Lexical Tools Briefing

The Lexical Systems Group

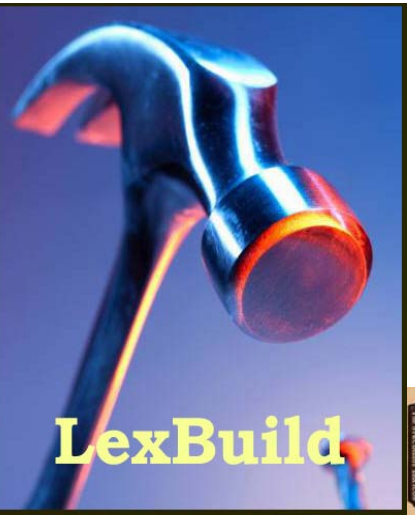
NLM. LHNCBC. CGSB

June, 2006

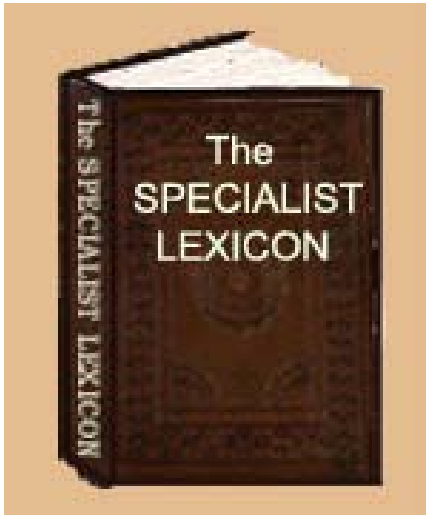
Table of Contents

Introduction

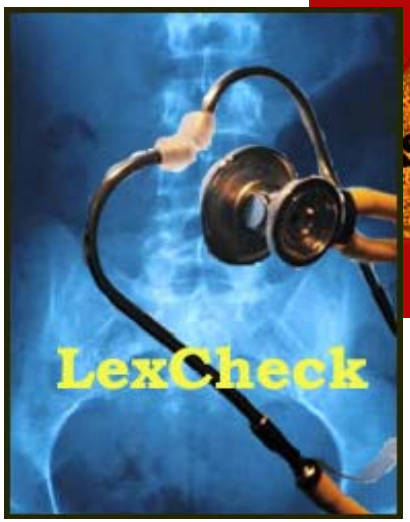

Introduction - LB



Introduction - Lexicon



Introduction - LC



Introduction - LA

Introduction - Numbers

Introduction - SCRT

Introduction – Lexical Tools

Introduction - GSpell

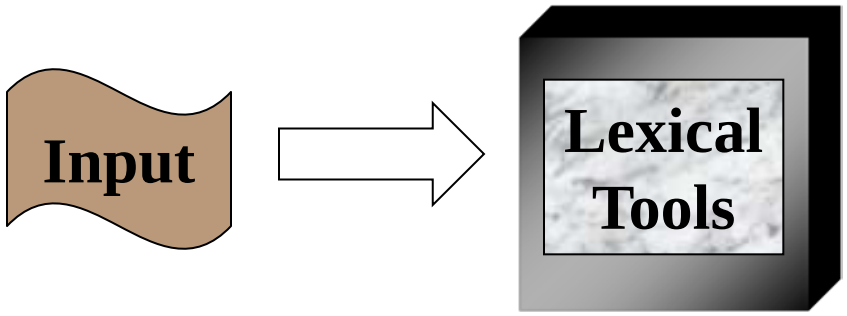
Introduction – Text Tools

Introduction - TC

Lexical Tools

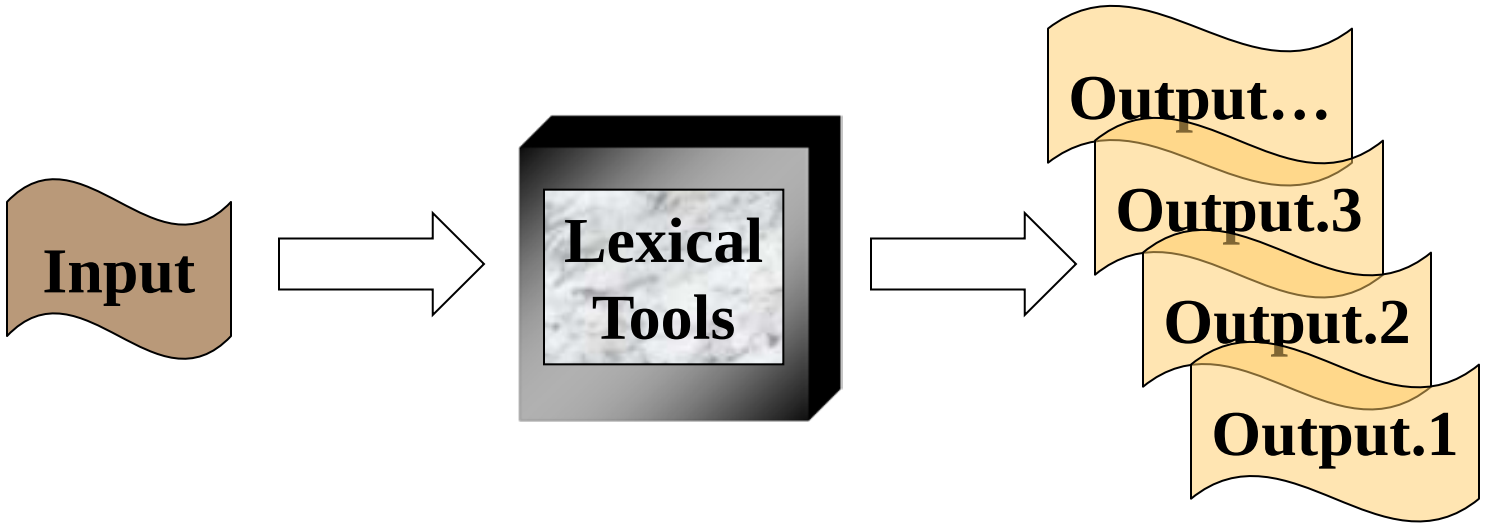
- A suite of text utilities

Lexical Tools



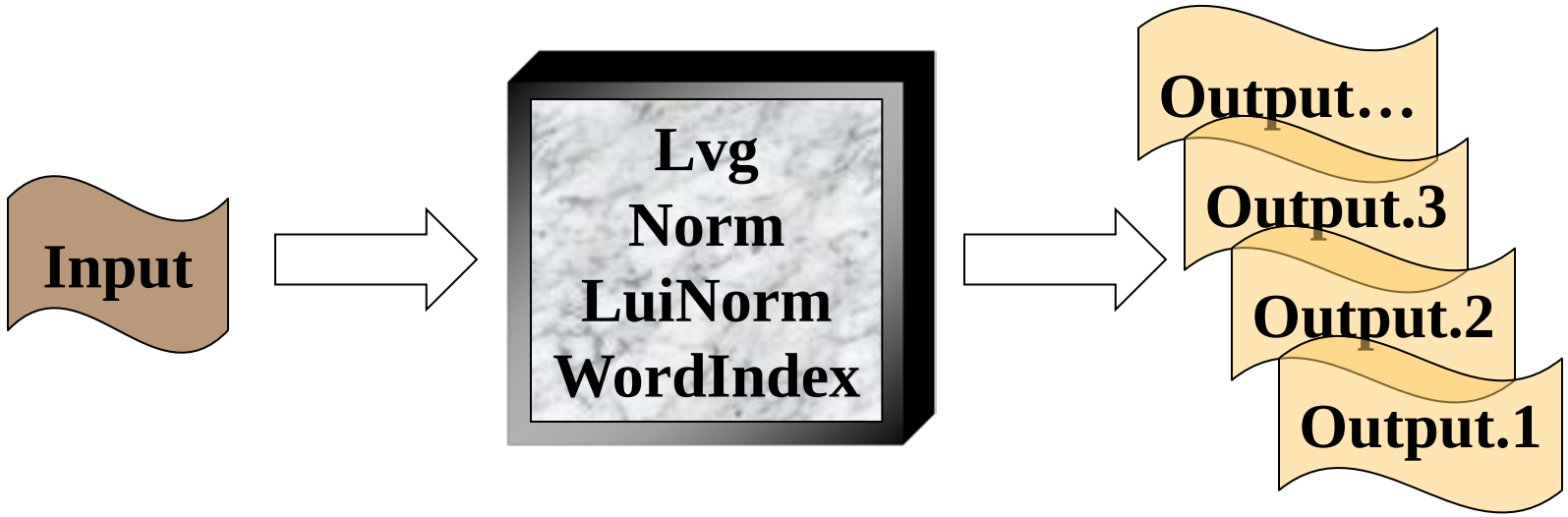
- A suite of text utilities take the given input

Lexical Tools



- A suite of text utilities that generate, mutate, and filter out lexical variants from the given input

Four Tools


Input
Lvg
Norm
LuiNorm
WordIndex
Output…
Output.3
Output.2
Output.1

Tool Types

- Command line tools
  - lvg (Lexical Variants Generation)
  - norm
  - luiNorm
  - wordInd
- Lexical Gui Tool (lgt)
- Web Tools
- Java API's

Functions

- Used in nature language processing for
  - aggressive text pattern matching
  - creating normalized and expanded terms
  - making word, term, phrase indexes
  - matching queries with indexed entries
  - increasing recall and/or precision

Facts

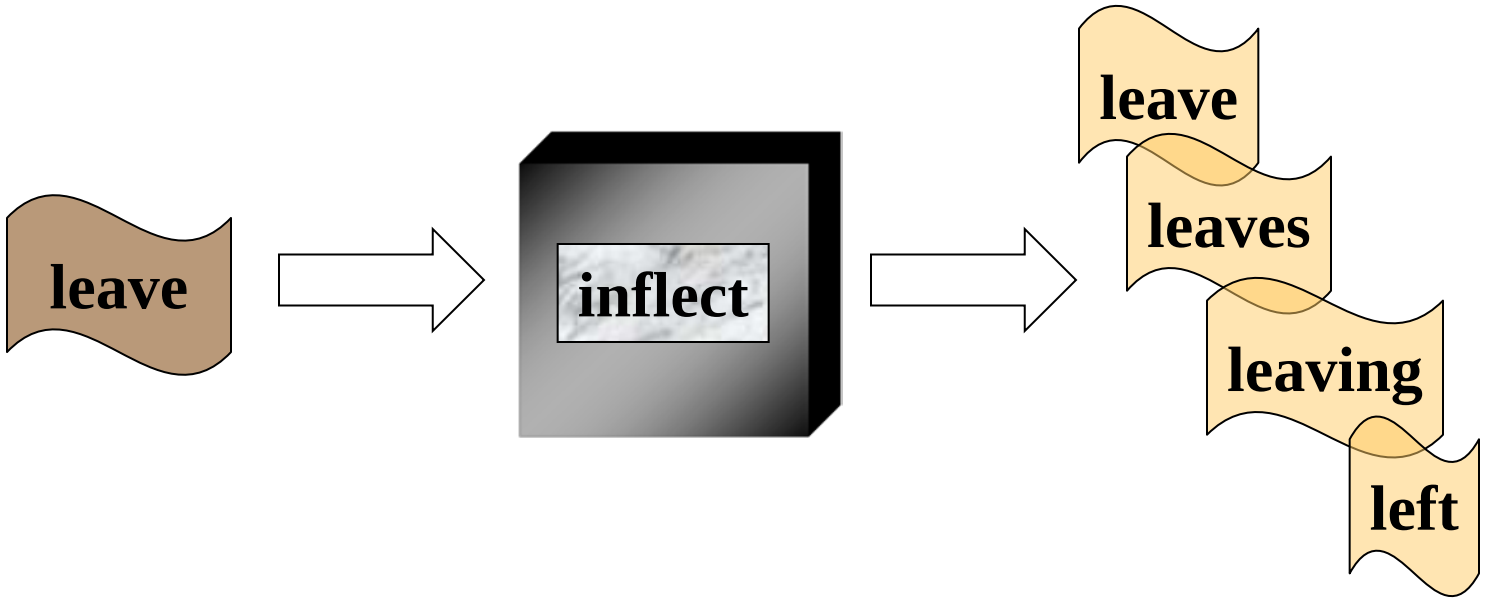
- Release annually
- 100% Java (since 2002)
- Free distributed with open source code
- Run on different platforms
- One complete package
- Documents & support

Lexical Variants Generation

LVG, 2006

- 58 flow components
- 37 options
  - input filter options (3)
  - global behavior options (13)
  - flow specific options (2)
  - output filter options (19)

Flow Components

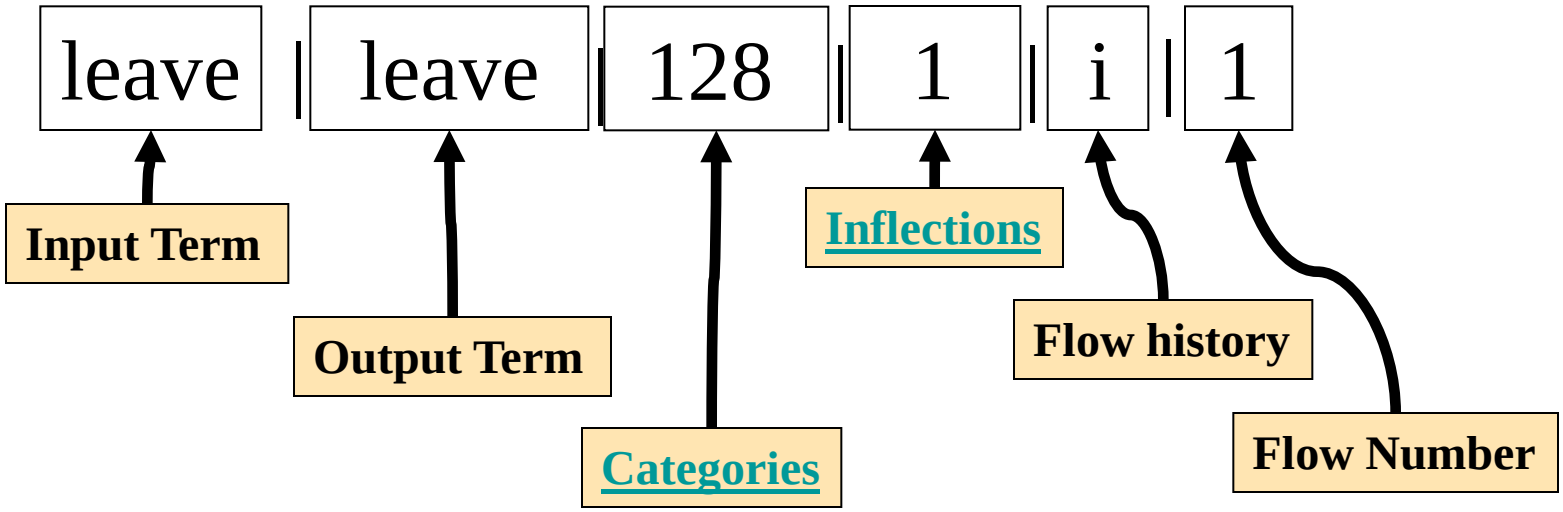


Command Line Tool

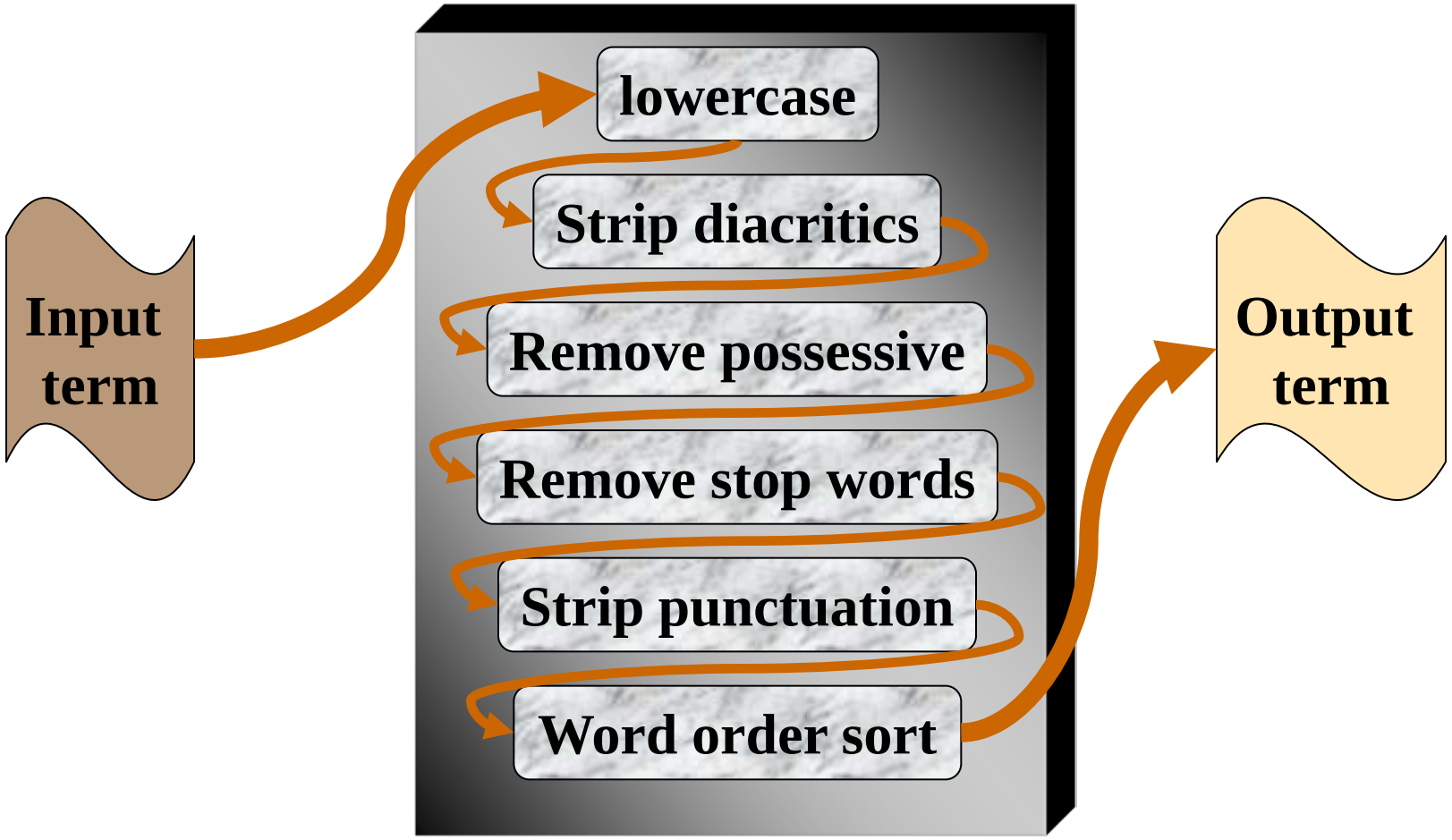
```
> lvg -f:i
leave
leave|leave|   |  |i|1|
leave|leave|128|512|i|1|
leave|leaves|128|8|i|1|
leave|left|1024|64|i|1|
leave|left|1024|32|i|1|
leave|leave|1024|1|i|1|
leave|leave|1024|262144|i|1|
leave|leave|1024|1024|i|1|
leave|leaves|1024|128|i|1|
leave|leaving|1024|16|i|1|
```

Fielded Output

> lvg –f:i
leave



A Serial Flow

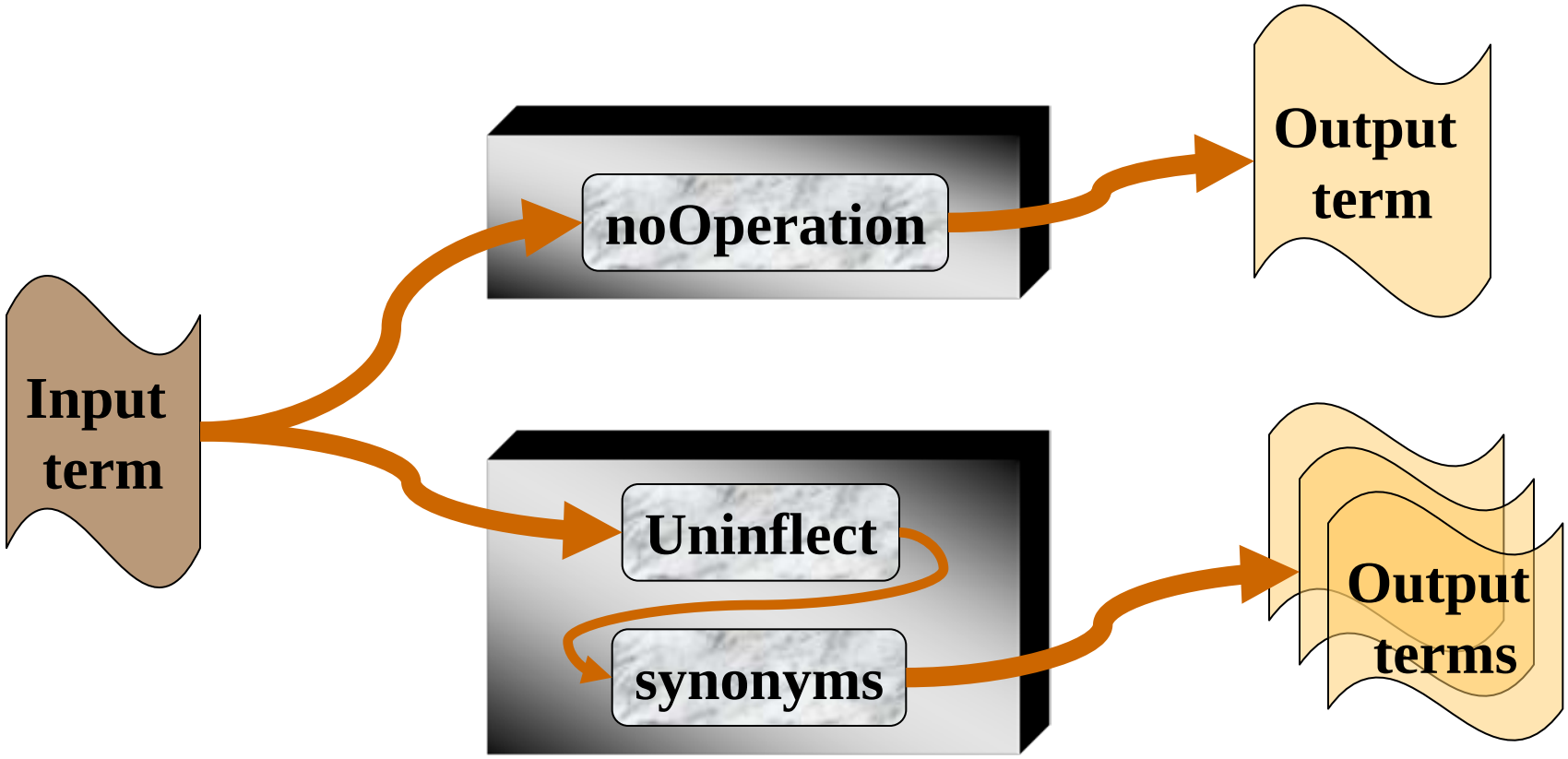


- Flow components can be arranged so that the output of one is the input to another.

A Serial Flow - Example

```
> 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|
```

Parallel Flows



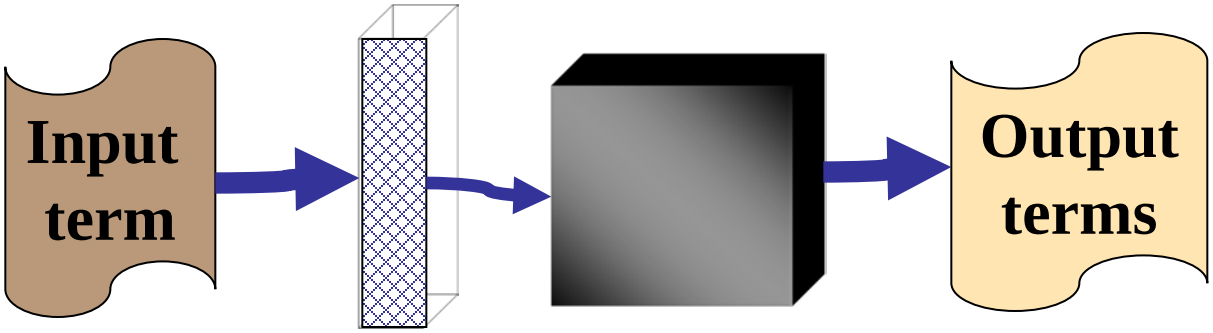
- Multiple flows can be defined

Parallel Flows - Example

```
> lvg –f:n –f:B:y
ear
ear|ear|2047|1048575|n|1|

ear|aural|1|1|B+y|2|
ear|auricularis|1|1|B+y|2|
ear|otic|1|1|B+y|2|
ear|otor|1|1|B+y|2|
```

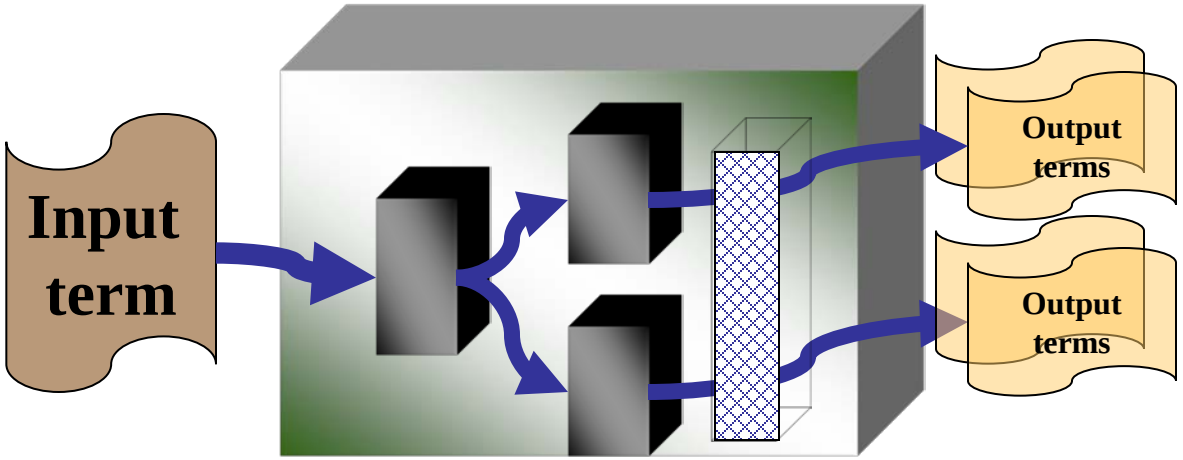
Input Filter Options



Take field 7 from the input

```
> lvg -f:u -t:7 -F:8:6
C0035440|ENG|S|L0035434|VW|S0003894|
   Rheumatic carditis, acute
acute Rheumatic carditis|S0003894
```

Global Behavior Options




```
> lvg -f:L –f:E -s:"\"
```

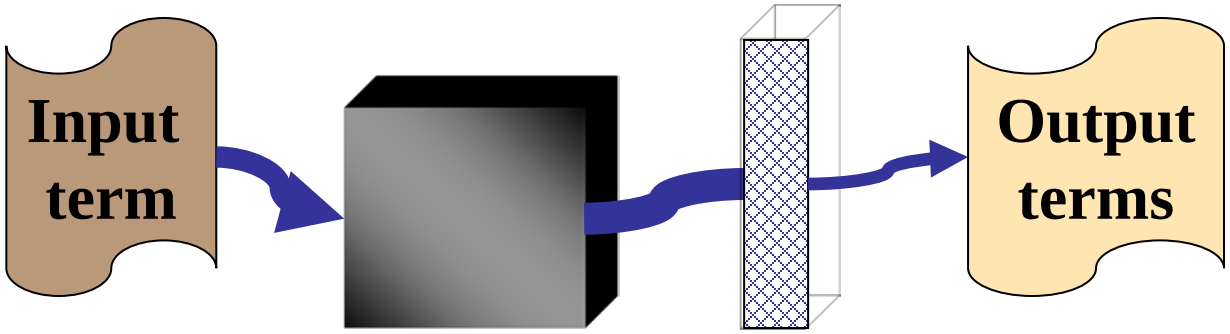
Change separator to "\"

```
otitis

otitis\otitis\128\513\L\1

otitis\E0044452\128\513\E\2
```

Output Filter Options




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

Show the category and inflection names

```
hot

hot|hot|<adj+verb>|<base+positive+infinitive+pres1p23p>|L|1|
```

Norm

- Composed of 11 Lvg flow components to abstract away from:
  - case
  - punctuation
  - possessive forms
  - inflections
  - spelling variants
  - stop words
  - diacritics & ligatures
  - word order

Norm

g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
q: strip diacritics
q2: split ligature
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
w: sort words by order
q4: get symbol names synonymy

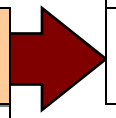
Norm

Hodgkin's Diseases, NOS

g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
q: strip diacritics
q2: split ligature
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
w: sort words by order
q4: get symbol names synonymy

Norm

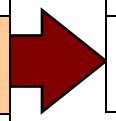
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
q: strip diacritics
q2: split ligature
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
w: sort words by order
q4: get symbol names synonymy



Hodgkin's Diseases, NOS
Hodgkin Diseases, NOS

Norm

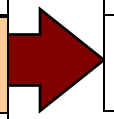
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
q: strip diacritics
q2: split ligature
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
w: sort words by order
q4: get symbol names synonymy



Hodgkin's Diseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases, NOS

Norm

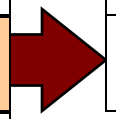
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
q: strip diacritics
q2: split ligature
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
w: sort words by order
q4: get symbol names synonymy



Hodgkin's Diseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases NOS

Norm

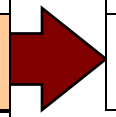
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
q: strip diacritics
q2: split ligature
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
w: sort words by order
q4: get symbol names synonymy



Hodgkin's Diseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases NOS
Hodgkin Diseases

Norm

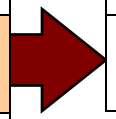
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
q: strip diacritics
q2: split ligature
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
w: sort words by order
q4: get symbol names synonymy



Hodgkin's Diseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases NOS
Hodgkin Diseases
Hodgkin Diseases

Norm

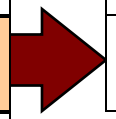
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
q: strip diacritics
q2: split ligature
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
w: sort words by order
q4: get symbol names synonymy



Hodgkin's Diseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases NOS
Hodgkin Diseases
Hodgkin Diseases
Hodgkin Diseases

Norm

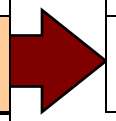
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
q: strip diacritics
q2: split ligature
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
w: sort words by order
q4: get symbol names synonymy



Hodgkin's Diseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases NOS
Hodgkin Diseases
Hodgkin Diseases
Hodgkin Diseases
hodgkin diseases

Norm

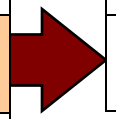
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
q: strip diacritics
q2: split ligature
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
w: sort words by order
q4: get symbol names synonymy



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

Norm

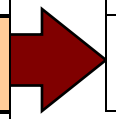
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
q: strip diacritics
q2: split ligature
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
w: sort words by order
q4: get symbol names synonymy



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

Norm

g: remove genitives	Hodgkin's Diseases, NOS
rs: remove parenthetic plural forms	Hodgkin Diseases, NOS
o: replace punctuation with spaces	Hodgkin Diseases, NOS
t: strip stop words	Hodgkin Diseases NOS
q: strip diacritics	Hodgkin Diseases
q2: split ligature	Hodgkin Diseases
l: lowercase	Hodgkin Diseases
B: uninflect each words in a term	hodgkin diseases
Ct: retrieve citations	hodgkin disease
w: sort words by order	hodgkin disease
q4: get symbol names synonymy	disease hodgkin

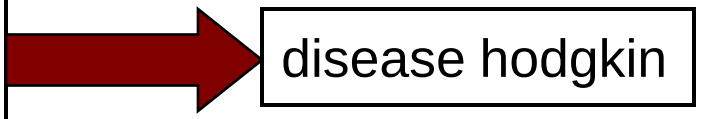


Norm

	Hodgkin's Diseases, NOS
g: remove genitives	Hodgkin Diseases, NOS
rs: remove parenthetic plural forms	Hodgkin Diseases, NOS
o: replace punctuation with spaces	Hodgkin Diseases NOS
t: strip stop words	Hodgkin Diseases
q: strip diacritics	Hodgkin Diseases
q2: split ligature	Hodgkin Diseases
l: lowercase	hodgkin diseases
B: uninflect each words in a term	hodgkin disease
Ct: retrieve citations	hodgkin disease
w: sort words by order	disease hodgkin
q4: get symbol names synonymy	disease hodgkin

Norm: Example

- 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



Text Categorization

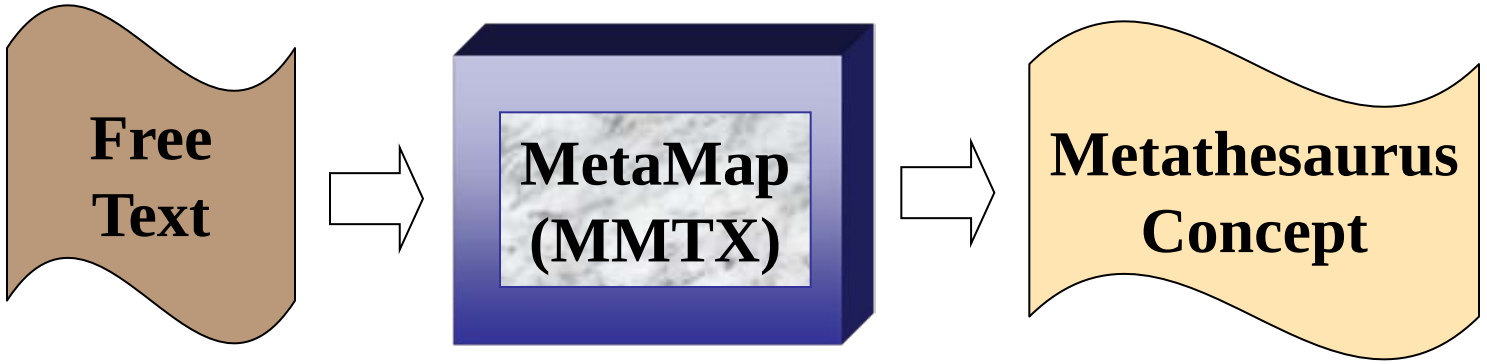
- Based on Journal Descriptor Indexing (JDI) methodology
- Uses a small set of high level descriptors, such as Journal Descriptors (JDs), Semantic Types (STs), Mesh subcategories, etc..
- For categorize text, index contents, retrieve records, and word sense disambiguation

Text Categorization

- Free distributed with open source code
- 100 % in Java
- Run on different platforms
- One complete package
- Documents & support
- Provides Java APIs, command line tools, GUI tools, and Web tools
- Planned first release, TC 2007

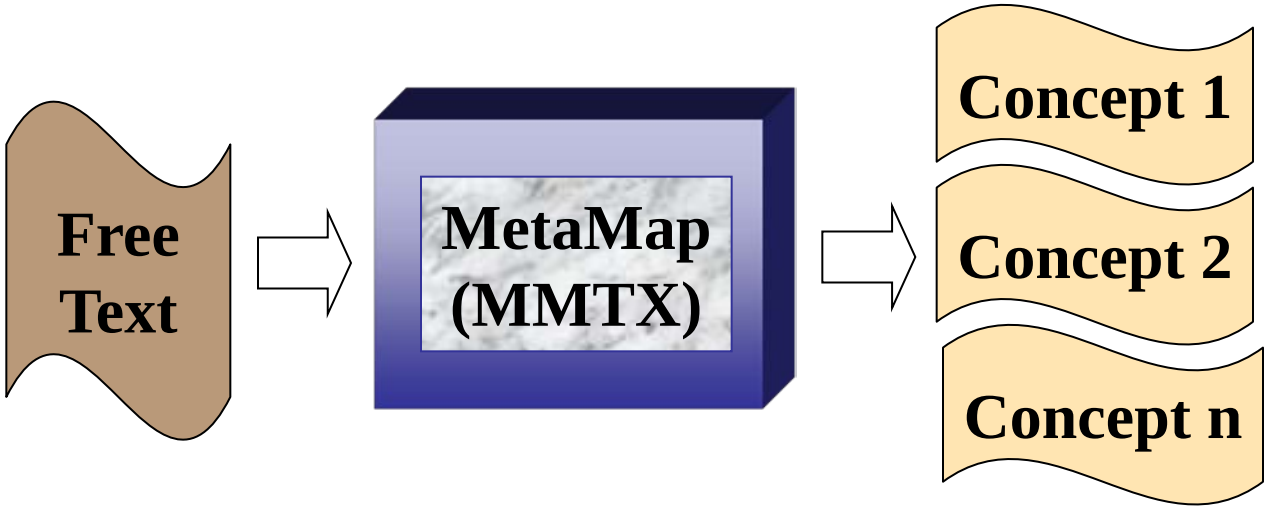
Text Categorization

- Words Senses disambiguation (WSD)



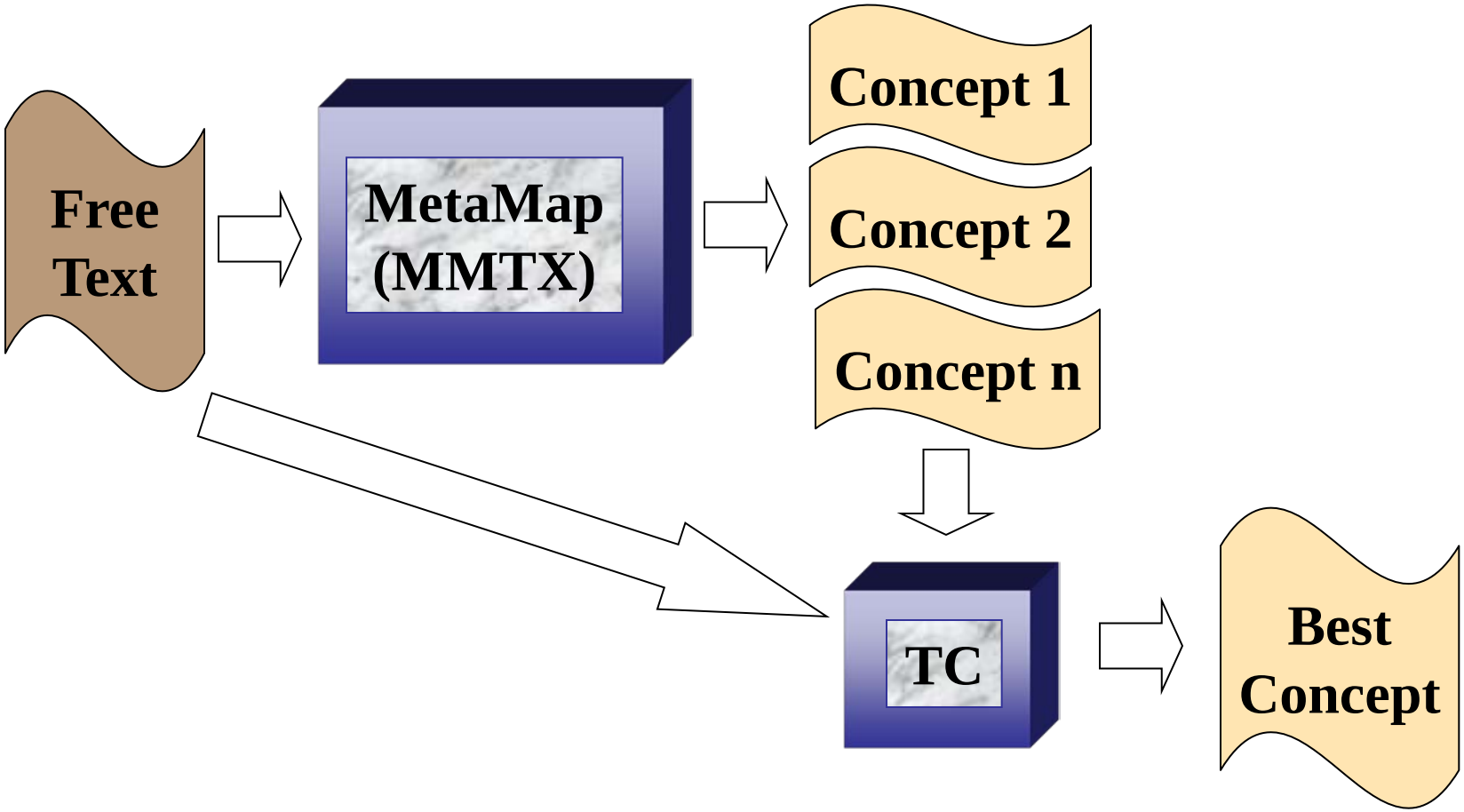
Text Categorization

- Words Senses disambiguation (WSD)



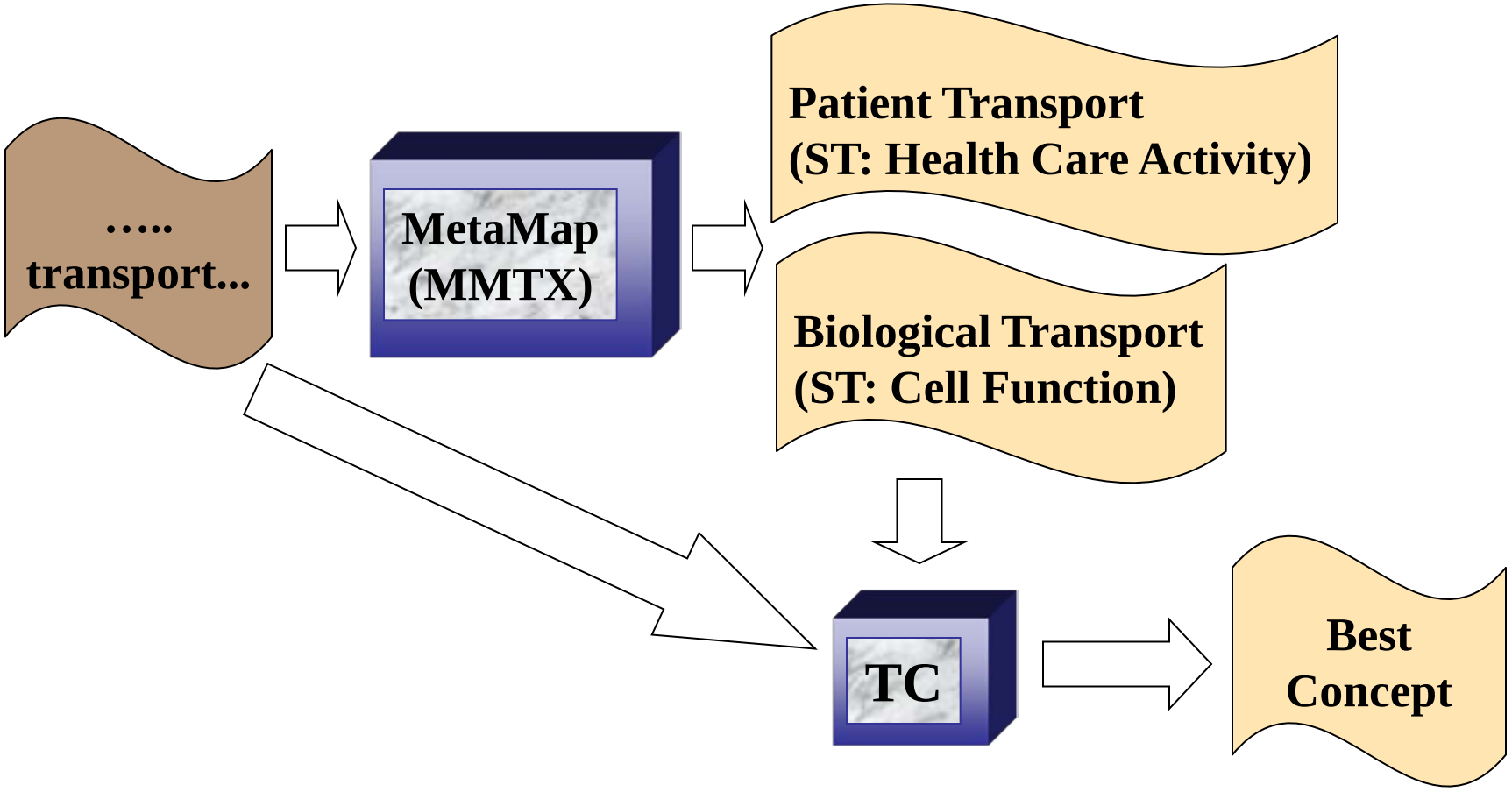
Text Categorization

- Words Senses disambiguation (WSD)



Text Categorization

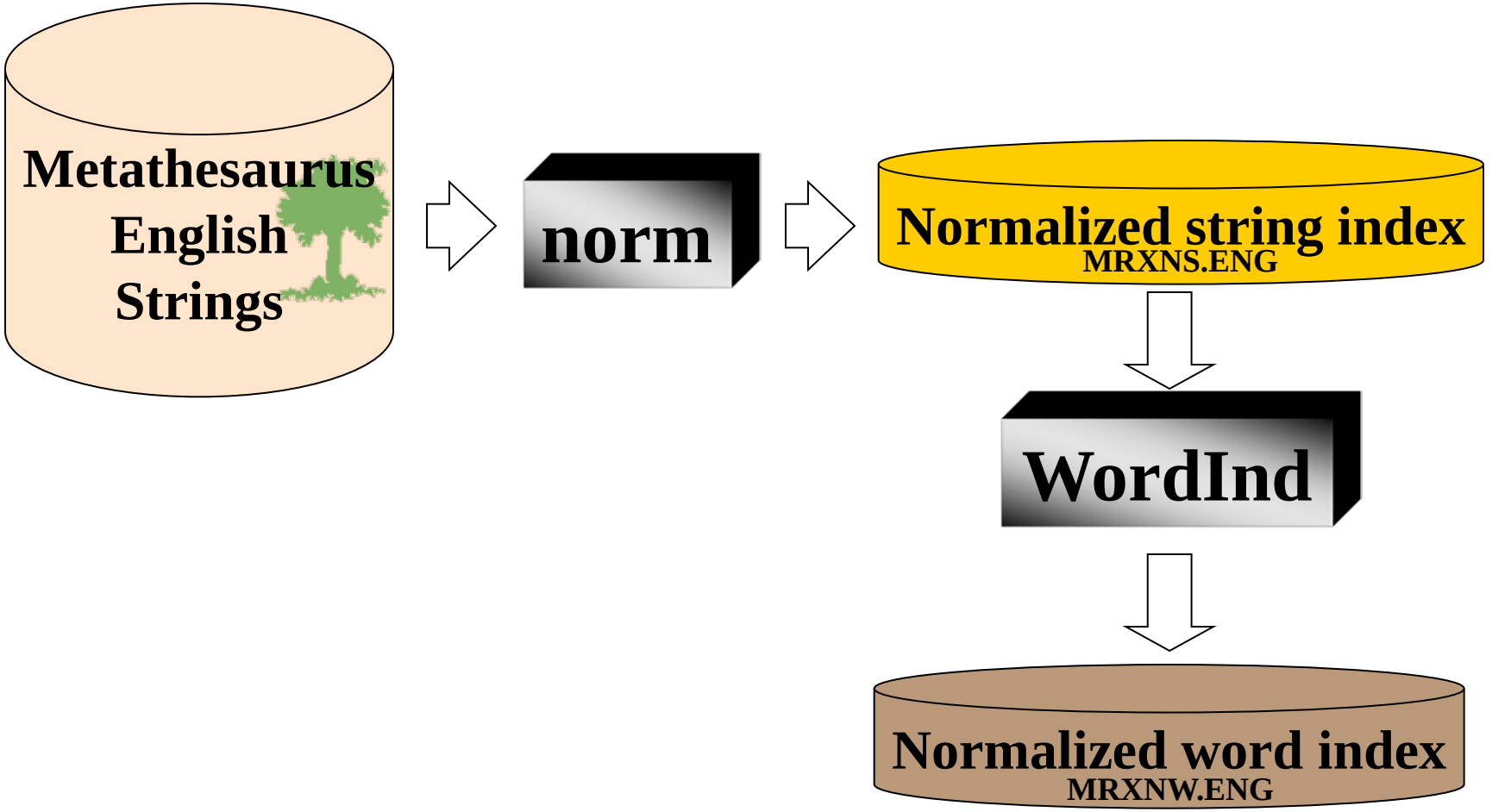
- Words Senses disambiguation (WSD)



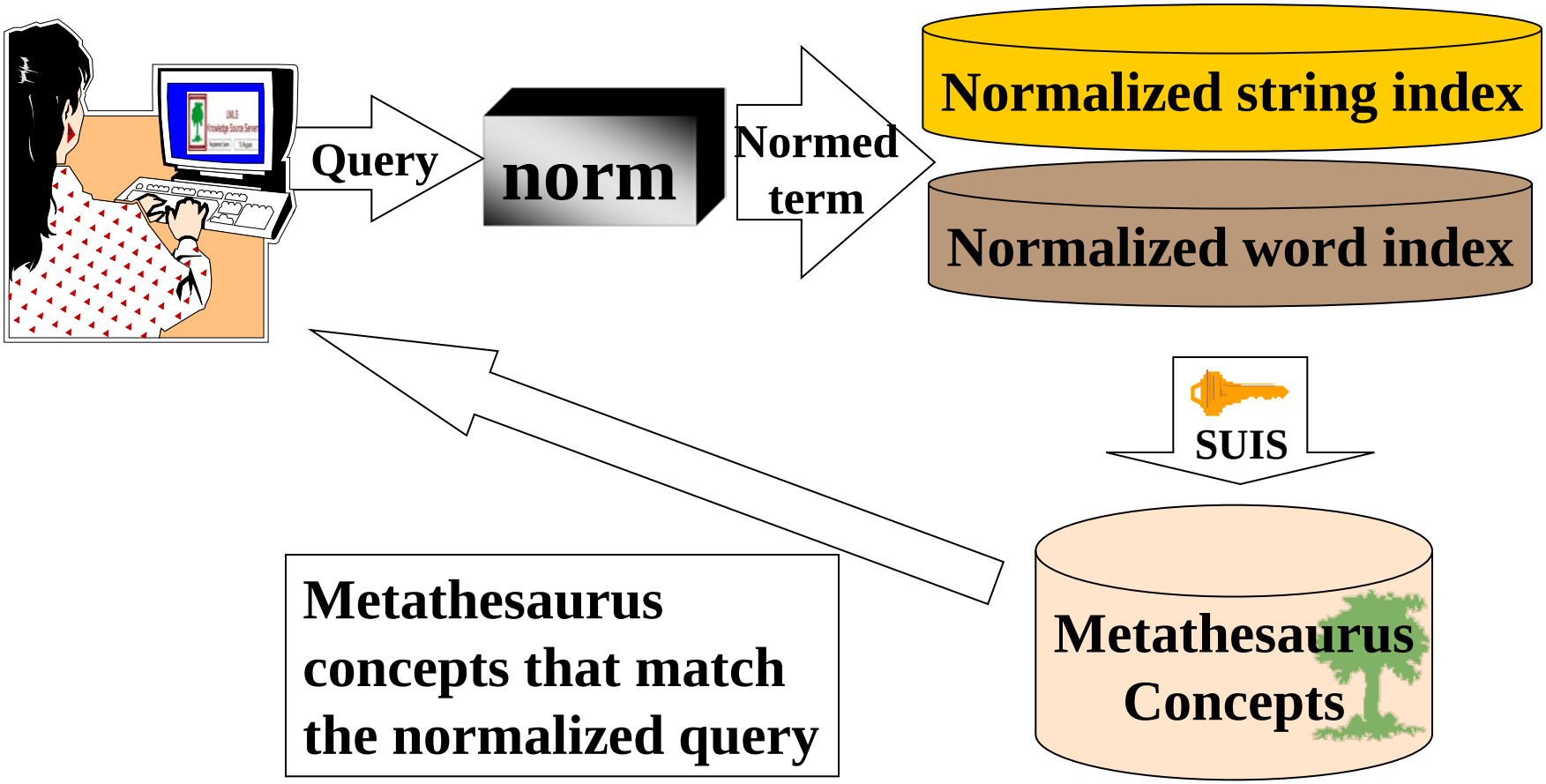
Questions

- Lexical Systems Group: http://umlslex.nlm.nih.gov
- Lexical Tools: http://umlslex.nlm.nih.gov/lvg

Application



Application



Example

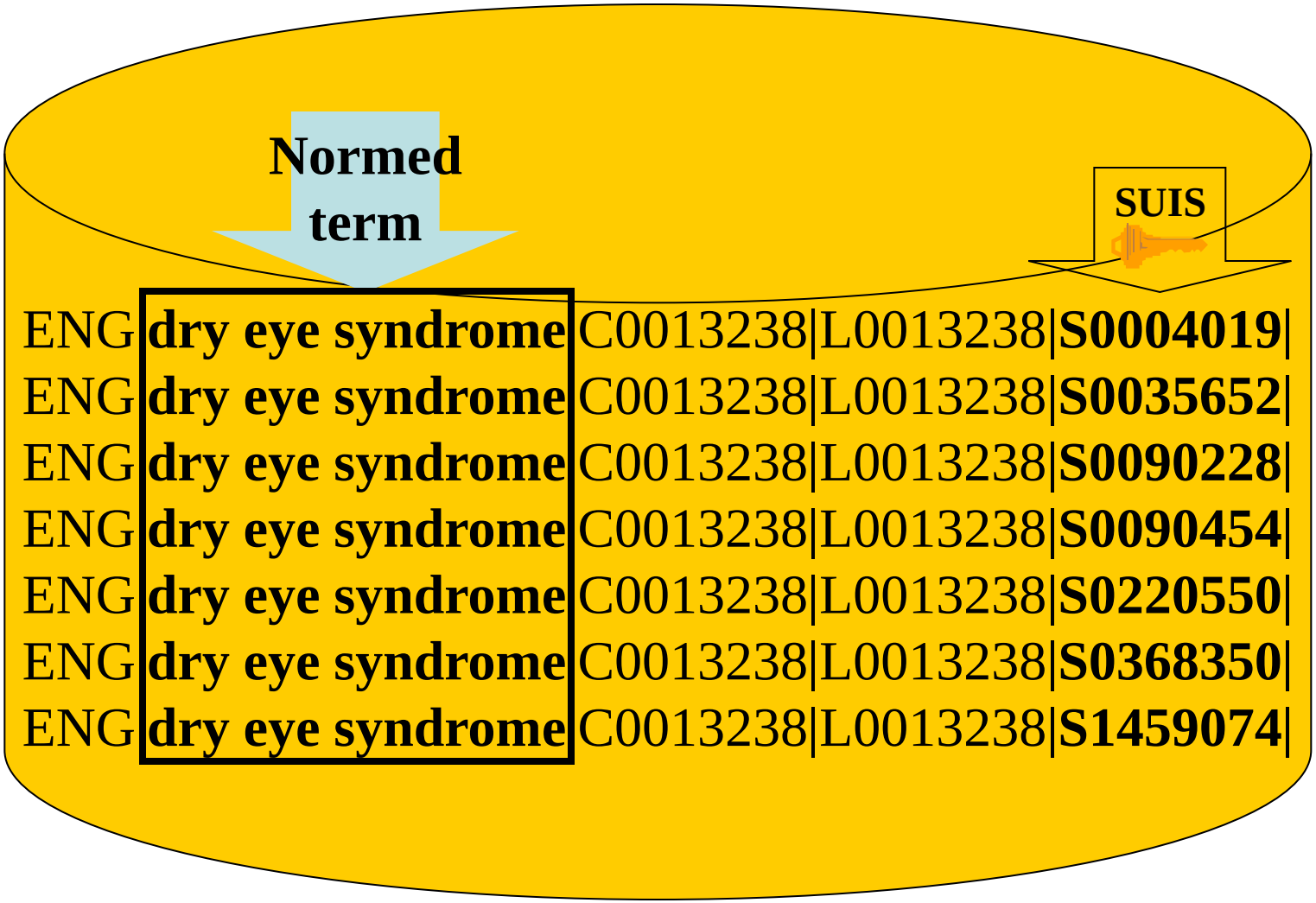
Query

norm

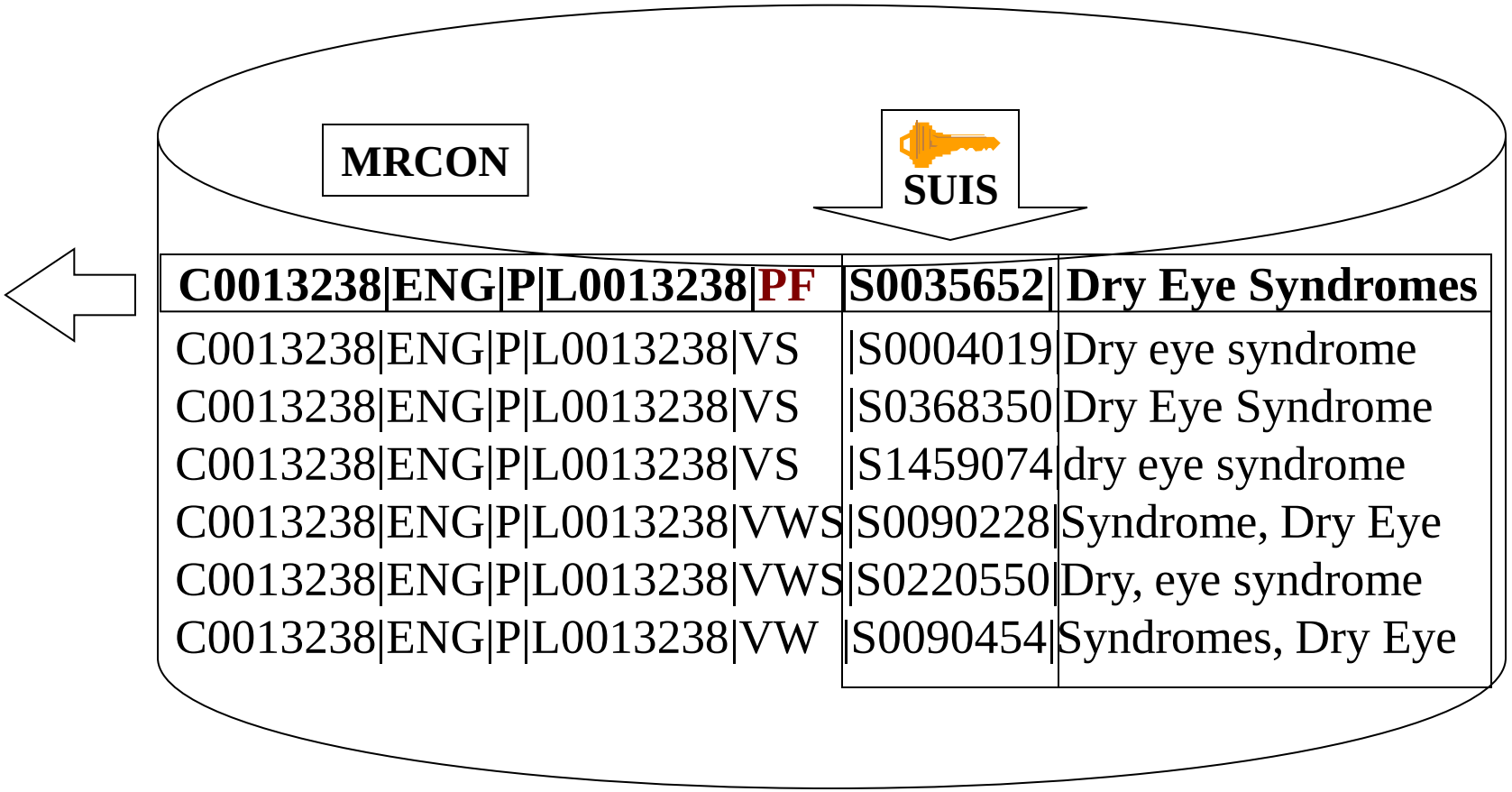
Normed term

dry eye syndrome

Example (Cont.)



Example (Cont.)



C0013238|ENG|P|L0013238|PF	S0035652|	Dry Eye Syndromes
C0013238|ENG|P|L0013238|VS	|S0004019	Dry eye syndrome
C0013238|ENG|P|L0013238|VS	|S0368350	Dry Eye Syndrome
C0013238|ENG|P|L0013238|VS	|S1459074	dry eye syndrome
C0013238|ENG|P|L0013238|VWS	|S0090228	Syndrome, Dry Eye
C0013238|ENG|P|L0013238|VWS	|S0220550	Dry, eye syndrome
C0013238|ENG|P|L0013238|VW	|S0090454	Syndromes, Dry Eye

Questions

- Lexical Systems Group: http://umlslex.nlm.nih.gov
- Lexical Tools: http://umlslex.nlm.nih.gov/lvg