

Manual of

Oak System

(version 0.1)

May. 15. 2002

Satoshi SEKINE

Computer Science Department
New York University

715 Broadway, 7th floor
New York NY 10003 USA

sekine@cs.nyu.edu
http://cs.nyu.edu/~sekine

LICENSE AGREEMENT

A nonexclusive license to use this software and its documentation for scholarly and research purposes only is hereby granted by New York University. There is no fee to use this software for these purposes. In using the software and documentation for scholarly and research purposes you may copy or modify them but (a) you must include the NYU copyright notice and the original authors' names on each copy or modification you make of the software and of the supporting documentation; and (b) you may not (i) distribute the software or the documentation (or any part of either) to anyone else, or (ii) use, rewrite, or adapt the software or the documentation (or any part of either) as the basis for any commercial software or hardware product, without in each instance obtaining the prior written consent of NYU or an appropriate license from NYU. You may not use this software and its documentation (or any part of either) for any other purpose without obtaining an appropriate license from NYU. To obtain any license or consent from NYU, please contact: Patrick Franc, Office of Industrial Liaison, New York University, 251 Mercer Street, New York, NY 10012.

Copyright 2002 by New York University. All rights reserved.

Contents

1	Introduction	4
2	Instration	5
2.1	Get the source and defreeze it	5
2.2	Directories and files	5
2.3	How to make it	6
2.4	How to run it	8
2.5	Examples of run	9
3	The things you must know	12
3.1	How to set the parameters	12
3.2	How to specify the process	12
3.3	How to specify the format	13
4	The things you may or may not want to know	14
4.1	Level	14
4.2	Format	21
5	The things most of you may NOT want to know	29
5.1	Dictionary	29
5.2	POStagger Rules	29
5.3	Chunking Rules	29
5.4	NE Hierarchy/Rules/Dictionary	29
A	List of parameters	30
A.1	User parameters	30
A.2	System parameters	32
B	List of commandline options	35
C	150 NE deifnition	36

1 Introduction

There have been many NLP tools for various languages. Many of them are very good and useful. However, I encountered a problem to combine them. For example, when I want to tag the SGML text with Named entities in parse trees, I wrote 4 perl scripts to transfer the data format, 2 perl scripts to modify the strings, and 2 scripts to combine different files. This is not productive, at all in the research if it takes more than 20% of the project time to write the perl scripts. This looks worse when you find the output formats of different programs are slightly different.

Based on this experience, I decided to make one single program which does it all. (The prototype of this program was called "*DoItAll*") In the program design, however, I take special care for not making large process if one needs only a part of this program. Yes, you have to download the entire files even if you want to use the sentence splitter of the program, but when you run it, the program loads only the necessary knowledge and the process size should be minimal.

So the **OAK** system is like a filter. It is a filter between text, splitted sentence, tokenized sentence, POS tagged sentence, chunked sentence, NE tagged sentence, dependency analyzed sentence and so on. The user can specify what level of input and output. Also, as the **OAK** system support many format, like Penn Treebank POS tagged format, bracket format, input format for Collin's parser, and so on, it is a filter between different format at different (or the same) level.

The author would like to thank many people who helped to make this tool. In particular, Professor Ralph Grishman gave me good comment and patients until I started trying to make the first cut of the program today. Also, I would like to tank colleagues of mine who encouraged to make this possible. Many requests from NLP researchers around world to make the system public after they look at the project homepage were also encouraging. I would like to express sincerery thanks to all of them.

2 Instration

2.1 Get the source and defreeze it

It depends on your computer environment (OS, ftp version etc), but you can, at least, get the latest information at the following URL. Or you may get the file from a CDROM.

```
http://www.cs.nyu.edu/cs/projects/teus/oak/
```

The downloaded file should be a tar + gzip file. Please make a directory for this (Let's call it `oak_dir` now on), and defreeze it by the following commands.

```
LINUX> cd oak_dir
LINUX> gzip -dc oak0_1.tgz | tar xvf -

OR

LINUX> tar xvzf - oak0_1.tgz
```

2.2 Directories and files

What you will get should be the following directories and files.

```

LINUX> cd oak_dir
LINUX> ls -F
data/ demo/ doc/ src/

LINUX> ls -F src/
Makefile      functagger.c  ne.c          print.o
analyze.c     functagger.o ne.o          read.c
analyze.o     function.h    oak*          read.o
chunker.c     global.h     oak.linux_pentium* splitter.c
chunker.o     load.c       oak.prm       splitter.h
default.h     load.o       oak.solaris* splitter.o
dep.c         macro.h      param.c       struct.h
dep.o         main.c       param.h       tokenizer.c
dictionary.c  main.o       param.o       tokenizer.h
dictionary.o  mainloop.c  postagger.c   tokenizer.o
eval.c        mainloop.o  postagger.o   util.c
eval.o        match.c     print.c       util.o
extern.h      match.o     print.h       var.h

LINUX> ls -F doc/
manual.aux  manual.log  manual.tex~
manual.dvi  manual.tex  manual.toc

LINUX> ls -F data/
NED_TOP/   WS002.htb  WS008.fnc  WS021.neh
WS002.chk  WS002.pos  WS021.dic  WS021.ner
WS002.chq  WS003.dep/ WS021.ned

```

2.3 How to make it

If you are using Linux on a Pentium machine or Solaris OS, and you are *VERY lucky*, the executable file at the bin/ directory may work. If you think you are one of those, please check it out, first!

```
LINUX> cd bin  
LINUX> oak
```

OR

```
SOLARIS> cd bin  
SOLARIS> mv oak.solaris oak  
SOLARIS> oak
```

Otherwise (if it does not work or you are using other OS or machine,) you have to make it by the following procedure.

```

LINUX> cd src
LINUX> make clean
/bin/rm -f main.o mainloop.o param.o util.o load.o dictionary.o re
ad.o print.o eval.o analyze.o match.o splitter.o tokenizer.o posta
gger.o chunker.o ne.o dep.o functagger.o oak *~
LINUX> make
gcc -Wall -g      -c -o main.o main.c
gcc -Wall -g      -c -o mainloop.o mainloop.c
gcc -Wall -g      -c -o param.o param.c
gcc -Wall -g      -c -o util.o util.c
gcc -Wall -g      -c -o load.o load.c
gcc -Wall -g      -c -o dictionary.o dictionary.c
gcc -Wall -g      -c -o read.o read.c
gcc -Wall -g      -c -o print.o print.c
gcc -Wall -g      -c -o eval.o eval.c
gcc -Wall -g      -c -o analyze.o analyze.c
gcc -Wall -g      -c -o match.o match.c
gcc -Wall -g      -c -o splitter.o splitter.c
gcc -Wall -g      -c -o tokenizer.o tokenizer.c
gcc -Wall -g      -c -o postagger.o postagger.c
gcc -Wall -g      -c -o chunker.o chunker.c
gcc -Wall -g      -c -o ne.o ne.c
gcc -Wall -g      -c -o dep.o dep.c
gcc -Wall -g      -c -o functagger.o functagger.c
gcc -Wall -g -o oak main.o mainloop.o param.o util.o load.o dictio
nary.o read.o print.o eval.o analyze.o match.o splitter.o tokenize
r.o postagger.o chunker.o ne.o dep.o functagger.o -lm -lc

```

2.4 How to run it

The default is set to be a chunker and NE tagger for normal sentence input (one sentence per a line). If you would like to try it, just try the following and type in whatever the sentence you like.

```

LINUX> cd src
LINUX> oak

```

2.5 Examples of run

In this subsection, some examples of oak runs will be explained. Please try some of them and learn how to use it for other purpose.

Sentence Splitter

```
LINUX> cat text
Pierre Vinken, 61 years old, will join the board as a
nonexecutive director Nov. 29. Mr. Vinken is chairman
of Elsevier N.V., the Dutch publishing group. Rudolph
Agnew, 55 years old and former chairman of Consolidated
Gold Fields PLC, was named a nonexecutive director of
this British industrial conglomerate.

LINUX> ../src/oak -i TEXT -s TEXT -o SENTENCE -O PLAIN -r text
Oak System (0.1)      May.15.2002   Satoshi Sekine (NYU)
-----
Loading Dictionary ... done
-----
Pierre Vinken, 61 years old, will join the board as a nonexecutiv
e director Nov. 29.
Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing gro
up.
Rudolph Agnew, 55 years old and former chairman of Consolidated G
old Fields PLC, was named a nonexecutive director of this British
industrial conglomerate.
```

Tokenizer

```
LINUX> oak -i SENTENCE -o TOKENIZED
Oak System (0.1)      May.15.2002   Satoshi Sekine (NYU)
-----
Loading Dictionary ... done
-----
> "I'm a boy."
" I 'm a boy . "
```

Stemmer

```
LINUX>oak -i SENTENCE -o POSTAG -O STEM
Oak System (0.1)      May.15.2002  Satoshi Sekine (NYU)
-----
Loading Dictionary ... done
Loading POS tagger rule ...done
-----
> Tables aren't broken.
table be not break .
```

POS tagger

```
LINUX> oak -i SENTENCE -o POSTAG -O PTB_TAG
Oak System (0.1)      May.15.2002  Satoshi Sekine (NYU)
-----
Loading Dictionary ... done
Loading POS tagger rule ...done
-----
> Prof. Sekine promised to create this program by December 2001.
Prof./NNP Sekine/NNP promised/VBD to/TO create/VB this/DT
program/NN by/IN December/NNP 2001/CD ./.
```

NE tagger

```
LINUX> oak -i SENTENCE -o NE -O MUC
Oak System (0.1)      May.15.2002   Satoshi Sekine (NYU)
-----
Loading Dictionary ... done
...
Loading NE rule ...done
-----
> Prof. Sekine promised to create this program by December 2001.
Prof. <ENAMEX TYPE=PERSON>Sekine</ENAMEX> promised to create this
program by <TIMEX TYPE=DATE>December 2001</TIMEX>.
```

Chunker

```
LINUX> oak -i SENTENCE -o CHUNK -O CONLL
Oak System (0.1)      May.15.2002   Satoshi Sekine (NYU)
-----
Loading Dictionary ... done
Loading POS tagger rule ...done
Loading chunker quadgram ...done
Loading chunker rule ...done
-----
> Prof. Sekine promised to create this program by December 2001.
Prof. NNP B-NP
Sekine NNP I-NP
promised VBD B-VP
to TO I-VP
create VB I-VP
this DT B-NP
program NN I-NP
by IN B-PP
December NNP B-NP
2001 CD I-NP
. . 0
```

3 The things you must know

3.1 How to set the parameters

Oak provides two methods to set the parameters.

1. Parameter file (default *oak.prm* or specify by `-p` option)
2. Command line options

If you set a parameter in both methods, the setting by command line option will be taken. This is the case even you set the parameter file name at the end of the command line options. The list of parameters you can specify is listed in the last section, and most important parameters will be explained in the next sections.

3.2 How to specify the process

The type of process to run is specified by **input level**, **start level** and **output level**.

- **input level**: level of input file
- **start level**: level of the process to start (i.e. you can start the process with the level more primitive than the input level in order to evaluate the system.)
- **output level**: level of output

For example, if **input level** and **start level** is SENTENCE and **output level** is POSTAG, then it runs like a POS tagger. The list of **level** is shown in Table 1. The level can be specified by

- command line options
e.g. `oak -i SENTENCE -o POSTAG` for POS tagger, or `oak -i TEXT -o NE` for NE tagger for non-sentence splitted text. You can use start level for the evaluation of the system; `oak -i POSTAG -s SENTENCE -o POSTAG` for POS tagger evaluation.
- OR in parameter file
e.g. Write `INPUT_LEVEL SENTENCE` or `OUTPUT_LEVEL POSTAG` for POS tagger.

If you don't specify any, it takes the default process, currently converting sentence to chunking and NE tagged sentence ¹

¹This may change in the future, you can see such information by running the system with `-H` option.

Level	Explanation
TEXT	Text
SENTENCE	Splitted sentences
TOKENIZED	Tokenized sentences
POSTAG	Each token is assigned by POS tag also with stemming
CHUNK	No recursive constituents are marked
NE	Named Entities are marked
CHUNK_NE	Both chunks and NE's are marked
DEPENDENCY	Dependencies between chunks are indicated (not yet implemented)
PARSE	Parse trees (not yet implemented)
FUNCTAGS	Parse trees with function tags (not yet implemented)
REGULARIZED	Regularized structures (not yet implemented)

Table 1: List of Level

3.3 How to specify the format

The format of input and output is specified by **input format** and **output format**. For example, **input format** can be PLAIN for the **input level** of TEXT, SENTENCE or TOKENIZED. The **output format** can be, for example, PTB_BRACKET, and PTB_TAG, MUC, TIPSTER or SGML for the **output level** of NE. The list of **level** is shown in Table 2. The detail of each format can be found in the next section.

Format	Explanation
PLAIN	Plain input without any meta information
PTB_BRACKET	Penn Treebank's cmb format
PTB_TAG	Penn Treebank's tag format
STEM	Stemmed sentence, each token is seperated by a space
STEM_TAG	Token, POS tag and stem are seperated by "/"
TIPSTER	Tipster architecture format (not yet implemented)
SGML	SGML format (not yet implemented)
TABLE	(not yet implemented)
COLLINS	Input format for the Collin's parser (POSTAG only)
CONLL	CONLL format (CHUNK, NE only)
MUC	MUC format (NE only)
DETAIL	Detail format. You can specify what to display in parameter file.

Table 2: List of Format

4 The things you may or may not want to know

In this section, each level and format will be explained in detail.

4.1 Level

Text

This is a plain text file. ASCII characters only, but it is not necessarily separated (line segmented) for each sentence. If there is an empty line, it indicates that the sentences before and the after the empty line are not in the same sentence. The input has to be a file and stdin is not implemented.

Sentence

One sentence per a line. Each sentence is separated by a “n” character.

Tokenized

Each token is separated by a space character.

POS tagged

Part-of-speech is tagged for each token. The definition of Part-of-speech is the same as the definition in Penn Treebank, shown in Table 3.

Chunk

A chunk is a non-recursive constituent in a sentence. In other words, it is a constituent which does not have any sub constituents in it. The kinds of constituents are the same as those defined at CONLL workshop on Chunking. There are 11 kinds of chunk labels as shown in Table refChunkdefinition. In the many occasion, we use BIO notation. B-label means that the word is the first

The Penn Treebank POS tagset

1.	CC	Coordinating conjunction	25.	TO	<i>to</i>
2.	CD	Cardinal number	26.	UH	Interjection
3.	DT	Determiner	27.	VB	Verb, base form
4.	EX	Existential <i>there</i>	28.	VBD	Verb, past tense
5.	FW	Foreign word	29.	VBG	Verb, gerund/present participle
6.	IN	Preposition/subord. conjunction	30.	VBN	Verb, past participle
7.	JJ	Adjective	31.	VBP	Verb, non-3rd ps. sing. present
8.	JJR	Adjective, comparative	32.	VBZ	Verb, 3rd ps. sing. present
9.	JJS	Adjective, superlative	33.	WDT	<i>wh</i> -determiner
10.	LS	List item marker	34.	WP	<i>wh</i> -pronoun
11.	MD	Modal	35.	WP\$	Possessive <i>wh</i> -pronoun
12.	NN	Noun, singular or mass	36.	WRB	<i>wh</i> -adverb
13.	NNS	Noun, plural	37.	#	Pound sign
14.	NNP	Proper noun, singular	38.	\$	Dollar sign
15.	NNPS	Proper noun, plural	39.	.	Sentence-final punctuation
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Colon, semi-colon
18.	PRP	Personal pronoun	42.	(Left bracket character
19.	PP\$	Possessive pronoun	43.)	Right bracket character
20.	RB	Adverb	44.	"	Straight double quote
21.	RBR	Adverb, comparative	45.	'	Left open single quote
22.	RBS	Adverb, superlative	46.	"	Left open double quote
23.	RP	Particle	47.	'	Right close single quote
24.	SYM	Symbol (mathematical or scientific)	48.	"	Right close double quote

Table 3: Definition of Part-of-speech

NP	Noun phrase
VP	Verb phrase
ADVP	Adverb phrase
SBAR	Clause introduced by a subordinating conjunction
ADJP	Adjective phrase
PRT	Particle phrase
CONJP	Conjunctive phrase
INTJ	Interjection phrase
LST	List marker phrase
UCP	Unlike Coordinated Phrase
O	Other (not in a chunk)

Table 4: Chunk definition

word of an chunk of label. I-label means that the word is second or later word if an chunk of label if the previous labels are B-label followed by one or more I-label. 0 is outside of any of chunks.

NE

This is Named Entity. It is initially defined at MUC, including 7 kinds of NE, organization, location, person, date, time, percent and money expressions. In the system, you can make your own NE definition (see following section for the detail), but the prepared knowledge includes 150 kinds of NE's (Shown in Appendix C). If you set @ALL at NE_LABEL in the parameter file, the system tries to tag all possible NE's. If you just want to have subset of the 150 NE's, you can do it by setting NE_LABEL in the parameter file. You can learn how to set the parameter by comparing three parameter files at the src directory, oak_allne.prm, oak_muc.prm and oak_ace.prm. Examples can be found in the followings.

```

LINUX> grep NE_LABEL oak_neall.prm | grep -v #
NE_LABEL                @ALL
LINUX> oak -p oak_neall.prm
Oak System (0.1)        May.15.2002   Satoshi Sekine (NYU)
-----
Loading Dictionary ... done
...
Loading NE rule ...done
-----
> New York University is in New York today.
<PI TYPE=SCHOOL>New York University</PI> is in <PI TYPE=CITY>New Y
ork</PI> <PI TYPE=DATE>today</PI>.

```

```

LINUX> grep NE_LABEL oak_muc.prm | grep -v #
NE_LABEL  ORG          "<ENAMEX TYPE=ORGANIZATION>" "</ENAMEX>"
NE_LABEL  LOC          "<ENAMEX TYPE=LOCATION>"   "</ENAMEX>"
NE_LABEL  GPE          "<ENAMEX TYPE=LOCATION>"   "</ENAMEX>"
NE_LABEL  PERSON       "<ENAMEX TYPE=PERSON>"     "</ENAMEX>"
NE_LABEL  FACILITY     "<ENAMEX TYPE=ORGANIZATION>" "</ENAMEX>"
NE_LABEL  DATE         "<TIMEX TYPE=DATE>"      "</TIMEX>"
NE_LABEL  TIME         "<TIMEX TYPE=TIME>"      "</TIMEX>"
NE_LABEL  PERCENT      "<NUMEX TYPE=PERCENT>"    "</NUMEX>"
NE_LABEL  MONEY        "<NUMEX TYPE=MONEY>"      "</NUMEX>"
LINUX> oak -p oak_muc.prm
Oak System (0.1)        May.15.2002   Satoshi Sekine (NYU)
-----
Loading Dictionary ... done
...
Loading NE rule ...done
-----
> New York University is in New York today.
<ENAMEX TYPE=ORGANIZATION>New York University</ENAMEX> is in <ENAM
EX TYPE=LOCATION>New York</ENAMEX> <TIMEX TYPE=DATE>today</TIMEX>.

```

```

LINUX> grep NE_LABEL oak_ace.prm | grep -v #
NE_LABEL  ORG      "<ENAMEX TYPE=ORGANIZATION>" "</ENAMEX>"
NE_LABEL  LOC      "<ENAMEX TYPE=LOCATION>"      "</ENAMEX>"
NE_LABEL  GPE      "<ENAMEX TYPE=GPE>"        "</ENAMEX>"
NE_LABEL  PERSON   "<ENAMEX TYPE=PERSON>"      "</ENAMEX>"
NE_LABEL  FACILITY "<ENAMEX TYPE=FACILITY>"   "</ENAMEX>"
LINUX> oak -p oak_ace.prm
Oak System (0.1)      May.15.2002   Satoshi Sekine (NYU)
-----
Loading Dictionary ... done
...
Loading NE rule ...done
-----
> New York University is in New York today.
<ENAMEX TYPE=FACILITY>New York University</ENAMEX> is in <ENAMEX T
YPE=GPE>New York</ENAMEX> today.

```

Chunk and NE

It has both chunk and NE information.

Dependency

(Not yet implemented)

Parse

(Not yet implemented)

The definition of nonterminal symbols are the same as that of Penn Treebank, shown in Table 5.

Function tag

(Not yet implemented)

S	Simple declarative clause, i.e. one that is not introduced by a subordinating conjunction or wh-word and that does not exhibit subject-verb inversion.
SBAR	Clause introduced by a (possibly empty) subordinating conjunction.
SBARQ	Direct question introduced by a wh-word or wh-phrase. Indirect questions and relative clauses should be bracketed as SBAR, not SBARQ.
SINV	Inverted declarative sentence, i.e. one in which the subject follows the tensed verb or modal.
SQ	Inverted yes/no question, or main clause of a wh-question, following the wh-phrase in SBARQ.
ADJP	Adjective Phrase. Phrasal category headed by an adjective (including comparative and superlative adjectives).
ADVP	Adverb Phrase. Phrasal category headed by an adverb (including comparative and superlative adverbs).
CONJP	Conjunction Phrase. Used to mark certain “multi-word” conjunctions, such as <i>as well as</i> , <i>instead of</i> .
FRAG	Fragment.
INTJ	Interjection. Corresponds approximately to the part-of-speech tag UH.
LST	List marker. Includes surrounding punctuation.
NAC	Not A Constituent; used to show the scope of certain prenominal modifiers within a noun phrase.
NP	Noun Phrase. Phrasal category that includes all constituents that depend on a head noun.
NX	Used within certain complex noun phrases to mark the head of the noun phrase. Corresponds very roughly to N-bar level but used quite differently.
PP	Prepositional Phrase. Phrasal category headed by a preposition.
PRN	Parenthetical.
PRT	Particle.
QP	Quantifier Phrase (i.e., complex measure/amount phrase); used within NP.
RRC	Reduced Relative Clause.
UCP	Unlike Coordinated Phrase.
VP	Verb Phrase. Phrasal category headed a verb.
WHADJP	Wh-adjective Phrase. Adjectival phrase containing a wh-adverb
WHADVP	Wh-adverb Phrase. Introduces a clause with an ADVP gap. May be null or lexical, containing a wh-adverb such as <i>how</i> or <i>why</i> .
WHNP	Wh-noun Phrase. Introduces a clause with an NP gap. May be null or lexical, containing some wh-word, e.g. <i>who</i> , <i>which book</i> , <i>whose daughter</i> , <i>none of which</i> , or <i>how many leopards</i> .
WHPP	Wh-prepositional Phrase. Prepositional phrase containing a wh-noun phrase (such as <i>of which</i> or <i>by whose authority</i>) that either introduces a PP gap or is contained by a WHNP.
X	Unknown, uncertain, or unbracketable. X is often used for bracketing typos and in bracketing <i>the...the</i> -constructions.

Table 5: Nonterminating symbol definition

Reguralized

(Not yet implemented)

4.2 Format

PLAIN (TEXT)

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29. Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group. Rudolph Agnew, 55 years old and former chairman of Consolidated Gold Fields PLC, was named a nonexecutive director of this British industrial conglomerate.

PLAIN (SENTENCE)

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.
Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.
Rudolph Agnew, 55 years old and former chairman of Consolidated Gold Fields PLC, was named a nonexecutive director of this British industrial conglomerate.

PLAIN (TOKENIZED)

Pierre Vinken , 61 years old , will join the board as a nonexecutive director Nov. 29 .
Mr. Vinken is chairman of Elsevier N.V. , the Dutch publishing group .
Rudolph Agnew , 55 years old and former chairman of Consolidated Gold Fields PLC , was named a nonexecutive director of this British industrial conglomerate .

PTB_TAG (POSTAG)

Pierre/NNP Vinken/NNP ,/, 61/CD years/NNS old/JJ ,/, will/MD join/
VB the/DT board/NN as/IN a/DT nonexecutive/JJ director/NN Nov./NNP
29/CD ./.

Mr./NNP Vinken/NNP is/VBZ chairman/NN of/IN Elsevier/NNP N.V./NNP
./, the/DT Dutch/JJ publishing/NN group/NN ./.

Rudolph/NNP Agnew/NNP ,/, 55/CD years/NNS old/JJ and/CC former/JJ
chairman/NN of/IN Consolidated/NNP Gold/NNP Fields/NNP PLC/NNP ,/,
was/VBD named/VBN a/DT nonexecutive/JJ director/NN of/IN this/DT
British/JJ industrial/JJ conglomerate/NN ./.

PTB_TAG (CHUNK)

[NP Pierre/NNP Vinken/NNP] ,/, [NP 61/CD years/NNS] [ADJP old/JJ
] ,/, [VP will/MD join/VB] [NP the/DT board/NN] [PP as/IN] [NP
a/DT nonexecutive/JJ director/NN] [NP Nov./NNP 29/CD] ./.

[NP Mr./NNP Vinken/NNP] [VP is/VBZ] [NP chairman/NN] [PP of/IN
] [NP Elsevier/NNP N.V./NNP] ,/, [NP the/DT Dutch/JJ publishing/N
N group/NN] ./.

[NP Rudolph/NNP Agnew/NNP] ,/, [NP 55/CD years/NNS old/JJ] and/C
C [ADJP former/JJ] [NP chairman/NN] [PP of/IN] [NP Consolidated
/NNP Gold/NNP Fields/NNP PLC/NNP] ,/, [VP was/VBD named/VBN] [NP
a/DT nonexecutive/JJ director/NN] [PP of/IN] [NP this/DT Britis
h/JJ industrial/JJ conglomerate/NN] ./.

PTB_TAG (NE)

<PERSON Pierre/NNP > Vinken/NNP ,/, 61/CD years/NNS old/JJ ,/, will/MD join/VB the/DT board/NN as/IN a/DT nonexecutive/JJ director/NN <DATE Nov./NNP 29/CD > ./.

Mr./NNP <PERSON Vinken/NNP > is/VBZ chairman/NN of/IN Elsevier/NNP N.V./NNP ,/, the/DT <PERSON Dutch/JJ > publishing/NN group/NN ./.

<PERSON Rudolph/NNP Agnew/NNP > ,/, 55/CD years/NNS old/JJ and/CC former/JJ chairman/NN of/IN Consolidated/NNP Gold/NNP Fields/NNP <ORG PLC/NNP > ,/, was/VBD named/VBN a/DT nonexecutive/JJ director/NN of/IN this/DT <GPE British/JJ > industrial/JJ conglomerate/NN ./.

PTB_TAG (CHUNK_NE)

[NP <PERSON Pierre/NNP > Vinken/NNP] ,/, [NP 61/CD years/NNS] [ADJP old/JJ] ,/, [VP will/MD join/VB] [NP the/DT board/NN] [PP as/IN] [NP a/DT nonexecutive/JJ director/NN] [NP <DATE Nov./NNP 29/CD >] ./.

[NP Mr./NNP <PERSON Vinken/NNP >] [VP is/VBZ] [NP chairman/NN] [PP of/IN] [NP Elsevier/NNP N.V./NNP] ,/, [NP the/DT <PERSON Dutch/JJ > publishing/NN group/NN] ./.

[NP <PERSON Rudolph/NNP Agnew/NNP >] ,/, [NP 55/CD years/NNS old/JJ] and/CC [ADJP former/JJ] [NP chairman/NN] [PP of/IN] [NP Consolidated/NNP Gold/NNP Fields/NNP <ORG PLC/NNP >] ,/, [VP was/VBD named/VBN] [NP a/DT nonexecutive/JJ director/NN] [PP of/IN] [NP this/DT <GPE British/JJ > industrial/JJ conglomerate/NN] ./.

PTB_BRACKET (POSTAG)

(TOP (NNP Pierre) (NNP Vinken) (, ,) (CD 61) (NNS years) (JJ old)
(, ,) (MD will) (VB join) (DT the) (NN board) (IN as) (DT a) (JJ nonexecutive) (NN director) (NNP Nov.) (CD 29) (. .))
(TOP (NNP Mr.) (NNP Vinken) (VBZ is) (NN chairman) (IN of) (NNP Elsevier) (NNP N.V.) (, ,) (DT the) (JJ Dutch) (NN publishing) (NN group) (. .))
(TOP (NNP Rudolph) (NNP Agnew) (, ,) (CD 55) (NNS years) (JJ old)
(CC and) (JJ former) (NN chairman) (IN of) (NNP Consolidated) (NNP Gold) (NNP Fields) (NNP PLC) (, ,) (VBD was) (VBN named) (DT a) (JJ nonexecutive) (NN director) (IN of) (DT this) (JJ British) (JJ industrial) (NN conglomerate) (. .))

PTB_BRACKET (CHUNK)

(TOP (NP (NNP Pierre) (NNP Vinken)) (, ,) (NP (CD 61) (NNS years)
) (ADJP (JJ old)) (, ,) (VP (MD will) (VB join)) (NP (DT the) (NN board)) (PP (IN as)) (NP (DT a) (JJ nonexecutive) (NN director)) (NP (NNP Nov.) (CD 29)) (. .))
(TOP (NP (NNP Mr.) (NNP Vinken)) (VP (VBZ is)) (NP (NN chairman)) (PP (IN of)) (NP (NNP Elsevier) (NNP N.V.)) (, ,) (NP (DT the)) (JJ Dutch) (NN publishing) (NN group)) (. .))
(TOP (NP (NNP Rudolph) (NNP Agnew)) (, ,) (NP (CD 55) (NNS years)
(JJ old)) (CC and) (ADJP (JJ former)) (NP (NN chairman)) (PP (IN of)) (NP (NNP Consolidated) (NNP Gold) (NNP Fields) (NNP PLC)) (, ,) (VP (VBD was) (VBN named)) (NP (DT a) (JJ nonexecutive) (NN director)) (PP (IN of)) (NP (DT this) (JJ British) (JJ industrial) (NN conglomerate)) (. .))

PTB_BRACKET (NE)

(TOP (PERSON (NNP Pierre)) (NNP Vinken) (, ,) (CD 61) (NNS years)
(JJ old) (, ,) (MD will) (VB join) (DT the) (NN board) (IN as) (D
T a) (JJ nonexecutive) (NN director) (DATE (NNP Nov.) (CD 29)) (.
.))
(TOP (NNP Mr.) (PERSON (NNP Vinken)) (VBZ is) (NN chairman) (IN o
f) (NNP Elsevier) (NNP N.V.) (, ,) (DT the) (PERSON (JJ Dutch)) (N
N publishing) (NN group) (. .))
(TOP (PERSON (NNP Rudolph) (NNP Agnew)) (, ,) (CD 55) (NNS years)
(JJ old) (CC and) (JJ former) (NN chairman) (IN of) (NNP Consolid
ated) (NNP Gold) (NNP Fields) (ORG (NNP PLC)) (, ,) (VBD was) (VB
N named) (DT a) (JJ nonexecutive) (NN director) (IN of) (DT this)
(GPE (JJ British)) (JJ industrial) (NN conglomerate) (. .))

STEM (POSTAG)

Pierre Vinken , 61 year old , will join the board as a nonexecutive
director Nov. 29 .
Mr. Vinken be chairman of Elsevier N.V. , the Dutch publishing group .
Rudolph Agnew , 55 year old and former chairman of Consolidated Gold
Fields PLC , be name a nonexecutive director of this British industrial
conglomerate .

STEM_TAG (POSTAG)

Pierre/NNP/Pierre Vinken/NNP/Vinken ,/,/, 61/CD/61 years/NNS/year old/JJ/old ,/,/, will/MD/will join/VB/join the/DT/the board/NN/boa
rd as/IN/as a/DT/a nonexecutive/JJ/nonexecutive director/NN/direct
or Nov./NNP/Nov. 29/CD/29 ././.

Mr./NNP/Mr. Vinken/NNP/Vinken is/VBZ/be chairman/NN/chairman of/IN
/of Elsevier/NNP/Elsevier N.V./NNP/N.V. ,/,/, the/DT/the Dutch/JJ/
Dutch publishing/NN/publishing group/NN/group ././.

Rudolph/NNP/Rudolph Agnew/NNP/Agnew ,/,/, 55/CD/55 years/NNS/year
old/JJ/old and/CC/and former/JJ/former chairman/NN/chairman of/IN/
of Consolidated/NNP/Consolidated Gold/NNP/Gold Fields/NNP/Fields P
LC/NNP/PLC ,/,/, was/VBD/be named/VBN/name a/DT/a nonexecutive/JJ/
nonexecutive director/NN/director of/IN/of this/DT/this British/JJ/
/British industrial/JJ/industrial conglomerate/NN/conglomerate ./.
/.

COLLINS (PSTAG)

18 Pierre NNP Vinken NNP , , 61 CD years NNS old JJ , , will MD jo
in VB the DT board NN as IN a DT nonexecutive JJ director NN Nov.
NNP 29 CD . .

13 Mr. NNP Vinken NNP is VBZ chairman NN of IN Elsevier NNP N.V. N
NP , , the DT Dutch JJ publishing NN group NN . .

26 Rudolph NNP Agnew NNP , , 55 CD years NNS old JJ and CC former
JJ chairman NN of IN Consolidated NNP Gold NNP Fields NNP PLC NNP
 , , was VBD named VBN a DT nonexecutive JJ director NN of IN this
DT British JJ industrial JJ conglomerate NN . .

CONLL (CHUNK)

```
Pierre NNP B-NP
Vinken NNP I-NP
, , 0
61 CD B-NP
years NNS I-NP
old JJ B-ADJP
, , 0
will MD B-VP
join VB I-VP
the DT B-NP
board NN I-NP
as IN B-PP
a DT B-NP
nonexecutive JJ I-NP
director NN I-NP
Nov. NNP B-NP
29 CD I-NP
. . 0

Mr. NNP B-NP
Vinken NNP I-NP
is VBZ B-VP
```

MUC (NE)

```
<ENAMEX TYPE=PERSON>Pierre</ENAMEX> Vinken, 61 years old, will jo
in the board as a nonexecutive director <TIMEX TYPE=DATE>Nov. 29<
/TIMEX>.
Mr. <ENAMEX TYPE=PERSON>Vinken</ENAMEX> is chairman of Elsevier N
.V., the <ENAMEX TYPE=PERSON>Dutch</ENAMEX> publishing group.
<ENAMEX TYPE=PERSON>Rudolph Agnew</ENAMEX>, 55 years old and form
er chairman of Consolidated Gold Fields <ENAMEX TYPE=ORGANIZATION
>PLC</ENAMEX>, was named a nonexecutive director of this <ENAMEX
TYPE=GPE>British</ENAMEX> industrial conglomerate.
```

DETAIL

This is output example of DETAIL fromat. Here input level is SENTENCE and the output level is CHUNK_NE. The flags in the parameter files for the following variables are set on (1). PD_SENTENCE, PD_WORD, PD_WORDADHERE, PD_WORDSTARTOFFSET, PD_WORDENDOFFSET, PS_WORDSTEM, PD_WORDPOS, PD_WORDCHUNK. Each line in the list has elements corresponding to the variables which were set to 1.

```
> I'm in New York now.
Sentence = I'm in New York now.
1  I          0  0  0  I          PRP  B-NP
2  'm        1  1  2  be        VBP  B-VP
3  in        0  4  5  in        IN   B-PP
4  New       0  7  9  New       NNP  B-NP
5  York      0  11 14  York      NNP  I-NP
6  now      0  16 18  now       RB   B-ADVP
7  .        1  19 19  .         .    0
```

5 The things most of you may NOT want to know

This section will be completed later.

5.1 Dictionary

5.2 POStagger Rules

5.3 Chunking Rules

5.4 NE Hierarchy/Rules/Dictionary

A List of parameters

A.1 User parameters

In this subsection, the user parameters, for which a normal user may want to modify to make the system what you want, will be explained. If there is a bracket at the each item header (like (-o)), it is the corresponding command line option.

- `DICTIONARY_FILENAME`
- `POSTAGGER_FILENAME`
- `CHUNKERQUAD_FILENAME`
- `CHUNKER_FILENAME`
- `NEHIERARCHY_FILENAME`
- `NEDICT_FILENAME`
- `NE_FILENAME`
- `DEPENDENCY_FILENAME`
- `HEADTABLE_FILENAME`
- `FUNCTAGS_FILENAME`
File names of knowledge files. The format of files are specified somewhere in this manual.
- `INPUT_LEVEL (-i)`
- `START_LEVEL (-s)`
- `OUTPUT_LEVEL (-o)`
The levels of the process are defined. Each of these should be one of the followings. `TEXT`, `SENTENCE`, `TOKENIZED`, `POSTAG`, `NE`, `CHUNK`, `CHUNK_NE`, `DEPENDENCY`, `PARSE`, `FUNCTAGS`, `REGULARIZED`.
- `INPUT_FORMAT (-I)`
- `OUTPUT_FORMAT (-O)`
The formats of input and output are specified. Each of these should be one of the following. `PLAIN`, `PTB_BRACKET`, `PTB_TAG`, `STEM`, `STEM_TAG`, `TIPSTER`, `SGML`, `TABLE`, `DETAIL`, `COLLINS (POSTAG)`, `CONLL (CHUNK)`, `MUC (NE)`.

- **DOCUMENT_TYPE**
Document type. If the input is written material, then it should be WSJ. If the input is transcribed material, in particular the one used in ACE project, then it should be set to ACE. In other cases it is better to set WSJ, which is default. If you set ACE, the sentence splitter and the tokenizer will act differently.
- **PROMPT**
The prompt string. The default if "> ".
- **WARNING**
Warning flag. (0: off, 1: on).
- **META_INPUT**
It specifies if the system recognize meta input, starting with character "*", like "*help", "*reset". (0: off, 1: on).
- **USE_TOKEN_BORDER**
Not yet implemented.
- **PRINT_EMPTY**
Not yet implemented.
- **PRINT_FUNC_TAGS**
Source to print function tags (0: off, 1: on).
- **PD_INPUT, PD_INPUTDETAIL, PD_SENTENCE, PD_SENTENCEOFFSET, PD_SENTENCEBORDER, PD_SENTENCENE, PD_SENTENCEDetail, PD_WORD, PD_WORDID, PD_WORDORIGINAL, PD_WORDDETAIL, PD_WORDBOSEOS, PD_WORDWID, PD_WORDADHERE, PD_WORDSTARTOFFSET, PD_WORDENDOFFSET, PD_WORDPOS, PD_WORDGOLDPOS, PD_WORDCHUNK, PD_WORDGOLDCHUNK, PD_WORDNE, PD_WORDGOLDNE, PD_WORDCHUNKID, PD_WORDNODEID, PD_WORDDICTIONARY, PD_WORDPOS_RULE_LIST, PD_WORDCHUNK_RULE_LIST, PD_WORDNE_RULE_LIST, PD_CHUNK, PD_CHUNKHEAD, PD_CHUNKHEAD_WORD, PD_CHUNKLABEL, PD_CHUNKSTART, PD_CHUNKEND, PD_CHUNKGOLD_HEAD, PD_CHUNKFLAG, PD_CHUNKWORD, PD_NODE**
Specify print out information, when `OUTPUT_FORMAT = DETAIL`. (0: off, 1: on).
- **WORD_STRING_CONVERT**
Word conversion for print out. The token specified at `source_string` will be printed out by `target_string`. This may be used when the user want to print special string for special symbols (like "right round bracket" etc).

Format: `WORD_STRING_CONVERT source_string target_string`
Example: `WORD_STRING_CONVERT (-LRB-`

- EVAL_POS, EVAL_CHUNK, EVAL_NE, EVAL_DEPENDENCY, EVAL_FUNCTAGS
Run evaluation process.
- ANALYZE_POS, ANALYZE_CHUNK, ANALYZE_NE, ANALYZE_FUNCTAGS
Not yet implemented.
- ANALYZE_POS_FILENAME, ANALYZE_CHUNK_FILENAME, ANALYZE_NE_FILENAME,
ANALYZE_FUNCTAGS_FILENAME
Not yet implemented.

A.2 System parameters

If you want to change the system dictionary, rules, etc, you might want to change parameters explained in this subsection. You need extra-caution if you are going to delete or change some of them. The system might not work without some of the parameters.

- WORD_UD, WORD_UK, WORD_BOS, WORD_EOS, WORD_EMPTY
Word label for 5 special words.
- CAT_LABEL_UD, CAT_LABEL_UK, CAT_LABEL_BOS, CAT_LABEL_EOS,
CAT_LABEL_EMPTY, CAT_LABEL_TOP
Category label (non-terminal syntactic label) for 6 special categories.
- CAT_LABEL
Define category labels.
- CAT_UNKNOWN_DEFAULT
Default category for unknown words. There is a unknown word category guessing process, but this is for the cases where even the process can't guess the category.
- CHUNK_LABEL_UD, CHUNK_LABEL_UK, CHUNK_LABEL_BOS, CHUNK_LABEL_EOS
Chunk label for 4 special chunks
- CHUNK_LABEL_DEFAULT
Default chunk label
- CHUNK_LABEL
Define chunk labels.
- NE_LABEL_UD, NE_LABEL_UK, NE_LABEL_BOS
- NE_LABEL_EOS, NE_LABEL_TOP
NE labels for 5 special NE

- **NE_LABEL_DEFAULT**
Default NE label.
- **NE_LABEL**
Define NE labels to be printed out. Labels starting @ are special case. @ALL means all NE defined in .neh file. Otherwise specify the label with the begining and ending SGML tags for MUC format output.

Format: NE_LABEL @label
NE_LABEL category "MUC-start-tag" "MUC-end-tag"
Example: NE_LABEL @ALL
NE_LABEL ORG "<ENAMEX TYPE=ORGANIZATION>" "</ENAMEX>"
- **MUC_READ_START_LABEL, MUC_READ_END_LABEL**
Start and end tag of MUC document to be read.
- **FUNCTAGS_LABEL_UD, FUNCTAGS_LABEL_UK**
FuncTags label for 2 special tags
- **FUNCTAGS_LABEL**
Define function tags.
- **CLASS_LABEL_UD**
Define class label for special 9 classes
- **TOK_CONCAT_LAST_PERIOD**
Tokenizer processing parameter If 1, the last period ending word will be a word i.e transform U.S.A. at the end of a sentence to U.S.A. (period for the acronym) rather than U.S.A and . (period for the sentence ending, which is default).
- **TREAT_FIRST_WORD**
- **FIRST_WORD_WEIGHT**
- **OPEN_QUOTE_WORD**
- **CLOSE_QUOTE_WORD**
- **TREAT_UNKNOWN_WORD**
- **POSTAGGER_ITEM_MATCH_FORCUS**
- **POSTAGGER_RELATIVE_FREQ_THRESHOLD**
- **POSTAGGER_RELATIVE_FREQ_THRESHOLD2**
POS processing parameter.

- **CHUNKER_ITEM_MATCH_FORCUS**
Chunker processing parameter.
- **NE_DYNAMIC_DICT_FLAG**
NE processing parameter.
 - 0: no use of dynamic dictionary
 - 1: Use dynamic dictionary, but ordinal dictionary is stronger
 - 2: Absolute power for dynamic dictionary
- **NE_DELETE_ISOLATED_CAPITAL**
NE processing parameter. If 1, NE tagged token(s) which is a part of capitalized word sequence, delete the NE tag
- **HEADTABLE_DIR_R_TO_L, HEADTABLE_DIR_L_TO_R**
Define HeadTable direction string.
- **DEP_DEFAULT_METHOD**
Dependency processing parameter.
 - 1: All chunks' head is the previous chunk
 - 2: All chunks' head is the final VP
 - 3: All chunks' head is prev/following towards the first VP
- **MATCH_ONCE_FUNCTAGS**
Function tagger processing parameter.

B List of commandline options

- **-h** : help
- **-p filename** : Parameter file
- **-i LEVEL** : Input level
- **-s LEVEL** : Start level
- **-o LEVEL** : Output level
- **-I FORMAT** : Input format
- **-O FORMAT** : Output format
- **-r filename** : Input file
- **-w filename** : Output file
- **-b** : Batch mode (no prompt)

C 150 NE deifinition

TOP

NAME

```
PERSON          # Bill Clinton, George W. Bush, Satoshi Sekine,
LASTNAME        # Clinton, Bush, Sekine,
MALE_FIRSTNAME  # Bill, George, Satoshi,
FEMALE_FIRSTNAME # Mary, Catherine, Ilene, Yoko

ORGANIZATION    # United Nations, NATO
COMPANY          # IBM, Microsoft
COMPANY_GROUP   # Star Alliance, Tokyo-Mitsubishi Group
MILITARY        # The U.S Navy
INSTITUTE       # the National Football League, ACL
MARKET          # New York Stock Exchange, NASDAQ
POLITICAL_ORGANIZATION #
GOVERNMENT      # Department of Education, Ministry of Finance
POLITICAL_PARTY # Republican Party, Democratic Party, GOP
PUBLIC_INSTITUTION # New York Post Office,
GROUP           # The Beatles, Boston Symphony Orchestra
SPORTS_TEAM     # the Chicago Bulls, New York Mets
ETHNIC_GROUP    # Han race, Hispanic
NATIONALITY     # American, Japanese, Spanish

LOCATION         # Times Square, Ground Zero
GPE            # Asia, Middle East, Palestine
CITY           # New York City, Los Angeles
COUNTY        # Westchester
PROVINCE       # State (US), Province (Canada), Prefecture (Japan)
COUNTRY        # the United States of America, Japan, England
REGION         # Scandinavia, North America, Asia, East coast
GEOLOGICAL_REGION # Altamira
LANDFORM       # Rocky Mountains, Manzano Peak, Matterhorn
WATER_FORM     # Hudson River, Fletcher Pond
SEA            # Pacific Ocean, Gulf of Mexico, Florida Bay
ASTRAL_BODY    # Halley's comet, the Moon
STAR           # Sirius, Sun, Cassiopeia, Centaurus
PLANET         # the Earth, Mars, Venus
ADDRESS        #
POSTAL_ADDRESS # 715 Broadway, New York, NY 10003
PHONE_NUMBER   # 212-123-4567
EMAIL         # sekine@cs.nyu.edu
URL           # http://www.cs.nyu/cs/projects/proteus
```

FACILITY	# Empire State Building, Hunter Mountain Ski Resort
GOE	# Pentagon, White House, NYU Hospital
SCHOOL	# New York University, Edgewood Elementary School
MUSEUM	# MOMA, the Metropolitan Museum of Art
AMUSEMENT_PARK	# Walt Disney World, Oakland Zoo
WORSHIP_PLACE	# Canterbury Cathedral, Westminster Abbey
STATION_TOP	#
AIRPORT	# JFK Airport, Narita Airport, Changi Airport
STATION	# Grand Central Station, London Victoria Station
PORT	# Port of New York, Sydney Harbour
CAR_STOP	# Port Authority Bus Terminal, Sydney Bus Depot
LINE	# Westchester Bicycle Road
RAILROAD	# Metro-North Harlem Line, New Jersey Transit
ROAD	# Lexington Avenue, 42nd Street
WATERWAY	# Suez Canal, Bering Strait
TUNNEL	# Euro Tunnel
BRIDGE	# Golden Gate Bridge, Manhattan Bridge
PARK	# Central Park, Hyde Park
MONUMENT	# Statue of Liberty, Brandenburg Gate
PRODUCT	# Windows 2000, Rosetta Stone
VEHICLE	# Vespa ET2, Honda Elite 50s
CAR	# Ford Escort, Audi 90, Saab 900, Civic, BMW 318i
TRAIN	# Acela, TGV, Bullet Train
AIRCRAFT	# F-14 Tomcat, DC-10, B-747
SPACESHIP	# Sputnik, Apollo 11, Space Shuttle Challenger, Mir
SHIP	# Titanic, Queen Elizabeth II, U.S.S. Enterprise
DRUG	# Pedialyte, Tylenol, Bufferin
WEAPON	# Patriot Missile, Pulser P-138
STOCK	# NABISCO stock
CURRENCY	# Euro, yen, dollar, peso,
AWARD	# Nobel Peace Prize, Pulitzer Prize
THEORY	# Newton's law, GB theory, Blum's Theory
RULE	# Kyoto Global Warming Pact, The U.S. Constitution
SERVICE	# Pan Am Flight 103, Acela Express 2190
CHARACTER	# Pikachu, Mickey Mouse, Snoopy
METHOD_SYSTEM	# New Deal program, Federal Tax
ACTION_MOVEMENT	# The U.N. Peace-keeping Operation
PLAN	# Manhattan Project, Star Wars Plan
ACADEMIC	# Sociology, Physics, Philosophy
CATEGORY	# Bantam Weight, 48kg class
SPORTS	# Men's 100 meter, Giant Slalom, ski, tennis
OFFENCE	# first-degree murder
ART	# Venus of Melos

PICTURE	# Night Watch, Monariza, Guernica
BROADCAST_PROGRAM	# Larry King Live, The Simpsons, ER, Friends
MOVIE	# E.T., Batman Forever, Jurassic Park, Star Wars
SHOW	# Les Miserables, Madam Butterfly
MUSIC	# The Star Spangled Banner, My Life, Your Song
PRINTING	# 2001 Consumer Survey
BOOK	# Master of the Game, 1001 Ways to Reward Employees
NEWSPAPER	# The New York Times, Wall Street Journal
MAGAZINE	# Newsweek, Time, National Business Employment Weekly
DISEASE	# AIDS, cancer, leukemia
EVENT	# Hanover Expo, Edinburgh Festival
GAMES	# Olympic, World Cup, PGA Championships
CONFERENCE	# APEC, Naples Summit
PHENOMENA	# El Nino
WAR	# World War II, Vietnam War, the Gulf War
NATURAL_DISASTER	# Kobe Earthquake, the Puu0o-Kupaianaha Eruption
CRIME	# Murder of Black Dahlia, the Oklahoma City bombing
TITLE	# Mr., Ms., Miss., Mrs,
POSITION_TITLE	# President, CEO, King, Prince, Prof., Dr.
LANGUAGE	# English, Spanish, Chinese, Greek
RELIGION	# Christianity, Islam, Buddhism
NATURAL_OBJECT	# mitochondria, shiitake mushroom
ANIMAL	# elephant, whale, pig, horse
VEGETABLE	# spinach, rice, daffodil
MINERAL	# Hydrogen, carbon monoxide,
COLOR	# black, white, red, blue
TIME_TOP	
TIMEX	
TIME	# 10 p.m., afternoon
DATE	# August 10, 2001, 10 Aug. 2001,
ERA	# Glacial period, Victorian age
PERIODX	# 2 semesters, summer vacation period
TIME_PERIOD	# 10 minutes, 15 hours, 50 hours
DATE_PERIOD	# 10 days, 50 days

WEEK_PERIOD	# 10 weeks, 50 weeks
MONTH_PERIOD	# 10 months, 50 months
YEAR_PERIOD	# 10 years, 50 years
NUMEX	# 100 pikel, 10 bits
MONEY	# \$10, 100 yen, 20 marks
STOCK_INDEX	# 26 5/8,
POINT	# 10 points
PERCENT	# 10%, 10 1/2%
MULTIPLICATION	# 10 times
FREQUENCY	# 10 times a day
RANK	# 1st prize, booby prize
AGE	# 36, 77 years old
MEASUREMENT	# 10 bytes, 10 Pa, 10 millibar
PHYSICAL_EXTENT	# 10 meters, 10 inches, 10 yards, 10 miles
SPACE	# 10 acres, 10 square feet,
VOLUME	# 10 cubic feet, 10 cubic yards
WEIGHT	# 10 milligrams, 10 ounces, 10 tons
SPEED	# 10 miles per hour, Mach 10
INTENSITY	# 10 lumina, 10 decibel
TEMPERATURE	# 60 degrees
CALORIE	# 10 calories
SEISMIC_INTENSITY	# 6.8 (on Richter scale)
COUNTX	
N_PERSON	# 10 biologists, 10 workers, 10 terrorists
N_ORGANIZATION	# 10 industry groups, 10 credit unions
N_LOCATION	# 10 cities, 10 areas, 10 regions, 10 states
N_COUNTRY	# 10 countries
N_FACILITY	# 10 buildings, 10 schools, 10 airports
N_PRODUCT	# 10 systems, 20 paintings, 10 supercomputers
N_EVENT	# 5 accidents, 5 interviews, 5 bankruptcies
N_ANIMAL	# 10 animals, 10 horses, 10 pigs
N_VEGETABLE	# 10 flowers, 10 daffodils
N_MINERAL	# 10 diamonds