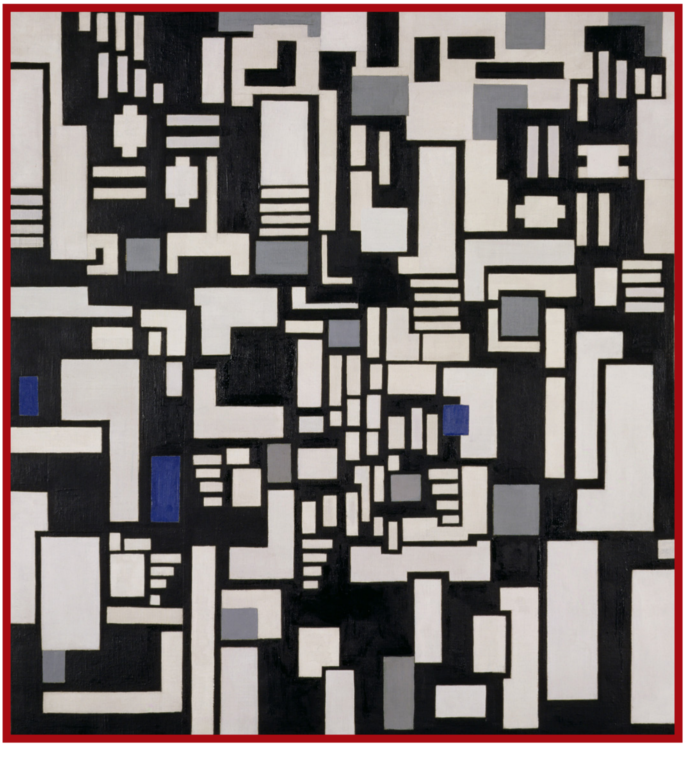
The Handbook of
Computational Linguistics and Natural Language Processing



Edited by

Alexander Clark, Chris Fox, and Shalom Lappin

WILEY-BLACKWELL

Praise for *The Handbook of Computational Linguistics and Natural Language Processing*

"All in all, this is very well compiled book, which effectively balances the width and depth of theories and applications in two very diverse yet closely related fields of language research."

Machine Translation

"This *Handbook* is exceptionally broad and exceptionally deep in its coverage. The contributions, by noted experts, cover all aspects of the field, from fundamental theory to concrete applications. Clark, Fox and Lappin have performed a great service by compiling this volume."

Richard Sproat, Oregon Health & Science University

Blackwell Handbooks in Linguistics

This outstanding multi-volume series covers all the major subdisciplines within linguistics today and, when complete, will offer a comprehensive survey of linguistics as a whole.

Already published:

The Handbook of Child Language
Edited by Paul Fletcher and Brian MacWhinney

The Handbook of Phonological Theory, Second Edition
Edited by John A. Goldsmith, Jason Riggle, and Alan C. L. Yu

The Handbook of Contemporary Semantic Theory
Edited by Shalom Lappin

The Handbook of Sociolinguistics
Edited by Florian Coulmas

The Handbook of Phonetic Sciences, Second Edition
Edited by William J. Hardcastle and John Laver

The Handbook of Morphology
Edited by Andrew Spencer and Arnold Zwicky

The Handbook of Japanese Linguistics
Edited by Natsuko Tsujimura

The Handbook of Linguistics
Edited by Mark Aronoff and Janie Rees-Miller

The Handbook of Contemporary Syntactic Theory
Edited by Mark Baltin and Chris Collins

The Handbook of Discourse Analysis
Edited by Deborah Schiffrin, Deborah Tannen, and Heidi E. Hamilton

The Handbook of Language Variation and Change
Edited by J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes

The Handbook of Historical Linguistics
Edited by Brian D. Joseph and Richard D. Janda

The Handbook of Language and Gender
Edited by Janet Holmes and Miriam Meyerhoff

The Handbook of Second Language Acquisition
Edited by Catherine J. Doughty and Michael H. Long

The Handbook of Bilingualism and Multilingualism, Second Edition
Edited by Tej K. Bhatia and William C. Ritchie

The Handbook of Pragmatics
Edited by Laurence R. Horn and Gregory Ward

The Handbook of Applied Linguistics
Edited by Alan Davies and Catherine Elder

The Handbook of Speech Perception
Edited by David B. Pisoni and Robert E. Remez

The Handbook of the History of English
Edited by Ans van Kemenade and Bettelou Los

The Handbook of English Linguistics
Edited by Bas Aarts and April McMahon

The Handbook of World Englishes
Edited by Braj B. Kachru; Yamuna Kachru, and Cecil L. Nelson

The Handbook of Educational Linguistics
Edited by Bernard Spolsky and Francis M. Hult

The Handbook of Clinical Linguistics
Edited by Martin J. Ball, Michael R. Perkins, Nicole Müller, and Sara Howard

The Handbook of Pidgin and Creole Studies
Edited by Silvia Kouwenberg and John Victor Singler

The Handbook of Language Teaching
Edited by Michael H. Long and Catherine J. Doughty

The Handbook of Language Contact
Edited by Raymond Hickey

The Handbook of Language and Speech Disorders
Edited by Jack S. Damico, Nicole Müller, Martin J. Ball

The Handbook of Computational Linguistics and Natural Language Processing
Edited by Alexander Clark, Chris Fox, and Shalom Lappin

The Handbook of Language and Globalization
Edited by Nikolas Coupland

The Handbook of Hispanic Linguistics
Edited by Manuel Díaz-Campos

The Handbook of Language Socialization
Edited by Alessandro Duranti, Elinor Ochs, and Bambi B. Schieffelin

The Handbook of Intercultural Discourse and Communication
Edited by Christina Bratt Paulston, Scott F. Kiesling, and Elizabeth S. Rangel

The Handbook of Historical Sociolinguistics
Edited by Juan Manuel Hernández-Campoy and Juan Camilo Conde-Silvestre

The Handbook of Hispanic Linguistics
Edited by José Ignacio Hualde, Antxon Olarrea, and Erin O'Rourke

The Handbook of Conversation Analysis
Edited by Jack Sidnell and Tanya Stivers

The Handbook of English for Specific Purposes
Edited by Brian Paltridge and Sue Starfield

The Handbook of Computational Linguistics and Natural Language Processing

Edited by

Alexander Clark, Chris Fox,
and Shalom Lappin

WILEY-BLACKWELL
A John Wiley & Sons, Ltd., Publication

This paperback edition first published 2013

Edition History: Blackwell Publishing Ltd (hardback, 2010)

Blackwell Publishing was acquired by John Wiley & Sons in February 2007. Blackwell's publishing program has been merged with Wiley's global Scientific, Technical, and Medical business to form Wiley-Blackwell.

Registered Office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Offices
350 Main Street, Malden, MA 02148-5020, USA
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, for customer services, and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication Data

The handbook of computational linguistics and natural language processing / edited by
Alexander Clark, Chris Fox, and Shalom Lappin.
p. cm. – (Blackwell handbooks in linguistics)
Includes bibliographical references and index.
ISBN 978-1-4051-5581-6 (hardcover : alk. paper) ISBN 978-1-118-34718-8 (paperback : alk. paper)
1. Computational linguistics. 2. Natural language processing (Computer science).
I. Clark, Alexander (Alexander Simon) II. Fox, Chris, 1965– III. Lappin, Shalom.
P98.H346 2010
410_.285–dc22
2010003116

A catalogue record for this book is available from the British Library.

Cover image: Theo van Doesburg, Composition IX, opus 18, 1917. Haags Gemeentemuseum,
The Hague, Netherlands / The Bridgeman Art Library.
Cover design by Workhaus.

Set in 10/12pt Palatino by SPi Publisher Services, Pondicherry, India

1 2013

For Camilla

לאחיי דוד ודניאל, ולאחותי נעמי באהבה ובהומור

Contents

List of Figures ix
List of Tables xiv
Notes on Contributors xv
Preface xxiii

Introduction 1

Part I Formal Foundations 9

1 Formal Language Theory 11
SHULY WINTNER
2 Computational Complexity in Natural Language 43
IAN PRATT-HARTMANN
3 Statistical Language Modeling 74
CIPRIAN CHELBA
4 Theory of Parsing 105
MARK-JAN NEDERHOF AND GIORGIO SATTA

Part II Current Methods 131

5 Maximum Entropy Models 133
ROBERT MALOUF
6 Memory-Based Learning 154
WALTER DAELEMANS AND ANTAL VAN DEN BOSCH
7 Decision Trees 180
HELMUT SCHMID
8 Unsupervised Learning and Grammar Induction 197
ALEXANDER CLARK AND SHALOM LAPPIN
9 Artificial Neural Networks 221
JAMES B. HENDERSON

10 Linguistic Annotation 238
MARTHA PALMER AND NIANWEN XUE
11 Evaluation of NLP Systems 271
PHILIP RESNIK AND JIMMY LIN

Part III Domains of Application 297

12 Speech Recognition 299
STEVE RENALS AND THOMAS HAIN
13 Statistical Parsing 333
STEPHEN CLARK
14 Segmentation and Morphology 364
JOHN A. GOLDSMITH
15 Computational Semantics 394
CHRIS FOX
16 Computational Models of Dialogue 429
JONATHAN GINZBURG AND RAQUEL FERNÁNDEZ
17 Computational Psycholinguistics 482
MATTHEW W. CROCKER

Part IV Applications 515

18 Information Extraction 517
RALPH GRISHMAN
19 Machine Translation 531
ANDY WAY
20 Natural Language Generation 574
EHUD REITER
21 Discourse Processing 599
RUSLAN MITKOV
22 Question Answering 630
BONNIE WEBBER AND NICK WEBB

References 655
Author Index 742
Subject Index 763

List of Figures

1.1 Chomsky’s hierarchy of languages. 39
2.1 Architecture of a multi-tape Turing machine. 45
2.2 A derivation in the Lambek calculus. 59
2.3 Productions of a DCG recognizing the language $\{a^n b^n c^n d^n e^n \mid n \geq 0\}$. 61
2.4 Derivation of the string aabbccddee in the DCG of Figure 2.3. 61
2.5 Semantically annotated CFG generating the language of the syllogistic. 66
2.6 Meaning derivation in a semantically annotated CFG. 67
2.7 Productions for extending the syllogistic with transitive verbs. 69
3.1 Recursive linear interpolation. 78
3.2 ARPA format for language model representation. 79
3.3 Partial parse. 82
3.4 A word-and-parse k-prefix. 83
3.5 Complete parse. 83
3.6 Before an adjoin operation. 84
3.7 Result of adjoin-left under NTlabel. 84
3.8 Result of adjoin-right under NTlabel. 84
3.9 Language model operation as a finite state machine. 85
3.10 SLM operation. 85
3.11 One search extension cycle. 89
3.12 Binarization schemes. 92
3.13 Structured language model maximum depth distribution. 98
3.14 Comparison of PPL, WER, labeled recall/precision error. 101
4.1 The CKY recognition algorithm. 108
4.2 Table $\mathcal{T}$ obtained by the CKY algorithm. 108
4.3 The CKY recognition algorithm, expressed as a deduction system. 109
4.4 The Earley recognition algorithm. 110
4.5 Deduction system for Earley’s algorithm. 111

4.6 Table $\mathcal{T}$ obtained by Earley's algorithm. 112

4.7 Parse forest associated with table $\mathcal{T}$ from Figure 4.2. 113

4.8 Knuth's generalization of Dijkstra's algorithm, applied to finding the most probable parse in a probabilistic context-free grammar $\mathcal{G}$. 115

4.9 The probabilistic CKY algorithm. 117

4.10 A parse of 'our company is training workers,' assuming a bilexical context-free grammar. 118

4.11 Deduction system for recognition with a 2-LCFG. We assume $w = a_1 \cdots a_n$, $a_{n+1} = \$$. 119

4.12 Illustration of the use of inference rules (f), (c), and (g) of bilexical recognition. 119

4.13 A projective dependency tree. 121

4.14 A non-projective dependency tree. 121

4.15 Deduction system for recognition with PDGs. We assume $w = a_1 \cdots a_n$, and disregard the recognition of $a_{n+1} = \$$. 123

4.16 Substitution (a) and adjunction (b) in a tree adjoining grammar. 124

4.17 The TAG bottom-up recognition algorithm, expressed as a deduction system. 125

4.18 A pair of trees associated with a derivation in a SCFG. 127

4.19 An algorithm for the left composition of a sentence w and a SCFG $\mathcal{G}$. 128

6.1 An example 2D space with six examples labeled white or black. 157

6.2 Two examples of the generation of a new hyper-rectangle in NGE. 168

6.3 An example of an induced rule in RISE, displayed on the right, with the set of examples that it covers (and from which it was generated) on the left. 169

6.4 An example of a family in a two-dimensional example space and ranked in the order of distance. 170

6.5 An example of family creation in Fambl. 171

6.6 Pseudo-code of the family extraction procedure in Fambl. 172

6.7 Generalization accuracies (in terms of percentage of correctly classified test instances) and F-scores, where appropriate, of MBL with increasing k parameter, and Fambl with $k = 1$ and increasing K parameter. 175

6.8 Compression rates (percentages) of families as opposed to the original number of examples, produced by Fambl at different maximal family sizes (represented by the x-axis, displayed at a log scale). 175

7.1 A simple decision tree for period disambiguation. 181

7.2 State of the decision tree after the expansion of the root node. 183

7.3 Decision tree learned from the example data. 183

7.4 Partitions of the two-dimensional feature subspace spanned by the features *'color'* and *'shape.'* 184

7.5 Data with overlapping classes and the class boundaries found by a decision tree. 186

7.6 Decision tree induced from the data in Figure 7.5 before and after pruning. 187
7.7 Decision tree with node numbers and information gain scores. 187
7.8 Decision tree with classification error counts. 188
7.9 Probabilistic decision tree induced from the data in Figure 7.5. 190
7.10 Part of a probabilistic decision tree for the nominative case of nouns. 194
9.1 A multi-layered perceptron. 223
9.2 Category probabilities estimated by an MLP. 226
9.3 A recurrent MLP, specifically a simple recurrent network. 227
9.4 A recurrent MLP unfolded over the sequence. 228
9.5 The SSN architecture, unfolded over a derivation sequence, with derivation decisions D^t and hidden layers S^t. 229
9.6 An SSN unfolded over a constituency structure. 232
10.1 An example PTB tree. 242
10.2 A labeled dependency structure. 243
10.3 OntoNotes: a model for multi-layer annotation. 257
12.1 Waveform (top) and spectrogram (bottom) of conversational utterance '*no right I didn't mean to imply that*.' 305
12.2 HMM-based hierarchical modeling of speech. 307
12.3 Representation of an HMM as a parameterized stochastic finite state automaton (left) and in terms of probabilistic dependences between variables (right). 307
12.4 Forward recursion to estimate $\alpha_t(q_j) = p(\mathbf{x}_1, \ldots, \mathbf{x}_t, q_t = q_j \mid \boldsymbol{\lambda})$. 309
12.5 Hidden Markov models for phonemes can be concatenated to form models for words. 311
12.6 Connected word recognition with a bigram language model. 319
12.7 Block processing diagram showing the AMI 2006 system for meeting transcription (Hain et al., 2006) 323
12.8 Word error rates (%) results in the NIST RT'06 evaluations of the AMI 2006 system on the evaluation test set, for the four decoding passes. 325
13.1 Example lexicalized parse-tree. 339
13.2 Example tree with complements distinguished from adjuncts. 340
13.3 Example tree containing a trace and the gap feature. 341
13.4 Example unlabeled dependency tree. 346
13.5 Generic algorithm for online learning taken from McDonald et al. (2005b). 347
13.6 The perceptron update. 348
13.7 Example derivation using forward and backward application. 353
13.8 Example derivation using type-raising and forward composition. 354
13.9 Example CCG derivation for the sentence *Under new features, participants can transfer money from the new funds.* 355
14.1 The two problems of word segmentation. 372
14.2 Word discovery from an MDL point of view. 378
14.3 A signature for two verbs in English. 383

14.4 Morphology discovery as local descent. 383
14.5 Building an FST from two FSAs. 390
15.1 Derivation of semantic representation with storage. 409
16.1 Basic components of a spoken dialogue system. 444
16.2 Finite state machine for a simple ticket booking application. 445
16.3 A simple frame. 445
16.4 Goal-oriented action schema. 446
16.5 A single utterance gives rise to distinct updates of the DGB for distinct participants. 469
17.1 Relative clause attachment ambiguity. 496
17.2 An example for the parse-trees generated by a probabilistic-context free grammar (PCFG) (adapted from Crocker & Keller 2006). 498
17.3 The architecture of the SynSem-Integration model, from Pado et al. (2009). 504
17.4 A simple recurrent network. 506
17.5 CIANet: a network featuring scene–language interaction with a basic attentional gating mechanism to select relevant events in a scene with respect to an unfolding utterance. 509
17.6 The competitive integration model (Spivey-Knowlton & Sedivy 1995). 510
18.1 Example dependency tree. 525
19.1 A sentence-aligned corpus. 533
19.2 A non-exact alignment. 533
19.3 In the word-based translation on the left we see that the noun–adjective reordering into English is missed. On the right, the noun and adjective are translated as a single phrase and the correct ordering is modeled in the phrase-based translation. 538
19.4 Merging source-to-target and target-to-source alignments (from Koehn 2010). 540
19.5 All possible source segmentations with all possible target translations (from Koehn 2004). 544
19.6 Hypothesis expansion via stack decoding (from Koehn 2004). 546
19.7 An aligned tree pair in DOT for the sentence pair: *he chose the ink cartridge, il a choisi la cartouche d'encre*. 552
19.8 Composition in tree-DOT. 563
20.1 Human and corpus wind descriptions for September 19, 2000. 576
20.2 An example literacy screener question (SkillSum input). 577
20.3 Example text produced by SkillSum. 577
20.4 Example SumTime document plan. 579
20.5 Example SumTime deep syntactic structure. 582
21.1 Example of the RST relation evidence. 607
22.1 Basic QA system architecture. 635
22.2 An ARDA scenario (from Small & Strzalkowski 2009). 645

22.3 An answer model for the question: *Where is Glasgow?* (Dalmas & Webber 2007), showing both Scotland and Britain as possible answers. 648
22.4 Example interaction taken from a live demonstration to the ARDA AQUAINT community in 2005. 649
22.5 Goal frame for the question: *What is the status of the Social Security system?* 649
22.6 Two cluster seed passages and their corresponding frames relative to the retirement clarification question. 650
22.7 Two cluster passages and their corresponding frames relative to the private accounts clarification question. 650

List of Tables

3.1 Headword percolation rules 91
3.2 Binarization rules 93
3.3 Parameter re-estimation results 96
3.4 Interpolation with trigram results 96
3.5 Maximum depth evolution during training 97
6.1 Examples generated for the letter–phoneme conversion task, from the word–phonemization pair *booking*–[bukIN], aligned as [b-ukI-N] 155
6.2 Number of extracted families at a maximum family size of 100, the average number of family members, and the raw memory compression, for four tasks 176
6.3 Two example families (represented by their members) extracted from the PP and CHUNK data sets respectively 177
7.1 Training data consisting of seven objects which are characterized by the features *'size,' 'color,'* and *'shape.'* The first four items belong to class '+,' the others to class '−' 182
8.1 Comparison of different tag sets on IPSM data 209
8.2 Cross-linguistic evaluation: 64 clusters, left all words, right $f \leq 5$ 212
11.1 Structure of a typical summary of evaluation results 280
11.2 Contingency table for a document retrieval task 283
16.1 NSUs in a subcorpus of the BNC 441
16.2 Comparison of dialogue management approaches 452
17.1 Conditional probability of a verb frame given a particular verb, as estimated using the Penn Treebank 499
19.1 Number of fragments for English-to-French and French-to-English HomeCentre experiments 564
20.1 Numerical wind forecast for September 19, 2000 576

Notes on Contributors

Ciprian Chelba is a Research Scientist with Google. Between 2000 and 2006 he worked as a Researcher in the Speech Technology Group at Microsoft Research.

He received his Diploma Engineer degree in 1993 from the Faculty of Electronics and Telecommunications at "Politehnica" University, Bucuresti, Romania, M.S. in 1996 and PhD in 2000 from the Electrical and Computer Engineering Department at the Johns Hopkins University.

His research interests are in statistical modeling of natural language and speech, as well as related areas such as machine learning and information theory as applied to natural language problems.

Recent projects include language modeling for large-vocabulary speech recognition (discriminative model estimation, compact storage for large models), search in spoken document collections (spoken content indexing, ranking and snipeting), as well as speech and text classification.

Alexander Clark is an Honorary Research Fellow in the Department of Computer Science at Royal Holloway, University of London. His first degree was in Mathematics from the University of Cambridge, and his PhD is from the University of Sussex. He did postdoctoral research at the University of Geneva. In 2007 he was a *Professeur invité* at the University of Marseille. He is on the editorial board of the journal *Research on Language and Computation*, and a member of the steering committee of the International Colloquium on Grammatical Inference. His research is on unsupervised learning in computational linguistics, and in grammatical inference; he has won several prizes and competitions for his research. He has co-authored with Shalom Lappin a book entitled *Linguistic Nativism and the Poverty of the Stimulus*, which is being published by Wiley-Blackwell in 2010.

Stephen Clark is a Senior Lecturer at the University of Cambridge Computer Laboratory where he is a member of the Natural Language and Information Processing Research Group. From 2004 to 2008 he was a University Lecturer at the Oxford University Computing Laboratory, and before that spent four years as a postdoctoral researcher at the University of Edinburgh's School of Informatics, working with Prof. Mark Steedman. He has a PhD in Artificial Intelligence from the University of Sussex and a first degree in Philosophy from the University of Cambridge. His main research interest is statistical parsing, with a focus on the grammar formalism combinatory categorial grammar. In 2009 he led a team at the Johns Hopkins University Summer Workshop working on "Large Scale Syntactic Processing: Parsing the Web." He is on the editorial boards of *Computational Linguistics* and the *Journal of Natural Language Engineering*, and is a Program Co-Chair for the 2010 Annual Meeting of the Association for Computational Linguistics.

Matthew W. Crocker obtained his PhD in Artificial Intelligence from the University of Edinburgh in 1992, where he subsequently held appointments as Lecturer in Artificial Intelligence and Cognitive Science and as an ESRC Research Fellow. In January 2000, Dr Crocker was appointed to a newly established Chair in Psycholinguistics, in the Department of Computational Linguistics at Saarland University, Germany. His current research brings together the experimental investigation of real-time human language processing and situated cognition in the development of computational cognitive models.

Matthew Crocker co-founded the annual conference on Architectures and Mechanisms for Language Processing (AMLaP) in 1995. He is currently an associate editor for *Cognition*, on the editorial board of Springer's *Studies in Theoretical Psycholinguistics*, and has been a member of the editorial board for *Computational Linguistics*.

Walter Daelemans (MA, University of Leuven, Belgium, 1982; PhD, Computational Linguistics, University of Leuven, 1987) held research and teaching positions at the Radboud University Nijmegen, the AI-LAB at the University of Brussels, and Tilburg University, where he founded the ILK (Induction of Linguistic Knowledge) research group, and where he remained part-time Full Professor until 2006. Since 1999, he has been a Full Professor at the University of Antwerp (UA), teaching Computational Linguistics and Artificial Intelligence courses and co-directing the CLiPS research center. His current research interests are in machine learning of natural language, computational psycholinguistics, and text mining. He was elected fellow of ECCAI in 2003 and graduated 11 PhD students as supervisor.

Raquel Fernández is a Postdoctoral Researcher at the Institute for Logic, Language and Computation, University of Amsterdam. She holds a PhD in Computer Science from King's College London for work on formal and computational modeling of dialogue and has published numerous peer-review articles on dialogue research. She has worked as Research Fellow in the Center for the Study of Language and Information (CSLI) at Stanford University and in the Linguistics Department at the University of Potsdam.

Dr **Chris Fox** is a Reader in the School of Computer Science and Electronic Engineering at the University of Essex. He started his research career as a Senior Research Officer in the Department of Language and Linguistics at the University of Essex. He subsequently worked in the Computer Science Department where he obtained his PhD in 1993. After that he spent a brief period as a Visiting Researcher at Saarbruecken before becoming a Lecturer at Goldsmiths College, University of London, and then King's College London. He returned to Essex in 2003. At the time of writing, he is serving as Deputy Mayor of Wivenhoe.

Much of his research is in the area of logic and formal semantics, with a particular emphasis on issues of formal expressiveness, and proof-theoretic approaches to characterizing intuitions about natural language semantic phenomena.

Jonathan Ginzburg is a Senior Lecturer in the Department of Computer Science at King's College London. He has previously held posts in Edinburgh and Jerusalem. He is one of the managing editors of the journal *Dialogue and Discourse*. He has published widely on formal semantics and dialogue. His monograph *The Interactive Stance: Meaning for Conversation* was published in 2009.

John A. Goldsmith is Edward Carson Waller Distinguished Service Professor in the Departments of Linguistics and Computer Science at the University of Chicago, where he has been since 1984. He received his PhD in Linguistics in 1976 from MIT, and taught from 1976 to 1984 at Indiana University. His primary interests are computational learning of natural language, phonological theory, and the history of linguistics.

Ralph Grishman is Professor of Computer Science at New York University. He has been involved in research in natural language processing since 1969, and since 1985 has directed the Proteus Project, with funding from DARPA, NSF, and other government agencies. The Proteus Project has conducted research in natural language text analysis, with a focus on information extraction, and has been involved in the creation of a number of major lexical and syntactic resources, including Comlex, Nomlex, and NomBank. He is a past President of the Association for Computational Linguistics and the author of the text *Computational Linguistics: An Introduction*.

Thomas Hain holds the degree Dipl.-Ing. with honors from the University of Technology, Vienna and a PhD from Cambridge University. In 1994 he joined Philips Speech Processing, which he left as Senior Technologist in 1997. He took up a position as Research Associate at the Speech, Vision and Robotics Group and Machine Intelligence Lab at the Cambridge University Engineering Department where he also received an appointment as Lecturer in 2001. In 2004 he joined the Department of Computer Science at the University of Sheffield where he is now a Senior Lecturer. Thomas Hain has a well established track record in automatic speech recognition, in particular involvement in best-performing ASR systems for participation in NIST evaluations. His main research interests are in speech recognition, speech and audio processing, machine learning, optimisation of large-scale statistical systems, and modeling of machine/machine interfaces. He is a member of the IEEE Speech and Language Technical Committee.

James B. Henderson is an MER (Research Professor) in the Department of Computer Science of the University of Geneva, where he is co-head of the interdisciplinary research group Computational Learning and Computational

Linguistics. His research bridges the topics of machine learning methods for structure-prediction tasks and the modeling and exploitation of such tasks in NLP, particularly syntactic and semantic parsing. In machine learning his current interests focus on latent variable models inspired by neural networks. Previously, Dr Henderson was a Research Fellow in ICCS at the University of Edinburgh, and a Lecturer in CS at the University of Exeter, UK. Dr Henderson received his PhD and MSc from the University of Pennsylvania, and his BSc from the Massachusetts Institute of Technology, USA.

Shalom Lappin is Professor of Computational Linguistics at King's College London. He does research in computational semantics, and in the application of machine learning to issues in natural language processing and the cognitive basis of language acquisition. He has taught at SOAS, Tel Aviv University, the University of Haifa, the University of Ottawa, and Ben Gurion University of the Negev. He was also a Research Staff member in the Natural Language group of the Computer Science Department at IBM T.J. Watson Research Center. He edited the *Handbook of Contemporary Semantic Theory* (1996, Blackwell), and, with Chris Fox, he co-authored *Foundations of Intensional Semantics* (2005, Blackwell). His most recent book, *Linguistic Nativism and the Poverty of the Stimulus*, co-authored with Alexander Clark, is being published by Wiley-Blackwell in 2010.

Jimmy Lin is an Associate Professor in the iSchool at the University of Maryland, affiliated with the Department of Computer Science and the Institute for Advanced Computer Studies. He graduated with a PhD in Computer Science from MIT in 2004. Lin's research lies at the intersection of information retrieval and natural language processing, and he has done work in a variety of areas, including question answering, medical informatics, bioinformatics, evaluation metrics, and knowledge-based retrieval techniques. Lin's current research focuses on "cloud computing," in particular, massively distributed text processing in cluster-based environments.

Robert Malouf is an Associate Professor in the Department of Linguistics and Asian/Middle Eastern Languages at San Diego State University. Before coming to SDSU, Robert held a postdoctoral fellowship in the Humanities Computing Department, University of Groningen (1999–2002). He received a PhD in Linguistics from Stanford University (1998) and BA in linguistics and computer science from SUNY Buffalo (1992). His research focuses on the application of computational techniques to understanding how language works, particularly in the domains of morphology and syntax. He is currently investigating the use of evolutionary simulation for explaining linguistic universals.

Prof. **Ruslan Mitkov** has been working in (applied) natural language processing, computational linguistics, corpus linguistics, machine translation, translation technology, and related areas since the early 1980s. His extensively cited research covers areas such as anaphora resolution, automatic generation of multiple-choice tests, machine translation, natural language generation, automatic summarization, computer-aided language processing, centering, translation memory, evaluation, corpus annotation, bilingual term extraction, question answering, automatic identification of cognates and false friends, and an NLP-driven corpus-based study of translation universals.

Mitkov is author of the monograph *Anaphora Resolution* (2002, Longman) and sole editor of *The Oxford Handbook of Computational Linguistics* (2005, Oxford University Press). Current prestigious projects include his role as Executive Editor of the *Journal of Natural Language Engineering* (Cambridge University Press) and Editor-in-Chief of the *Natural Language Processing* book series (John Benjamins Publishing). Ruslan Mitkov received his MSc from the Humboldt University in Berlin, his PhD from the Technical University in Dresden and he worked as a Research Professor at the Institute of Mathematics, Bulgarian Academy of Sciences, Sofia. Prof. Mitkov is Professor of Computational Linguistics and Language Engineering at the School of Humanities, Languages and Social Sciences at the University of Wolverhampton which he joined in 1995, where he set up the Research Group in Computational Linguistics. In addition to being Head of the Research Group in Computational Linguistics, Prof. Mitkov is also Director of the Research Institute in Information and Language Processing.

Dr **Mark-Jan Nederhof** is a Lecturer in the School of Computer Science at the University of St Andrews. He holds a PhD (1994) and MSc (1990) in computer science from the University of Nijmegen. Before coming to St Andrews in 2006, he was Senior Researcher at DFKI in Saarbrücken and Lecturer in the Faculty of Arts at the University of Groningen. He has served on the editorial board of *Computational Linguistics* and has been a member of the programme committees of EACL, HLT/EMNLP, and COLING-ACL.

His research covers areas of computational linguistics and computer languages, with an emphasis on formal language theory and computational complexity. He is also developing tools for use in philological research, and especially the study of Ancient Egyptian.

Martha Palmer is an Associate Professor in the Linguistics Department and the Computer Science Department of the University of Colorado at Boulder, as well as a Faculty Fellow of the Institute of Cognitive Science. She was formerly an Associate Professor in Computer and Information Sciences at the University of Pennsylvania. She has been actively involved in research in natural language processing and knowledge representation for 30 years and did her PhD in Artificial Intelligence at the University of Edinburgh in Scotland. She has a life-long interest in the use of semantic representations in natural language processing and is dedicated to the development of community-wide resources. She was the leader of the English, Chinese, and Korean PropBanks and the Pilot Arabic PropBank. She is now the PI for the Hindi/Urdu Treebank Project and is leading the English, Chinese, and Arabic sense-tagging and PropBanking efforts for the DARPA-GALE OntoNotes project. In addition to building state-of-the-art word-sense taggers and semantic role labelers, she and her students have also developed VerbNet, a public-domain rich lexical resource that can be used in conjunction with WordNet, and SemLink, a mapping from the PropBank generic arguments to the more fine-grained VerbNet semantic roles as well as to FrameNet Frame Elements. She is a past President of the Association for Computational Linguistics, and a past Chair of SIGHAN and SIGLEX, where she was instrumental in getting the Senseval/Semeval evaluations under way.

Ian Pratt-Hartmann studied Mathematics and Philosophy at Brasenose College, Oxford, and Philosophy at Princeton and Stanford Universities, gaining his PhD from Princeton in 1987. He is currently Senior Lecturer in the Department of Computer Science at the University of Manchester.

Ehud Reiter is a Reader in Computer Science at the University of Aberdeen in Scotland. He completed a PhD in natural language generation at Harvard in 1990 and worked at the University of Edinburgh and at CoGenTex (a small US NLG company) before coming to Aberdeen in 1995. He has published over 100 papers, most of which deal with natural language generation, including the first book ever written on applied NLG. In recent years he has focused on data-to-text systems and related "language and the world" research challenges.

Steve Renals received a BSc in Chemistry from the University of Sheffield in 1986, an MSc in Artificial Intelligence in 1987, and a PhD in Speech Recognition and Neural Networks in 1990, both from the University of Edinburgh. He is a Professor in the School of Informatics, University of Edinburgh, where he is the Director of the Centre for Speech Technology Research. From 1991 to 1992, he was a Postdoctoral Fellow at the International Computer Science Institute, Berkeley, CA, and was then an EPSRC Postdoctoral Fellow in Information Engineering at the University of Cambridge (1992–4). From 1994 to 2003, he was a Lecturer then Reader at the University of Sheffield, moving to the University of Edinburgh in 2003. His research interests are in the area of signal-based approaches to human communication, in particular speech recognition and machine learning approaches to modeling multi-modal data. He has over 150 publications in these areas.

Philip Resnik is an Associate Professor at the University of Maryland, College Park, with joint appointments in the Department of Linguistics and the Institute for Advanced Computer Studies. He completed his PhD in Computer and Information Science at the University of Pennsylvania in 1993. His research focuses on the integration of linguistic knowledge with data-driven statistical modeling, and he has done work in a variety of areas, including computational psycholinguistics, word-sense disambiguation, cross-language information retrieval, machine translation, and sentiment analysis.

Giorgio Satta received a PhD in Computer Science in 1990 from the University of Padua, Italy. He is currently a Full Professor at the Department of Information Engineering, University of Padua. His main research interests are in computational linguistics, mathematics of language and formal language theory.

For the years 2009–10 he is serving as Chair of the European Chapter of the Association for Computational Linguistics (EACL). He has joined the standing committee of the Formal Grammar conference (FG) and the editorial boards of the journals *Computational Linguistics*, *Grammars* and *Research on Language and Computation*. He has also served as Program Committee Chair for the Annual Meeting of the Association for Computational Linguistics (ACL) and for the International Workshop on Parsing Technologies (IWPT).

Helmut Schmid works as a Senior Scientist at the Institute for Natural Language Processing in Stuttgart with a focus on statistical methods for NLP. He developed a range of tools for tokenization, POS tagging, parsing, computational morphology, and statistical clustering, and he frequently used decision trees in his work.

Antal van den Bosch (MA, Tilburg University, The Netherlands, 1992; PhD, Computer Science, Universiteit Maastricht, The Netherlands, 1997) held Research Assistant positions at the experimental psychology labs of Tilburg University and the Université Libre de Bruxelles (Belgium) in 1993 and 1994. After his PhD project at the Universiteit Maastricht (1994–7), he returned to Tilburg University in 1997 as a postdoc researcher. In 1999 he was awarded a Royal Dutch Academy of Arts and Sciences fellowship, followed in 2001 and 2006 by two consecutively awarded Innovational Research funds of the Netherlands Organisation for Scientific Research. Tilburg University appointed him as Assistant Professor (2001), Associate Professor (2006), and Full Professor in Computational Linguistics and AI (2008). He is also a Guest Professor at the University of Antwerp (Belgium). He currently supervises five PhD students, and has graduated seven PhD students as co-supervisor. His research interests include memory-based natural language processing and modeling, machine translation, and proofing tools.

Prof. **Andy Way** obtained his BSc (Hons) in 1986, MSc in 1989, and PhD in 2001 from the University of Essex, Colchester, UK. From 1988 to 1991 he worked at the University of Essex, UK, on the Eurotra Machine Translation project. He joined Dublin City University (DCU) as a Lecturer in 1991 and was promoted to Senior Lecturer in 2001 and Associate Professor in 2006. He was a DCU Senior Albert College Fellow from 2002 to 2003, and has been an IBM Centers for Advanced Studies Scientist since 2003, and a Science Foundation Ireland Fellow since 2005. He has published over 160 peer-reviewed papers. He has been awarded grants totaling over €6.15 million since 2000, and over €6.6 million in total. He is the Centre for Next Generation Localisation co-ordinator for Integrated Language Technologies (ILT). He currently supervises eight students on PhD programs of study, all of whom are externally funded, and has in addition graduated 10 PhD and 11 MSc students. He is currently the Editor of the journal *Machine Translation*, President of the European Association for Machine Translation, and President-Elect of the International Association for Machine Translation.

Nick Webb is a Senior Research Scientist in the Institute for Informatics, Logics and Security Studies, at the University at Albany, SUNY, USA. Previously he was a Research Fellow in the Natural Language Processing Group at the University of Sheffield, UK, and a Research Officer at the University of Essex, UK, where he obtained a BSc in Computer Science (with a focus on Artificial Intelligence) and an MSc (in Computational Linguistics). His PhD from Sheffield concerns the analysis of dialogue corpora to build computational models of dialogue-act classification, and his research interests concern intelligent information access, including interactive question answering and dialogue systems.

Bonnie Webber was a Researcher at Bolt Beranek and Newman while working on the PhD she received from Harvard University in 1978. She then taught in the Department of Computer and Information Science at the University of Pennsylvania for 20 years before joining the School of Informatics at the University of Edinburgh. Known for research on discourse and on question answering, she is a Past President of the Association for Computational Linguistics, co-developer (with Aravind Joshi, Rashmi Prasad, Alan Lee, and Eleni Miltsakaki) of the Penn Discourse TreeBank, and co-editor (with Annie Zaenen and Martha Palmer) of the new electronic journal, *Linguistic Issues in Language Technology*.

Shuly Wintner is a Senior Lecturer at the Department of Computer Science, University of Haifa, Israel. His research spans various areas in computational linguistics, including formal grammars, morphology, syntax, development of language resources and machine translation, with a focus on Semitic languages. He has published over 60 scientific papers in computational linguistics. Dr Wintner is the Editor-in-Chief of the journal *Research in Language and Computation*.

Nianwen Xue is an Assistant Professor of Languages & Linguistics and Computer Science at Brandeis University. His research interests include syntactic and semantic parsing, machine translation, temporal representation and inference, Chinese-language processing, and linguistic annotation (Chinese Treebank, Chinese Proposition Bank, OntoNotes). He serves on the ACL SIGANN committee and co-organized the Linguistic Annotation Workshops (LAW II and LAW III) and the 2009 CoNLL Shared Task on Syntactic and Semantic Dependencies in Multiple Languages. He got his PhD in linguistics from the University of Delaware.

Preface

We started work on this handbook three years ago and, while bringing it to fruition has involved a great deal of work, we have enjoyed the process. We are grateful to our colleagues who have contributed chapters to the volume. Its quality is due to their labor and commitment. We appreciate the considerable time and effort that they have invested in making this venture a success. It has been a pleasure working with them.

We owe a debt of gratitude to our editors at Wiley-Blackwell, Danielle Descoteaux and Julia Kirk, for their unstinting support and encouragement throughout this project. We wish that all scientific-publishing projects were blessed with publishers of their professionalism and good nature.

Finally, we must thank our families for enduring the long period of time that we have been engaged in working on this volume. Their patience and good will has been a necessary ingredient for its completion.

The best part of compiling this handbook has been the opportunity that it has given each of us to observe in detail and in perspective the wonderful burst of creativity that has taken hold of our field in recent years.

Alexander Clark, Chris Fox, and Shalom Lappin
London and Wivenhoe
September 2009

Introduction

The field of computational linguistics (CL), together with its engineering domain of natural language processing (NLP), has exploded in recent years. It has developed rapidly from a relatively obscure adjunct of both AI and formal linguistics into a thriving scientific discipline. It has also become an important area of industrial development. The focus of research in CL and NLP has shifted over the past three decades from the study of small prototypes and theoretical models to robust learning and processing systems applied to large corpora. This handbook is intended to provide an introduction to the main areas of CL and NLP, and an overview of current work in these areas. It is designed as a reference and source text for graduate students and researchers from computer science, linguistics, psychology, philosophy, and mathematics who are interested in this area.

The volume is divided into four main parts. Part I contains chapters on the formal foundations of the discipline. Part II introduces the current methods that are employed in CL and NLP, and it divides into three subsections. The first section describes several influential approaches to Machine Learning (ML) and their application to NLP tasks. The second section presents work in the annotation of corpora. The last section addresses the problem of evaluating the performance of NLP systems. Part III of the handbook takes up the use of CL and NLP procedures within particular linguistic domains. Finally, Part IV discusses several leading engineering tasks to which these procedures are applied.

In Chapter 1 Shuly Wintner gives a detailed introductory account of the main concepts of formal language theory. This subdiscipline is one of the primary formal pillars of computational linguistics, and its results continue to shape theoretical and applied work. Wintner offers a remarkably clear guide through the classical language classes of the Chomsky hierarchy, and he exhibits the relations between these classes and the automata or grammars that generate (recognize) their members.

While formal language theory identifies classes of languages and their decidability (or lack of such), complexity theory studies the computational resources

The Handbook of Computational Linguistics and Natural Language Processing, First Edition.
Edited by Alexander Clark, Chris Fox and Shalom Lappin.

in time and space required to compute the elements of these classes. Ian Pratt-Hartmann introduces this central area of computer science in Chapter 2, and he takes up its significance for CL and NLP. He describes a series of important complexity results for several prominent language classes and NLP tasks. He also extends the treatment of complexity in CL/NLP from classical problems, like syntactic parsing, to the relatively unexplored area of computing sentence meaning and logical relations among sentences.

Statistical modeling has become one of the primary tools in CL and NLP for representing natural language properties and processes. In Chapter 3 Ciprian Chelba offers a clear and concise account of the basic concepts involved in the construction of statistical language models. He reviews probabilistic n-gram models and their relation to Markov systems. He defines and clarifies the notions of perplexity and entropy in terms of which the predictive power of a language model can be measured. Chelba compares n-gram models with structured language models generated by probabilistic context-free grammars, and he discusses their applications in several NLP tasks.

Part I concludes with Mark-Jan Nederhof and Giorgio Satta's discussion of the formal foundations of parsing in Chapter 4. They illustrate the problem of recognizing and representing syntactic structure with an examination of (non-lexicalized and lexicalized) context-free grammars (CFGs) and tabular (chart) parsing. They present several CFG parsing algorithms, and they consider probabilistic CFG parsing. They then extend their study to dependency grammar parsers and tree adjoining grammars (TAGs). The latter are mildly context sensitive, and so more formally powerful than CFGs. This chapter provides a solid introduction to the central theoretical concepts and results of a core CL domain.

Robert Malouf opens the first section of Part II with an examination of maximum entropy models in Chapter 5. These constitute an influential machine learning technique that involves minimizing the bias in a probability model for a set of events to the minimal set of constraints required to accommodate the data. Malouf gives a rigorous account of the formal properties of MaxEnt model selection, and exhibits its role in describing natural languages. He compares MaxEnt to support vector machines (SVMs), another ML technique, and he looks at its usefulness in part of speech tagging, parsing, and machine translation.

In Chapter 6 Walter Daelemans and Antal van den Bosch give a detailed overview of memory-based learning (MBL), an ML classification model that is widely used in NLP. MBL invokes a similarity measure to evaluate the distance between the feature vectors of stored training data and those of new events or entities in order to construct classification classes. It is a highly versatile and efficient learning framework that constitutes an alternative to statistical language modeling methods. Daelemans and van den Bosch consider modified and extended versions of MBL, and they review its application to a wide variety of NLP tasks. These include phonological and morphological analysis, part of speech tagging, shallow parsing, word disambiguation, phrasal chunking, named entity recognition, generation, machine translation, and dialogue-act recognition.

Helmut Schmid surveys decision trees in Chapter 7. These provide an efficient procedure for classifying data into descending binary branching subclasses, and they can be quickly induced from large data samples. Schmid points out that simple decision trees often exhibit instability because of their sensitivity to small changes in feature patterns of the data. He considers several modifications of decision trees that overcome this limitation, specifically bagging, boosting, and random forests. These methods combine sets of trees induced for a data set to achieve a more robust classifier. Schmid illustrates the application of decision trees to natural language tasks with discussions of grapheme conversion to phonemes, and POS tagging.

Alex Clark and Shalom Lappin characterize grammar induction as a problem in unsupervised learning in Chapter 8. They compare supervised and unsupervised grammar inference, from both engineering and cognitive perspectives. They consider the costs and benefits of both learning approaches as a way of solving NLP tasks. They conclude that, while supervised systems are currently more accurate than unsupervised ones, the latter will become increasingly influential because of the enormous investment in resources required to annotate corpora for training supervised classifiers. By contrast, large quantities of raw text are readily available online for unsupervised learning. In modeling human language acquisition, unsupervised grammar induction is a more appropriate framework, given that the primary linguistic data available to children is not annotated with sample classifications to be learned. Clark and Lappin discuss recent work in unsupervised POS tagging and grammar inference, and they observe that the most successful of these procedures are beginning to approach the performance levels achieved by state-of-the-art supervised taggers and parsers.

Neural networks are one of the earliest and most influential paradigms of machine learning. James B. Henderson concludes the first section of Part II with an overview in Chapter 9 of neural networks and their application to NLP problems. He considers multi-layered perceptrons (MLPs), which contain hidden units between their inputs and outputs, and recurrent MLPs, which have cyclic links to hidden units. These cyclic links allow the system to process unbounded sequences by storing copies of hidden unit states and feeding them back as input to units when they are processing successive positions in the sequence. In effect, they provide the system with a memory for processing sequences of inputs. Henderson shows how a neural network can be used to calculate probability values for its outputs. He also illustrates the application of neural networks to the tasks of generating statistical language models for a set of data, learning different sorts of syntactic parsing, and identifying semantic roles. He compares them to other machine learning methods and indicates certain equivalence relations that hold between neural networks and these methods.

In the second section (Chapter 10), Martha Palmer and Nianwen Xue address the central issue of corpus annotation. They compare alternative systems for marking corpora and propose clear criteria for achieving adequate results across distinct annotation tasks. They look at a number of important types of linguistic information that annotation encodes including, *inter alia*, POS tagging, deep and shallow syntactic parsing, coreference and anaphora relations, lexical meanings, semantic roles, temporal connections among propositions, logical entailments among propositions, and discourse structure. Palmer and Xue discuss the problems of securing reasonable levels of annotator agreement. They show how a sound and well-motivated annotation scheme is crucial for the success of supervised machine learning procedures in NLP, as well as for the rigorous evaluation of their performance.

Philip Resnik and Jimmy Lin conclude Part II with a discussion in the last section (Chapter 11) of methods for evaluating NLP systems. They consider both intrinsic evaluation of a procedure's performance for a specified task, and external assessment of its contribution to the quality of a larger engineering system in which it is a component. They present several ways to formulate precise quantitative metrics for grading the output of an NLP device, and they review testing sequences through which these metrics can be applied. They illustrate the issues of evaluation by considering in some detail what is involved in assessing systems for word-sense disambiguation and for question answering. This chapter extends and develops some of the concerns raised in the previous chapter on annotation. It also factors out and addresses evaluation problems that emerged in earlier chapters on the application of machine learning methods to NLP tasks.

Part III opens with Steve Renals and Thomas Hain's comprehensive account in chapter 12 of current work in automatic speech recognition (ASR). They observe that ASR plays a central role in NLP applications involving spoken language, including speech-to-speech translation, dictation, and spoken dialogue systems. Renals and Hain focus on the general task of transcribing natural conversational speech to text, and present the problem in terms of a statistical framework in which the problem of the speech recogniser is to find the most likely word sequence given the observed acoustics. The focus of the chapter is acoustic modeling based on hidden Markov models (HMMs) and Gaussian mixture models. In the first part of the chapter they develop the basic acoustic modeling framework that underlies current speech recognition systems, including refinements to include discriminative training and the adaptation to particular speakers using only small amounts of data. These components are drawn together in the description of a state-of-the-art system for the automatic transcription of multiparty meetings. The final part of the chapter discusses approaches that enable robustness for noisier or less constrained acoustic environments, the incorporation of multiple sources of knowledge, the development of sequence models that are richer than HMMs, and issues that arise when developing large-scale ASR systems.

In Chapter 13 Stephen Clark discusses statistical parsing as the probabilistic syntactic analysis of sentences in a corpus, through supervised learning. He traces the development of this area from generative parsing models to discriminative frameworks. Clark studies Collins' lexicalized probabilistic context-free grammars (PCFGs) as a particularly successful instance of these models. He examines the parsing algorithms, procedures for parse ranking, and methods for parse optimization that are commonly used in generative parse models like PCFG. Discriminative parsing does not model sentences, but provides a way of modeling