

Contents

1	Introduction	1
	Peter Spyns	
1.1	Context.....	1
1.2	STEVIN Projects	4
1.3	Mission Accomplished	10
1.4	Organisation of This Volume	15
	References.....	15

Part I How It Started

2	The STEVIN Programme: Result of 5 Years Cross-border HLT for Dutch Policy Preparation	21
	Peter Spyns and Elisabeth D’Halleweyn	
2.1	Context.....	21
2.2	Historical Background	22
2.3	The STEVIN Programme	25
2.4	Discussion	35
2.5	Conclusion	37
	References.....	38

Part II HLT Resource-Project Related Papers

3	The JASMIN Speech Corpus: Recordings of Children, Non-natives and Elderly People	43
	Catia Cucchiarini and Hugo Van hamme	
3.1	Introduction	43
3.2	Potential Users of HLT Applications	44
3.3	The Need for Dedicated Corpora	45
3.4	JASMIN-CGN: Aim of the Project	46
3.5	Material and Methods	47

- 3.6 Results 53
- 3.7 Discussion 57
- 3.8 Related Work and Contribution to the State of the Art 57
- References 58
- 4 Resources Developed in the Autonomata Projects 61**
 - Henk van den Heuvel, Jean-Pierre Martens, Gerrit Bloothoof, Marijn Schraagen, Nanneke Konings, Kristof D’hanens, and Qian Yang
 - 4.1 Introduction 61
 - 4.2 The Autonomata Spoken Names Corpus (ASNC) 62
 - 4.3 The Autonomata Transcription Toolbox 67
 - 4.4 The Autonomata P2P Converters 74
 - 4.5 The Autonomata TOO POI Corpus 74
 - References 78
- 5 STEVIN Can Praat 79**
 - David Weenink
 - 5.1 Introduction 79
 - 5.2 The KlattGrid Acoustic Synthesiser 80
 - 5.3 Vowel Editor 89
 - 5.4 Robust Formant Frequency Analysis 90
 - 5.5 Availability of the Mathematical Functions in the GNU Scientific Library 92
 - 5.6 Search and Replace with Regular Expressions 92
 - 5.7 Software Band Filter Analysis 93
 - 5.8 Conclusion 93
 - References 94
- 6 SPRAAK: Speech Processing, Recognition and Automatic Annotation Kit 95**
 - Patrick Wambacq, Kris Demuyne, and Dirk Van Compernelle
 - 6.1 Introduction 95
 - 6.2 Intended Use Scenarios of the SPRAAK Toolkit 96
 - 6.3 Features of the SPRAAK Toolkit 100
 - 6.4 SPRAAK Performance 108
 - 6.5 SPRAAK Requirements 108
 - 6.6 SPRAAK Licensing and Distribution 109
 - 6.7 SPRAAK in the STEVIN Programme 109
 - 6.8 Future Work 110
 - 6.9 Conclusions 111
 - References 112

7	COREA: Coreference Resolution for Extracting Answers for Dutch	115
	Iris Hendrickx, Gosse Bouma, Walter Daelemans, and Véronique Hoste	
	7.1 Introduction	115
	7.2 Related Work	116
	7.3 Material and Methods	117
	7.4 Evaluation	122
	7.5 Conclusion	125
	References	126
8	Automatic Tree Matching for Analysing Semantic Similarity in Comparable Text	129
	Erwin Marsi and Emiel Krahmer	
	8.1 Introduction	129
	8.2 Analysing Semantic Similarity	130
	8.3 DAESO Corpus	132
	8.4 Memory-Based Graph Matcher	133
	8.5 Experiments	134
	8.6 Related Work	141
	8.7 Conclusions	143
	References	144
9	Large Scale Syntactic Annotation of Written Dutch: Lassy	147
	Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste	
	9.1 Introduction	147
	9.2 Annotation and Representation	148
	9.3 Querying the Treebanks	151
	9.4 Using the Lassy Treebanks	157
	9.5 Validation	160
	9.6 Conclusion	161
	References	163
10	Cornetto: A Combinatorial Lexical Semantic Database for Dutch ...	165
	Piek Vossen, Isa Maks, Roxane Segers, Hennie van der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke	
	10.1 Introduction	165
	10.2 Related Work	167
	10.3 The Design of the Database	168
	10.4 Building the Database	174
	10.5 Editing the Cornetto Database	176
	10.6 Qualitative and Quantitative Results	177
	10.7 Acquisition Toolkits	180

10.8	Further Development of Cornetto	181
10.9	Conclusion	182
	References	183
11	Dutch Parallel Corpus: A Balanced Parallel Corpus for Dutch-English and Dutch-French	185
	Hans Paulussen, Lieve Macken, Willy Vandeweghe, and Piet Desmet	
11.1	Introduction	185
11.2	Corpus Design and Data Acquisition	186
11.3	Corpus Processing	189
11.4	Corpus Exploitation	192
11.5	Conclusion	197
	References	198
12	Identification and Lexical Representation of Multiword Expressions	201
	Jan Odijk	
12.1	Introduction	201
12.2	Multiword Expressions	202
12.3	Identification of MWEs and Their Properties	203
12.4	Lexical Representation of MWEs	207
12.5	The DuELME Lexical Database	210
12.6	Concluding Remarks	214
	References	215
13	The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch	219
	Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman	
13.1	Introduction	219
13.2	Corpus Design and Data Acquisition	221
13.3	Corpus (Pre)Processing	227
13.4	Corpus Annotation	230
13.5	Concluding Remarks	242
	References	244
Part III HLT-Technology Related Papers		
14	Lexical Modeling for Proper name Recognition in Autonomata Too	251
	Bert Réveil, Jean-Pierre Martens, Henk van den Heuvel, Gerrit Bloothoof, and Marijn Schraagen	
14.1	Introduction	251
14.2	Formerly Proposed Approaches	252
14.3	Potential for Further Improvement	256
14.4	A Novel Pronunciation Modeling Approach	257

14.5	Experimental Validation	261
14.6	Conclusions.....	268
	References.....	269
15	N-Best 2008: A Benchmark Evaluation for Large Vocabulary Speech Recognition in Dutch.....	271
	David A. van Leeuwen	
15.1	Introduction.....	271
15.2	The N-Best Project.....	273
15.3	The N-Best Evaluation.....	276
15.4	Results.....	279
15.5	Discussion and Conclusions.....	285
	References.....	287
16	Missing Data Solutions for Robust Speech Recognition.....	289
	Yujun Wang, Jort F. Gemmeke, Kris Demuynck, and Hugo Van hamme	
16.1	Introduction.....	289
16.2	Missing Data Techniques.....	290
16.3	Material and Methods: Sparse Imputation.....	291
16.4	Experiments: Sparse Imputation.....	292
16.5	Material and Methods: Gaussian-Dependent Imputation.....	295
16.6	Experiments: Gaussian-Dependent Imputation.....	298
16.7	Discussion and Conclusions.....	301
	References.....	302
17	Parse and Corpus-Based Machine Translation.....	305
	Vincent Vandeghinste, Scott Martens, Gideon Kotzé, Jörg Tiedemann, Joachim Van den Bogaert, Koen De Smet, Frank Van Eynde, and Gertjan van Noord	
17.1	Introduction.....	305
17.2	Syntactic Analysis.....	307
17.3	The Transduction Grammar.....	308
17.4	The Transduction Process.....	311
17.5	Generation.....	314
17.6	Evaluation.....	315
17.7	Conclusions and Future Work.....	316
	References.....	317

Part IV HLT Application Related Papers

18	Development and Integration of Speech Technology into Courseware for Language Learning: The DISCO Project.....	323
	Helmer Strik, Joost van Doremalen, Jozef Colpaert, and Catia Cucchiarini	
18.1	Introduction.....	323
18.2	DISCO: Aim of the Project.....	324

18.3	Material and Methods: Design.....	325
18.4	Results	332
18.5	Related Work and Contribution to the State of the Art	335
18.6	Discussion and Conclusions	337
	References.....	337
19	Question Answering of Informative Web Pages:	
	How Summarisation Technology Helps	339
	Jan De Belder, Daniël de Kok, Gertjan van Noord, Fabrice Nauze, Leonoor van der Beek, and Marie-Francine Moens	
19.1	Introduction.....	339
19.2	Problem Definition	340
19.3	Cleaning and Segmentation of Web Pages	341
19.4	Rhetorical Classification	344
19.5	Sentence Compression	346
19.6	Sentence Generation	350
19.7	Proof-of-Concept Demonstrator	353
19.8	Conclusions.....	355
	References.....	355
20	Generating, Refining and Using Sentiment Lexicons	359
	Maarten de Rijke, Valentin Jijkoun, Fons Laan, Wouter Weerkamp, Paul Ackermans, and Gijs Geleijnse	
20.1	Introduction.....	359
20.2	Related Work	361
20.3	Generating Topic-Specific Lexicons	363
20.4	Data and Experimental Setup.....	367
20.5	Qualitative Analysis of Lexicons	367
20.6	Quantitative Evaluation of Lexicons	368
20.7	Bootstrapping Subjectivity Detection	370
20.8	Mining User Experiences from Online Forums.....	373
20.9	Conclusion.....	375
	References.....	376
Part V And Now		
21	The Dutch-Flemish HLT Agency: Managing the Lifecycle of STEVIN's Language Resources	381
	Remco van Veenendaal, Laura van Eerten, Catia Cucchiari, and Peter Spyns	
21.1	Introduction.....	381
21.2	The Flemish-Dutch HLT Agency.....	382
21.3	Managing the Lifecycle of STEVIN Results.....	384
21.4	Target Groups and Users	390
21.5	Challenges Beyond STEVIN	391

21.6	Conclusions and Future Perspectives	392
	References	393
22	Conclusions and Outlook to the Future	395
	Jan Odijk	
22.1	Introduction	395
22.2	Results of the STEVIN Programme	395
22.3	Desiderata for the Near Future	397
22.4	Future	399
22.5	Concluding Remarks	403
	References	403
	Index	405



<http://www.springer.com/978-3-642-30909-0>

Essential Speech and Language Technology for Dutch
Results by the STEVIN-programme

Spyns, P.; Odijk, J. (Eds.)

2013, XVII, 413 p. 79 illus., 29 illus. in color., Hardcover

ISBN: 978-3-642-30909-0