

# Summary of Contents

Preface .....	xxiii
1 Introduction .....	1
<b>I Words</b>	
2 Regular Expressions and Automata .....	17
3 Words & Transducers .....	45
4 N-grams .....	83
5 Part-of-Speech Tagging .....	123
6 Hidden Markov and Maximum Entropy Models .....	173
<b>II Speech</b>	
7 Phonetics .....	215
8 Speech Synthesis .....	249
9 Automatic Speech Recognition .....	287
10 Speech Recognition: Advanced Topics .....	337
11 Computational Phonology .....	365
<b>III Syntax</b>	
12 Formal Grammars of English .....	389
13 Parsing with Context-Free Grammars .....	431
14 Statistical Parsing .....	465
15 Features and Unification .....	495
16 Language and Complexity .....	537
<b>IV Semantics and Pragmatics</b>	
17 Representing Meaning .....	553
18 Computational Semantics .....	593
19 Lexical Semantics .....	627
20 Computational Lexical Semantics .....	653
21 Computational Discourse .....	697
<b>V Applications</b>	
22 Information Extraction .....	741
23 Question Answering and Summarization .....	783
24 Dialogue and Conversational Agents .....	829
25 Machine Translation .....	879
<b>Bibliography</b>	<b>929</b>
<b>Index</b>	<b>981</b>



# Contents

## Preface

xxiii

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Knowledge in Speech and Language Processing . . . . .	2
1.2	Ambiguity . . . . .	4
1.3	Models and Algorithms . . . . .	5
1.4	Language, Thought, and Understanding . . . . .	6
1.5	The State of the Art . . . . .	8
1.6	Some Brief History . . . . .	9
1.6.1	Foundational Insights: 1940s and 1950s . . . . .	9
1.6.2	The Two Camps: 1957–1970 . . . . .	10
1.6.3	Four Paradigms: 1970–1983 . . . . .	11
1.6.4	Empiricism and Finite State Models Redux: 1983–1993 . . . . .	12
1.6.5	The Field Comes Together: 1994–1999 . . . . .	12
1.6.6	The Rise of Machine Learning: 2000–2007 . . . . .	13
1.6.7	On Multiple Discoveries . . . . .	13
1.6.8	A Final Brief Note on Psychology . . . . .	14
1.7	Summary . . . . .	15
	Bibliographical and Historical Notes . . . . .	15

## I Words

<b>2</b>	<b>Regular Expressions and Automata</b>	<b>17</b>
2.1	Regular Expressions . . . . .	17
2.1.1	Basic Regular Expression Patterns . . . . .	18
2.1.2	Disjunction, Grouping, and Precedence . . . . .	21
2.1.3	A Simple Example . . . . .	22
2.1.4	A More Complex Example . . . . .	23
2.1.5	Advanced Operators . . . . .	24
2.1.6	Regular Expression Substitution, Memory, and ELIZA . . . . .	25
2.2	Finite-State Automata . . . . .	26
2.2.1	Using an FSA to Recognize Sheeptalk . . . . .	26
2.2.2	Formal Languages . . . . .	30
2.2.3	Another Example . . . . .	31
2.2.4	Non-Deterministic FSAs . . . . .	32
2.2.5	Using an NFSA to Accept Strings . . . . .	33
2.2.6	Recognition as Search . . . . .	35
2.2.7	Relating Deterministic and Non-Deterministic Automata . . . . .	38
2.3	Regular Languages and FSAs . . . . .	38
2.4	Summary . . . . .	40
	Bibliographical and Historical Notes . . . . .	42
	Exercises . . . . .	42

<b>3</b>	<b>Words &amp; Transducers</b>	<b>45</b>
3.1	Survey of (Mostly) English Morphology . . . . .	47
3.1.1	Inflectional Morphology . . . . .	48
3.1.2	Derivational Morphology . . . . .	50
3.1.3	Cliticization . . . . .	51
3.1.4	Non-concatenative Morphology . . . . .	52
3.1.5	Agreement . . . . .	52
3.2	Finite-State Morphological Parsing . . . . .	53
3.3	Building a Finite-State Lexicon . . . . .	54
3.4	Finite-State Transducers . . . . .	57
3.4.1	Sequential Transducers and Determinism . . . . .	59
3.5	FSTs for Morphological Parsing . . . . .	60
3.6	Transducers and Orthographic Rules . . . . .	63
3.7	Combining FST Lexicon and Rules . . . . .	65
3.8	Lexicon-Free FSTs: The Porter Stemmer . . . . .	68
3.9	Word and Sentence Tokenization . . . . .	69
3.9.1	Segmentation in Chinese . . . . .	70
3.10	Detecting and Correcting Spelling Errors . . . . .	72
3.11	Minimum Edit Distance . . . . .	74
3.12	Human Morphological Processing . . . . .	77
3.13	Summary . . . . .	79
	Bibliographical and Historical Notes . . . . .	80
	Exercises . . . . .	81
<b>4</b>	<b>N-grams</b>	<b>83</b>
4.1	Counting Words in Corpora . . . . .	84
4.2	Simple (Unsmoothed) <i>N</i> -grams . . . . .	86
4.3	Training and Test Sets . . . . .	91
4.3.1	<i>N</i> -gram Sensitivity to the Training Corpus . . . . .	92
4.3.2	Unknown Words: Open versus closed vocabulary tasks . . . . .	94
4.4	Evaluating <i>N</i> -grams: Perplexity . . . . .	95
4.5	Smoothing . . . . .	97
4.5.1	Laplace Smoothing . . . . .	98
4.5.2	Good-Turing Discounting . . . . .	101
4.5.3	Some advanced issues in Good-Turing estimation . . . . .	102
4.6	Interpolation . . . . .	103
4.7	Backoff . . . . .	105
4.7.1	Advanced: Details of computing Katz backoff $\alpha$ and $P^*$ . . . . .	106
4.8	Practical Issues: Toolkits and Data Formats . . . . .	107
4.9	Advanced Issues in Language Modeling . . . . .	109
4.9.1	Advanced Smoothing Methods: Kneser-Ney Smoothing . . . . .	109
4.9.2	Class-based <i>N</i> -grams . . . . .	111
4.9.3	Language Model Adaptation and Using the Web . . . . .	111
4.9.4	Using Longer Distance Information: A Brief Summary . . . . .	112
4.10	Advanced: Information Theory Background . . . . .	113
4.10.1	Cross-Entropy for Comparing Models . . . . .	116

4.11	Advanced: The Entropy of English and Entropy Rate Constancy . .	117
	Bibliographical and Historical Notes . . . . .	119
4.12	Summary . . . . .	120
	Exercises . . . . .	121
<b>5</b>	<b>Part-of-Speech Tagging</b>	<b>123</b>
5.1	(Mostly) English Word Classes . . . . .	124
5.2	Tagsets for English . . . . .	130
5.3	Part-of-Speech Tagging . . . . .	133
5.4	Rule-Based Part-of-Speech Tagging . . . . .	137
5.5	HMM Part-of-Speech Tagging . . . . .	139
5.5.1	Computing the most-likely tag sequence: An example . . .	142
5.5.2	Formalizing Hidden Markov Model taggers . . . . .	144
5.5.3	The Viterbi Algorithm for HMM Tagging . . . . .	145
5.5.4	Extending the HMM algorithm to trigrams . . . . .	149
5.6	Transformation-Based Tagging . . . . .	151
5.6.1	How TBL Rules Are Applied . . . . .	152
5.6.2	How TBL Rules Are Learned . . . . .	152
5.7	Evaluation and Error Analysis . . . . .	153
5.7.1	Error Analysis . . . . .	156
5.8	Advanced Issues in Part-of-Speech Tagging . . . . .	157
5.8.1	Practical Issues: Tag Indeterminacy and Tokenization . . .	157
5.8.2	Unknown Words . . . . .	158
5.8.3	Part-of-Speech Tagging for Other Languages . . . . .	160
5.8.4	Combining Taggers . . . . .	163
5.9	Advanced: The Noisy Channel Model for Spelling . . . . .	163
5.9.1	Contextual Spelling Error Correction . . . . .	167
5.10	Summary . . . . .	168
	Bibliographical and Historical Notes . . . . .	169
	Exercises . . . . .	171
<b>6</b>	<b>Hidden Markov and Maximum Entropy Models</b>	<b>173</b>
6.1	Markov Chains . . . . .	174
6.2	The Hidden Markov Model . . . . .	177
6.3	Computing Likelihood: The Forward Algorithm . . . . .	179
6.4	Decoding: The Viterbi Algorithm . . . . .	184
6.5	Training HMMs: The Forward-Backward Algorithm . . . . .	187
6.6	Maximum Entropy Models: Background . . . . .	193
6.6.1	Linear Regression . . . . .	194
6.6.2	Logistic regression . . . . .	197
6.6.3	Logistic regression: Classification . . . . .	199
6.6.4	Advanced: Learning in logistic regression . . . . .	200
6.7	Maximum Entropy Modeling . . . . .	201
6.7.1	Why do we call it Maximum Entropy? . . . . .	205
6.8	Maximum Entropy Markov Models . . . . .	207
6.8.1	Decoding and Learning in MEMMs . . . . .	210

6.9 Summary . . . . .	212
Bibliographical and Historical Notes . . . . .	212

## II Speech

<b>7 Phonetics</b>	<b>215</b>
7.1 Speech Sounds and Phonetic Transcription . . . . .	216
7.2 Articulatory Phonetics . . . . .	218
7.2.1 The Vocal Organs . . . . .	218
7.2.2 Consonants: Place of Articulation . . . . .	220
7.2.3 Consonants: Manner of Articulation . . . . .	221
7.2.4 Vowels . . . . .	222
7.3 Phonological Categories and Pronunciation Variation . . . . .	225
7.3.1 Phonetic Features . . . . .	227
7.3.2 Predicting Phonetic Variation . . . . .	228
7.3.3 Factors Influencing Phonetic Variation . . . . .	229
7.4 Acoustic Phonetics and Signals . . . . .	230
7.4.1 Waves . . . . .	231
7.4.2 Speech Sound Waves . . . . .	231
7.4.3 Frequency and Amplitude; Pitch and Loudness . . . . .	233
7.4.4 Interpreting Phones from a Waveform . . . . .	236
7.4.5 Spectra and the Frequency Domain . . . . .	236
7.4.6 The Source-Filter Model . . . . .	241
7.5 Phonetic Resources . . . . .	241
7.6 Advanced: Articulatory and Gestural Phonology . . . . .	244
7.7 Summary . . . . .	245
Bibliographical and Historical Notes . . . . .	246
Exercises . . . . .	247
<b>8 Speech Synthesis</b>	<b>249</b>
8.1 Text Normalization . . . . .	250
8.1.1 Sentence Tokenization . . . . .	251
8.1.2 Non-Standard Words . . . . .	253
8.1.3 Homograph Disambiguation . . . . .	256
8.2 Phonetic Analysis . . . . .	257
8.2.1 Dictionary Lookup . . . . .	258
8.2.2 Names . . . . .	259
8.2.3 Grapheme-to-Phoneme . . . . .	259
8.3 Prosodic Analysis . . . . .	263
8.3.1 Prosodic Structure . . . . .	263
8.3.2 Prosodic prominence . . . . .	264
8.3.3 Tune . . . . .	267
8.3.4 More sophisticated models: ToBI . . . . .	267
8.3.5 Computing duration from prosodic labels . . . . .	270
8.3.6 Computing F0 from prosodic labels . . . . .	271
8.3.7 Final result of text analysis: Internal Representation . . . . .	272

8.4	Diphone Waveform synthesis . . . . .	273
8.4.1	Building a diphone database . . . . .	274
8.4.2	Diphone concatenation and TD-PSOLA for prosody . . . . .	275
8.5	Unit Selection (Waveform) Synthesis . . . . .	278
8.6	Evaluation . . . . .	282
	Bibliographical and Historical Notes . . . . .	283
	Exercises . . . . .	285
<b>9</b>	<b>Automatic Speech Recognition</b>	<b>287</b>
9.1	Speech Recognition Architecture . . . . .	289
9.2	Applying the Hidden Markov Model to Speech . . . . .	293
9.3	Feature Extraction: MFCC vectors . . . . .	297
9.3.1	Preemphasis . . . . .	298
9.3.2	Windowing . . . . .	298
9.3.3	Discrete Fourier Transform . . . . .	300
9.3.4	Mel filter bank and log . . . . .	301
9.3.5	The Cepstrum: Inverse Discrete Fourier Transform . . . . .	302
9.3.6	Deltas and Energy . . . . .	304
9.3.7	Summary: MFCC . . . . .	304
9.4	Computing Acoustic Likelihoods . . . . .	305
9.4.1	Vector Quantization . . . . .	305
9.4.2	Gaussian PDFs . . . . .	308
9.4.3	Probabilities, log probabilities and distance functions . . . . .	315
9.5	The Lexicon and Language Model . . . . .	316
9.6	Search and Decoding . . . . .	316
9.7	Embedded Training . . . . .	326
9.8	Evaluation: Word Error Rate . . . . .	330
9.9	Summary . . . . .	332
	Bibliographical and Historical Notes . . . . .	333
	Exercises . . . . .	336
<b>10</b>	<b>Speech Recognition: Advanced Topics</b>	<b>337</b>
10.1	Multipass Decoding: $N$ -best lists and lattices . . . . .	338
10.2	A* ('Stack') Decoding . . . . .	343
10.3	Context-Dependent Acoustic Models: Triphones . . . . .	347
10.4	Discriminative Training . . . . .	351
10.4.1	Maximum Mutual Information Estimation . . . . .	352
10.4.2	Acoustic Models based on Posterior Classifiers . . . . .	354
10.5	Modeling Variation . . . . .	355
10.5.1	Environmental Variation and Noise . . . . .	355
10.5.2	Speaker and Dialect Adaptation: Variation due to speaker differences . . . . .	356
10.5.3	Pronunciation Modeling: Variation due to Genre . . . . .	357
10.6	Metadata: Boundaries, Punctuation, and Disfluencies . . . . .	359
10.7	Speech Recognition by Humans . . . . .	361
10.8	Summary . . . . .	363

Bibliographical and Historical Notes . . . . .	363
Exercises . . . . .	364
<b>11 Computational Phonology</b>	<b>365</b>
11.1 Finite-State Phonology . . . . .	365
11.2 Advanced Finite-State Phonology . . . . .	369
11.2.1 Harmony . . . . .	369
11.2.2 Templatic Morphology . . . . .	370
11.3 Computational Optimality Theory . . . . .	371
11.3.1 Finite-State Transducer Models of Optimality Theory . . . .	373
11.3.2 Stochastic Models of Optimality Theory . . . . .	375
11.4 Syllabification . . . . .	376
11.5 Learning Phonology & Morphology . . . . .	380
11.5.1 Learning Phonological Rules . . . . .	380
11.5.2 Learning Morphology . . . . .	381
11.5.3 Learning in Optimality Theory . . . . .	385
11.6 Summary . . . . .	386
Bibliographical and Historical Notes . . . . .	386
Exercises . . . . .	388
 <b>III Syntax</b>	
<b>12 Formal Grammars of English</b>	<b>389</b>
12.1 Constituency . . . . .	390
12.2 Context-Free Grammars . . . . .	391
12.2.1 Formal definition of context-free grammar . . . . .	395
12.3 Some Grammar Rules for English . . . . .	396
12.3.1 Sentence-Level Constructions . . . . .	396
12.3.2 Clauses and Sentences . . . . .	398
12.3.3 The Noun Phrase . . . . .	398
12.3.4 Agreement . . . . .	403
12.3.5 The Verb Phrase and Subcategorization . . . . .	404
12.3.6 Auxiliaries . . . . .	406
12.3.7 Coordination . . . . .	407
12.4 Treebanks . . . . .	408
12.4.1 Example: The Penn Treebank Project . . . . .	408
12.4.2 Using a Treebank as a Grammar . . . . .	410
12.4.3 Searching Treebanks . . . . .	412
12.4.4 Heads and Head Finding . . . . .	413
12.5 Grammar Equivalence and Normal Form . . . . .	416
12.6 Finite-State and Context-Free Grammars . . . . .	417
12.7 Dependency Grammars . . . . .	418
12.7.1 The Relationship Between Dependencies and Heads . . . .	419
12.7.2 Categorical Grammar . . . . .	420
12.8 Spoken Language Syntax . . . . .	421
12.8.1 Disfluencies and Repair . . . . .	422



12.8.2	Treebanks for Spoken Language . . . . .	423
12.9	Grammars and Human Processing . . . . .	423
12.10	Summary . . . . .	425
	Bibliographical and Historical Notes . . . . .	426
	Exercises . . . . .	428
<b>13</b>	<b>Parsing with Context-Free Grammars</b>	<b>431</b>
13.1	Parsing as Search . . . . .	432
13.1.1	Top-Down Parsing . . . . .	433
13.1.2	Bottom-Up Parsing . . . . .	434
13.1.3	Comparing Top-Down and Bottom-Up Parsing . . . . .	435
13.2	Ambiguity . . . . .	436
13.3	Search in the Face of Ambiguity . . . . .	438
13.4	Dynamic Programming Parsing Methods . . . . .	439
13.4.1	CKY Parsing . . . . .	440
13.4.2	The Earley Algorithm . . . . .	447
13.4.3	Chart Parsing . . . . .	452
13.5	Partial Parsing . . . . .	454
13.5.1	Finite-State Rule-Based Chunking . . . . .	455
13.5.2	Machine Learning-Based Approaches to Chunking . . . . .	456
13.5.3	Evaluating Chunking Systems . . . . .	459
13.6	Summary . . . . .	460
	Bibliographical and Historical Notes . . . . .	461
	Exercises . . . . .	462
<b>14</b>	<b>Statistical Parsing</b>	<b>465</b>
14.1	Probabilistic Context-Free Grammars . . . . .	466
14.1.1	PCFGs for Disambiguation . . . . .	467
14.1.2	PCFGs for Language Modeling . . . . .	469
14.2	Probabilistic CKY Parsing of PCFGs . . . . .	470
14.3	Learning PCFG Rule Probabilities . . . . .	473
14.4	Problems with PCFGs . . . . .	474
14.4.1	Independence assumptions miss structural dependencies between rules . . . . .	474
14.4.2	Lack of sensitivity to lexical dependencies . . . . .	475
14.5	Improving PCFGs by Splitting and Merging Nonterminals . . . . .	477
14.6	Probabilistic Lexicalized CFGs . . . . .	479
14.6.1	The Collins Parser . . . . .	481
14.6.2	Advanced: Further Details of the Collins Parser . . . . .	483
14.7	Evaluating Parsers . . . . .	485
14.8	Advanced: Discriminative Reranking . . . . .	486
14.9	Advanced: Parser-Based Language Modeling . . . . .	488
14.10	Human Parsing . . . . .	489
14.11	Summary . . . . .	491
	Bibliographical and Historical Notes . . . . .	492
	Exercises . . . . .	494

<b>15 Features and Unification</b>	<b>495</b>
15.1 Feature Structures . . . . .	496
15.2 Unification of Feature Structures . . . . .	499
15.3 Feature Structures in the Grammar . . . . .	503
15.3.1 Agreement . . . . .	504
15.3.2 Head Features . . . . .	507
15.3.3 Subcategorization . . . . .	508
15.3.4 Long-Distance Dependencies . . . . .	513
15.4 Implementing Unification . . . . .	513
15.4.1 Unification Data Structures . . . . .	514
15.4.2 The Unification Algorithm . . . . .	515
15.5 Parsing with Unification Constraints . . . . .	519
15.5.1 Integrating Unification into an Earley Parser . . . . .	520
15.5.2 Unification-Based Parsing . . . . .	526
15.6 Types and Inheritance . . . . .	528
15.6.1 Advanced: Extensions to Typing . . . . .	531
15.6.2 Other Extensions to Unification . . . . .	532
15.7 Summary . . . . .	532
Bibliographical and Historical Notes . . . . .	533
Exercises . . . . .	534
<b>16 Language and Complexity</b>	<b>537</b>
16.1 The Chomsky Hierarchy . . . . .	538
16.2 How to Tell if a Language Isn't Regular . . . . .	540
16.2.1 The Pumping Lemma . . . . .	541
16.2.2 Are English and Other Natural Languages Regular Languages? . . . . .	543
16.3 Is Natural Language Context-Free? . . . . .	546
16.4 Complexity and Human Processing . . . . .	548
16.5 Summary . . . . .	550
Bibliographical and Historical Notes . . . . .	551
Exercises . . . . .	552
<b>IV Semantics and Pragmatics</b>	
<b>17 Representing Meaning</b>	<b>553</b>
17.1 Computational Desiderata for Representations . . . . .	555
17.1.1 Verifiability . . . . .	555
17.1.2 Unambiguous Representations . . . . .	556
17.1.3 Canonical Form . . . . .	557
17.1.4 Inference and Variables . . . . .	559
17.1.5 Expressiveness . . . . .	559
17.2 Model-Theoretic Semantics . . . . .	560
17.3 First-Order Logic . . . . .	563
17.3.1 Basic Elements of First Order Logic . . . . .	563
17.3.2 Variables and Quantifiers . . . . .	566
17.3.3 Lambda Notation . . . . .	568

17.3.4	The Semantics of First-Order Logic . . . . .	569
17.3.5	Inference . . . . .	570
17.4	Representing Events and States . . . . .	572
17.4.1	Representing Time . . . . .	575
17.4.2	Aspect . . . . .	578
17.5	Related Representational Approaches . . . . .	581
17.5.1	Description Logics . . . . .	582
17.6	Alternative Approaches to Meaning . . . . .	587
17.6.1	Meaning as Action . . . . .	588
17.7	Summary . . . . .	588
	Bibliographical and Historical Notes . . . . .	589
	Exercises . . . . .	590
<b>18</b>	<b>Computational Semantics</b>	<b>593</b>
18.1	Syntax-Driven Semantic Analysis . . . . .	593
18.2	Semantic Augmentations to Context-Free Grammar Rules . . . . .	595
18.3	Quantifier Scope Ambiguity and Underspecification . . . . .	602
18.3.1	Store and Retrieve Approaches . . . . .	602
18.4	Unification-Based Approaches to Semantic Analysis . . . . .	604
18.5	Semantic Attachments for a Fragment of English . . . . .	610
18.5.1	Sentences . . . . .	610
18.5.2	Noun Phrases . . . . .	612
18.5.3	Verb Phrases . . . . .	615
18.5.4	Prepositional Phrases . . . . .	617
18.6	Integrating Semantics into the Earley Parser . . . . .	619
18.7	Idioms and Compositionality . . . . .	621
18.8	Summary . . . . .	622
	Bibliographical and Historical Notes . . . . .	623
	Exercises . . . . .	624
<b>19</b>	<b>Lexical Semantics</b>	<b>627</b>
19.1	Word Senses . . . . .	628
19.2	Relations between Senses . . . . .	631
19.2.1	Synonymy and Antonymy . . . . .	631
19.2.2	Hyponymy . . . . .	632
19.2.3	Semantic Fields . . . . .	633
19.3	WordNet: A Database of Lexical Relations . . . . .	633
19.4	Event Participants: Semantic Roles and Selectional Restrictions . . . . .	635
19.4.1	Thematic Roles . . . . .	636
19.4.2	Diathesis Alternations . . . . .	637
19.4.3	Problems with Thematic Roles . . . . .	639
19.4.4	The Proposition Bank . . . . .	640
19.4.5	FrameNet . . . . .	641
19.4.6	Selectional Restrictions . . . . .	643
19.5	Primitive Decomposition . . . . .	645
19.6	Advanced concepts: Metaphor . . . . .	647

19.7 Summary . . . . .	648
Bibliographical and Historical Notes . . . . .	649
Exercises . . . . .	650
<b>20 Computational Lexical Semantics</b>	<b>653</b>
20.1 Word Sense Disambiguation: Overview . . . . .	654
20.2 Supervised Word Sense Disambiguation . . . . .	655
20.2.1 Extracting Feature Vectors for Supervised Learning . . . . .	656
20.2.2 Naive Bayes and Decision List Classifiers . . . . .	657
20.3 WSD Evaluation, Baselines, and Ceilings . . . . .	660
20.4 WSD: Dictionary and Thesaurus Methods . . . . .	662
20.4.1 The Lesk Algorithm . . . . .	662
20.4.2 Selectional Restrictions and Selectional Preferences . . . . .	664
20.5 Minimally Supervised WSD: Bootstrapping . . . . .	666
20.6 Word Similarity: Thesaurus Methods . . . . .	668
20.7 Word Similarity: Distributional Methods . . . . .	674
20.7.1 Defining a Word's Co-occurrence Vectors . . . . .	675
20.7.2 Measures of Association with Context . . . . .	676
20.7.3 Defining similarity between two vectors . . . . .	679
20.7.4 Evaluating Distributional Word Similarity . . . . .	683
20.8 Hyponymy and other word relations . . . . .	684
20.9 Semantic Role Labeling . . . . .	687
20.10 Advanced: Unsupervised Sense Disambiguation . . . . .	690
Bibliographical and Historical Notes . . . . .	691
Exercises . . . . .	695
<b>21 Computational Discourse</b>	<b>697</b>
21.1 Discourse Segmentation . . . . .	700
21.1.1 Unsupervised Discourse Segmentation . . . . .	700
21.1.2 Supervised Discourse Segmentation . . . . .	702
21.1.3 Evaluating Discourse Segmentation . . . . .	704
21.2 Text Coherence . . . . .	705
21.2.1 Rhetorical Structure Theory . . . . .	706
21.2.2 Automatic Coherence Assignment . . . . .	708
21.3 Reference Resolution . . . . .	711
21.4 Reference Phenomena . . . . .	714
21.4.1 Five Types of Referring Expressions . . . . .	714
21.4.2 Information Status . . . . .	716
21.5 Features for Pronominal Anaphora Resolution . . . . .	717
21.6 Three algorithms for pronominal anaphora resolution . . . . .	720
21.6.1 Pronominal Anaphora Baseline: The Hobbs Algorithm . . . . .	720
21.6.2 A Centering Algorithm for Anaphora Resolution . . . . .	722
21.6.3 A Log-Linear model for Pronominal Anaphora Resolution . . . . .	724
21.6.4 Features . . . . .	725
21.7 Coreference Resolution . . . . .	726
21.8 Evaluating Coreference Resolution . . . . .	728

21.9	Advanced: Inference-Based Coherence Resolution . . . . .	728
21.10	Psycholinguistic Studies of Reference and Coherence . . . . .	734
21.11	Summary . . . . .	735
	Bibliographical and Historical Notes . . . . .	736
	Exercises . . . . .	738

## V Applications

<b>22</b>	<b>Information Extraction</b>	<b>741</b>
22.1	Named Entity Recognition . . . . .	743
22.1.1	Ambiguity in Named Entity Recognition . . . . .	745
22.1.2	NER as Sequence Labeling . . . . .	745
22.1.3	Evaluating Named Entity Recognition . . . . .	749
22.1.4	Practical NER Architectures . . . . .	750
22.2	Relation Detection and Classification . . . . .	751
22.2.1	Supervised Learning Approaches to Relation Analysis . . . . .	752
22.2.2	Lightly Supervised Approaches to Relation Analysis . . . . .	755
22.2.3	Evaluating Relation Analysis Systems . . . . .	758
22.3	Temporal and Event Processing . . . . .	759
22.3.1	Temporal Expression Recognition . . . . .	760
22.3.2	Temporal Normalization . . . . .	762
22.3.3	Event Detection and Analysis . . . . .	765
22.3.4	TimeBank . . . . .	766
22.4	Template-Filling . . . . .	768
22.4.1	Statistical Approaches to Template-Filling . . . . .	769
22.4.2	Finite-State Template-Filling Systems . . . . .	770
22.5	Advanced: Biomedical Information Extraction * . . . . .	773
22.5.1	Biological Named Entity Recognition . . . . .	774
22.5.2	Gene Normalization . . . . .	775
22.5.3	Biological Roles and Relations . . . . .	776
22.6	Summary . . . . .	778
	Bibliographical and Historical Notes . . . . .	778
	Exercises . . . . .	779
<b>23</b>	<b>Question Answering and Summarization</b>	<b>783</b>
23.1	Information Retrieval . . . . .	785
23.1.1	The Vector Space Model . . . . .	786
23.1.2	Term Weighting . . . . .	789
23.1.3	Term Selection and Creation . . . . .	790
23.1.4	Evaluating Information Retrieval Systems . . . . .	790
23.1.5	Homonymy, Polysemy, and Synonymy . . . . .	794
23.1.6	Improving User Queries . . . . .	795
23.2	Factoid Question Answering . . . . .	796
23.2.1	Question Processing . . . . .	798
23.2.2	Passage Retrieval . . . . .	801
23.2.3	Answer Processing . . . . .	802

23.2.4	Evaluation of Factoid Answers . . . . .	805
23.3	Summarization . . . . .	805
23.3.1	Summarizing Single Documents . . . . .	808
23.4	Multi-Document Summarization . . . . .	814
23.4.1	Content Selection in Multi-Document Summarization . . . . .	815
23.4.2	Information Ordering in Multi-Document Summarization . . . . .	816
23.5	Between Question Answering and Summarization: Query-Focused Summarization . . . . .	819
23.6	Summarization Evaluation . . . . .	823
23.7	Summary . . . . .	825
	Bibliographical and Historical Notes . . . . .	826
	Exercises . . . . .	828
<b>24</b>	<b>Dialogue and Conversational Agents</b> . . . . .	<b>829</b>
24.1	Properties of Human Conversations . . . . .	831
24.1.1	Turns and Turn-Taking . . . . .	832
24.1.2	Language as Action: Speech Acts . . . . .	833
24.1.3	Language as Joint Action: Grounding . . . . .	834
24.1.4	Conversational Structure . . . . .	836
24.1.5	Conversational Implicature . . . . .	837
24.2	Basic Dialogue Systems . . . . .	839
24.2.1	ASR component . . . . .	839
24.2.2	NLU component . . . . .	841
24.2.3	Generation and TTS components . . . . .	844
24.2.4	Dialogue Manager . . . . .	845
24.2.5	Dialogue Manager Error Handling: Confirmation/Rejection . . . . .	849
24.3	VoiceXML . . . . .	851
24.4	Dialogue System Design and Evaluation . . . . .	854
24.4.1	Designing Dialogue Systems . . . . .	854
24.4.2	Dialogue System Evaluation . . . . .	855
24.5	Information-state & Dialogue Acts . . . . .	857
24.5.1	Dialogue Acts . . . . .	859
24.5.2	Interpreting Dialogue Acts . . . . .	860
24.5.3	Detecting Correction Acts . . . . .	862
24.5.4	Generating Dialogue Acts: Confirmation and Rejection . . . . .	863
24.6	Markov Decision Process Architecture . . . . .	865
24.7	Advanced: Plan-based Dialogue Agents . . . . .	869
24.7.1	Plan-Inferential Interpretation and Production . . . . .	869
24.7.2	The Intentional Structure of Dialogue . . . . .	872
24.8	Summary . . . . .	874
	Bibliographical and Historical Notes . . . . .	875
	Exercises . . . . .	876
<b>25</b>	<b>Machine Translation</b> . . . . .	<b>879</b>
25.1	Why is Machine Translation So Hard? . . . . .	882
25.1.1	Typology . . . . .	882

25.1.2	Other Structural Divergences . . . . .	885
25.1.3	Lexical Divergences . . . . .	885
25.2	Classical MT & the Vauquois Triangle . . . . .	887
25.2.1	Direct Translation . . . . .	887
25.2.2	Transfer . . . . .	890
25.2.3	Combining direct and transfer approaches in classic MT . . . . .	892
25.2.4	The Interlingua Idea: Using Meaning . . . . .	893
25.3	Statistical MT . . . . .	895
25.4	$P(F E)$ : the Phrase-Based Translation Model . . . . .	897
25.5	Alignment in MT . . . . .	900
25.5.1	IBM Model 1 . . . . .	901
25.5.2	HMM Alignment . . . . .	903
25.6	Training Alignment Models . . . . .	905
25.6.1	EM for Training Alignment Models . . . . .	906
25.7	Symmetrizing Alignments for Phrase-based MT . . . . .	908
25.8	Decoding for Phrase-Based Statistical MT . . . . .	910
25.9	MT Evaluation . . . . .	914
25.9.1	Using Human Raters . . . . .	915
25.9.2	Automatic Evaluation: Bleu . . . . .	915
25.10	Advanced: Syntactic Models for MT . . . . .	918
25.11	Advanced: IBM Model 3 for fertility-based alignment . . . . .	920
25.11.1	Training for Model 3 . . . . .	923
25.12	Advanced: Log-linear Models for MT . . . . .	924
	Bibliographical and Historical Notes . . . . .	925
	Exercises . . . . .	927
	<b>Bibliography</b>	<b>929</b>
	<b>Index</b>	<b>981</b>

# Foreword

TO BE ADDED

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

DRAFT



# 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 explosion of Web-based language techniques, and the commercial availability of telephone-based dialogue systems, and speech synthesis and speech recognition have provided an important impetus for the 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 machine translation in natural language processing courses in computer science departments; and computational morphology, phonology and pragmatics 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, artificial intelligence, machine learning, statistics, or 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 log-linear models) can be applied to important real-world problems: spell checking, text document search, machine translation, speech recognition, information extraction from the web, and spoken-language dialogue. 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. Most chapters therefore include an evaluation section describing modern empirical methods for evaluating systems

and performing error analysis, 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, Fisher, 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 be 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 five parts in addition to an introduction and end matter. Part I, "Words", introduces concepts related to the processing of words and simple word sequences: word segmentation, word morphology, word edit distance, parts-of-speech, and the algorithms used to process them: regular expressions, finite automata, finite transducers,  $N$ -grams, Hidden Markov Models, and log-linear models. Part II, "Speech", introduces linguistic phonetics, and then covers speech synthesis, speech recognition, and linguistic topics in computational phonology. Part III, "Syntax", introduces phrase structure grammars for English and gives essential algorithms for processing structured syntactic relationships among words: the CKY and Earley algorithms for parsing, statistical parsing, unification and typed feature structures, and analytical tools like the Chomsky hierarchy and the pumping lemma. Part IV, "Semantics and Pragmatics", introduces first order predicate calculus and other ways of representing meaning, the lambda calculus, lexical semantics, lexical semantic resources such as WordNet, PropBank, and FrameNet, and computational models of lexical semantics for word similarity, word sense disambiguation, and discourse topics like coreference and coherence. Part V, "Applications", covers information extraction, machine translation and dialog and conversational agents.

## 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:

<b>NLP 1 quarter</b>	<b>NLP 1 semester</b>	<b>Speech + NLP 1 semester</b>	<b>Comp. Ling. 1 quarter</b>
1. Intro	1. Intro	1. Intro	1. Intro
2. Regex, FSA	2. Regex, FSA	2. Regex, FSA	2. Regex, FSA
4. N-grams	4. N-grams	4. N-grams	3. Morph., FST
5. POS tagging	5. POS tagging	5. POS tagging	4. N-grams
12. CFGs	6. HMMS	6. HMMs	5. POS tagging
13. Parsing	12. CFGs	8. TTS	13. Parsing
14. Stat. Parsing	13. Parsing	9. ASR	14. Stat. Parsing
19. Lex. Semantics	14. Stat. Parsing	12. CFGs	15. Complexity
20. Comp. Lex. Sem.	17. Semantics	13. Parsing	16. Unification
23. QA & Summar.	18. Comp. Semantics	14. Stat. Parsing	20. Comp. Lex. Sem.
25. MT	19. Lex. Semantics	17. Semantics	21. Discourse
	20. Comp. Lex. Sem.	19. Lex. Sem.	
	21. Discourse	20. Comp. Lex. Sm.	
	22. IE	22. IE	
	23. QA & Summar.	24. Dialog	
	25. MT	25. MT.	

Selected chapters from the book could also be used to augment courses in Artificial Intelligence, Cognitive Science, Information Retrieval, or Electrical Engineering-oriented courses in Speech Processing.

Resources associated with the book such as online versions of figures and pointers to web content can be found at the book home-page:

<http://www.cs.colorado.edu/~martin/slp.html>.

## Acknowledgments

Andy Kehler wrote the Discourse chapter for the first edition; we took Andy's chapter as a starting point for this second edition chapter. Similarly, Nigel Ward wrote most of the MT chapter for the first edition, and we used this as the starting point for the MT chapter of this second edition. Kevin Bretonnel Cohen wrote section 22.5 on biomedical information extraction. Keith Vander Linden wrote the Generation chapter in the first edition.

Dan would like to thank his parents for encouraging him to do the right thing, do 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 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 teaching him to always go look at the data, Robert Wilensky, his dissertation advisor, for teaching him the importance of collaboration and group spirit in research, Chris Manning for being a superb collaborator at Stanford, and of course all his former colleagues at Boulder.

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 dissertation 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; Martha Palmer, Tammy Sumner and Wayne Ward for being wonderful collaborators at Boulder. Finally, Jim would like to thank his wife Linda for all her support and patience through the years, and his daughter Katie who has waited her entire life for the completion of this edition.

Boulder and Stanford have been very rewarding places to work on speech and language processing. We'd like to thank our departments, our colleagues, and our students at both places for their collaborations, which have greatly influenced our research and teaching.

We are grateful to the many people who helped enormously on the first edition of the book. This second edition has also benefited from our many readers and from their course-testing. Special thanks for extraordinarily helpful comments and ideas on wide swaths of the book to Regina Barzilay, Philip Resnik, Emily Bender, and Adam Przepiórkowski. Our editor Tracy Dunkelberger, 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, Stanford University, and the LSA Summer Institutes at the University of Illinois at Urbana-Champaign (1999), MIT (2005), and Stanford (2007), as well as:

Rieks op den Akker, Kayra Akman, Angelos Alexopoulos, Robin Aly, S. M. Niaz Arifin, Nimar S. Arora, Tsz-Chiu Au, G. W. Blackwood, Bai Xiaojing, Ellie Baker, Jason Baldridge, Clay Beckner, Jonathan Boiser, Marion Bond, Marco Aldo Piccolino Boniforti, Onn Brandman, Chris Brew, Tore Bruland, Denis Bueno, Sean M. Burke, Bill Byrne, Kai-Uwe Carstensen, Alejandro Cdebaca, Pichuan Chang, Grace Chung, Andrew Clausen, Kevin B. Cohen, Frederik Coppens, Stephen Cox, Heriberto Cuayáhuitl, Martin Davidsson, Paul Davis, Jon Dehdari, Franz Deuzer, Mike Dillinger, Bonnie Dorr, Jason Eisner, John Eng, Ersin Er, Hakan Erdogan, Gülsen Eryiğit, Barbara Di Eugenio, Eric Fosler-Lussier, Olac Fuentes, Dale Gerdemann, Dan Gildea, Filip Ginter, Cynthia Girand, Anthony Gitter, John A. Goldsmith, Michelle Gregory, Rocio Guillen, Jeffrey S. Haemer, Adam Hahn, Patrick Hall, Harald Hammarström, Mike Hammond, Eric Hansen, Marti Hearst, Paul Hirschbühler, Julia Hirschberg, Julia Hockenmaier, Jeremy Hoffman, Greg Hullender, Gaja Jarosz, Eric W. Johnson, Chris Jones, Edwin de Jong, Bernadette Joret, Fred Karlsson, Graham Katz, Stefan Kaufmann, Andy Kehler, Manuel Kirschner, Sheldon Klein, Kevin Knight, Jean-Pierre Koenig, Kimmo Koskenniemi, Alexander Kostyrkin, Valerie Krugler, Mikko Kurimo, Mike LeBeau, Chia-Ying Lee, Jaeyong Lee, Scott Leishman, Szymon Letowski, Liuyang Li, Marc Light, Greger Lind'en, Pierre Lison, Diane Litman, Chao-Lin Liu, Feng Liu, Roussanka Louka, Artyom Lukanin, Jean Ma, Maxim Makatchev, Inderjeet Mani, Steve Marmon, Marie-Catherine de Marneffe, Hendrik Maryns, Jon May, Dan Melamed, Johanna Moore, Nelson Morgan, Emad Nawfal, Mark-Jan Nederhof, Hwee Tou Ng, John Niekrasz, Rodney Nielsen, Yuri Niyazov, Kris Nuttycombe, Mike O'Connell, Robert Oberbreckling, Scott Olsson, Woodley Packard, Gabor Palagyi, Gerald Penn, Rani Pinchuk, Sameer Pradhan, Kathryn Pruitt, Drago Radev, William J. Rapaport, Dan Ramage, Ron Regan, Ehud Reiter, Steve Renals, Chang-han Rhee, Dan Rose, Mike Rosner, Deb Roy, Teodor Rus, William Gregory Sakas, Murat Saraclar, Stefan Schaden, Anna Schapiro, Zhang Sen, Matt Shannon, Stuart C. Shapiro, Ilya Sherman, Lokesh Shrestha, Nathan Silberman, Otakar Smrz, Rion Snow, Niyue Tan, Frank Yung-Fong Tang, Ahmet Cüneyd Tantuğ, Paul Taylor, Lorne Temes,

Rich Thomason, Almer S. Tigelaar, Richard Trahan, Antoine Trux, Clement Wang, Nigel Ward, Rachel Weston, Janyce Wiebe, Lauren Wilcox, Ben Wing, Dean Earl Wright III, Dekai Wu, Lei Wu, Eric Yeh, Alan C. Yeung, Margalit Zabludowski, Menno van Zaanen, Sam Shaojun Zhao, and Xingtao Zhao.

Daniel Jurafsky  
Stanford, California  
James H. Martin  
Boulder, Colorado