

This excerpt from

Foundations of Statistical Natural Language Processing.  
Christopher D. Manning and Hinrich Schütze.  
© 1999 The MIT Press.

is provided in screen-viewable form for personal use only by members of MIT CogNet.

Unauthorized use or dissemination of this information is expressly forbidden.

If you have any questions about this material, please contact [cognetadmin@cognet.mit.edu](mailto:cognetadmin@cognet.mit.edu).

## *List of Figures*

1.1	Zipf's law.	26
1.2	Mandelbrot's formula.	27
1.3	Key Word In Context (KWIC) display for the word <i>showed</i> .	32
1.4	Syntactic frames for <i>showed</i> in <i>Tom Sawyer</i> .	33
2.1	A diagram illustrating the calculation of conditional probability $P(A B)$ .	42
2.2	A random variable $X$ for the sum of two dice.	45
2.3	Two examples of binomial distributions: $b(r; 10, 0.7)$ and $b(r; 10, 0.1)$ .	52
2.4	Example normal distribution curves: $n(x; 0, 1)$ and $n(x; 1.5, 2)$ .	53
2.5	The entropy of a weighted coin.	63
2.6	The relationship between mutual information $I$ and entropy $H$ .	67
2.7	The noisy channel model.	69
2.8	A binary symmetric channel.	69
2.9	The noisy channel model in linguistics.	70
3.1	An example of recursive phrase structure expansion.	99
3.2	An example of a prepositional phrase attachment ambiguity.	108
4.1	Heuristic sentence boundary detection algorithm.	135
4.2	A sentence as tagged according to several different tag sets.	140
5.1	Using a three word collocational window to capture bigrams at a distance.	158



5.2	Histograms of the position of <i>strong</i> relative to three words.	160
7.1	Bayesian disambiguation.	238
7.2	The Flip-Flop algorithm applied to finding indicators for disambiguation.	240
7.3	Lesk's dictionary-based disambiguation algorithm.	243
7.4	Thesaurus-based disambiguation.	245
7.5	Adaptive thesaurus-based disambiguation.	246
7.6	Disambiguation based on a second-language corpus.	249
7.7	Disambiguation based on "one sense per collocation" and "one sense per discourse."	252
7.8	An EM algorithm for learning a word sense clustering.	254
8.1	A diagram motivating the measures of precision and recall.	268
8.2	Attachments in a complex sentence.	285
8.3	A document-by-word matrix $A$ .	297
8.4	A word-by-word matrix $B$ .	297
8.5	A modifier-by-head matrix $C$ .	297
9.1	A Markov model.	319
9.2	The crazy soft drink machine, showing the states of the machine and the state transition probabilities.	321
9.3	A section of an HMM for a linearly interpolated language model.	323
9.4	A program for a Markov process.	325
9.5	Trellis algorithms.	328
9.6	Trellis algorithms: Closeup of the computation of forward probabilities at one node.	329
9.7	The probability of traversing an arc.	334
10.1	Algorithm for training a Visible Markov Model Tagger.	348
10.2	Algorithm for tagging with a Visible Markov Model Tagger.	350
10.3	The learning algorithm for transformation-based tagging.	364
11.1	The two parse trees, their probabilities, and the sentence probability.	385
11.2	A Probabilistic Regular Grammar (PRG).	390
11.3	Inside and outside probabilities in PCFGs.	391
12.1	A word lattice (simplified).	408

<i>List of Figures</i>	<i>xxiii</i>
12.2 A Penn Treebank tree.	413
12.3 Two CFG derivations of the same tree.	421
12.4 An LC stack parser.	425
12.5 Decomposing a local tree into dependencies.	430
12.6 An example of the PARSEVAL measures.	433
12.7 The idea of crossing brackets.	434
12.8 Penn trees versus other trees.	436
13.1 Different strategies for Machine Translation.	464
13.2 Alignment and correspondence.	469
13.3 Calculating the cost of alignments.	473
13.4 A sample dot plot.	476
13.5 The pillow-shaped envelope that is searched.	480
13.6 The noisy channel model in machine translation.	486
14.1 A single-link clustering of 22 frequent English words represented as a dendrogram.	496
14.2 Bottom-up hierarchical clustering.	502
14.3 Top-down hierarchical clustering.	502
14.4 A cloud of points in a plane.	504
14.5 Intermediate clustering of the points in figure 14.4.	504
14.6 Single-link clustering of the points in figure 14.4.	505
14.7 Complete-link clustering of the points in figure 14.4.	505
14.8 The K-means clustering algorithm.	516
14.9 One iteration of the K-means algorithm.	517
14.10 An example of using the EM algorithm for soft clustering.	519
15.1 Results of the search “glass pyramid” Pei Louvre’ on an internet search engine.	531
15.2 Two examples of precision-recall curves.	537
15.3 A vector space with two dimensions.	540
15.4 The Poisson distribution.	546
15.5 An example of a term-by-document matrix $A$ .	555
15.6 Dimensionality reduction.	555
15.7 An example of linear regression.	558
15.8 The matrix $T$ of the SVD decomposition of the matrix in figure 15.5.	560
15.9 The matrix of singular values of the SVD decomposition of the matrix in figure 15.5.	560



15.10	The matrix $D$ of the SVD decomposition of the matrix in figure 15.5.	561
15.11	The matrix $B = S_{2 \times 2} D_{2 \times n}$ of documents after rescaling with singular values and reduction to two dimensions.	562
15.12	Three constellations of cohesion scores in topic boundary identification.	569
16.1	A decision tree.	578
16.2	Geometric interpretation of part of the tree in figure 16.1.	579
16.3	An example of a Reuters news story in the topic category “earnings.”	580
16.4	Pruning a decision tree.	585
16.5	Classification accuracy depends on the amount of training data available.	587
16.6	An example of how decision trees use data inefficiently from the domain of phonological rule learning.	588
16.7	The Perceptron Learning Algorithm.	598
16.8	One error-correcting step of the perceptron learning algorithm.	600
16.9	Geometric interpretation of a perceptron.	602

This excerpt from

Foundations of Statistical Natural Language Processing.  
Christopher D. Manning and Hinrich Schütze.  
© 1999 The MIT Press.

is provided in screen-viewable form for personal use only by members of MIT CogNet.

Unauthorized use or dissemination of this information is expressly forbidden.

If you have any questions about this material, please contact [cognetadmin@cognet.mit.edu](mailto:cognetadmin@cognet.mit.edu).