



Lexical Tools for UMLS Developers

November 4, 2001

Allen C. Browne, Guy Divita,
Chris Lu

Lister Hill National Center for Biomedical Communications

National Library of Medicine

Metathesaurus Concept Search - Netscape

File Edit View Go Communicator Help



- Home
- Metathesaurus**
- Semantic Network
- SPECIALIST Lexicon
- Expert Search
- Download Results
- Comments
- Help

UMLS Knowledge Source Server

CONCEPT SEARCH

Term:

hand-foot-mouth diseases

Submit Reset

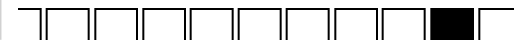
Restrict to:

ALL SOURCES

Matching criteria:

- Normalized string index
- Normalized word index
- Approximate matching
- Word index
- Left truncation
- Right truncation

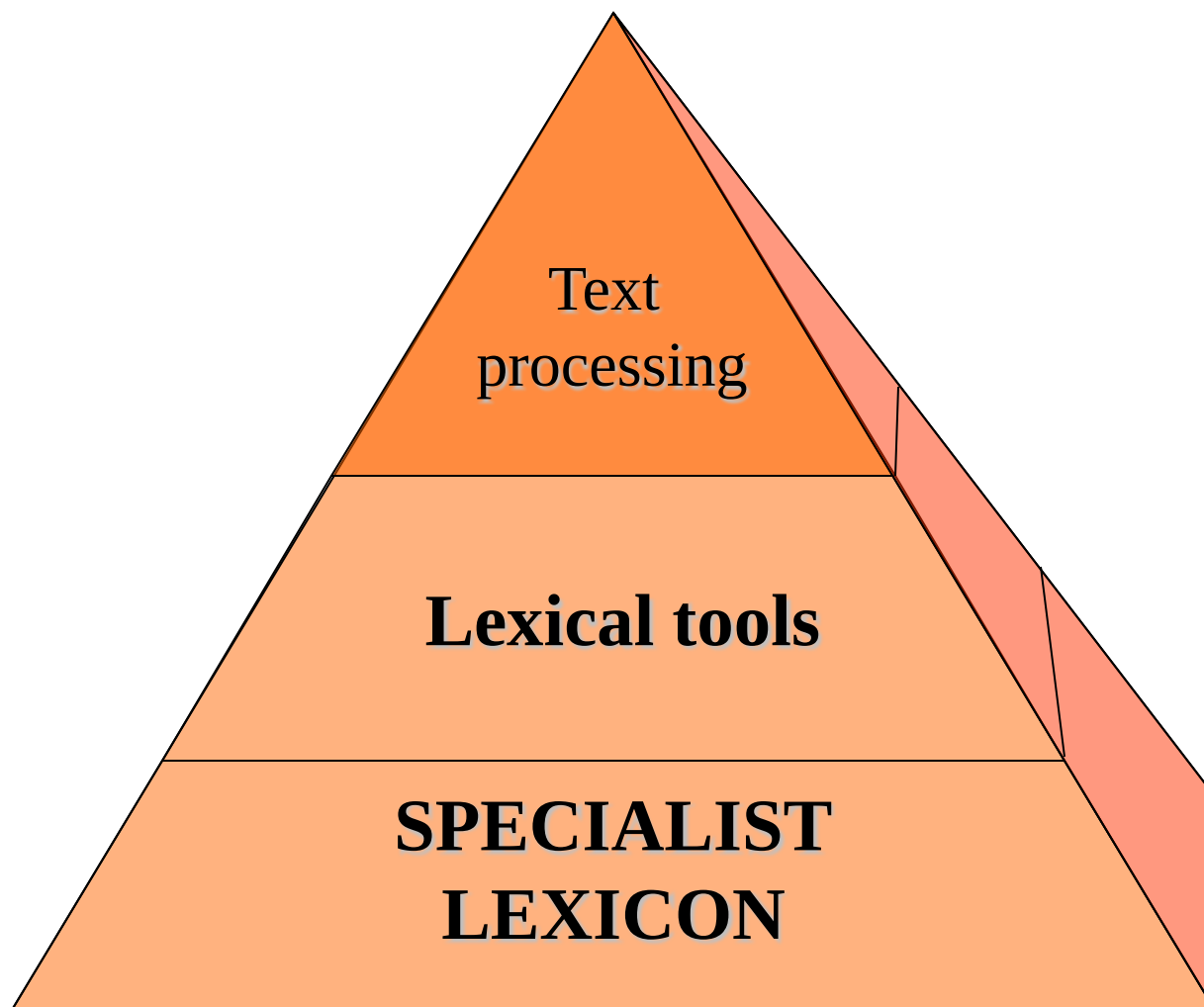
Document: D



Lexical Tools for UMLS Developers

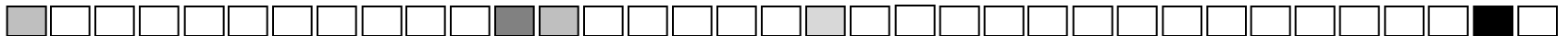
- The SPECIALIST lexicon – Browne
- The lexical tools – Divita/Lu
- Coffee Break– 10:00 - 10:30
- Lexical tools cont. – Divita/Lu





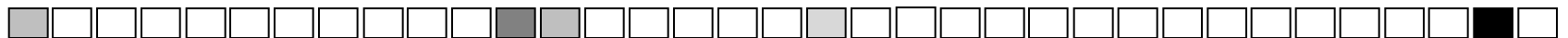
The SPECIALIST Lexicon

- A syntactic lexicon
- Biomedical and general English
- Over 160,000 records

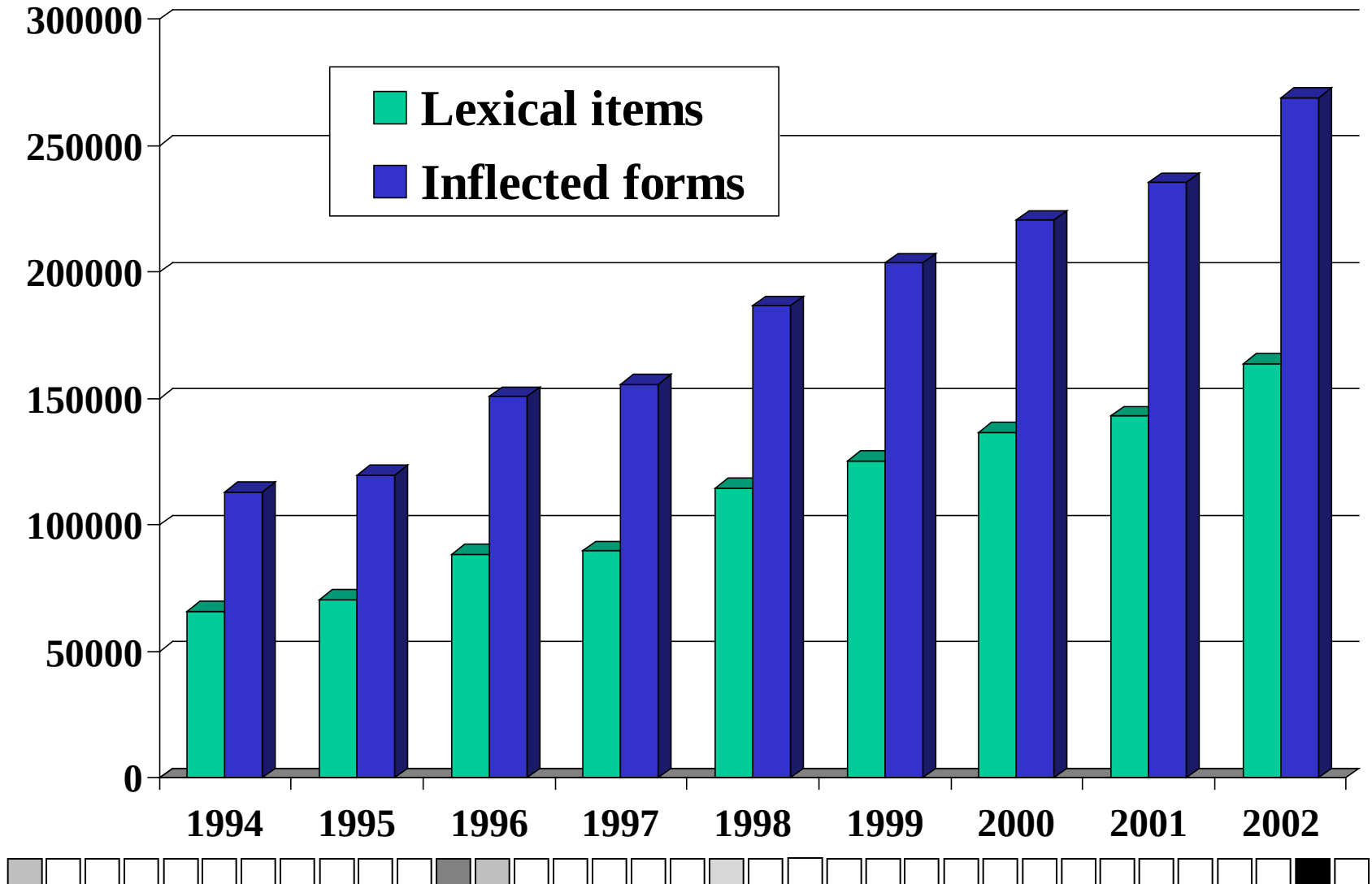


The SPECIALIST Lexicon

- General English:
- 10,000 most frequent words from the American Heritage word frequency list
- 2,000 words used by Longman's Dictionary of Contemporary English
- Verbs and adjectives identified by heuristics



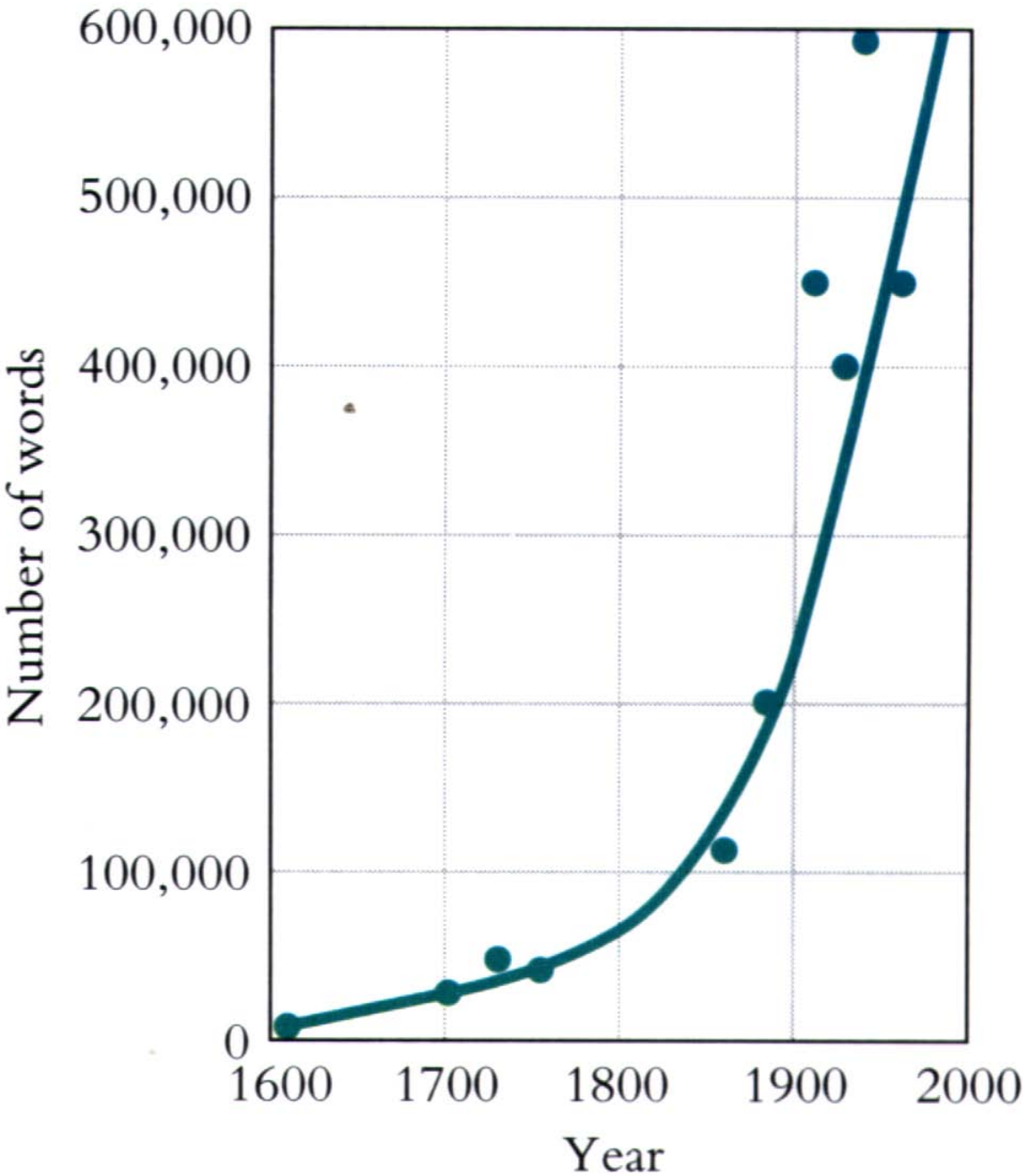
Lexicon Growth



George A.
Miller

The Science
of Words

1991



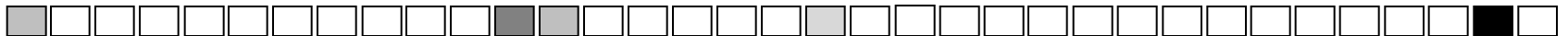
The SPECIALIST Lexicon

- Morphology
 - Inflection
 - Derivation
- Orthography
 - Spelling variants
- Syntax
 - Complementation for verbs, nouns, and adjectives



Morphology

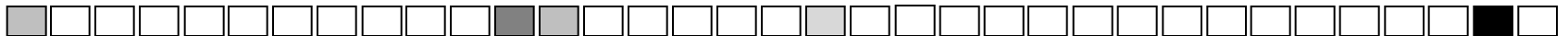
- Inflectional
 - nucleus -- nuclei
 - cauterize, cauterizes, cauterized, cauterizing
 - red, redder reddest
- Derivational
 - laryngeal -- larynx
 - transport -- transportation



Orthography

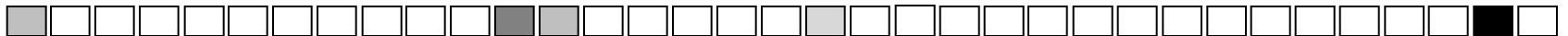
Spelling Variation

- align -- aline
- Grave's disease -- Graves's disease -- Graves' disease
- anesthetize -- anesthetise
- esophagus -- oesophagus



British and American Spelling

- Criticise -- criticize
- naturalise --naturalize
- centre -- center
- foetus -- fetus



Try This Test

Which of the variant spellings below do you, accept as standard American English? After making your choices, consult the pages of this book to see what the dictionaries say. You are due for some astonishments.

A Comparative Study of Spellings in four major collegiate dictionaries,

**by Lee C Deighton,
Hardscrabble Press 1972.**



Syntax -- Verb Complements

- Intran
 - I'll treat.
- tran=np
 - He treated the patient.
- ditran=np,pphr(with,np)
 - She treated the patient with the drug.

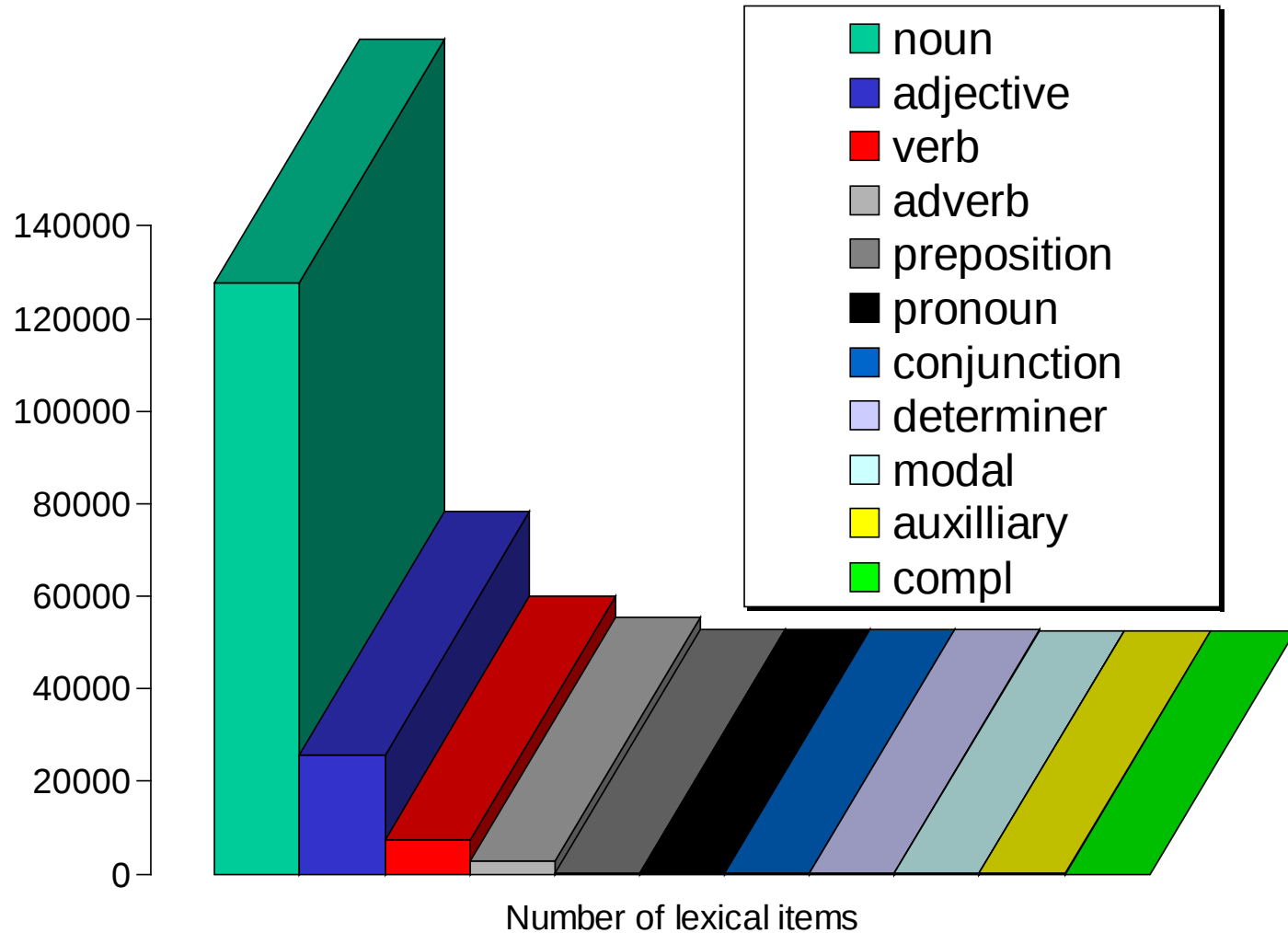


Syntax -- Verb Complements

```
{base=treat
entry=E0061964
    cat=verb
    variants=reg
    intran
    tran=np
    tran=pphr(with,np)
    tran=pphr(of,np)
    ditran=np,pphr(to,np)
    ditran=np,pphr(with,np)
    ditran=np,pphr(for,np)
    cplxtran=np,advbl
    nominalization=treatment|noun|E0061968
}
```



The 2002 SPECIALIST Lexicon



village
square the circle
square
square
fair and square root

Lexicon Unit Records

{**base**=Kaposi's sarcoma
spelling_variant=Kaposi sarcoma
entry=E0003576
 cat=noun
 variants=uncount
 variants=reg
 variants=glreg
}

{**base**=chronic
entry=E0016869
 cat=adj
 variants=inv
 position=attrib(1)
 position=pred
 stative
}

{**base**=aspirate
entry=E0010803
 cat=verb
 variants=reg
 tran=np
 nominalization=aspiration|noun|E0010804
}

{**base**=in
entry=E0033870
 cat=prep
}



Noun Variants

```
{base=Kaposi's sarcoma  
spelling_variant=Kaposi sarcoma  
entry=E0003576  
    cat=noun  
    variants=uncount  
    variants=reg  
    variants=glreg  
}
```

- Kaposi's sarcoma
- Kaposi's sarcomas
- Kaposi's sarcomata
- Kaposi sarcoma
- Kaposi sarcomas
- Kaposi sarcomata

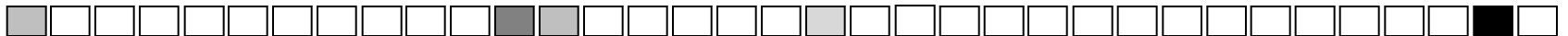


Regular Nouns

The plural suffix is *s*.

y becomes *ie* following a consonant before *s*.

e is inserted before *s* if the base ends in *s*, *z*, *x*, *ch*, or *s*



Regular Nouns

Base ends with	Plural ends with	Examples
Cy	Cies	fly: flies
-s	-ses	illness: illnesses
-z	-zes	waltz: waltzes
-x	-xes	box: boxes
-ch	-ches	match: matches
-sh	-shes	splash: splashes
X	Xs	book: books



Greco-latin Regular nouns

singular ends with:	plural ends with:	Examples
-us	-i	focus/foci
-ma	-mata	trauma/traumata
-a	-ae	larva/larvae
-um	-a	ilium/ilia
-on	-a	taxon/taxa
-sis	-ses	analysis/analyses
-is	-ides	cystis/cystides
-men	-mina	foramen/foramina
-ex	-ices	index/indices
-x	-ces	matrix/matrices



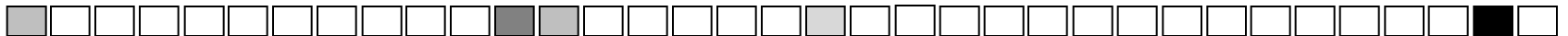
Uncount Nouns

(abstract or mass)

```
{base=smallpox  
entry=E0056359  
  cat=noun  
  variants=uncount  
}
```

```
{base=potassium  
entry=E0049387  
  cat=noun  
  variants=uncount  
}
```

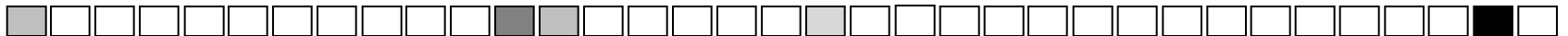
- * a smallpox
- * two smallpoxes
- much smallpox
- * a potassium
- * two potassiums
- much potassium



Fixed Plural Nouns

```
{base=police  
entry=E0048616  
  cat=noun  
  variants=plur  
}
```

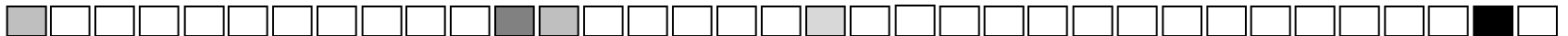
```
{base=scissors  
entry=E0054633  
  cat=noun  
  variants=plur  
}
```



Irregular Nouns

```
{base=corpus  
entry=E0019113  
  cat=noun  
  variants=irreg|corpora|  
  variants=reg  
}
```

```
{base=larynx  
entry=E0036919  
  cat=noun  
  variants=irreg|larynges|  
  variants=reg  
}
```



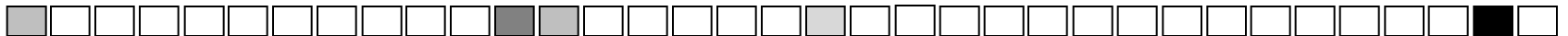
Regular Verbs

- The third person present tense suffix is *s*.
 - *y* becomes *ie* following a consonant before *s*.
 - *e* is inserted between *z*, *x*, *ch*, or *sh* and *s*.
- The past tense suffix is *ed*.
 - *y* becomes *ie* following a consonant before *ed*.
 - Final *e* is deleted before *ed*.



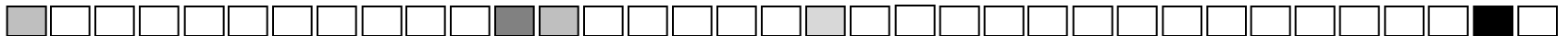
Regular Verbs

- dismiss:dismisses, dismissed, dismissing
- agree: agrees; agreed; agreeing
- dry: dries, dried, drying



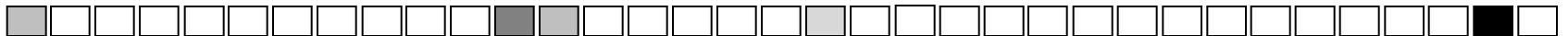
Regular Doubling Verbs

- End in a CVC pattern
- Double the final consonant before *ed* and *ing*.
- Are otherwise regular
- variants=regd
- e.g. control: controls, controlled, controlling

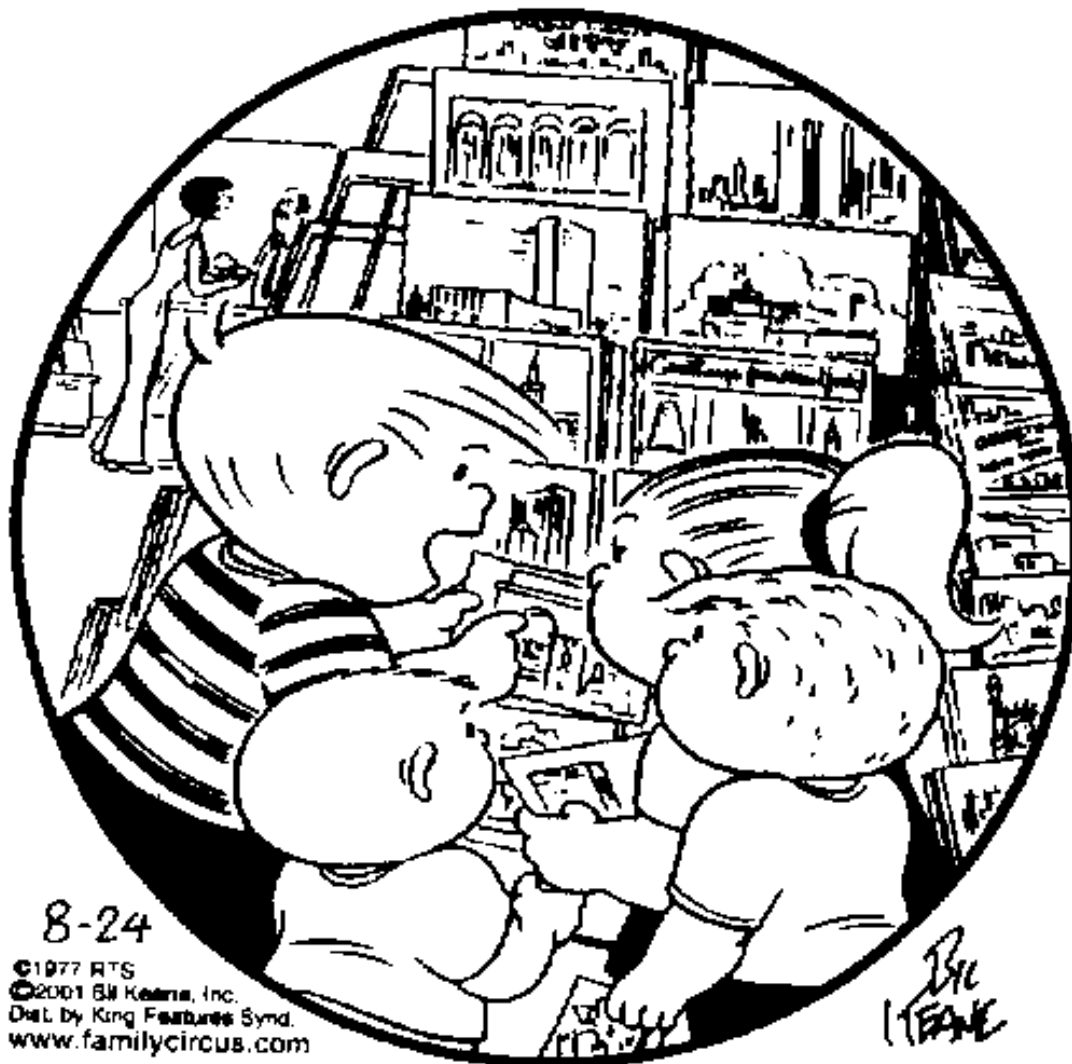


Irregular Verbs

```
{base=dive
cat=verb
  variants=reg
  variants=irreg|dives|dove|dove|diving|
  intran
  intran;part(in)
  ...
}
```



THE FAMILY CIRCUS BIL KEANE



8-24

©1977 RTS
©2001 Bil Keane, Inc.
Dist. by King Features Synd.
www.familycircus.com

"There's the place we sightsaw
yesterday."

Defective Verb

sightsee

*sightsees

*sightsaw

*sightseen

sightseeing



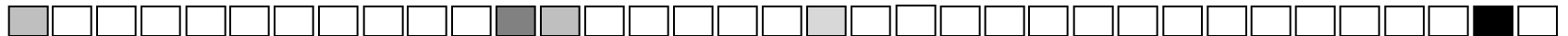
Regular Adjectives and Adverbs

- The comparative suffix is *er*.
- The superlative suffix is *est*.
 - *y* become *ie* after a consonant before *er* or *est*.
 - Final *e* is deleted before *er* or *est*.
- e.g. green: greener, greenest



Regular Doubling Adjectives and Adverbs

- CVC final pattern
- Final consonant is doubled before ed or est.
- Otherwise regular
- e.g. red: redder, reddest



Ancillary Data Bases

- Synonymy
 - sm.db
- Derivation
 - dm.db, dm.rules
- Inflection
 - im.rules
- Neoclassical compounds
 - nc.db



Derivational Facts and Rules

dm.facts

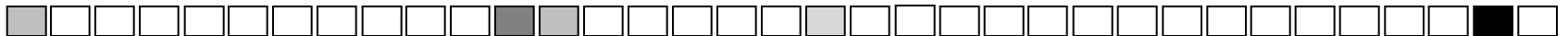
treatment|noun|treat|verb

prohibition|noun|prohibitive|adj

cell lineage|noun|cell line|noun

photochemotherapeutic|adj|photochemotherapy|noun

pharmacotherapeutic|adj|pharmacotherapy|noun



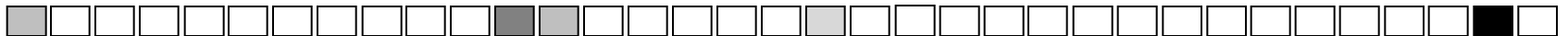
Derivational Facts and Rules

dm.rules

e.g. alienation|alienate

ation\$|noun|ate|verb

ration|rate; station|state;



Inflectional Facts and Rules

im.rules

Noun rules (ggreg)

us\$|noun|singular|i\$|noun|plural

antus|anti;

ma\$|noun|singular|mata\$|noun|plural

a\$|noun|singular|ae\$|noun|plural

um\$|noun|singular|a\$|noun|plural

on\$|noun|singular|a\$|noun|plural

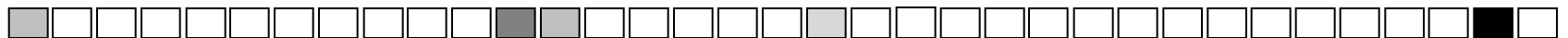
sis\$|noun|singular|ses\$|noun|plural

is\$|noun|singular|ides\$|noun|plural

men\$|noun|singular|mina\$|noun|plural

ex\$|noun|singular|ices\$|noun|plural

x\$|noun|singular|ces\$|noun|plural



Neoclassical compounds

nc.db

abdomin(o)|abdomen|root

ab|away from|prefix

acanth(o)|prickle|root

acar(o)|mite|root

acetabul(o)|acetabulum|root

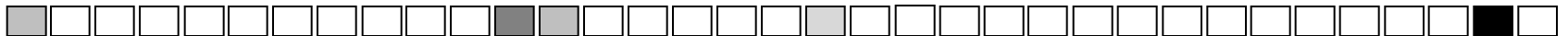
ad|towards|prefix

agogue|inducing|terminal

albumin(o)|albumin|root

sis|condition|terminal

stomy|surgical opening|terminal



Synonyms

sm.db

alar|adj|wing|noun

amygdaline|adj|tonsil|noun

articular|adj|joint|noun

bulbar|adj|medulla oblongata|noun

fununcular|adj|boil|noun

genicular|adj|knee|noun

hepatocellular|adj|liver cells|noun

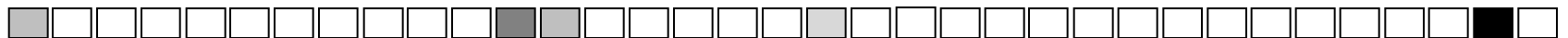
lazar|adj|leprosy|noun

lenticular|adj|crystalline lens|noun

ypsiliiform|adj|upsiloid|adj

wolfram|noun|tungsten|noun

double vision|noun|diplopia|noun



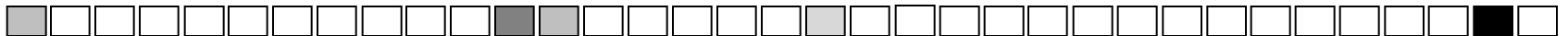
Relational Tables

- One line records
- Pipe separated Fields -- “|”
- Keyed to EUI
- LRAGR matches forms to EUIs
- Word index: LRWD



Relational Tables

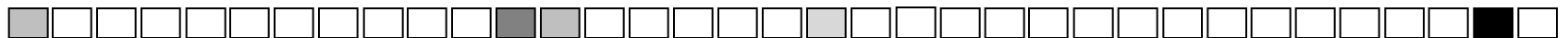
- LRAGR - Agreement
- LRCMP - Complements
- LRFIL - Files
- LRFLD - Fields
- LRMOD - Modification
- LRNOM - Nominalization
- LRPRN - Pronouns
- LRPRP - Properties
- LRSPL - Spelling
- LRTRM - Trademarks
- LRWD - Word index



LRAGR

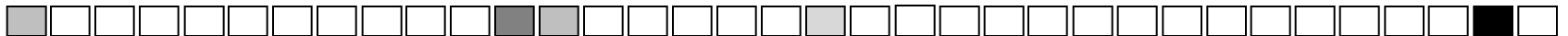
Agreement and Inflection

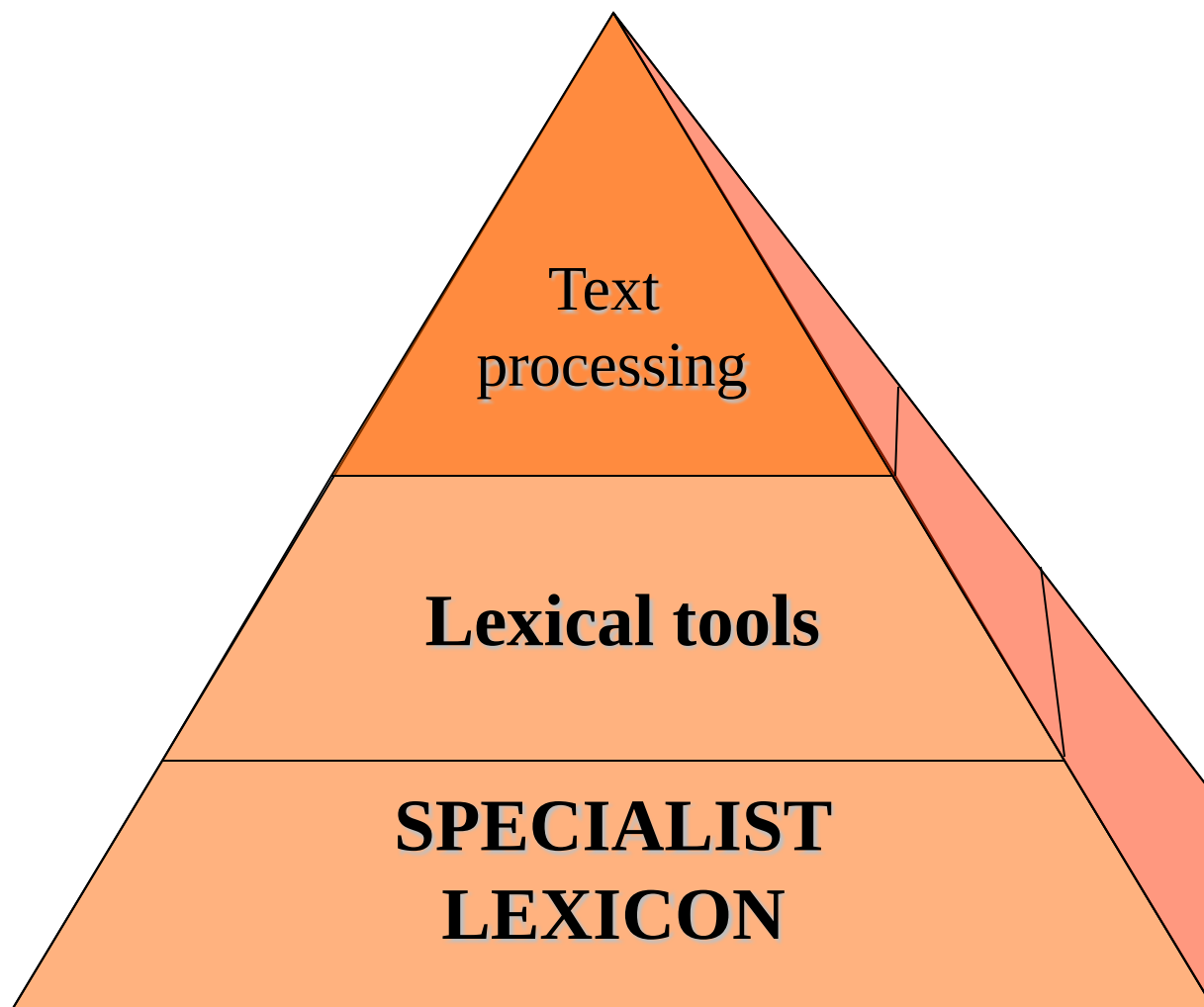
- EUI - Entry ID
- STR - Inflected form
- SCA - Syntactic category
- AGR - agreement information
- BAS - Base form (morphological)
- CIT - Citation form (base=)



LRAGR

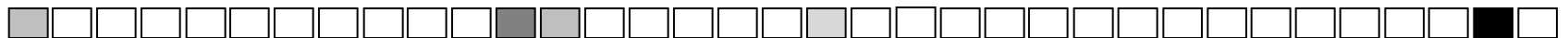
E0003576| Kaposi sarcomas| noun| count(thr_plur)| Kaposi sarcoma| Kaposi's sarcoma|
E0003576| Kaposi sarcomata| noun| count(thr_plur)| Kaposi sarcoma| Kaposi's sarcoma|
E0003576| Kaposi sarcoma| noun| count(thr_sing)| Kaposi sarcoma| Kaposi's sarcoma|
E0003576| Kaposi sarcoma| noun| uncount(thr_sing)| Kaposi sarcoma| Kaposi's sarcoma|
E0003576| Kaposi's sarcomas| noun| count(thr_plur)| Kaposi's sarcoma| Kaposi's sarcoma|
E0003576| Kaposi's sarcomata| noun| count(thr_plur)| Kaposi's sarcoma| Kaposi's sarcoma|
E0003576| Kaposi's sarcoma| noun| count(thr_sing)| Kaposi's sarcoma| Kaposi's sarcoma|
E0003576| Kaposi's sarcoma| noun| uncount(thr_sing)| Kaposi's sarcoma| Kaposi's
sarcoma|





Lexical Tools

- Wordind -- breaks strings into words
 - Produces the Metathesaurus word indexes (MRXW)
- LVG -- performs various lexical transformations
- NORM -- a selection of LVG transformations,
 - Used for Metathesaurus indexing
 - Produces the Metathesaurus Normalized word and string indexes (MRXNW & MRXNS)
 - Used to access those indexes



Normalization

- **Hodgkin Disease**
- **HODGKINS DISEASE**
- **Hodgkin's Disease**
- **Disease, Hodgkin's**
- **HODGKIN'S DISEASE**
- **Hodgkin's disease**
- **Hodgkins Disease**
- **Hodgkin's disease NOS**
- **Hodgkin's disease, NOS**
- **Disease, Hodgkins**
- **Diseases, Hodgkins**
- **Hodgkins Diseases**
- **Hodgkins disease**
- **hodgkin's disease**
- **Disease;Hodgkins**
- **Disease, Hodgkin**
- **Disease hodgkin**



Lexical Tools



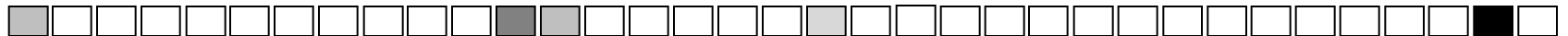
The Lexical Tools

- Introduction
- Norm/WordInd/Lvg
- Details, Details and More Details ...
- Installation
- Embedding This into Your Application
- Using The Lexical Tools with The Metathesaurus
- Building an Index Using The Lexical Tools
- Plans for 2003 and Beyond



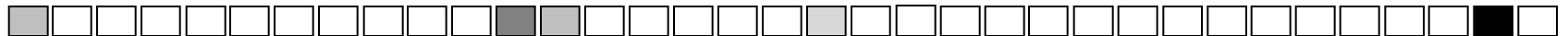
Lexical Tools: Introduction

- These tools:
 - Includes Norm, WordInd, and Lvg
 - Pure Java based
 - Command line tools
 - Java APIs



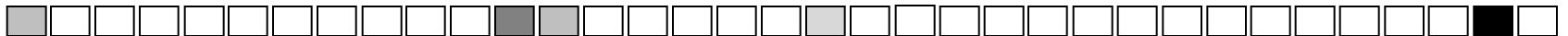
Lexical Tools: Introduction

- These tools are good for
 - aggressive text pattern matching
 - making word, term, phrase indexes
 - matching queries with indexed entries
 - increasing recall and/or precision

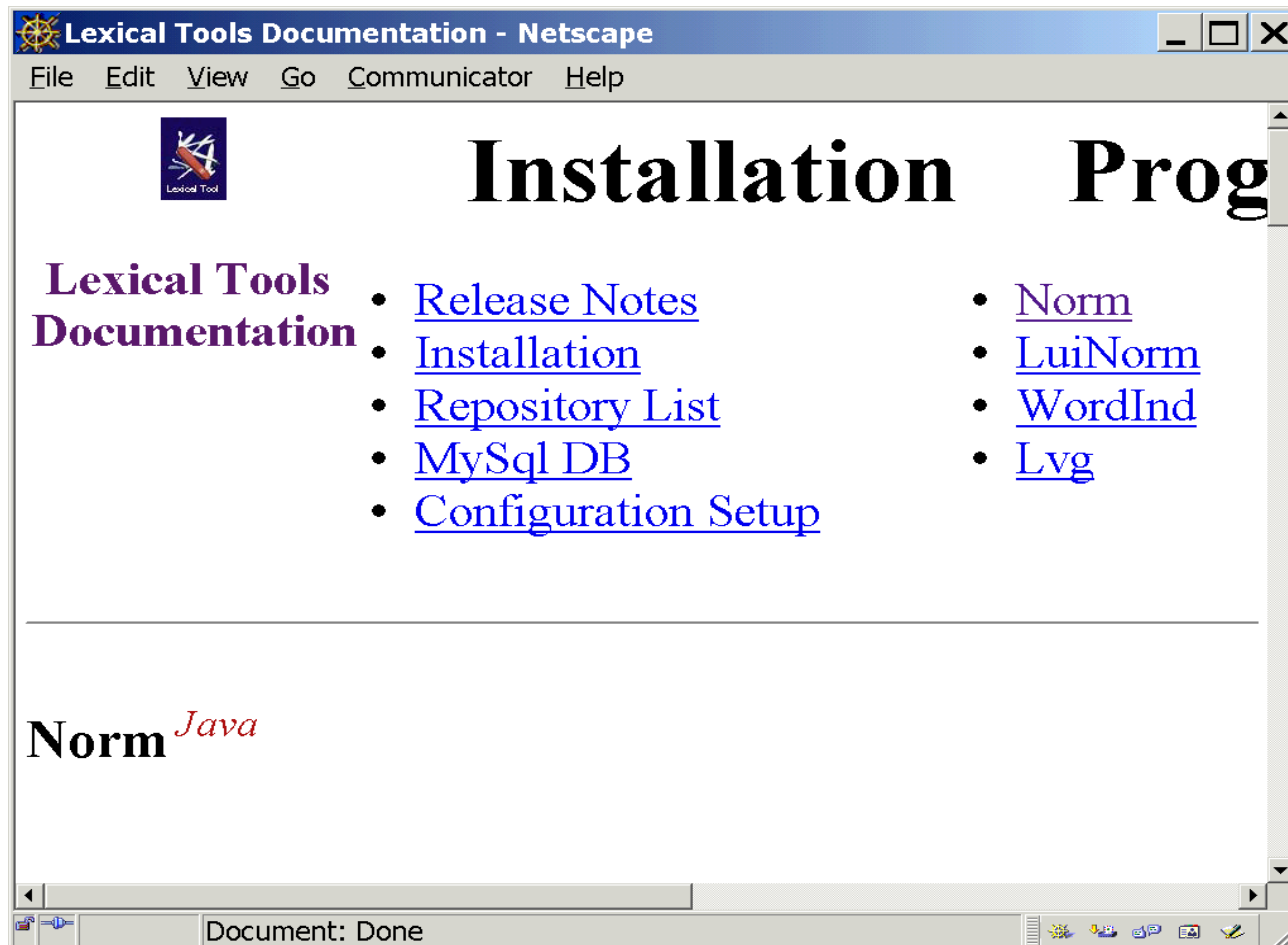


Lexical Tools: Introduction

- Characteristics of all the command line tools
 - take input from the screen or a file
 - put their results to the screen or a file
 - Interpret fielded text
 - Can be told which fields contain what type of information

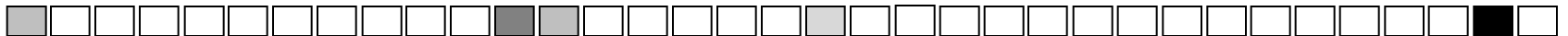


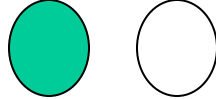
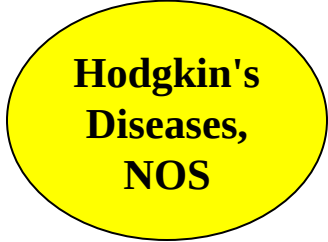
Lexical Tools: Norm



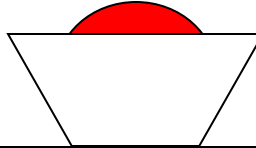
Lexical Tools: Norm

- Norm abstracts away from:
 - case
 - Punctuation
 - word order
 - possessive forms
 - inflectional variation





Lexical Tools: Norm



remove genitives

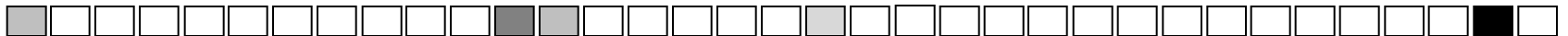
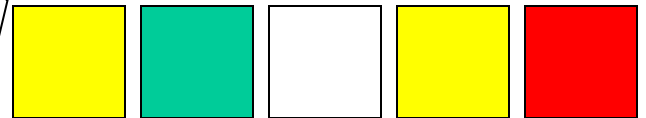
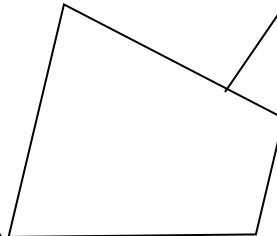
replace punctuation with spaces

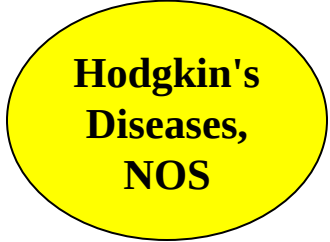
remove stop words

lowercase

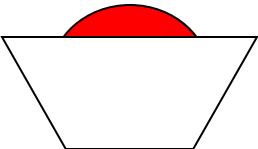
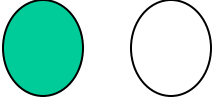
uninflect each word

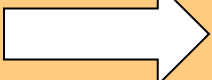
word order sort





Lexical Tools: Norm



remove genitives 

Hodgkin's Diseases, NOS
Hodgkin Diseases, NOS

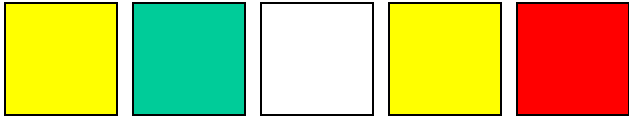
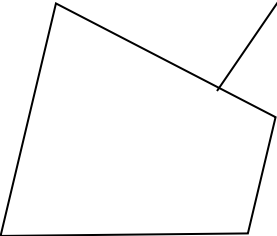
replace punctuation with spaces

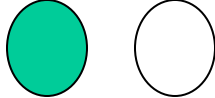
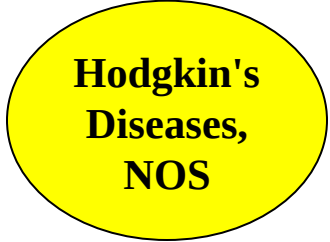
remove stop words

lowercase

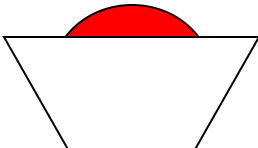
uninflect each word

word order sort

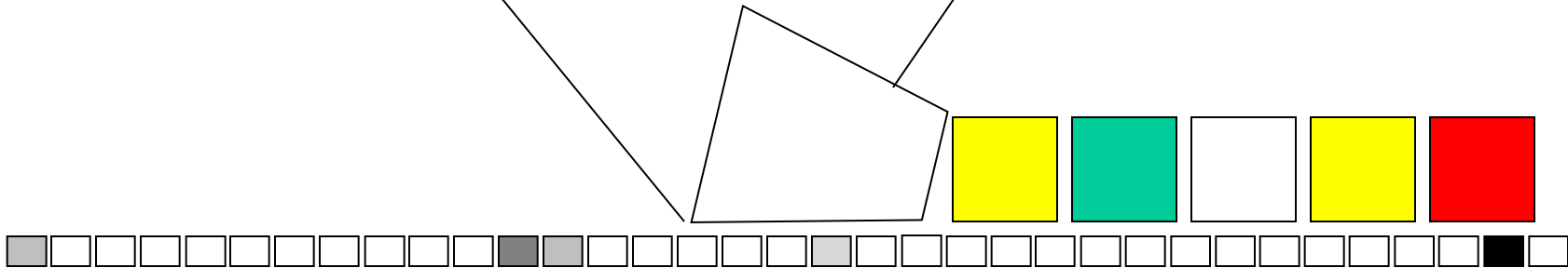
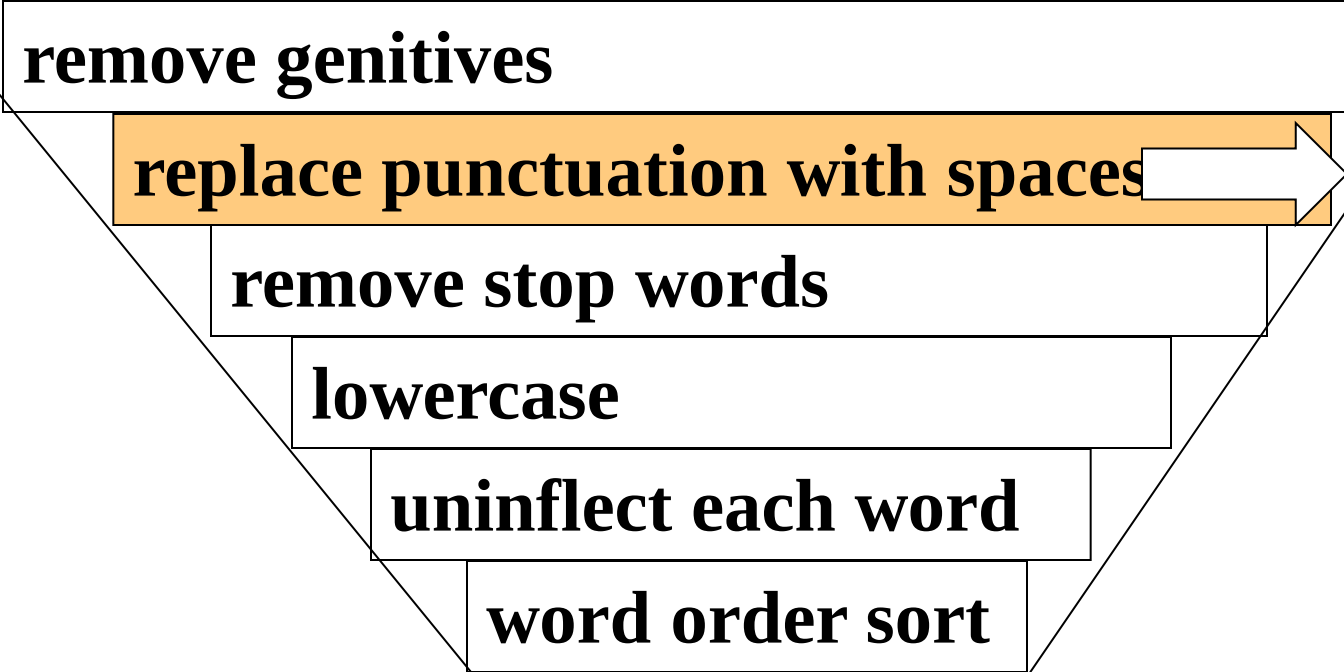


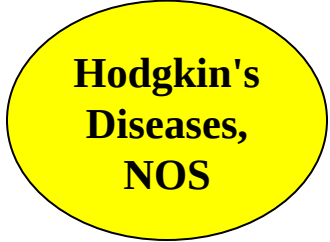


Lexical Tools: Norm

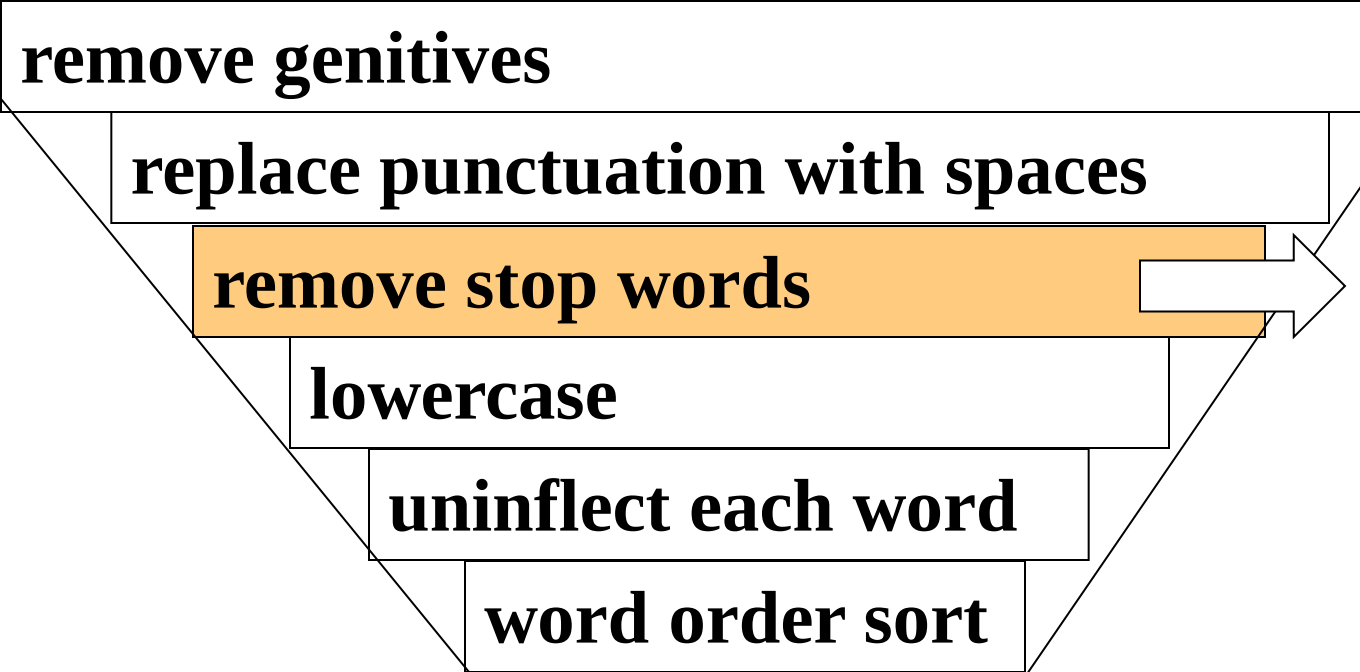
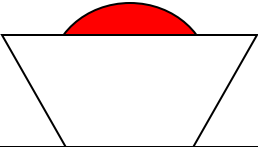


Hodgkin'sDiseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases NOS

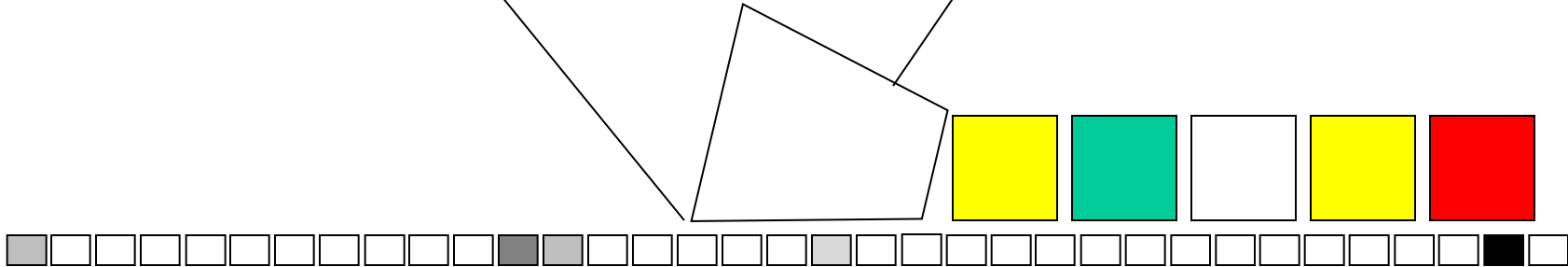


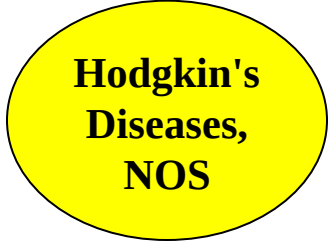


Lexical Tools: Norm

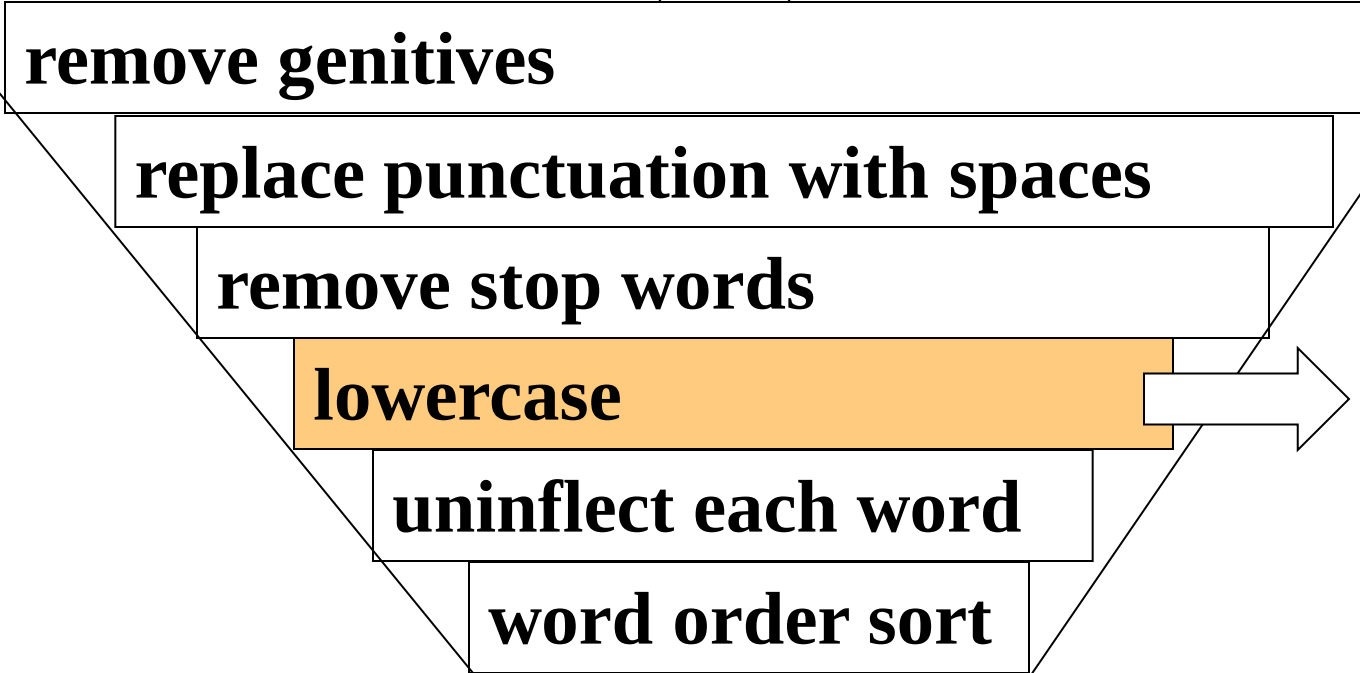
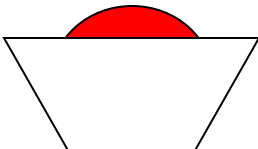
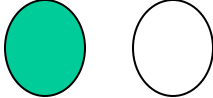


Hodgkin'sDiseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases NOS
Hodgkin Diseases

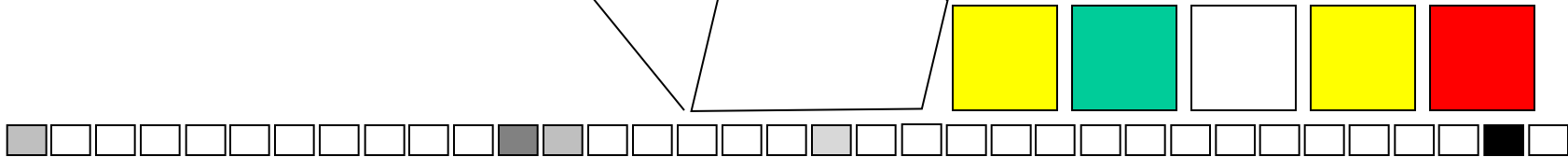


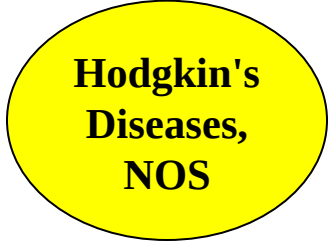


Lexical Tools: Norm

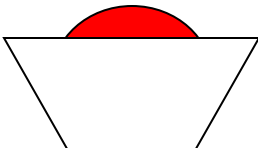
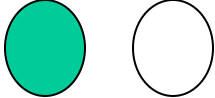


Hodgkin'sDiseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases NOS
Hodgkin Diseases
hodgkin diseases





Lexical Tools: Norm



remove genitives

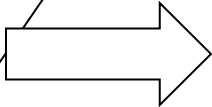
replace punctuation with spaces

remove stop words

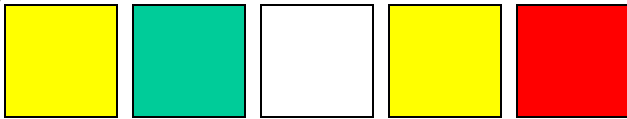
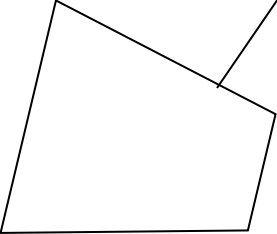
lowercase

uninflect each word

word order sort



Hodgkin'sDiseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases NOS
Hodgkin Diseases
hodgkin diseases
hodgkin disease



**Hodgkin's
Diseases,
NOS**

Lexical Tools: Norm

remove genitives

replace punctuation with spaces

remove stop words

lowercase

uninflect each word

word order sort

**Hodgkin'sDiseases,
NOS**

**Hodgkin Diseases,
NOS**

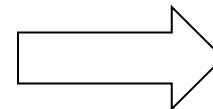
**Hodgkin Diseases
NOS**

Hodgkin Diseases

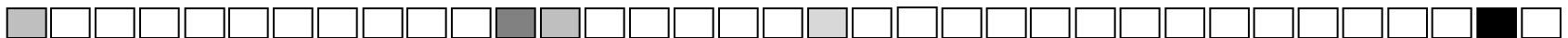
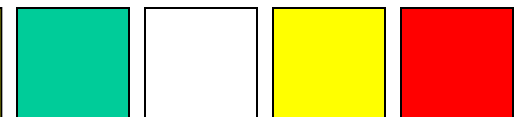
hodgkin diseases

hodgkin disease

disease hodgkin



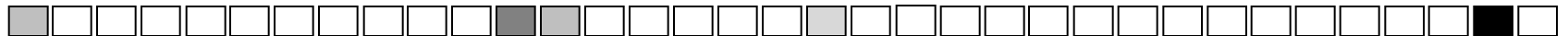
**disease
hodgkin**



Lexical Tools: Norm

Down's Syndrome Down Syndrome	down syndrome
Acetolyses acetolysis	acetolysis
Lung cancer Cancer, lung	cancer lung

**Norm
to**



Lexical Tools: Norm

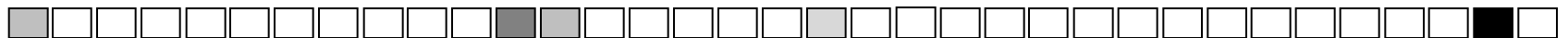
> norm

Paget's disease-scapula

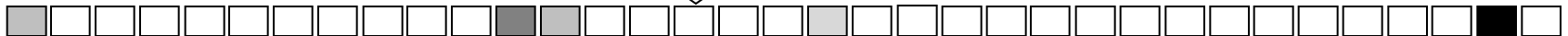
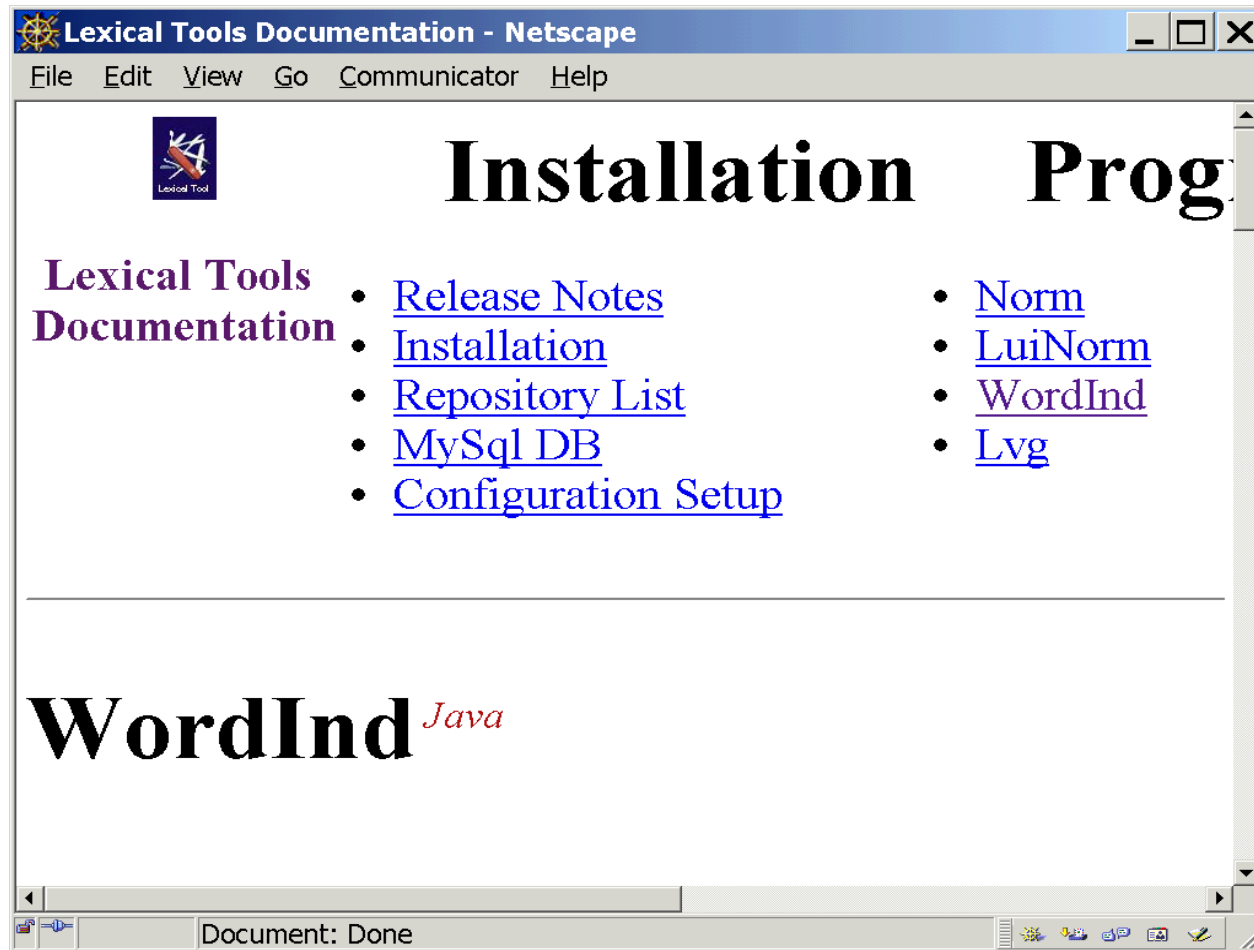
Paget's disease-scapula | **disease paget scapula**

Scapula, Paget Disease

Scapula, Paget Disease | **disease paget scapula**

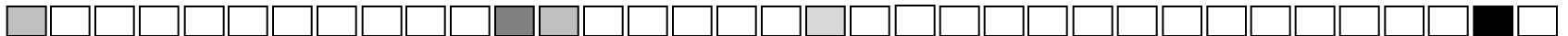


Lexical Tools: WordInd



Lexical Tools: WordInd

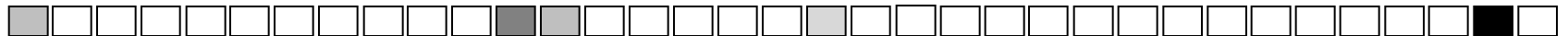
- Breaks words into tokens
- Passes other fields to output, untouched
- Lowercases
- Removes white space and punctuation



Lexical Tools: WordInd

Useful command line options for wordInd

-t[:Num]	Defines what field to tokenize
-f[:Num[:Num]]	Defines what fields get passed through



Lexical Tools: WordInd

> wordInd -t:7 -F:1:6

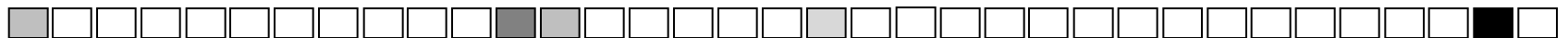
C0185495 ENG P L0223844 PF S0298948 Denis-Browne splint strapping 3
--

C0185495|S0298948|**denis**

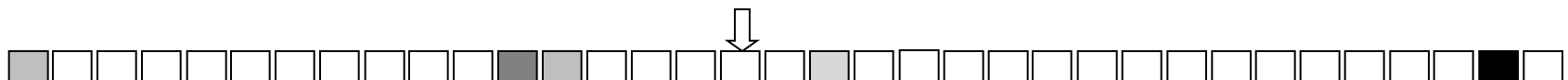
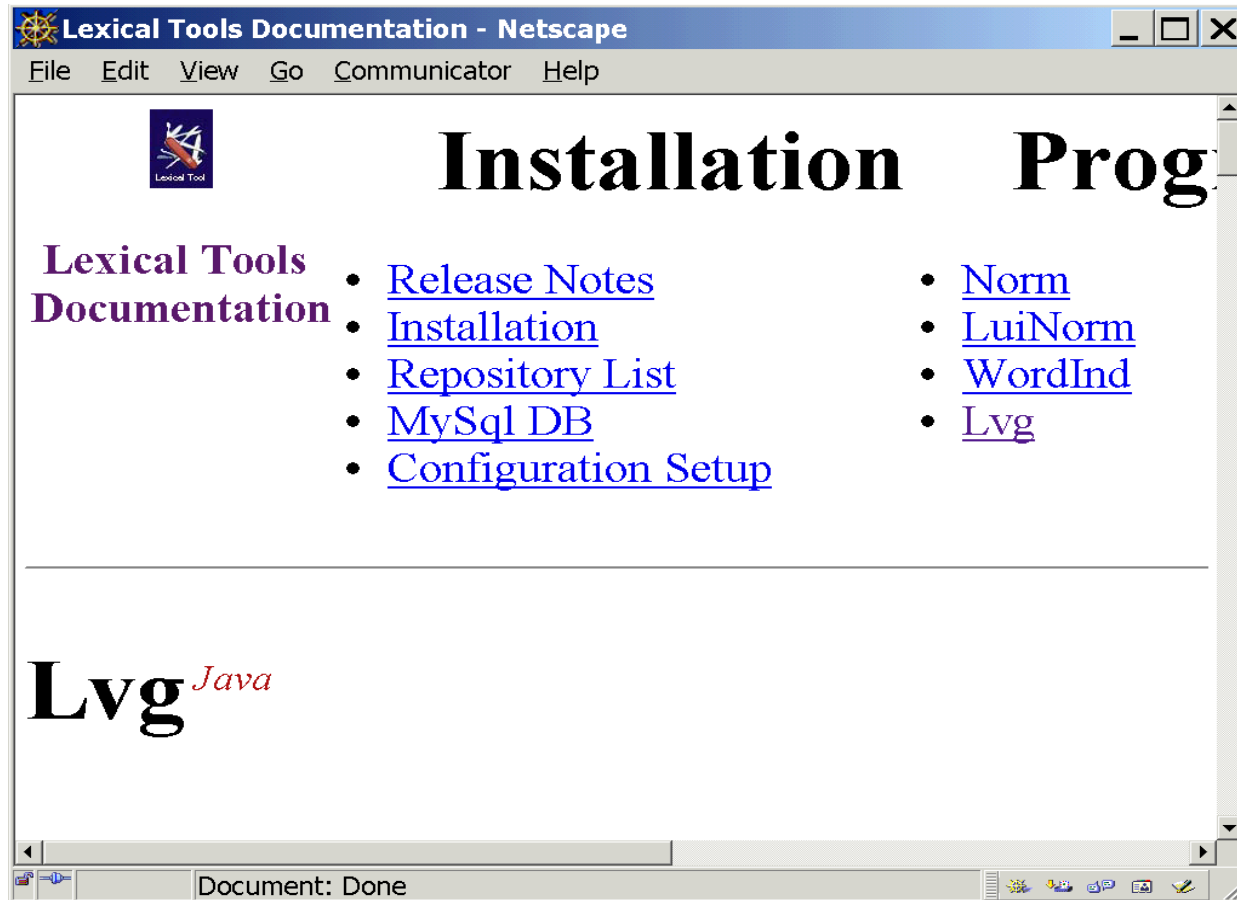
C0185495|S0298948|**browne**

C0185495|S0298948|**splint**

C0185495|S0298948|**strapping**

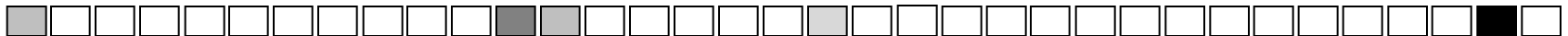


Lexical Tools: Lvg



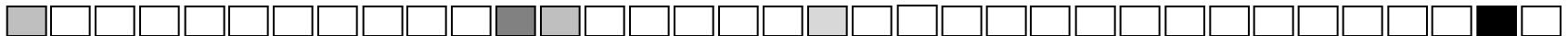
Lexical Tools: Flow Components

Mnemonic	Tool
A	<u>Return known acronyms</u>
a	<u>Return known acronym expansions</u>
B	<u>Uninflect words in a term</u>
b	<u>Uninflect a term</u>
C	<u>Canonicalize</u>
Ct	<u>Retrieve the citation term</u>



Lexical Tools: Flow Components

Mnemonic	Tool
c	<u>Tokenize a term into "words"</u>
ca	<u>Tokenize, keep everything</u>
ch	<u>Tokenize without breaking hyphens</u>
d	<u>Generate derivational variants</u>
dc~N	<u>Generate derivational variants with specifying output categories</u>



Lexical Tools: Flow Components

Mnemonic	Tool
E	<u>Retrieve the unique EUI for a term</u>
f	<u>Filter output to contain only forms from lexicon</u>
Gn	<u>Generate known fruitful variants</u>
g	<u>Remove genitive</u>
i	<u>Generate inflectional variants</u>
ici~Cats+Infls	<u>Generate inflectional variants, by Categories and inflections</u>

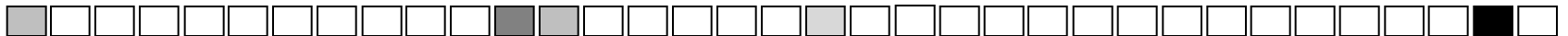


Lexical Tools: Flow Components

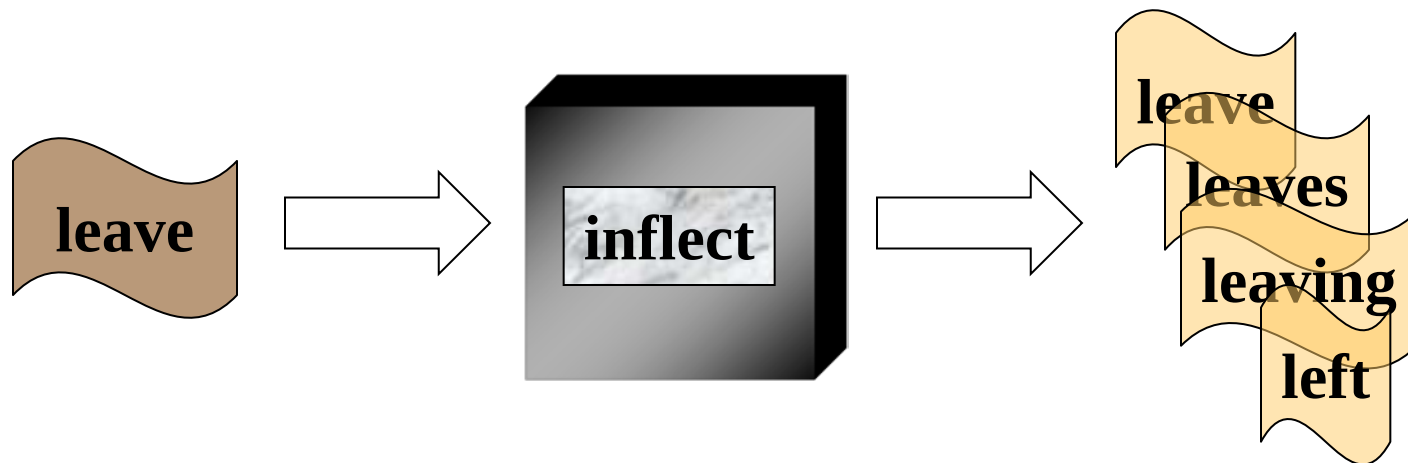
Mnemonic	Tool
L	<u>Retrieve category and inflection for a term</u>
l	<u>Lowercase</u>
N	<u>Normalize the input text in a non-canonical way (Norm)</u>
o	<u>Replace punctuations with spaces</u>
p	<u>Strip Punctuation</u>
q	<u>Strip diacritics</u>
R	<u>Generate derivational variants, recursively</u>
r	<u>Generate synonyms, recursively</u>

Lexical Tools: Flow Components

Mnemonic	Tool
s	<u>Generate known spelling variants</u>
t	<u>Strip stop words</u>
u	<u>Uninvert the input phrase around commas</u>
w	<u>Sort words by order</u>
y	<u>Generate synonyms</u>



Lexical Tools: Flows

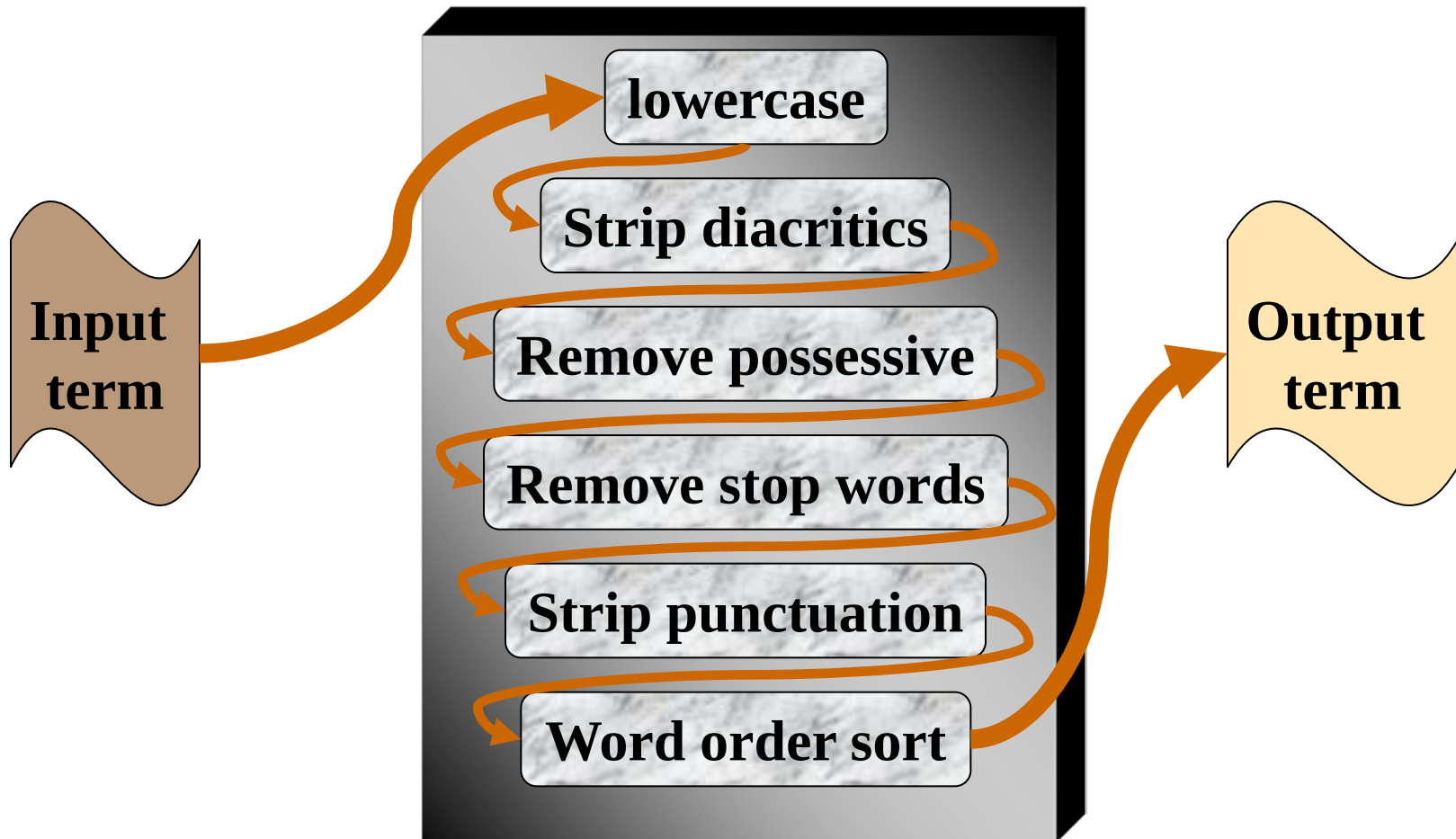


Lexical Tools: Flows

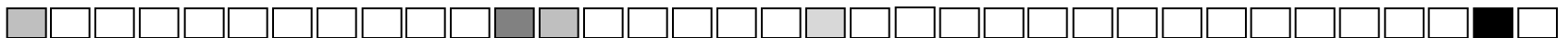
```
> lvg -f:i
leave
leave|leave|128|1|i|1|
leave|leave|128|512|i|1|
leave|leaves|128|8|i|1|
leave|left|1024|64|i|1|
leave|left|1024|32|i|1|
leave|leave|1024|1|i|1|
leave|leave|1024|262144|i|1|
leave|leave|1024|1024|i|1|
leave|leaves|1024|128|i|1|
leave|leaving|1024|16|i|1|
```



Lexical Tools: A Serial Flow



Flow components can be arranged so that the output of one is the input to another.



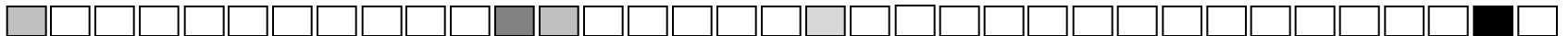
Lexical Tools: A Serial Flow

> lvg -f:l:q:g:t:p:w

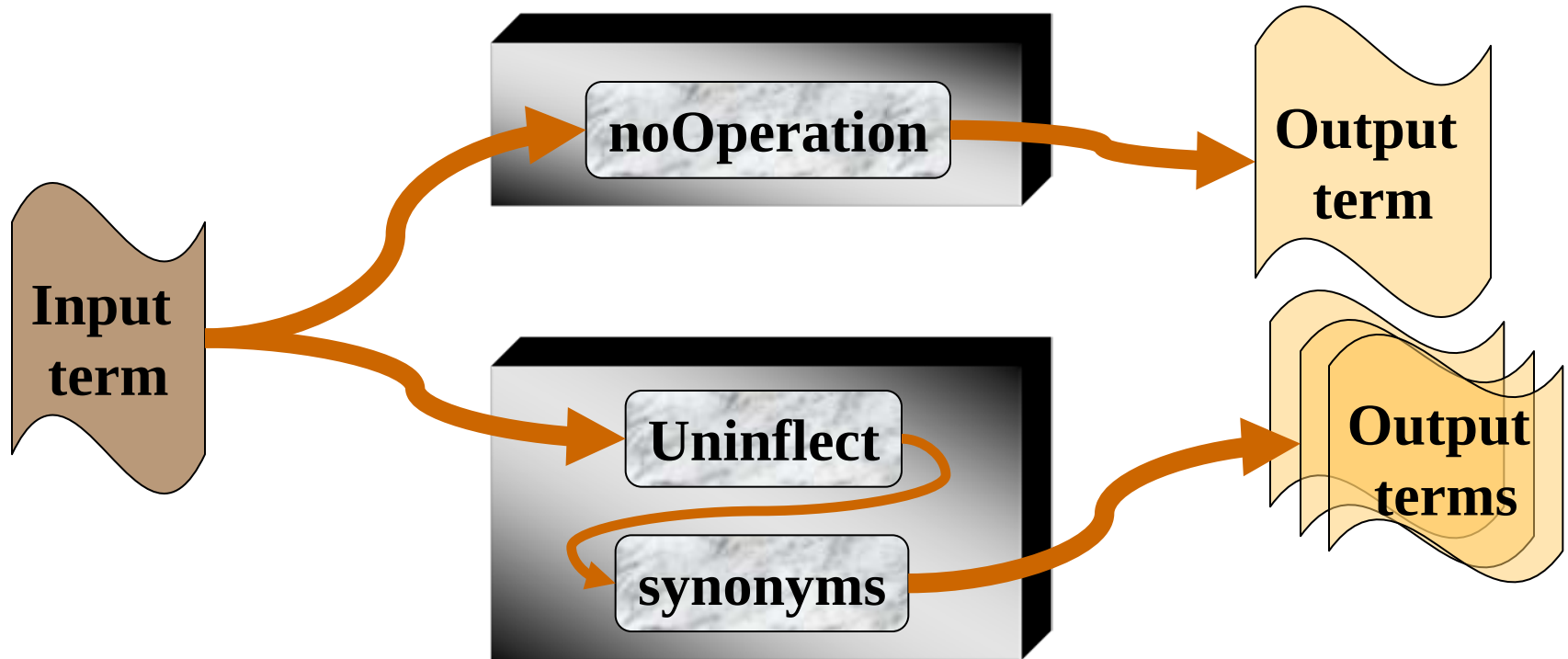
The Gougerot-Sjögren's Syndrome

The Gougerot-Sjögren's Syndrome

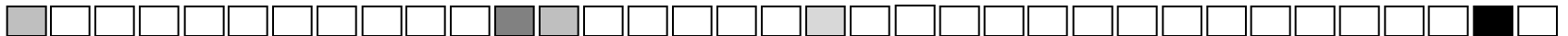
gougerotsjogren syndrome|2047|16777215|
l+q+g+t+p+w|1|



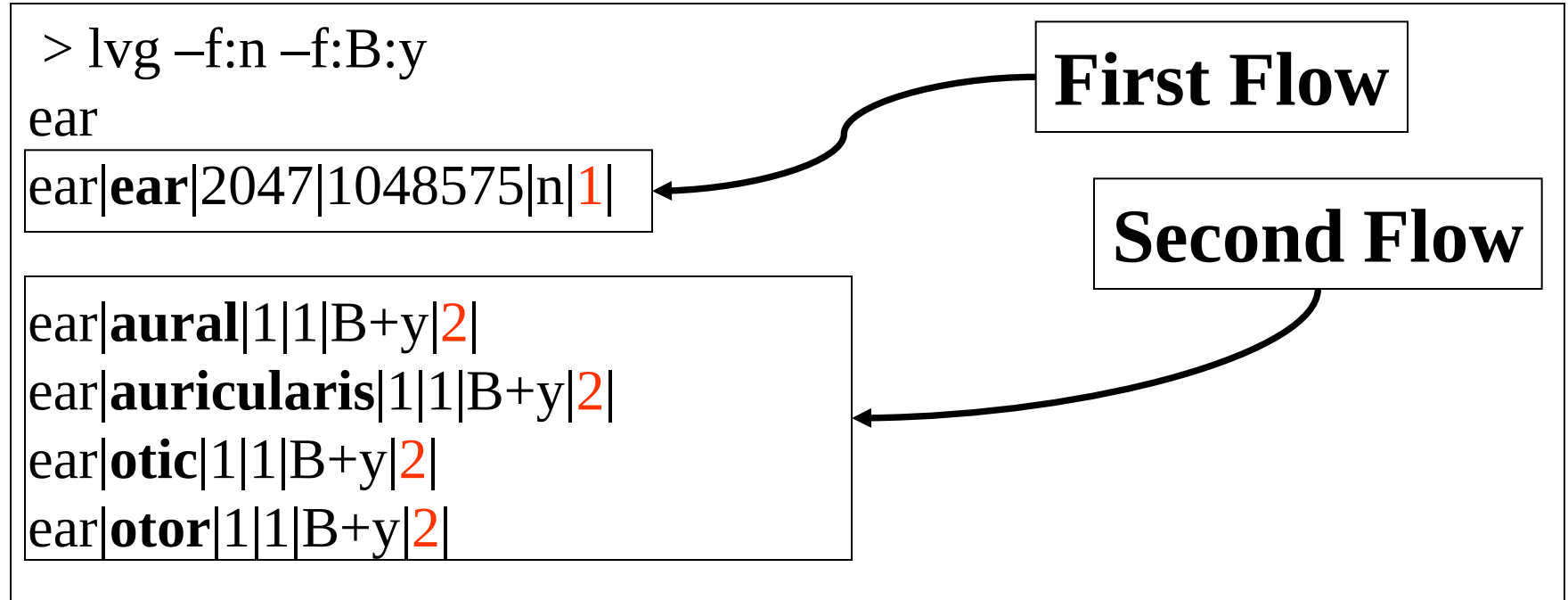
Lexical Tools: Parallel Flows



Multiple flows can be defined

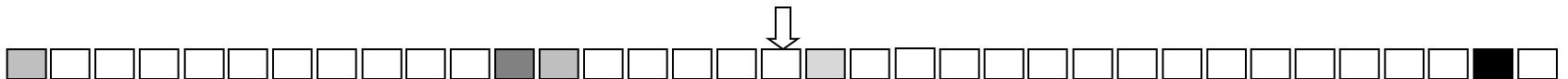
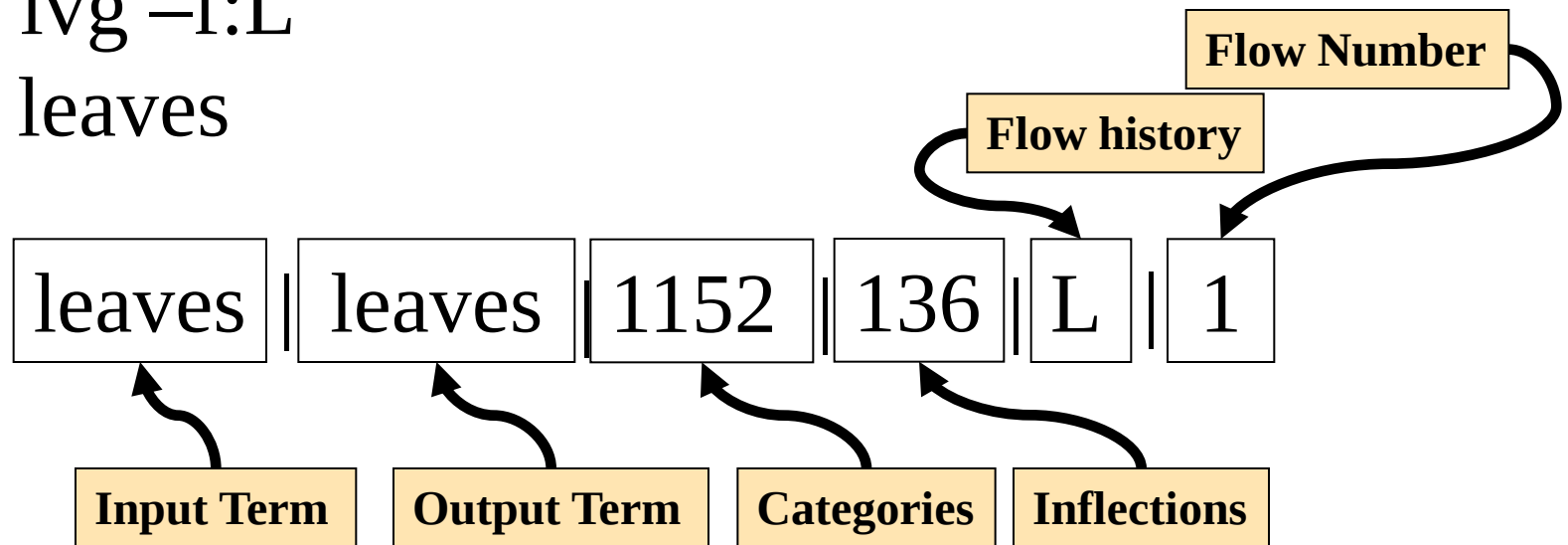


Lexical Tools: Parallel Flows

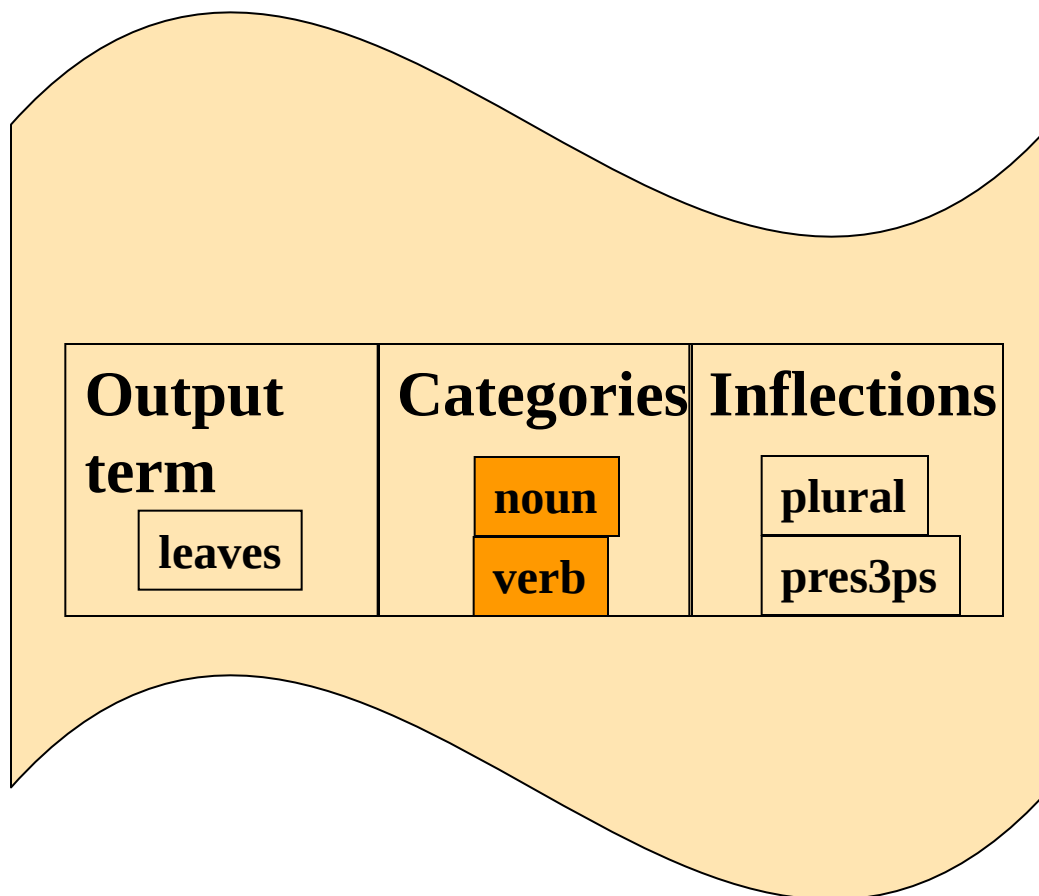


Lexical Tools: Fielded Output

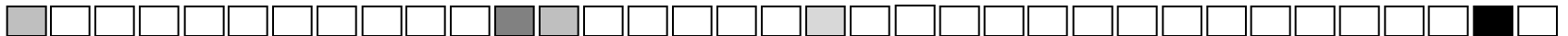
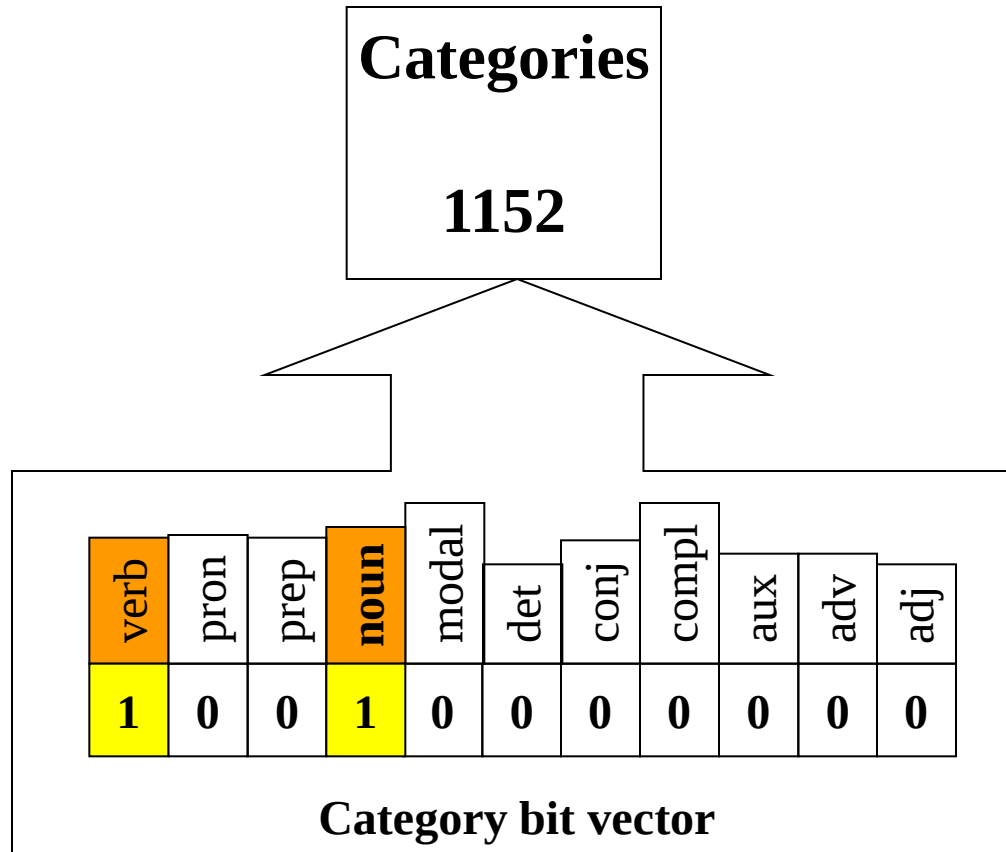
> lvg -f:L
leaves



Lexical Tools: Fielded Output



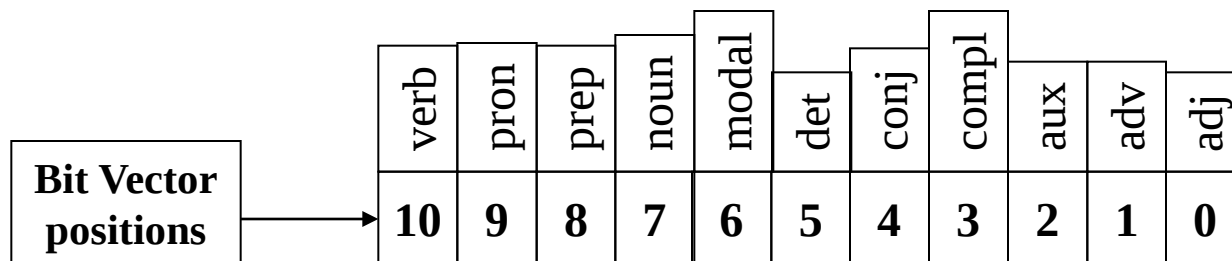
Lexical Tools: Categories



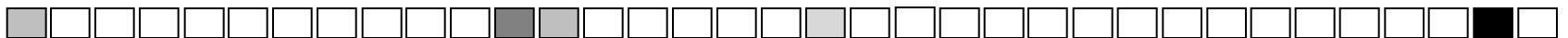
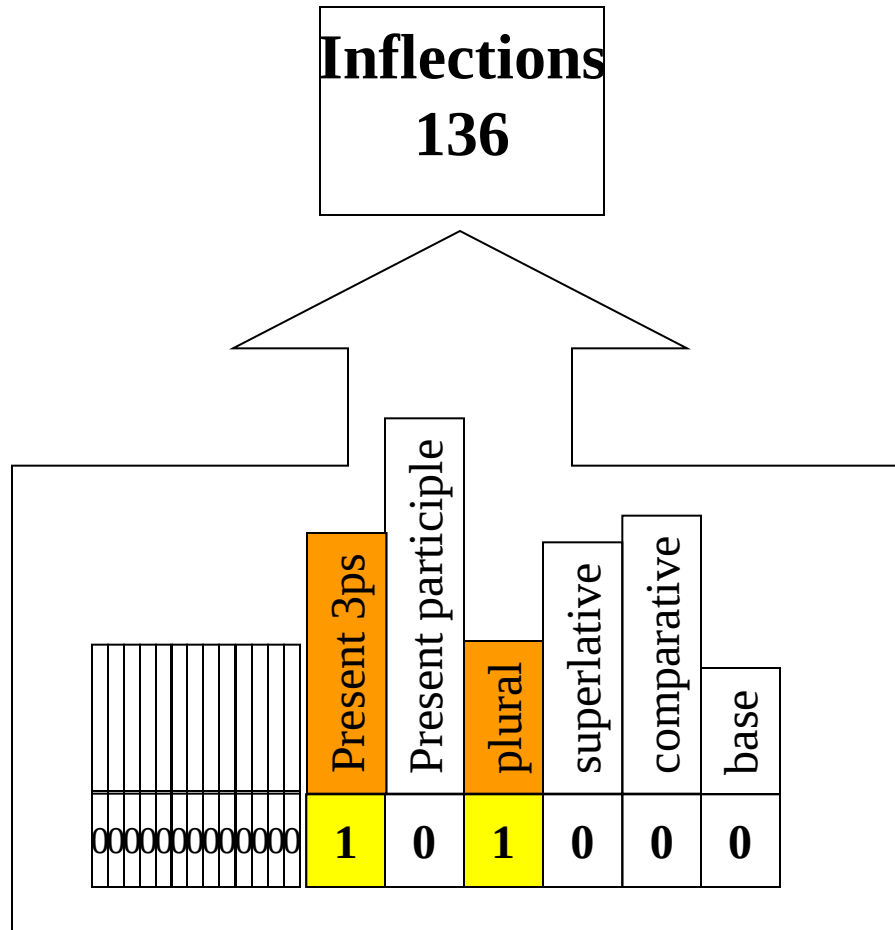
Lexical Tools: Categories

Adjective	1
Adverb	2
Auxiliary	4
Complement	8
Conjunction	16
Determiner	32

Modal	64
Noun	128
Preposition	256
Pronoun	512
Verb	1024



Lexical Tools: Inflections



Lexical Tools: Inflections

Base	1
Comparative	2
Superlative	4
Plural	8
Present Participle	16
Past	32
Past Participle	64
Present 3 rd Person Singular	128

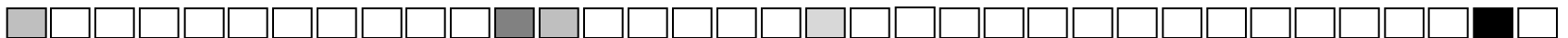
Positive	256
Singular	512
Infinitive	1024
Pres 123p	2048
Past Neg	4096
pres123pNeg	8192
Pres1s	16384
past1p23pNeg	32768
past1s3sNeg	65536

pres1p23p	131072
pres1p23p	262144
pres1p23pNeg	524288
past1s3s	1048576
pres	2097152
pres3sNeg	4194304
presNeg	8388608
all	16777215

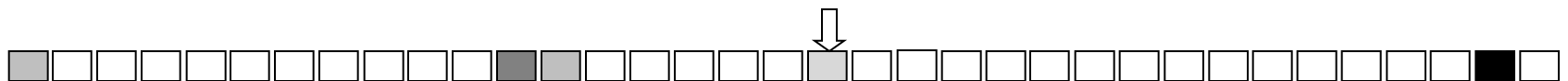
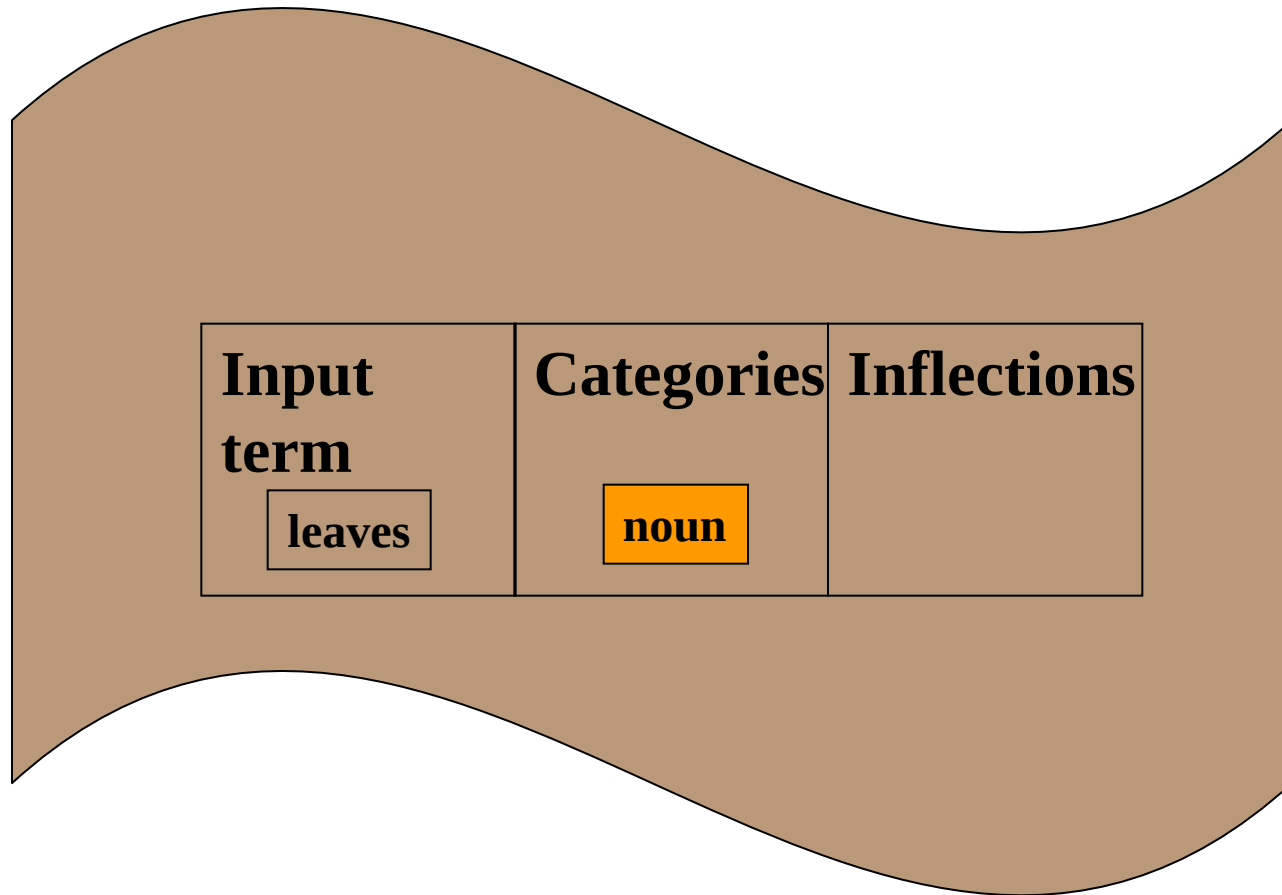


Lexical Tools: Fielded Output

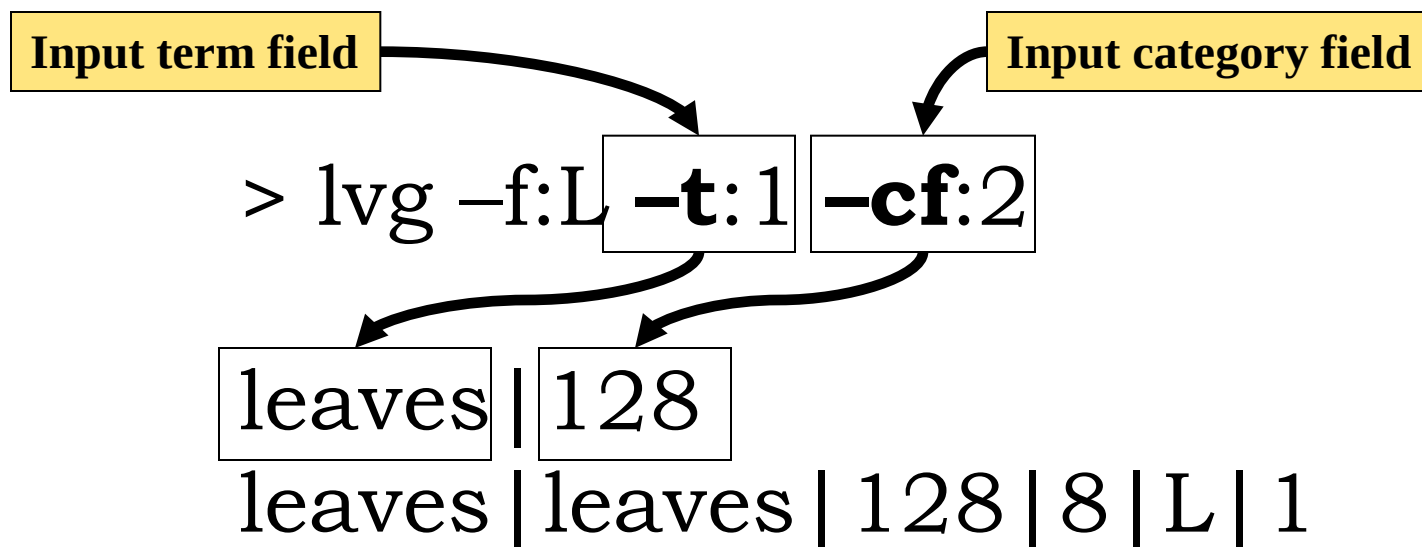
Input term	Output term	Categories	Inflections	Flow history	Flow number
leaves	leaves	1152	136	L	1



Lexical Tools: Fielded Input

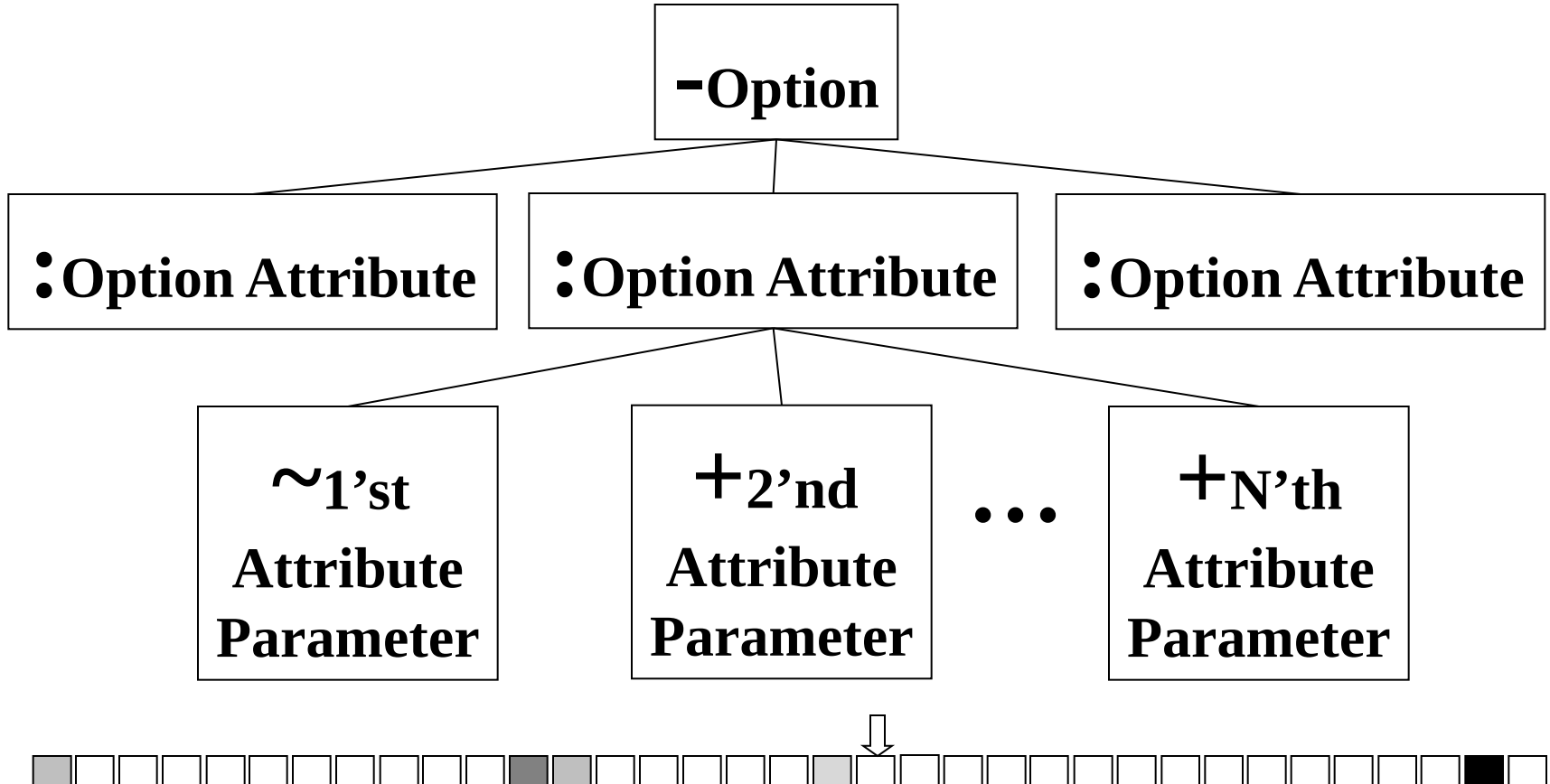


Lexical Tools: Fielded Input



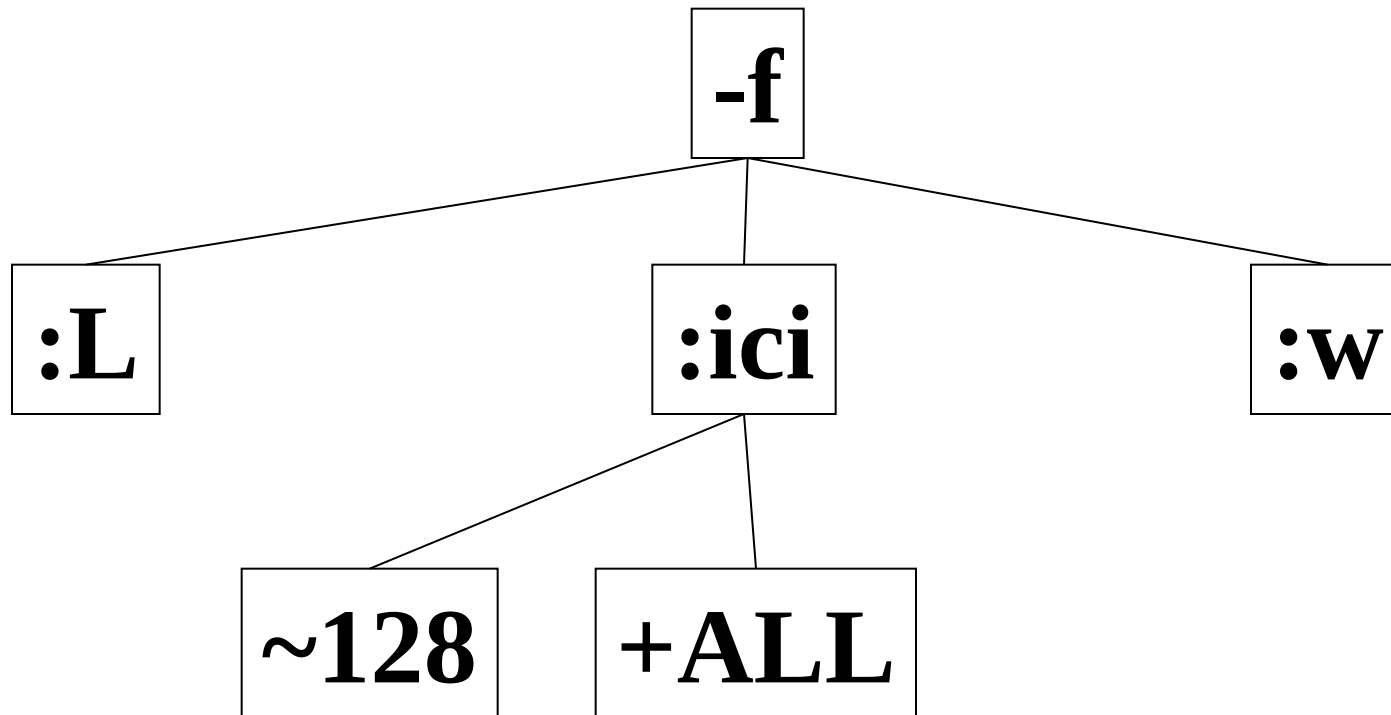
Lexical Tools: Command Line Syntax

- Hierarchical structure

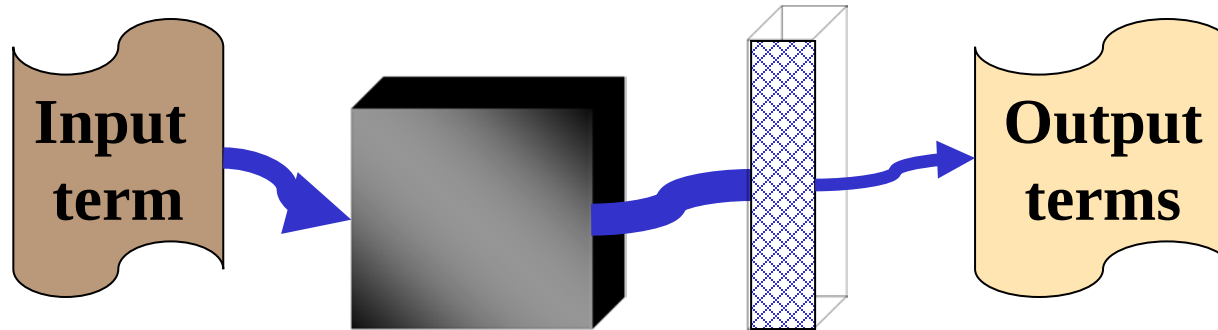


Lexical Tools: Command Line Syntax

-f:L:ici~128+ALL:w



Lexical Tools: Post Flow Options



SC	<u>Show category names</u>
SI	<u>Show inflection names</u>
ccgi	<u>Mark the end of the set of variants returned</u>
F:Int[:Int]	<u>Specify fields for outputs</u>
ti	<u>Display the only input term in the output when using fielded input</u>
R:Int	<u>Restrict the number of variants returned</u>



Lexical Tools: Post Flow Options

Show category names

Show inflection names

> lvg -f:L **-SC -SI**

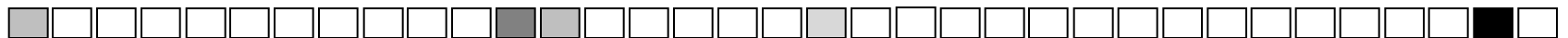
phosphoprotein

phosphoprotein | phosphoprotein | **<noun>** | **<base+singular>** | L | 1 |

sclerosing

sclerosing | sclerosing | **<adj+verb>** | **<base+presPart+positive>** | L | 1 |

Show the category and
inflection names



Lexical Tools: Post Flow Options

Mark the end of the set of variants returned

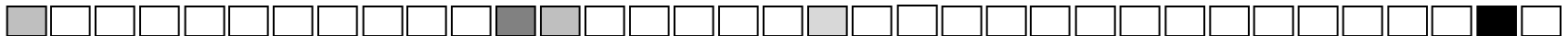
Mark the end of processing

> lvg -f:L **-ccgi**

behavior

behavior | behavior | 128 | 513 | L | 1 |

__THE_END__



Lexical Tools: Post Flow Options

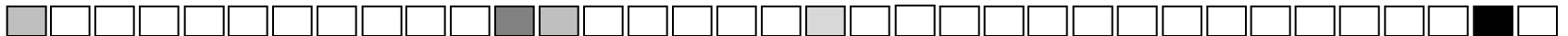
Specify fields for outputs

Display only the 8th and 6th field from the output

> lvg -f:u -t:7 **-F:8:6**

C0035440|ENG|S|L0035434|VW|S0003894|Rheumatic carditis, acute

acute Rheumatic carditis|S0003894



Lexical Tools: Post Flow Options

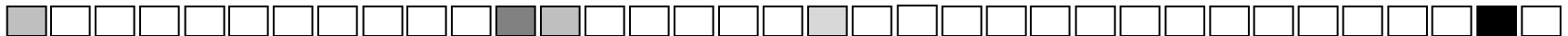
Display only the input term field when using fielded input

> lvg -f:u -t:7 **-ti**

Display only the input term
from the fielded input to
the output

C0035440 | S0003894 | ***Rheumatic carditis, acute***

Rheumatic carditis, acute | acute Rheumatic carditis | 2047 | 16777215 | u | 1 |



Lexical Tools: Post Flow Options

Restrict the number of variants returned

```
> lvg -f:i -R:2
```

Limit the number of
output terms to 2

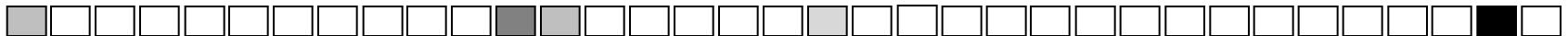
```
foo
```

```
foo | foo | 128 | 1 | i | 1 |
```

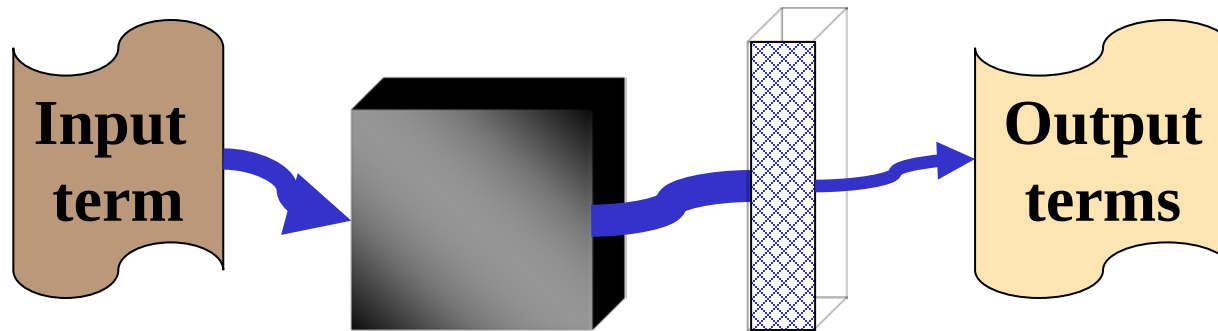
```
foo | foos | 128 | 8 | i | 1 |
```

Note: Dangerous!
Do not try this at home!

Note: The unrestricted output would
have produced 12 rows otherwise



Lexical Tools: Post Flow Options



EC:Long	<u>Display variants: exclude categories specified</u>
EI:Long	<u>Display variants: exclude inflections specified</u>
DC:Long	<u>Display variants that only contain the categories specified</u>
DI:Long	<u>Display variants that only contain the inflections specified</u>



Lexical Tools: Post Flow Options

Display variants: exclude categories specified

> lvg -f:i **-EC:1919**

sleep

sleep | sleep | 128 | 1 | i | 1

sleep | sleep | 128 | 512 | i | 1

Display variants, but exclude all terms other than nouns

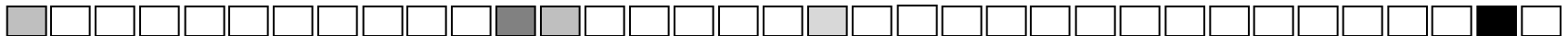


Lexical Tools: Post Flow Options

Display variants exclude inflection specified

Display variants, but
exclude base forms

> lvg -f:i -EC:1919 **-EI:1**
sleep
sleep | sleep | 128 | 512 | i | 1

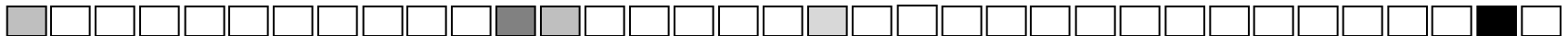


Lexical Tools: Post Flow Options

Display variants contain category specified

```
> lvg -f:i -DC:128  
sleep  
sleep | sleep | 128 | 1 | i | 1  
sleep | sleep | 128 | 512 | i | 1
```

Display
variants, but
only include
nouns in the
output



Lexical Tools: Post Flow Options

Display variants contain inflection specified

> lvg -f:i **-DI:255**

Display variants, but
only include “simplified”
inflections

sleep

sleep | sleep | 128 | 1 | i | 1 |

sleep | sleep | 1024 | 1 | i | 1 |

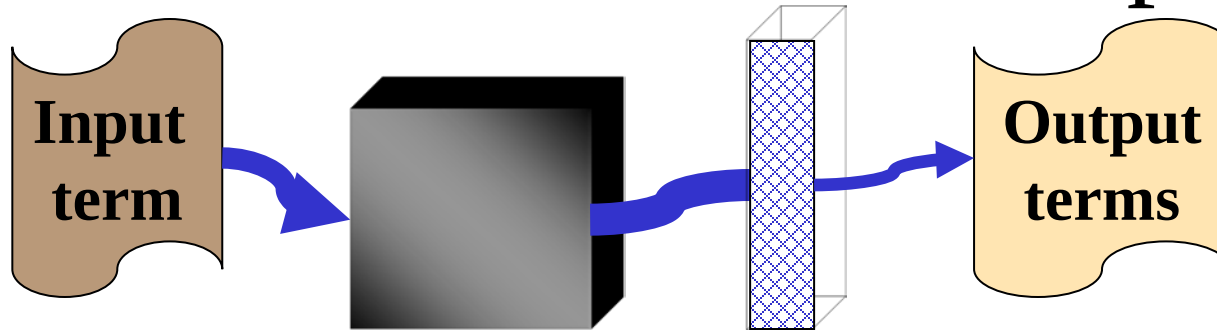
sleep | slept | 1024 | 64 | i | 1 |

sleep | slept | 1024 | 32 | i | 1 |

sleep | sleeps | 1024 | 128 | i | 1 |

sleep | sleeping | 1024 | 16 | i | 1 |

Lexical Tools: Post Flow Options



CR:o	<u>Combine record by output term</u>
CR:oc	<u>Combine record by output term and category</u>
CR:oi	<u>Combine record by output term and inflection</u>
St:o	<u>Sort outputs by terms in an alphabetical order</u>
St:oc	<u>Sort outputs by term and category</u>
St:oci	<u>Sort outputs by term, category, and inflection</u>



Lexical Tools: Post Flow Options

Combine record by output term

Combine records by term

> lvg -f:i **-CR:o**

Note: this is a noun+verb

sleep

sleep | sleep | **1152** | 263681 | i | 1 |

sleep | slept | 1024 | 96 | i | 1 |

sleep | sleeps | 1024 | 128 | i | 1 |

sleep | sleeping | 1024 | 16 | i | 1 |

Lexical Tools: Post Flow Options

Combine record by output term and category

> lvg -f:i **-CR:oc**

Combine records by term
and category.

sleep

Note: this is both the
base+singular inflections
combined

sleep | sleep | 128 | 513 | i | 1 |

sleep | sleep | 1024 | 263169 | i | 1 |

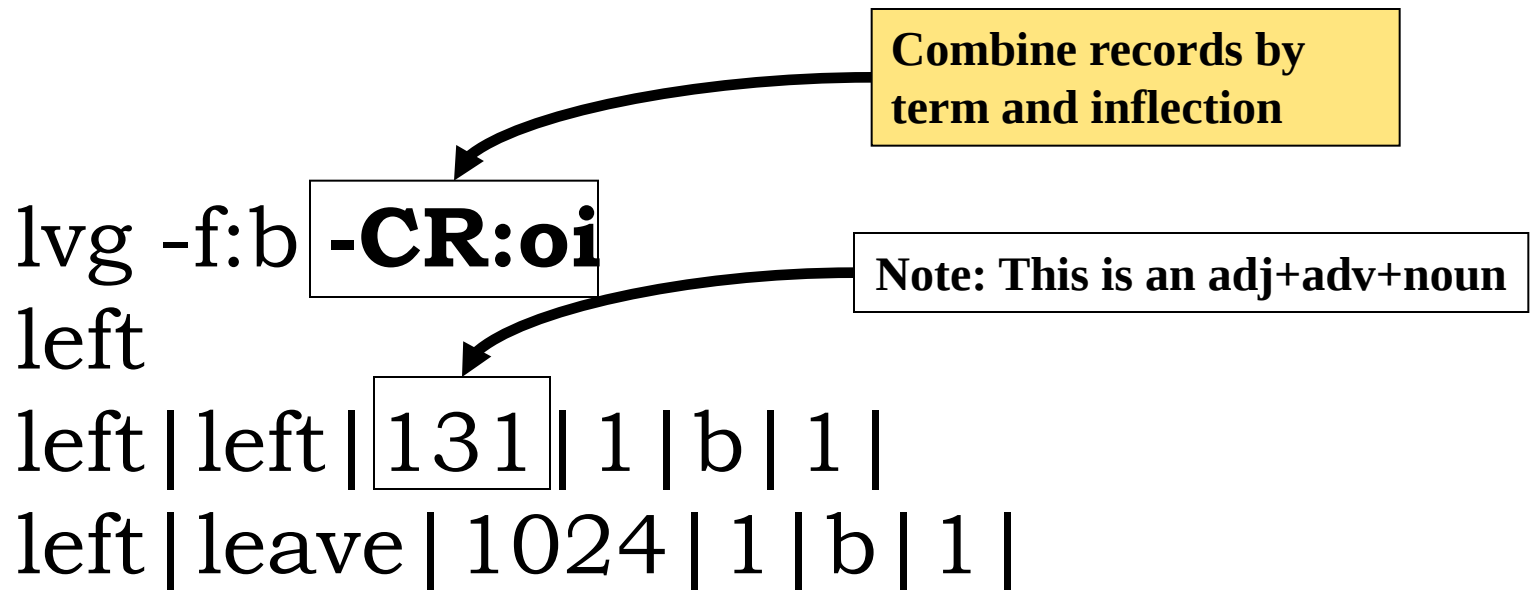
sleep | slept | 1024 | 96 | i | 1 |

sleep | sleeps | 1024 | 128 | i | 1 |

sleep | sleeping | 1024 | 16 | i | 1 |

Lexical Tools: Post Flow Options

Combine record by output term and inflection



Lexical Tools: Post Flow Options

Sort outputs by terms in an alphabetical order

```
>lvg -f:i -St:o
```

Sort by the output term

```
see
```

```
see|saw|1024|32|i|1|
```

```
see|see|1024|1|i|1|
```

```
see|see|1024|262144|i|1|
```

```
see|see|1024|1024|i|1|
```

```
see|seeing|1024|16|i|1|
```

```
see|seen|1024|64|i|1|
```

```
see|sees|1024|128|i|1|
```

Lexical Tools: Post Flow Options

Sort outputs by term and category

```
> lvg -f:i -St:oc
```

```
left
```

```
...
```

```
left|left|1|1|i|1|
```

```
left|left|1|256|i|1|
```

```
left|left|2|1|i|1|
```

```
left|left|2|256|i|1|
```

```
left|left|128|1|i|1|
```

```
left|left|128|512|i|1|
```

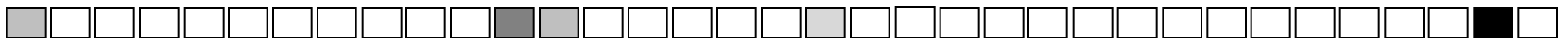
```
left|left|128|8|i|1|
```

```
left|left|1024|64|i|1|
```

```
left|left|1024|32|i|1|
```

```
left|lefts|128|8|i|1|
```

Sort by the output term
and category



Lexical Tools: Post Flow Options

Sort outputs by term, category, and inflection

```
> lvg -f:i -St:oci
```

Sort by the output
term, category and
inflection

see

see | saw | 1024 | 32 | i | 1 |

see | see | 1024 | 1 | i | 1 |

see | see | 1024 | 1024 | i | 1 |

see | see | 1024 | 262144 | i | 1 |

see | seeing | 1024 | 16 | i | 1 |

see | seen | 1024 | 64 | i | 1 |

see | sees | 1024 | 128 | i | 1 |

Lexical Tools:

Why Base+Singular

- Base form for Nouns:

- Singular form

- There are exceptions:

- Police

lv9 -f:i -SC -SI -DC:128

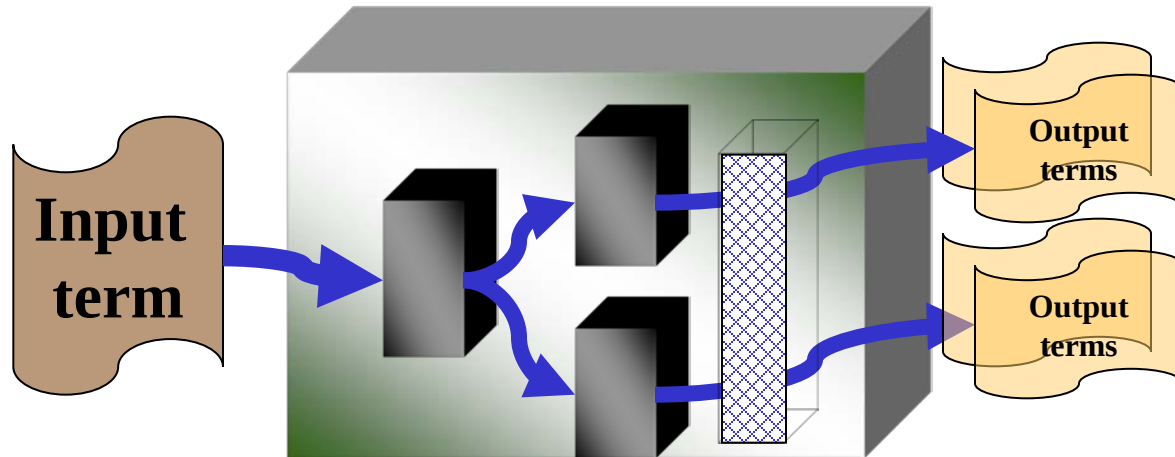
police

police|police|<noun>|<base>|i|1|

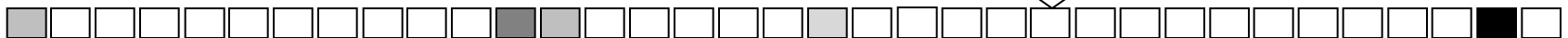
police|police|<noun>|<plural>|i|1|



Lexical Tools: Global Behaviors



<i>i:filename</i>	<u>Define input file name</u>
<i>o:filename</i>	<u>Define output file name</u>
<i>x:filename</i>	<u>Loading an alternative configuration file</u>
<i>p</i>	<u>Interactive prompt</u>
<i>m</i>	<u>Print extra information of flow mutations</u>
<i>s:Str</i>	<u>Defines a field separator.</u>



Lexical Tools: Global Behaviors

Interactive prompt

> lvg -f:s -CR:oc

-p

Set the prompt on

- Please input a term (type "Ctl-d" to Quit) >

color

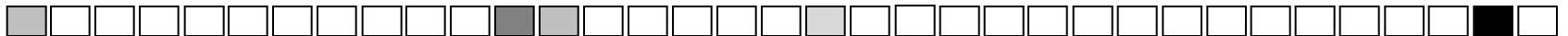
color | color | 128 | 513 | s | 1 |

color | color | 1024 | 263169 | s | 1 |

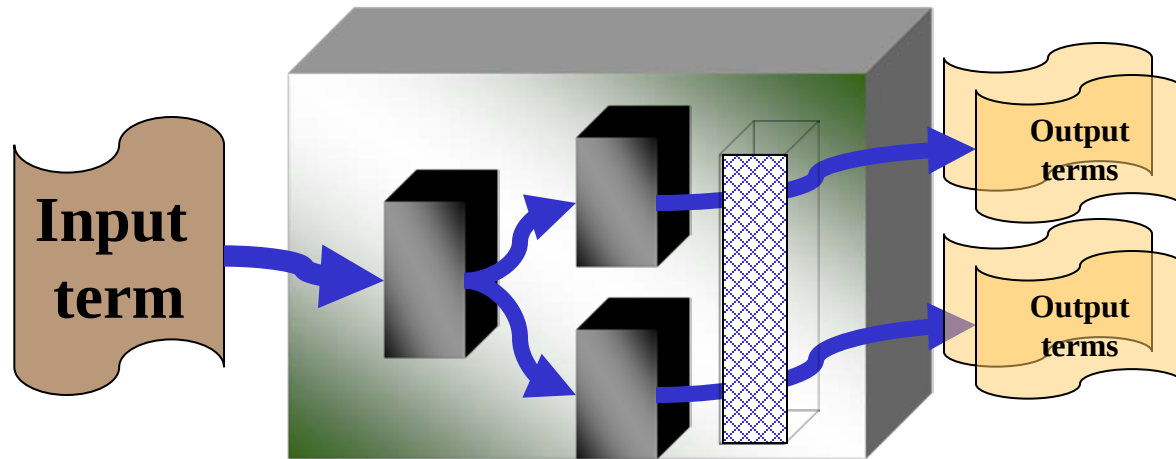
color | colour | 128 | 513 | s | 1 |

color | colour | 1024 | 263169 | s | 1 |

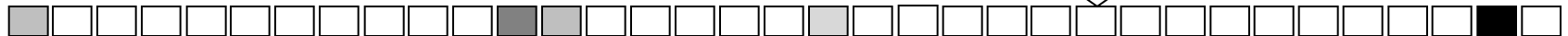
- Please input a term (type "Ctl-d" to Quit) >



Lexical Tools: Global Behaviors



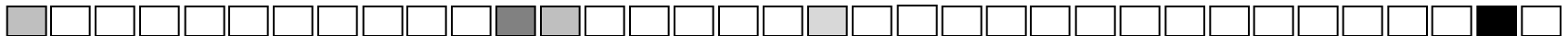
K:dInt	<u>Configure the derivation morphology behavior</u>
K:iInt	<u>Configure the inflection morphology behavior</u>



Lexical Tools: Global Behaviors

Configure the derivation morphology behavior

1	Restrict the output to those variants which are known to the lexicon (default).
2	Restrict the output to those variants which are known to the lexicon, unless none of the variants are found in the lexicon, in which case the entire (rule-generated) list is returned.
3	No restriction on the output of the morphology. Both facts and rules generated variants are displayed.



Lexical Tools: Global Behaviors

Configure the derivation morphology behavior

Restrict the derivations to terms that are known to the Lexicon

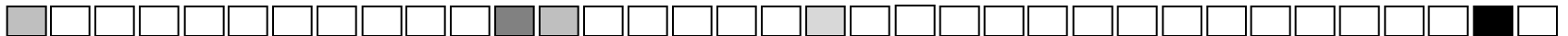
> lvg -f:d **-kd:1** -m

breath

breath | breather | 128 | 1 | d | 1 | RULE | \$ | verb | base | er\$ | noun | base |

breath | breathy | 1 | 1 | d | 1 | RULE | \$ | noun | base | y\$ | adj | base |

breath | breathless | 1 | 1 | d | 1 | FACT | breath | 128 | breathless | 1 |



Lexical Tools: Global Behaviors

Configure the inflection morphology behavior

1	Restrict the output to those variants which are known to the lexicon.
2	Restrict the output to those variants which are known to the lexicon, unless none of the variants are found in the lexicon, in which case the entire (rule-generated) list is returned (Default).
3	No restriction on the output of the morphology. Both facts and rules generated variants are displayed



Lexical Tools: Global Behaviors

Configure the inflection morphology behavior

> lvg -f:i -m **-ki:3** -R:5 -F:1:2:7:3:4 -SC -SI

Amish

Amish | Amish | FACT | <noun> | <base>

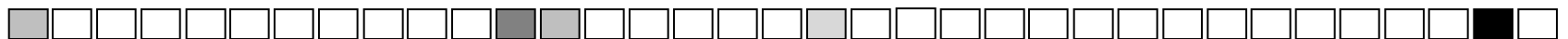
Amish | Amish | FACT | <noun> | <plural>

Amish | Amishs | RULE | <noun> | <plural>

Amish | Amishes | RULE | <noun> | <plural>

Amish | Amishs | RULE | <verb> | <pres>

Return all
inflected
variants,
both fact
and rule
generated



Lexical Tools: No Operation

-f:n

**Copies the input term to the output
with no transformation**

> lvg **-f:n** -f:d -f:y -SC -SI

force

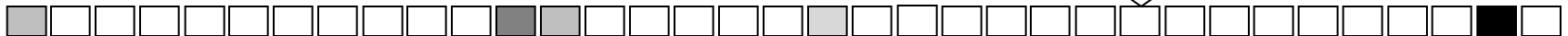
force | **force** | <all> | <all> | **n** | 1 |

force | forcefully | <adv> | <base> | d | 2 |

force | forceful | <adj> | <base> | d | 2 |

force | forcible | <adj> | <base> | d | 2 |

force | dynamic | <adj> | <base> | y | 3 |



Lexical Tools: Inflect

-f:i

Generate inflectional variants

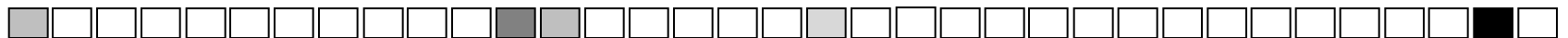
> lvg **-f:i** -SC -SI

West Nile Virus

West Nile Virus | **West Nile virus** | <noun> | <base> | i | 1 |

West Nile Virus | **West Nile virus** | <noun> | <singular> | i | 1 |

West Nile Virus | **West Nile viruses** | <noun> | <plural> | i | 1 |



Lexical Tools: Inflect

-f:ici~cats+infls

**Generate inflections, filter by cat
and/or inflection**

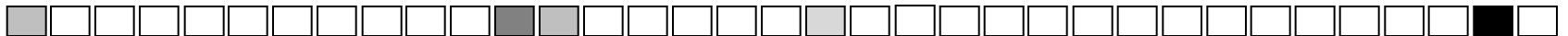
> lvg **-f:ici~128+ALL** -SC -SI

bioassay

bioassay | **bioassay** | <noun> | <base> | ici | 1 |

bioassay | **bioassay** | <noun> | <singular> | ici | 1 |

bioassay | **bioassays** | <noun> | <plural> | ici | 1 |



Lexical Tools: Uninflect by Term

-f:b

Uninflect by term

> lvg **-f:b** -SC -SI

left atria

left atria | **left atrium** | <noun> | <base> | b | 1 |



Lexical Tools: Derivations

-f:d

Generate derivations

> lvg **-f:d** -SC -SI

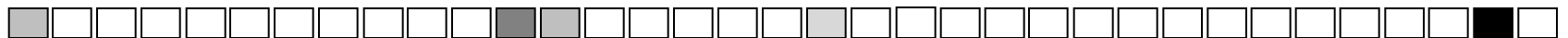
diagnostic

diagnostic | **diagnosis** | <noun> | <base> | d | 1 |

diagnostic | **diagnostics** | <noun> | <base> | d | 1 |

diagnostic | **diagnose** | <verb> | <base> | d | 1 |

diagnostic | **diagnostical** | <adj> | <base> | d | 1 |



Lexical Tools: Derivations

-f:dc~cats

Generate derivations, filter by category

> lvg **-f:dc~129** -SC -SI

reduce

reduce | **reduction** | <noun> | <base> | d | 1 |

reduce | **reducible** | <adj> | <base> | d | 1 |



Lexical Tools: Synonyms

-f:y

Generate synonyms

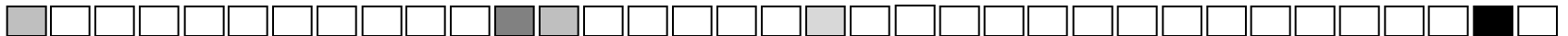
> lvg -f:y -SC -SI

kidney

kidney | **nephric** | <adj> | <base> | y | 1 |

kidney | **nephritic** | <adj> | <base> | y | 1 |

kidney | **renal** | <adj> | <base> | y | 1 |



Lexical Tools: Normalize (norm)

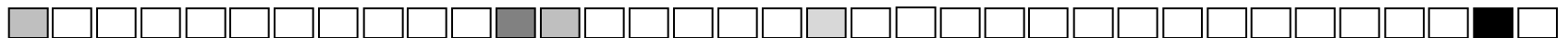
-f:N

Remove stop words, then remove genitives, then replace punctuation with spaces, then lowercase, then uninflect each word, then take each of the uninflected words, then word order sort.

> lvg -f:N

Syndrome, Dry Eyes

Syndrome, Dry Eyes|**dry eye syndrome**|2047|1|g+o+t+l+B+w|1|



Lexical Tools: Installation

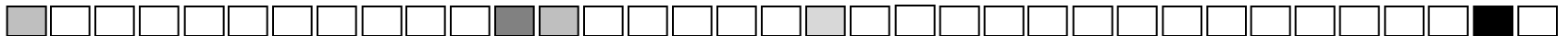
- Requirements
 - 1.6 gigabytes of space
 - Solaris/NT/Linux
 - Tar, gzip or WinZip
 - Minimum of 36 MB Memory
- Java JREs included



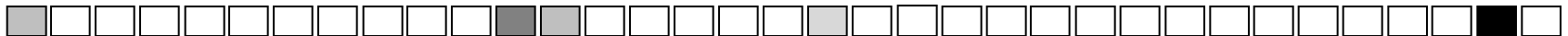
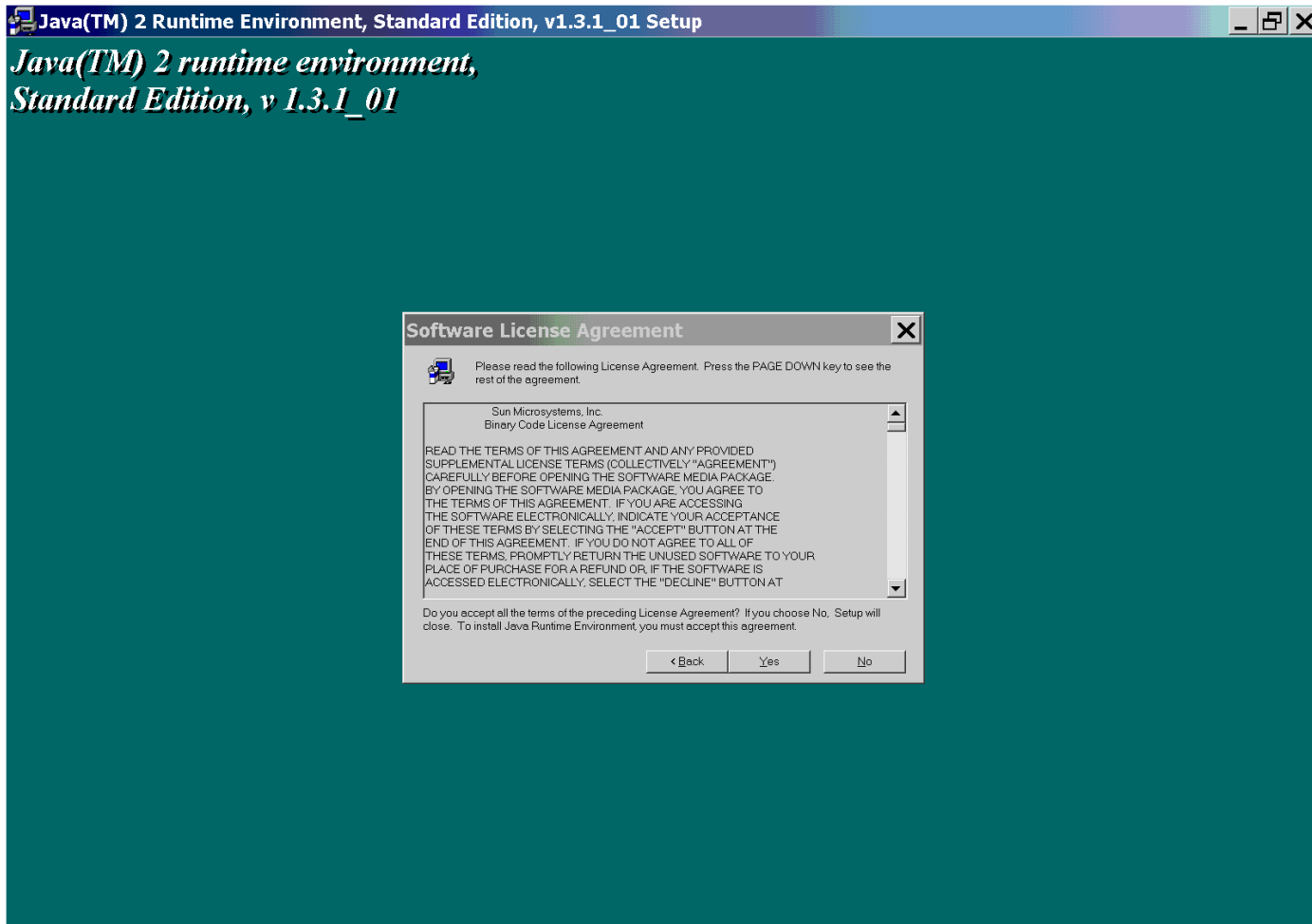
Lexical Tools: Installation

```
> ./install/bin/install_solaris_sparc.sh
```

```
> .\install\bin\install_win
```



Lexical Tools: Installation



Lexical Tools: Installation

Welcome to the Java Lexical Tools Installation!

Please read the *installationNotes.html* prior to invoking this script.

This script will configure the `_LVG_DIR_/data/config/lvg.properties` file.

This script will create configured `lvg`, `norm`, `luiNorm`, and `wordind` scripts in the `_LVG_DIR_` directory.

This script, as an option, will configure and load the lexical tools into a pre-existing MySQL database. This option is only available if you already have a MySQL database running.

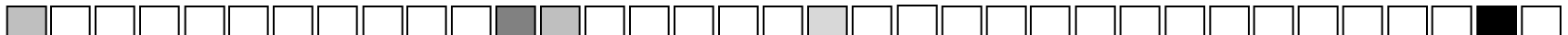
~~~~~

~~~~~  
The Java Lexical tools use or (may use) the following third party software packages:

- Instant DB from Enhydra.org
(see <http://instantdb.enhydra.org>)
This package is covered under an Enhydra Public License, which allows for redistribution and use with a "use notice".
- The installation script uses some GNU CYGWIN programs that have been distributed along with this script. These commands are only used when this installation is on a Windows platform. The commands that have been distributed include `tar`, `gunzip`, `rm`, `find`, `hosthame`, `tee`, and `cmp`.

The CYGWIN package is licensed under the GNU Public License. The entire CYGWIN package, along with the sources to these commands can be found at <http://sources.redhat.com/cygwin/>

Enhydra is a trademark of Lutris Technologies, Inc.



Lexical Tools: Installation

+-----Note-----+

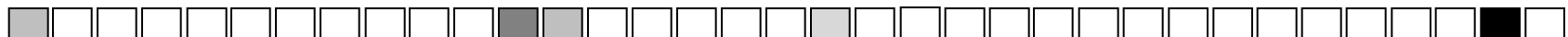
| Although we do not support this feature, and it |
| is very likely that we will change the underlying |
| implementation, we have found it useful for our |
| applications to use an existing database system |
| other than IDB. We often use Mysql as the underlying |
| database for our applications. Since we needed a |
| way to get the lexical tools data into MySQL, we |
| figured that we would incorporate the MYSQL loader |
| into this install. |

| This option requires that an existing mysql database |
| (version 3.23 or higher) have already been installed, |
| that you have database root access privileges, |
| that the database has enough room to load this |
| data, and that the database is running now. |

+-----+

+----- **Question** -----+

| Do you want to have the lexical tools use |
| your mysql database [y/n] [n]? **n**



Lexical Tools: Installation

```
+-----+
| Verifying the installation ... |
+-----+
```

Enhydra InstantDB - Version 3.26

The Initial Developer of the Original Code is Lutris Technologies Inc. Portions created by Lutris are Copyright (C) 1997-2001 Lutris Technologies, Inc. All Rights Reserved.

Database sample is shutting down...

Database sample shutdown complete.

~~~~~

~~~~~



Lexical Tools: Installation

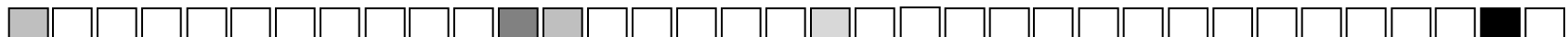
Congratulations!

This script has completed configuring the java lexical tools. You may invoke these tools from a command line. These tools are found in the `__LVG_DIR__/bin` directory.

You can add this `__LVG_DIR__/bin` path to your `$PATH` environment variable. This would enable you to find and run these tools from any location. In UNIX, this would be done by adding this path to your `~/.cshrc` or `~/.profile` startup script.

In Windows, this would be done by appending this path to the `PATH` variable from the control panel/System/Environment variables menus.

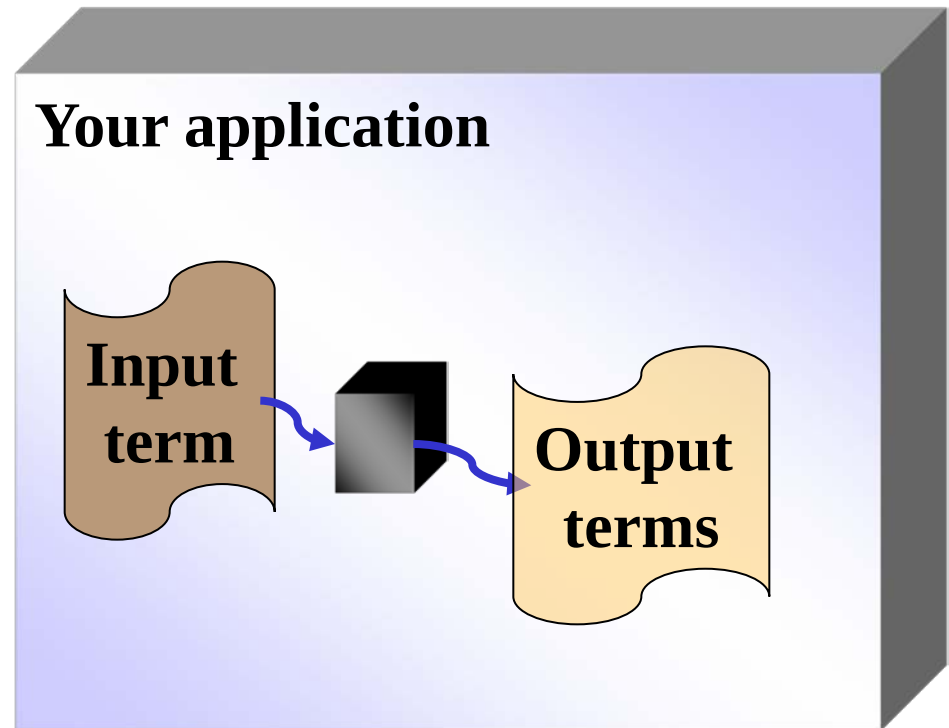
The lexical tools are ready to be used!



Lexical Tools:

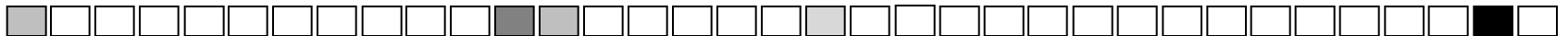
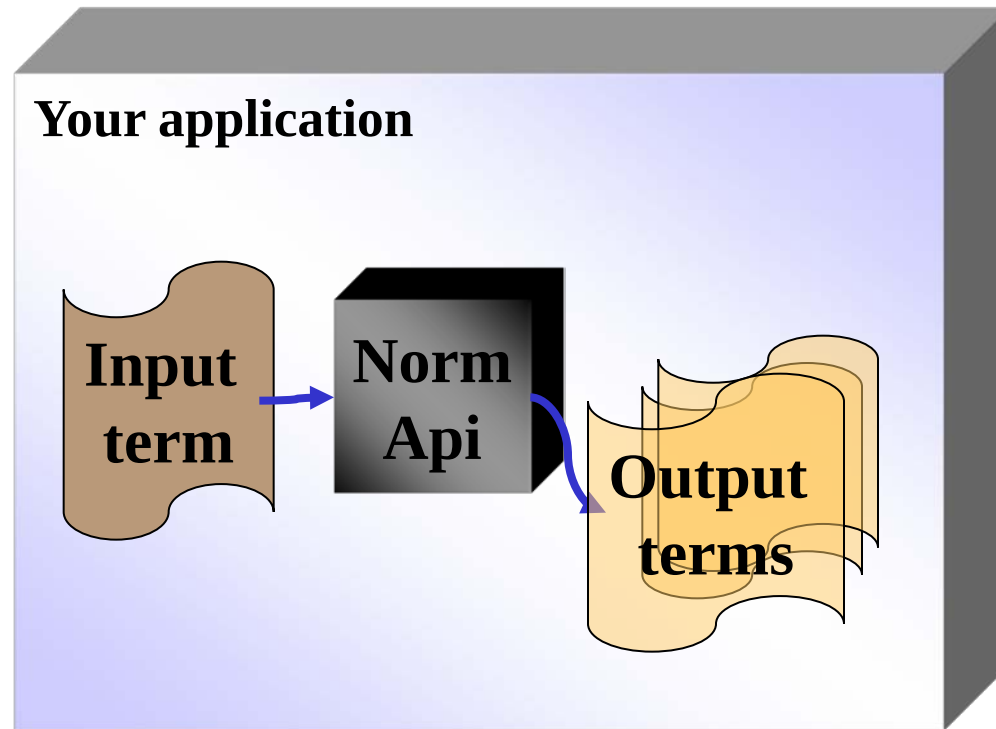
Embedding These Tools into Your Application

- Classpath
- NormApi()
- LvgCmdApi()
- LexItem
- LvgLexItemApi()



Lexical Tools:

Embedding Norm into Your Application



Lexical Tools:

Embedding These Tools into Your Application

CLASSPATH = \${CLASSPATH}:

\${LVG_DIR}:

\${LVG_DIR}/classes/*lvg2002.jar*:

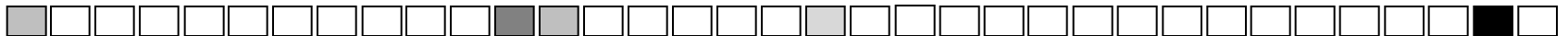
If you are
using IDB

\${LVG_DIR}/classes/IDB/jta-spec1_0_1.jar:

\${LVG_DIR}/classes/IDB/idb.jar:

If you
are using
MySQL

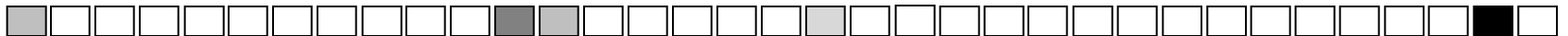
\${LVG_DIR}/classes/jdbcDrivers/mm.mysql-2.0.6



Lexical Tools:

Embedding Norm into Your Application

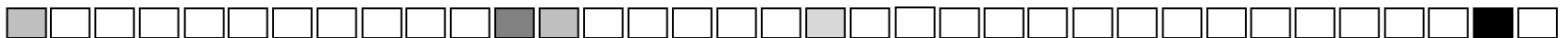
```
import Lvg.Api.*;  
  
NormApi    normalize = new NormApi();  
String      input2Norm = null;  
Vector      outputFromNorm = null;
```



Lexical Tools:

Embedding Norm into Your Application

```
while ( (input2Norm = stdIn.readLine() ) != null ) {  
    outputFromNorm= normalize.Mutate(input2Norm);  
    for ( int i = 0; i < outputFromNorm.size(); i++ ) {  
        System.out.println((String) outputFromNorm.get(i));  
    }  
}  
normalize.CleanUp();
```



Lexical Tools:

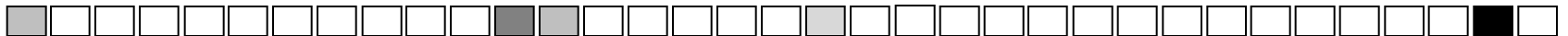
Embedding Lvg into Your Application



Input term

LvgCmd Api

Output terms



Lexical Tools:

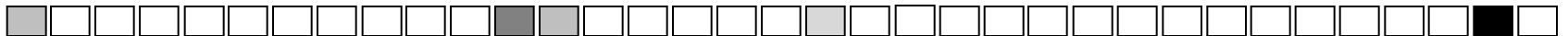
Embedding Lvg into Your Application

```
import Lvg.Api.*;
```

```
LvgCmdApi lvgApi = new LvgCmdApi("-f:b -CR:o -SC -SI");
```

```
String      input2Lvg = null;
```

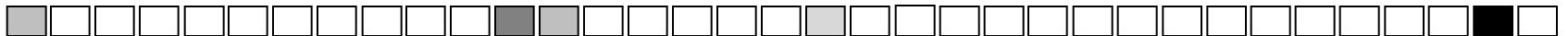
```
Vector  outputFromLvg = null;
```



Lexical Tools:

Embedding Lvg into Your Application

```
while ( (input2Lvg = stdIn.readLine() ) != null ) {  
    outputFromLvg= lvgApi.Mutate(input2Lvg);  
    for ( int i = 0; i < outputFromLvg.size(); i++ ) {  
        System.out.println((String) outputFromLvg.get(i));  
    }  
}  
  
lvgApi.CleanUp();
```



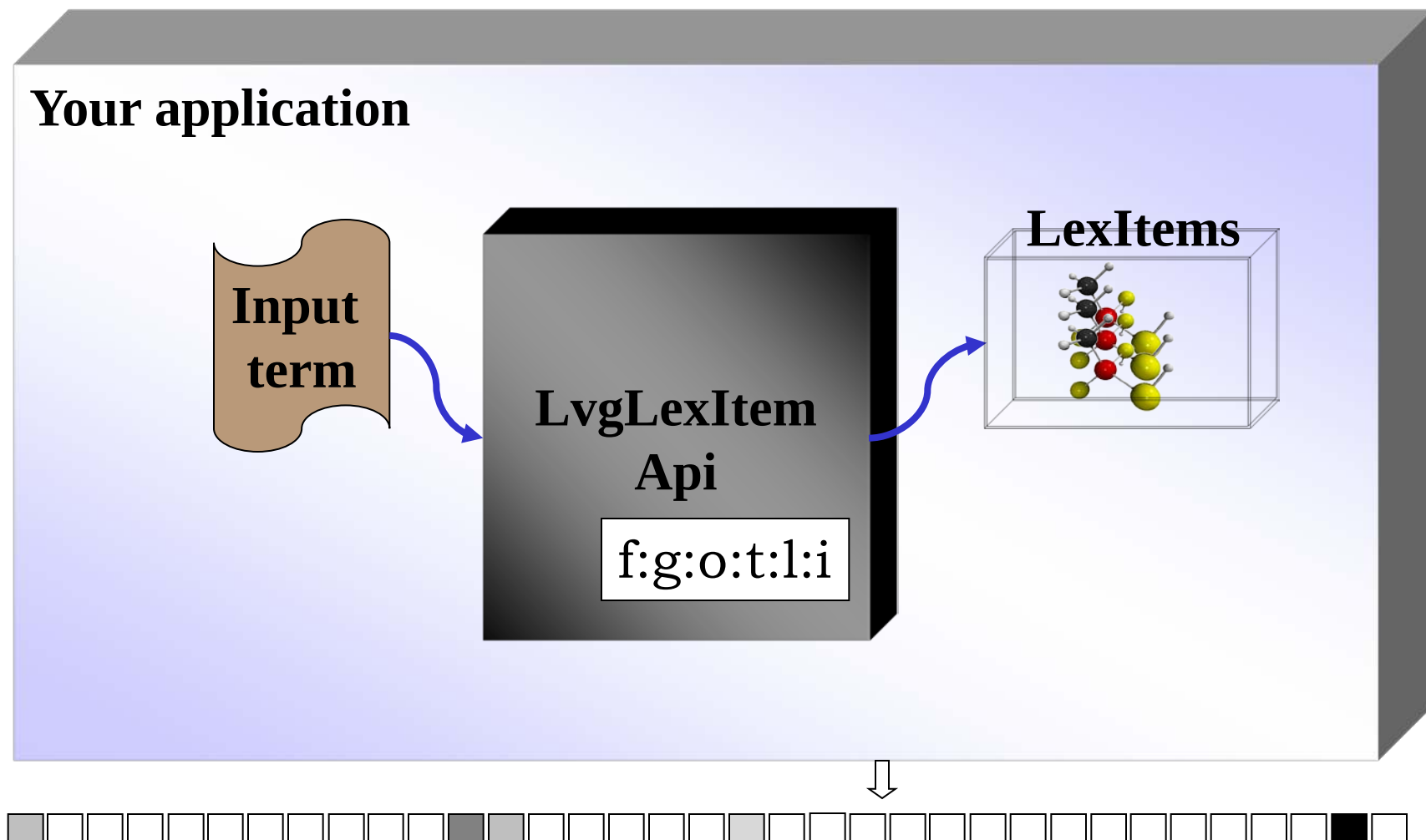
Lexical Tools:

Embedding Lvg into Your Application



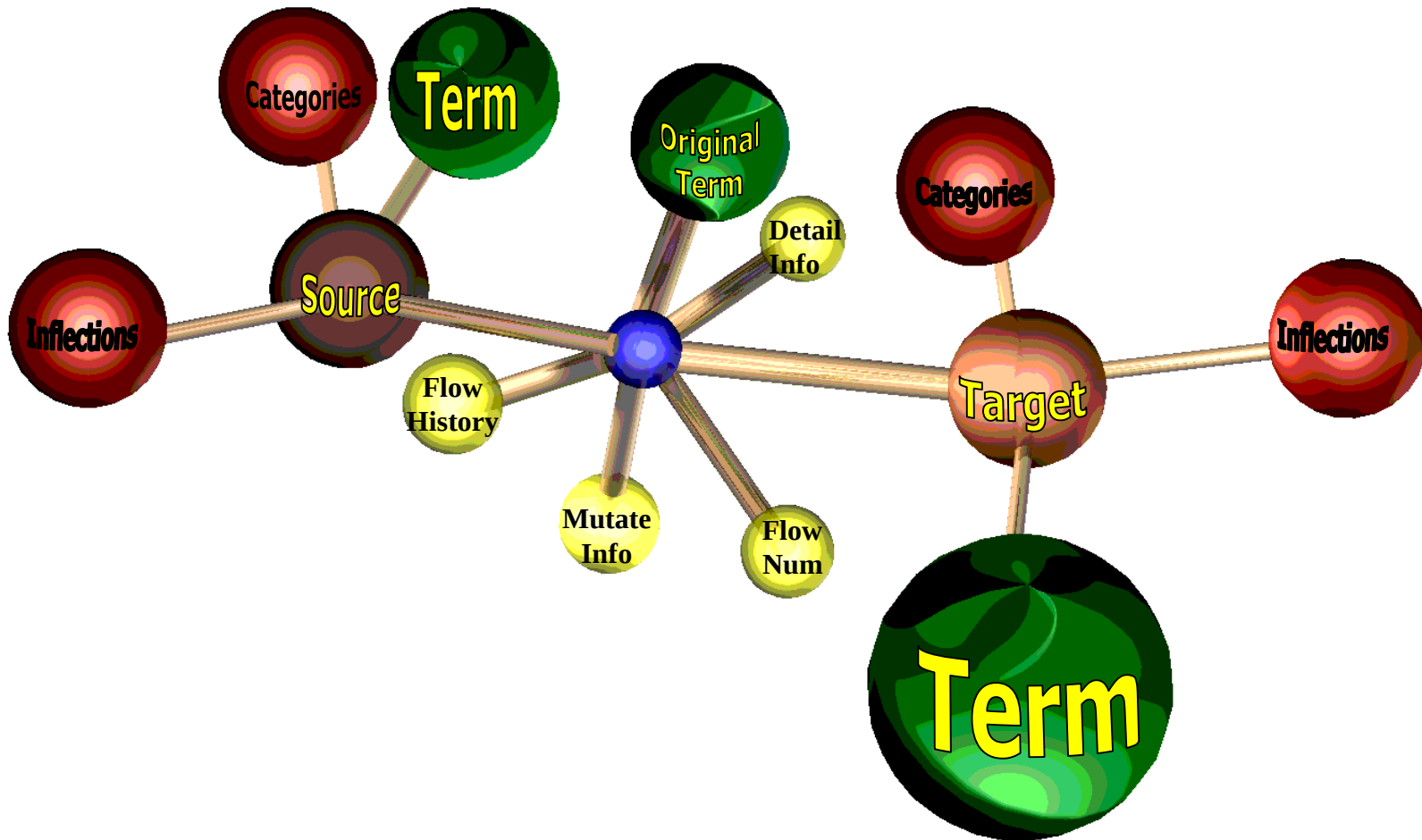
Lexical Tools:

Embedding Lvg into Your Application



Lexical Tools:

The LexItem Class

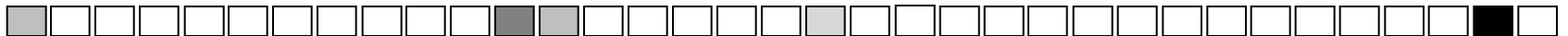


Lexical Tools:

Embedding Lvg into Your Application

```
import Lvg.Api.*;  
import Lvg.Lib.*;
```

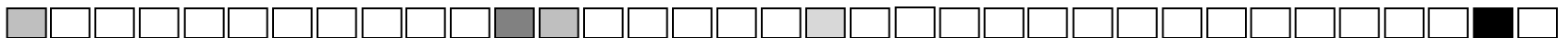
```
LvgCmdApi  lvgApi = new LvgLexItemApi("-f:g:o:t:l:i");  
String     input2Lvg = null;  
Vector     outputFromLvg = null;  
LexItem    aLexItem = null;
```



Lexical Tools:

Embedding Lvg into Your Application

```
while ( (input2Lvg = stdIn.readLine() ) != null ) {  
  
    outputFromLvg=lvglApi.MutateLexItem(input2Lvg);  
  
    for ( int i = 0; i < outputFromLvg.size(); i++ ) {  
        aLexItem = (LexItem) outputFromLvg.get(i);  
        System.out.println(aLexItem.GetSourceTerm() + "|" +  
            aLexItem.GetTargetTerm() + "|" +  
            aLexItem.GetTargetCategory().GetName()+ "|" +  
            aLexItem.GetTargetInflection().GetName() + "|" +  
            aLexItem.GetTargetInflection().GetValue() );  
    }  
}  
lvglApi.CleanUp();
```



Lexical Tools:

Java API Documentation

Packages	
<u>Lvg.Api</u>	This package provides APIs
<u>Lvg.CmdLineSyntax</u>	Provides Java classes necessary to create a command line system.
<u>Lvg.Db</u>	Provides a higher level interface to LVG database.
<u>Lvg.Flows</u>	This package provides API of all Lvg flow components.
<u>Lvg.Lib</u>	Contains LVG general library classes of BitMaskBase, Category, Gender, Inflection, Configuration, CombineRecords, OutputFilter, Flow, GlobalBehavior, LexItem, LexItemComparator
<u>Lvg.Trie</u>	Provides the classes necessary to generate inflections, uninflections, and derivations using LVG rules tries
<u>Lvg.Util</u>	Contains LVG general utility classes of comparators, Bit operations, Case, In, Out, Strip operations, token operation, Char, Str, and Word.



Lexical Resources



- Home
- Metathesaurus
- Semantic Network
- SPECIALIST Lexicon
- Expert Search
- Download Results
- Comments
- Help

 UMLS Knowledge Source Server

Other UMLS Resources

[Knowledge Sources](#)

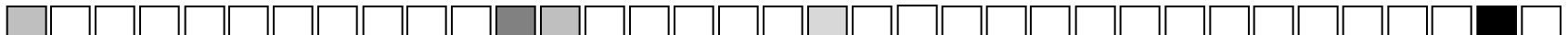
Source files for the Metathesaurus and Semantic Network and the SPECIALIST Lexicon.

[Knowledge Source Server API](#)

The UMLS Knowledge Source Server API and command line interface.

[Lexical Tools](#)

The Lexical Tools are utilities which manipulate lexical data in order to abstract away from a various kinds of lexical variation; inflection, inversion, alphabetic case, etc.



References

McCray AT, Loane RF, Browne AC, Bangalore AK. Terminology Issues in User Access to Web-based Medical Information. Proc AMIA Symp. 1998;;107-111

Divita G, Browne AC, Rindflesch T. Evaluating Lexical Variant Generation to Improve Information Retrieval, Proc AMIA Symp. 1998;;775-9

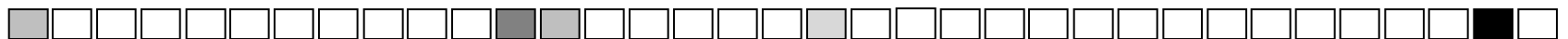
McCray AT, Browne AC. Discovering the modifiers in a terminology data set. Proc AMIA Symp. 1998;;780-4

McCray AT. The nature of lexical knowledge. Methods Inf Med. 1998 Nov;37(4-5):353-60.

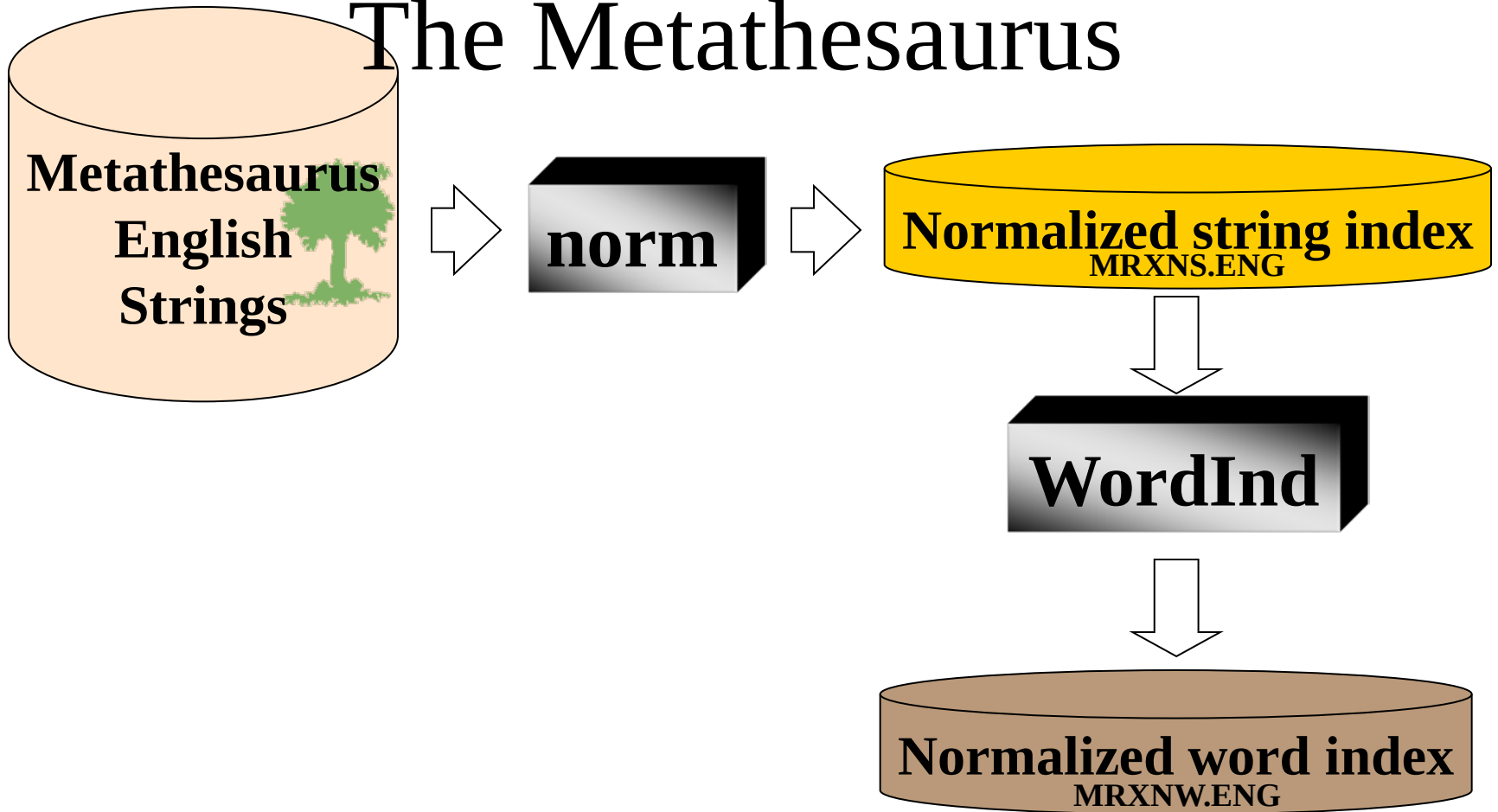
McCray AT, Cheh ML, Bangalore AK, Rafei K, Razi AM, Divita G, Stavri PZ. Conducting the NLM/AHCPR Large Scale Vocabulary Test: a distributed Internet-based experiment. Proc AMIA Annu Fall Symp. 1997;;560-4.

McCray AT, Razi, AM, Bangalore AK, Browne AC, Stavri PZ. The UMLS Knowledge Source Server: A versatile Internet-based research tool. Proc AMIA Symp. 1996;164-8.

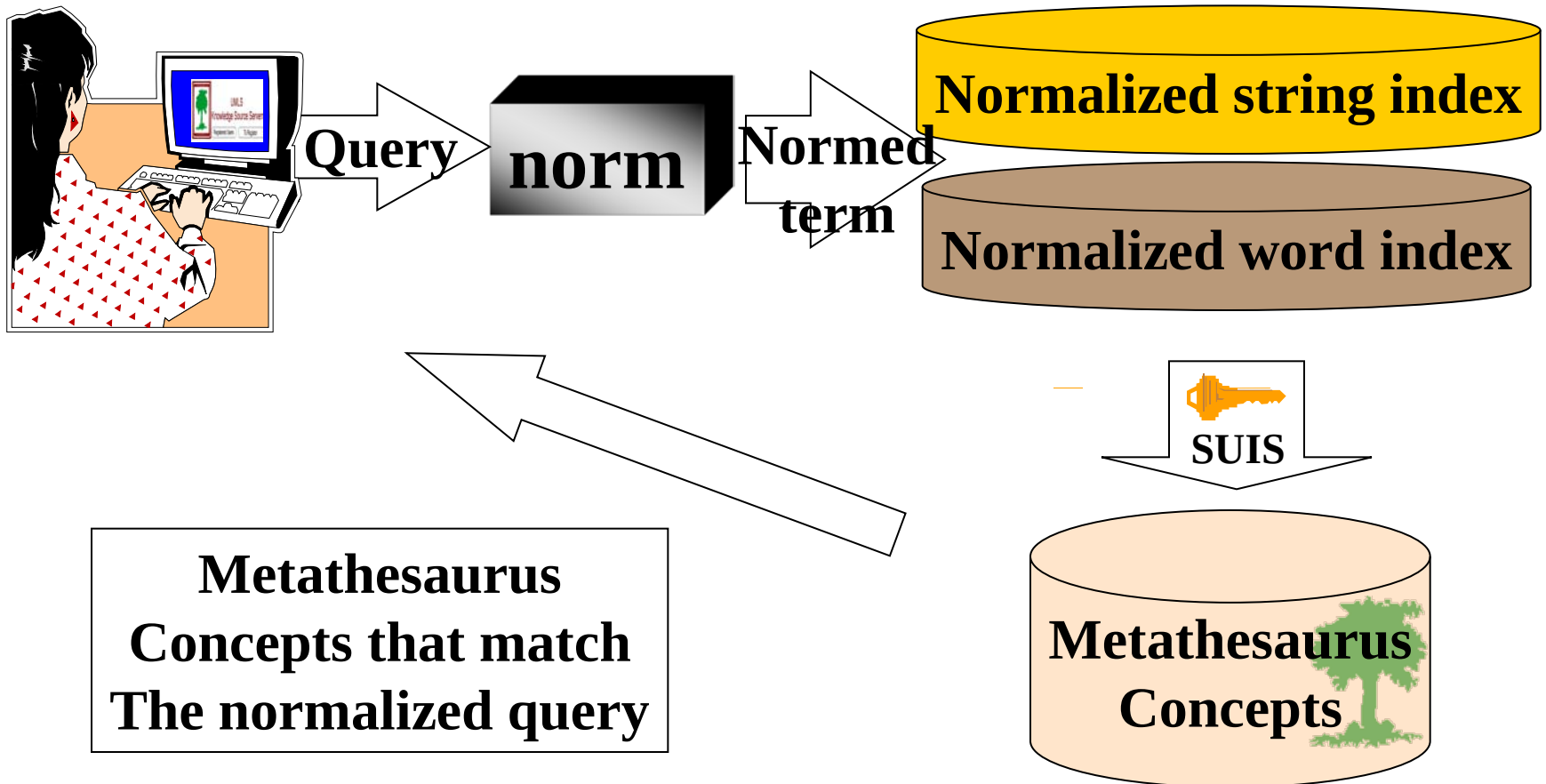
McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care. 1994;;235-9.



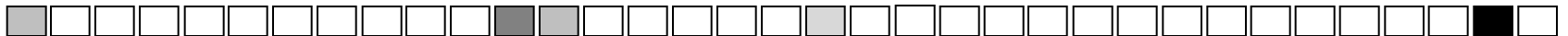
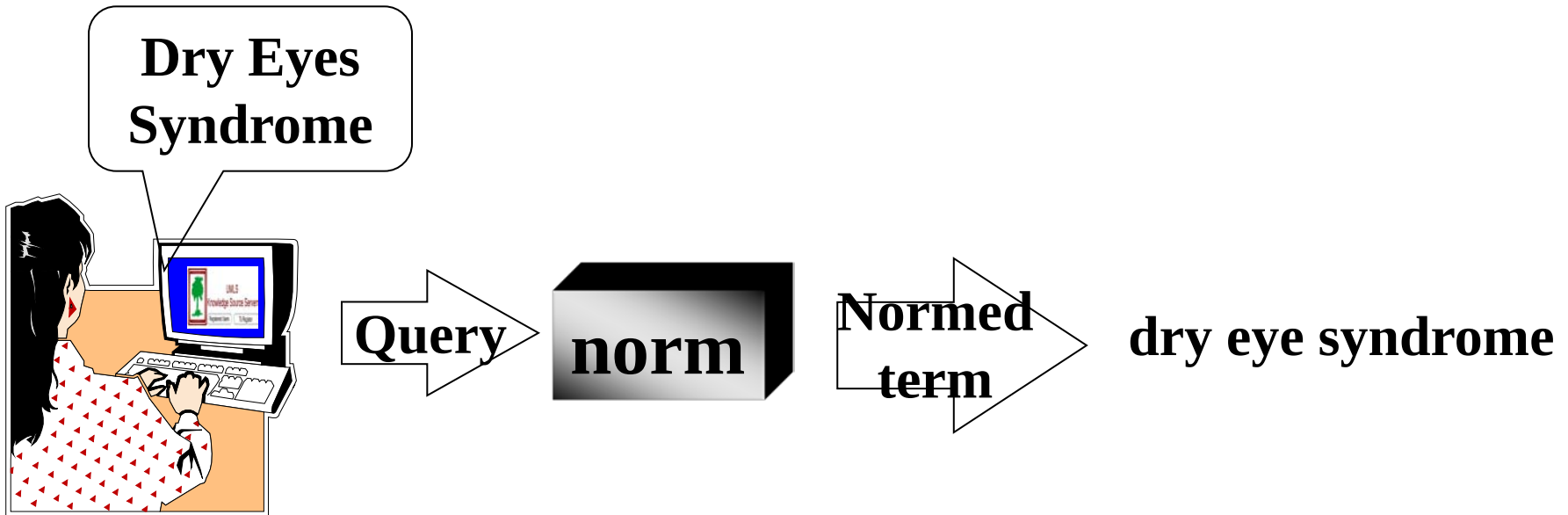
Using The Lexical Tools with The Metathesaurus



Using The Lexical Tools with The Metathesaurus



Using The Lexical Tools with The Metathesaurus

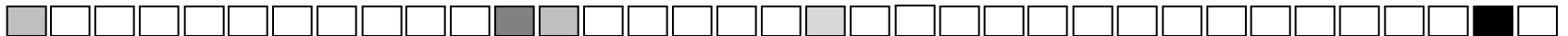


Using The Lexical Tools with The Metathesaurus

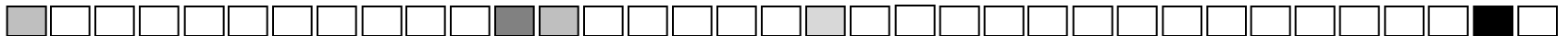
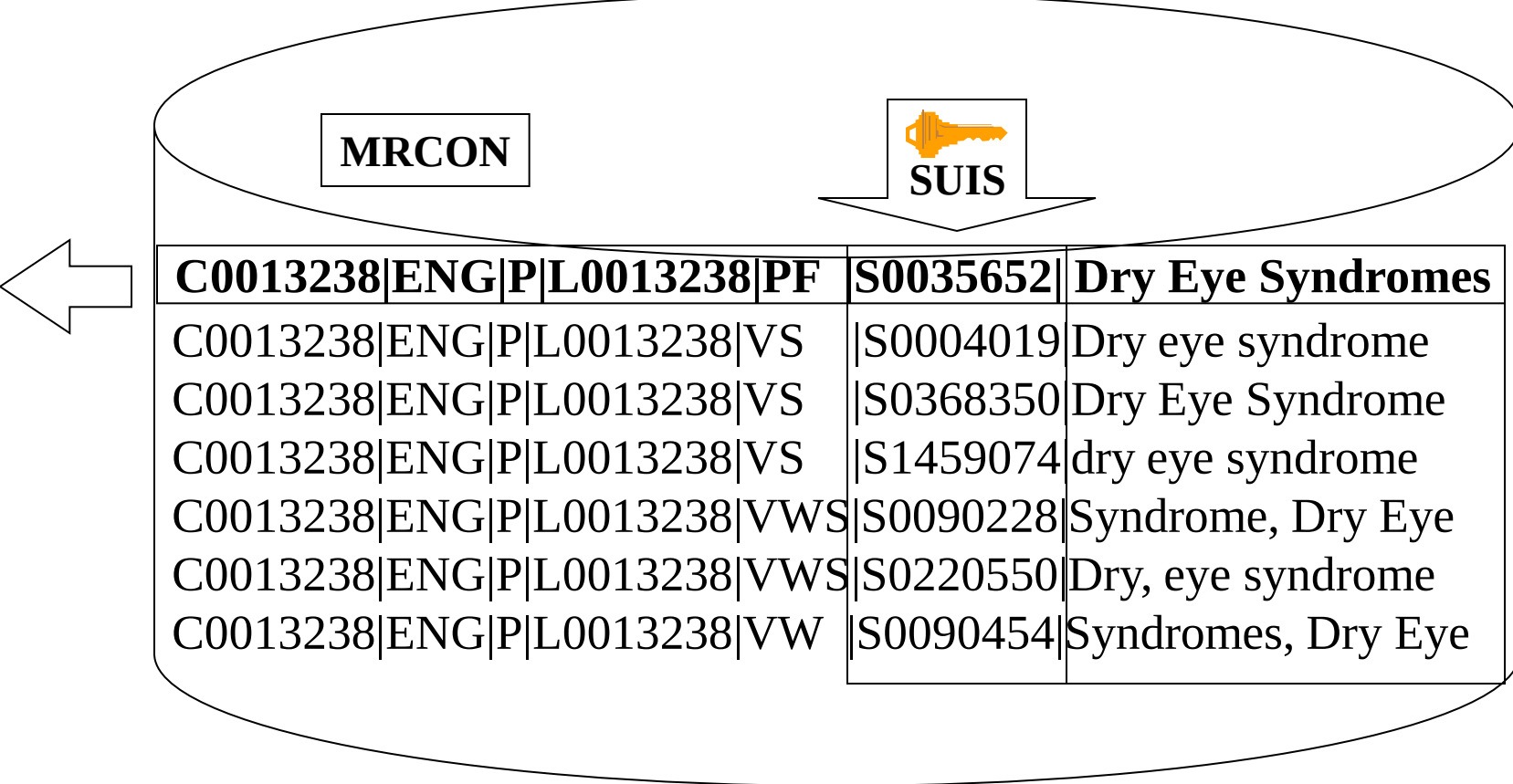
Normed term

SUIS

ENG	dry eye syndrome	C0013238	L0013238	S0004019
ENG	dry eye syndrome	C0013238	L0013238	S0035652
ENG	dry eye syndrome	C0013238	L0013238	S0090228
ENG	dry eye syndrome	C0013238	L0013238	S0090454
ENG	dry eye syndrome	C0013238	L0013238	S0220550
ENG	dry eye syndrome	C0013238	L0013238	S0368350
ENG	dry eye syndrome	C0013238	L0013238	S1459074



Using The Lexical Tools with The Metathesaurus



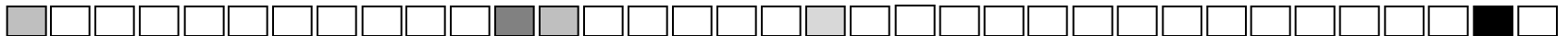
Building an Index Using The Lexical Tools

- Can we build a tool that increases recall?
- Can we build a tool that increases precision?



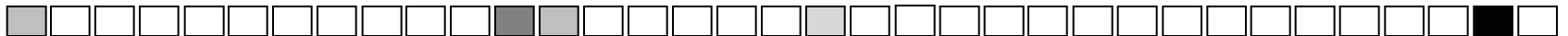
Building an Index Using The Lexical Tools

- Can we build a tool that increases precision?
 - Case
 - Constrain by part of speech
 - Filter to the lexicon

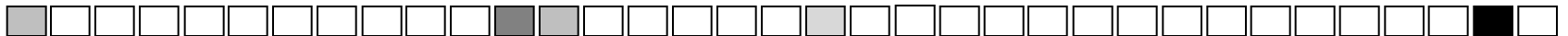
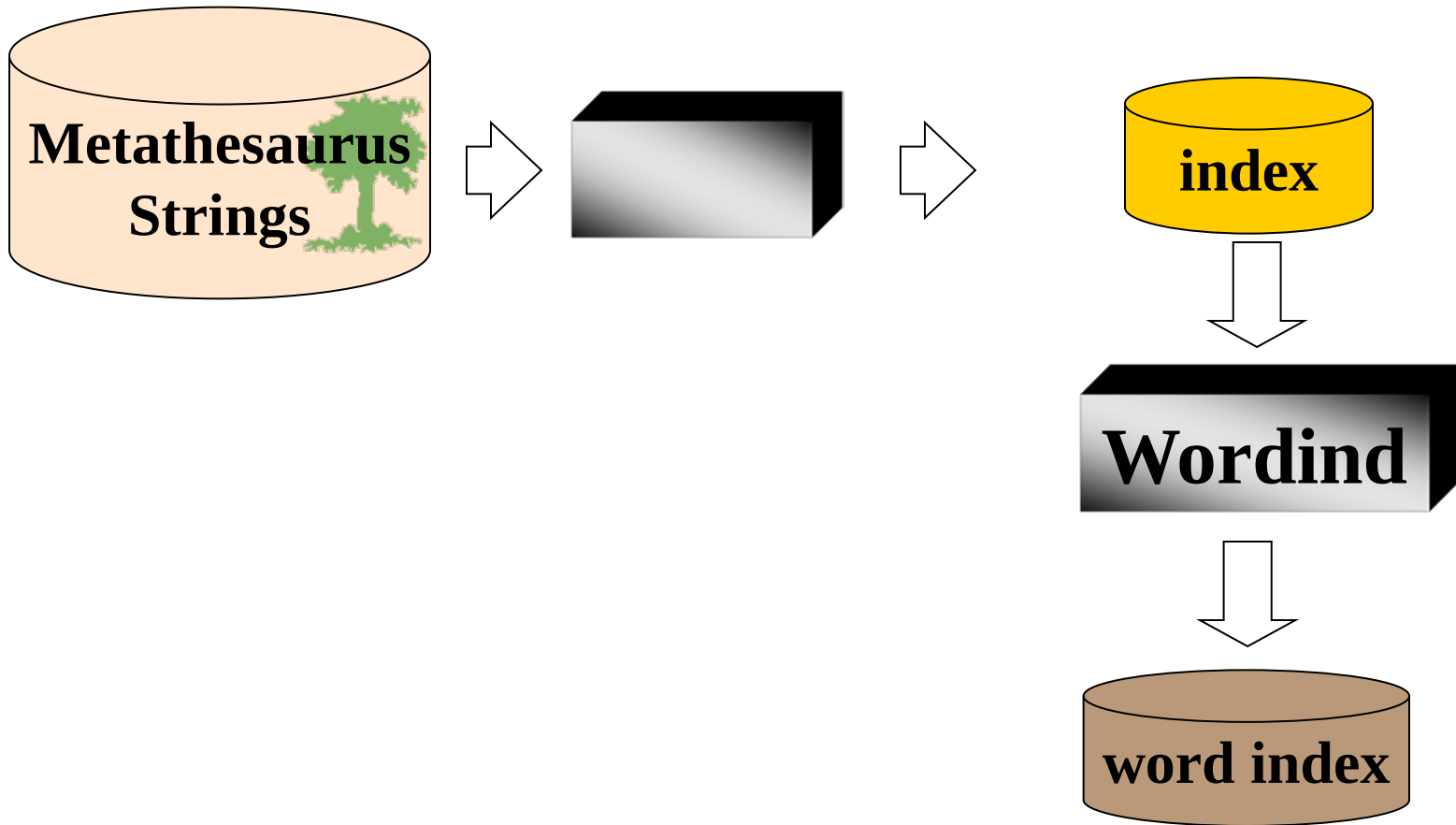


Building an Index Using The Lexical Tools

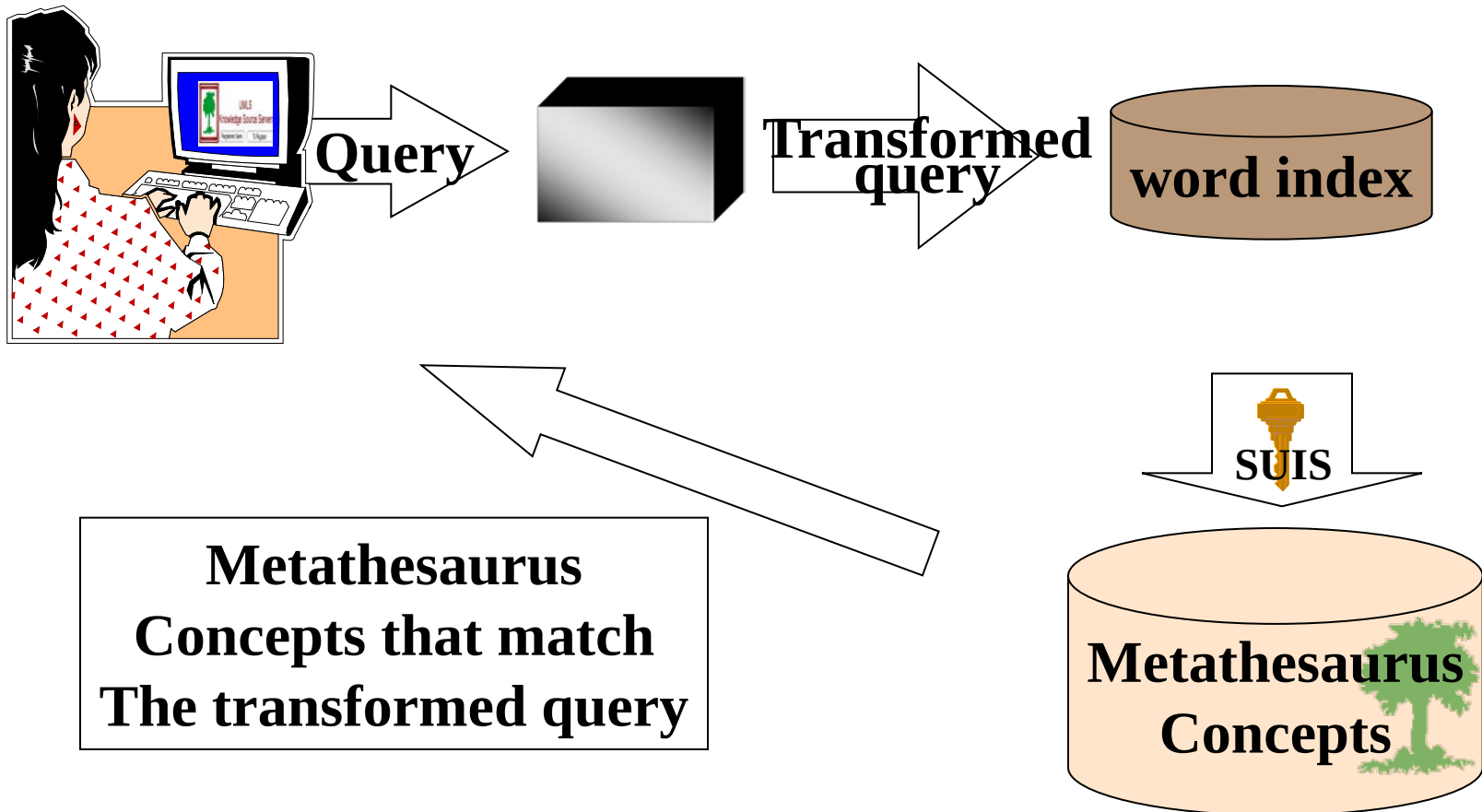
- Can we a tool that increases recall?
 - Include
 - synonyms
 - derivations
 - acronyms and their expansions
 - spelling variants



Building an Index Using The Lexical Tools



Building an Index Using The Lexical Tools



Lexical Tools:

2003 and Beyond

- Performance issues addressed
- Graphic User Interface
 - Lexical GUI tool
 - On-line Web based tool
- Enhance flow components
 - Norm improved
- New flow components
 - Nominalize, Complementation
- Additional functionalities
 - XML output option





Lexical Tools for UMLS Developers

November 4, 2001

Allen C. Browne, Guy Divita,
Chris Lu

Lister Hill National Center for Biomedical Communications

National Library of Medicine

Email: umlslex.nlm.nih.gov

Knowledge Source Server: <http://umlsks.nlm.nih.gov>

UMLS Information: <http://umlsInfo.nlm.nih.gov>



Appendix

NormExample.java

LvgExampleEasy.java

LvgExampleHarder.java

LvgExampleEvenHarder.java

