1983 election in Britain (in the English corpus) and the other is the Israeli invasion of Beirut and South Lebanon (in the Arabic corpus).

While the collection of articles manifestly comprise a variety of topics, each corpus is nevertheless uniform as regard the language used. It seems therefore reasonable to assume that they represent written standard English and Arabic.

5.3.2.3 Texts Excluded

In selecting the corpus, certain types of texts have been excluded. The criteria for exclusion have been as follows:

1. All anonymous articles and contributions are left out. Selection is limited to articles that bear the name of author.

2. All translated articles as well as articles written by non-native (i.e. non-British or non-Arab) writers have been excluded.

3. Excluded also are texts that represent telegrams from news agencies. Such texts, particularly in Arabic, are often translated from other languages.

4. All advertisements are omitted. These often represent their own genre and deserve a separate study devoted entirely to investigating their textual properties.

5. Narrative texts have been avoided. Selection has focused on expository prose, particularly argumentative texts.

6. Sports articles and art reviews are excluded on the grounds that in Arabic newspapers such texts are normally left anonymous. In addition, Arabic art and sports reviews sometimes include extensive translation of reviews of international sports or art events published in foreign journals. This has cast some doubt on the representativeness of the textual content of these articles of Arabic stylistic/rhetorical organisation.

7. Light and informal writing has been left out. Selection is based on the more "serious" type of writing. But even here we may occasionally have some informal, substandard or odd expressions. These have been retained, since they reflect the author's individual choice.

8. Articles with instances of meta-language or with long quotations have been left out.

9. Texts that represent letters from readers have been excluded. We were not certain of the authenticity of such texts, nor of the amount of editorial intervention involved.

10. Excluded also are all texts that have missing material (words or lines that cannot be reconstituted).

5.3.3 Synchronicity

To satisfy this criterion, a period of time had to be determined for text collection. It has therefore been decided that each corpus should be based on texts selected from newspapers published between the period of October 1982 and July 1983. The texts thus reflect the most current mode of writing and, particularly in the case of Arabic, cannot be accused of having outdated styles.

5.3.4 Size

Each corpus comprises ca 250,000 words of running text. We feel that this volume of material is large enough to a) yield data that help characterise the textual patterning of most connectives, and b) permit a practical understanding of the statistical behaviour of connectives. However, we have to admit that the corpus, as a sample of the universe, is still too small to provide definite facts about infrequent or rare connectives, or about certain possible patterns of some more common ones. But in order to increase the number of types so that rare connectives or usages can be accommodated, we have to enlarge the volume of the corpus to the extent that we would not be in a position to produce the detailed analyses, both qualitative and quantitative, that we have now made. Hence, for statistical and pragmatic reasons, we have to accept the present size of the corpus as a compromise.

Each corpus is a collection of texts. We have avoided segmenting the texts or using text portions of equal size to build up the corpus, a procedure that is followed in assembling the Brown corpus (Kučera and Francis 1967, Francis and Kučera 1982) and the LOB corpus (Johansson 1978, Johansson and Hofland 1980). Instead, we have decided to include the full length of the text. It is true that this procedure has the disadvantage of creating discrepancies both in text size within the corpus itself and in number of texts across the two corpora. Yet, this procedure has the advantage of retaining text unity, an essential requirement in the investigation of textuality, particularly cohesion and coherence. The textual role of connectives can then be observed in the whole length of a text, rather than in an arbitrary selection of a portion. In addition, this will create a better basis for studying macro-structures, rhetorical factors in text development or the integration of textual components, as a few possible tasks for future work.

The English corpus is composed of 254 texts. The shortest text is 187 words, while the longest is 2904 words. The average text length is 1007 words. In comparison, the Arabic corpus comprises 220 texts. The shortest is 173 words, the longest is 4872 words. The average text length is 1165 words. Lists of text sizes (in words) for the corpora are given in Appendices (8) and (9).

5.3.5 Organisation

There are different possible ways of organising the texts in the corpus. Some of these are:

a) arrangement according to the alphabetic ordering of text authors' names;

b) chronological arrangement (according to date of publication);

c) arrangement according to newspaper;

d) arrangement according to text length in terms of number of words or sentences (either ascending, starting with the shortest and ending with the longest, or descending starting with the longest to the shortest).

These possibilities were considered and we have opted for an overall arrangement by newspapers, ordered themselves alphabetically. Accordingly, texts in the English corpus are arranged in this order: the Guardian, the Observer, the Daily Telegraph, the Sunday Telegraph, the Times and the Sunday Times. Texts in the Arabic corpus are arranged in this order: Al-Anba', Al-Ahram, Al-Thawra, Al-Jumhuriyya, Al-Sharq Al-Awsat, and Al-Arab. Within each group, texts are arranged chronologically according to date of publication.

5.3.6 Automatic accessibility

It was decided at the outset that the two corpora are to be assembled in machine-readable form. This procedure is partly related to the aims of the study (see Chapter 1); but there are two other factors that are related to the performance of the investigatory apparatus: efficiency and economy (see also the discussion in Gillow 1979 and Sinclair 1982). In general, the automation of the two corpora can serve a number of purposes that are now well known for computer users:

1. Provision of storage: Magnetic storage has made it possible to store large volumes of text in a convenient and efficient way. Some relevant features of this type of storage are: permanent retention of data, reduced overall physical size, and speed of access.

2. Retrieval: Having a text data-base facilitates the extraction of various types of data to produce a certain piece of information (e.g. exemplification, occurrence of a pattern or recurrence of a function). The same information when produced manually will be cumbersome, time-consuming, and prone to errors.

3. Computerised analysis: This is related to the aims of the project. Computerisation of analysis is only feasible when the texts to be analysed are structured as a data-base. Programs and packages can then be run to achieve the variety of tasks required by the analysis.

4. Use in further research: a corpus in machine readable form can be used for future work, e.g. expansion of the present project or initiation of other corpus-based projects.

5.4 Sampling

There are two sets of methods often used in linguistics for deciding how to select texts for sampling:

1) simple random sampling

2) stratified sampling

A sample is said to be a simple random sample when in making it each text (member of the universe) has the same chance of being selected and if subsequent selections are independent of each other. The method often involves the use of a table of random numbers for determining which numbers to select. However, using random numbers in this project runs into two types of problems:

a) Since the method assumes the availability of a large volume of text that satisfies selection criteria, it is particularly useful when extracts (e.g. pages or lines) from a fairly long text (e.g. a book) are to be selected. In this project conditions are different. The universe of the quantitative description does not comply in every respect with the requirements of statistical collectives for which simple random sampling would be adequate. Our sample comprises whole texts selected under restrictive criteria (See 5.3 above). Problems start when, in applying random numbers to the source of population, we arrive at texts that do not satisfy these criteria and have therefore to be rejected.

b) Occasionally, we have also to reject texts sampled via random numbers for a number of unpredicted reasons, e.g. when such texts contain confused paragraphs or missing material that is too hard to reconstitute (a problem that is not rare in newspaper texts), or when, in English, a text contains too many dirty, blotted or blurred marks for optical scanning to operate efficiently.

For these reasons, random numbers were not used. We then considered the second type of sampling method: stratified sampling.(6) Here, according to Herdan (1966), the idea is to spread the sample in a methodical way, i.e. as uniformly as possible, over the whole of the work. By way of exemplification, Herdan states (p.96) that

> "we may adopt some arbitrary numerical rule for choosing lines or words from each page, for instance, 1st, 3rd, 5th, etc. line or word, or we may collect for our sample the first and the last n lines or words per randomly selected page."

He, then, demonstrates, by referring to a word count of Pushkin's famous story "The Captain's Daughter", that stratified sampling of this kind does not produce statistically inadmissible deviations, compared with results produced by random sampling.(7)

This method essentially relies on the assumption that a spread sample of extracts from a long continuous text (e.g. a novel) is more representative than it would be if large contiguous pieces of writing (whole chapters or texts) were selected. Since, again, the universe of population in this project does not comply in all instances to the requirements of stratified sampling, we have to opt for a compromise: a method that simultaneously makes use of randomisation and stratification in sampling.

The procedure we have used can be summarised in the following steps:

1. All texts that satisfy our selection criteria were detached from their sources and arranged without imposing a specific pre-determined order.

2. Texts were then divided into 3 groups - an arbitrary number.

3. Texts in each group (A,B,C) were serialised.

4. In each group sampling was made in a methodical way: in group A, each alternate text was selected (1st, 3rd, 5th, etc.). In group B, each 3rd text was sampled, starting with the first one (e.g. 1st, 4th, 7th, etc.). And in group C, each 4th text was sampled, starting with the first one (1st, 5th, 9th, etc.).

5. Texts that had been selected were then given new serial numbering and input in that order on the computer. Programs that keep track of word token counts made it easy to stop inputting after a pre-determined number of word tokens was reached (ca 250,000 words). This step will be discussed in detail later. The rest of the texts were discarded.

6. Texts that had been input on the computer were then ordered according to the organisational scheme suggested in 5.3.5 above.

7. Steps 1-6 above are followed in sampling both the English and Arabic corpora.

Basically, this procedure reflects a kind of stratification whereby sampling is made on the basis of some arbitrary numerical rule. In addition, we would like to claim that randomisation exists and is inherent in the procedure itself, despite the fact that no use of tables of random numbers has been attempted.

In arguing for the random nature of this procedure we have relied on views first suggested by Kendall and Babington Smith (1938) and expanded by Herdan (1960). Kendall and Babington Smith define the operational concept of random sampling as a method of sampling for a specified characteristic such that the method is independent of the characteristic itself. Additionally, they maintain that a random method cannot be considered apart from the universe whose members are to be selected. This is justified on the grounds that within the same universe, a method which is random in respect of one characteristic is not necessarily random in respect of another.

Herdan elaborates these views further by arguing that "provided the method of sampling was independent of the characteristic for which we sample, it does not matter how regular or systematic the procedure is by which the individual samples are selected" (1960 p.115). This is, Herdan explains, what von Mises (1939) has called the "Prinzip des ausgeschlossenen Spielsystems" (the principle of the impossibility of a gambling system), which he (i.e. von Mises) considers as the most important criterion of a random series. Herdan adds (loc. cit.):

> "If the universe is a random aggregate of events which may occur in two alternatives, say _A_ and _B_, then no matter what elaborate system one were to devise for getting the better of chance, _provided only that the method did not imply a knowledge of the characteristics A and B themselves,_ the resulting selection of items would again be a random series or random sample ...
>
> My contention is that the linear sequences of linguistic forms in written texts or speech are random series with respect to certain quantitative characteristics, and any sampling procedures, be it by disconnected linguistic units, or by continuous pieces of text, by pages or by chapters etc., will give a random sample of such a quantitative characteristic, provided only that the sampling method is in no way connected with the characteristic; that is, provided that it does not consist in a direct or indirect selection of categories of just the characteristic one is sampling for." (His emphasis)

In the light of the above views,[8] we would like to maintain that our sampling method would produce a random sample since it does not involve direct or indirect selection of connectives (the object we would ultimately like to describe) whereby some connectives are included and others neglected. That is to say, there is no bias with regard to the type and frequency of connectives which are comprised in the sample.

Having discussed the general criteria for selection and the nature and steps of sampling methods, we next describe the conversion of the corpora into machine-readable form, discussing the problems that we have faced and the solutions suggested. This will be preceded by a short description of the computer facilities used for inputting and processing the corpora.

5.5 Hardware Facilities

Hardware available at Aston University that has been used for automating the corpora and for various steps in texts processing include the following:

1. Harris system, comprising two computers, a Harris 500 and Harris 800, each running the VOS operating system. Each processor has its own systems disks but both access the two 0.6785 Gigabyte disks used for user files. The H500 is rated for power at approx. 0.75 Mips and the H800 at approx. 1.5 Mips, a total of 2.25 Mips. On this system, the corpora were first assembled, proofread and edited and the experimental and initial stages of processing, particularly using OCP (see below), were performed.[9]

2. The new VAX Cluster, comprising two VAX 8650 computers and 9.6 Gigabytes of disk storage, and running the VMS operating system. The total power of the system is approx. 13 Mips. The cluster system was installed during the first quarter of 1986, in time to do all the final programming and to have some major OCP runs.(10)

3. VAX 11/750 Supermini computer has 0.5 Gigabytes of filestore and is rated at 0.7 Mips. This system was used for testing and performing some SPITBOL programs (see below), particularly before the installation of VAX Cluster.

4. KDEM (Kurzweil Data Entry Machine): an optical character reader provided with a built-in lexicon of about 33,000 English words. The machine was used as an input device for the English corpus (see below).(11)

5. IBM PC-XT with a 512K RAM, 360K floppy and a 10Mbyte Winchester, used mainly for inputting certain parts of the corpora and for initial editing and proofreading. It has also been used extensively for testing some SNOBOL programs. This machine, among several others, constitutes the main equipment in the IBM laboratory in the Department of Modern Languages at Aston University and were made available during the early stages of the project (particularly in 1983).

6. Several Superbrain-QD machines with 64K RAM and two 338K floppy disks used at the early input stages of the project, and later as terminals for certain interactive tasks with the mainframes.

7. Apple Macintosh with a 512 RAM and 800K floppy used mainly for producing statistical diagrams. Two packages have been used for this purpose: Statswork and Microsoft Chart. Output has been printed on a laser printer.

5.6 Data Input

5.6.1 Preliminaries

The first step in computerising a corpus of text is its input on the computer, a requirement that is not free of some special problems. The task itself used to be time-consuming in the days when data coding was performed manually. At present, with the arrival of optical scanning readers, a substantial amount of chore has been eliminated. Texts can now be input rapidly and efficiently with minimum requirement for editing and proofreading. However, there are particular problems, some inherent in the physical appearance of the text that constitute the corpus, i.e. newspaper text, and others in coding the Arabic script.

This section intends to give a short account of the steps that we followed in encoding the data input. We shall consider four stages: automatic input, manual input, resolution of the problem of the word, and finally editing and proofreading.

5.6.2 Automatic Input

5.6.2.1 Use of OCR

As described in 5.5 above, the optical scanning reader used to input the English corpus is the KDEM. This is a powerful piece of equipment, unique in that it is an "intelligent" scanner. It can read printed material in a variety of typefaces and can be trained to distinguish the shapes of characters in a text, not only those of the Roman script, but a variety of other scripts, e.g. Cyrillic, Greek or Hebrew.

The machine itself is composed of five basic hardware components: a scanner, CPU, disk drive, terminal and magnetic tape drive. Its operation comprises three distinct stages:

1) Training: In this stage the machine learns to distinguish characters in terms of shape, size, spacing, degree of blackness, etc. As each character is identified, the operator verifies it, deleting those that are atypical, badly formed or wrongly identified and saving the typical and correctly recognised ones. The KDEM stores what it has learnt on disk on a "training set", which can then be employed for scanning future texts of the same typeface. A different typeface will require retraining.

2) Data entry: In this stage, the scanner starts reading the texts. The reading head of the scanner, called the camera, moves along the lines of the text and displays what has been scanned on the screen of the terminal. The operator can then intervene to verify, alter, insert or delete, as necessary. The scanner often uses guess-work to identify a string, a procedure which, depending on other factors, may lead to ample subsequent operator's intervention and editing. To improve the operation and minimise guess-work, KDEM is provided with a built-in dictionary that it regularly consults to verify the components of strings of

characters. If a string is not found in the dictionary, and KDEM cannot guess it, the operator is then queried and asked to intervene.

3. Storing output: the edited texts is then stored in a large buffer on the KDEM disk ready for output. In our project, the texts were output onto the Harris system and stored there to enable subsequent steps of proofreading and editing.

5.6.2.2 Limitation of KDEM

KDEM's reading capability is, in general, very good indeed. However, the machine has some limitations. The points we shall mention below are based on our own experience of using the machine for this project. (For a more detailed assessment of KDEM's capabilities and limitations see Hockey 1986a).

1. Typescripts: KDEM is extremely sensitive to the quality of typescripts: its font, size and style. It can scan fast and efficiently the typescript it has been trained to read. It is therefore ideal for scanning printed books, where typescript is fairly unified, or similar printed material, such as texts printed by an electric typewriter using a golf ball. A newspaper text often has a multiplicity of typescripts: an introductory paragraph may be printed in a typescript different in pitch size or font from the rest of the text, or occasionally, a final part of a text that appears on a different page (as a continuation) may be in a different typescript. Furthermore, a text differs from one another, across the various newspapers or across different issues (published at different dates) of the same paper. Consequently, scanning

requires constant training and retraining and continuous operator's intervention. This, in effect, slows down the operation to a considerable extent and renders it cumbersome and inefficient.

2. Type of Text: KDEM is most efficient when reading modern printed books in which the paper is smooth and free of accidental ink spots. Unfortunately newspaper texts are printed on fuzzy paper where ink has run slightly or where some words or even whole lines are not imprinted perfectly: either too dark or too light. Furthermore, the newspaper text is often smudgy and this creates extra characters that are not part of the text (KDEM reads every ink spot as a character and tries to match it with the nearest character shape in its store). Additionally, the text background is often rather grey, which adds another problematic dimension to scanning. All these factors generate constant intervention for correcting or disambiguating. As a result, scanning is often slow and inefficient, to the extent that, in our project, some texts had to be replaced.

3. Text layout: Newspaper texts are printed in columns that may not be equal in length, often interspersed with spaces, usually in a frame and distributed unevenly, for main views in the text, comments, illustrations, graphs, or simply the author's name and title. Although KDEM supports an electronic tablet which is useful when only part of a page is to be scanned, experience has shown that, given the odd ways in which text columns are organised, setting up a text on the tablet is fiddly and time-consuming. A better procedure is to perform a physical editing to the text layout whereby each text is cut into columns, each column is then pasted to

a sheet of paper in a convenient, easy to recognise shape, and all material that is not a genuine part of the text is removed. This, while facilitating the operation of scanning, creates ample preparatory work.

4. Nature of the operation: In identifying characters, the KDEM treats as a character each shape that is surrounded by white space. The machine is, therefore, suitable for all scripts that use separate characters (not joined to the preceding or following ones), but it is not compatible for scanning Arabic script, whose characters usually join to form units that may represent either whole or part of orthographic words. Accordingly, a word such as كَتَبَ (kataba) will be identified as made up of three blocks or shapes: كٮٮ and the two dots of ت (= ta) and the dot under ب (= ba). Since this creates an infinite number of shapes, depending on the number of combinations in each word in the corpus, and since the character set that can be stored and applied is limited, the machine is not capable of coding Arabic script.

5.6.3 Manual Input

Our search for compatible software or alternative OCR capable of automatic scanning of Arabic script has been negative.[13] Decisions were then taken to convert the Arabic script by using a transliteration scheme, which made it mandatory to code the entire corpus manually.

The transliteration scheme has been made as informative as possible. In this respect three major decisions are taken:

1. Transliteration is to reflect the graphemic rather than the phonemic values of the Arabic script. This has required adopting a scheme that is peculiar to this purpose.

2. The corpus is to be fully vocalised, i.e. all short vowels, "tanwin", gemination, and other markers of case, mood, tense, and voice are to be inserted.

3. The entity of the word in Arabic is to be resolved. This is achieved by putting forward a linguistically competent analysis for the concept of word as a computational unit.

Decisions (1) and (2) are discussed in detail in Appendix (4). Decision (3) is discussed in Appendix (1).

5.7 Editing and Proofreading

The next stage in the processing of the corpora is editing and proofreading. This is composed of four steps in which different editing, verification and correction are performed:

1. First Step: This concerns mainly the English corpus and includes the following operations:

a) Serialisation: Texts are given a serial number according to their order of arrangement in the corpus.

b) Titles and date of publications are added to each text

c) Paragraph indentation is made consistent.

d) Soft hyphens are removed and words divided by a soft hyphen at the end of lines are reconstituted. This operation is first performed automatically by a program, followed later by manual disambiguation (in case the removed hyphen happens to be a genuine one, where it is re-inserted).

2. Second Step: This involves correction of misspellings, omissions, syntactic errors of case or mood (in Arabic) and other types of errors. This was done in three operations:

a. A spelling program is run on successive short portions of the English corpus and a customised dictionary is gradually built up for all lexical items that appear in the corpus but not in the original computerised dictionary.

b. Alphabetic word lists are produced for the Arabic corpus, which are checked for errors and ambiguities. Erratic words are corrected and ambiguous expressions are subjected to operation c.

c. To resolve ambiguities (In English and Arabic), concordances are produced with sufficient contexts. Corrections are then applied to the corpora.

d. A further check is undertaken by reading a hard copy of large selected portions of the coded text with the original. Omissions, extraneous inclusions and errors of coding are noted and corrected on the corpora.

Most corrections are introduced on the Harris system, using its editor, and the Harris Word Processor (Muse) in conjunction to some utility programs such as "Find" and "Compare". Other corrections were introduced on the VAX Cluster system.

3. Third Step: Since newspaper texts are coded in lines of

various lengths (representing newspaper columns), it is felt necessary to have a consistent line length to make the corpus layout more presentable. Programs are, therefore, written (in SPITBOL, see 5.9 below) to block up text. The constant line length specified is 45 characters. If the 45th column in a line happens to be within a word, that word is allowed to be completed before a new line is initiated.

4. Fourth Step: This is a final check-up made against new alphabetic word lists of both corpora and verified via the use of concordances. When all corrections are made, the corpora are ready for processing. There are two stages involved in processing. The first involves production of general word lists and concordances and outputting them on microfiche. The second involves automation and processing of connectives. These two stages will be discussed each at a time next.

5.8 Text Processing: Word List and Concordance Generation

5.8.1 Uses of Word Lists and Concordances

The first stage in processing the corpus is the production of global word lists and concordances. This is motivated by a number of factors, most of them are familiar to concordance compilers, but are here treated as peculiar to the project.

1. Word list and concordances give a global profile of the lexical distribution of a text corpus which can be manipulated for making a statistical statement on the corpus.

2. The lists can be used as a basis for a frequency dictionary.

In English there is a proliferation of lists (Brown, LOB, Birmingham lists, to mention a few examples). In Arabic, however, lists are sparse. One early and comprehensive list that was prepared manually is Landau's frequency word count of Arabic written prose (1954; see Appendix 1 for a critique). Another more recent list has been prepared by a team in the German Democratic Republic from a corpus of 79,561 running words (see Fromm 1982).

3. Word lists and particularly concordances can supply a neat as well as practical lay-out for observational work of the key items. They are manipulated in this section for inspecting and verifying whether particular items have the cohesive role of a connective in the text (see details of this operation later).

4. The indexing information and the statistical frequency information that is furnished for each key item provide easy access within the text corpus itself.

5. Initial word lists and concordances, as discussed earlier, are used as techniques for observing and identifying errors of coding in both corpora.

5.8.2 Some Decisions

Word list and concordance compilers are aware of a number of decisions that have to be taken; some are predominantly linguistic while others are related to layout. When implemented, they make up general specifications of the final output. We shall discuss below those related to this project.

A. Linguistic

1. The decision to exclude case and mood in Arabic from determining membership of V (see Appendix 1 for details) requires provision of means by which words having different case and mood marking, however complex, are to be grouped together. For instance, words such as "mu'allifūna" [authors] (in nominative case), "mu'allifīna" (in accusative and genitive case) and "mu'allifū" and "mu'allifī" their respective variants when entering in "iḍāfa" construction, are to be grouped under one key word. Verbs such as "yarmī" [throw] (indicative), "yarmiya" (subjunctive) and "yarmi" (jussive) are also grouped together. Similarly, words, mainly perfect verbs and some prepositions, that undergo vocalic change because they are in proximity to some personal pronouns are classified under one form. For instance, the verb "qāla" [said], "qul" (in qul-tu, qul-nā) and qāl (in qal-ā, qal-ū) are grouped under one form.

2. No serious attempt at lemmatisation of entries have been made, apart from the groupings suggested in (1) above (which can be considered as elementary lemmatisation). Procedures for lemmatisation can be planned as part of future research.

3. A hyphen is treated as a character (a diacritic) and thus can affect sorting of key words. However, hyphenated words in the English corpus pose a problem of consistency. Words such as counter-weight, break-through, half-way, door-step, house-keeping, inner-cities, are used as hyphenated in some texts and without hyphen in others. To unify the form of these words, three dictionaries have been consulted "Collins English Dictionary",

"Chamber's 20th Century Dictionary", and "Longman Dictionary of Contemporary English". In case of disagreement, personal preference is used.

4. The apostrophe in English is treated as a character and therefore members and members' have two different entries. Similarly cannot and can't are two different words.

5. A serious problem in word list and concordance generation by computer is the distinction of homographs. A successful achievement of this task requires extensive human intervention based on careful appraisal of a number of linguistic factors: lexical and contextual. This, first, demands setting criteria for meaning distinction. Then careful inspection has to be made of every word to determine whether homographs exist. Later, techniques have to be adopted to help the computer resolve the problem (probably via tagging procedures). All this is time-consuming and is unlikely to be error-free. Accordingly, we have decided to ignore the problem of homographs in generating the global word list and concordances. However, in compiling concordances of the connectives, the problem of homographs is more acute and demands specific measures to resolve it.

6. The word lists and concordances include every word in the corpora. No exclusion has been made for words of high frequency (such as "the", "a", and, "in"). However, there are two types of exclusions. Subtitles, a usual feature in most newspaper texts, have been ignored though they exist in the input text. The second exclusion concerns a very small number of vernacular words used in the Arabic corpus, such as "ḥīṭa" (from "ḥā'iṭ" [wall]).

7. Since the Arabic transliteration scheme is graphemically oriented, as mentioned in 5.6 above and discussed in Appendix (4), variant shapes of the same letter have been given different characters. For instance, the hamza is represented by "Q,O,G,J", and the "alif", as a vowel, is represented by "A", "V", and "′". Variant forms of the same letters need to be equated so that words that are identical except in these forms may be sorted together. In English, a similar procedure is needed for equating upper and lower case characters.

8. Proper nouns and figures are not masked or excluded and are therefore sorted in the concordances. An extension of this project intends to locate these and tag them so that they can be ignored. The decimal point (.) creates a problem. As the computer treats a dot as a full-stop and therefore as a separator, a figure having the decimal point will be treated as two words. This problem is avoided in the word lists and concordances.

B. Layout

1. The word lists will be arranged in columns, each column will have the entry followed by a figure denoting its frequency. The concordance is in KWIC format.

2. In the concordances a context of 50 characters on each side of the key word is provided. This context, with the necessary references (see 3 next) and the keyword will occupy a 132-character line.

3. Reference information will be restricted to the text serial number and line number of the corpus.

4. There are two types of word lists according to sorting arrangement of key words: alphabetic and by descending frequency. In the concordances, key words are sorted alphabetically.

5. Word lists and concordances are produced on microfiche and these form the microfiche appendices to this thesis.

Having made these decisions on the content and format of word lists and concordances, we now consider briefly some available tools used for achieving various types of text analysis tasks, including generating word counts and concordances.

5.9 Software Tools for Linguistic Analysis

Researchers involved in computer-aided linguistic investigations have at their disposal a number of software tools that assist at the various stages of the analyses. We shall briefly mention some that are used in corpus-based research.

5.9.1 Programming Languages

Several researchers have opted for widely-available general purpose procedural languages such as FORTRAN, PASCAL and ALGOL 68. However, these languages, in general, have poor facilities for representing and manipulating strings, lists and trees, and, moreover, require considerable skill and experience for achieving some standard tasks in linguistic analysis. Two more recent languages ADA and Small Talk have been developed with syntax very much like natural languages (Huntsman 1982). ADA is an ALGOL-family language (related to PASCAL and C); Small Talk is an object-oriented language (as distinct from procedure-oriented languages like PASCAL and FORTRAN) that attempts to replicate human interaction. Unfortunately, as ADA is very rich and complex, ADA compilers are slow (Atwell 1985). In addition, both languages have poor string-, list- and tree-handling capabilities.

Two more efficient language-oriented languages are the functional language LISP and the logic language PROLOG. These have widely been used in research in artificial intelligence and computational linguistics. They have the advantage of straightforward string- and list- processing facilities, and also the ability to evaluate a string or list as a piece of program code.

However, the two languages that are specifically designed for string handling are SNOBOL4 and ICON. SNOBOL4 has been recognised as an attractive language for text analysis. It has several features that contribute to its efficiency (see Gimpel 1967, Griswold et al. 1973, Griswold and Griswold 1973, 1986, Newsted 1975, Maurer 1976, Burnard 1978, 1979, Tucker 1979, Griswold 1982, Day 1984, Butler 1985b, Hockey 1985):

1. SNOBOL4 treats strings as individual objects instead of requiring the programmer to deal with arrays of characters.

2. Its pattern-matching facility permits a great number of string analysis and transformation operations to be formulated at a high level of abstraction.

3. Management of storage is entirely automatic, hence the programmer is not required to specify storage requirements when the program is written.

4. A number of features such as associative tables are very useful for specific tasks of linguistic processing.

SNOBOL4, however, has some weaknesses (Griswold 1982). These are related to its limited control structures, lack of convenient facilities for numerical operations, lack of a repertoire of low-level string processing functions, and its large and slow implementation on most computers. However, the durability of SNOBOL4 in the face of such deficiencies is partly due to its attractive features, but more specifically "because other programming languages do not offer better alternatives" (Griswold 1982 p.8).

ICON has been introduced as an alternative to SNOBOL4, with the aim of incorporating the better features of SNOBOL4 "while correcting some of its deficiencies and adding the results of knowledge gained over the past decade" (Griswold 1982 p.8). Syntactically, Icon is different from SNOBOL4. The latter is a statement-oriented language with fields for a label, subject operation and goto. Icon, in comparison, is an expression-oriented language with many conventional control structures that may be nested to reflect the logical structure of algorithms. In his evaluation of Icon, Griswold (1982) states that despite its implementation on a number of different computers, it is still not well known or widely available. Furthermore, considerable effort is needed to master Icon unless its user has prior experience with other programming languages. Comparisons of programs of substantial size show that SNOBOL is faster for some kinds of processing while Icon is faster for others. Griswold concludes that Icon must be considered speculative at this time. "While it has had considerable use and is currently being used for production work, it is not yet - and may never be - a replacement for SNOBOL4" (ibid p.16).

In this project, nearly all programs are written in SPITBOL, a SNOBOL compiler (developed by Dewar and Belcher at Illinois Institute of Technology) that contains a number of features that are not in standard SNOBOL4 (see Maurer 1976 for a short discussion). A suite of 50 programs has been written, each in two versions to handle the English and Arabic corpora. These programs accomplish a variety of text processing tasks ranging from simple reformatting of text to tagging complex expressions. Some programs have been designed to produce numerical profiles of two types: global (representing the whole corpora) or specific (e.g. representing connectives). However, for generating concordances and main word lists we have found more economical to use general-purpose packages. This is discussed next.

5.9.2 Packages

The production of word lists, indexes and concordances is one of the earliest and most obvious applications of computers in linguistic and literary research. The need for an efficient computer routine for this application has led to the development of a number of machine-independent packages in an effort to lift some of the programming burden. In the design of these routines, consideration has been given to the choice of a common programming language, the various types of input requirements and the availability of optional features within the routines themselves. As a result, most packages are capable of being installed on different computers and in different institutions.

One excellent early example of a well-conceived, general purpose program is COCOA (Berry-Rogghe and Crawford 1973), which was developed at the Atlas Laboratory in Britain and completed in 1973. It is written in Standard FORTRAN and can produce concordances arranged in KWIC format, word counts and word frequency information for material in any language as long as the characters can be encoded on the implementing computer.

Another package is JEUDEMO, which was developed at the University of Montreal. It aims at providing a simple but at the same time highly flexible tool for generating concordances, indexes, and vocabulary counts, and for performing a variety of statistical calculations (cf. Bratley et al. 1974). It is directly descended form COCOA with some additional features borrowed from various information retrieval systems.

A program that is written in PL/1 is CLAS, reported in Borden and Watts (1971). This concordance generator has been used with some success in a number of literary studies.

Another package that has to be mentioned in this brief review is CONSTANT. This is a text-processing computer program that had its origin in 1968 in NASA space programme (cf. Ule 1975) and has been developed as a general-purpose literary research tool at the Universities of California, Los Angeles, and Southern California. The package is capable of producing KWOC-type index and provides many textual discriminants, such as word-length distribution, and a variety of calculations (e.g. number and length of sentences and paragraphs, Yule's mean, standard deviation, covariance and characteristic K).

Two other highly useful packages are EYEBALL and OXEYE. These are designed to take an unedited English text and analyses its sentences linguistically into parts of speech and syntactic functions (via parsing routines) as an aid to stylistic analysis. They are, therefore, not concordance generators, but they are capable of producing word indexes and statistical calculations. EYEBALL is a FORTRAN program written by Ross and Rasche (see Ross and Rasche 1972 and Ross 1981 for details). OXEYE is a version of EYEBALL that is entirely re-written in SPITBOL at Oxford University. It retains many features of the original, but has been enhanced in flexibility and has been provided with some additional facilities.

We conclude this review by mentioning one of the most recent and powerful concordance generators: CLOC. The package was developed at Birmingham University by Alan Reed (see Reed 1977, 1978, and Reed and Schonfelder 1978 for details). It was originally written in a subset of ALGOL68, but a FORTRAN version has been released recently. The program comprises facilities for the production of vocabulary lists, the printing of concordances and the automatic production of collocations. The last function is lacking in most of the other programs,[14] and has been the main motivation behind the design of CLOC.[15] Indeed, the acronym itself is derived from the word "collocation".

Each of these packages has its own merits. We have, however, opted to use OCP (the Oxford Concordance Program) for the purposes of this project. The features and our manipulation of this package will be discussed next.

5.10 Using OCP: Problems, Procedures and Evaluation

5.10.1 Preliminaries

OCP is a machine-independent general purpose computer program which produces word lists and concordances in a variety of languages and alphabets. It is a powerful package of facilities that can be used for various text analysis applications including the investigation of style, vocabulary distribution, grammatical forms, rhyme schemes, and text editing. The program was developed at Oxford University during 1979-1980 (Hockey and Marriott 1979, 1980a, and 1980b).[16] Its design makes use of ideas borrowed from many sources, particularly from COCOA and CLOC.

5.10.2 Accessing OCP:

In order to access the package, the user must provide two sets of information: the text to be analysed (the corpus) and a set of commands describing the analysis to be performed. These commands are grouped into four sections depending on their function. The sections must be arranged in this order: "Input", "Words", "Action", and "Format". The final instruction is "Go", which is used to indicate that the end of the commands has been reached. Refer to Appendices (5A, 5B, and 6) for a detailed discussion of the commands that we have used and their main requirements.

5.10.3 OCP Output

A full KWIC concordance is produced for each corpus, where key words are sorted alphabetically. In addition, two general word lists are produced for each corpus, one alphabetically sorted and the other sorted according to descending frequency. Later OCP runs requested concordances and word lists of connectives (partially or fully tagged, see 5.12 below). Final output is made on microfiche (see microfiche appendices).

5.10.4 OCP Performance: Some Comments

The comments that are made here reflect the user's, and not the programmer's, evaluation of OCP performance. They are, therefore, rather general in nature and do not include any technical specification.

It has been mentioned (5.10.1 above) that OCP is a powerful concordance generator. It has three essential features that contribute to its success:

a. It is machine independent and is therefore portable.

b. It is general-purpose and can therefore serve a variety of tasks ranging from producing general word lists to specific tasks of text editing of linguistic/literary analysis. Additionally, it can serve all languages provided the input text is Romanised.

c. OCP is easy to operate and its documentation (particularly the Users' Manual) is comprehensive, easy to understand and full of exemplifications.

But despite these strengths, there are some weaknesses that, we hope, may be removed from future versions. We shall mention only those problematic features that we have encountered while using this package throughout the various stages of the research. They are grouped under three classes: layout, processing and requirements.

1. Layout: We have encountered two problems in connection with the concordance layout:

a. Reference information of key items has posed a special problem in producing a general concordance. The command "References COCOA", while indicating reference information, ignore the number of records/lines on which these references are placed and start or resume registering line number count from the first line of the text. This is useful in cases where the input texts are plays or poems. The references in this case are restricted to act/scene or title/author specifications and actual text lines start after that. However, in our corpus this facility has not been very helpful. Ignoring reference line numbers (i.e. treating text references as if they did not exist) has caused OCP to give wrong line number locations of key words, though text serial numbers are accurate. If "Reference" command is not given in the Input section, all words in the references will be treated as part of the text and so the word lists are unnecessarily lengthened and confused. If text references are specified in the "select except between" command, or declared as "text comments", line numbers of key words will be correct, but no other text references are obtainable (e.g. source or serial number of text). We suggest that the "Reference" command should give the researcher the choice either to include or ignore numbering the lines on which the text references are placed.

b. If the length of the textual context requested for key words is more than can be fitted on a printer line (e.g. requesting 100 letter on each side of the key word), the key word as used in the context will lose its highlight feature. One has to read the whole context to locate where the word is. A facility is therefore required whereby words are highlighted via, for example, underscoring, bracketing, bold type, etc.

2. Processing: This type groups problems related to the "Action" section of OCP.

a. The "Pick suffixes" command, as will be discussed in more detail in Appendix (6) (see also 5.11 below), detaches not only suffixes but any final letters that are not declared as suffixes.

b. The dummy symbols that are used for producing skeletal forms or patterns of words need to be increased. At present only two dummy symbols exist: "*" to stand for any number of letters including none, and "@" to stand for any one letter. Some further dummy symbols that are useful particularly for analysis of Arabic include:

i. a dummy symbol to stand for a particular letter that can be specified by the user.

ii. a dummy symbol to stand for a user-specified set of characters in a fixed order.

iii. a dummy symbol to stand for a user-specified set of characters in a variable order.

c. Options are required for producing more comprehensive statistics of the input text including, for instance, number of sentences and paragraphs, distribution curves for the type-token, standard deviation, covariance, distribution curves of word, sentence and paragraph lengths, and distribution of the letters of the alphabet.

3. Processing Requirement: Depending on the length of the input text, OCP work files can grow to a substantial size, which requires a large space on the disks. On a number of occasions, OCP jobs crashed because of insufficient disk space quota. In addition, OCP is very slow in processing long texts, and thus require a massive CPU time allocation. Unless one has at his disposal unlimited CPU time and unlimited disk quota and provided that there is sufficient space left on the system disks, one has to fragment the OCP job into smaller jobs and run them consecutively.

These comments are here intended as suggestions for improvement and should by no means diminish the strengths of OCP. A new version is under preparation, which will make use of the numerous suggestions that users have made (Hockey and Martin 1987). Another version is being prepared for use on PCs (Hockey 1986b)

5.11 Tagging Procedures

5.11.1 Preliminaries

Tagging has been used extensively as a means for rendering a computerised corpus of text more useful for linguistic analysis. The operation consists of placing one or more of predetermined tags so as to assist computer search and information retrieval. Tags that are used for categorising grammatical categories in a text corpus promote its investigative value and assist the linguist more readily in observing and studying the grammatical repertoire of language in use. For instance, grammatical tags may give the linguist easy access to all "noun + noun" constructions, all sequences of verb + subject, all adverbs or conjunctions of a particular type, etc. Similarly, tags representing prosodic features of a corpus of spoken texts are highly useful in computerised phonetic/phonological descriptions.

5.11.2 Taxonomy of Tagging Systems

There has been an increasing number of projects that require tagging procedures. Linguists, sometimes seemingly unaware of each other's work, have used a variety of systems to tag their corpora, and, in some cases, have applied different methods of analysis to the same material. It is, of course, impossible to survey the various types of systems or projects used. Such a task goes beyond the scope of this work. What is feasible and convenient, however, is to make a brief and general taxonomy with some exemplification.

Generally speaking, tagging systems differ in relation to three basic factors: aim, method and typology of tag. These are outlined below:

1. Aim: As far as aim is concerned there are two types of systems: general (or neutral), and specific.

a. General: These are intended for general use and are concerned with developing a source of data as neutral as possible with respect to future applications. To this type belong the tagging systems used in the Brown, LOB and Dutch corpora; the systems are based on fairly uncontroversial, traditional, "surface" analysis.(17) Also to this type belongs the tagging system of the London-Lund corpus (the survey of English Usage) where syntactic as well as prosodic tags are used.(18)

b. Specific: This type includes systems that are designed for specific linguistic projects. The nature of the project often dictates the extent of their specificity. Ellegård, for instance, adopts a very detailed system applied to a portion of the Brown corpus with the aim of investigating the syntactic features of four categories of English texts (for the results of this investigation see Ellegård 1978). In contrast stands the Copenhagen project, described in Faerch (1979), where tagging is restricted to the assignment of word-class labels. The project aims to produce a grammatical analysis of a corpus of learners' language collected in Denmark and consisting of English as spoken and written by Danes. Another project that is different in aim is the "Zagreb version" of the Brown corpus (described in a paper by Rudolf Filipović and reported in ICAME NEWS no.3 1979). Half the Brown corpus is selected and translated into Serbo-Croatian with the object of providing a source of data for contrastive analysis. The tagging system is partly similar to Ellegård's and partly to the Dutch project.[19]

2. Method: According to the method of assigning tags, tagging systems can be divided into three: manual, automatic and semi-automatic.

a. Manual: where assigning information is performed by manual methods. The process is normally tedious and time-consuming but does allow the linguist to use an elaborate system of tags to represent various linguistic information. Manual methods are resorted to when the nature of the tagging system makes it cumbersome or difficult to perform by automatic means. Examples of such systems is Ellegård's, and the systems employed in the Zagreb project and in the Dutch CCPP (Computer Corpus Pilot Project).

b. Automatic: This involves syntactic analysis that is performed with no (or minimum) manual intervention and relies on using some parsing techniques. One example is the system developed by Geens (reported in ICAME News 3 1979; see also Geens 1984). The system has been implemented on the Leuven Drama Corpus and includes a "syntactic recognition procedure" that is used to produce a rough characterisation of the apparent syntactic structure of each sentence, and a "syntactic analysis procedure". Within this category of system we may include Sager's computer grammar of English which is discussed in her (1981) volume. Strictly speaking, this is more a natural language parser than a tagging system and it has its own lexicon and string computable grammar.

c. Semi-automatic: Most systems are designed to use automatic procedures and then, for disambiguation or performing additional tasks, resort to human intervention. For instance, the Brown corpus uses the routines developed by Greene and Rubin (1971) to assign the various tags and then has to rely on the linguists for manual resolution of all kind of ambiguities (see Francis 1979b). The LOB corpus is tagged by a suite of programs collectively known as CLAWS (Constituent-Likelihood Automatic Word-tagging System). The human intervention is minimal and is restricted to the first stage of the procedure (Blackwell 1985, cf. Elliott 1985). CLAWS, using probabilistic methods, is claimed to be fairly successful, achieving a ca 96% correct result (Leech 1985). A third example is the tagging system of the London-Lund corpus of spoken English. The orthographic transcription and the prosodic marking are achieved manually. The subsequent syntactic tagging was performed in a semi-automatic way, whereby tone-units are taken as the basis of grammatical analysis, word-class tags are chosen from a general-purpose dictionary, phrase structure rules are applied extensively and an interactive mode of analysis is adopted (Quirk and Svartvik 1979, Svartvik 1980).

3. Typology of tags: Tagging systems have adopted two types of tags:

a. A set of letters and some non-alphabet symbols: These are attached to the front or back position, or both, of each word. Most systems have adopted this type of tags.(20)

b. A set of digits: The Dutch CCPP (Computer Corpus Pilot Project) has been syntactically analysed by use of a four digit code that was assigned to each word (Aarts and Heuvel 1980, see also Johansson's comments in his 1979 paper).(21) The first two digits contain information about word classes, while the second pair of digits marks constituent boundaries. This type of tags is adopted with some modification in the Zagreb project.

5.11.3 The Tagging System in this Project

5.11.3.1 Features

The main features of the tagging system of this project will be discussed, as we have done above in the taxonomy of systems, from three perspectives: aim, method and types of tags.

1. Aim: There are two distinct aims for employing a tagging system in this project:

a. To mark inflections in the Arabic corpus in such a way that OCP can ignore all inflectional variations related to verb moods, noun cases and word adjacency. The ultimate aim is to produce word lists where the concept of word complies to the theorisation we have offered in Appendix (1).

b. To mark connectives, the object of this study, for various levels of detail (see 5.12 below). This will subsequently help us perform the proposed textual analysis in both corpora.

These aims are specific in nature and hence demands a tagging system that is customised for the purposes of this work. However, we can envisage some general application whereby the system of tags in the Arabic corpus can assist in studying some aspects of the morphological system of the Arabic language, or, at least, represents a first step towards devising and implementing a more general system of morphological or syntactic tagging. It can also serve as a preliminary stage in a project for lemmatisation of Arabic.

2. Method: To meet the two aims that are specified above, we require two types of tagging operations. The first one is applied to the Arabic corpus only and involves using tags for morphological identification, mainly for the purpose of generating concordances of Arabic via OCP. The other operation is applied to both corpora and involves recognition and analysis of textual connectives.

At the outset of the project, our aim was to make the tagging procedure as automatic as possible. However, given that no fully automatic system exists for tagging English, let alone Arabic, connectives, and since, to the best of my knowledge, no automatic systems are currently available for performing a general syntactic analysis of Arabic, we propose to implement an interactive semi-manual mode of analysis. This methodological decision is justified on the basis of the following two factors:

a. This method proffers an insight into the textual properties of connectives in English and Arabic not only as a result of, but also in the course of the analysis. In addition, an understanding is gained into the means of handling morphological patterns in a computerised analysis of Arabic.

b. We can gain in accuracy and flexibility by reducing the amount of effort involved in subsequent verification, correction and disambiguation.

3. Types of Tags: The tags used in the first tagging operation are kept to the minimum despite the intricacy of the Arabic morphological system. It consists of a small number of non-alphabetic characters of the ASCII set. These are: " [ ] { } >". The second tagging operation (connective tagging) demands a set of tags that represent: a) functionality, b) range of operationality. (These will be discussed in more detail in 5.12). No digital tags are used for either operation.

5.11.3.2 Tag Assignment

In this section we discuss tag assignment as used in the first tagging operation. This involves five types of procedures that collectively tag morphological patterns in the Arabic corpus. The tags are assigned to inflectional variations that are caused by changes in case, mood and juxtaposition (details are in Appendix 1). The procedures consist of: morphological identification, automatic and manual tag assignment, proofreading and disambiguation, and treatment of some special cases. These procedures are outlined below.

1. Morphological Identification

The first procedure is to identify the type of inflectional variations that accompanies changes in case (nominative to accusative or genitive), mood (indicative, subjunctive or jussive) or juxtaposition (adjacency) of certain words. These inflectional variations are identified in two steps:

a. A comprehensive list of all possible inflectional patterns is assembled from Arabic reference grammars. This list is used for checking the inflections in the corpus.

b. OCP is requested to produce a reverse word list that are sorted alphabetically.

c. Words are examined for the endings. All morphological endings are put on a second list and checked against the first. As a result of this check a third list is produced that classifies the inflectional patterns of the second list. The classification is made on the basis of the three factors mentioned above: case, mood and juxtaposition.

d. The next step is to identify and categorise the internal morpho-phonemic changes that defective verbs undergo when they are in the subjunctive or jussive moods or when they are juxtaposed (orthographically connected) to personal pronouns. This is achieved as follows:

i. OCP is requested to produce a word list of defective verbs when in the subjunctive or jussive. The skeletal patterns that OCP provides in the PICK command (i.e. the asterisk to represent any word or part of a word, and the symbol @ to represent any one character declared in the alphabet) are useful for this purpose. The morpho-phonemic changes are noted and classified in a separate (fourth) list.

ii. OCP is requested to produce "phrases", each consisting of a defective verb and an orthographically connected personal pronoun (such as tu, ti, ta, nā, nī, etc.). The types of morpho-phonemic changes are noted, classified and added to the previous (fourth) list.

iii. The third and fourth lists of morphological patterns detected in the corpus are then unified to produce a final list that is used for the next set of procedures.

2. Tag Specification

In this step, tags are associated with the various identified classes of inflectional patterns. These are discussed in three main categories: case, mood and juxtaposition.

a. Case

Changes in morphological patterns related to case are briefly discussed in Appendix (1). Here we discuss the way the tags are assigned in the operation. The general procedure is to equate the various case suffixes so that they do not affect the sorting operation during the execution of OCP. Equating case suffixes is performed in these steps:

i. Case markers, i.e. damma, fatha, kasrah and their tanwin equivalent (transliterated in the computerised scheme as "u, a, i, W, N, @, M" respectively), are declared as suffixes. This is performed by using the "Pick suffixes" command in the Action section of OCP. This command causes these suffixes to be detached from words before they are sorted, but retained in the context. However, we have discovered during implementing this step that OCP has a weakness, probably due to unobserved bug in the system. OCP will not only detach the declared suffixes, but any final letter from any word. For instance, while u and a are detached from the end of "qalamu" and "qalama" so that they are sorted together under the keyword "qalam", it also detaches the last letters of such words as min, fatā, kitāb (in kitab-i), producing such nonsense key words as mi, fat, kitā. This produces a messy and totally confused word lists.

To overcome this problem in the system, we have written a SPITBOL program that creates a dummy suffix for all words in the corpus that do not end with any of the declared case suffixes. This is accomplished by tagging the end of these words with the character "[", and then including this character in the declared list of suffixes. For instance, in the previous example the words will be min[, fatā[, kitāb[.

ii. The dual case markers "āni, ā (nominative), ayni, ay (accusative, genitive)" are tagged with the symbol " ` " as follows: (the exemplified word is the dual of "ra'is" [president]; note that the second column makes use of the computerised transliteration)

ra'isāni	raGIsA` ni
ra'isayni	raGIsay` ni
ra'isay	raGIsay`
ra'isā	(not tagged)

The tagged case markers are then equated in the alphabet command with "ā".

iii. The sound masculine plural markers "ūna, ū (nominative), īna, ī (accusative, genitive" are tagged with the symbol "]" as follows:

(the exemplified word is the plural of "mufakkir" [thinker])

mufakkirūna	mafak#irUn]a
mufakkirīna	mafuk#irIn]a
mufakkirī	mufak#irI]
mufakkirū	(not tagged)

The tagged case markers are then equated in the alphabet command with "ū".

iv. Defective nouns that end with "ī" have special case marking. These are tagged with the symbol "]" as follows:

(the exemplified word is "nadī" [club]; note that the equivalent tagged forms are in the computerised scheme symbols)

nominative and	no marking	nadī	no tag
genitive:	substitute in	nadin	nAdin]
accusative	substitute iya	nadiya	nAdiy]a
	substitute iyan	nadiyan	nAdiy]N

The tagged suffixes are equated in the alphabet command with "ī".

b. Mood

The inflectional changes that verb mood brings about vary according to verb class: whether it is strong or inform, and whether the infirm verb is defective or hollow (see Appendix 1 for details). The tags are therefore specified in accordance with the inflectional patterns dictated by these classes.

i. Strong verbs: The mood markers of the strong verb are the damma suffix and fatha suffix to mark the indicative and subjunctive moods respectively, and the 0-suffix (absence of marker) to mark the jussive. The OCP command "Pick suffixes", used for noun cases, will equate these forms by detaching the declared suffixes.

ii. Infirm verbs:

1) Defective verbs, i.e. verbs of which the third radical letter is a vowel. The jussive and subjunctive inflections are equated with the indicative as classified below (the first column in the classification refers to mood, the second gives a typical example and the third gives the type of tag).

a) defective verbs with the third radical ā:

indicative and subjunctive	yalqā	no tag
jussive	yalqa	yalqa`

The character set "a\` " is equated with "A" in the alphabet command.

b) defective verbs with the third radical u:

indicative	yabdu	no tag
subjunctive	yabduwa	yabduw]a
jussive	yabdu	yabdu]

The character sets "uw]" and "u]" is equated with "U" in the alphabet command.

c) defective verbs with the third radical "ī" (i.e. I in the computerised scheme):

indicative	nabnī	no tag
subjunctive	nabniya	nabniy]a
jussive	nabni	nabni]

The character sets "iy]" and "i]" are equated with "ī" in the alphabet command.

2) Hollow verbs, i.e. verbs of which the second radical letter is a vowel. The jussive form undergoes an internal morpho-phonemic change, whereby the medial long vowel is shortened. This short vowel is tagged with "]" and equated with its original vowel in the alphabet command. For example:

indicative	jussive	tagged form
yaqūlu	yaqul	yaqu]l
yamīlu	yamil	yami]l
yanāmu	yanam	yana]m

Thus, in the alphabet command, the character sets "u]", "i]", and "a]" are equated with "ū", "ī" and "ā" respectively.

c. Word Juxtaposition

The influence of word juxtaposition is discussed in Appendix (1). It simply refers to the variety of vocalic changes that certain words, particularly infirm verbs, undergo when they are juxtaposed on some (orthographically connected) personal pronouns. Tags are specified and used in accordance with the type of vocalic change involved and can be classified as follows:

i. Defective verbs:

1) Perfect verbs whose last letter is "ā" (e.g. 'iltaqā) undergo the following changes. The vowel "ā" becomes "ay" ('iltaqay) when in juxtaposition with all connected subject pronouns (-tu, -ta, nā, etc.) except u. In the latter case it is shortened to "a" ('iltaqa ū). These forms are tagged with " " (e.g. 'iltaqay`

and 'iltaqa\`). The character sets "ay" and "à" are then equated with the long vowel "A" and "V" in the alphabet command.

2) Imperfect verbs that end with ā or ī (e.g. yatarajja, yabtagi) lose these vowels when juxtaposed to "una" or "ina". The tags > and } are used respectively to represent these two missing vowels (e.g. yatarajj> ūna, yabtag} ūna). In the alphabet command, the tag ">" is equated with "ā" (i.e. "A" and "V" in the computerised scheme), while the tag "}" is equated with "ī".

ii. Hollow verbs:

Perfect hollow verbs (e.g. qāla, nāma, 'ixtāra) undergo inter-vocalic changes where the long vowel ā becomes "u", "i" or "a" according to a complex set of morphological rules. These vowels are tagged with "\`" (e.g. qu\`l-tu, ni\`m tu, 'ixta\`r-tu). The character sets "u\`", "i\`" and "a\`" are equated with "ā" (i.e. "A"in the computerised scheme) in the alphabet command.

Other types of vocalic changes related to juxtaposition are treated as special cases (see below).

3. Automatic tagging:

After the inflectional patterns are identified and the necessary tags specified, computer techniques involving programming are used to achieve an automatic tag assignment. The program relies on search and pattern matching of particular letter combinations (representing inflectional patterns) that are grouped in a special suffix dictionary assembled for this task (for dictionary lookup techniques see below). The operation has been performed with a high degree of accuracy.

4. Manual Tagging:

Certain tagging tasks, particularly those involving some hollow verbs are performed manually. This decision is justified on the grounds that some hollow verbs require specific inspection of types of roots and patterns before a tag is selected.

5. Disambiguation

The next step is to proofread the tagged words, correct any errors, and resolve ambiguities. This is a manual task and is performed first by creating word lists (via OCP) of all tagged expressions. Key words are then checked against the lists compiled in procedures 1 and 2 above. When all correction and disambiguation is completed, a final check is made on all words containing inflectional patterns that have not, for one reason or another, been tagged. A small list of such patterns is compiled and the entries are either manually tagged in the text, or left to be treated as special cases.

6. Treatment of Special cases

Certain inflectional patterns require some special consideration. These have been grouped and equated by using OCP special facility "Pick headword", which causes all specified word strings to be grouped under a given head word (see Appendix 5B). These patterns are of 7 groups.

a. Forms of some orthographically connected personal pronouns:

i. Dual pronouns "āni","ā".

ii. Masculine third person plural pronoun forms "ūna", "ū" (with or without a silent alif, symbolised as "L").

iii. Two first person pronoun forms, one with and the other without "protection n": nī, ī.

iv. Feminine second/third person singular: "ī", "īna".

b. Some orthographically connected personal pronouns that are different only in their vocalised form.

i. Dual third person pronouns "humā" and "himā".

ii. Masculine third person plural pronouns "hum" and"him".

iii. Feminine third person plural pronoun "hunna", "hinna".

c. The defective verb "laysa" which becomes "las" when in juxtaposition with connected subject pronouns (e.g. ta, tu, ti, nā, ū, tuma, tum).

d. Prepositions and adverbs ending with "'alif maqṣūra" such as "'ilā", "<alā" and "ladā", where the "alif maqṣūra" changes to "ay" when the word is juxtaposed onto an orthographically connected pronoun.

e. Case variations of the "five nouns". These end with "ū" in the nominative, which changes to "ā" in the accusative and "ī" in the genitive, e.g. ('abū, 'abā, 'abī; 'axū, 'axā, 'axī; ḏū, ḏā, ḏī).

f. Case variation of "kilā", "kiltā" [respectively masculine and feminine equivalent of "both"] (in the nominative) where the "ā" changes to "ay" in the accusative and genitive.

g. Variations in the form of the perfect infirm solid verb that are due to juxtaposition with orthographically connected subject pronouns. The infirm solid verb (see Appendix 1) is one where the second and third root radicals are the same elements. In the usual form of the verb these two letters are geminated. When in juxtaposition to subject pronouns, they are separated by a fatha "a", which annuls the gemination. The following are examples: 'aḥassa, 'aḥasas (-tu); wadda, wadad (-tu), 'aḥabba, 'aḥbab (-tu).

5.12 Automation and Processing of Connectives:

5.12.1 Preliminaries:

The culminative point at this stage of the work involves the automation and processing of connectives. This task is performed in four sequential phases:

1. Identification: Connectives are recognised in the two corpora and tagging procedures are used to mark them.
2. Patterning: This includes a set of procedures aimed to facilitate later study of connective patterns.
3. Categorisation: This involves using tagging procedures to classify connectives according to functionality and textual range of operation.
4. Quantification: This involves preparing programs for producing a calculus of connectives.

One essential requirement is the preparation of a suite of programs to achieve the various tasks involved in the four phases. Manual intervention is inevitable and is reserved to those tasks that for reasons of economy and efficiency are more appropriately performed that way. Those phases are discussed in more detail next.

5.12.2 Identification:

This phase comprises a set of procedures aimed at identifying and tagging connectives and based on using dictionary lookup strategies. These are outlined below.

5.12.2.1 Assembling the Dictionary:

The first procedure that we have used for the identification of connectives is to set up a dictionary, i.e. lists that comprise all expressions that are likely to function as connectives. The lists are compiled in the following way:

A. Dictionary of English Connectives

1. Lists are borrowed from authoritative grammars. These include lists of conjunctions, sentence modifiers, some adverbials and sentence connectors. Most of these grammars are reviewed, or at least mentioned, in Chapter 4 and also in 6.1.

2. Adapted are also lists compiled by various scholars in their studies on cohesion in general or connectives in particular.

3. We have adopted with some modification the comprehensive lists prepared by Gardner and his team and reported in Gardner (1977).

4. Lists are also borrowed from studies of the adverbials in English, particularly Greenbaum (1969) and Jacobson (1975, 1978).

5. Additional lists are constructed by consulting the available word lists and concordance and by using the concordance to observe the patterns of some potential items.

6. Special concordances are produced for the following types of

expressions·and all key items are carefully inspected to isolate those with a cohesive role:

i. adverbs ending with -ly.

ii. prepositional phrases, particularly those containing a referential item, e.g. at this stage, on the other hand.

iii. Expressions that are related in form or meaning to established connectives.

B. Dictionary of Arabic connectives

This has constituted a major problem. Lists of Arabic connectives are sparse (if not non-existent) and, when available, represent mainly one type of connectives: coordinators and some subordinators. The following steps are therefore taken to assemble the dictionary.

1. Lists of conjunctions are borrowed from authoritative Arabic grammar references (written in Arabic) both Medieval and modern.

2. Lists are complemented by consulting Arabic grammar books written in English.

3. The main part of the lists is compiled at the first stage of coding the Arabic corpus on the computer. The manual nature of that operation has permitted the isolation of potential connectives.

4. Arabic word lists have been inspected and all items that may act, in isolation or in association with other items, as connectives are checked against the concordance. In the cases where the contextualisation provided by the concordance for each key word is not enough, special concordances have been obtained with sufficient contextual information (for instance in some cases OCP is instructed

to give more than 400 words of text to contextualise the key item).

5. More potential connectives are obtained via translating English connectives. These are then checked against special concordances and, where an item exists and manifests a cohesive function, it is added to the list.

5.12.2.2 Editing the Dictionary

The number of entries and the choice of words that make them up is important for an automated dictionary. The two dictionaries that have been assembled, using the procedures explained above, suffer from two drawbacks: a) the number of entries is unnecessarily big, and, b) a substantial number of entries can hardly be accepted as "real" connectives.

Accordingly an editing operation is deemed necessary to remove all entries that conflict with the definition and description of connectives as adopted in this study (see Chapter 4). This can only appropriately be achieved by examining the contexts in which the "suspect" entries occur in the text. This procedure requires production of special concordances for those entries and sufficient contextualisation is to be requested. Where the context given in the concordance is not enough, more information has to be derived from the input text.

The result of implementing this operation is that a number of entries have had to be removed, which rendered the dictionary shorter, better representative and therefore more efficient.

The entries that constitute each dictionary are of three types according to their structure: a) single-word connectives (henceforward labelled "simple connectives"), e.g. and, but, however; b) multi-word connectives (labelled henceforward "compound connectives"), e.g. in addition, as a result, on the other hand; c) correlative connectives, made up of more than one element separated by text, e.g. either ... or, as ... so, so .. that.

Since this variety in the structure of the entries may complicate the tagging procedures, we have resolved to divide the dictionary into two versions: Dictionary A contains simple connectives and correlates while Dictionary A consists of compound connectives. This means that two dictionary lookup and tagging operations need to be performed.

5.12.2.3 Dictionary Lookup

Systems for dictionary lookup (22) depend on two principal operations: the search strategy used to locate a dictionary entry, and matching and assigning the necessary information. To perform these two operations, a SPITBOL program is written, whereby the text and dictionary A are the input files. The program uses a simple "word in hand match" against the associative table. Search is accomplished by storing the dictionary in the buffer and then scanning each word form in the text and compare it to the forms in the dictionary. When a match fails, the next word form in the text is picked. Provision is made to ignore lines beginning with "((" and "//" which respectively indicate text references and subtitles. If a match is successful, the word is assigned the tag "^" before its first character so that the tag is an integral part of the word and represents its first character. When the last word form in the text is examined, the whole corpus including the tag is written on an output file.

It should be noted that only the first element of a correlative connective is tagged. The second element is left without any specific marking. This is a methodological decision and is made for reasons of convenience and ease of calculation.

5.12.2.4 Problems of Multi-word Connectives

Compound connectives raise a problem of two dimensions. The first is computational and concerns the search and match operations; the second is linguistic and concerns which element in the group receives the tag. To resolve the linguistic aspect of the problem we have to seek advice from lexicography. First we have to assume the existence of a control centre within the compound connective that acts as a core. This represents the entry in the lexicon where the compound connective is likely to be attached to. For instance, for the compound connective as a result it is more likely that it is classified in the lexicon under the entry "result" than under "as" or "a". Similarly, the connective on the one hand is more likely to be classified in a dictionary under the entry "hand" than any of the other elements. Accordingly, result and hand are considered control centres or cores within the above connectives and will, therefore, receive the tag.(23)

The computational aspect of this problem is resolved by writing a SPITBOL program that would identify and glue together (with the underscore sign) the elements of a multi-word connective and then give it a tag. The search is accomplished by successively selecting the entries of dictionary B, which have been ordered by descending length and using the entry involved as the search profile within the text. The algorithm includes suitable break-points so as to avoid diminishing returns on a considerable amount of pattern-matching effort.

5.12.2.5 Disambiguation

The last procedure within this phase is the resolution of any ambiguities that may have been created by the tagging programs. Generally, we have observed three sources of ambiguity:

1. Multiple function: In some cases ambiguity results when the tagged connective fails to meet the specifications we have set in Chapter 4 for delimiting the concept of "connective". This is true when an entity such as and, but, or or as (in English) and wa or 'aw (in Arabic) functions as a "phrasal connective" (in van Dijk's (1977b) use of the term, i.e. a connective that combines two nouns, adjectives, adverbs or prepositions, hence a connective that does not function as an inter-clausal or inter-sentential connective).

2. Homography: Ambiguity due to homography results from the identity of word forms: from the coincidence of assigned head words. For instance, assigning a tag to an entry "when" leads to tagging not only the subordinator but also the question-word.

3. Polysemy: Occasionally, a tagged form can have a number of related meanings, one of which satisfies the specifications of a

connective. For instance, the form certainly has a number of related meanings in the corpus, one of which has a cohesive force, i.e. it relates the subsequent stretch of text to the current one, and is therefore treated as a connective. And yet the programs would tag all occurrences of certainly. As a result, disambiguating such forms demands careful consideration of meaning and context, and is therefore more delicate and more time consuming than resolving the other two types of ambiguity.

The resolution of ambiguities is performed by dissociating the tags from the problematic expressions. This task has been achieved manually.

5.12.3 Patterning

This is an intermediate phase with a very specific aim: to examine the textual context of connectives. It is performed in two procedures:

1. Connective deletion: A SPITBOL program is written that would delete all occurrences of connectives in the corpus and replace them with < > (see a sample in Appendix 7B). The output provides the opportunity of examining the text without its connectives. More specifically, we are interested in two related questions; one concerns the cohesive force of connectives and the other concerns their positioning. These questions are:

a. What impact has the presence or absence of connectives on the cohesiveness of a text and on its organisation?

b. What type of positions do connectives occupy within the structure of the sentence and are they rhetorically determined?

2. Connective retention: A SPITBOL program is written to delete every word in the text except connectives. Each word deleted is replaced by hyphens, each standing for a letter. All text references, punctuation marks, paragraph indentations are retained. The result is a map where connectives are explicitly located along the text. The procedure offers a quick examination of the patterns of distribution of connectives within and across the paragraphs (see a sample of output in Appendix 7C).

5.12.4 Categorisation

5.12.4.1 Preliminaries

Procedures of the next phase aim to categorise connectives for functionality and range of operation. This is an essential phase in the project and is justified by the intricacy of the semantic relations that connectives signal in the text. Features of the categorisation scheme are discussed in Chapter 6 below, but we would like here to outline levels and steps in the process of categorisation.

5.12.4.2 Levels of Functionality

Categorisation is applied to three levels of functionality, each receiving its own distinctive tags. Although the procedures for tagging the three levels have been performed simultaneously, we have followed a bottom-up approach by starting with the lowest, i.e. most delicate, level and then moving up to the more general, more enclosing one. These levels are:

1. Level one: This is the most delicate level in our categorisation scheme. Connectives are tagged for the type of specific semantic relation that they signal. Note that it is possible to have one or more lower levels than this one, depending on the semantic/rhetorical depth that the investigator would like to probe. However, for pragmatic reasons, we have opted to start at this level.

2. Level two: The next level is more general and represents grouping of level one semantic relations into larger classes, each with similar characteristics of functionality. A separate tag is allocated for each class and placed next to level one tag.

3. Level three: The third level is the broadest in the categorisation scheme and represents further grouping according to textual/rhetorical functions. Three categories are posited and each is given its distinctive tag. (Details of these categories are in Chapter 10).

5.12.4.3 Range of Operationality

An additional feature in categorisation is tagging the range of textual operationality of each connective. The concept of range will be discussed in more detail in Chapter 9 (cf. also Weise 1982). Each connective is examined and a tag is attached to it indicating the type of range. Four types of range have been specified and tagged.

1. Immediate range: This is tagged as "GI" and refers to

connectives that operate within the sentence boundaries.

2. Short range: Connectives of this type are inter-sentential: they relate two clauses/clause-complexes each being autonomous within a sentence boundary. The two related entities (sentences) are next to each other and are not separated by text. The tag used is "GS".

3. Medium Range: This label subsumes connectives that relate two sentences separated by text but occurring within the paragraph boundary. The tag used is "GM".

4. Long range: We use this label to refer to connectivity that is operational across the paragraph boundaries. The tag used is "GL".

5.12.4.4 Connective Tags: An Example

An example of a connective that is fully tagged is "^First[XAn`GL]". The tags fall in three parts. The first is the tag "^" placed at the front of a word to label it as a connective or a connective core (in the case of a compound connective) or the first element in a correlative connective. In the example, the word "First" is classified as a connective.

The other two parts refer to functionality and range. They are enclosed in square brackets [ ] placed at the back of the connective and is thus conveniently separated from the actual word form. The lower case letter "n" indicates the first (low) level of functionality: the connective is identified as an enumerative (see Chapter 6). The tag "A" indicates the category to which enumeratives belong, i.e. additive connectives, and, therefore, represents the second level of functionality. The third, broad, level is indicated by the tag "X" which, as well be discussed in Chapter 10, represents a grouping of certain functions under the label "textual extension". A list of these tags are given in Chapter 6.

The tag "` " is a separator and is used to separate the tag of functionality from the range tag. The "GL" is the third type of tags and indicates range of connectivity , i.e. that this connective is used to list the first of a number of views, points, instances, etc. related to a previous paragraph.

5.12.4.5 Some Reserved Tags

In addition to these tags, a few others have been reserved to assign more information to the connectives. The tag "P" is added to the functional tag to indicate that the connective is compound, i.e. multi-word. For instance the connective First of all is tagged [XAnP GL].

Another tag is "C" which is, again, placed after the functional tag to indicate that the tagged word is the first element of a correlative connective. Thus the tag used with "either" indicates that the connective is "either ... or"; if it is not used, then "either" is a simple single-word connective.

Another reserved tag is "X" used in two positions with a small number of connectives:

a. It is added to the functionality tag "NCn" to indicate that the connective (particularly then in English or fa in Arabic) expresses a result of a condition and is used with a conditional connective, particularly if or 'iḏā.

b. "X" is also used instead of the range tag to indicate that the connective, because of its exophoric meaning, cannot be tagged for a specific range. Such connectives include again and once more.

5.12.4.6 Editing and Verification of Tags

Since tagging the various categories of connectives is basically a manual operation, though some simple computer routines are written to tag some obvious connectives, errors and discrepancies are bound to occur. The next step is to verify the tags, correct any error and resolve any ambiguity. This is performed by producing word lists and concordances and then carefully checking the tags against the connective textual function and its range of operationality. Occasionally a large textual context is required to verify a particular tag or check delicate discrepancies in the scheme. This is obtained by instructing OCP to produce concordances with large contexts for key words (about 400 words).

Another aspect of this procedure is removing some categorial inconsistencies so that categories are better specified. This is related to the fact that each category of semantic function represents a continuum and not a discrete class, which in effect maximises the difficulty of establishing coherent classes of meaning. This is a problem that lexicographers, for instance, are well aware of. Accordingly, in some cases, two connectives that have been tagged as belonging to two different categories of semantic relations may on closer inspection be considered as belonging to one category but on two different places of the continuum. This results in readjustment of some categories and reconsideration of their main textual features.

5.12.4.7 Word Lists and Concordances of Connectives

The next step is instructing OCP to produce word lists and concordances for tagged connectives. These are intended to be used for a discussion of the functionality and patterning of connectives as will be discussed in Chapter 6. We have requested two types of word lists and concordances. The first is general, i.e. all connectives are picked as key words and sorted alphabetically. The second one is category-specific, i.e. only connectives belonging to a particular category are picked. These are, again, of two types. The first comprises word lists and concordances of connectives of main categories (of level two) while the second consists of those of more specific categories (of level one). Thus while we have one concordance for all additive connectives, we have several others for connectives of additive subcategories.(24)

5.12.5 Quantification

The last phase in the processing of connectives consists of procedures for constructing the calculus of connectives (see Chapter 1), i.e. finding the statistical profile of connectives of the two corpora so as to provide a basis for the later contrastive statements. The procedures consist of designing a suite of SPITBOL programs for statistical analysis and running them on the two tagged corpora. No manual work or intervention is necessary at this stage.

However, in implementing these procedures we have been aware of two pitfalls that, if present, may reduce what has started as a careful piece of research into a crude form of empiricism. These pitfalls are:

a. Uninformed use: Within quantitative linguistics there exist a large number of statistical procedures available to the linguist. The danger lurks where procedures are utilised with little understanding of their assumptions or mathematical bases, which may lead to erroneous or at least irrelevant analysis or synthesis. It has, therefore, been important within this phase of processing to identify only those procedures that are essential for producing a meaningful description of the quantitative properties of connectives.

b. Over-access: Computer techniques are capable of producing large amounts of information on a single pass of a data file. This can easily lure the researcher to pile more statistical information than is required which may, as a result, either divert the course of the research or slow down its progress or both. We have, therefore, requested only the information for which there is theoretical expectation (as expressed in the functioning of the investigatory apparatus in Chapters 1 and 2).

Accordingly, the measurements to be performed on the two corpora have to be sufficiently relevant and theoretically necessary for the aims of the project. All irrelevant or remotely relevant measurements of various types of characteristics will be avoided. Needless to say that the measurements, and the programs that implement them, should be mathematically accurate in order to proffer a correct profile of the object investigated. In this project, measurements consist of a global statistical profile of connectives with special reference to their distribution, growth and interval/gap calculation. The results of these researches will be discussed in Chapters 7 and 8.

5.13 Conclusion

The observational approach adopted in this project for the identification and categorisation of the cohesive role of connectives and the description of their quantitative properties across English and Arabic requires two corpora of naturally produced material to operate on. Particular attention has to be paid to the selection of the texts that make up each corpus and therefore a set of criteria has to be defined and consistently followed.

The corpora are based on journalistic texts selected from quality morning newspapers published during the period during the period October 1982 to July 1983. Each corpus is made up of ca. 250,000 words of running texts. Selection is accomplished via a combination of random and stratified sampling procedures that ensure that the selection and subsequent ordering and organisation of text do not determine or imply knowledge of the phenomenon under investigation.

Early in the project, and this is ascertained in a feasibility experiment, a decision has been taken to assemble the two corpora in machine-readable form in order to facilitate the use of computer techniques for investigatory purposes. Automatic means, via an OCR, have been used with success for inputting the English corpus on the computer. The Arabic corpus, because of the intricacy of the script and the incompatibility of the OCR software, cannot be input in the same way. A decision is then taken to input the Arabic corpus manually.

This decision prompted a reconsideration of the available Romanisation schemes of the corpus. Two schemes have been introduced, one for general use in citations of various lengths that appear in the thesis, while the other is reserved for the computerised representation, and is used in all stages of the processing.

The first stage in text processing involves the production of global word lists and concordances. Consideration of the status of the word as a unit of linguistic measurement has made it mandatory to adopt a tagging system in order to help the computer ignore inflectional variations of case, mood and word juxtaposition that, otherwise, adversely affect word sorting and the general statistics of the corpus.

The next stage involves identification of connectives and requires as a first step the assembling of lists of connectives to be used as associative tables in pattern matching and dictionary lookup techniques. The lists are culled from a number of sources, which represent the various codifications of linguistic forms:

grammars, concordances and lists appearing as by-products in some limited studies on cohesion and connectives.

After connectives are identified and tagged in the corpora, three computer-aided analyses are conducted, whereby connective patterns are studied, their semantic role categorised and their quantitative properties calculated. These tasks have required a comprehensive tagging system to represent functionality and range of operationality of connectives, and a detailed suite of SPITBOL programs to carry out the analyses.

Among the immediate results of these researches is the generation of word lists and concordances of connectives, both general (no categorisation of function and range indicated) and specific (generated and sorted according to functional categories). These are used for examining and discussing the cohesive role of connectives, which is described in the next chapter. The other direct result is the production of a statistical profile of connectives in English and Arabic. This is described in detail in Chapters 7 and 8.

Footnotes to Chapter 5

(1) Probably the most prominent centres of Baghdad, Beirut-Damascus, Cairo, and Tangier-Fez-Marrakesh constitute the loci of the prestige dialects at the present time (cf. Snow 1965).

(2) Beeston (1970 p.11) uses the term "vernaculars" to designate these "dialects". These "form a continuous spectrum of variation, of which the extremities, Moroccan and Iraqi, differ to the point of mutual unintelligibility" (loc. cit.). However, there is a form of "educated" spoken Arabic that, despite variations, can enable communication across the Arabic speaking world. A project conducted in Leeds University under the supervision of Professor T. Mitchell studied this form of language as used by educated speakers of Arabic of various dialectal backgrounds.

3) Traditional Arab grammarians rely heavily, for the purposes of setting the rules and offering illustrative examples, on a corpus of poetry, orations and quotations from what is traditionally termed the "age of exemplification" as well as on the text of the Holy Qur'an and the texts of the Prophet's teachings. The "age of exemplification" covers, roughly, pre-Islamic and early Islamic periods up to the end of the Second Century of the Hijra calendar (up to the early Ninth Century A.D)

(4) Other names currently in use are "Modern Literary Arabic" in spite of the fact that many of its manifestations, e.g. newspapers or educational textbooks, have nothing to do with literature; "Modern Written Arabic", although it can be used for formal addresses; "Classical Arab", although this confuses it with the variety used in the 6th-12th centuries; and, less frequently, "Contemporary Arabic" or "Contemporary Standard Arabic". In default of a more satisfactory term, "Modern Standard Arabic" is used here.

(5) It has been noted (cf. for instance Snow 1965) that this situation presents a classical case of "diglossia". Ferguson (1959 p.336) defines this term in the following way:

> "DIGLOSSIA is a relatively stable language situation in which, in addition to the primary dialects of the language (which may include a standard or regional standards), there is a very divergent, highly codified (often grammatically complex) superposed variety, the vehicle of a large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely by formal education and is used for most written and formal spoken purposes but is not used by any sector of the community for ordinary conversation."

Another view on this situation is expressed by Professor Roger Allen of the University of Pennsylvania (Personal Communication in 1987) who believes that the term diglossia is unfortunate in as far as it is applied to Arabic. The language, he believes, falls on a continuum with "fuṣḥa" on one end and the vernacular on the other.

Education and proper schooling helps the speaker to have a wider range of mastery on the continuum, approaching the "fuṣḥa" end.

(6) Stratified sampling was used by Bathurst in assembling his corpus of Arabic (see Bathurst 1969, 1976).

(7) On account of this (and other pieces of) evidence, Herdan makes an interesting conclusion, that "the word order in a linguistic text is, partly at least, governed by chance, as against the popular view held by psychologists and linguists that it is completely deliberate, or determined by acts of decision on the part of the writer" (1967 p.99).

(8) For treatment of statistical choice of language material, see Tesitelova 1972 a-b. See also the procedures in Bathurst (1970), Kraus (1972) and Knowles (1981).

(9) Because of the size of the program jobs, provision was permitted for unlimited disk space and unlimited CPU time to facilitate the completion of major computer runs.

(10) The first major run to test the Cluster system during installation was made using OCP to produce complete word lists for this project.

(11) Aston University Computing Service was the first few academic institutions to install a KDEM machine in the U.K. The equipment was purchased in 1982 and installed in 1983 for the purpose of providing a data entry service for University departments and for academic institutions in Britain. Our English corpus was the first major text to be scanned with this machine.

(12) This is the main reason for excluding texts from the Financial Times from the English corpus. The paper colour makes the texts extremely difficult to read by KDEM, which normally functions best when the paper colour is white so that the image of the characters is more readily and accurately recognised.

(13) Correspondence with the manufacturers of KDEM has revealed that it is possible to create software compatible for scanning Arabic if sufficient funds are offered.

(14) On this, Reed and Schonfelder (1978 p.59) state that "use of COCOA for studies where investigation of word association is important are possible only by clumsy and somewhat artificial means".

(15) The design aim of CLOC is five-fold: a) CLOC is to have a simple readable command structure; b) the processes of text reading and analysis are to be separated; c) extensive and flexible key word selection is to be provided; d) CLOC is to be able to produce word lists, concordances and collocations; and, e) there are to be no explicit limits to either text or vocabulary size.

(16) A new version of OCP is to be released shortly (see Hockey

1986c for a short description). Another version is being prepared for use on PCs.

(17) Details of the tagging system of the Brown corpus are given in Greene and Rubin (1971), Francis (1979a, reprinted in Johansson 1982) and Francis (1979b). Details of the tagging systems of the LOB corpus are given in Garside and Leech (1982), Johansson and Jahr (1982), Leech, Garside and Atwell (1983) and Atwell, Leech and Garside (1984). The Dutch corpus of modern English texts and its tagging system is reported in Aarts and Heuvel (1980).

(18) Description of the London-Lund corpus and its tagging system are reported in Quirk and Svartvik (1979), Svartvik (1980), Svartvik and Eeg-Olofsson (1982), Svartvik et al. (1982).

(19) Other description of tagging procedures are given in Haan (1984) and Stenström (1984).

(20) As an illustrative example of this type, we reproduce this excerpt from the LOB. The untagged text runs like this:

```
       "  'Stop Electing Life Peers'
          By Trevor Williams
A move to stop Mr. Gaitskell from nominating more Labour
life peers is to be made at a meeting of Labour MPs tomorrow.  "
```

A horizontal format of the tagged version of the same text looks like this:

```
A01  2  ^ *'_*' stop_VB electing_VBG life_NN peers_NNS **'_**' ._.
A01  3  ^ by_IN Trevor_NP Williams_NP ._.
A01  4    ^ a-AT move_NN to_TO stop_VB \0Mr_NPT Gaitskell_NP from_IN
A01  4  nominating_VBG any_DTI more_AP labour_NN
A01  5  life_NN peers_NNS is_BEZ to_TO be_BE made_VBN at_IN a_AT
        meeting_NN
A01  5  of_IN labour_NN \0MPs_NPTS tomorrow_NR ._.
```

The vertical format has each word on a separate record:

```
 A01  2  001  ----- ----------------------
 A01  2  002  *'    *'               H
 A01  2  010  VB    stop             H
 A01  2  020  VBG   electing         H
 A01  2  030  NN    life             H
 A01  2  040  NNS   peers            H
 A01  2  041  **'   **'              H
 A01  2  042  .     .                H    @
 A01  3  001  ----- ----------------------
 A01  3  010  IN    by               H
 A01  3  020  NP    Trevor           H
 A01  3  030  NP    Williams         H
 A01  3  031  .     .                H    @
 A01  4  001  ----- ----------------------
 A01  4  010  AT    a                      P
 A01  4  020  NN    move
 A01  4  030  TO    to
 A01  4  040  VB    stop
 A01  4  050  NPT   \0Mr       \0
 A01  4  060  NP    Gaitskell
 A01  4  070  IN    from
 A01  4  080  VBG   nominating
 A01  4  090  DTI   any
 A01  4  100  AP    more
 A01  4  110  NN    labour         N
 A01  5  010  NN    life
 A01  5  020  NNS   peers          N
etc..
```

(21) Below is an example from the Dutch CCPP (mentioned in Aarts and Heuvel 1980 p.2) to illustrate the use of a four-digit code for tagging word class and constituent boundary.

The	21	(determiner, def. art)	02	
roof	31	(common noun, sg, common case)	02	
of	91	(preposition)	03	
the	21		04	
house	31		01	(02,03,04)
collapsed	A3	(verb intransitive, pa. tense)	01	
when	63	(conjunction, subordinator)	02	
a	25	(determiner, indef. art)	04	
bomb	31		03	(04)
fell	A3		00	(01,02,03)

The first digit of the word class code indicates the major word class, the second carries information about subclasses or about morphological characteristics. For verbs there is a letter, instead of a digit, in the first position of the word class code. This letter indicates that the word in question is a verb. In the constituent boundary code, the end of a sentence is marked by assigning 00 to the last word. The last word in each immediate constituent is marked 01 (see further details in Aarts and Heuvel op. cit.).

(22) A common strategy for assigning information in lists to individual items is that of table searching. When the keys to the entries in the tables are written words or similar linguistic elements, as is the case in the lists of connectives, the technique is referred to as dictionary lookup (cf. Hays 1967, Mepham 1973).

(23) This arrangement has been followed carefully. However, in a small number of cases, it is difficult to determine with confidence which element is the core of the compound connective. Selecting and tagging a core in such cases is governed by pragmatic considerations (such as methodological convenience or consistency).

(24) Note that general and specific concordance of main categories are included in the microfiche appendices. Concordances of subcategories are not included. They are, however, available for inspection from the library of the Department of Modern Languages, Aston University, or from the author personally.