

Cognitive Technologies

Managing Editors: D. M. Gabbay J. Siekmann

Editorial Board: A. Bundy J. G. Carbonell
M. Pinkal H. Uszkoreit M. Veloso W. Wahlster
M. J. Wooldridge

Advisory Board:

Luigia Carlucci Aiello
Franz Baader
Wolfgang Bibel
Leonard Bolc
Craig Boutilier
Ron Brachman
Bruce G. Buchanan
Anthony Cohn
Artur d'Avila Garcez
Luis Fariñas del Cerro
Koichi Furukawa
Georg Gottlob
Patrick J. Hayes
James A. Hendler
Anthony Jameson
Nick Jennings
Aravind K. Joshi
Hans Kamp
Martin Kay
Hiroaki Kitano
Robert Kowalski
Sarit Kraus
Maurizio Lenzerini
Hector Levesque
John Lloyd

Alan Mackworth
Mark Maybury
Tom Mitchell
Johanna D. Moore
Stephen H. Muggleton
Bernhard Nebel
Sharon Oviatt
Luis Pereira
Lu Ruqian
Stuart Russell
Erik Sandewall
Luc Steels
Oliviero Stock
Peter Stone
Gerhard Strube
Katia Sycara
Milind Tambe
Hidehiko Tanaka
Sebastian Thrun
Junichi Tsujii
Kurt VanLehn
Andrei Voronkov
Toby Walsh
Bonnie Webber

For further volumes:
<http://www.springer.com/series/5216>

Laura Kallmeyer

Parsing Beyond Context-Free Grammars

 Springer

PD Dr. Laura Kallmeyer
SFB 833
Universität Tübingen
Nauklerstr. 35
72074 Tübingen
Germany
lk@sfs.uni-tuebingen.de

Managing Editors

Prof. Dr. Dov M. Gabbay
Augustus De Morgan Professor of Logic
King's College London
Dept. Computer Science
London WC2R 2LS
United Kingdom

Prof. Dr. Jörg Siekmann
Forschungsbereich Deduktions- und
Multiagentensysteme, DFKI
Stuhlsatzenweg 3, Geb. 43
66123 Saarbrücken, Germany

Cognitive Technologies ISSN 1611-2482
ISBN 978-3-642-14845-3 e-ISBN 978-3-642-14846-0
DOI 10.1007/978-3-642-14846-0
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010933793

ACM Computing Classification (1998): I.2.7, F.4, J.5

© Springer-Verlag Berlin Heidelberg 2010

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: KünkelLopka GmbH, Heidelberg

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Given that context-free grammars cannot adequately describe natural languages, grammar formalisms beyond CFG that are still computationally tractable are of central interest for computational linguists. However, despite the considerable interest in such formalisms and in their various parsing algorithms, a coherent textbook that allows access to the large body of knowledge on polynomial-time parsing beyond context-free grammars has not been available so far. Textbooks on parsing covered mainly context-free grammars while mentioning more powerful formalisms only very briefly.

This want of a detailed presentation of grammar formalisms and parsing beyond CFG is addressed with this book. The book provides an extensive overview of the formal language landscape between CFG and PTIME. It moves from Tree Adjoining Grammars to Multiple Context-Free Grammars and then to Range Concatenation Grammars while explaining available parsing techniques for these formalisms. The text is enriched with many illustrations and examples coming with the different formalisms and algorithms. This makes the book accessible to anybody familiar with basic notions of CFG parsing. It is useful both for researchers and students in computational linguistics and in formal language theory.

Tübingen,
June 2010

Laura Kallmeyer

Acknowledgments

First of all and most importantly, I want to thank my colleague Wolfgang Maier. We taught two courses at the University of Tübingen and one course at the European Summer School in Logic, Language and Information (ESSLLI) in 2008 in Hamburg, all of them covering the topic of parsing beyond context-free grammars. The idea to write this textbook arose out of these courses and much material from the course slides was reused when writing the book. The course preparations and the related discussions of the subject were crucial for achieving a good understanding of the topic and being able to cover it in a textbook. Also, when writing the book, I frequently discussed its content and structure with Wolfgang. Therefore one can say that without Wolfgang's help the book would not look as it does and, furthermore, it would very probably not exist at all.

The suggestion to write a textbook on parsing beyond context-free grammars came from Carl Vogel who participated in our ESSLLI course on this topic. I am grateful for this suggestion; it made me for the first time seriously consider the idea of covering the course material in a book.

While writing this book, I was financed by an Emmy Noether Grant from the German Research Foundation DFG (Deutsche Forschungsgemeinschaft).

Contents

1	Introduction	1
1.1	Formal Grammars and Natural Languages	1
1.2	Parsing Beyond CFGs	5
1.3	What This Book Is Not About	7
1.4	Overview of the Book	8
1.4.1	Grammar Formalisms for Natural Languages	8
1.4.2	Parsing: Preliminaries	8
1.4.3	Tree Adjoining Grammars	8
1.4.4	MCFG and LCFRS	9
1.4.5	Range Concatenation Grammars	9
1.4.6	Automata	10
1.5	Some Basic Definitions	10
1.5.1	Languages	10
1.5.2	Context-Free Grammars	11
1.5.3	Automata	12
1.5.4	Trees	14
2	Grammar Formalisms for Natural Languages	17
2.1	Context-Free Grammars and Natural Languages	17
2.1.1	The Generative Capacity of CFGs	17
2.1.2	CFGs and Lexicalization	20
2.1.3	Mild Context-Sensitivity	23
2.2	Grammar Formalisms Beyond CFG	26
2.2.1	Tree Adjoining Grammars	26
2.2.2	Linear Indexed Grammars	31
2.2.3	Linear Context-Free Rewriting Systems	33
2.2.4	Multicomponent Tree Adjoining Grammars	33
2.2.5	Multiple Context-Free Grammars	36
2.2.6	Range Concatenation Grammars	36
2.3	Summary	38

3	Parsing: Preliminaries	41
3.1	Parsing as Deduction	41
3.1.1	Motivation	41
3.1.2	Items	42
3.1.3	Deduction Rules	44
3.2	Implementation Issues	44
3.2.1	Dynamic Programming	44
3.2.2	Chart Parsing and Tabulation	46
3.2.3	Hypergraphs	47
3.3	Properties of Parsing Algorithms	48
3.3.1	Soundness and Completeness	48
3.3.2	Complexity	49
3.3.3	Valid Prefix Property	51
3.4	Summary	51
4	Tree Adjoining Grammars	53
4.1	Introduction to Tree Adjoining Grammars	53
4.1.1	Definition of TAG	53
4.1.2	Formal Properties	58
4.1.3	Linguistic Principles for TAG	63
4.1.4	Extended Domain of Locality and Factoring of Recursion	65
4.1.5	Constituency and Dependencies	68
4.2	Equivalent Formalisms	70
4.2.1	Tree-Local MCTAG	70
4.2.2	Linear Indexed Grammars	72
4.2.3	Combinatory Categorical Grammars	72
4.3	Summary	74
5	Parsing Tree Adjoining Grammars	77
5.1	A CYK Parser for TAG	77
5.1.1	The Recognizer	77
5.1.2	Complexity	82
5.2	An Earley Parser for TAG	82
5.2.1	Introduction	82
5.2.2	Items	83
5.2.3	Inference Rules	85
5.2.4	Extending the Algorithm to Substitution	88
5.2.5	The Parser	91
5.2.6	Properties of the Algorithm	92
5.2.7	Prefix Valid Earley Parsing	93
5.3	An LR Parser for TAG	96
5.3.1	Introduction	96
5.3.2	Construction of the Automaton	99
5.3.3	The Recognizer	101
5.3.4	Valid Prefix Property	107

5.4	Summary	107
6	Multiple Context-Free Grammars and Linear Context-Free Rewriting Systems	109
6.1	Introduction to MCFG, LCFRS and Simple RCG	109
6.1.1	MCFG and LCFRS	110
6.1.2	Formal Properties	117
6.1.3	Applications	122
6.2	Equivalent Formalisms	125
6.2.1	Set-Local Multicomponent TAG	125
6.2.2	Minimalist Grammars	126
6.2.3	Finite-Copying LFG	126
6.3	Summary	128
7	Parsing MCFG, LCFRS and Simple RCG	131
7.1	CYK Parsing of MCFG	131
7.1.1	The Basic Algorithm	131
7.1.2	The Naïve Algorithm	134
7.1.3	The Active Algorithm	136
7.1.4	The Incremental Algorithm	139
7.1.5	Prediction Strategies	141
7.2	Simplifying Simple RCGs	142
7.2.1	Eliminating Useless Rules	142
7.2.2	Eliminating ε -Rules	143
7.2.3	Ordered Simple RCG	145
7.2.4	Binarization of the Rules	147
7.3	An Incremental Earley Parser for Simple RCG	149
7.3.1	The Algorithm	149
7.3.2	Filters	154
7.4	Summary	155
8	Range Concatenation Grammars	157
8.1	Introduction to Range Concatenation Grammars	157
8.1.1	Definition of RCG	157
8.1.2	Applications	164
8.2	Relations to Other Formalisms	167
8.2.1	Literal Movement Grammars	167
8.2.2	CFG, TAG and MCFG	170
8.3	Summary	173
9	Parsing Range Concatenation Grammars	177
9.1	Basic RCG Parsing	177
9.1.1	CYK Parsing with Passive Items	178
9.1.2	Non-directional Top-Down Parsing	179
9.1.3	Directional Top-Down Parsing	180

9.1.4	Optimizations	183
9.2	Parsing with Constraint Propagation	184
9.2.1	Range Constraints	185
9.2.2	CYK Parsing with Active Items	186
9.2.3	Earley Parsing	188
9.3	Summary	190
10	Automata	193
10.1	Embedded Push-Down Automata	193
10.1.1	Definition of EPDA	193
10.1.2	EPDA and TAG	197
10.1.3	Bottom-Up Embedded Push-Down Automata	197
10.1.4	k -Order EPDA	199
10.2	Two-Stack Automata	200
10.2.1	General Definition	200
10.2.2	Strongly-Driven Two-Stack Automata	202
10.3	Thread Automata	204
10.3.1	Idea	204
10.3.2	General Definition of TA	206
10.3.3	Constructing a TA for a TAG	208
10.3.4	Constructing a TA for an Ordered SRCG	209
10.4	Summary	213
	Appendix A: Hierarchy of Grammar Formalisms	215
	Appendix B: List of Acronyms	217
	Solutions	219
	References	235
	Index	245