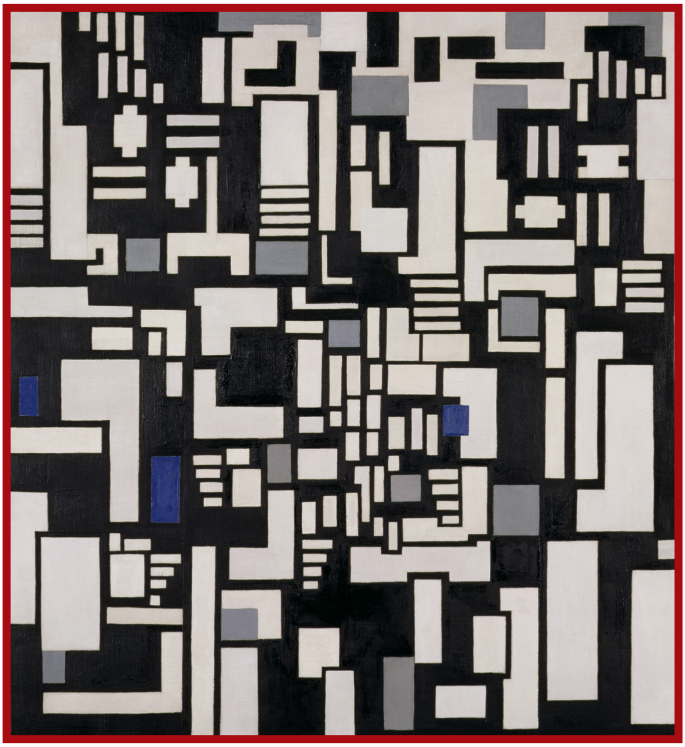
The Handbook of Computational Linguistics and Natural Language Processing



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

© 2013 Blackwell Publishing Ltd except for editorial material and organization
© 2013 Alexander Clark, Chris Fox, and Shalom Lappin

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.

The right of Alexander Clark, Chris Fox, and Shalom Lappin to be identified as the authors of the editorial material in this work has been asserted in accordance with the UK Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

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 $\alpha_t(q_j) = p(\mathbf{x}_1, \dots, \mathbf{x}_t, q_t = q_j \mid \boldsymbol{\lambda})$. 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