

Chapter 1

Appendix: NLTK Modules and Corpora

Corpora and Corpus Samples Distributed with NLTK (starred items with NLTK-Lite)		
Corpus	Compiler	Contents
Brown Corpus	Francis, Kucera	15 genres, 1.15M words, tagged
CoNLL 2000 Chunking Data	Tjong Kim Sang	270k words, tagged and chunked
Genesis Corpus	Misc web sources	6 texts, 200k words, 6 languages
Project Gutenberg (sel)	Hart, Newby, et al	14 texts, 1.7M words
NIST 1999 Info Extr (sel)	Garofolo	63k words, newswire and named-entity SGML markup
Lexicon Corpus		Words, tags and frequencies from Brown Corpus and WSJ
Names Corpus	Kantrowitz, Ross	8k male and female names
PP Attachment Corpus	Ratnaparkhi	28k prepositional phrases, tagged as noun or verb modifiers
Presidential Addresses	Ahrens	485k words, formatted text
Roget's Thesaurus	Project Gutenberg	200k words, formatted text
SEMCOR	Rus, Mihalcea	880k words, part-of-speech and sense tagged
SENSEVAL 2 Corpus	Ted Pedersen	600k words, part-of-speech and sense tagged
Stopwords Corpus	Porter et al	2,400 stopwords for 11 languages
Penn Treebank (sel)	LDC	40k words, tagged and parsed
TIMIT Corpus (sel)	NIST/LDC	audio files and transcripts for 16 speakers
Wordlist Corpus	OpenOffice.org et al	960k words and 20k affixes for 8 languages

Table 1.1:

About this document...

This chapter is a draft from *Introduction to Natural Language Processing*, by Steven Bird, Ewan Klein and Edward Loper, Copyright © 2007 the authors. It is distributed with the *Natural Language Toolkit* [<http://nltk.sourceforge.net>], Version 0.7.5, under the terms of the *Creative Commons Attribution-ShareAlike License* [<http://creativecommons.org/licenses/by-sa/2.5/>].

This document is Revision: 4518 Wed May 16 20:08:28 EST 2007