Lexical Tools Briefing

Dr. Chris J. Lu

The Lexical Systems Group

NLM. LHNCBC. CGSB

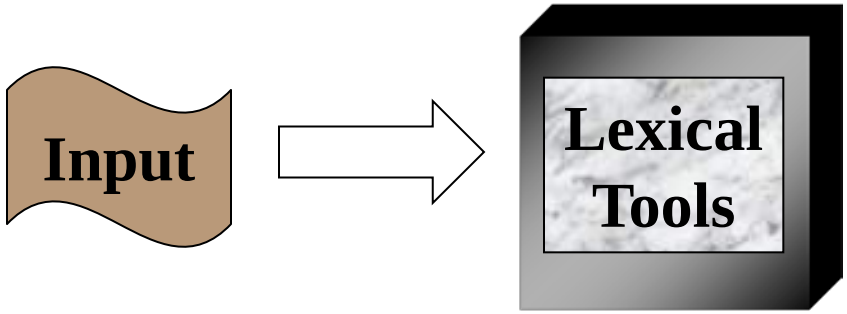
Feb., 2013

Table of Contents

Lexical Tools

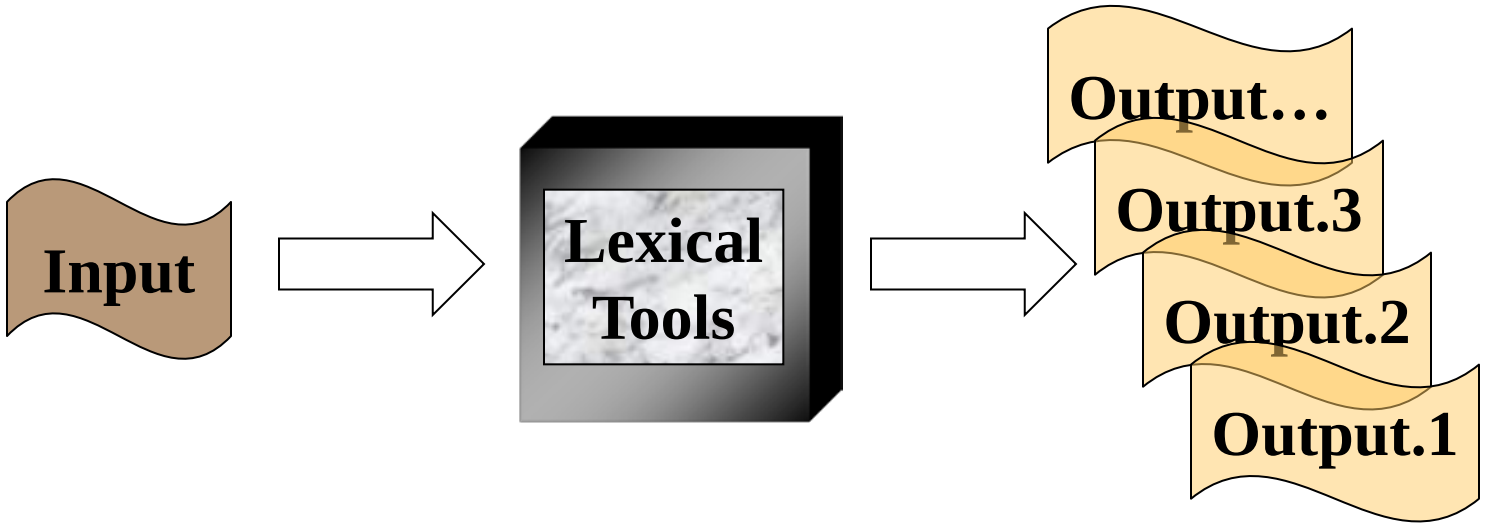
- A suite of text utilities

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

Six Tools


Input
lvg
norm
fields
toAscii
luiNorm
wordIndex
Output…
Output.3
Output.2
Output.1

Tool Types

- Command line tools
  - lvg (Lexical Variants Generation)
  - norm
  - luiNorm
  - wordInd
  - toAscii
  - fields
- 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
  - increasing recall and/or precision
  - making word, term, phrase indexes
  - matching queries with indexed entries
  - …

Facts

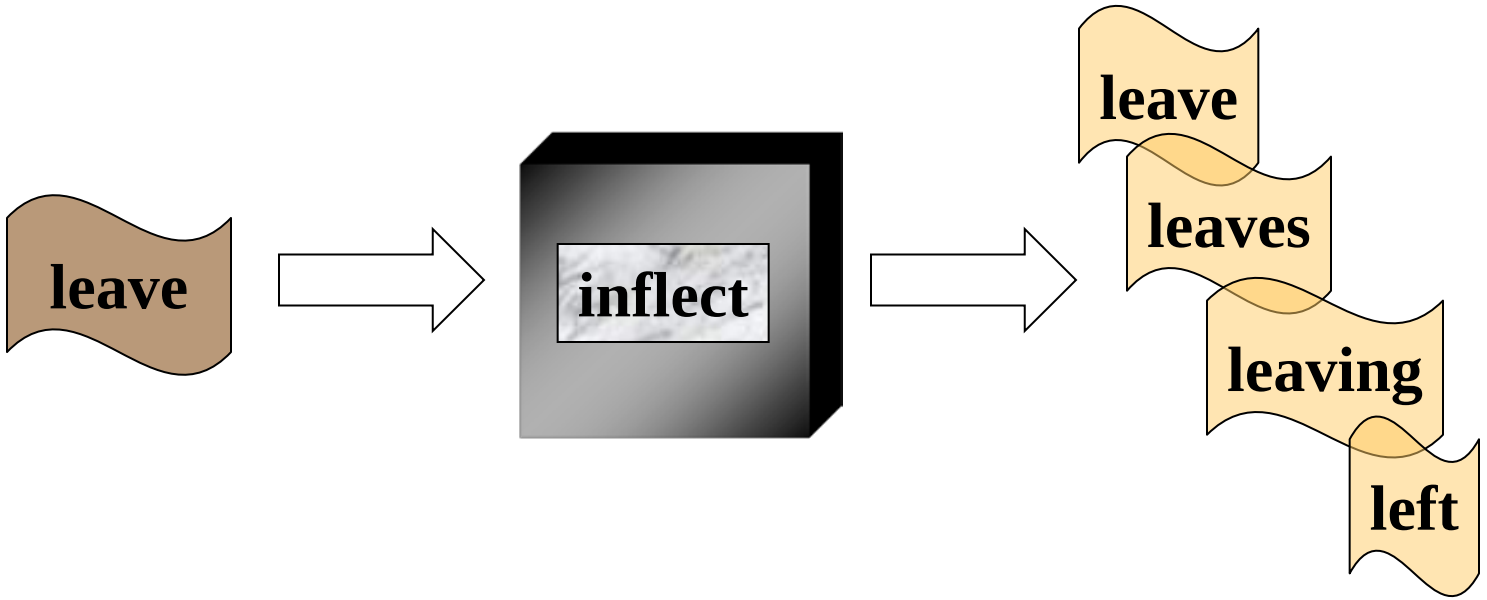
- Release annually
- 100% Java (since 2002)
- Free distributed with open source code
- Run on different platforms
- One complete package
- Documents & support

Lexical Variants Generation

LVG

- 62 flow components
- 34 options
  - input filter options (4)
  - global behavior options (14)
  - flow specific options (3)
  - output filter options (13)

Flow Components

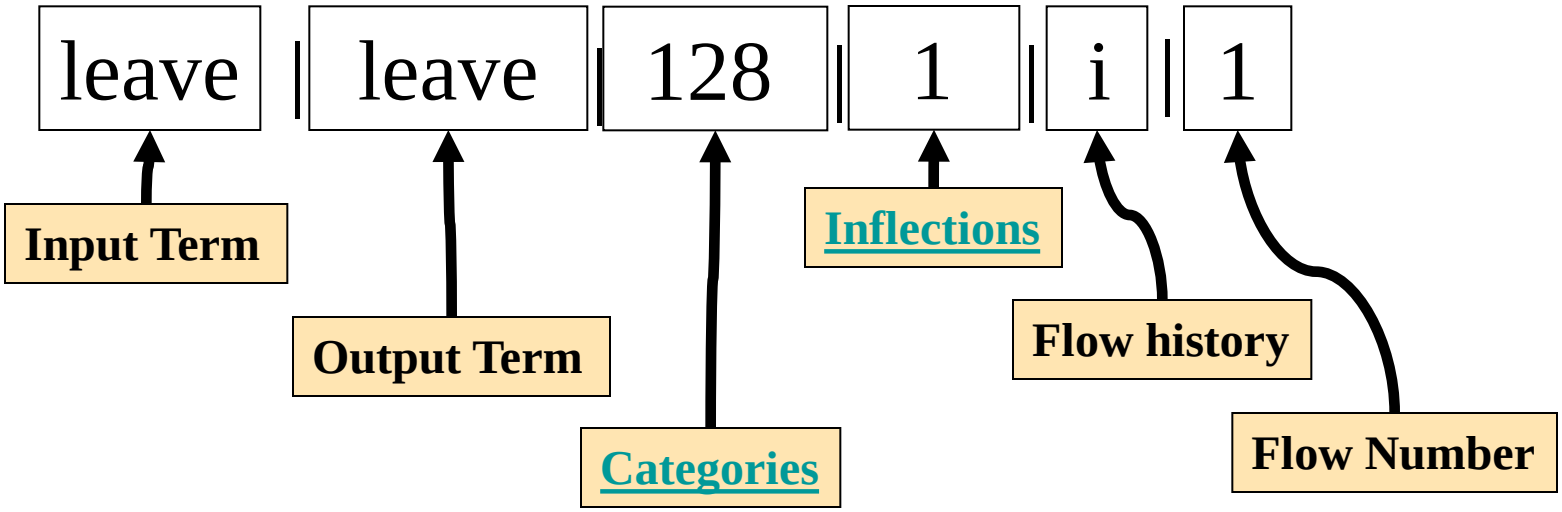


Command Line Tool

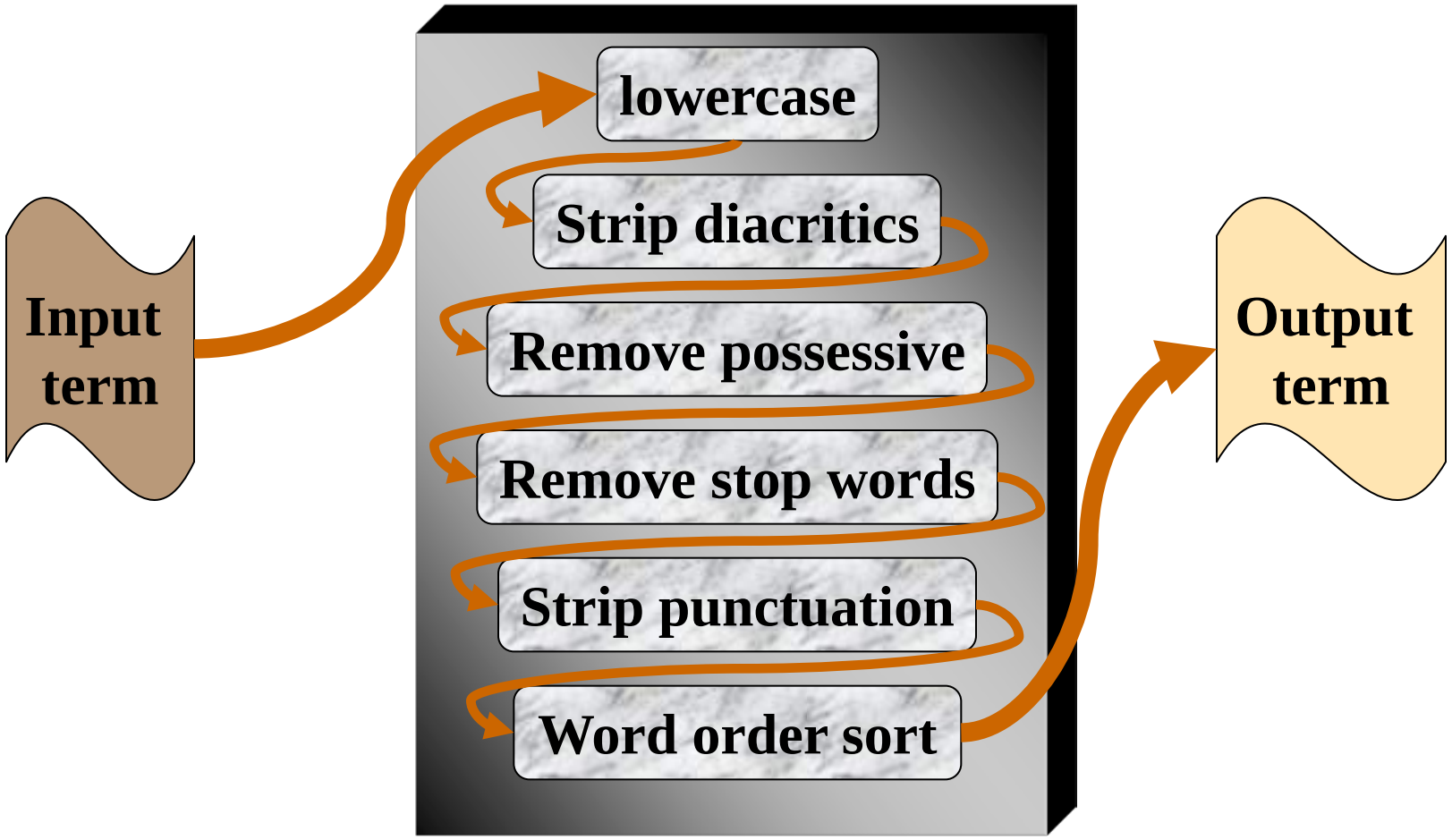
```
> lvg -f:i
leave
leave|leave|128|1|i|1|
leave|leave|128|512|i|1|
leave|leaves|128|8|i|1|
leave|left|1024|64|i|1|
leave|left|1024|32|i|1|
leave|leave|1024|1|i|1|
leave|leave|1024|262144|i|1|
leave|leave|1024|1024|i|1|
leave|leaves|1024|128|i|1|
leave|leaving|1024|16|i|1|
```

Fielded Output

> lvg –f:i
leave



A Serial Flow

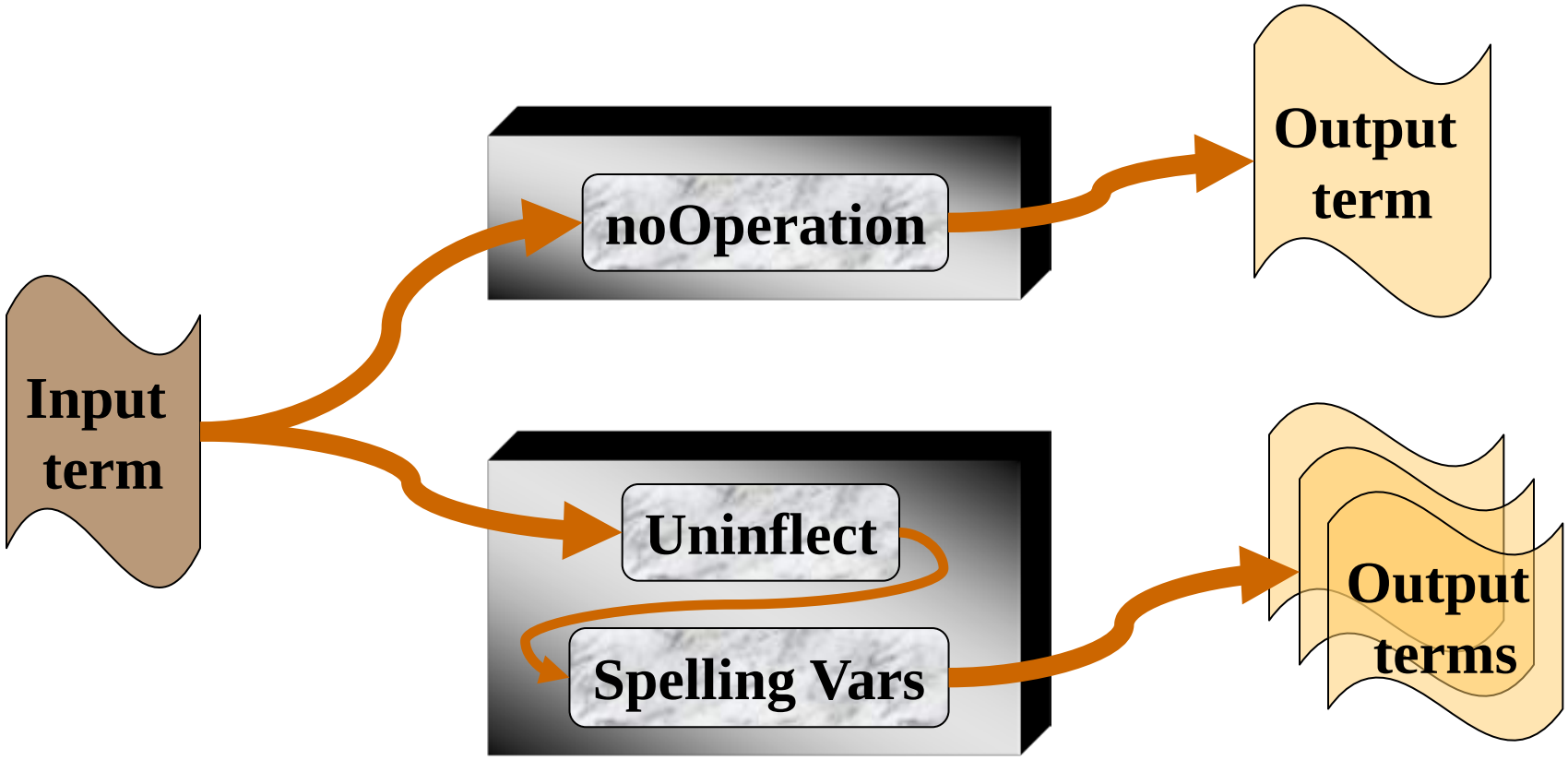


- Flow components can be arranged so that the output of one is the input to another.

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



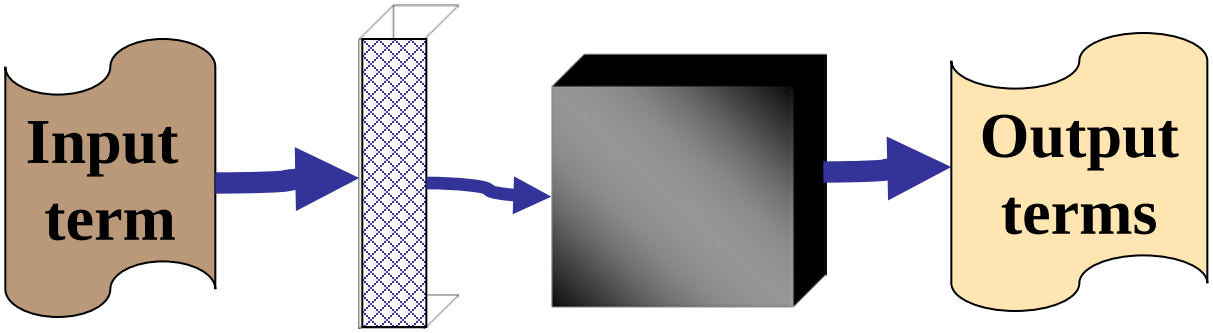
- Multiple flows can be defined

Parallel Flows - Example

```
> lvg -f:n -f:B:s
color
color|color|2047|16777215|n|1|

color|color|128|1|B+s|2|
color|color|1024|1|B+s|2|
color|colour|128|1|B+s|2|
color|colour|1024|1|B+s|2|
```

Input Filter Options



Take field 7 from the input

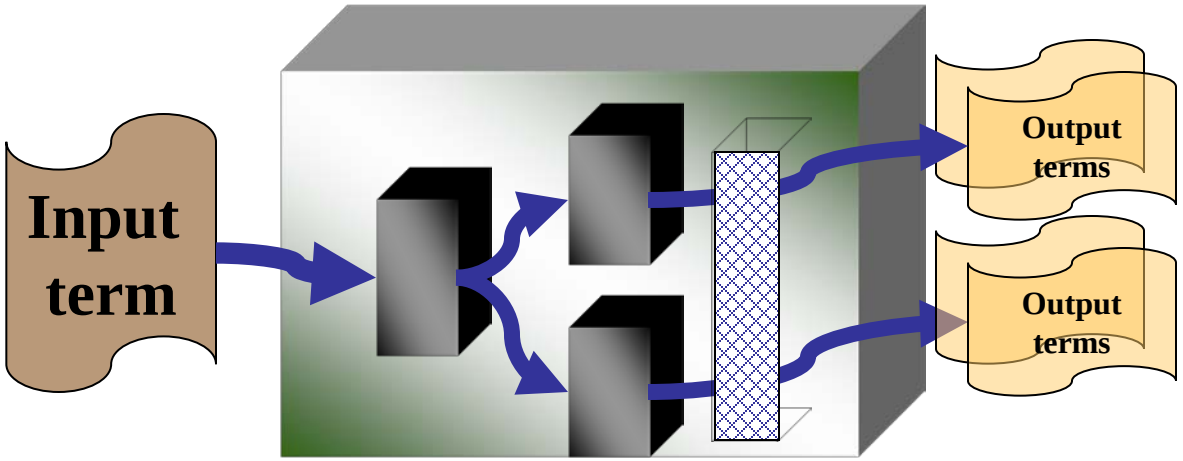
```
> lvg -f:u -t:7 -F:6

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

   Rheumatic carditis, acute

acute Rheumatic carditis|S0003894
```

Global Behavior Options




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

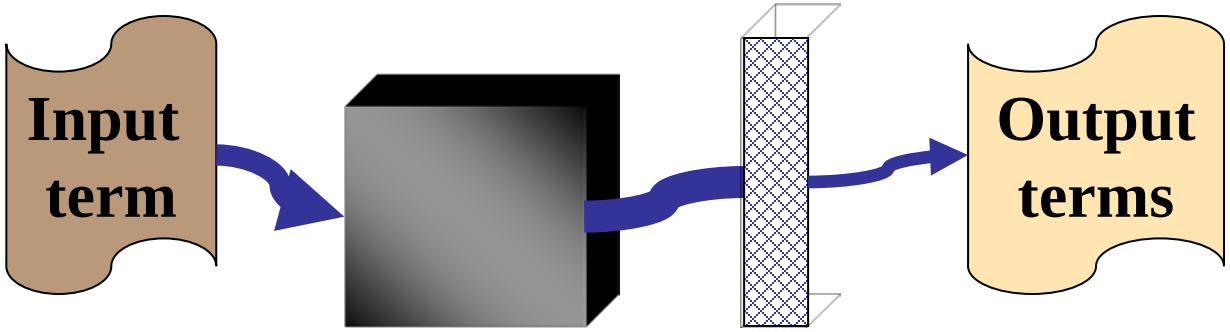
Change separator to "\"

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 (non-ASCII Unicode)
  - word order

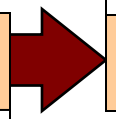
Norm

q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order

“Fœtoproteins α’s, NOS“

Norm

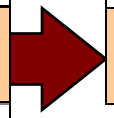
q0: map symbols to ASCII	"Fœtoproteins α's, NOS"
g: remove genitives	
rs: remove parenthetic plural forms	
o: replace punctuation with spaces	
t: strip stop words	
l: lowercase	
B: uninflect each words in a term	
Ct: retrieve citations	
q7: Unicode core Norm	
q8: strip or map Unicode to ASCII	
w: sort words by order	



"Fœtoproteins α's, NOS"

Norm

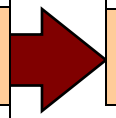
q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order



“Fœtoproteins α’s, NOS“
"Fœtoproteins α**’s**, NOS"
"Fœtoproteins α, NOS"

Norm

q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order



“Fœtoproteins α’s, NOS“
"Fœtoproteins α’s, NOS"
"Fœtoproteins α, NOS"
"Fœtoproteins α, NOS"

Norm

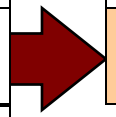
q0: map symbols to ASCII	“Fœtoproteins α’s, NOS“
g: remove genitives	"Fœtoproteins α’s, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	"Fœtoproteins α, NOS"
t: strip stop words	Fœtoproteins α NOS
l: lowercase	
B: uninflect each words in a term	
Ct: retrieve citations	
q7: Unicode core Norm	
q8: strip or map Unicode to ASCII	
w: sort words by order	

Norm

q0: map symbols to ASCII	“Fœtoproteins α’s, NOS“
g: remove genitives	"Fœtoproteins α’s, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	"Fœtoproteins α, NOS"
t: strip stop words	Fœtoproteins α NOS
l: lowercase	Fœtoproteins α
B: uninflect each words in a term	
Ct: retrieve citations	
q7: Unicode core Norm	
q8: strip or map Unicode to ASCII	
w: sort words by order	

Norm

q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order



“Fœtoproteins α’s, NOS“
"Fœtoproteins α’s, NOS"
"Fœtoproteins α, NOS"
"Fœtoproteins α, NOS"
Fœtoproteins α NOS
Fœtoproteins α
fœtoproteins α

Norm

q0: map symbols to ASCII	“Fœtoproteins α’s, NOS“
g: remove genitives	"Fœtoproteins α’s, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	"Fœtoproteins α, NOS"
t: strip stop words	Fœtoproteins α NOS
l: lowercase	Fœtoproteins α
B: uninflect each words in a term	fœtoproteins α
Ct: retrieve citations	fœtoprotein α
q7: Unicode core Norm	
q8: strip or map Unicode to ASCII	
w: sort words by order	

Norm

q0: map symbols to ASCII	“Fœtoproteins α’s, NOS“
g: remove genitives	"Fœtoproteins α’s, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	"Fœtoproteins α, NOS"
t: strip stop words	Fœtoproteins α NOS
l: lowercase	Fœtoproteins α
B: uninflect each words in a term	fœtoproteins α
Ct: retrieve citations	fœtoprotein α
q7: Unicode core Norm	fetoprotein α
q8: strip or map Unicode to ASCII	
w: sort words by order	

Norm

q0: map symbols to ASCII	“Fœtoproteins α’s, NOS“
g: remove genitives	"Fœtoproteins α’s, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	"Fœtoproteins α, NOS"
t: strip stop words	Fœtoproteins α NOS
l: lowercase	Fœtoproteins α
B: uninflect each words in a term	fœtoproteins α
Ct: retrieve citations	fœtoprotein α
q7: Unicode core Norm	fetoprotein α
q8: strip or map Unicode to ASCII	fetoprotein α
w: sort words by order	

Norm

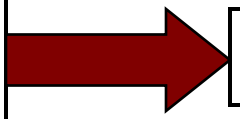
q0: map symbols to ASCII	“Fœtoproteins α’s, NOS“
g: remove genitives	"Fœtoproteins α’s, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	"Fœtoproteins α, NOS"
t: strip stop words	Fœtoproteins α NOS
l: lowercase	Fœtoproteins α
B: uninflect each words in a term	fœtoproteins α
Ct: retrieve citations	fœtoprotein α
q7: Unicode core Norm	fetoprotein α
q8: strip or map Unicode to ASCII	fetoprotein α
w: sort words by order	fetoprotein alpha

Norm

q0: map symbols to ASCII	“Fœtoproteins α’s, NOS“
g: remove genitives	"Fœtoproteins α’s, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	"Fœtoproteins α, NOS"
t: strip stop words	Fœtoproteins α NOS
l: lowercase	Fœtoproteins α
B: uninflect each words in a term	fœtoproteins α
Ct: retrieve citations	fœtoprotein α
q7: Unicode core Norm	fetoprotein α
q8: strip or map Unicode to ASCII	fetoprotein α
w: sort words by order	fetoprotein alpha
	alpha fetoprotein

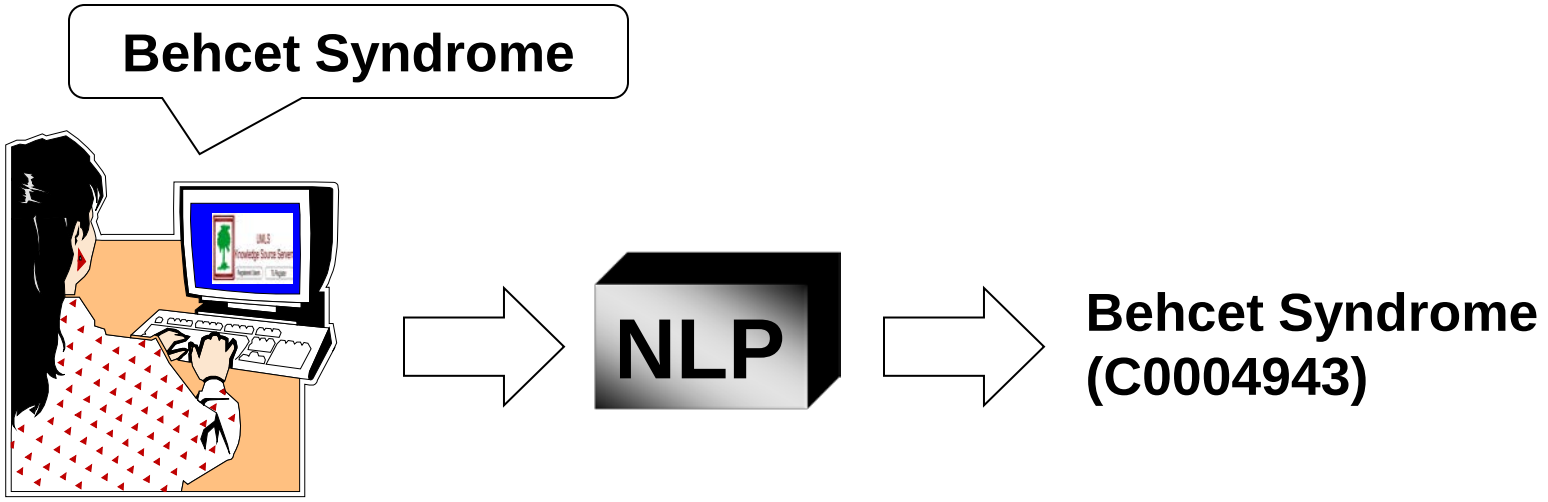
Norm: Example

alpha Fetoprotein
alpha Fetoproteins
alpha-Fetoprotein
alpha-Fetoproteins
Alpha fetoproteins
alpha fetoprotein
alpha Foetoprotein
alpha foetoprotein
alpha fetoproteins
Alpha-fetoprotein
alpha-fetoprotein
Alpha Fetoproteins
Alpha-Fetoprotein
Alpha-fetoprotein NOS
Alpha Fetoprotein
alpha-fetoprotein
ALPHA-FETOPROTEIN
Alpha Fœtoprotein
…



alpha fetoprotein

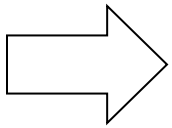
NLP Application - 1



NLP Application - 1

- UMLS Lexical Variants:

Behcet Syndrome
(C0004943|L0004943)

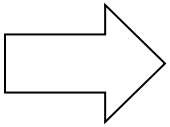


Behcet Syndrome|S0018227
Behcet's Syndrome|S0018228
Behcets Syndrome|S0018229
BEHCET SYNDROME|S0357516
Behcet's syndrome|S0358685
Behcet's syndrome, NOS|S0472369
behcet syndrome|S11855409
behcet's syndrome|S11855411
behcets syndrome|S11855414
syndrome behcet's|S11943704
Behcet syndrome|S1624748
Behçet's syndrome|S3638540

NLP Application - 1

- Other Lexical Variants:

Behcet Syndrome
(C0004943|L0004943)



Behcet Syndrome|S0018227
Behcet's Syndrome|S0018228
Behcets Syndrome|S0018229
BEHCET SYNDROME|S0357516
Behcet's syndrome|S0358685
Behcet's syndrome, NOS|S0472369
behcet syndrome|S11855409
behcet's syndrome|S11855411
behcets syndrome|S11855414
syndrome behcet's|S11943704
Behcet syndrome|S1624748
Behçet's syndrome|S3638540

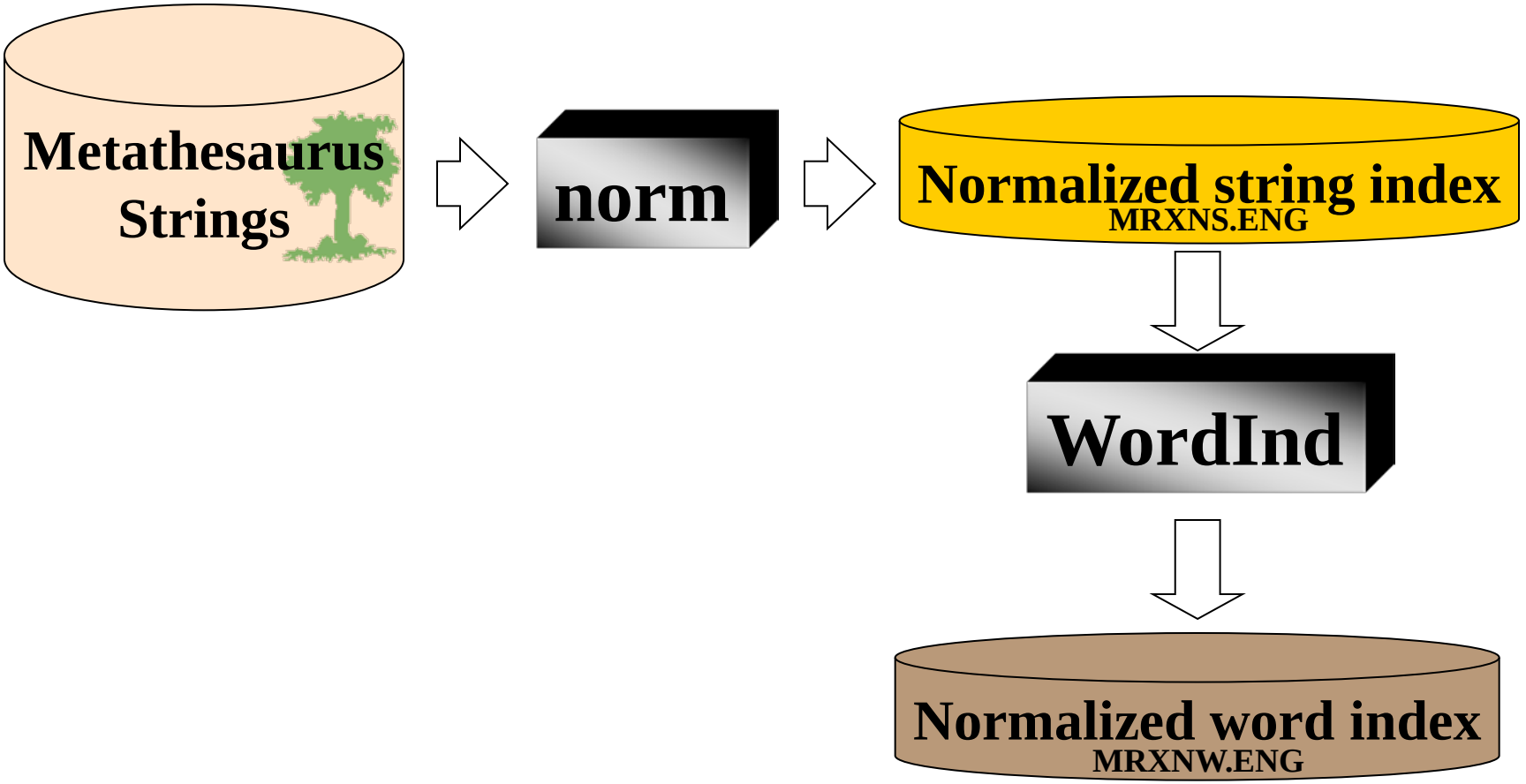
Behçet syndrome
behçet syndrome
BEHÇET SYNDROME
…

NLP Application - 1

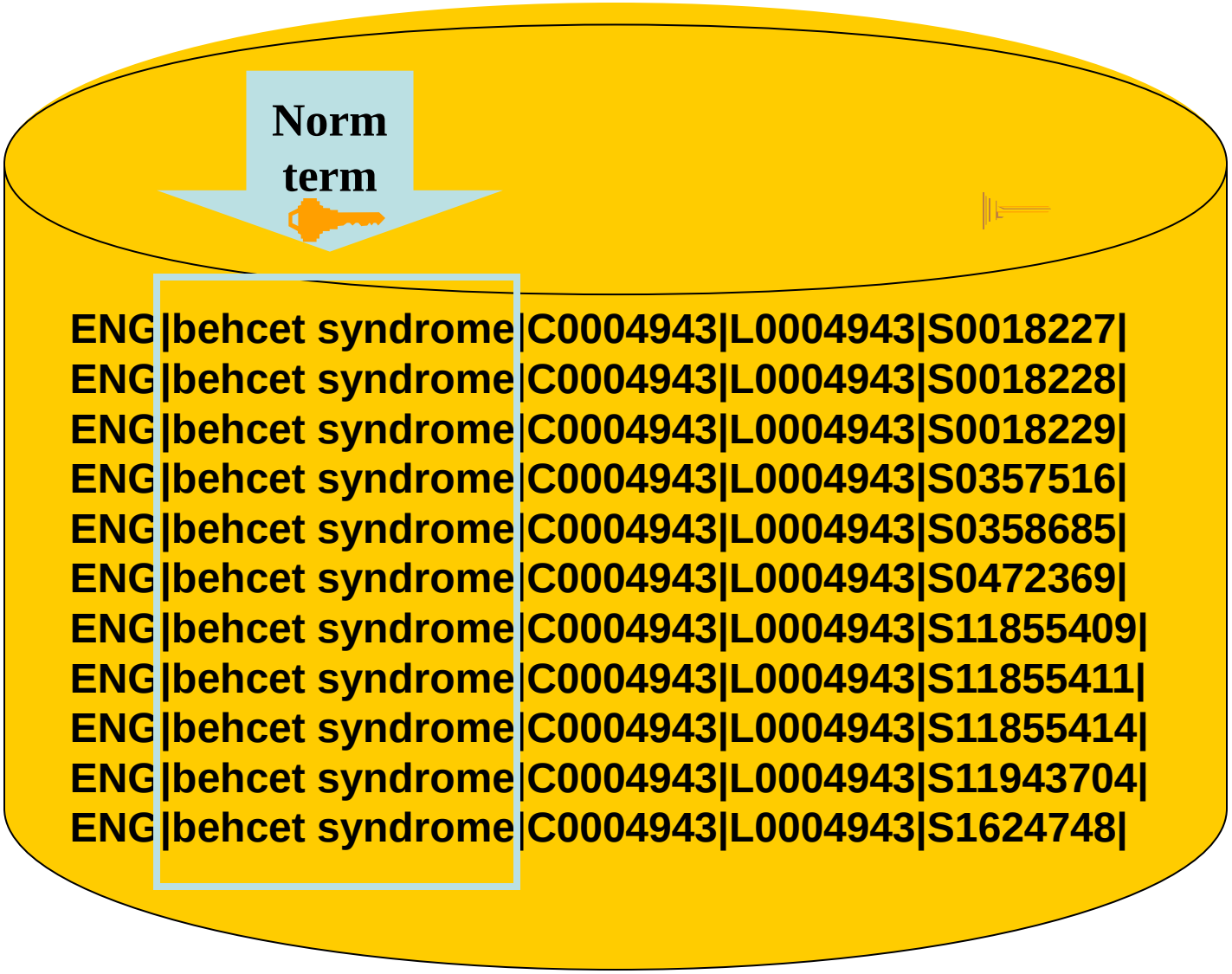
Behcet Syndrome
Behcet's Syndrome
Behcets Syndrome
BEHCET SYNDROME
Behcet's syndrome
Behcet's syndrome, NOS
behcet syndrome
behcet's syndrome
behcets syndrome
syndrome behcet's
Behcet syndrome
Behçet's syndrome
…

Behcet Syndrome (C0004943)

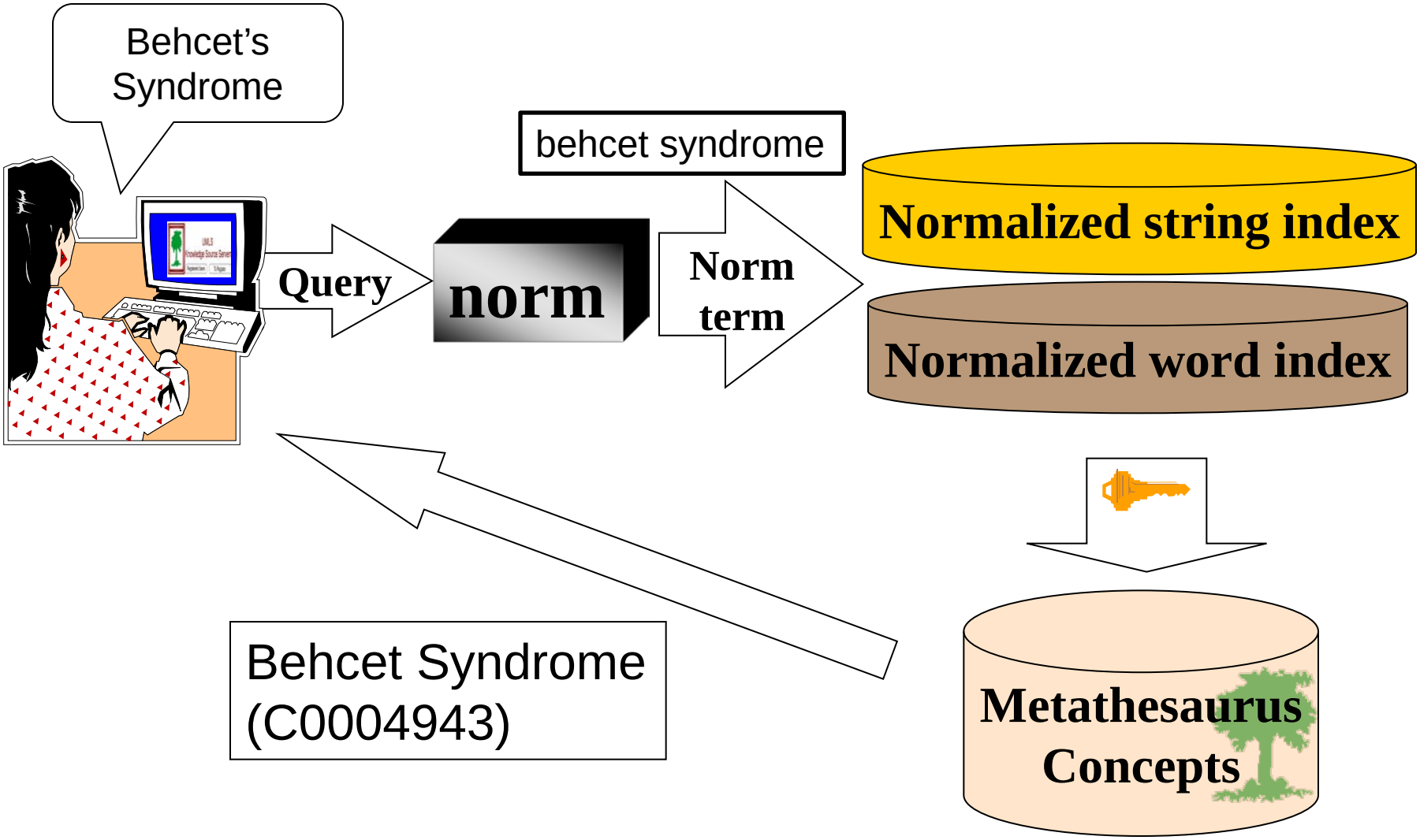
NLP Application - 1



NLP Application - 1



NLP Application - Norm



NLP Application – 2

- uricosuric
- infected ear
- Pneumonia due to other virus
- …

?

NLP Application – 2

- uricosuric
- infected ear
- Pneumonia due to other virus
- …

- uricosuria
- infection ear
- Pneumonia due to other viral
- …

- uricosuria
- ear infection
- other pneumonia viral
- …

lvg

Norm

Normalized string index

Metathesaurus Concepts

- uricosuric
=> C0151582|Uric acid crystalluria
- infected ear
=> C0699744|Infection of ear
- Pneumonia due to other virus
=> C0348677|Other viral pneumonia
- …

NLP Application – 2

(Query Expansion - STMT)

Input term

- synonyms
- acronyms
- expansions
- inflections
- derivations
- …

lvg

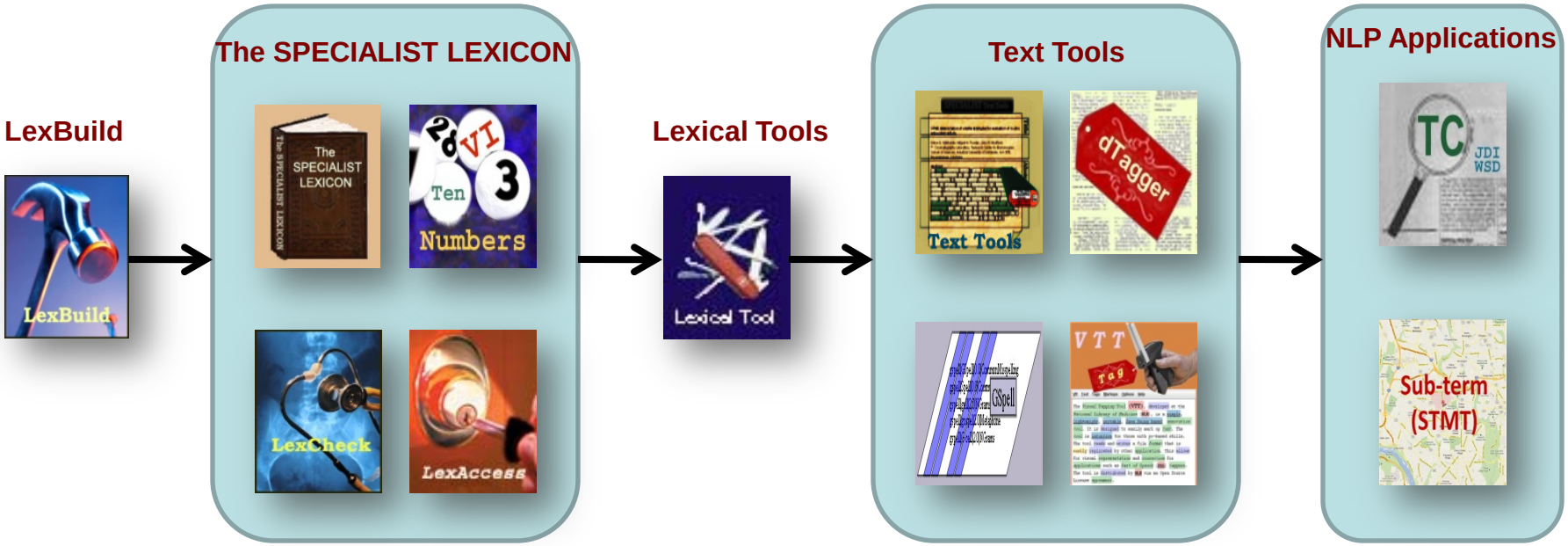
Norm

Normalized string index

Metathesaurus Concepts

Result: concept

The SPECIALIST NLP Tools



- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov

Demo

Questions

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