

Lexical Tools Briefing

[The Lexical Systems Group](#)

[NLM](#). [LHNCBC](#). [CGSB](#)

May, 2006

Table of Contents

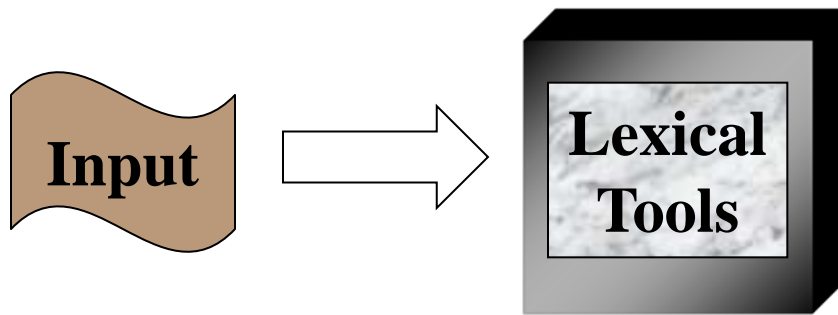
- Introduction
- Lvg
- Norm
- LuiNorm
- Application Example
- Users
- Annual Release Cycle
- Tests
- Questions

Introduction – Lexical Tools



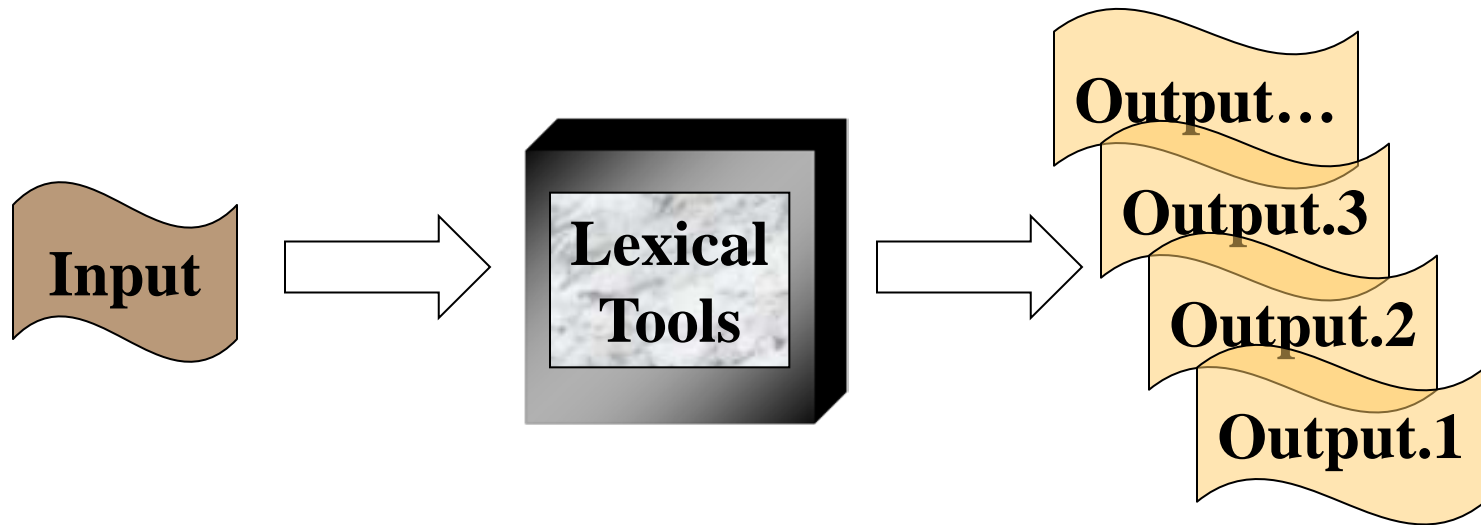
- A suite of text utilities

Introduction – Lexical Tools



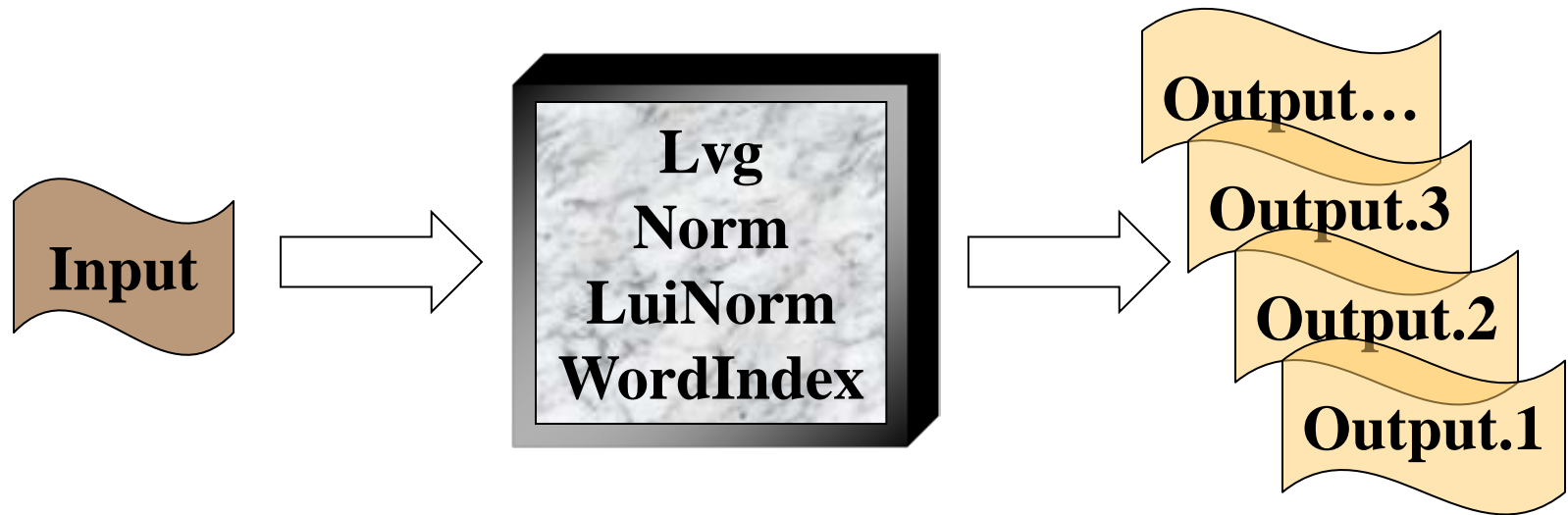
- A suite of text utilities take the given input

Introduction – Lexical Tools



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

Four Tools



Tool Types

- Command line tools
 - [lvq](#) (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

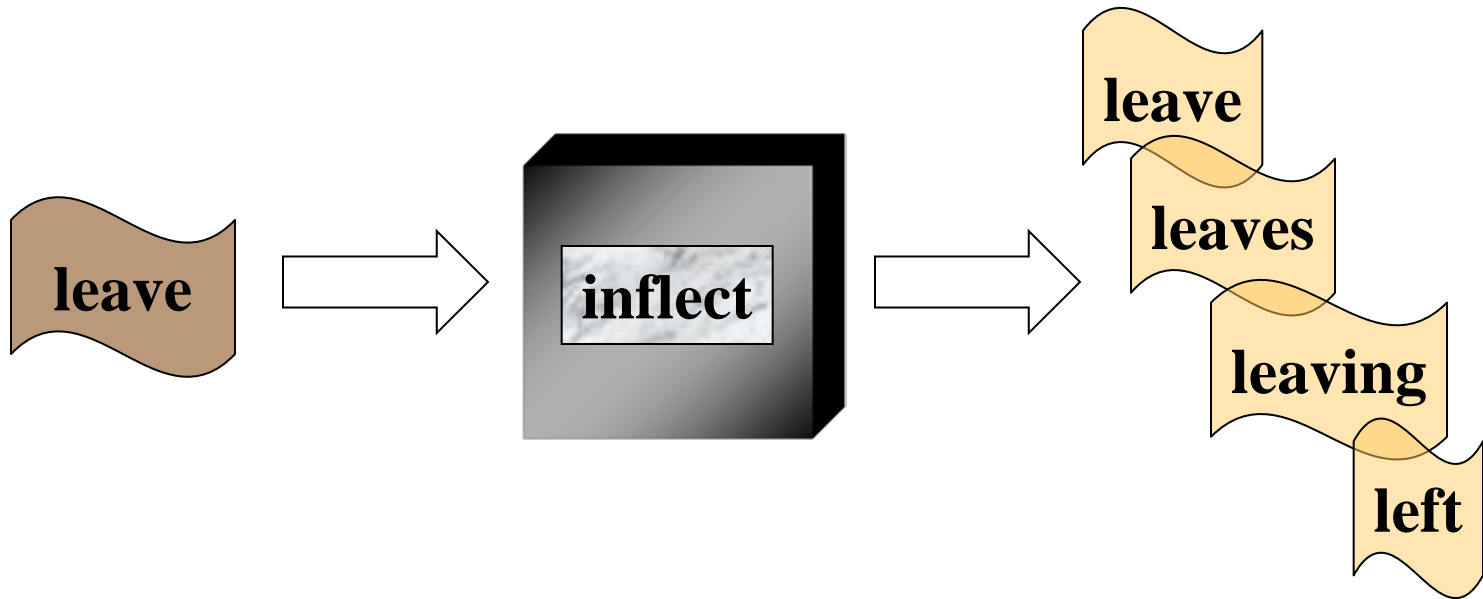
Lexical Variants Generation



LVG

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

Flow Components

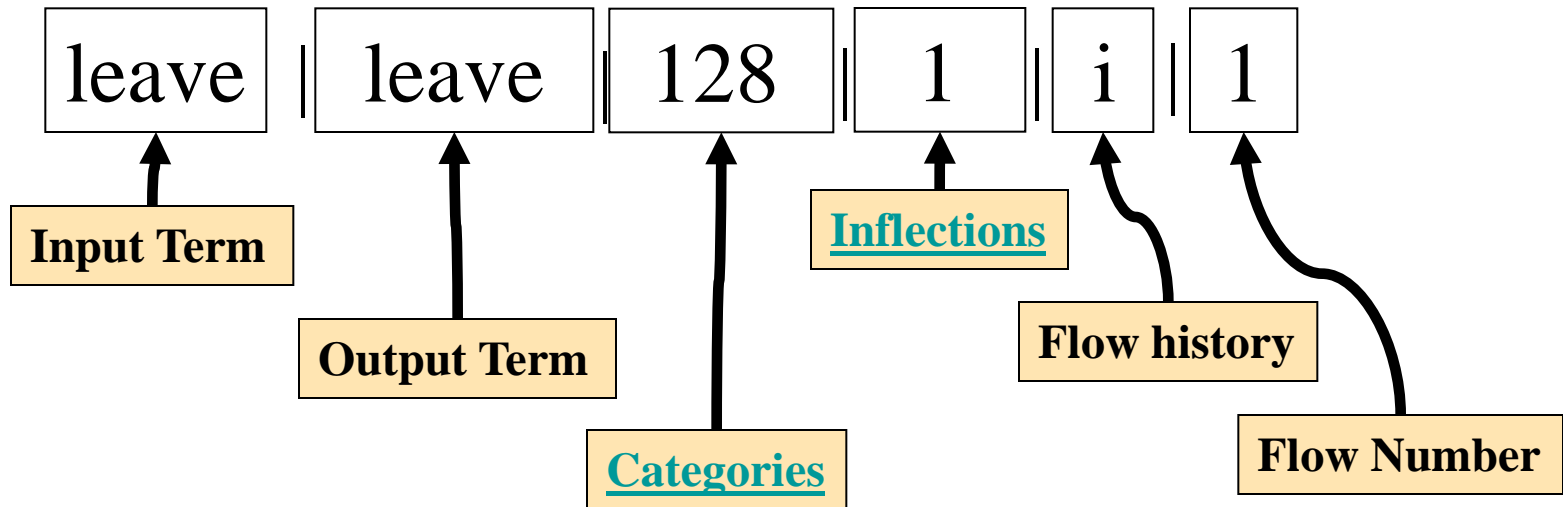


Command Line Tool

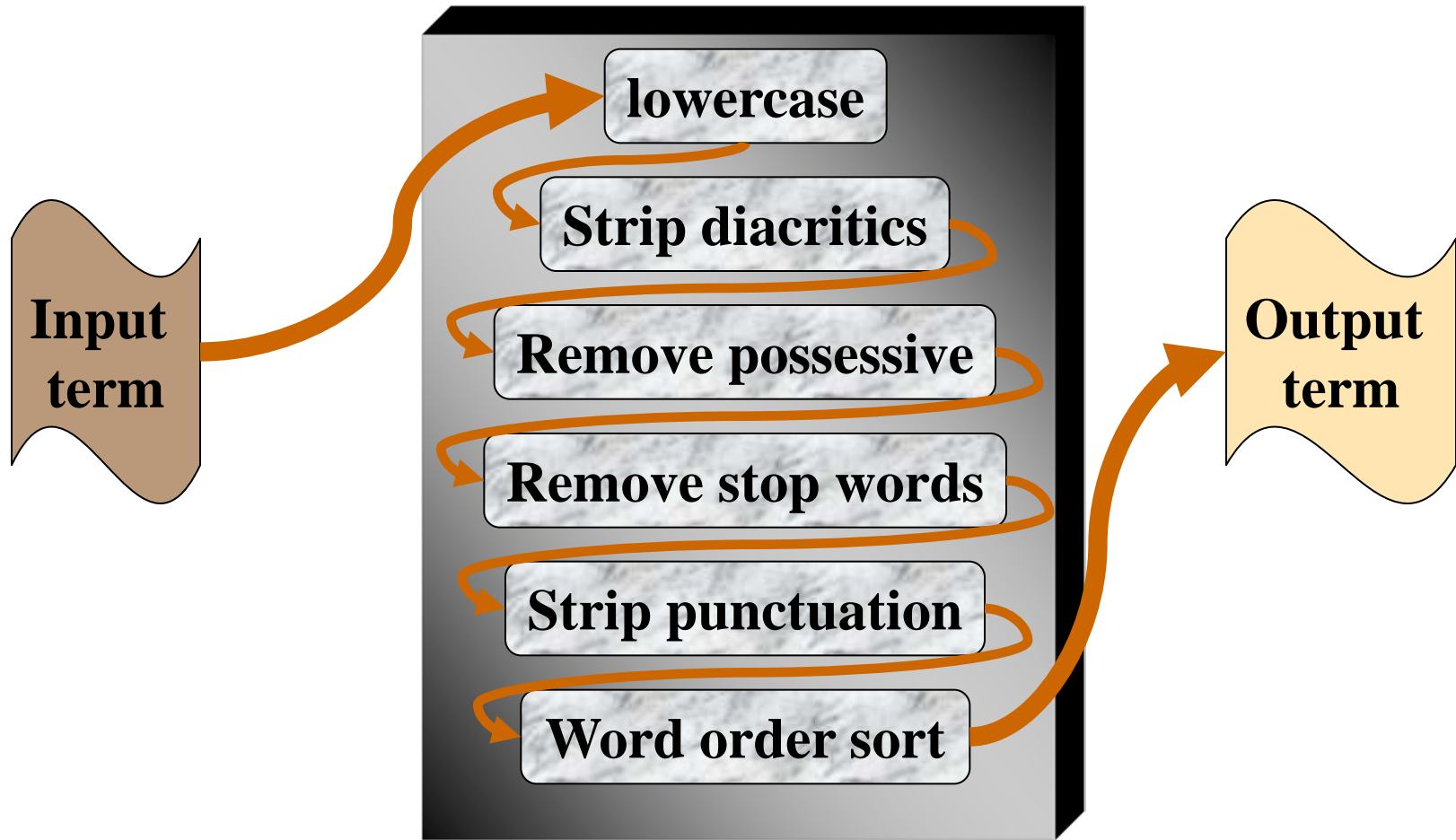
```
> lvg -f:i  
leave  
leave | leave | | | i | 1 |  
leave | leave | 128 | 512 | i | 1 |  
leave | leaves | 128 | 8 | i | 1 |  
leave | left | 1024 | 64 | i | 1 |  
leave | left | 1024 | 32 | i | 1 |  
leave | leave | 1024 | 1 | i | 1 |  
leave | leave | 1024 | 262144 | i | 1 |  
leave | leave | 1024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

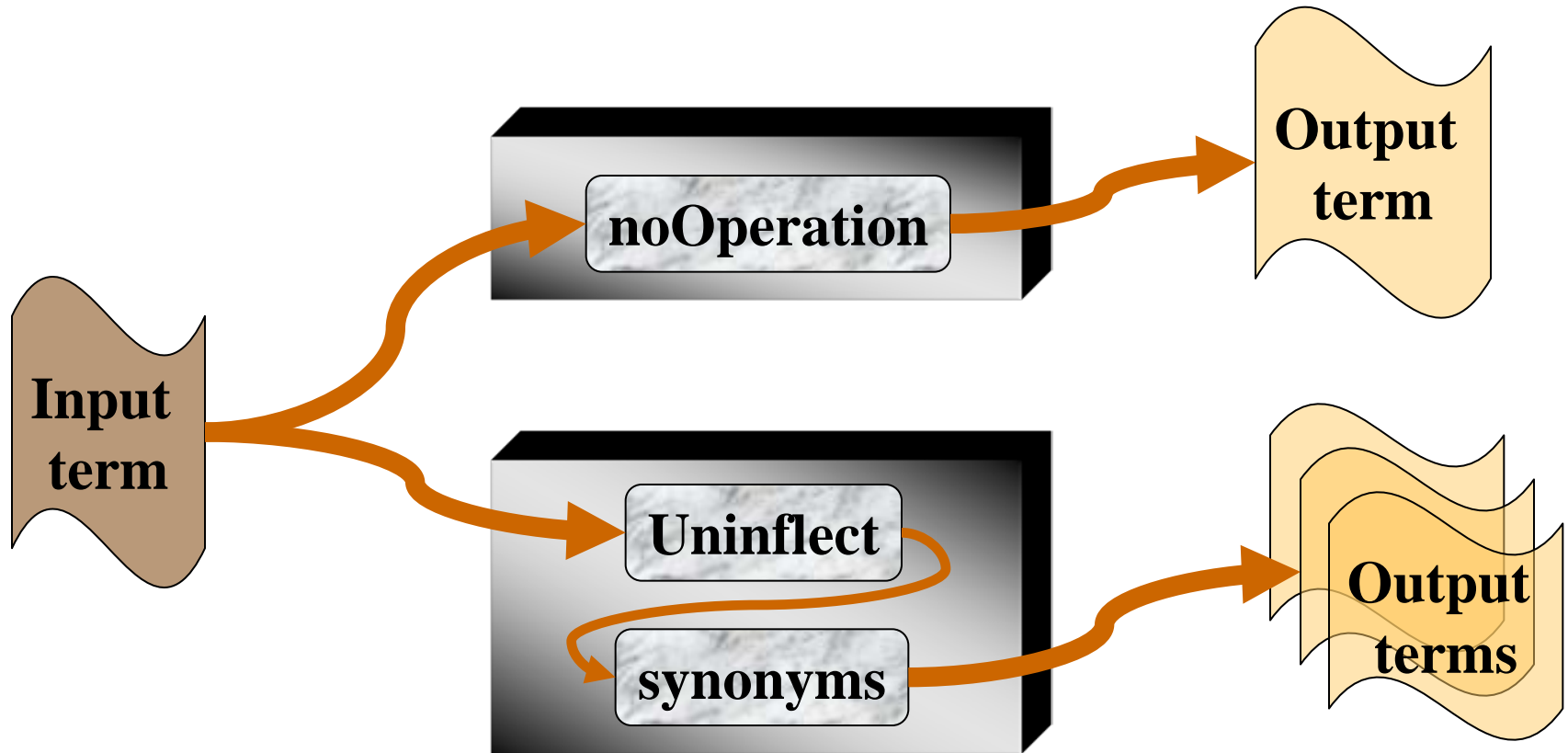
```
The Gougerot-Sjögren's Syndrome
```

```
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 |
```

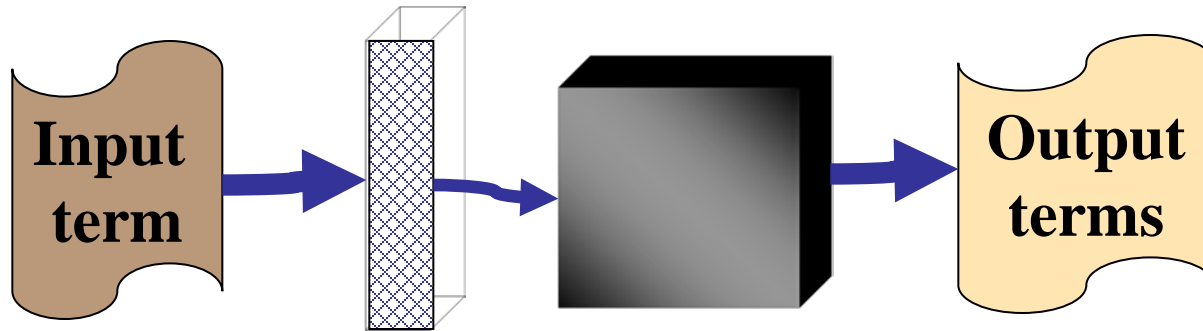
```
ear | aural | 1 | 1 | B+y | 2 |
```

```
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

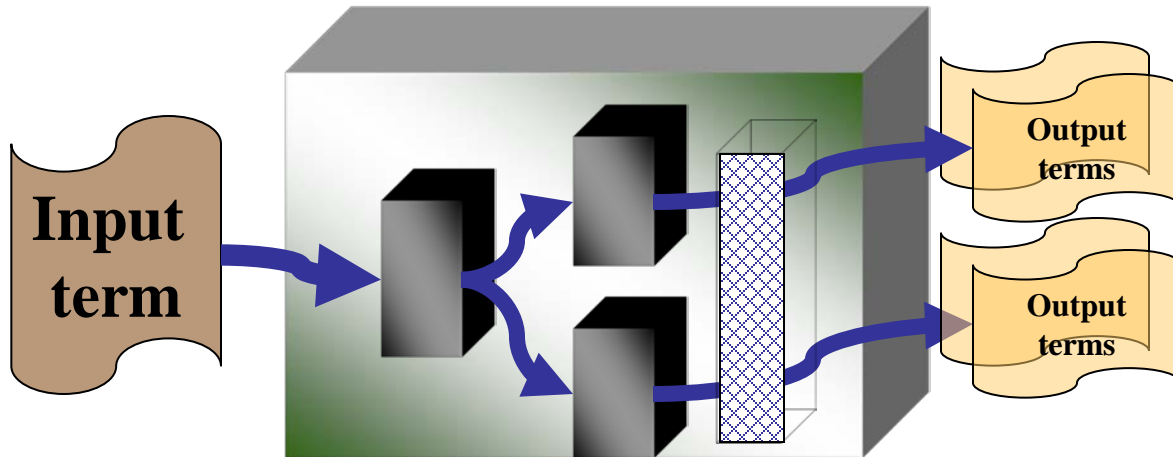
```
> lvg -f:u -t:7 -F:8:6
```

```
C0035440 | ENG | S | L0035434 | VW | S0003894 |
```

```
Rheumatic carditis, acute
```

```
acute Rheumatic carditis | S0003894
```

Global Behavior Options



```
> lvg -f:L -f:E -s:"\"
```

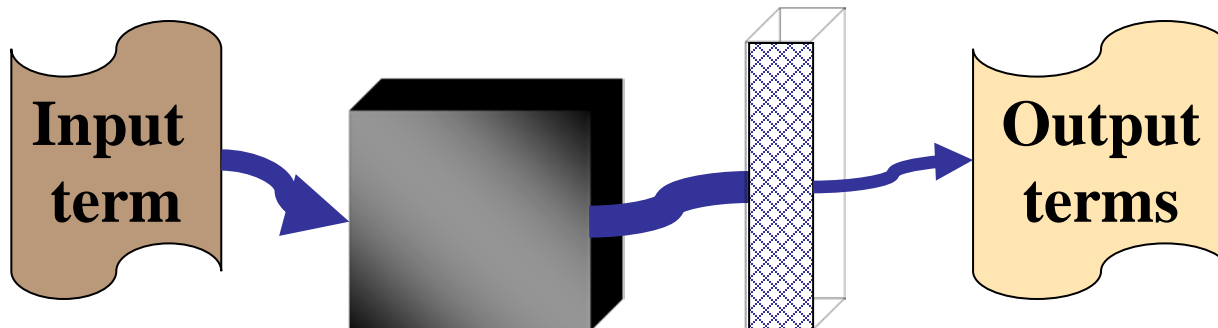
Change separator to “\”

```
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 parenthetical 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 parenthetical 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

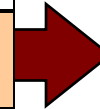
q4: get symbol names synonymy

Hodgkin's Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases NOS



Norm

g: remove genitives

rs: remove parenthetical 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

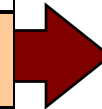
Hodgkin's Diseases, NOS

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

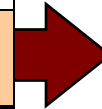
Hodgkin Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases NOS

Hodgkin Diseases

Hodgkin Diseases



Norm

g: remove genitives

rs: remove parenthetical 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

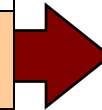
Hodgkin Diseases, NOS

Hodgkin Diseases NOS

Hodgkin Diseases

Hodgkin Diseases

Hodgkin Diseases



Norm

g: remove genitives

rs: remove parenthetical 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

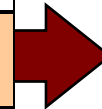
Hodgkin Diseases NOS

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 parenthetical 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

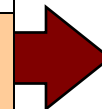
Hodgkin Diseases

Hodgkin Diseases

hodgkin diseases

hodgkin disease

hodgkin disease



Norm

g: remove genitives

rs: remove parenthetical 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

disease hodgkin

Norm

g: remove genitives

rs: remove parenthetical 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

disease hodgkin

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



disease hodgkin

LuiNorm

- A special version of Norm
- Used in the UMLS Metathesaurus
- Composed of 11 lvg flow components
- Replace $-f:Ct$ (in norm) to $-f:C$
- Provide one to one correspondence between an input and an output

LuiNorm

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

C: retrieve canonical form

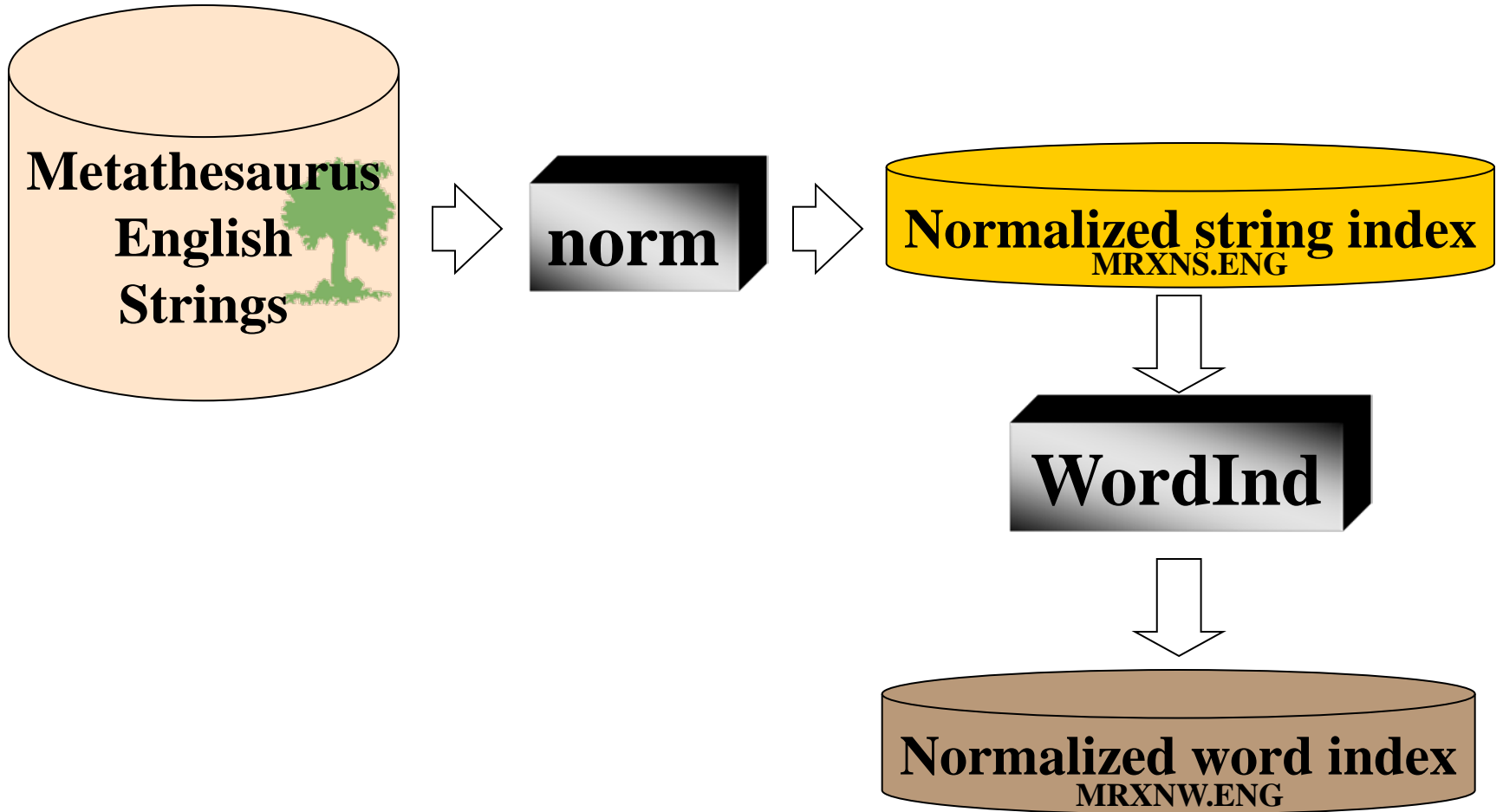
w: sort words by order

q4: get symbol names synonymy

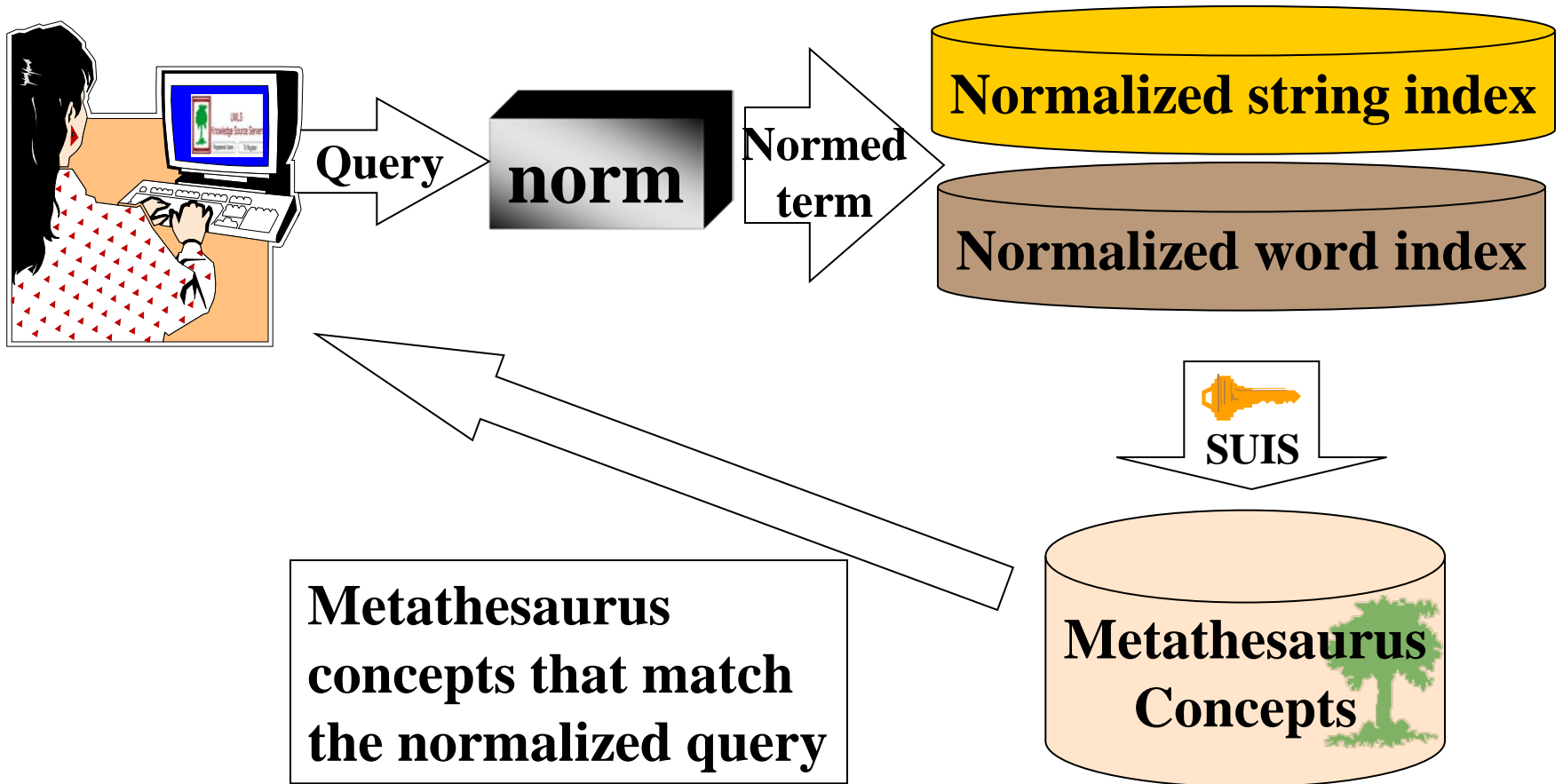
Canonical Form

- To manage ambiguity generated by uninflection
 - “left” is uninflected to “left” (adj) or “leave” (verb)
- A Canonical class includes terms have same inflections or spelling variants
 - “left”, “leave”, and “leaf” have same inflections “leaves”
 - “analog” and “analogue” are spelling variants
- Canonical form is an arbitrarily chosen member of a Canonical class
 - alphabetical order
 - shortest member
 - in The SPECIALIST LEXICON

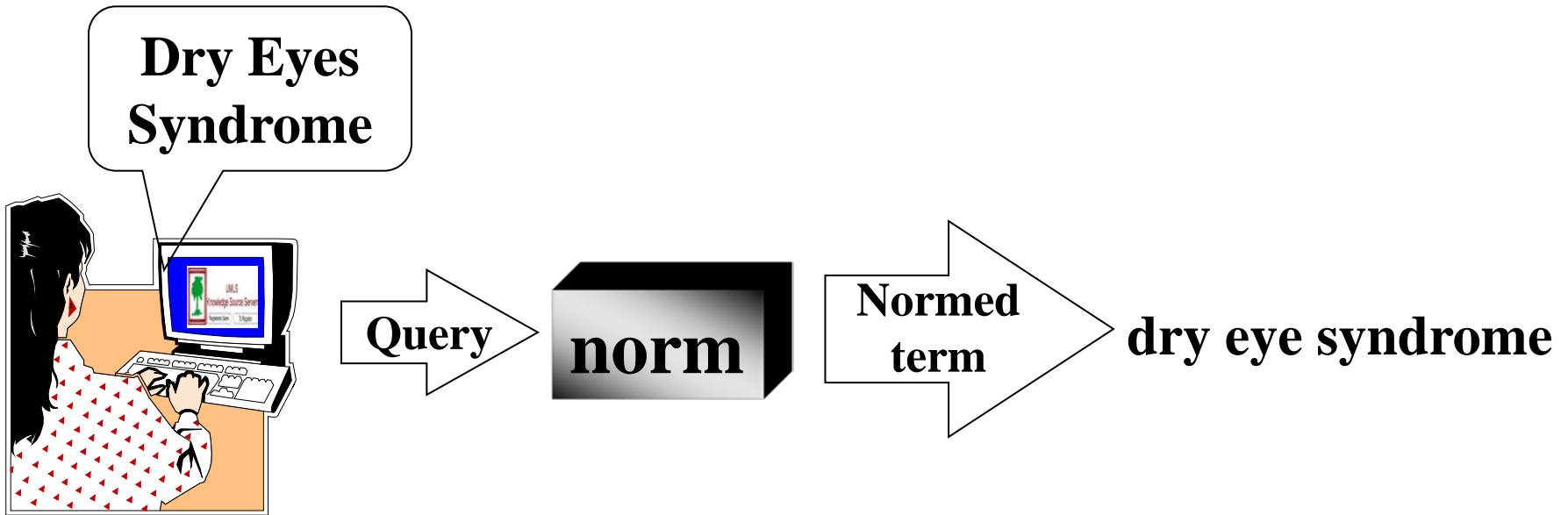
Application



Application



Example



Example (Cont.)

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

Example (Cont.)

MRCON



SUIS

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

Users

- Internal NLM Users

- Lexical Systems Group
- UMLS Group (Apelon)
- MMTX (MetaMap): map text phrases to Metathesaurus concept
- UMLS Knowledge Source Server
- Clinical Trial
- Indexing Initiative
- Semantic Knowledge Representation
- Terminology Server
- Medical Ontology
- Word Sense Disambiguation
- ...

Users (Cont.)

- Public Users (USA, edu)
 - [University of North Carolina, USA](#)
 - [University of Washington, USA](#)
 - [Mayo Clinic, USA](#)
 - [Iowa State University, USA](#)
 - [University of Texas, Medical Center, USA](#)
 - [The University of Arizona, USA](#)
 - [Columbia University, USA](#)
 - [Harvard University, USA](#)
 - [Johns Hopkins Medical Institutions, USA](#)
 - [Johns Hopkins University, USA](#)
 - [Medical informatics UC Davis, USA](#)
 - [Medical College of Wisconsin, USA](#)
 - [Stanford University, USA](#)
 - ...

Users (Cont.)

- Public Users (USA, non-edu)
 - [Schering-Plough, USA](#)
 - [Mayo Clinic, USA](#)
 - [Translational Genomics Research Institute, USA](#)
 - [Emergint, USA](#)
 - [MedTopia, USA](#)
 - [Mitre, USA](#)
 - [NICHD, USA](#)
 - [American College of Physicians, USA](#)
 - ...

Users (Cont.)

- Public Users (international)
 - [Vienna University of Technology, Austria](#)
 - [GlaxoSmithKline Research and Development, worldwide](#)
 - National Institute of Hospital Administration, China
 - [University of Manchester, UK](#)
 - [National Health Service, UK](#)
 - [The University of Western Ontario, Canada](#)
 - [Taipei Medical University, Taiwan](#)
 - [Université Paris, France](#)
 - [Bioinformatics Group, Japan](#)
 - [Seoul National University Hospital, Korea](#)
 - [Myong Ji University, Korea](#)
 - [Hôpital Charles Nicolle, France](#)
 - [Universitaetsklinikum Freiburg, Germany](#)
 - ...

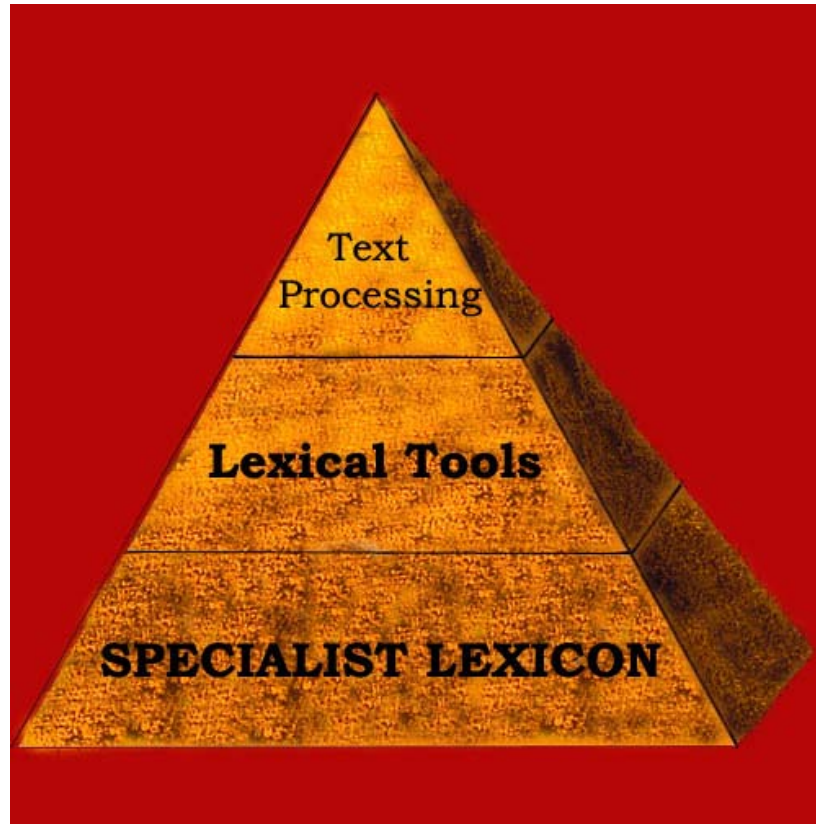
Annual Release Cycle

- Release with UMLS Resources (Jan.)
- Provide technical support and open SCRs
- Create a new release baseline
- Complete SCRs (Jun.)
- Tests (begin)
- Integrate with new LEXICON (Jul.)
- Update all software components: Gui tool & examples
- Internal release (Oct.)
- Update all documents: apiDocs, userDocs, designDocs
- Update web sites and web tools
- Tests (end)
- Build, pack, release, and deploy (Dec.)

Tests

- Unit Test (black box test):
 - new software components
 - flows components
 - options
- Integration Test
 - Gui tool & Web tools
 - other applications
- Distribution test
 - platforms: Linux, Unix, Window NT
- Performance Test
 - norm
 - luiNorm

Questions



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