

# Lexical Tools



# The Lexical Tools

- Introduction
- Norm
- WordInd
- Lvg
- Additional Tools Developed by NLM

# Lexical Tools: Introduction

- Command line tools
  - norm
  - lvg
  - wordInd
- Web GUI, Lexical Gui Tool (lgt)
- Embeddable Java API's

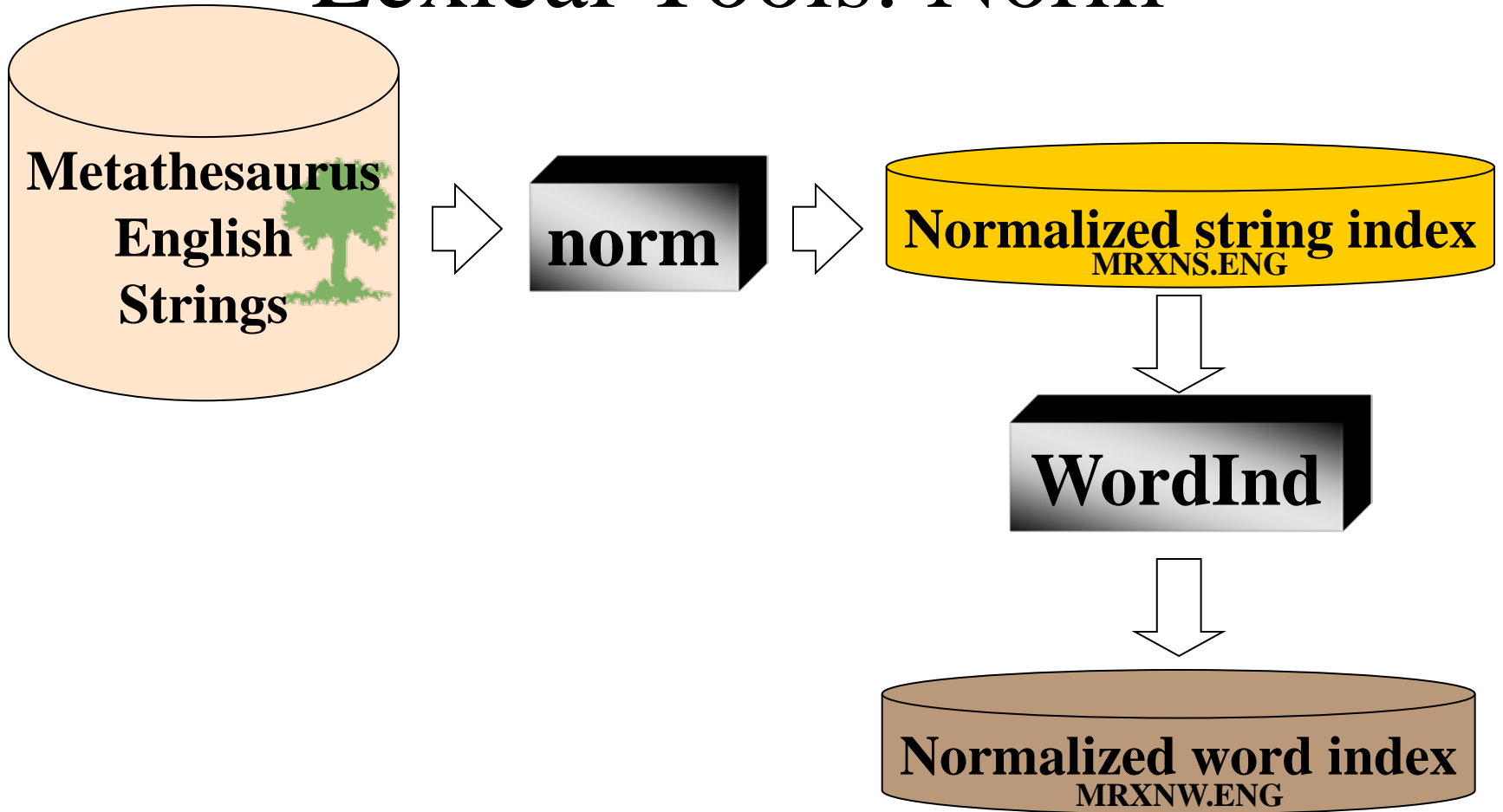
# Lexical Tools: Introduction

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

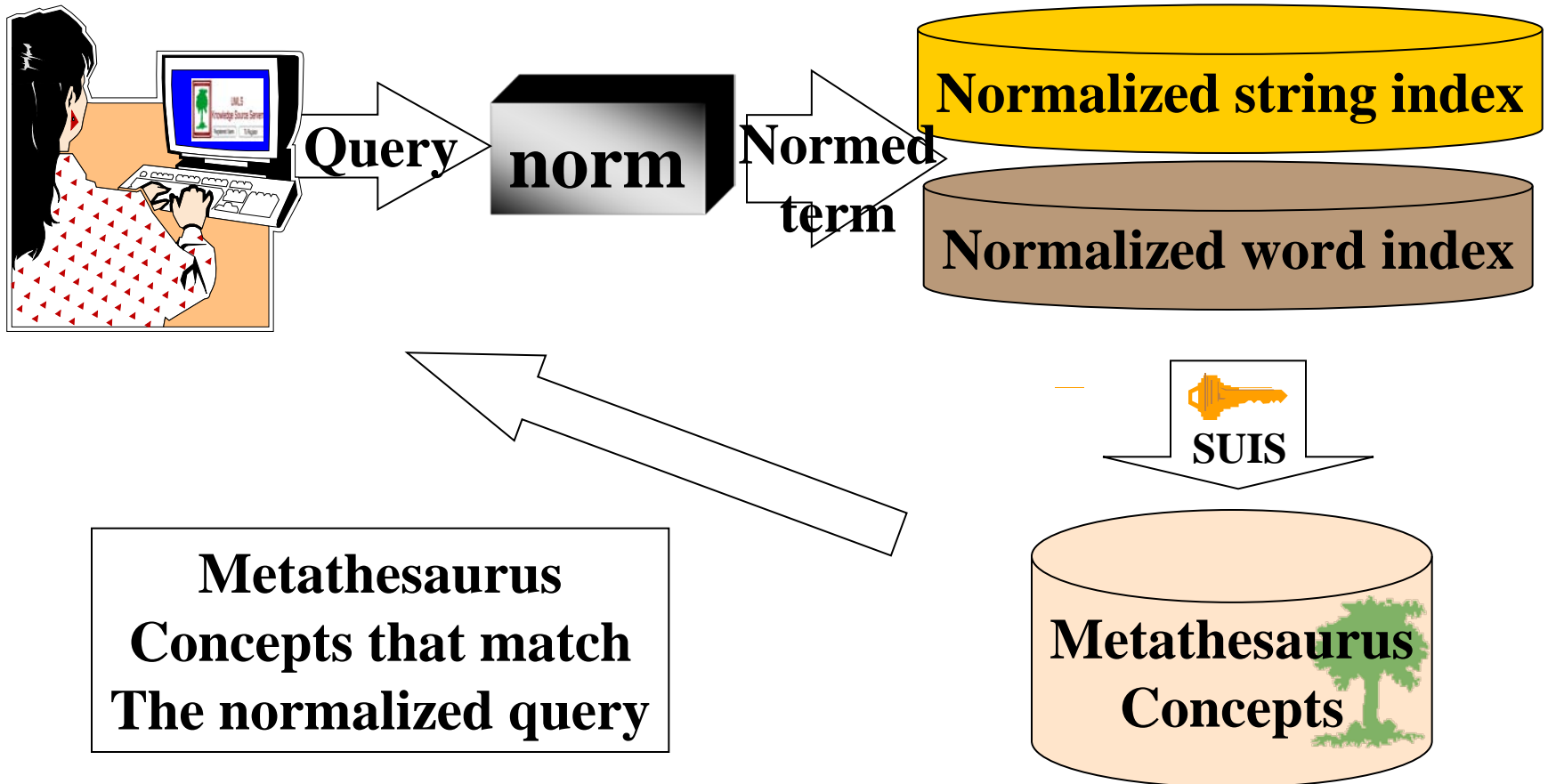
# Lexical Tools: Introduction

- 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



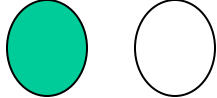
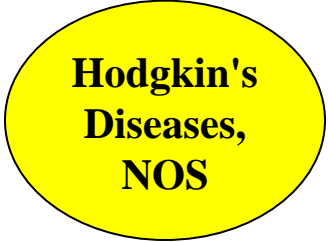
# Lexical Tools: Norm



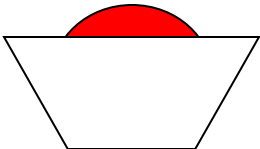
# Lexical Tools: Norm

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





# Lexical Tools: Norm



**remove genitives**

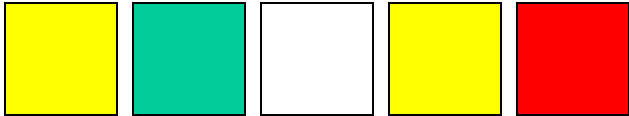
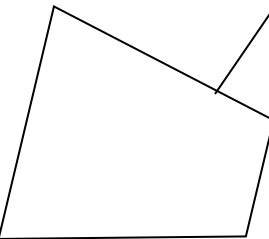
**replace punctuation with spaces**

**remove stop words**

**lowercase**

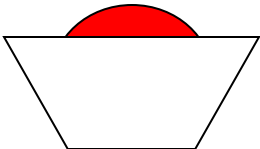
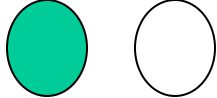
**uninflect each word**

**word order sort**



**Hodgkin's  
Diseases,  
NOS**

# Lexical Tools: Norm



<b>Hodgkin'sDiseases, NOS</b>
<b>Hodgkin Diseases, NOS</b>

**remove genitives** 

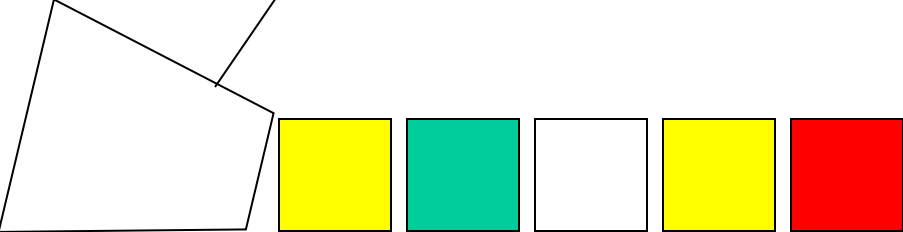
**replace punctuation with spaces**

**remove stop words**

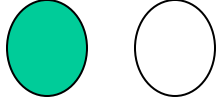
**lowercase**

**uninflect each word**

**word order sort**



**Hodgkin's  
Diseases,  
NOS**



# Lexical Tools: Norm



<b>Hodgkin's Diseases, NOS</b>
<b>Hodgkin Diseases, NOS</b>
<b>Hodgkin Diseases NOS</b>

**remove genitives**

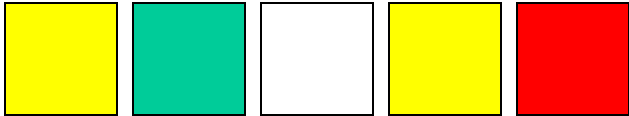
**replace punctuation with spaces** 

**remove stop words**

**lowercase**

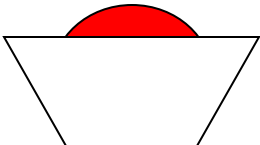
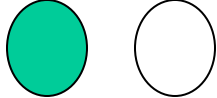
**uninflect each word**

**word order sort**



**Hodgkin's  
Diseases,  
NOS**

# Lexical Tools: Norm



<b>Hodgkin'sDiseases, NOS</b>
<b>Hodgkin Diseases, NOS</b>
<b>Hodgkin Diseases NOS</b>
<b>Hodgkin Diseases</b>

**remove genitives**

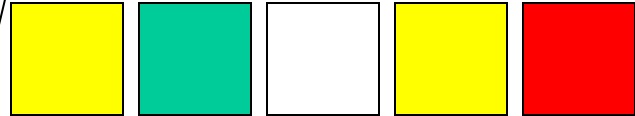
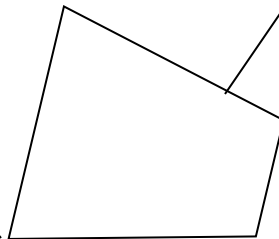
**replace punctuation with spaces**

**remove stop words**

**lowercase**

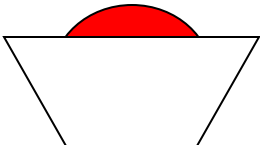
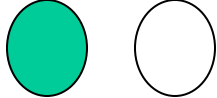
**uninflect each word**

**word order sort**



**Hodgkin's  
Diseases,  
NOS**

# Lexical Tools: Norm



<b>Hodgkin'sDiseases, NOS</b>
<b>Hodgkin Diseases, NOS</b>
<b>Hodgkin Diseases NOS</b>
<b>Hodgkin Diseases</b>
<b>hodgkin diseases</b>

**remove genitives**

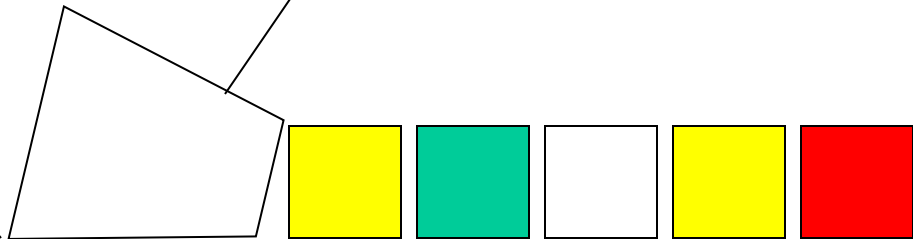
**replace punctuation with spaces**

**remove stop words**

**lowercase** 

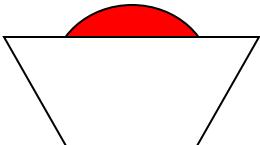
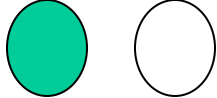
**uninflect each word**

**word order sort**



**Hodgkin's  
Diseases,  
NOS**

# Lexical Tools: Norm



**remove genitives**

**replace punctuation with spaces**

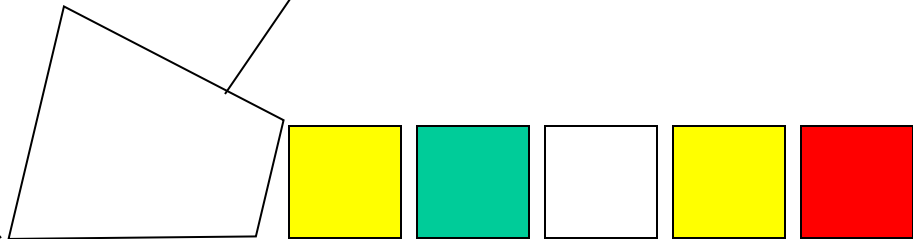
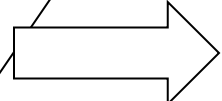
**remove stop words**

**lowercase**

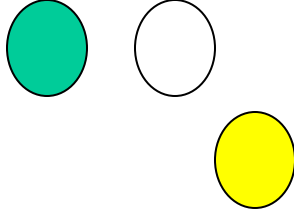
**uninflect each word**

**word order sort**

<b>Hodgkin'sDiseases, NOS</b>
<b>Hodgkin Diseases, NOS</b>
<b>Hodgkin Diseases NOS</b>
<b>Hodgkin Diseases</b>
<b>hodgkin diseases</b>
<b>hodgkin disease</b>



**Hodgkin's  
Diseases,  
NOS**



# Lexical Tools: Norm

**remove genitives**

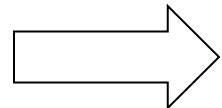
**replace punctuation with spaces**

**remove stop words**

**lowercase**

**uninflect each word**

**word order sort**

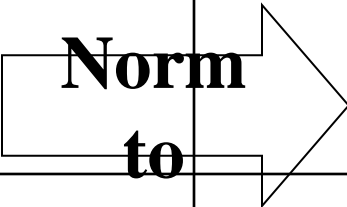


<b>Hodgkin'sDiseases, NOS</b>
<b>Hodgkin Diseases, NOS</b>
<b>Hodgkin Diseases NOS</b>
<b>Hodgkin Diseases</b>
<b>hodgkin diseases</b>
<b>hodgkin disease</b>
<b>disease hodgkin</b>



# Lexical Tools: Norm

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

**Norm**  
**to** 



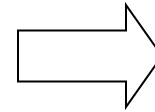
N A T E P A T R I S S U M M I Q U I T E L A F E R O C I A F R A N C I S  
 D A M I H I R I T E C R U C I S U I C T R I C I A C A R M I N A F A R I  
 N A C A E L E S T A N I M A L M I <sup>mirro</sup> V O L E T O R E I O H A N N E S  
 T R A N S P E N E T R A <sup>1. relecte</sup> <sup>2. rualcom</sup> <sup>folcom</sup> A E T U R O O M I N E U I D I T  
 E O U M S O L E U E R <sup>1. relecte</sup> P O L O R U M H O C  
 G R A T I A S I C Q U A N N I A N T E O M N E S U I U I D A F A T U  
 D O N A T A U T Q U A R T H O M I N I S I B I S U M E R E T A U C T O R  
 S C R I P S E R I T A T Q U E P I O D E U S U N E T A L M  
 S E P E R C U P A T R E Q I T A Q U E S A L U S Q U E  
 S I T N A T U S F A C T U S Q U E C A R O D O M I N A T O R I N O R B E  
 H U N C L E O H U N C U I T U L U S R E C E D A N T P O N T I F I C E Q  
 U T L E O Q U I F O R T I S R E T U L I T C E R T A M I N E P R A E D A  
 H O S T I A E T O B T U L E R A T S U M M U S S E R I T E S A C E R D O S  
 M Y S T I C A D O N A S U I S C O N S O R T I B O P T I M E D O N A N S  
 N E M P E D A T O R U M F A T O S E P T E M E T P I E P A N U M  
 D A T O T U Q U O Q U E U I U I D E C A N T U  
 H O C I N T E R F I A P I E C A S  
 D A T R H E A P O N C I F I C E Q U E  
 F O R M A N T I M A N U S F A R M I N A B I L I S I N F A N S  
 I N B E T H L E E G E N I T M A R E M I R A B I L I S I N F A N S  
 I P S E S A T U S M A R I A M T I D O T I S S I M A C U R A H U I C  
 N O B I L I S A T Q U E P U E R P E R S O N A V E T U S T A D I E R U M  
 Q U I V E N I T D I E D O M D E B O S R A H I C U E S T E C R U E N T A  
 C A L C A T U R U S E R A T Q U I S O L U S T O R C U L A R A U C T O R  
 I N C R U C E P E N S A N D Q U I S U S T I N E T A S T R A S U P E R N  
 U T C R U X A L M A F O R E T D I U I N O H A I C M U N E R E D I U E S  
 N A S C R I B E N S B E N E M A T I B E U S D E D I T O R D I N E P R I M  
 Q U I I N F A C I E F I R M A T I B O R I G I N E D A I D  
 P R O G E N I T E S S E B O R I G I N E D A I D  
 C U M M O N S T R A U T I B O R I G I N E D A I D  
 Q U O D G E N H O C D E D E R I T P I S F I D A B R A A  
 E X P U L S O N A M R I T E P R A T I O N I S O M N E R E T E X T U M  
 C O N T I N E T H O C V E R R A T I O N I S C R E D E R E S I G N U M  
 N E M P E D E C E T D U D U M C R I S T U S Q I A N A S C I E R I L L A  
 P R O M I S S U S S T I R P E S T S A L U A T O R M A X I M U S O R B I

This eighth-century Latin manuscript was carefully inscribed, but without spaces to mark word boundaries.

# Lexical Tools: WordInd

**Wordind** is a tool to break terms into words.

It is used to take a row from a **Metathesaurus** table that contains a term, sentence, paragraph, story, and break the text part of that row into its constituent words.



**wordind**  
is  
a  
tool  
to  
break  
terms  
into  
words  
it  
is  
used  
to

# 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

<code>-t[:Num]</code>	Defines what field to tokenize
<code>-f[:Num[:Num]]</code>	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



The screenshot shows a Netscape browser window with the following elements:

- Address Bar:** Location: <http://umlslex.nlm.nih.gov:8888/WebLvg/jsp/lvg/lvg.jsp?type=Lvg>
- Main Content:**
  - Logo: 
  - Text: *Lexical Web Tools Java*
  - Text: *- Lvg 2003 -*
  - Navigation bar: [Home](#) | [Norm](#) | [LuiNorm](#) | [WordInd](#) | [Lvg](#) | [Contact Us](#) | [Releases](#) | [About](#)
  - Options: [Options](#): [Input](#) | [Global Behavior](#) | [Flow Setup](#) | [Output](#) | [Version](#) | [Reset](#)
- Status Bar:** Document: Done

# 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>
c	<u>Tokenize a term into "words"</u>
ca	<u>Tokenize, keep everything</u>
ch	<u>Tokenize without breaking hyphens</u>

# Lexical Tools: Flow Components

Mnemonic	Tool
Ct	<u>Retrieve the citation term</u>
d	<u>Generate derivational variants</u>
dc~N	<u>Generate derivational variants with specifying output categories</u>
R	<u>Generate derivational variants, recursively</u>
E	<u>Retrieve the unique EUI for a term</u>
f	<u>Filter output to contain only forms from lexicon</u>



# Lexical Tools: Flow Components

Mnemonic	Tool
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>
is	<u>Generate inflectional variants, by Categories and inflections</u>
Si	<u>Inflections to simplified inflections</u>

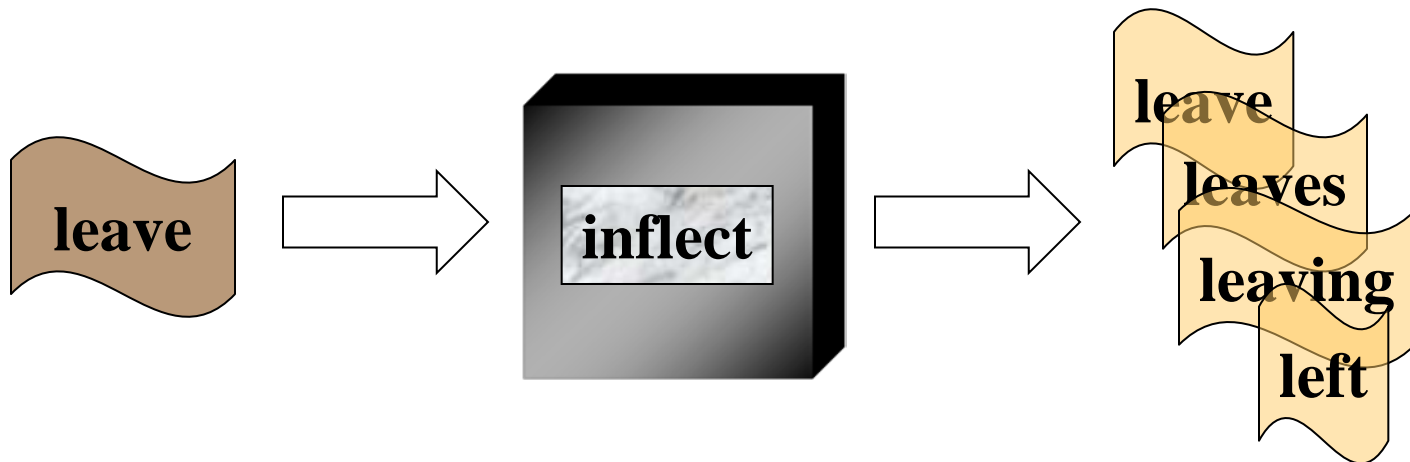
# Lexical Tools: Flow Components

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>

# 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>
r	<u>Generate synonyms, recursively</u>

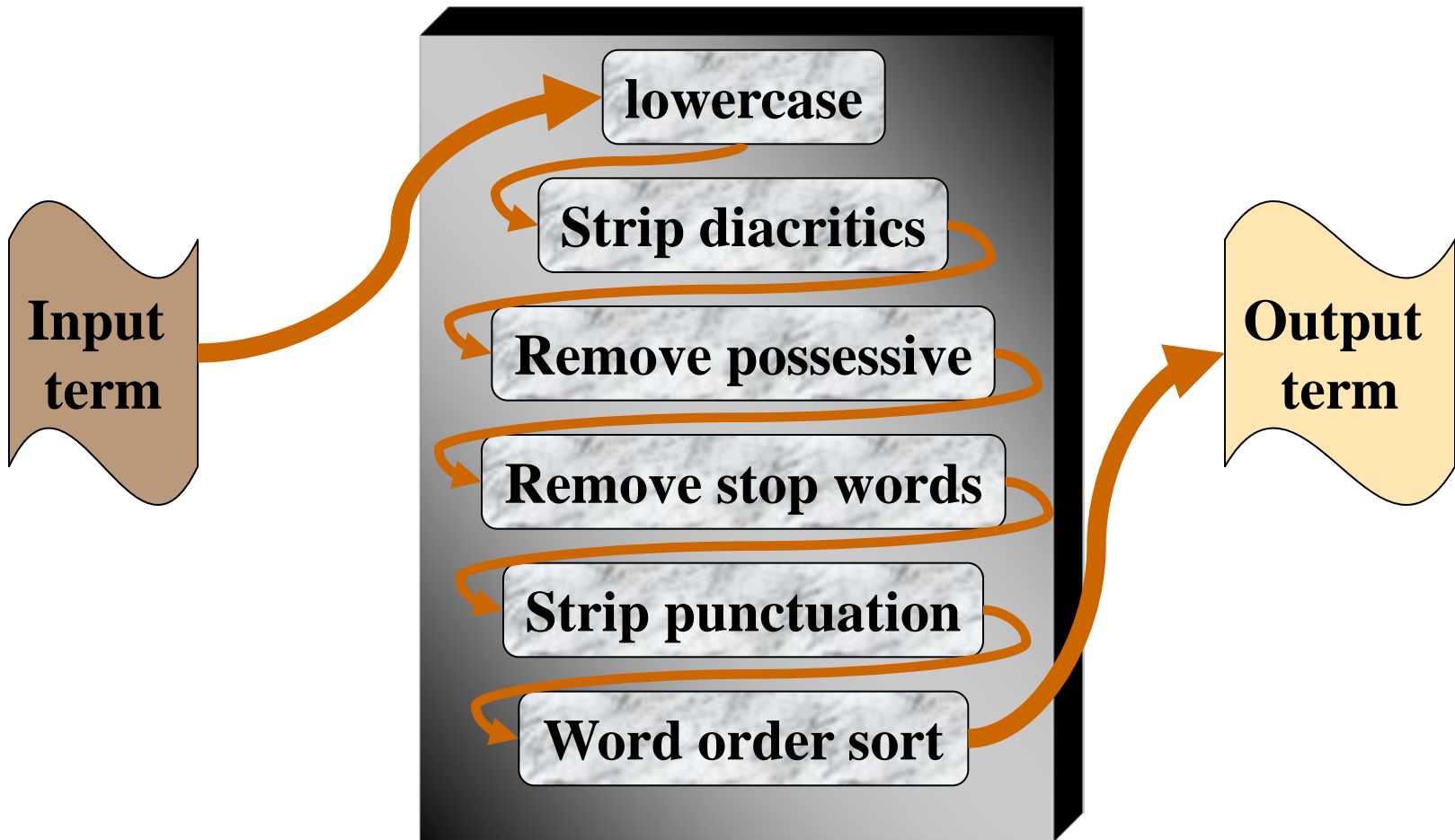
# Lexical Tools: Flows



# 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

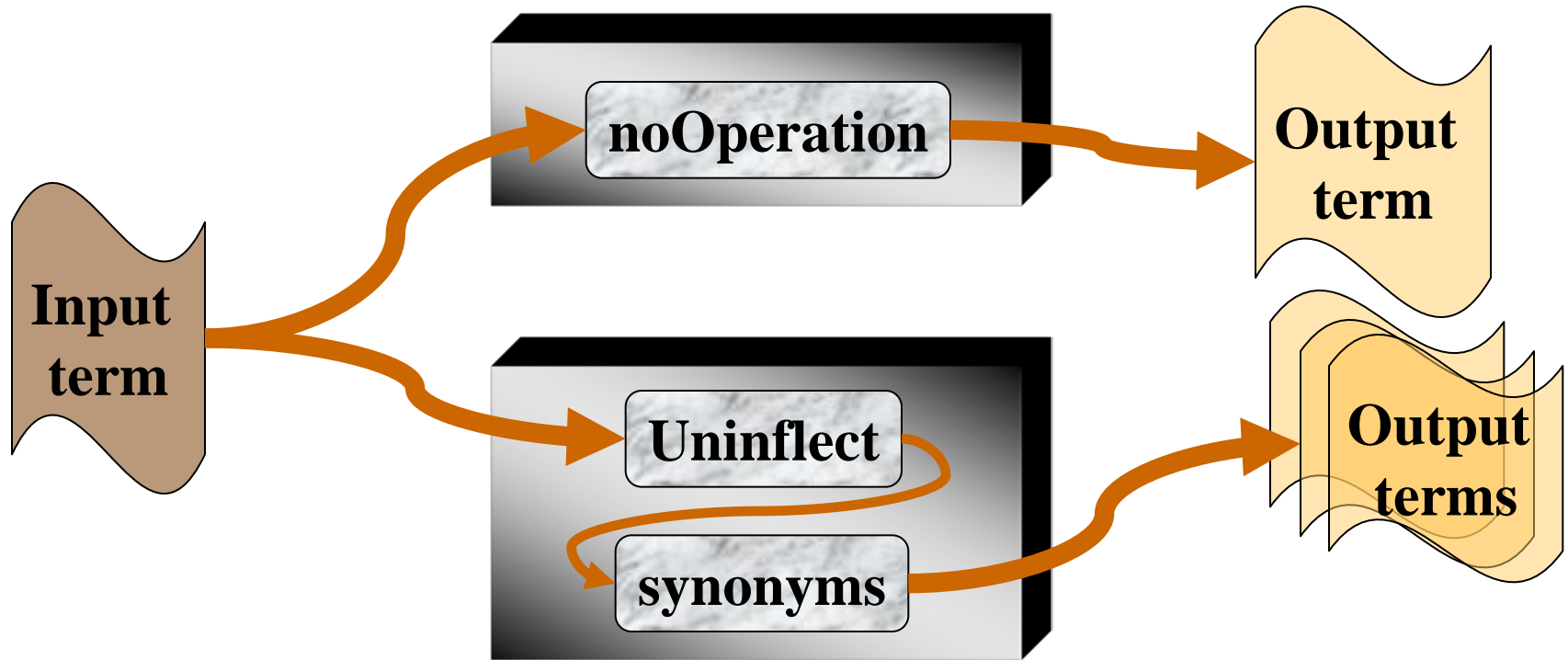
**The Gougerot-Sjögren's Syndrome**

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

```
> 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|
```

**First Flow**

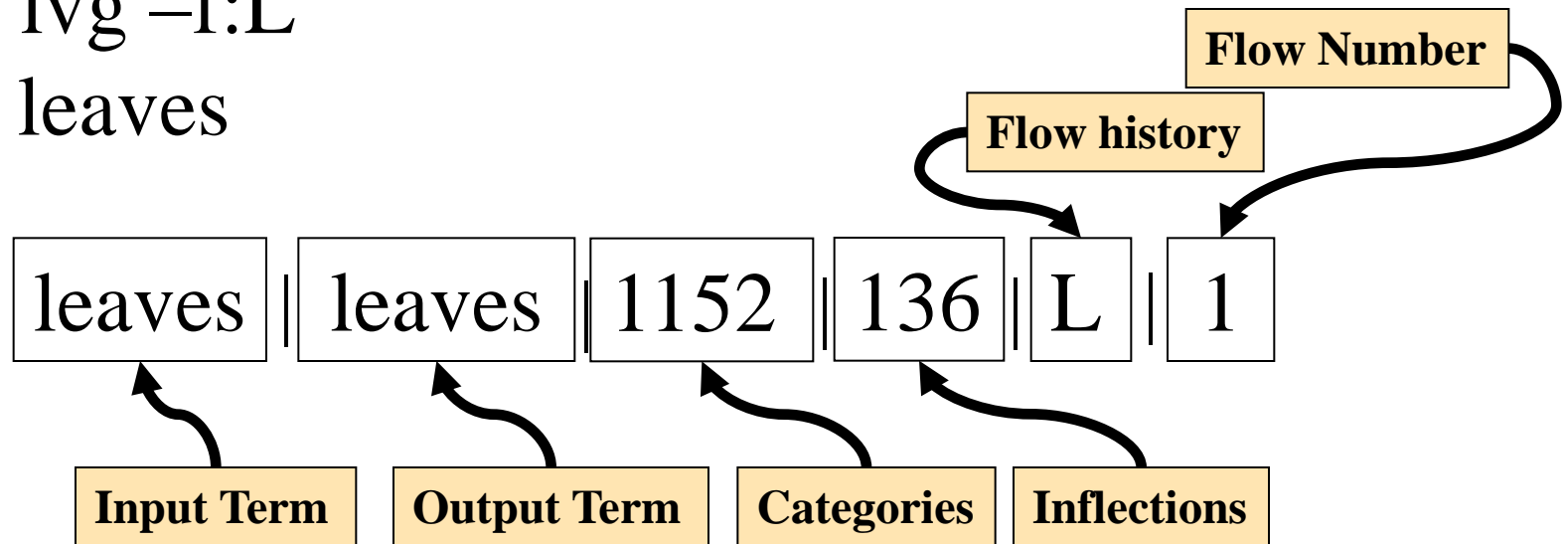


**Second Flow**

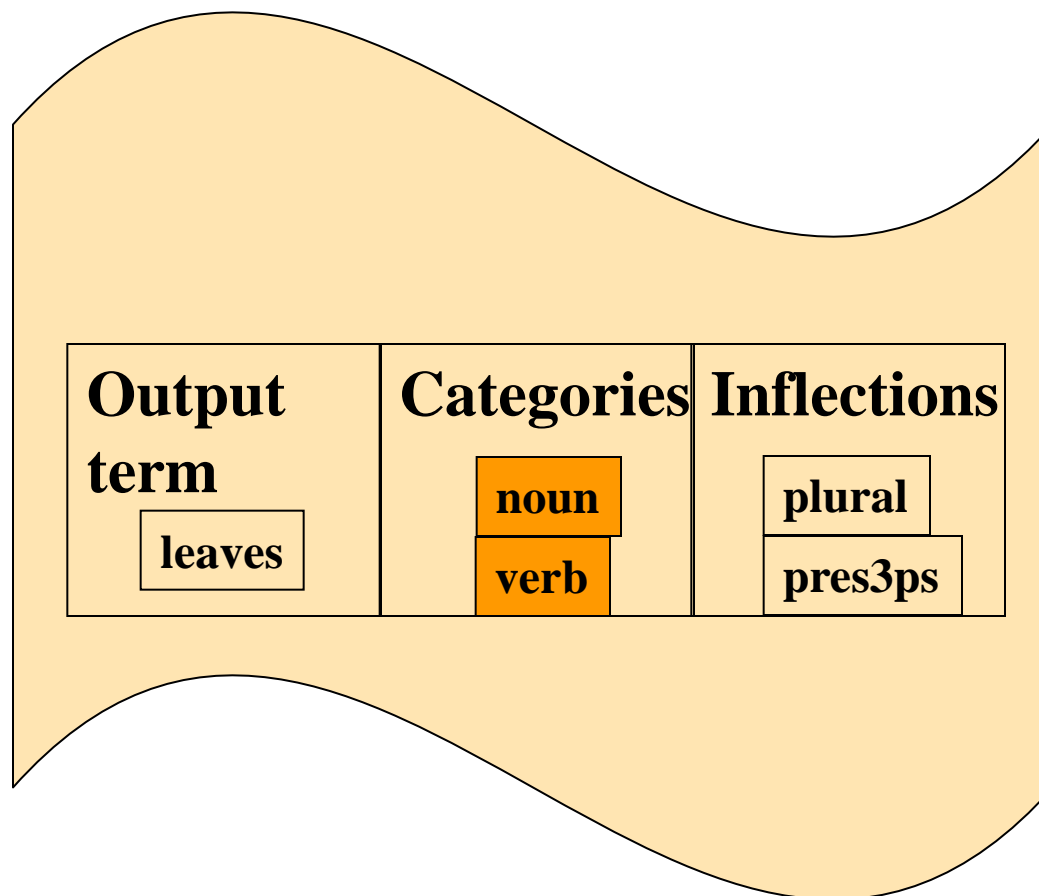


# Lexical Tools: Fielded Output

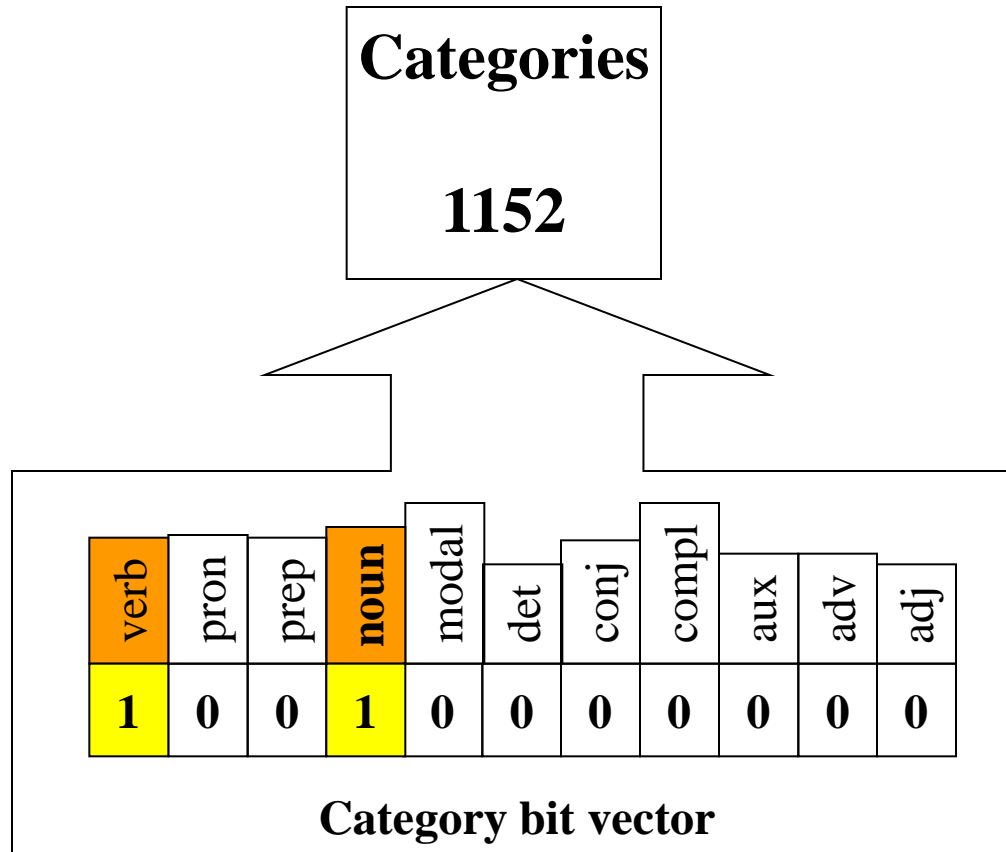
> lvg -f:L  
leaves



# Lexical Tools: Fielded Output



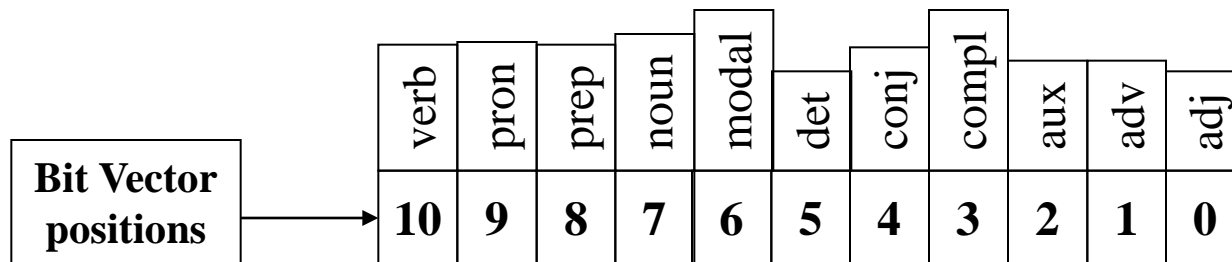
# Lexical Tools: Categories



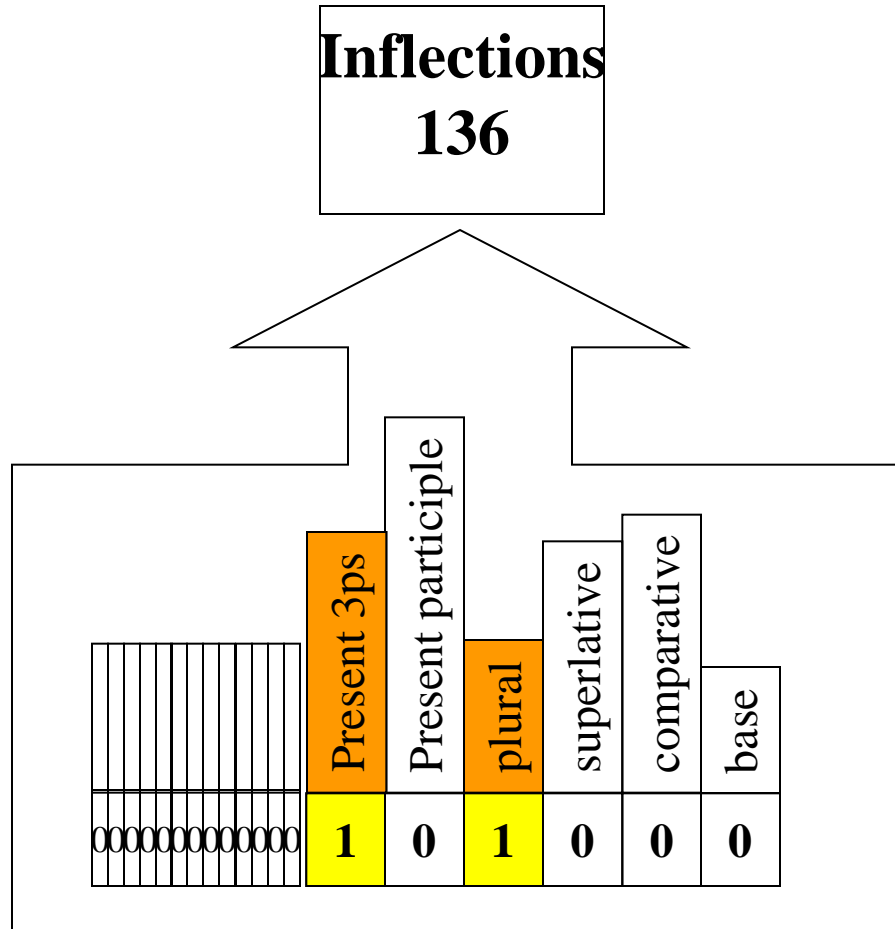
# Lexical Tools: Categories

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 <sup>rd</sup> Person Singular	128

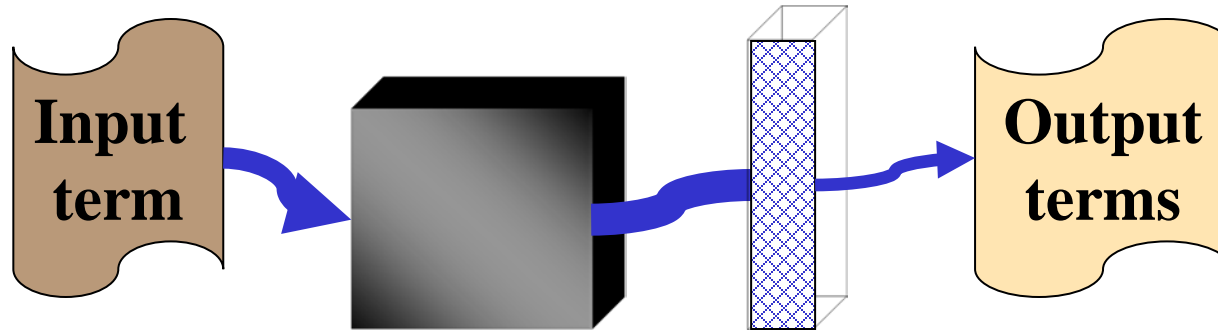
# Lexical Tools: Inflections

Inflection	Expanded	Mapped
Positive	256	1
Singular	512	1
Infinitive	1024	1
Pres 123p	2048	128
Past Neg	4096	32
pres123pNeg	8192	128
Pres1s	16384	128
past1p23pNeg	32768	32
past1s3sNeg	65536	32

Inflection	Expanded	Mapped
Pres1p23p	131072	32
pres1p23p	262144	128
pres1p23pNeg	524288	128
past1s3s	1048576	32
pres	2097152	128
pres3sNeg	4194304	128
presNeg	8388608	128
all	16777215	255



# Lexical Tools: Post Flow Options



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

# Lexical Tools: Post Flow Options

Show category names

Show inflection names

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

Show the category and  
inflection names

*phosphoprotein*

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

*sclerosing*

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

# 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 8<sup>th</sup> and 6<sup>th</sup> 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
```

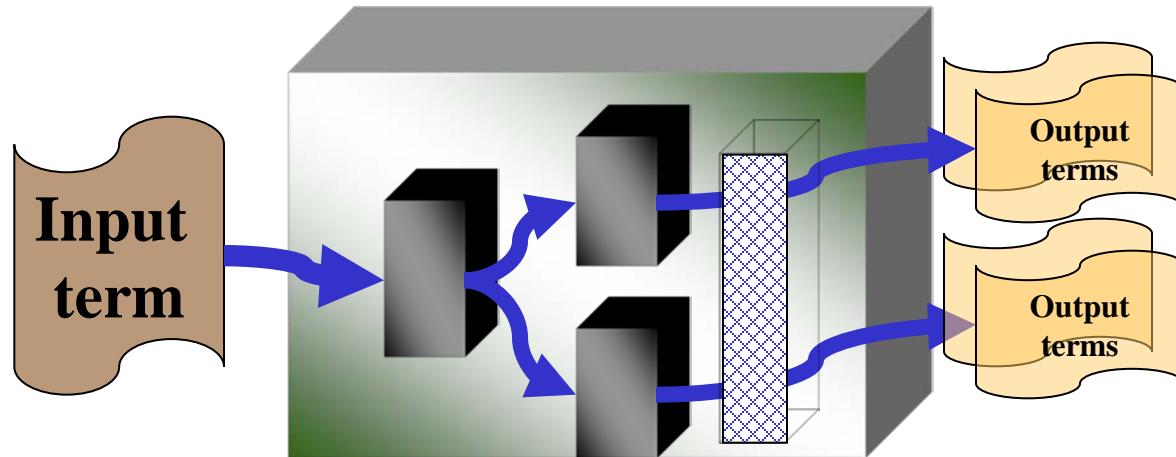
```
foo | foo | 128 | 1 | i | 1 |
```

```
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: Global Behaviors



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

# 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: Inflect Simple

**-f:is**

**Generate inflections noting simplified inflection tags**

> lvg **-f:is**

cut|cut|1024|1|is|1|

cut|cutting|1024|16|is|1|

cut|cut|1024|32|is|1|

cut|cut|1024|64|is|1|

cut|cut|1024|128|is|1|

cut|cuts|1024|128|is|1|

cut|cut|128|1|is|1|

cut|cuts|128|8|is|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 | **reducer** | <noun> | <base> | d | 1 |

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 Resources



## UMLS Knowledge Source Server (UMLSKS)

UMLSKS Version 2.1

UMLS Releases: 2002 2002AB

**Metathesaurus**

**Semantic Network**

**SPECIALIST Lexicon**

[Documentation](#)

[Resources](#)

[Views/Profiles](#)

[Logout](#)

The following resources are provided as part of the UMLSKS suite of tools.

[Download UMLS Knowledge Sources](#)

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

[Metathesaurus Resources](#)

Tools and resources for the UMLS Metathesaurus.

[NLP and Lexical Resources](#)

Generic configurable Natural Language Processing tools and utilities for manipulating lexical data.

# 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