NAME

grind − process WordNet lexicographer files

SYNOPSIS

grind [**−v**] [**−s**] [**−L***logfile*] [**−a**] [**−d**] [**−i**] [**−o**] [**−n**] *filename* [*filename...*]

DESCRIPTION

grind( ) processes WordNet lexicographer files, producing database files suitable for use with the WordNet search and interface code and other applications. The syntactic and structural integrity of the input files is verified. Warnings and errors are reported via **stderr** and a run-time log is produced on **stdout**. A database is generated only if there are no errors.

Input Files

Input files correspond to the syntactic categories implemented in WordNet − **noun**, **verb**, **adjective** and **adverb**. Each input lexicographer file consists of a list of synonym sets (*synsets*) for one part of speech. Although the basic synset syntax is the same for all of the parts of speech, some parts of the syntax only apply to a particular part of speech. See **wninput**(5WN) for a description of the input file format.

Each *filename* specified is of the form:

pathname/*pos***.***suffix*

where *pathname* is optional and *pos* is either **noun**, **verb**, **adj** or **adv**. *suffix* may be used to separate groups of synsets into different files, for example **noun.animal** and **noun.plant**. One or more input files, in any combination of syntactic categories, may be specified. See **lexnames**(5WN) for a list of the lexicographer files used to build the complete WordNet database.

Output Files

grind( ) produces the following output files:

Filename	**Description**
index.*pos*	Index file for each syntactic category
data.*pos*	Data file for each syntactic category
index.sense	Sense index

See **wndb**(5WN) for a description of the database file formats.

Each time **grind( )** is run, any existing database files are overwritten with the database files generated from the specified input files. If no input files from a syntactic category are specified, the corresponding database files are not overwritten.

Sense Numbers

Senses 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 in an arbitrary order. Note that this ordering is only an estimate based on usage in a small corpus.

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** file provided with the database lists the number of times each sense is tagged in the semantic concordances. **grind( )** uses the data from **cntlist** 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.

OPTIONS

−v	Verify integrity of input without generating database.
−s	Suppress generation of warning messages. Usually **grind** is run with this option until all syntactic and structural errors are corrected since the warning messages may make it difficult to spot error messages.
−L*logfile*	Write all messages to *logfile* instead of **stderr**.
−a	Generate statistical report on input files processed.
−d	Generate distribution of senses by string length report on input files processed.
−i	Generate sense index file.
−o	Order senses using **cntlist**.
−n	Generate nominalization (derivational morphology) links in database.
filename	Input file of the form described in **Input Files.**

FILES

*pos***.***	lexicographer files to use to build database
cntlist	file of combined semantic concordance **cntlist** files. Used to assign sense numbers in WordNet database

SEE ALSO

cntlist(5WN), **lexnames**(5WN), **senseidx**(5WN), **wndb**(5WN), **wninput**(5WN), **uniqbeg**(7WN), **wngloss**(7WN).

DIAGNOSTICS

Exit status is normally 0. Exit status is -1 if non-specific error occurs. If syntactic or structural errors exist, exit status is number of errors detected.

usage: grind [−v] [−s] [−Llogfile] [−a] [−d] [−i] [−o] [−n] filename [filename...]
Invalid options were specified on the command line.

No input files processed.
None of the filenames specified were of the appropriate form.

n **syntactic errors found.**
Syntax errors were found while parsing the input files.

n **structural errors found.**
Pointer errors were found that could not be automatically corrected.

BUGS

Please report bugs to **wordnet@princeton.edu**.