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

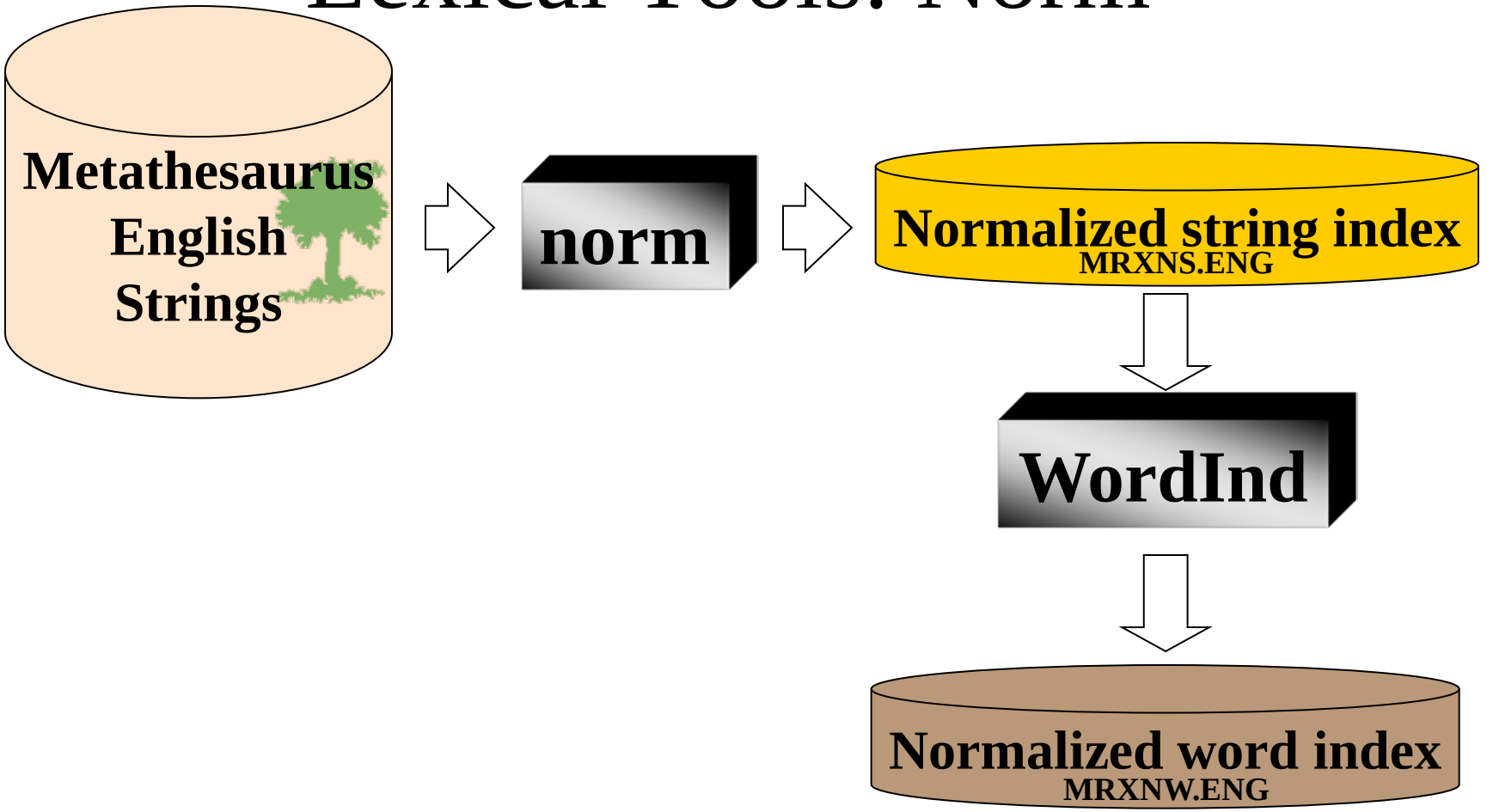
Lexical Tools: Introduction

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

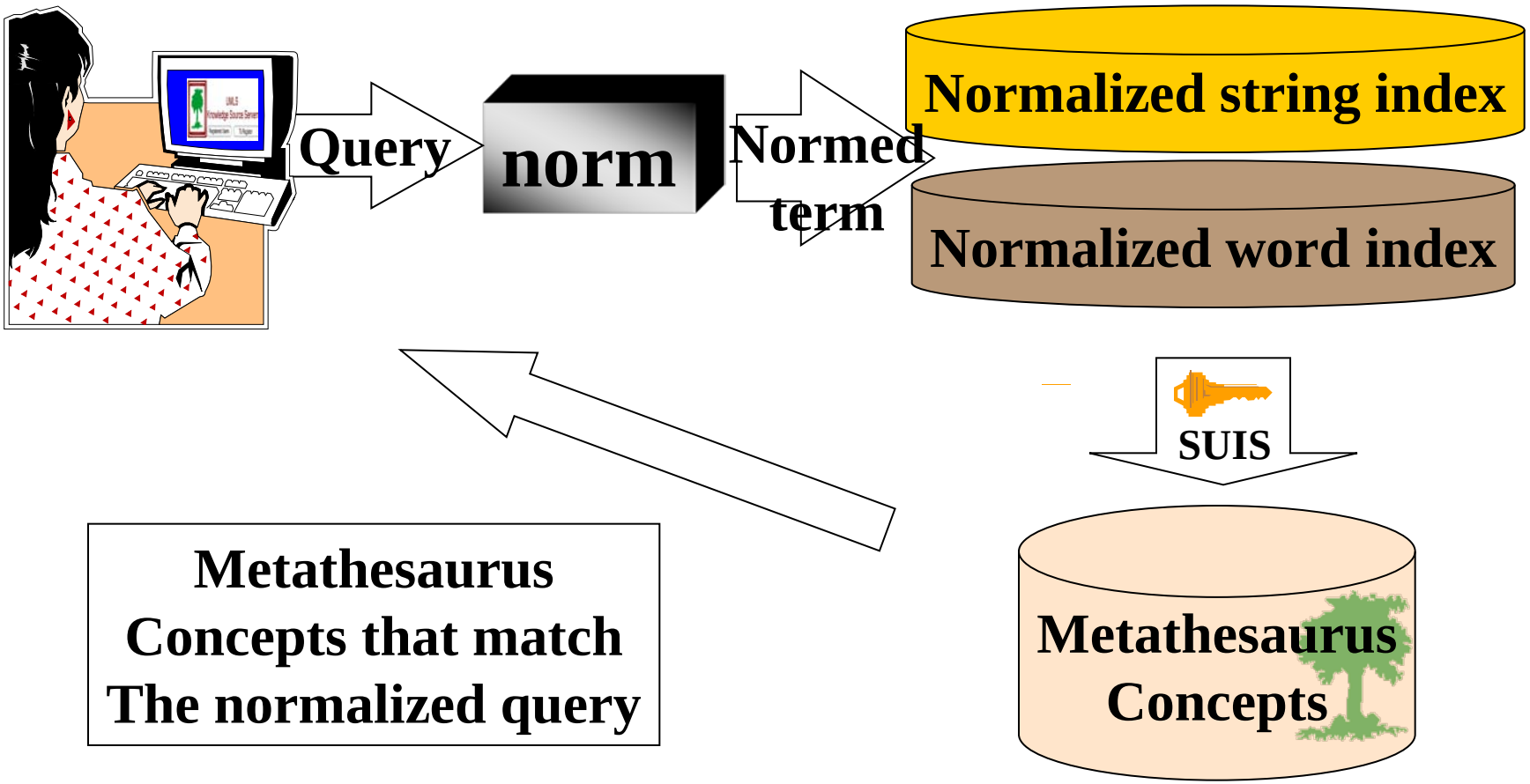
Lexical Tools: Introduction

- 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

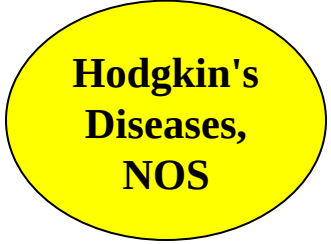


Lexical Tools: Norm

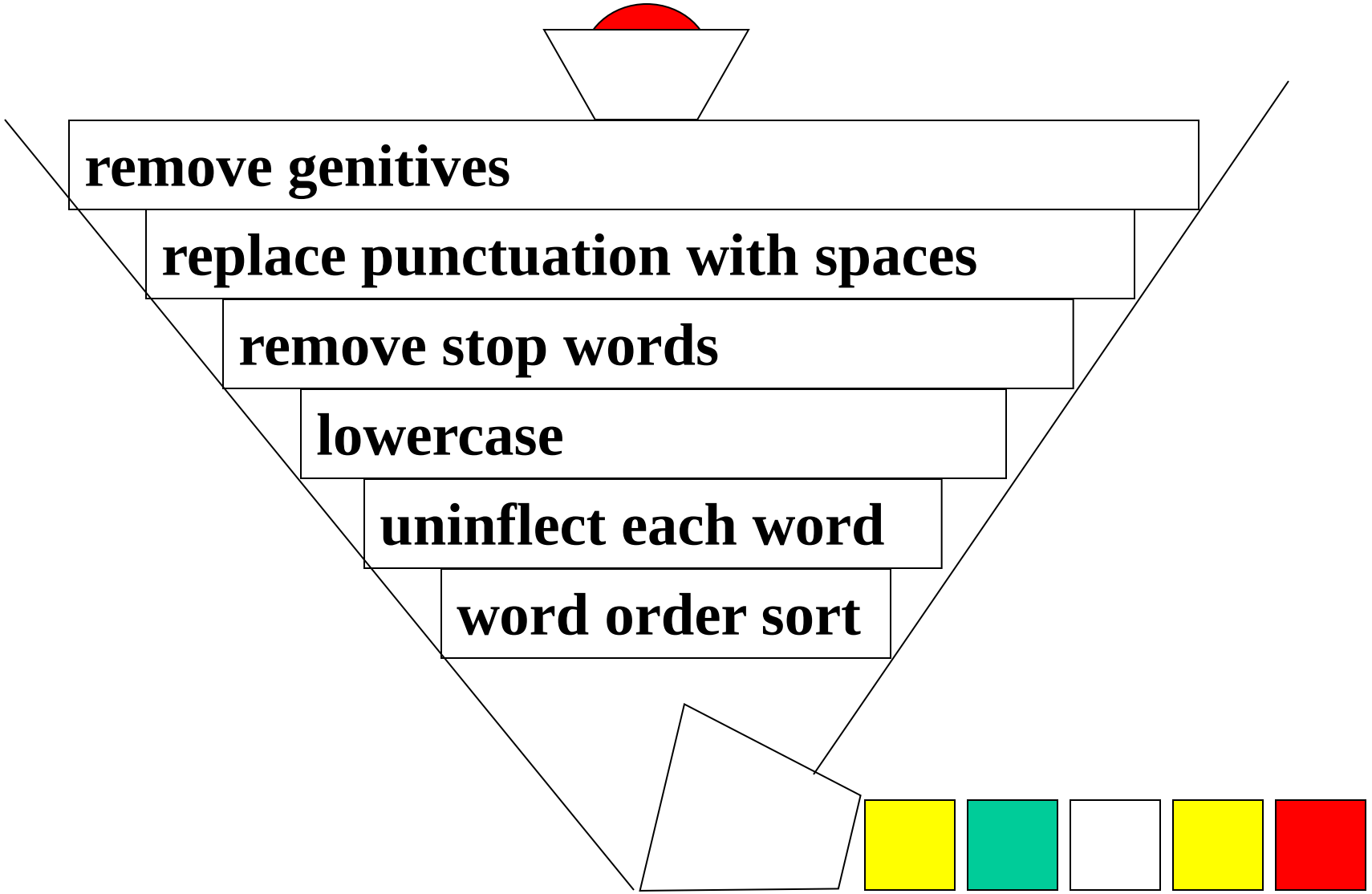


Lexical Tools: Norm

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

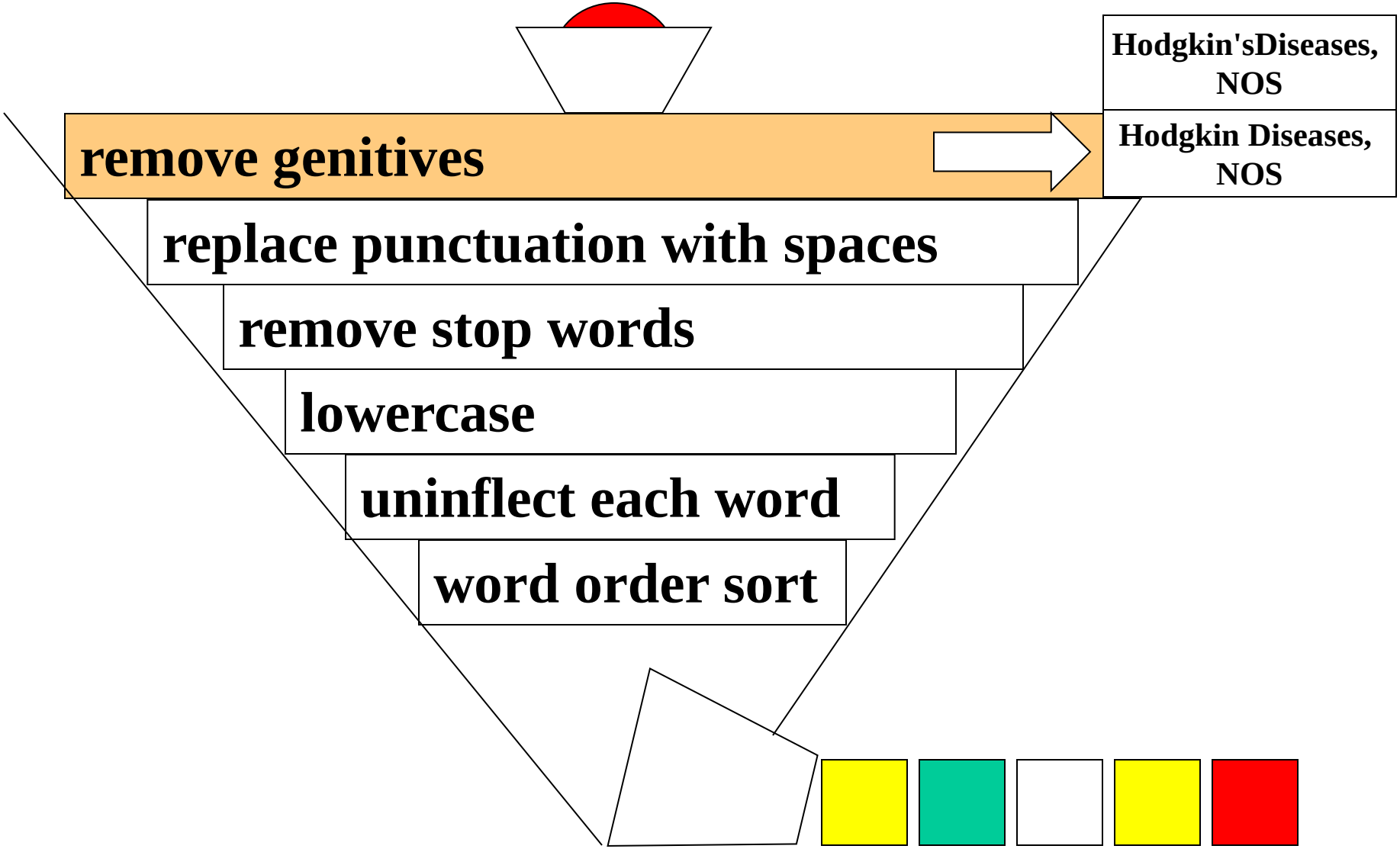


Lexical Tools: Norm



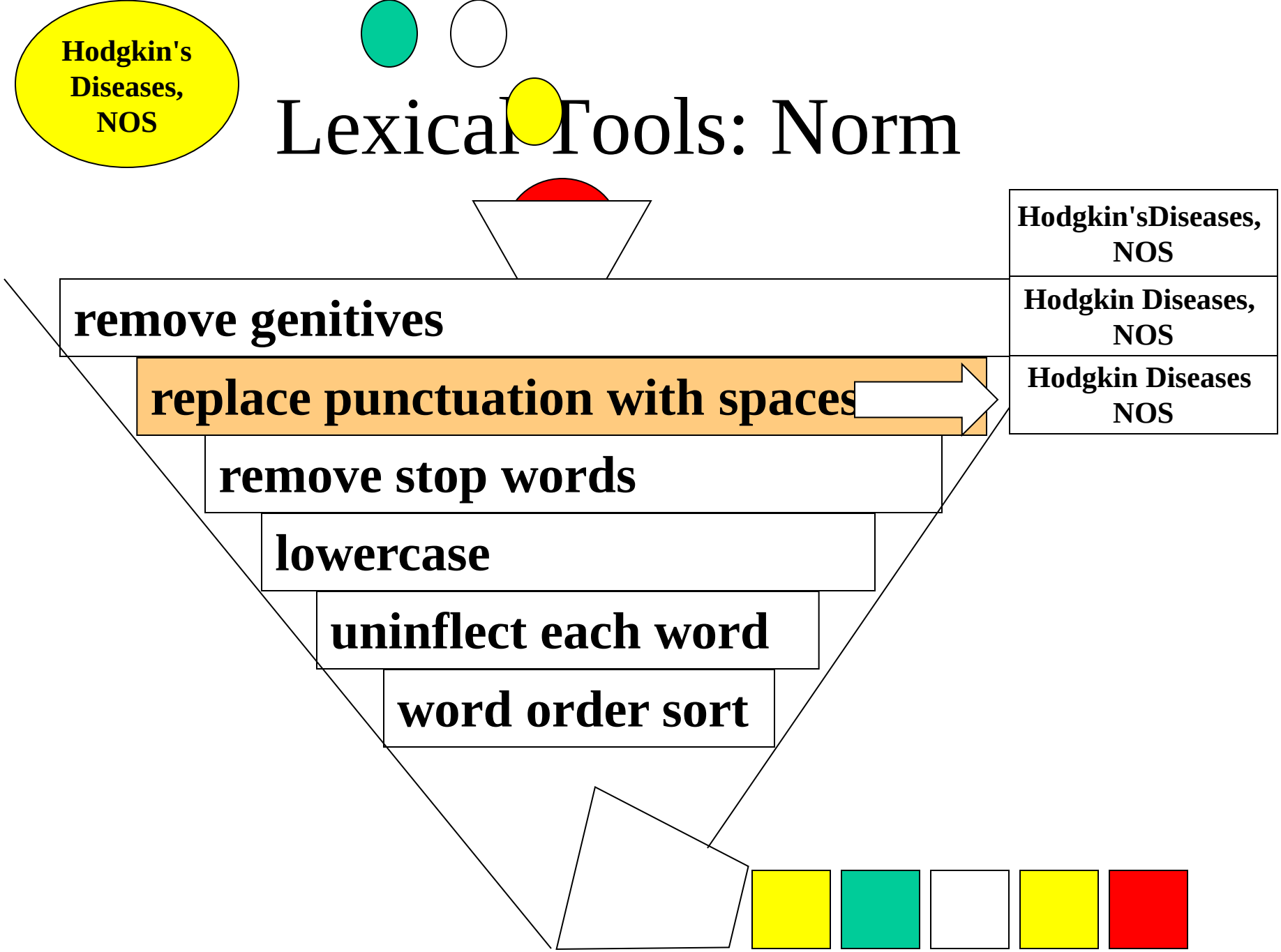
Hodgkin's Diseases, NOS

Lexical Tools: Norm



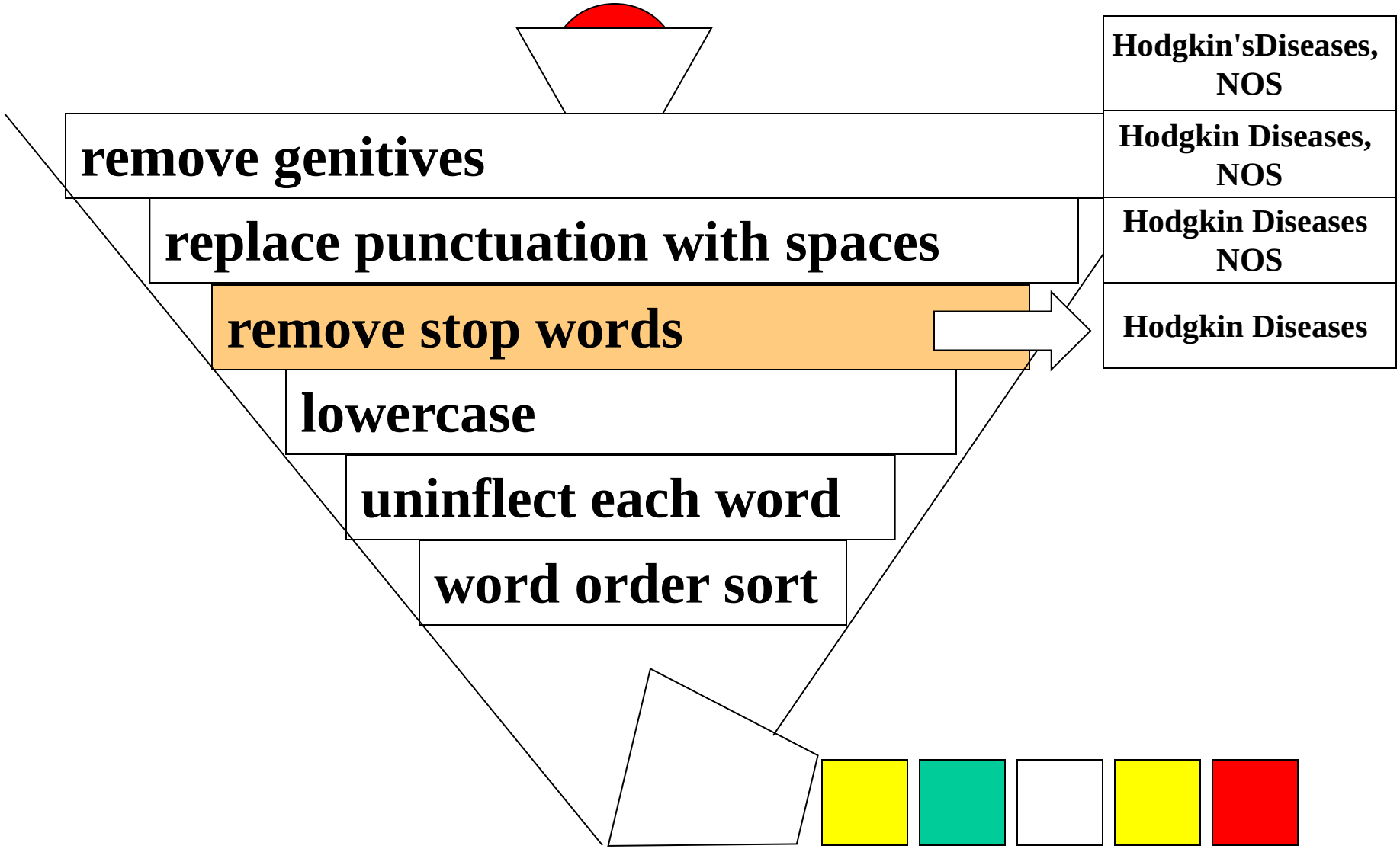
Hodgkin's Diseases, NOS

Lexical Tools: Norm



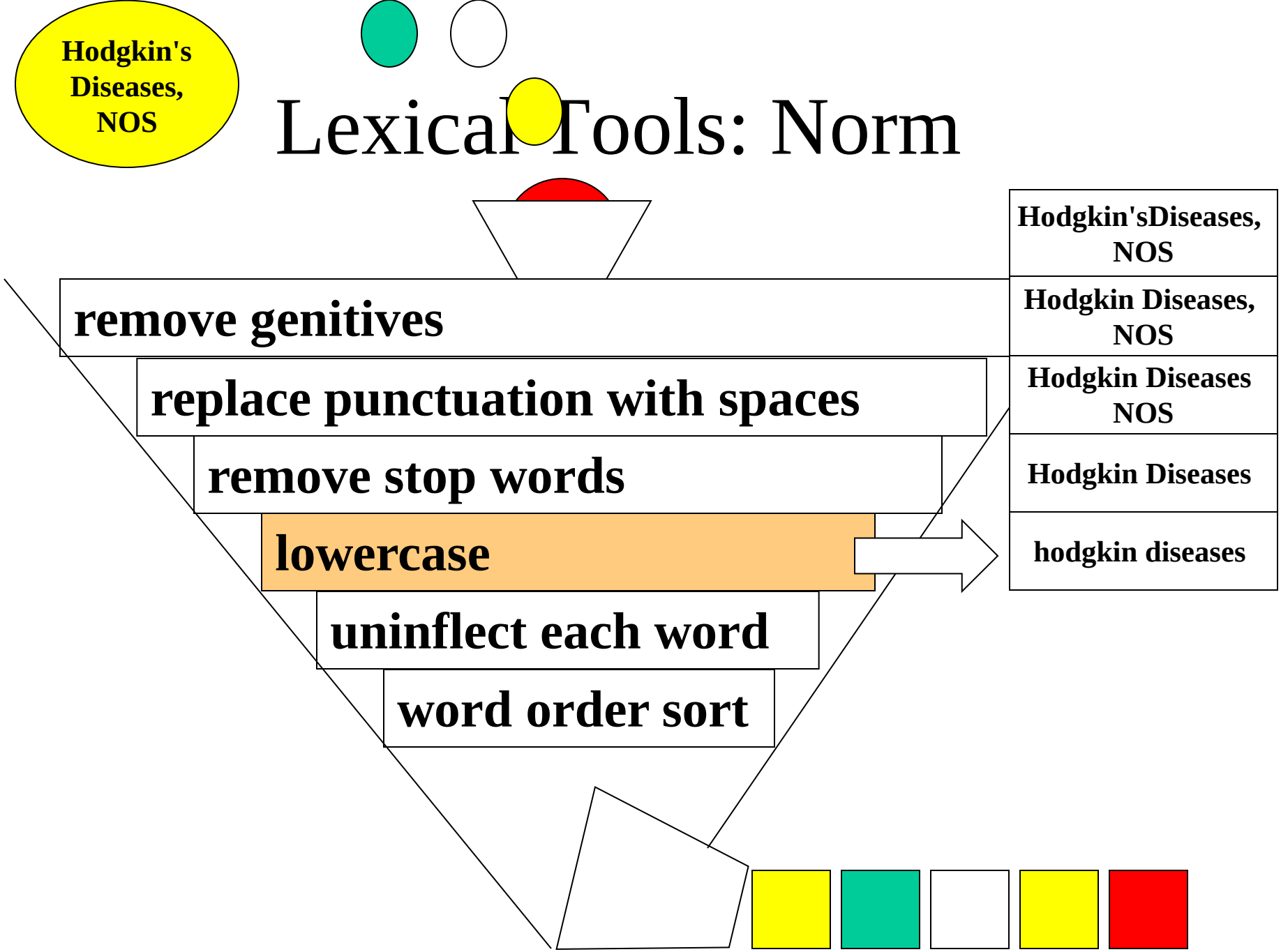
Hodgkin's Diseases, NOS

Lexical Tools: Norm



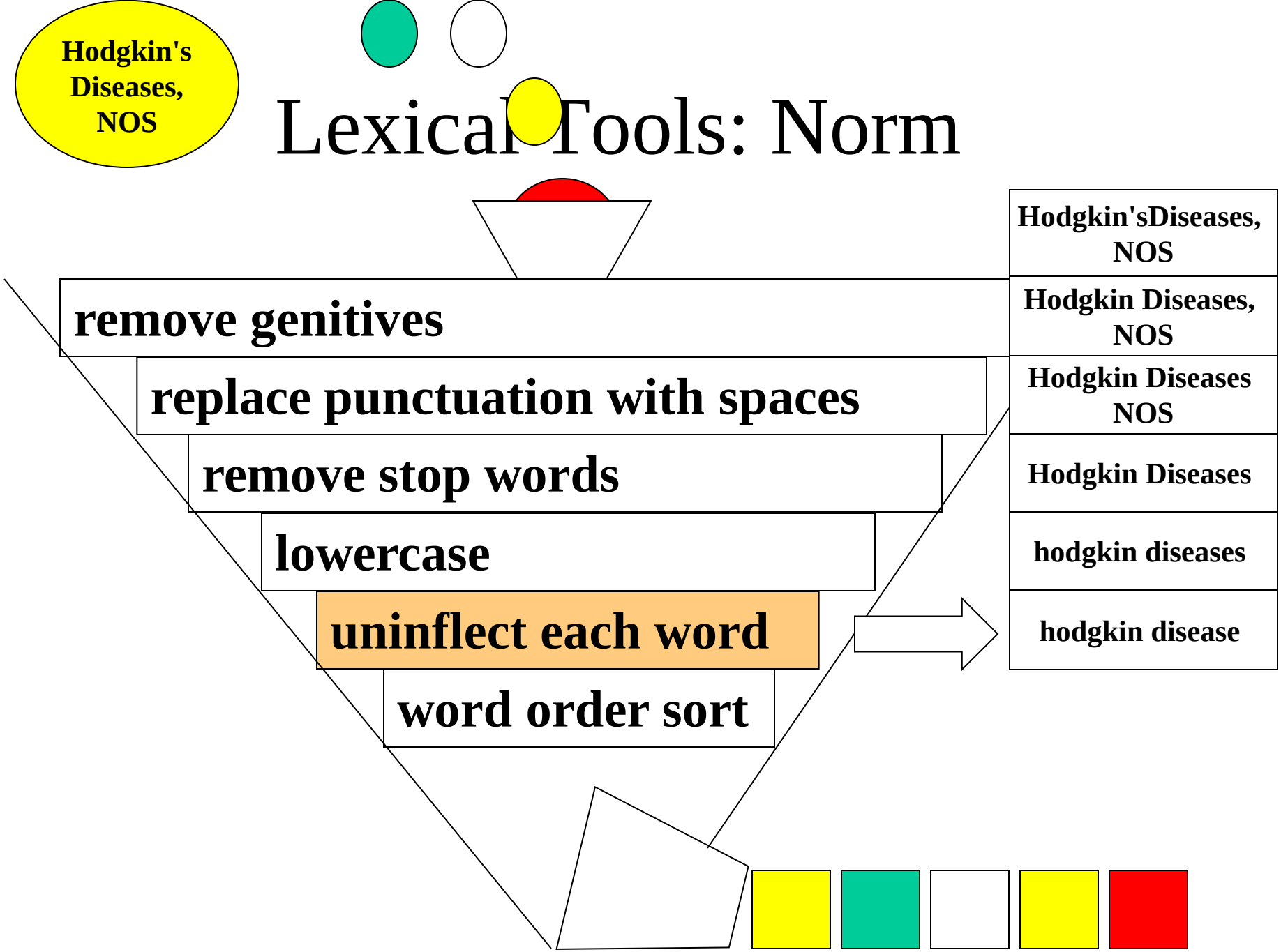
Hodgkin's Diseases, NOS

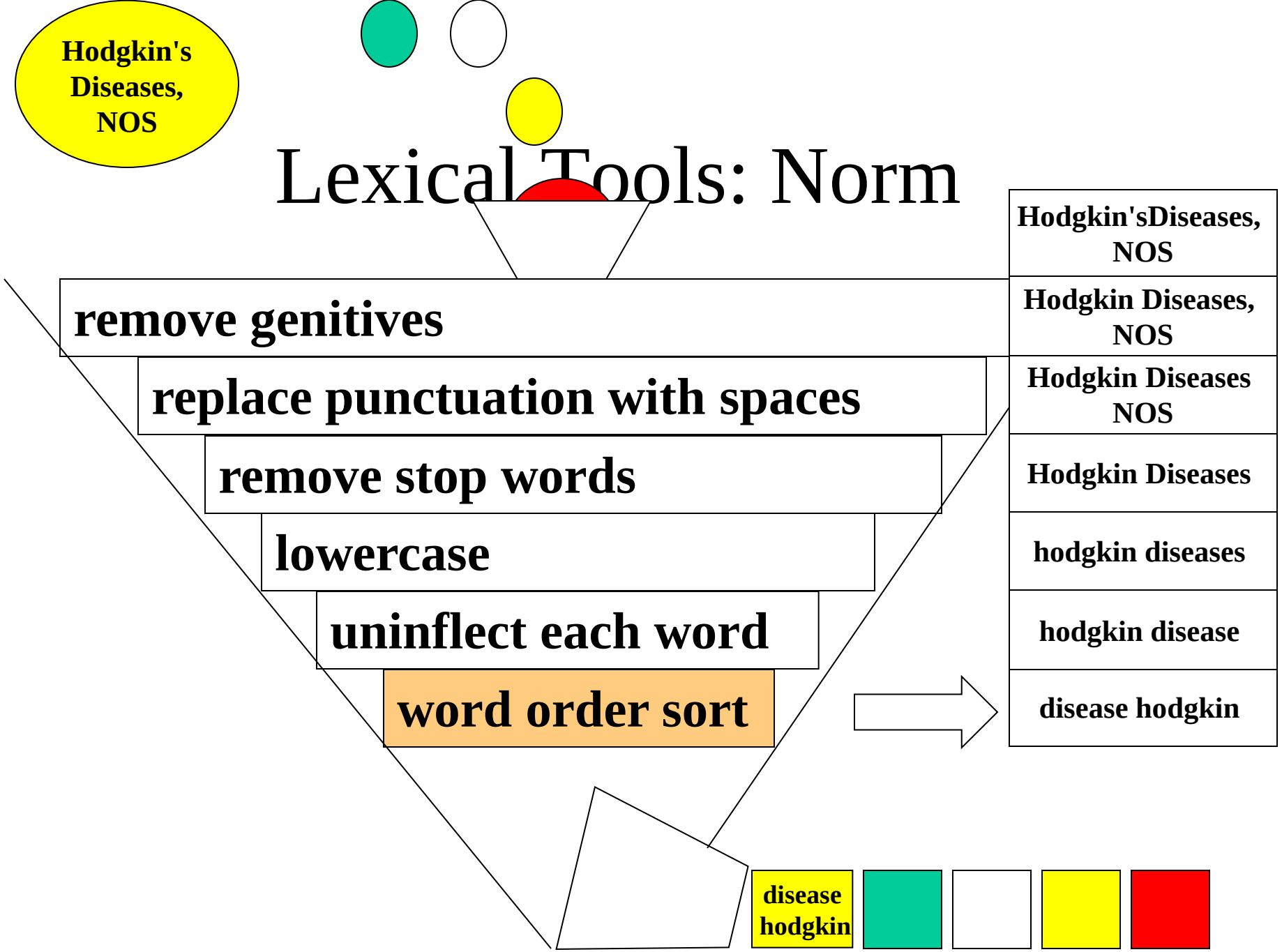
Lexical Tools: Norm



Hodgkin's Diseases, NOS

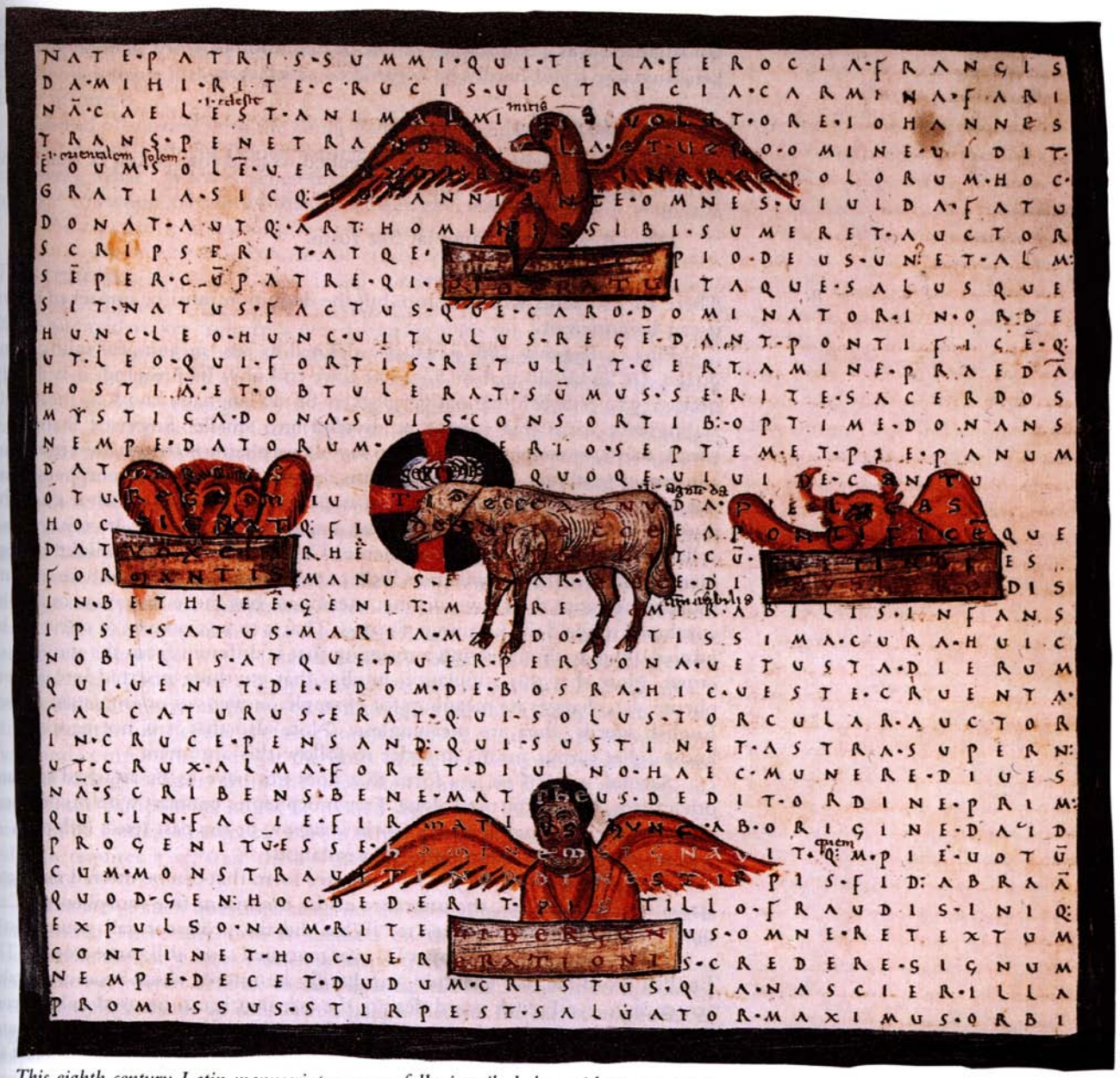
Lexical Tools: Norm




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 acetolyses	**Norm to**	acetolysis
Lung cancer Cancer, lung		cancer lung



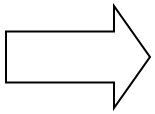
This eighth-century Latin manuscript was carefully inscribed, but without spaces to mark word boundaries.

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 it's 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

-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	**Return known acronyms**
a	**Return known acronym expansions**
B	**Uninflect words in a term**
b	**Uninflect a term**
C	**Canonicalize**
c	**Tokenize a term into "words"**
ca	**Tokenize, keep everything**
ch	**Tokenize without breaking hyphens**

Lexical Tools: Flow Components

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

Lexical Tools: Flow Components

Mnemonic	Tool
Gn	**Generate known fruitful variants**
g	**Remove genitive**
i	**Generate inflectional variants**
ici~Cats+Infls	**Generate inflectional variants, by Categories and inflections**
is	**Generate inflectional variants, by Categories and inflections**
Si	**Inflections to simplified inflections**

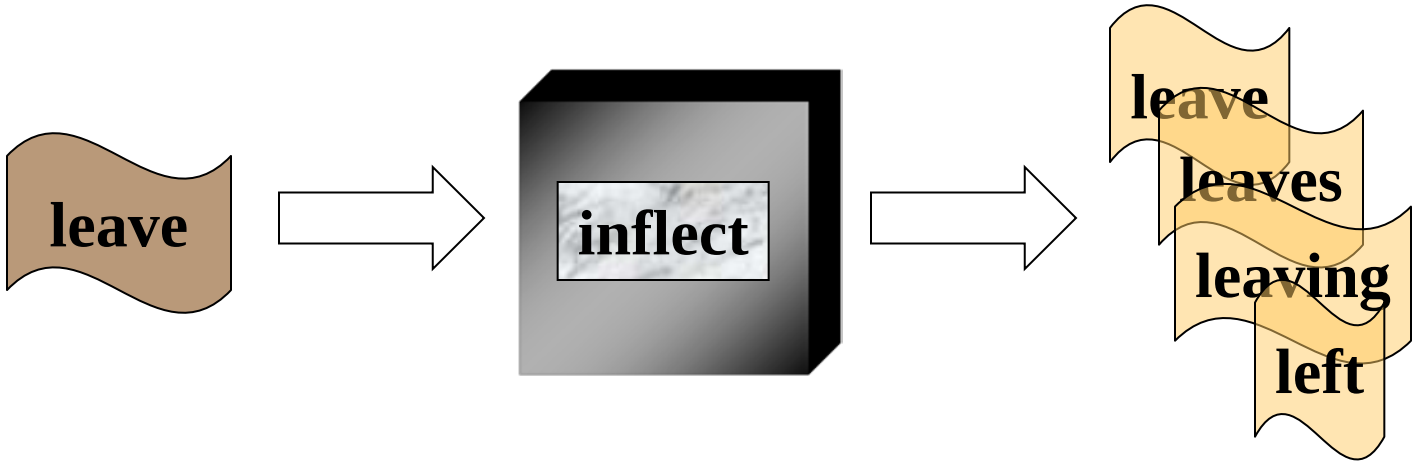
Lexical Tools: Flow Components

Mnemonic	Tool
L	**Retrieve category and inflection for a term**
l	**Lowercase**
N	**Normalize the input text in a non-canonical way (Norm)**
o	**Replace punctuations with spaces**
p	**Strip Punctuation**
q	**Strip diacritics**

Lexical Tools: Flow Components

Mnemonic	Tool
s	**Generate known spelling variants**
t	**Strip stop words**
u	**Uninvert the input phrase around commas**
w	**Sort words by order**
y	**Generate synonyms**
r	**Generate synonyms, recursively**

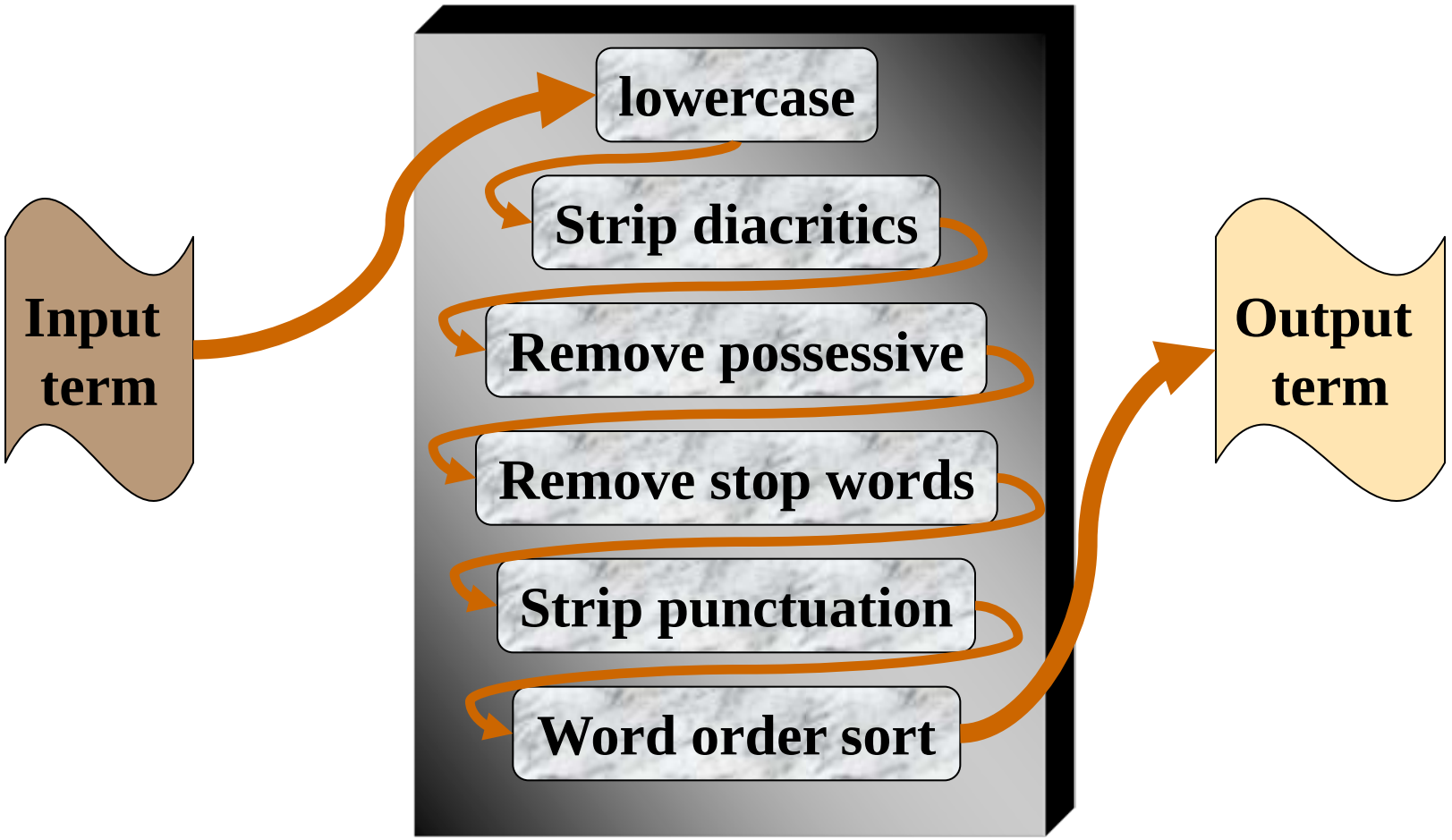
Lexical Tools: Flows



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|

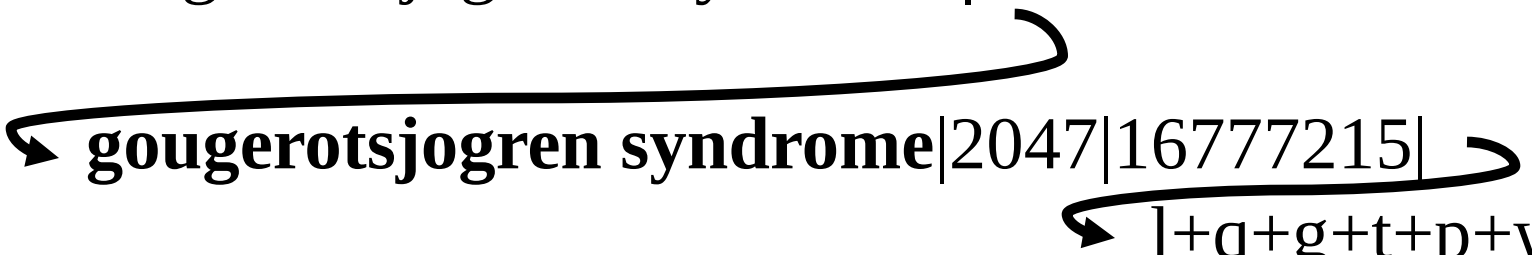
Lexical Tools: A Serial Flow



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

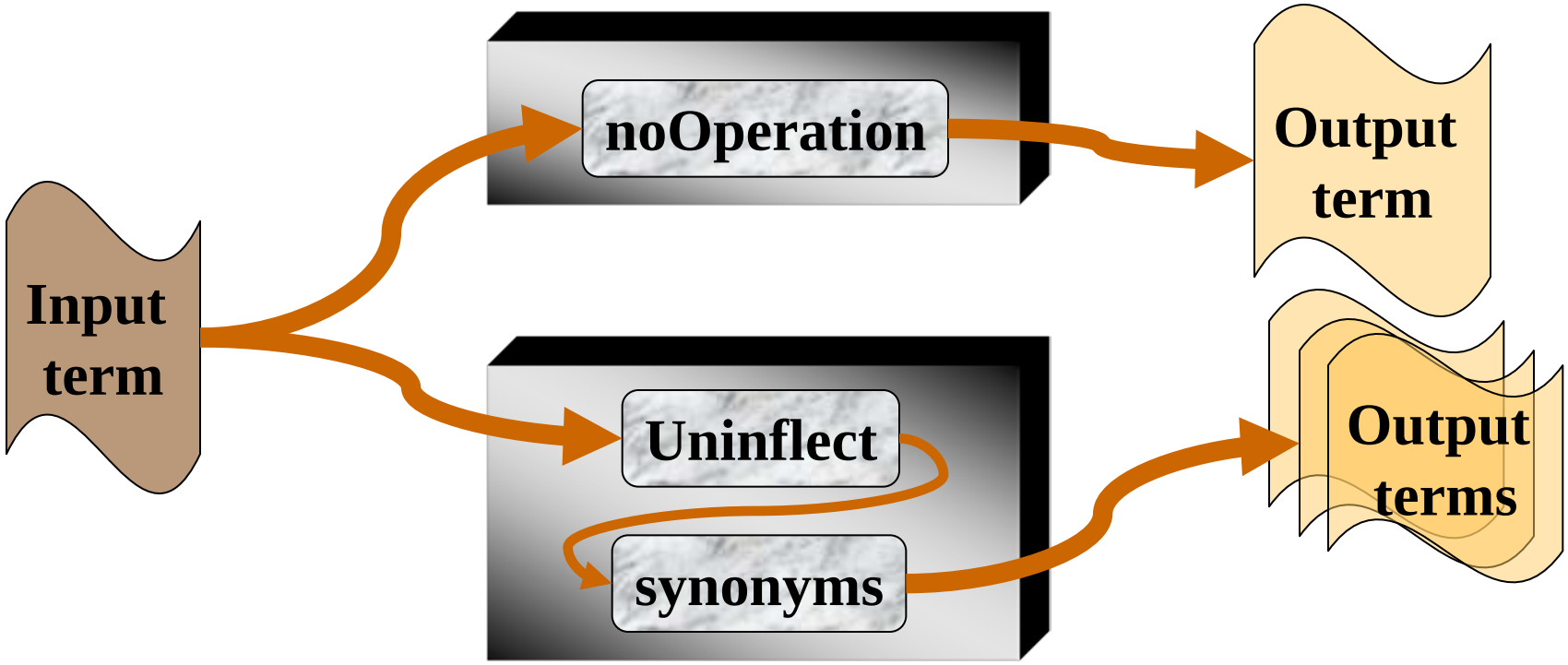
Lexical Tools: A Serial Flow

> lvg –f:l:q:g:t:p:w
The Gougerot-Sjögren's Syndrome
The Gougerot-Sjögren's Syndrome|



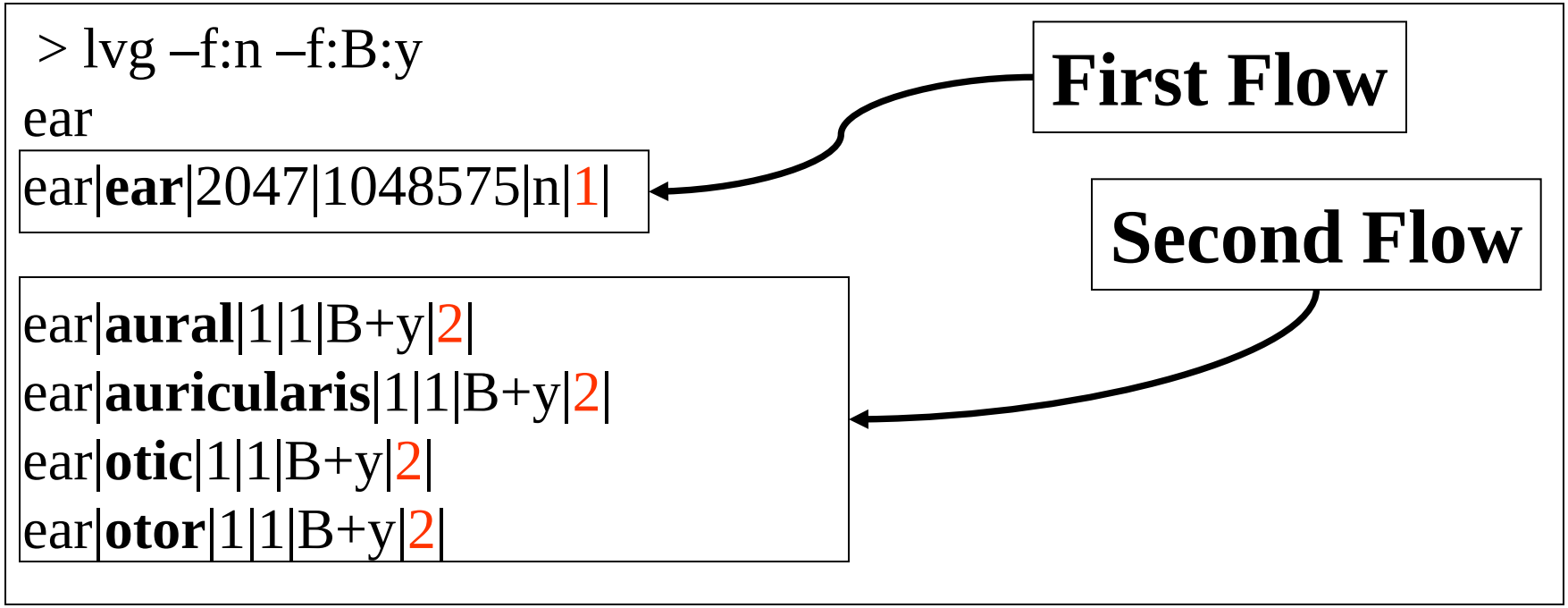
gougerotsjogren syndrome|2047|16777215|
l+q+g+t+p+w|1|

Lexical Tools: Parallel Flows



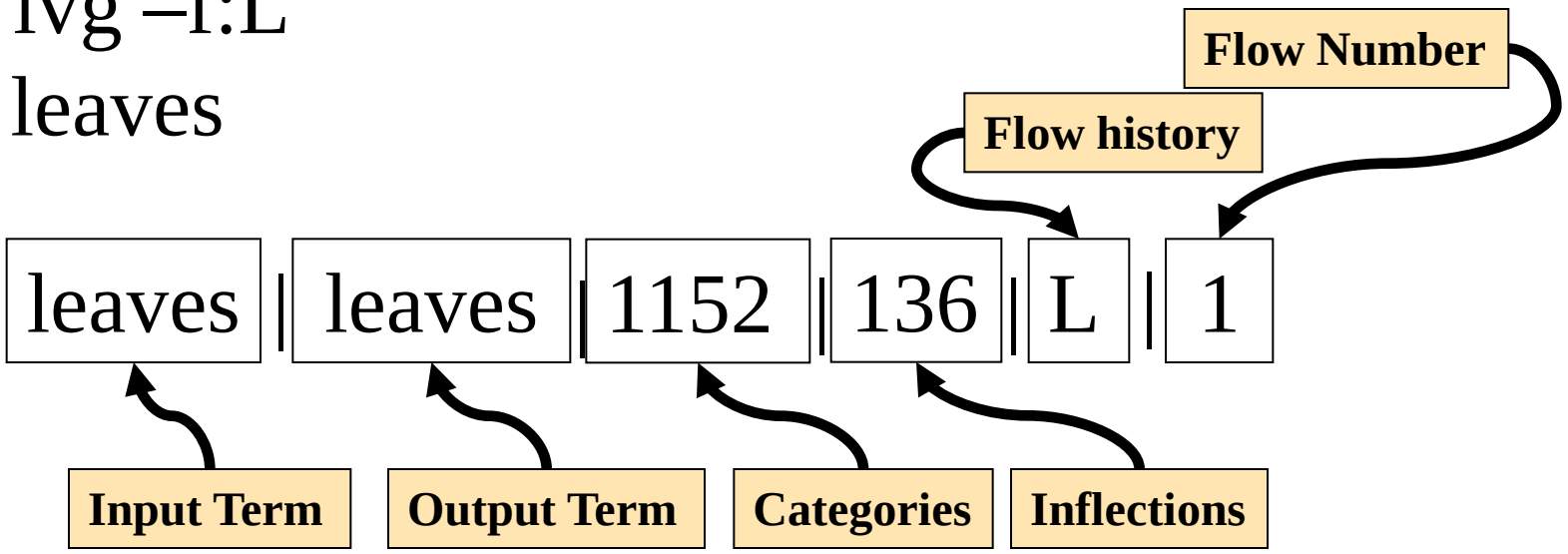
Multiple flows can be defined

Lexical Tools: Parallel Flows

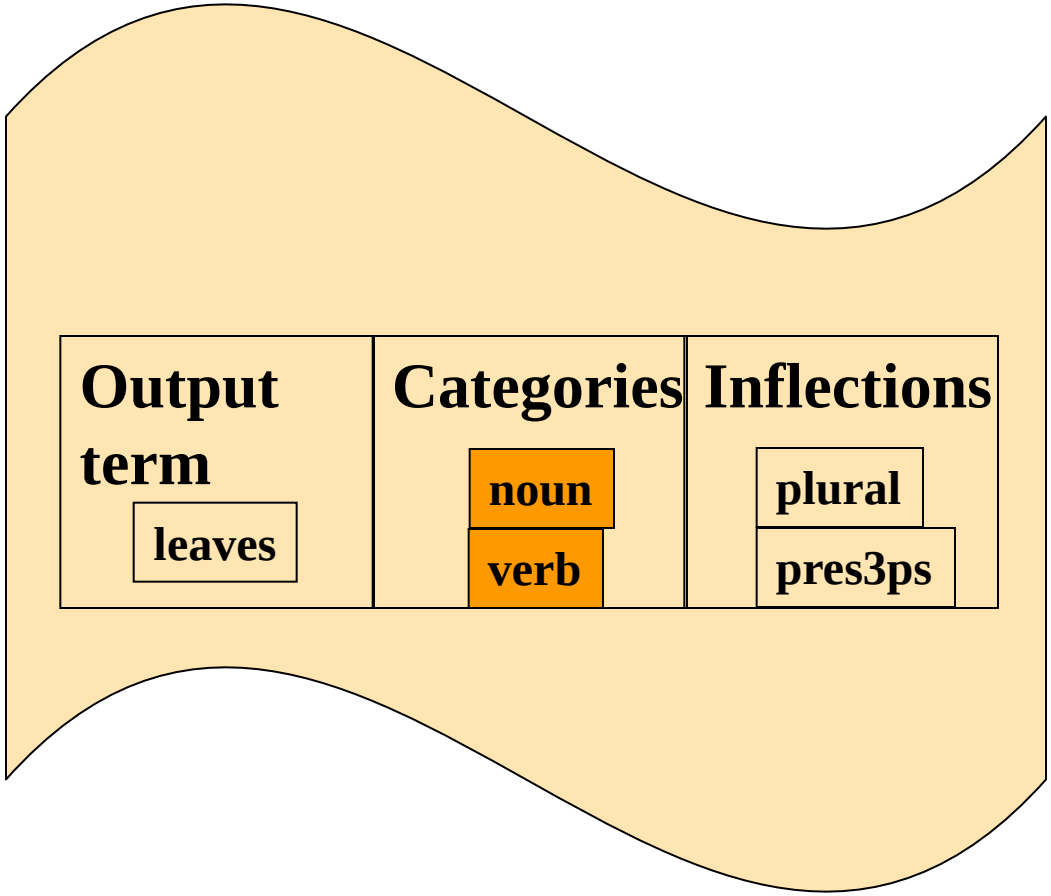


Lexical Tools: Fielded Output

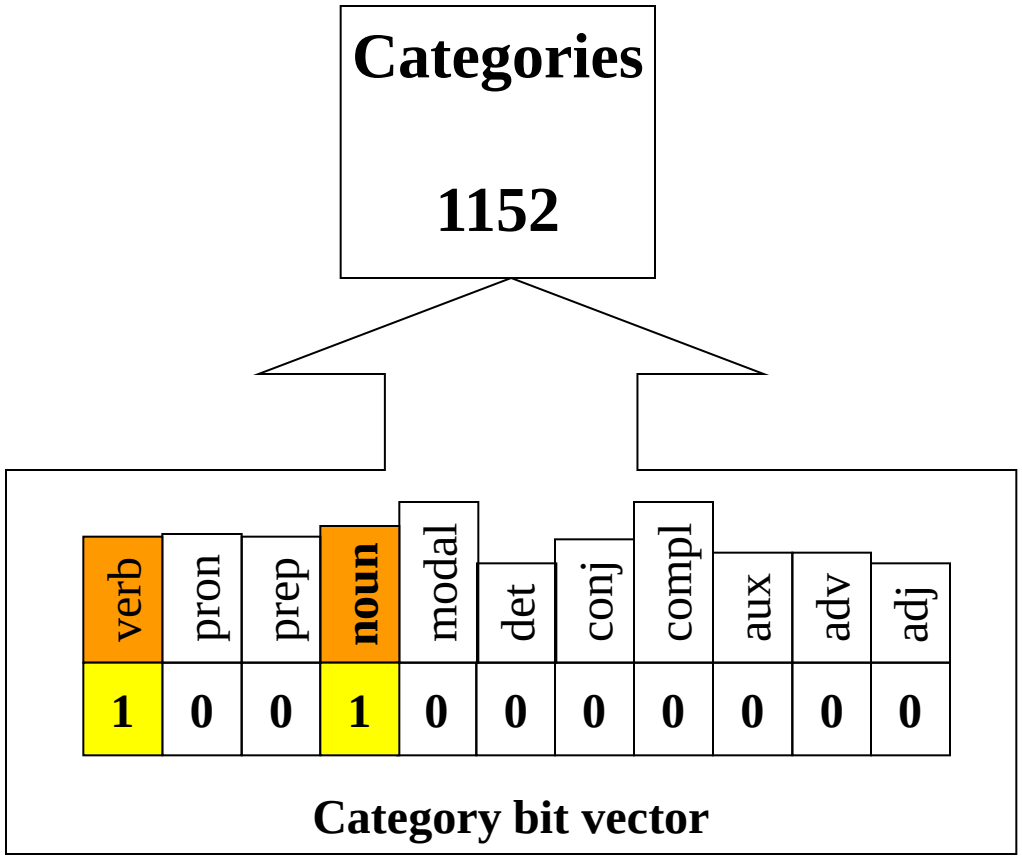
> lvg –f:L
leaves



Lexical Tools: Fielded Output



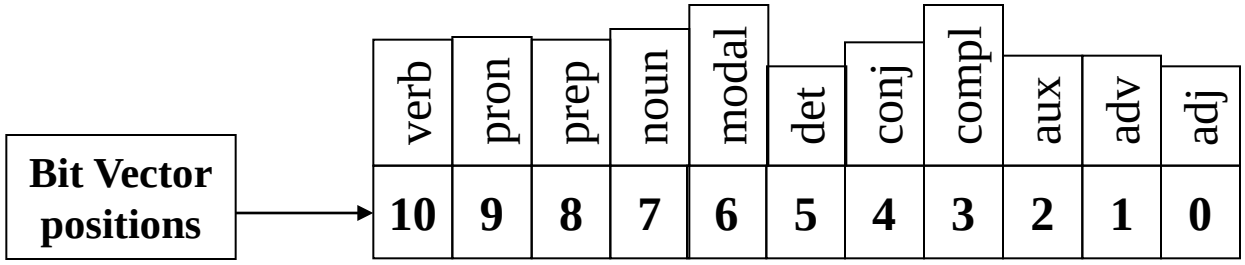
Lexical Tools: Categories



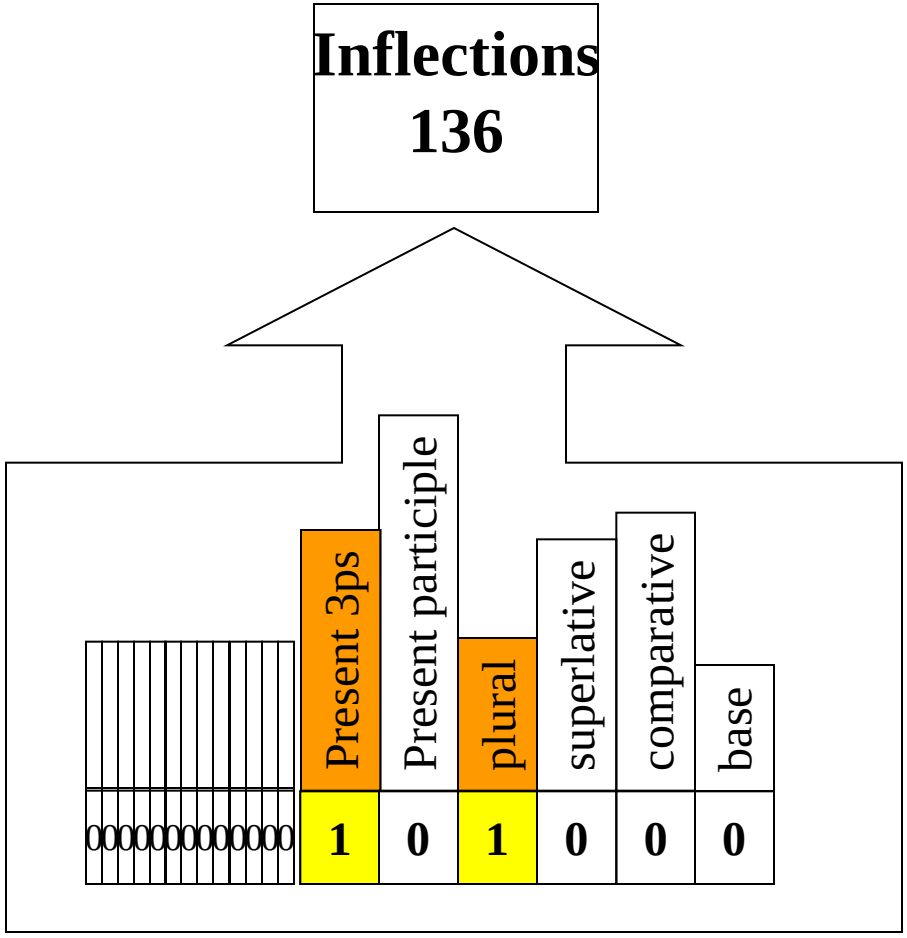
Lexical Tools: Categories

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

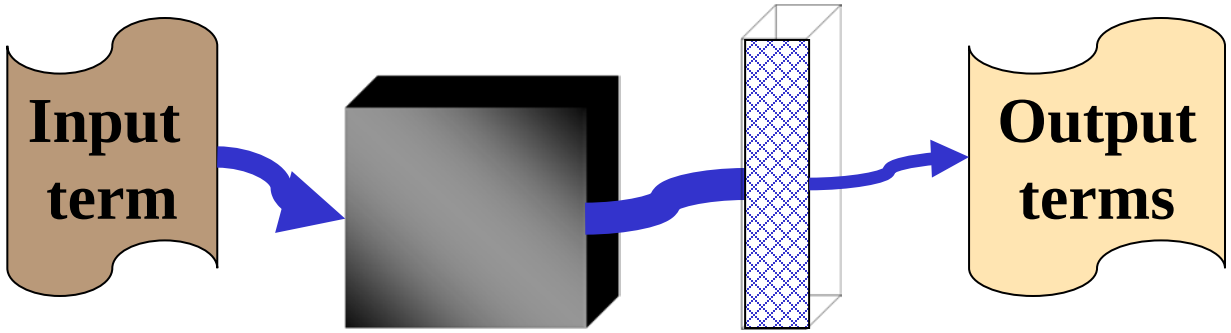
Base	1
Comparative	2
Superlative	4
Plural	8
Present Participle	16
Past	32
Past Participle	64
Present 3rd Person Singular	128

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



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

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|

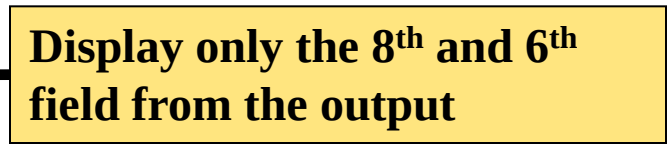
Lexical Tools: Post Flow Options

Mark the end of the set of variants returned


```
> lvg -f:L -ccgi
behavior
behavior|behavior|128|513|L|1|
__THE_END__
```

Lexical Tools: Post Flow Options

Specify fields for outputs



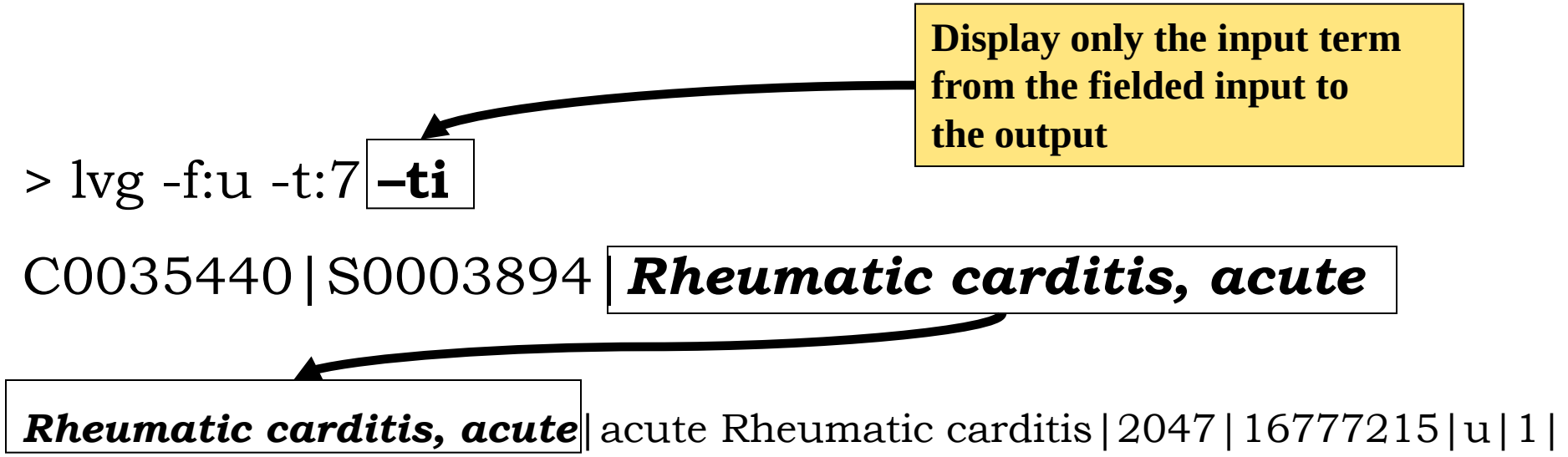
> lvg -f:u -t:7 **-F:8:6**

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

acute Rheumatic carditis|S0003894

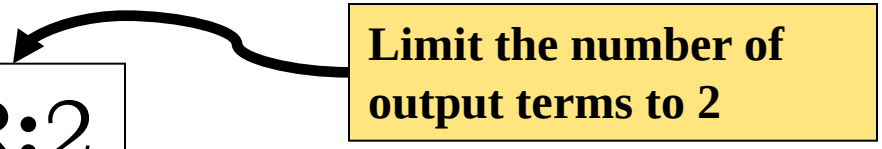
Lexical Tools: Post Flow Options

Display only the input term field when using fielded input



Lexical Tools: Post Flow Options

Restrict the number of variants returned

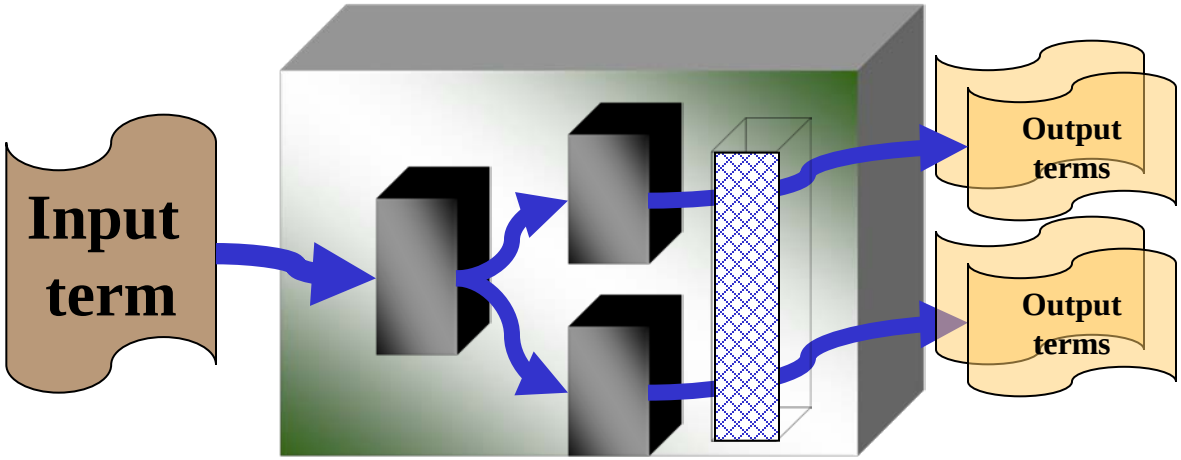



```
> lvg -f:i -R:2
foo
foo|foo|128|1|i|1|
foo|foos|128|8|i|1|
```

Note: Dangerous!
Do not try this at home!

Note: The unrestricted output would have produced 12 rows otherwise

Lexical Tools: Global Behaviors



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

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