Chapter 1

Appendix: NLTK Modules and Corpora

NLTK Organization: NLTK is organized into a collection of task-specific packages. Each package is a combination of data structures for representing a particular kind of information such as trees, and implementations of standard algorithms involving those structures such as parsers. This approach is a standard feature of *object-oriented design*, in which components encapsulate both the resources and methods needed to accomplish a particular task.

The most fundamental NLTK components are for identifying and manipulating individual words of text. These include: `tokenize`, for breaking up strings of characters into word tokens; `tag`, for adding part-of-speech tags, including regular-expression taggers, n-gram taggers and Brill taggers; and the Porter stemmer.

The second kind of module is for creating and manipulating structured linguistic information. These components include: `tree`, for representing and processing parse trees; `featurestructure`, for building and unifying nested feature structures (or attribute-value matrices); `cfg`, for specifying context-free grammars; and `parse`, for creating parse trees over input text, including chart parsers, chunk parsers and probabilistic parsers.

Several utility components are provided to facilitate processing and visualization. These include: `draw`, to visualize NLP structures and processes; `probability`, to count and collate events, and perform statistical estimation; and `corpora`, to access tagged linguistic corpora.

A further group of components is not part of NLTK proper. These are a wide selection of third-party contributions, often developed as student projects at various institutions where NLTK is used, and distributed in a separate package called *NLTK Contrib*. Several of these student contributions, such as the Brill tagger and the HMM module, have now been incorporated into NLTK. Although these contributed components are not maintained, they may serve as a useful starting point for future student projects.

In addition to software and documentation, NLTK provides substantial corpus samples. Many of these can be accessed using the `corpora` module, avoiding the need to write specialized file parsing code before you can do NLP tasks. These corpora include: Brown Corpus — 1.15 million words of tagged text in 15 genres; a 10% sample of the Penn Treebank corpus, consisting of 40,000 words of syntactically parsed text; a selection of books from Project Gutenberg totally 1.7 million words; and other corpora for chunking, prepositional phrase attachment, word-sense disambiguation, text categorization, and information extraction.

Corpora and Corpus Samples Distributed with NLTK		
Corpus	**Compiler**	**Contents**
Alpino Dutch Treebank	van Noord	140k words, tagged and parsed (Dutch)
Australian ABC News	Bird	2 genres, 660k words, sentence-segmented
Brown Corpus	Francis, Kucera	15 genres, 1.15M words, tagged, categorized
CESS-CAT Catalan Treebank	CLiC-UB et al	500k words, tagged and parsed
CESS-ESP Spanish Treebank	CLiC-UB et al	500k words, tagged and parsed
CMU Pronouncing Dictionary	CMU	127k entries
CoNLL 2000 Chunking Data	Tjong Kim Sang	270k words, tagged and chunked
CoNLL 2002 Named Entity	Tjong Kim Sang	700k words, pos- and named-entity-tagged (Dutch, Spanish)
Floresta Treebank	Diana Santos et al	9k sentences (Portuguese)
Genesis Corpus	Misc web sources	6 texts, 200k words, 6 languages
Gutenberg (sel)	Hart, Newby, et al	14 texts, 1.7M words
Indian POS-Tagged Corpus	Kumaran et al	60k words, tagged (Bangla, Hindi, Marathi, Telugu)
MacMorpho Corpus	NILC, USP, Brazil	1M words, tagged (Brazilian Portuguese)
Movie Reviews	Pang, Lee	Sentiment Polarity Dataset 2.0
Names Corpus	Kantrowitz, Ross	8k male and female names
NIST 1999 Info Extr (sel)	Garofolo	63k words, newswire and named-entity SGML markup
NPS Chat Corpus	Forsyth, Martell	10k IM chat posts, POS-tagged and dialogue-act tagged
PP Attachment Corpus	Ratnaparkhi	28k prepositional phrases, tagged as noun or verb modifiers
Presidential Addresses	Ahrens	485k words, formatted text
Proposition Bank	Palmer	113k propositions, 3300 verb frames
Question Classification	Li, Roth	6k questions, categorized
Reuters Corpus	Reuters	1.3M words, 10k news documents, categorized
Roget's Thesaurus	Project Gutenberg	200k words, formatted text
RTE Textual Entailment	Dagan et al	8k sentence pairs, categorized
SEMCOR	Rus, Mihalcea	880k words, part-of-speech and sense tagged
SENSEVAL 2 Corpus	Ted Pedersen	600k words, part-of-speech and sense tagged
Shakespeare XML texts (sel)	Jon Bosak	8 books
Stopwords Corpus	Porter et al	2,400 stopwords for 11 languages
Switchboard Corpus (sel)	LDC	36 phonecalls, transcribed, parsed
Univ Decl of Human Rights	■	480k words, 300+ languages

Corpora and Corpus Samples Distributed with NLTK		
Corpus	**Compiler**	**Contents**
US Pres Addr Corpus	Ahrens	480k words
Penn Treebank (sel)	LDC	40k words, tagged and parsed
TIMIT Corpus (sel)	NIST/LDC	audio files and transcripts for 16 speakers
VerbNet 2.1	Palmer et al	5k verbs, hierarchically organized, linked to WordNet
Wordlist Corpus	OpenOffice.org et al	960k words and 20k affixes for 8 languages
WordNet 3.0 (English)	Miller, Fellbaum	145k synonym sets

Table 1.1: