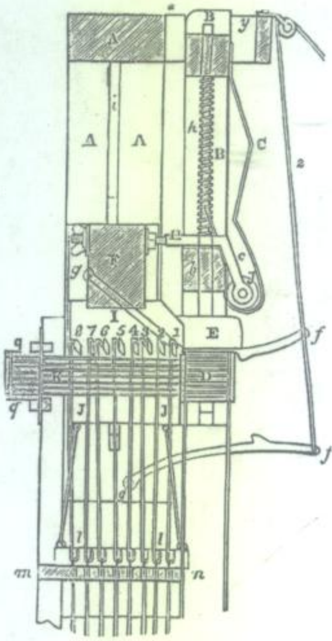
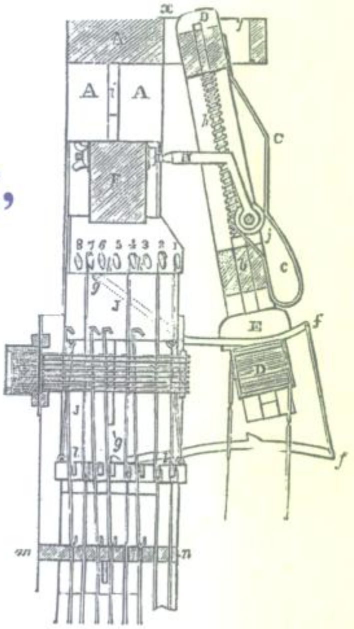


SPEECH and LANGUAGE PROCESSING

An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition



DANIEL JURAFSKY & JAMES H. MARTIN

Speech and Language Processing

An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

Daniel Jurafsky and James H. Martin

University of Colorado, Boulder

Contributing writers:
Andrew Kehler, Keith Vander Linden, and Nigel Ward

Prentice Hall

Prentice Hall
Upper Saddle River, New Jersey 07458

Library of Congress Cataloging-in-Publication Data

Jurafsky, Daniel S. (Daniel Saul)
Speech and Langauge Processing / Daniel Jurafsky, James H. Martin.
p. cm.

Includes bibliographical references and index.
ISBN 0-13-095069-6

Editor-in-Chief: *Marcia Horton*
Publisher: *Alan Apt*
Editorial/production supervision: *Scott Disanno*
Editorial assistant: *Toni Holm*
Executive managing editor: *Vince O'Brien*
Cover design director: *Heather Scott*
Cover design execution: *John Christiana*
Manufacturing manager: *Trudy Pisciotti*
Manufacturing buyer: *Pat Brown*
Assistant vice-president of production and manufacturing: *David W. Riccardi*

Cover design: *Daniel Jurafsky, James H. Martin, and Linda Martin.* The front cover drawing is the action for the Jacquard Loom (Usher, 1954). The back cover drawing is Alexander Graham Bell's Gallows telephone (Rhodes, 1929).

This book was set in Times-Roman, TIPA (IPA), and PMC (Chinese) by the authors using LaTeX2e.

Pearson Higher Education
Upper Saddle River, New Jersey 07458

The author and publisher of this book have used their best efforts in preparing this book. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness. The author and publisher shall not be liable in any event for incidental or consequential damages in connection with, or arising out of, the furnishing, performance, or use of these programs.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1
ISBN 0-13-095069-6

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada, Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Prentice-Hall Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

Speech and Language Processing

"This book is an absolute necessity for instructors at all levels, as well as an indispensible reference for researchers. Introducing NLP, computational linguistics, and speech recognition comprehensively in a single book is an ambitious enterprise. The authors have managed it admirably, paying careful attention to traditional foundations, relating recent developments and trends to those foundations, and tying it all together with insight and humor. Remarkable."

– Philip Resnik, University of Maryland

"... ideal for ... linguists who want to learn more about computational modeling and techniques in language processing; computer scientists building language applications who want to learn more about the linguistic underpinnings of the field; speech technologists who want to learn more about language understanding, semantics and discourse; and all those wanting to learn more about speech processing. For instructors ... this book is a dream. It covers virtually every aspect of NLP... What's truly astounding is that the book covers such a broad range of topics, while giving the reader the depth to understand and make use of the concepts, algorithms and techniques that are presented... ideal as a course textbook for advanced undergraduates, as well as graduate students and researchers in the field.

– Johanna Moore, University of Edinburgh

"*Speech and Language Processing* is a comprehensive, reader-friendly, and up-to-date guide to computational linguistics, covering both statistical and symbolic methods and their application. It will appeal both to senior undergraduate students, who will find it neither too technical nor too simplistic, and to researchers, who will find it to be a helpful guide to the newly established techniques of a rapidly growing research field."

– Graeme Hirst, University of Toronto

"The field of human language processing encompasses a diverse array of disciplines, and as such is an incredibly challenging field to master. This book does a wonderful job of bringing together this vast body of knowledge in a form that is both accessible and comprehensive. Its encyclopedic coverage makes it a must have for people already in the field, while the clear presentation style and many examples make it an ideal textbook."

– Eric Brill, Microsoft Research

This is quite simply the most complete introduction to natural language and speech technology ever written. Virtually every topic in the field is covered, in a prose style that is both clear and engaging. The discussion is linguistically informed, and strikes a nice balance between theoretical computational models, and practical applications. It is an extremely impressive achievement.

– Richard Sproat, AT&T Labs – Research"

PRENTICE HALL SERIES
IN ARTIFICIAL INTELLIGENCE
Stuart Russell and Peter Norvig, Editors

GRAHAM	ANSI *Common Lisp*
RUSSELL & NORVIG	*Artificial Intelligence: A Modern Approach*
JURAFSKY & MARTIN	*Speech and Language Processing*

For my parents, Ruth and Al Jurafsky — D.J.

For Linda — J.M.

Foreword

Linguistics has a hundred-year history as a scientific discipline, and computational linguistics has a forty-year history as a part of computer science. But it is only in the last five years that language understanding has emerged as an industry reaching millions of people, with information retrieval and machine translation available on the internet, and speech recognition becoming popular on desktop computers. This industry has been enabled by theoretical advances in the representation and processing of language information.

Speech and Language Processing is the first book to thoroughly cover language technology, at all levels and with all modern technologies. It combines deep linguistic analysis with robust statistical methods. From the point of view of levels, the book starts with the word and its components, moving up to the way words fit together (or syntax), to the meaning (or semantics) of words, phrases and sentences, and concluding with issues of coherent texts, dialog, and translation. From the point of view of technologies, the book covers regular expressions, information retrieval, context free grammars, unification, first-order predicate calculus, hidden Markov and other probabilistic models, rhetorical structure theory, and others. Previously you would need two or three books to get this kind of coverage. *Speech and Language Processing* covers the full range in one book, but more importantly, it relates the technologies to each other, giving the reader a sense of how each one is best used, and how they can be used together. It does all this with an engaging style that keeps the reader's interest and motivates the technical details in a way that is thorough but not dry. Whether you're interested in the field from the scientific or the industrial point of view, this book serves as an ideal introduction, reference, and guide to future study of this fascinating field.

Peter Norvig & Stuart Russell, Editors
Prentice Hall Series in Artificial Intelligence

Preface

This is an exciting time to be working in speech and language processing. Historically distinct fields (natural language processing, speech recognition, computational linguistics, computational psycholinguistics) have begun to merge. The commercial availability of speech recognition and the need for Web-based language techniques have provided an important impetus for development of real systems. The availability of very large on-line corpora has enabled statistical models of language at every level, from phonetics to discourse. We have tried to draw on this emerging state of the art in the design of this pedagogical and reference work:

1. *Coverage*
   In attempting to describe a unified vision of speech and language processing, we cover areas that traditionally are taught in different courses in different departments: speech recognition in electrical engineering; parsing, semantic interpretation, and pragmatics in natural language processing courses in computer science departments; and computational morphology and phonology in computational linguistics courses in linguistics departments. The book introduces the fundamental algorithms of each of these fields, whether originally proposed for spoken or written language, whether logical or statistical in origin, and attempts to tie together the descriptions of algorithms from different domains. We have also included coverage of applications like spelling-checking and information retrieval and extraction as well as areas like cognitive modeling. A potential problem with this broad-coverage approach is that it required us to include introductory material for each field; thus linguists may want to skip our description of articulatory phonetics, computer scientists may want to skip such sections as regular expressions, and electrical engineers skip the sections on signal processing. Of course, even in a book this long, we didn't have room for everything. Thus this book should not be considered a substitute for important relevant courses in linguistics, automata and formal language theory, or, especially, statistics and information theory.
2. *Emphasis on practical applications*
   It is important to show how language-related algorithms and techniques (from HMMs to unification, from the lambda calculus to transformation-based learning) can be applied to important real-world problems: spelling checking, text document search, speech recogni-

tion, Web-page processing, part-of-speech tagging, machine translation, and spoken-language dialogue agents. We have attempted to do this by integrating the description of language processing applications into each chapter. The advantage of this approach is that as the relevant linguistic knowledge is introduced, the student has the background to understand and model a particular domain.

3. *Emphasis on scientific evaluation*
   The recent prevalence of statistical algorithms in language processing and the growth of organized evaluations of speech and language processing systems has led to a new emphasis on evaluation. We have, therefore, tried to accompany most of our problem domains with a **Methodology Box** describing how systems are evaluated (e.g., including such concepts as training and test sets, cross-validation, and information-theoretic evaluation metrics like perplexity).
4. *Description of widely available language processing resources*
   Modern speech and language processing is heavily based on common resources: raw speech and text corpora, annotated corpora and treebanks, standard tagsets for labeling pronunciation, part-of-speech, parses, word-sense, and dialogue-level phenomena. We have tried to introduce many of these important resources throughout the book (e.g., the Brown, Switchboard, callhome, ATIS, TREC, MUC, and BNC corpora) and provide complete listings of many useful tagsets and coding schemes (such as the Penn Treebank, CLAWS C5 and C7, and the ARPAbet) but some inevitably got left out. Furthermore, rather than include references to URLs for many resources directly in the textbook, we have placed them on the book's Web site, where they can more readily updated.

The book is primarily intended for use in a graduate or advanced undergraduate course or sequence. Because of its comprehensive coverage and the large number of algorithms, the book is also useful as a reference for students and professionals in any of the areas of speech and language processing.

Overview of the Book

The book is divided into four parts in addition to an introduction and end matter. Part I, "Words", introduces concepts related to the processing of words: phonetics, phonology, morphology, and algorithms used to process them: finite automata, finite transducers, weighted transducers, N-grams, and Hidden Markov Models. Part II, "Syntax", introduces parts-of-speech and phrase structure grammars for English and gives essential algorithms for processing word classes and structured relationships among words: part-of-speech taggers based on HMMs and transformation-based learning, the CYK and Earley algorithms for parsing, unification and typed feature structures, lexicalized and probabilistic parsing, and analytical tools like the Chomsky hierarchy and the pumping lemma. Part III, "Semantics", introduces first order predicate calculus and other ways of representing meaning, several approaches to compositional semantic analysis, along with applications to information retrieval, information extraction, speech understanding, and machine translation. Part IV, "Pragmatics", covers reference resolution and discourse structure and coherence, spoken dialogue phenomena like dialogue and speech act modeling, dialogue structure and coherence, and dialogue managers, as well as a comprehensive treatment of natural language generation and of machine translation.

Using this Book

The book provides enough material to be used for a full-year sequence in speech and language processing. It is also designed so that it can be used for a number of different useful one-term courses:

NLP	NLP	Speech + NLP	Comp. Linguistics
1 quarter	1 semester	1 semester	1 quarter
1. Intro	1. Intro	1. Intro	1. Intro
2. Regex, FSA	2. Regex, FSA	2. Regex, FSA	2. Regex, FSA
8. POS tagging	3. Morph., FST	3. Morph., FST	3. Morph., FST
9. CFGs	6. N-grams	4. Comp. Phonol.	4. Comp. Phonol.
10. Parsing	8. POS tagging	5. Prob. Pronun.	10. Parsing
11. Unification	9. CFGs	6. N-grams	11. Unification
14. Semantics	10. Parsing	7. HMMs & ASR	13. Complexity
15. Sem. Analysis	11. Unification	8. POS tagging	16. Lex. Semantics
18. Discourse	12. Prob. Parsing	9. CFGs	18. Discourse
20. Generation	14. Semantics	10. Parsing	19. Dialogue
	15. Sem. Analysis	12. Prob. Parsing	
	16. Lex. Semantics	14. Semantics	
	17. WSD and IR	15. Sem. Analysis	
	18. Discourse	19. Dialogue	
	20. Generation	21. Mach. Transl.	
	21. Mach. Transl.		

Selected chapters from the book could also be used to augment courses in Artificial Intelligence, Cognitive Science, or Information Retrieval.

Acknowledgments

The three contributing writers for the book are Andy Kehler, who wrote Chapter 18 (Discourse), Keith Vander Linden, who wrote Chapter 20 (Generation), and Nigel Ward, who wrote most of Chapter 21 (Machine Translation). Andy Kehler also wrote Section 19.4 of Chapter 19. Paul Taylor wrote most of Section 4.7 and Section 7.8.

Dan would like to thank his parents for encouraging him to do a really good job of everything he does, finish it in a timely fashion, and make time for going to the gym. He would also like to thank Nelson Morgan, for introducing him to speech recognition and teaching him to ask "but does it work?"; Jerry Feldman, for sharing his intense commitment to finding the right answers and teaching him to ask "but is it really important?"; Chuck Fillmore, his first advisor, for sharing his love for language and especially argument structure, and teaching him to always go look at the data, (and all of them for teaching by example that it's only worthwhile if it's fun); and Robert Wilensky, his dissertation advisor, for teaching him the importance of collaboration and group spirit in research. He is also grateful to the CU Lyric Theater program and the casts of *South Pacific*, *Gianni Schicchi*, *Guys and Dolls*, *Gondoliers*, *Iolanthe*, and *Oklahoma*, and to Doris and Cary, Elaine and Eric, Irene and Sam, Susan and Richard, Lisa and Mike, Mike and Fia, Erin and Chris, Eric and Beth, Pearl and Tristan, Bruce and Peggy, Ramon and Rebecca, Adele and Ali, Terry, Kevin, Becky, Temmy, Lil, Lin and Ron and David, Mike, and Jessica and Bill, and all their families for providing lots of emotional support and often a place to stay during the writing.

Jim would like to thank his parents for encouraging him and allowing him to follow what must have seemed like an odd path at the time. He would also like to thank his thesis advisor, Robert Wilensky, for giving him his start in NLP at Berkeley; Peter Norvig, for providing many positive examples along the way; Rick Alterman, for encouragement and inspiration at a critical time; and Chuck Fillmore, George Lakoff, Paul Kay, and Susanna Cumming for teaching him what little he knows about linguistics. He'd also like to thank Michael Main for covering for him while he shirked his departmental duties. Finally, he'd like to thank his wife Linda for all her support and patience through all the years it took to complete this book.

Boulder is a very rewarding place to work on speech and language processing. We'd like to thank our colleagues here for their collaborations, which have greatly influenced our research and teaching: Alan Bell, Barbara Fox, Laura Michaelis and Lise Menn in linguistics; Clayton Lewis, Gerhard

Fischer, Mike Eisenberg, Mike Mozer, Liz Jessup, and Andrzej Ehrenfeucht in computer science; Walter Kintsch, Tom Landauer, and Alice Healy in psychology; Ron Cole, John Hansen, and Wayne Ward in the Center for Spoken Language Understanding, and our current and former students in the computer science and linguistics departments: Marion Bond, Noah Coccaro, Michelle Gregory, Keith Herold, Michael Jones, Patrick Juola, Keith Vander Linden, Laura Mather, Taimi Metzler, Douglas Roland, and Patrick Schone.

This book has benefited from careful reading and enormously helpful comments from a number of readers and from course-testing. We are deeply indebted to colleagues who each took the time to read and give extensive comments and advice, which vastly improved large parts of the book, including Alan Bell, Bob Carpenter, Jan Daciuk, Graeme Hirst, Andy Kehler, Kemal Oflazer, Andreas Stolcke, and Nigel Ward. Our editor Alan Apt, our series editors Peter Norvig and Stuart Russell, and our production editor Scott DiSanno made many helpful suggestions on design and content. We are also indebted to many friends and colleagues who read individual sections of the book or answered our many questions for their comments and advice, including the students in our classes at the University of Colorado, Boulder, and in Dan's classes at the University of California, Berkeley, and the LSA Summer Institute at the University of Illinois at Urbana-Champaign, as well as

> Yoshi Asano, Todd M. Bailey, John Bateman, Giulia Bencini, Lois Boggess, Michael Braverman, Nancy Chang, Jennifer Chu-Carroll, Noah Coccaro, Gary Cottrell, Gary Dell, Jeff Elman, Robert Dale, Dan Fass, Bill Fisher, Eric Fosler-Lussier, James Garnett, Susan Garnsey, Dale Gerdemann, Dan Gildea, Michelle Gregory, Nizar Habash, Jeffrey Haemer, Jorge Hankamer, Keith Herold, Beth Heywood, Derrick Higgins, Erhard Hinrichs, Julia Hirschberg, Jerry Hobbs, Fred Jelinek, Liz Jessup, Aravind Joshi, Terry Kleeman, Jean-Pierre Koenig, Kevin Knight, Shalom Lappin, Julie Larson, Stephen Levinson, Jim Magnuson, Jim Mayfield, Lise Menn, Laura Michaelis, Corey Miller, Nelson Morgan, Christine Nakatani, Mike Neufeld, Peter Norvig, Mike O'Connell, Mick O'Donnell, Rob Oberbreckling, Martha Palmer, Dragomir Radev, Terry Regier, Ehud Reiter, Phil Resnik, Klaus Ries, Ellen Riloff, Mike Rosner, Dan Roth, Patrick Schone, Liz Shriberg, Richard Sproat, Subhashini Srinivasin, Paul Taylor, Wayne Ward, Pauline Welby, Dekai Wu, and Victor Zue.

We'd also like to thank the Institute of Cognitive Science and the Departments of Computer Science and Linguistics for their support over the years. We are also very grateful to the National Science Foundation: Dan Jurafsky's time on the book was supported in part by NSF CAREER Award IIS-9733067 and Andy Kehler was supported in part by NSF Award IIS-9619126.

Daniel Jurafsky
James H. Martin
Boulder, Colorado

Summary of Contents

Contents