NAME

index.noun, data.noun, index.verb, data.verb, index.adj, data.adj, index.adv, data.adv − WordNet database files (default file names)

noun.idx, noun.dat, verb.idx, verb.dat, adj.idx, adj.dat, adv.idx, adv.dat − WordNet database files (Windows)

noun.exc, verb.exc. adj.exc adv.exc − morphology exception lists

sentidx.vrb, sents.vrb − files used by search code to display sentences illustrating the use of some specific verbs

(The remainder of this manual page refers to database files by their default file names.)

DESCRIPTION

For each syntactic category, two files are needed to represent the contents of the WordNet database − **index.***pos* and **data.***pos*, where *pos* is **noun**, **verb**, **adj** and **adv**. The other auxiliary files are used by the WordNet library's searching functions and are needed to run the various WordNet browsers.

Each index file is an alphabetized list of all the words found in WordNet in the corresponding part of speech. On each line, following the word, is a list of byte offsets (*synset_offset*s) in the corresponding data file, one for each synset containing the word. Words in the index file are in lower case only, regardless of how they were entered in the lexicographer files. This folds various orthographic representations of the word into one line enabling database searches to be case insensitive. See **wninput**(5WN) for a detailed description of the lexicographer files

A data file for a syntactic category contains information corresponding to the synsets that were specified in the lexicographer files, with relational pointers resolved to *synset_offset*s. Each line corresponds to a synset. Pointers are followed and hierarchies traversed by moving from one synset to another via the *synset_offset*s.

The exception list files, *pos***.exc**, are used to help the morphological processor find base forms from irregular inflections.

The files **sentidx.vrb** and **sents.vrb** contain sentences illustrating the use of specific senses of some verbs. These files are used by the searching software in response to a request for verb sentence frames. Generic sentence frames are displayed when an illustrative sentence is not present.

The various database files are in ASCII formats that are easily read by both humans and machines. All fields, unless otherwise noted, are separated by one space character, and all lines are terminated by a newline character. Fields enclosed in italicized square brackets may not be present.

See **wngloss**(7WN) for a glossary of WordNet terminology and a discussion of the database's content and logical organization.

Index File Format

Each index file begins with several lines containing a copyright notice, version number and license agreement. These lines all begin with two spaces and the line number so they do not interfere with the binary search algorithm that is used to look up entries in the index files. All other lines are in the following format. In the field descriptions, **number** always refers to a decimal integer unless otherwise defined.

lemma pos synset_cnt p_cnt [ptr_symbol...] sense_cnt tagsense_cnt synset_offset [synset_offset...]

lemma lower case ASCII text of word or collocation. Collocations are formed by joining individual words with an underscore (_) character.

pos Syntactic category: **n** for noun files, **v** for verb files, **a** for adjective files, **r** for adverb files.

All remaining fields are with respect to senses of *lemma* in *pos*.

synset_cnt Number of synsets that *lemma* is in. This is the number of senses of the word in WordNet. See **Sense Numbers** below for a discussion of how sense numbers are assigned and the order of *synset_offset*s in the index files.

p_cnt Number of different pointers that *lemma* has in all synsets containing it.

ptr_symbol A space separated list of *p_cnt* different types of pointers that *lemma* has in all synsets containing it. See **wninput**(5WN) for a list of *pointer_symbol*s. If all senses of *lemma* have no pointers, this field is omitted and *p_cnt* is **0**.

sense_cnt Same as *sense_cnt* above. This is redundant, but the field was preserved for compatibility reasons.

tagsense_cnt Number of senses of *lemma* that are ranked according to their frequency of occurrence in semantic concordance texts.

synset_offset Byte offset in **data.***pos* file of a synset containing *lemma*. Each *synset_offset* in the list corresponds to a different sense of *lemma* in WordNet. *synset_offset* is an 8 digit, zero-filled decimal integer that can be used with **fseek**(3) to read a synset from the data file. When passed to **read_synset**(3WN) along with the syntactic category, a data structure containing the parsed synset is returned.

Data File Format

Each data file begins with several lines containing a copyright notice, version number and license agreement. These lines all begin with two spaces and the line number. All other lines are in the following format. Integer fields are of fixed length, and are zero-filled.

synset_offset lex_filenum ss_type w_cnt word lex_id [word lex_id...] p_cnt [ptr...] [frames...] | *gloss*

synset_offset Current byte offset in the file represented as an 8 digit decimal integer.

lex_filenum Two digit decimal integer corresponding to the lexicographer file name containing the synset. See **lexnames**(5WN) for the list of filenames and their corresponding numbers.

ss_type One character code indicating the synset type:

n	NOUN
v	VERB
a	ADJECTIVE
s	ADJECTIVE SATELLITE
r	ADVERB

w_cnt Two digit hexadecimal integer indicating the number of words in the synset.

word ASCII form of a word as entered in the synset by the lexicographer, with spaces replaced by underscore characters (_). The text of the word is case sensitive, in contrast to its form in the corresponding **index.***pos* file, that contains only lower-case forms. In **data.adj**, a *word* is followed by a syntactic marker if one was specified in the lexicographer file. A syntactic marker is appended, in parentheses, onto *word* without any intervening spaces. See **wninput**(5WN) for a list of the syntactic markers for adjectives.

lex_id One digit hexadecimal integer that, when appended onto *lemma*, uniquely identifies a sense within a lexicographer file. *lex_id* numbers usually start with **0**, and are incremented as additional senses of the word are added to the same file, although there is no requirement that the numbers be consecutive or begin with **0**. Note that a value of **0** is the default, and therefore is not present in lexicographer files.

p_cnt Three digit decimal integer indicating the number of pointers from this synset to other synsets. If *p_cnt* is **000** the synset has no pointers.

ptr A pointer from this synset to another. *ptr* is of the form:

pointer_symbol synset_offset pos source/target

where *synset_offset* is the byte offset of the target synset in the data file corresponding to *pos*.

The *source/target* field distinguishes lexical and semantic pointers. It is a four byte field, containing two two-digit hexadecimal integers. The first two digits indicates the word number in the current (source) synset, the last two digits indicate the word number in the target synset. A value of **0000** means that *pointer_symbol* represents a semantic relation between the current (source) synset and the target synset indicated by *synset_offset*.

A lexical relation between two words in different synsets is represented by non-zero values in the source and target word numbers. The first and last two bytes of this field indicate the word numbers in the source and target synsets, respectively, between which the relation holds. Word numbers are assigned to the *word* fields in a synset, from left to right, beginning with **1**.

See **wninput**(5WN) for a list of *pointer_symbol*s, and semantic and lexical pointer classifications.

frames In **data.verb** only, a list of numbers corresponding to the generic verb sentence frames for *word*s in the synset. *frames* is of the form:

f_cnt + f_num w_num [+ f_num w_num...]

where *f_cnt* a two digit decimal integer indicating the number of generic frames listed, *f_num* is a two digit decimal integer frame number, and *w_num* is a two digit hexadecimal integer indicating the word in the synset that the frame applies to. As with pointers, if this number is **00**, *f_num* applies to all *word*s in the synset. If non-zero, it is applicable only to the word indicated. Word numbers are assigned as described for pointers. Each *f_num w_num* pair is preceded by a **+**. See **wninput**(5WN) for the text of the generic sentence frames.

gloss Each synset contains a gloss. A *gloss* is represented as a vertical bar (**|**), followed by a text string that continues until the end of the line. The gloss may contain a definition, one or more example sentences, or both.

Sense Numbers

Senses in WordNet are generally ordered from most to least frequently used, with the most common sense numbered **1**. Frequency of use is determined by the number of times a sense is tagged in the various semantic concordance texts. Senses that are not semantically tagged follow the ordered senses. The *tagsense_cnt* field for each entry in the **index.***pos* files indicates how many of the senses in the list have been tagged.

The **cntlist**(5WN) file provided with the database lists the number of times each sense is tagged in the semantic concordances. The data from **cntlist** is used by **grind**(1WN) to order the senses of each word. When the **index**.*pos* files are generated, the *synset_offset*s are output in sense number order, with sense 1 first in the list. Senses with the same number of semantic tags are assigned unique but consecutive sense numbers. The WordNet **OVERVIEW** search displays all senses of the specified word, in all syntactic categories, and indicates which of the senses are represented in the semantically tagged texts.

Exception List File Format

Exception lists are alphabetized lists of inflected forms of words and their base forms. The first field of each line is an inflected form, followed by a space separated list of one or more base forms of the word. There is one exception list file for each syntactic category.

Note that the noun and verb exception lists were automatically generated from a machine-readable dictionary, and contain many words that are not in WordNet. Also, for many of the inflected forms, base forms could be easily derived using the standard rules of detachment programmed into Morphy (See **morph**(7WN)). These anomalies are allowed to remain in the exception list files, as they do no harm.

Verb Example Sentences

For some verb senses, example sentences illustrating the use of the verb sense can be displayed. Each line of the file **sentidx.vrb** contains a *sense_key* followed by a space and a comma separated list of example sentence template numbers, in decimal. The file **sents.vrb** lists all of the example sentence templates. Each line begins with the template number followed by a space. The rest of the line is the text of a template example sentence, with **%s** used as a placeholder in the text for the verb. Both files are sorted alphabetically so that the *sense_key* and template sentence number can be used as indices, via **binsrch**(3WN), into the appropriate file.

When a request for **FRAMES** is made, the WordNet search code looks for the sense in **sentidx.vrb**. If found, the sentence template(s) listed is retrieved from **sents.vrb**, and the **%s** is replaced with the verb. If the sense is not found, the applicable generic sentence frame(s) listed in *frames* is displayed.

NOTES

Information in the **data.***pos* and **index.***pos* files represents all of the word senses and synsets in the WordNet database. The *word*, *lex_id*, and *lex_filenum* fields together uniquely identify each word sense in WordNet. These can be encoded in a *sense_key* as described in **senseidx**(5WN). Each synset in the database can be uniquely identified by combining the *synset_offset* for the synset with a code for the syntactic category (since it is possible for synsets in different **data.***pos* files to have the same *synset_offset*).

The WordNet system provide both command line and window-based browser interfaces to the database. Both interfaces utilize a common library of search and morphology code. The source code for the library and interfaces is included in the WordNet package. See **wnintro**(3WN) for an overview of the WordNet source code.

ENVIRONMENT VARIABLES

WNHOME Base directory for WordNet. Unix default is **/usr/local/WordNet-1.7.1**, Windows default is **C:\Program Files\WordNet\1.7.1**.

WNSEARCHDIR Directory in which the WordNet database has been installed. Unix default is **WNHOME/dict**, Windows default is **WNHOME\dict**.

FILES

All files are in the directory **WNSEARCHDIR**.

index.*pos* database index files (Unix)

*pos***.idx** database index files (Windows)

data.*pos* database data files (Unix)

pos.**dat**	database data files (Windows)
∗**.vrb**	files of sentences illustrating the use of verbs
pos.**exc**	morphology exception lists

SEE ALSO

grind(1WN), **wn**(1WN), **wnb**(1WN), **wnintro**(3WN), **binsrch**(3WN), **wnintro**(5WN), **cntlist**(5WN), **lexnames**(5WN), **senseidx**(5WN), **wninput**(5WN), **morphy**(7WN), **wngloss**(7WN), **wngroups**(7WN), **wnstats**(7WN).