The SPECIALIST NLP Tools

Dr. Chris J. Lu

The Lexical Systems Group

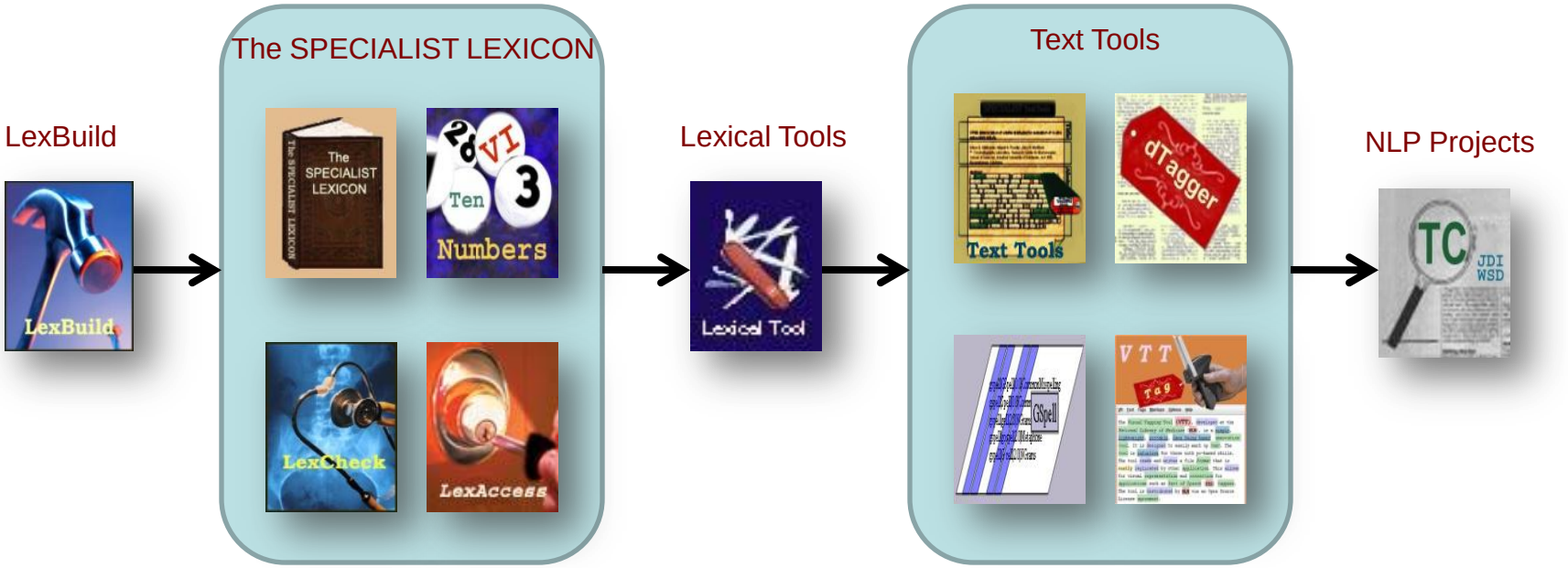
NLM. LHNCBC. CGSB

Nov., 2011

NLP Tools

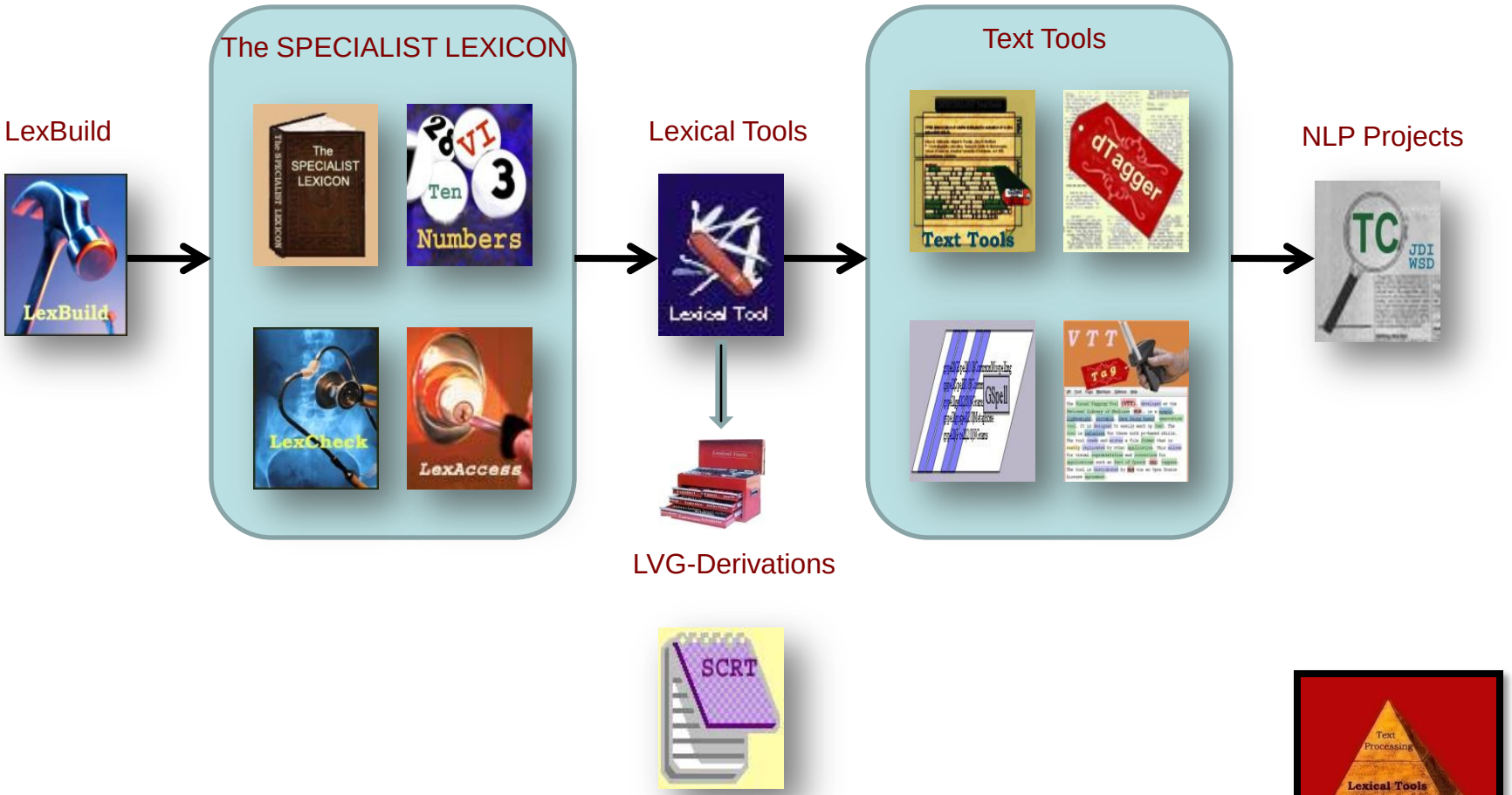
- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov

The SPECIALIST NLP Tools



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

The SPECIALIST NLP Tools



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

Table of Contents

Lexicon To Lexical Tools

- How to use Lexicon (lexical records)?
- Lexical record(s) / Lexicon:
  - Text
  - Tables
- LexCheck package
  - XML
  - APIs: Java object(s)

Example: Spelling Variant

- Example: color & colour
- Lexical record in text & Java API:

```
{base=color
spelling_variant=colour
entry=E0017902
    cat=noun
    variants=uncount
    variants=reg
}
```

Step 1: Convert to Java Object
- LexRecord ToJavaObjFromText(String text)
- Vector<LexRecord> ToJavaObjsFromText(String text)
- Vector<LexRecord> ToJavaObjsFromTextFile(String inFile)

…

Step 2: Retrieve information from LexRecord Java Object
- Vector<String> GetSpellingVars()
- String GetBase()
- String GetCategory()

…

- LRSPL table:

EUI	Spelling Variant	Base form
…	…	…
E0017902	colour	color
…	…	…

Example: Inflectional Variant

- Lexical record in text & Java API:

```
{base=color
spelling_variant=colour
entry=E0017902
        cat=noun
        variants=uncount
        variants=reg
}
```

Step 1: Convert to Java Object

- Vector<LexRecord> ToJavaObjsFromText(String text)
- Vector<LexRecord> ToJavaObjsFromTextFile(String inFile)

…

Step 2: Retrieve information from LexRecord Java Object

- InflVarsAndAgreements GetInflVarsAndAgreements()
  - Vector<InflVar> GetInflValues()
    - String GetInflection()
- String GetBase()
- String GetCategory()

…

- LRAGR table:

EUI	Infl Var	Category	Agreement	Citation Form	Base Form
…	…		…		
E0017902	color	noun	count(sing)	color	color
E0017902	color	noun	uncount(sing)	color	color
E0017902	colors	noun	count(plur)	color	color
E0017902	colour	noun	count(sing)	color	colour
E0017902	colour	noun	uncount(sing)	color	colour
E0017902	colours	noun	count(plur)	color	colour
…	…		…		

Lexical Variations (Lexicon)

- Spelling variants (-f:s):
  - color|colour (noun|E0017902, verb|E0017903)
  - grey|gray (adj|E0030394, noun|E0030395, verb|E0030396)
  - heart burn|heart-burn|heartburn (noun|E0030961)
  - hemostasis|haemostasis (noun|E0030684)
- Inflectional/uninflectional variants (-f:I, -f:b, -f:B):
  - heart burn (noun|E0030961)
  - color|colors (noun|E0017902)
  - color|colored|colored|colors|coloring (verb|E0017903)
  - see|sees|saw|seen|seeing (verb|E0055007)
  - saw|saws|sawed|sawn|sawing (verb|E0054444)
- Acronyms/abbreviations, expansions (-f:a, -f:A, -f:fa):
  - ER|emergency room
  - ER|enhancement ratio
  - ER|eye research
  - 20+ known acronyms …
- Nominalization (-f:nom):
  - active|adj|activity|noun
  - active|adj|activeness|noun
- ProperNoun (-f:fp):
  - Clinton
  - Virginia
  - University of Virginia

Basic Variations

- Lowercase (-f:l):
  - AIDS|aids
  - ÀÁÂÃÄÅ ÈÉÊË ÌÍÎÏ ÒÓÔÕÖ Ø ÙÚÛÜ|àáâãäå èéêë ìíîï òóôõö ø ùúûü
- Strip punctuation (-f:o, -f:p, -f:P)
  - St. John's|St Johns
- Strip stopwords (-f:t)
  - Remove "of", "and", "with", "for", "nos", "to", "in", "by", "on", "the", "(non mesh)", etc.
  - Academy of Physical Medicine|Academy Physical Medicine
- Remove genitive (-f:g):
  - Down's Syndrome|Down Syndrome
- Remove parenthetic plural form of (s), (es), (ies) (-f:rs)
  - Burn(s);skin|Burn;skin
  - 9(s)-erythromycylamine|9(s)-erythromycylamine
- Strip ambiguity tags (-f:T)
  - cold <1>|cold
- Sort words (-f:w)
  - Cancer, Lung|Cancer Lung
  - Lung Cancer|Cancer Lung
- Word size filter (-f:ws)
  - Academy of Physical Medicine|Academy Physical Medicine
- …

Others Variations

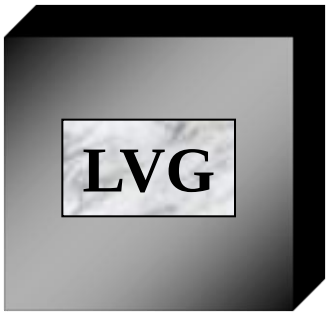
- Derivational variants (-f:d, -f:R):
  - gene|noun|genetic|adj
  - gene|noun|genic|adj
  - hyperuricemic|adj|hyperuricemia|noun
  - hyperplastic|adj|hyperplasia|noun
- Synonyms (-f:y):
  - otitis|ear inflammation (C0029877)
  - kidney|renal| (C0022646)
  - eartburn|pyrosis|brash (C0018834)
- Canonical Form (-f:C): used in LuiNorm for Lui assignment
  - color|color
  - colour|color
  - colored|color
  - coloued|color
- …

Complicated Variations

- ASCII Conversion (-f:q, -f:q0, -f:q1, … -f:q8):
  - resumé|resume
  - spælsau|spaelsau
  - ⅝|5/8
  - “Quote” |"Quote"
  - α-Best™|alpha-Best
  - …
- Norm (-f:N):
  - Hodgkin's diseases, NOS|disease hodgkin
  - proofread|proof read
  - proof-read|proof read
  - proof read|proof read
  - left|left
  - left|leave
- LuiNorm (-f:N3):
  - left|leaf
- AntiNorm (Use for Approximate match in Lexicon):
  - Abrami disease|Abrami's disease
  - Abrami disease|Abrami's diseases
- …

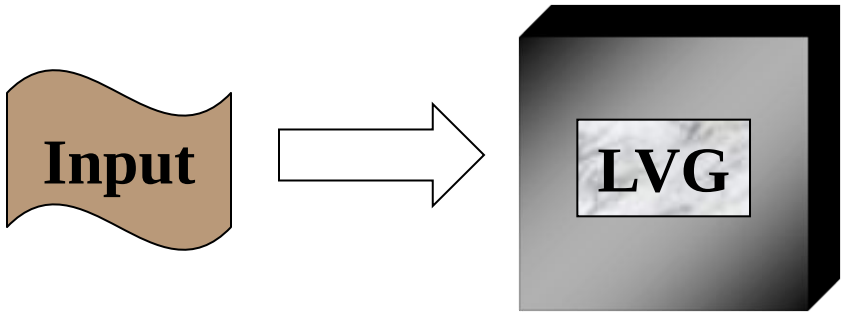
Lexical Variant Generation (LVG)

Lexical Tools - LVG



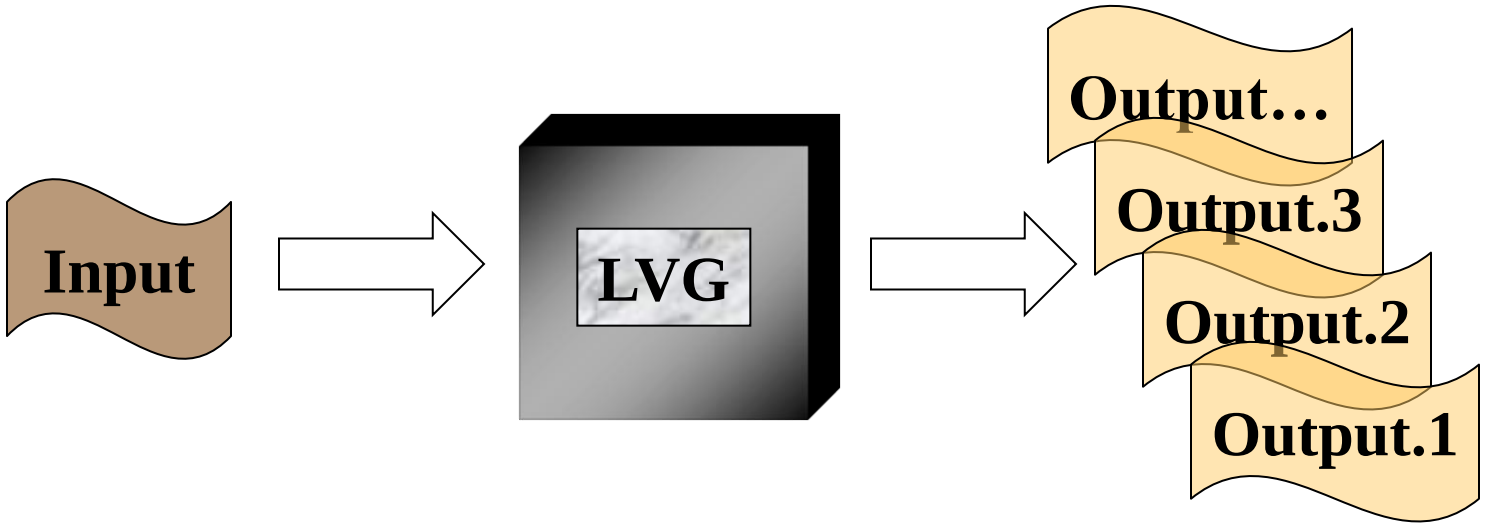
- A suite of text utilities

Lexical Tools - LVG



- A suite of text utilities take the given input

Lexical Tools - LVG

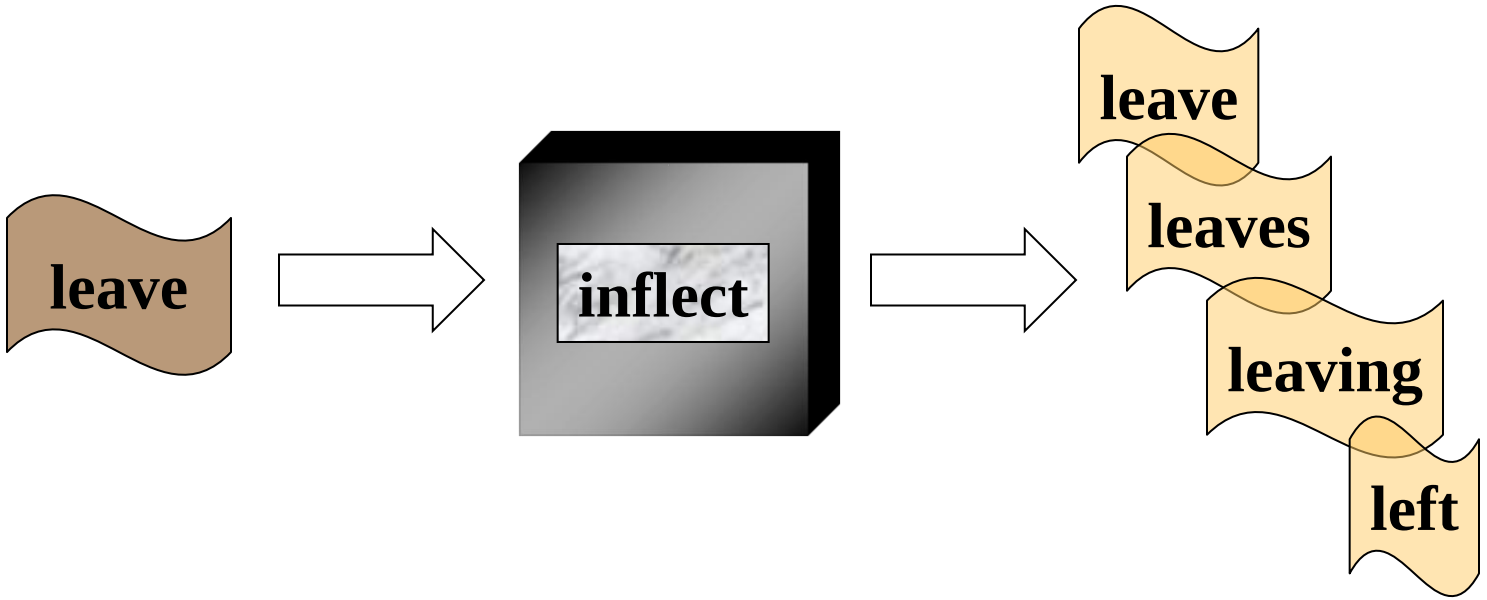


- A suite of text utilities that generate, mutate, and filter out lexical variants from the given input

LVG - 2012

- 62 flow components
- 37 options
  - input filter options (3)
  - global behavior options (12)
  - flow specific options (2)
  - output filter options (20)

Flow Components

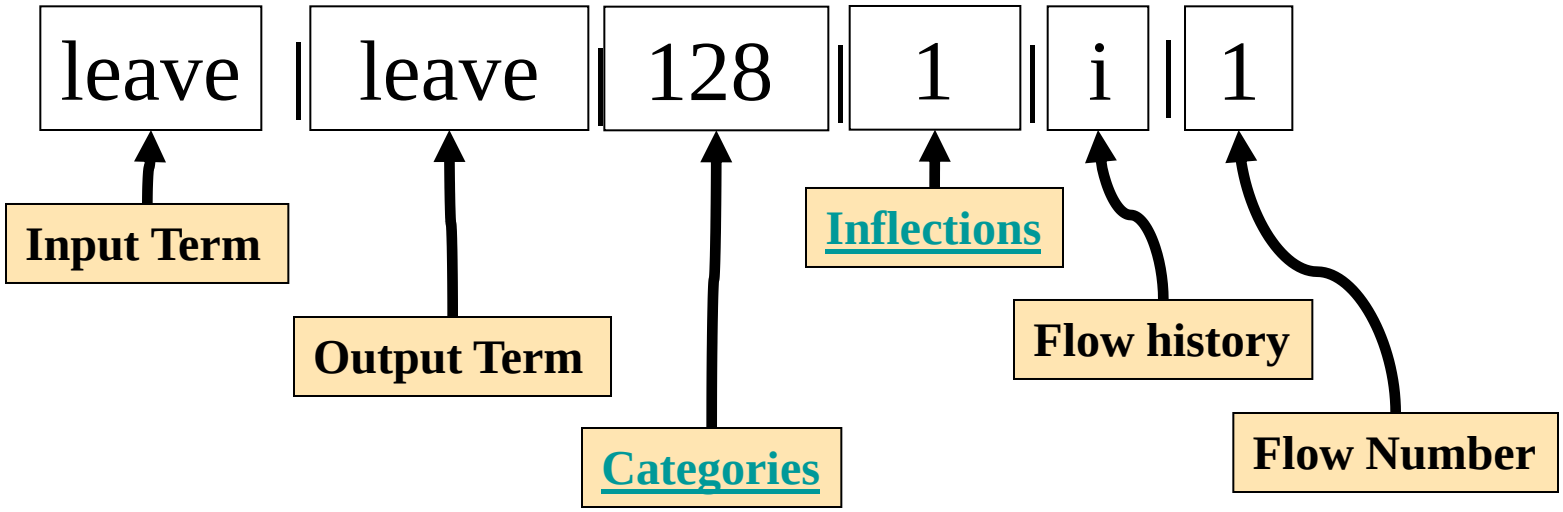


Command Line Tool

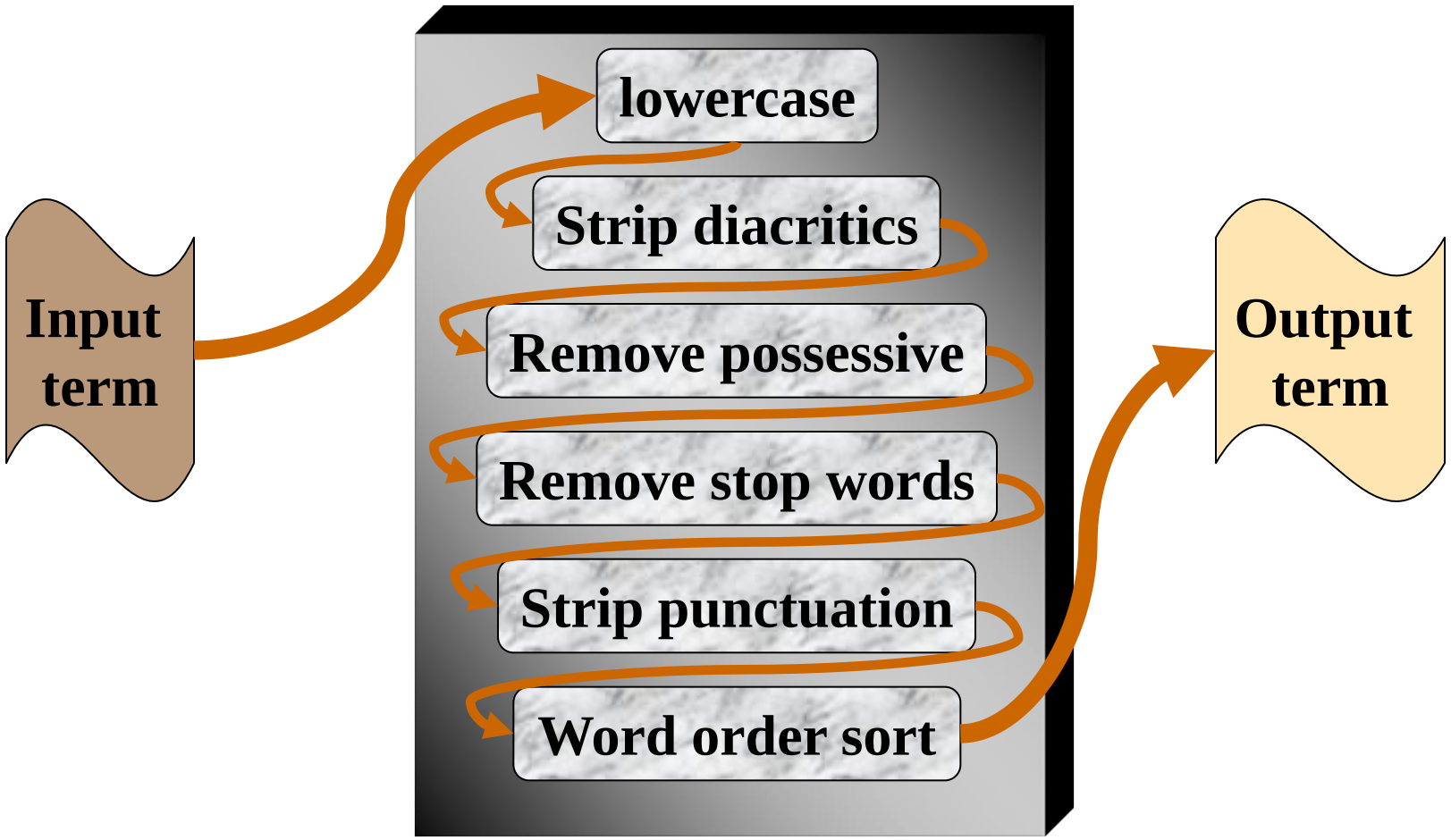
```
shell> lvg -f:i
leave
leave|leave|128|1|i|1|
leave|leave|128|512|i|1|
leave|leaves|128|8|i|1|
leave|left|1024|64|i|1|
leave|left|1024|32|i|1|
leave|leave|1024|1|i|1|
leave|leave|1024|262144|i|1|
leave|leave|1024|1024|i|1|
leave|leaves|1024|128|i|1|
leave|leaving|1024|16|i|1|
```

Fielded Output

> lvg –f:i
leave



A Serial Flow

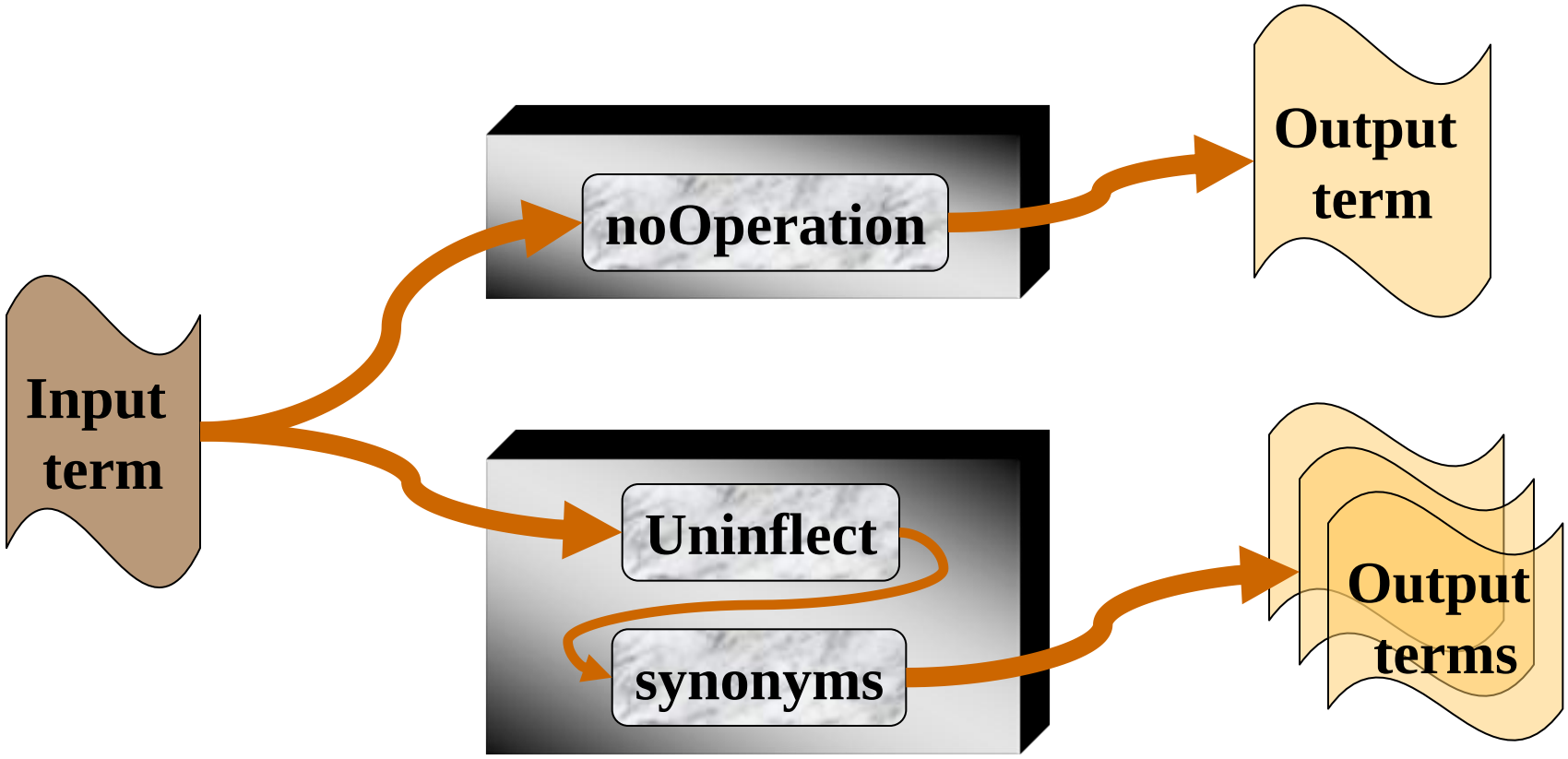


- Flow components can be arranged so that the output of one is the input to another.

A Serial Flow - Example

```
shell> 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



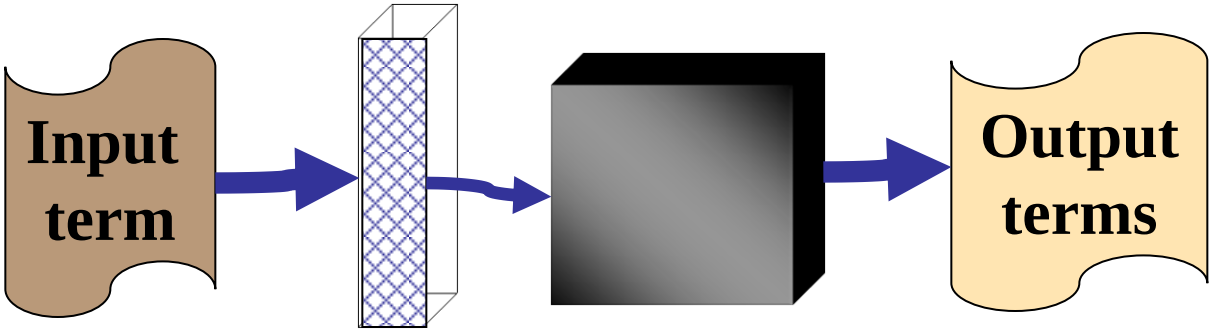
- Multiple flows can be defined

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



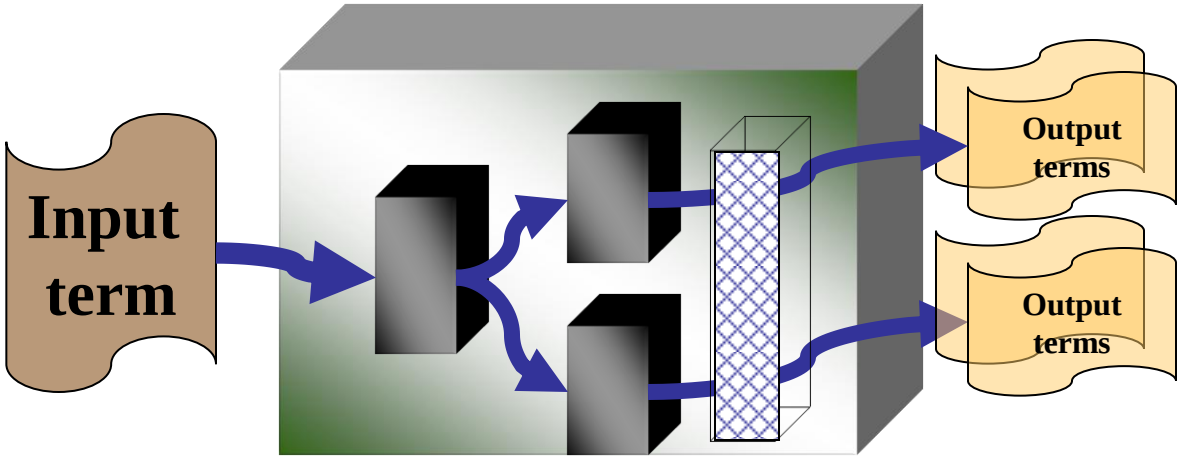
Take field 7 from the input

```
> 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:"\"
```

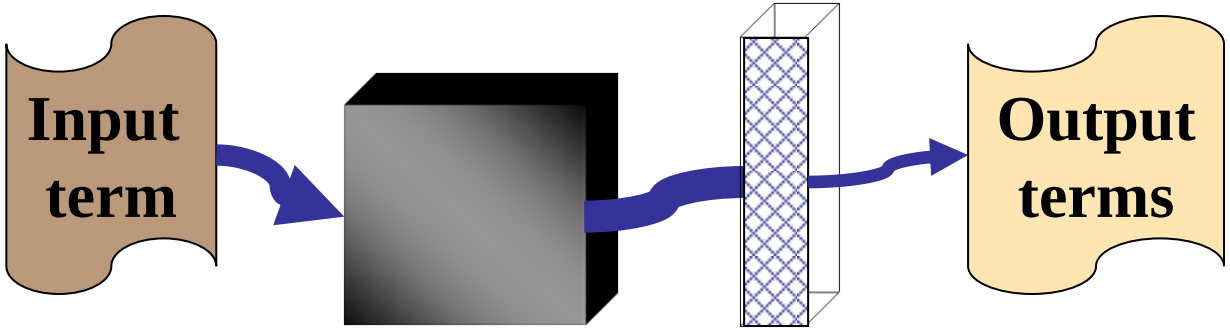
Change separator to "\"

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|

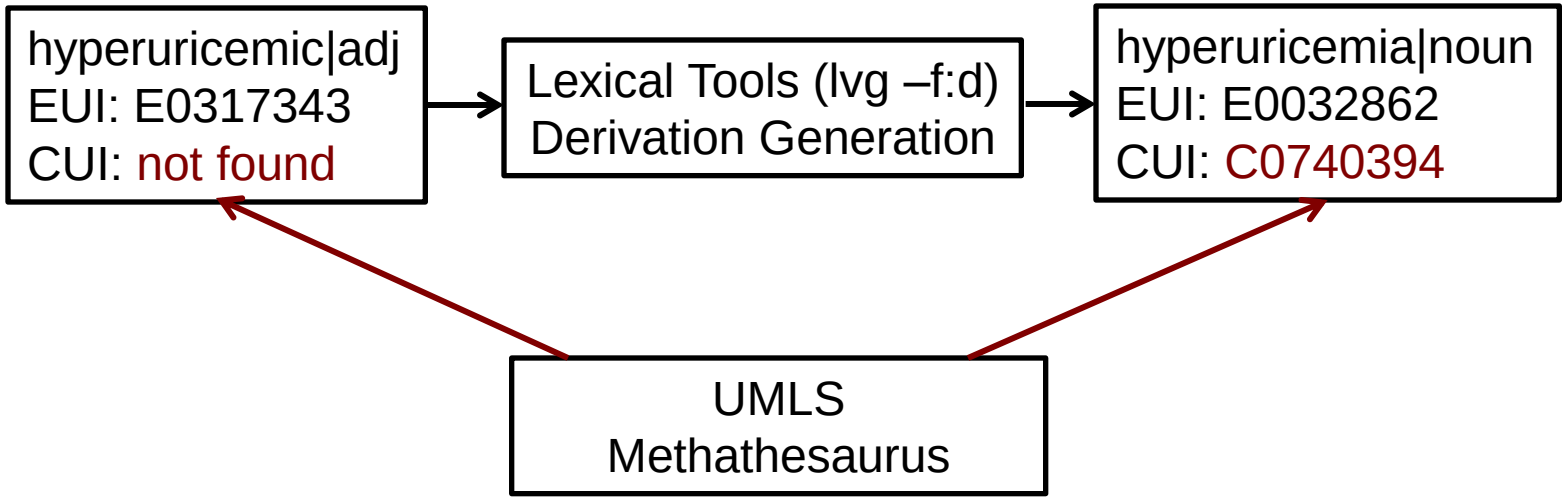
Derivational Variants

- Words related by a derivational process
  - Derivational process: suffix and prefix
  - Used to create new words based on existing words
  - Meaning change
  - Category change

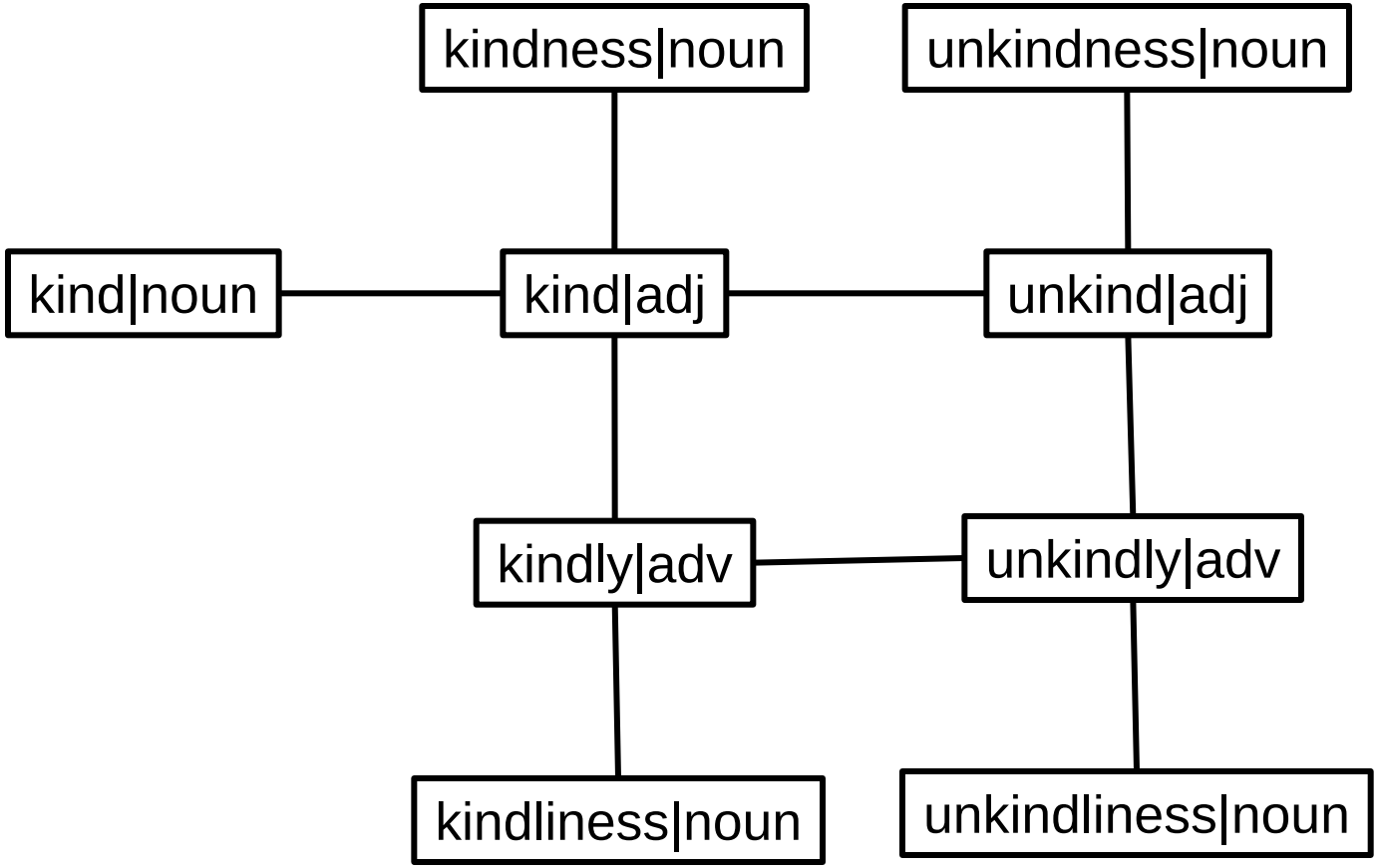
- Focus on relatedness (no direction)

Derivations Application

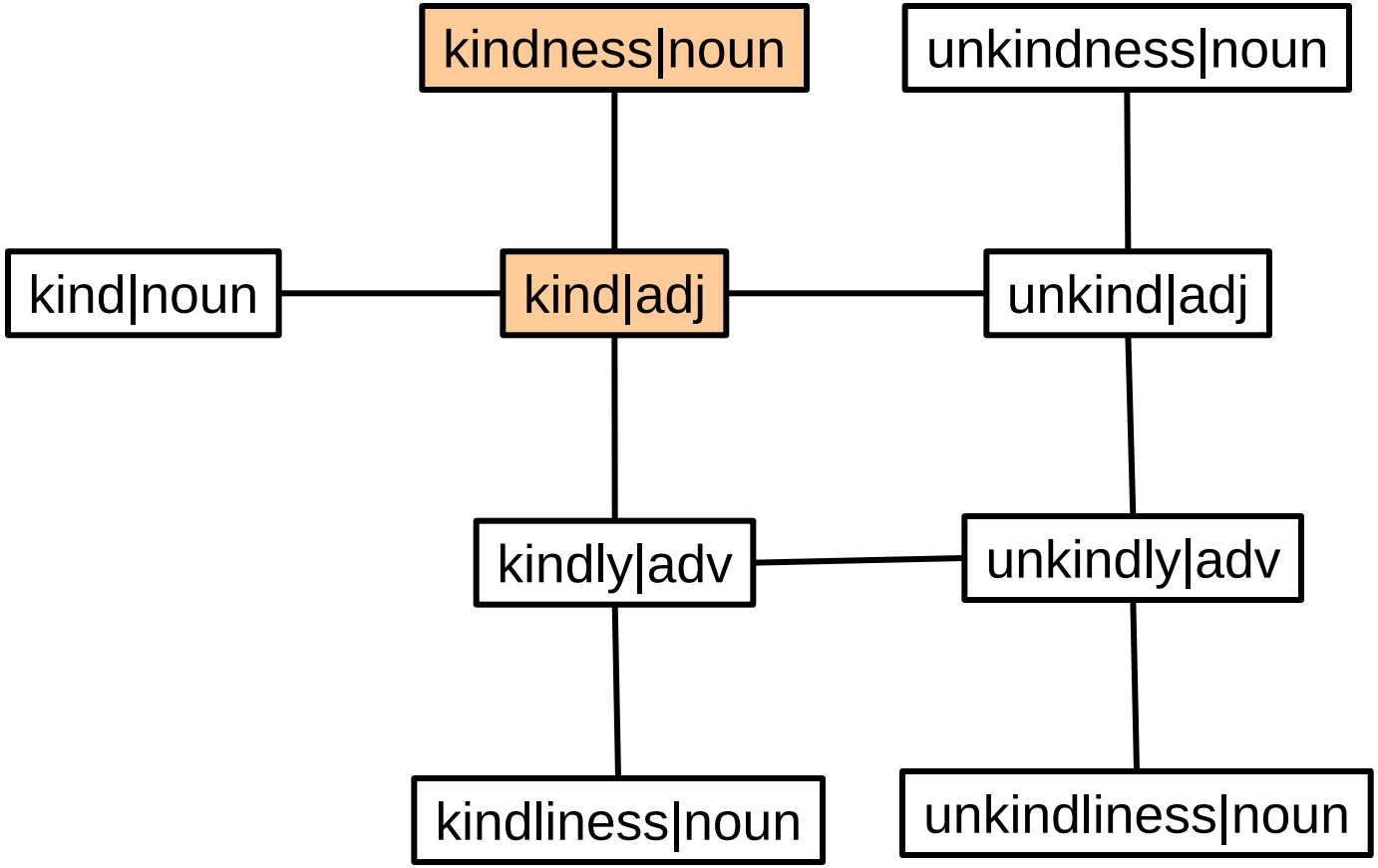
- hyperuricemic|adj, E0317343, no CUI
- hyperuricemia|noun, E0032862,
  is a UMLS Metathesaurus term (C0740394)



Derivational Network



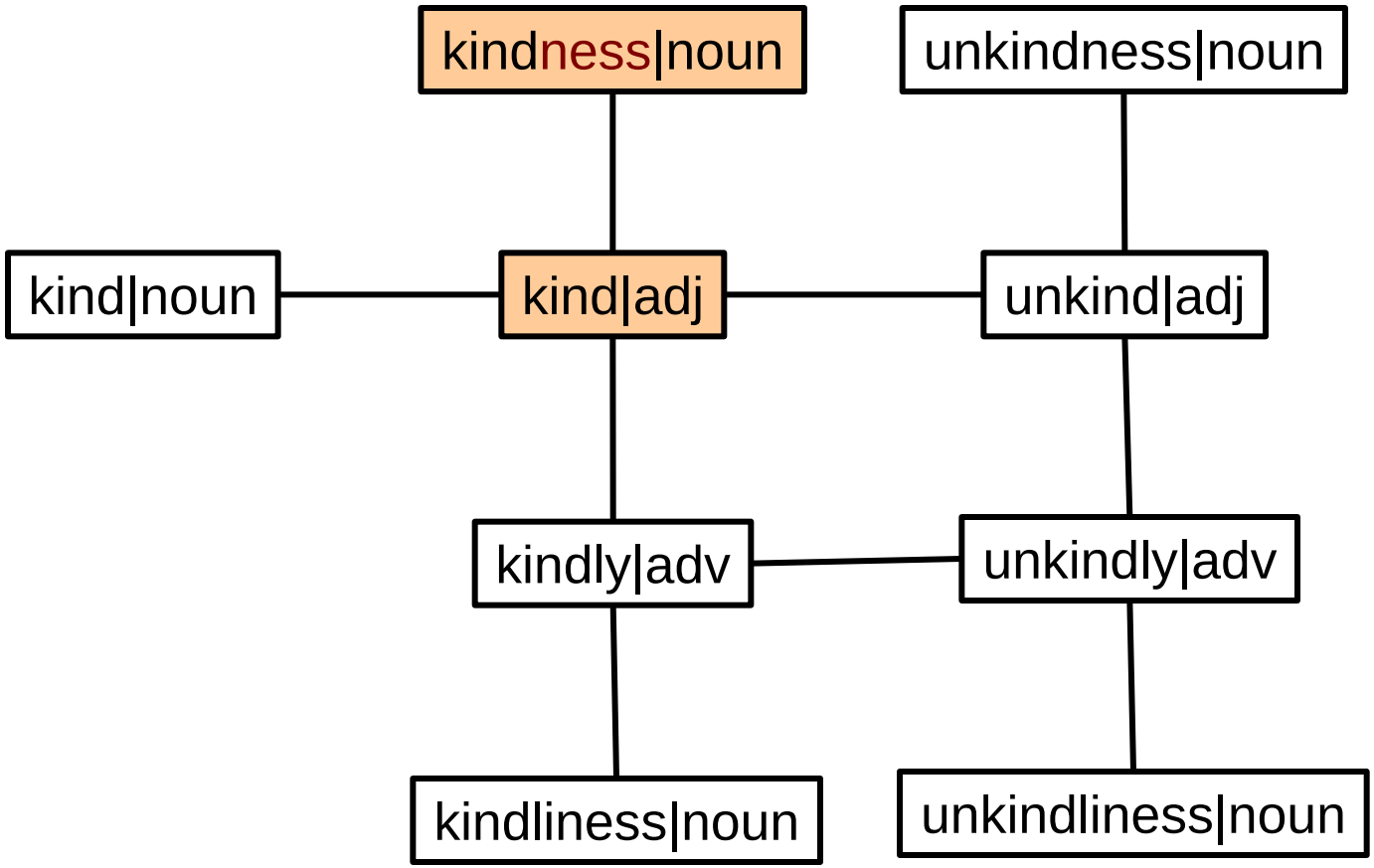
Derivational Pair



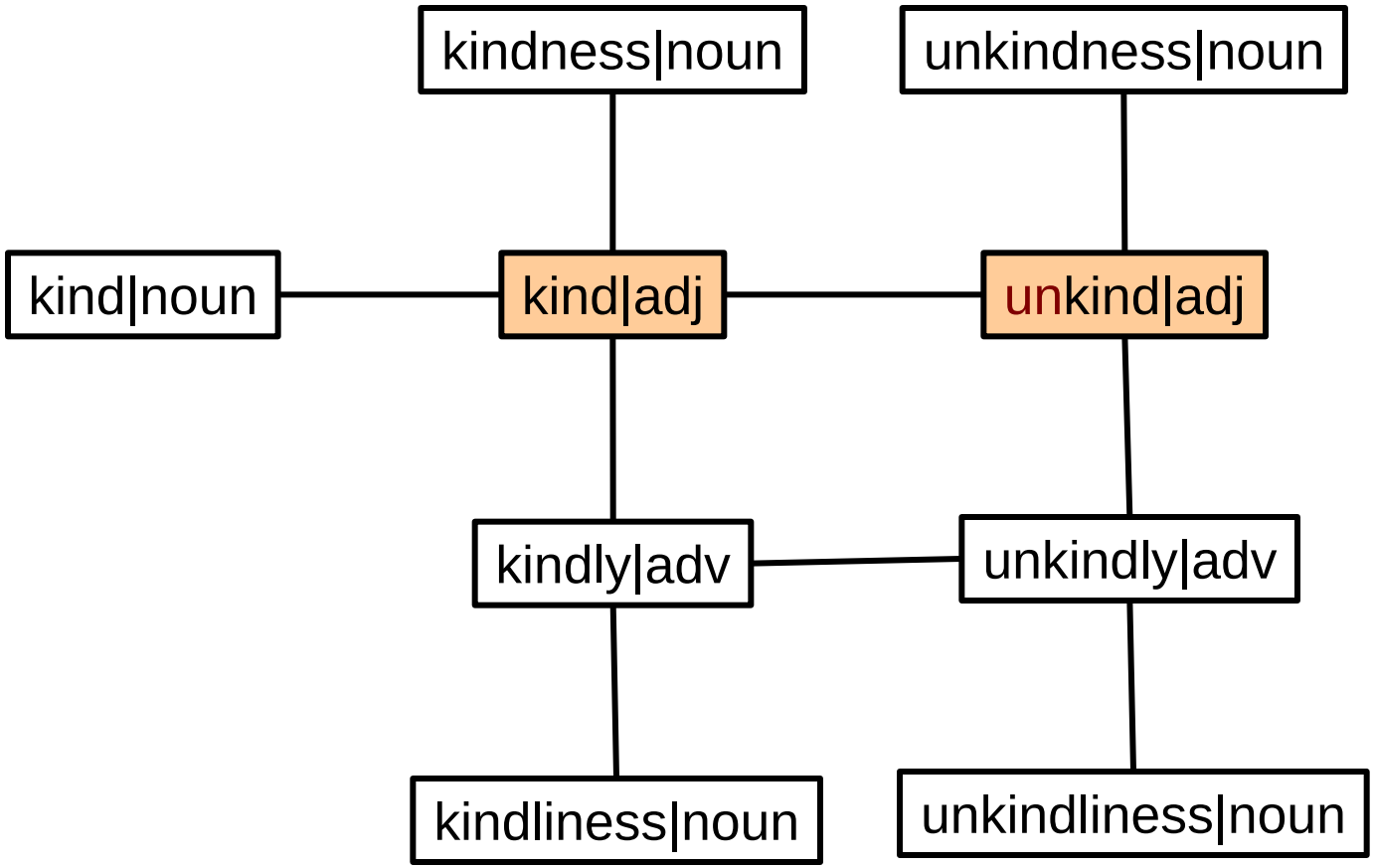
Derivational Pair

- Each link and the associated two nodes in derivational network define a derivational pair
- Includes base forms and syntactic category information
- Bi-directional
- Only involves one or none derivational affix
- Lvg format: base 1|category 1|base 2|category 2
- Examples:
  - kind|adj|kindness|noun
  - kind|adj|kindly|adv
  - kind|adj|unkind|adj
  - kind|adj|kind|noun

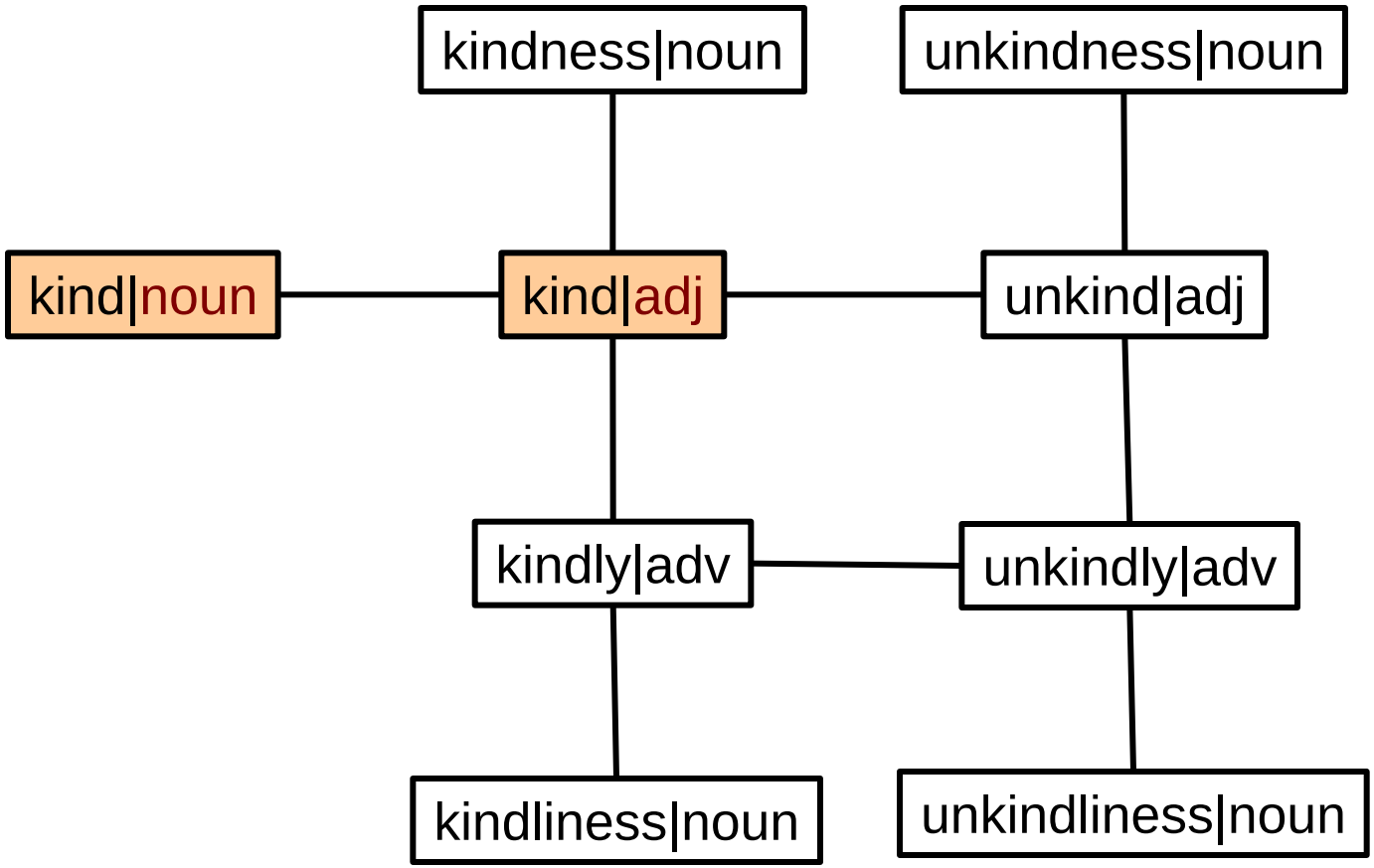
Suffix Derivation (SD) Pair



Prefix Derivation (PD) Pair

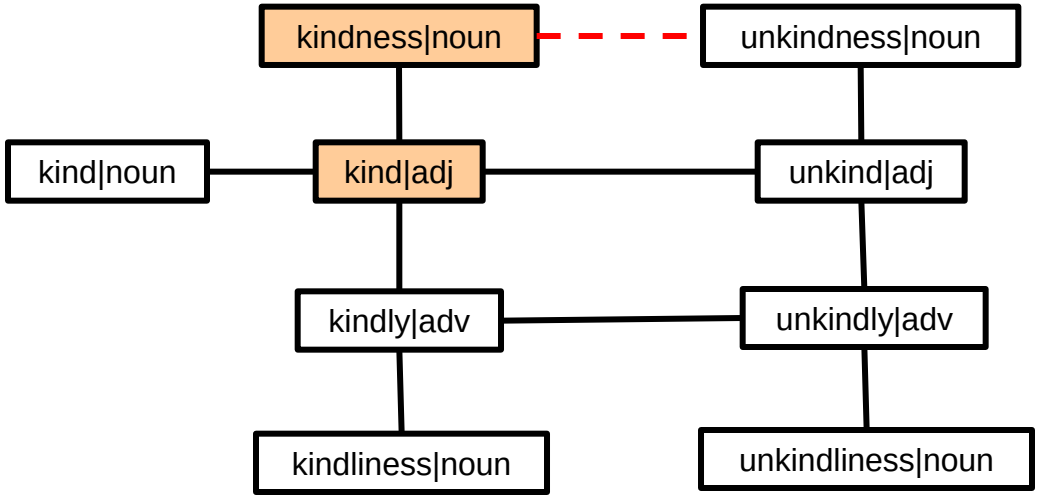


Zero Derivation (ZD) Pair



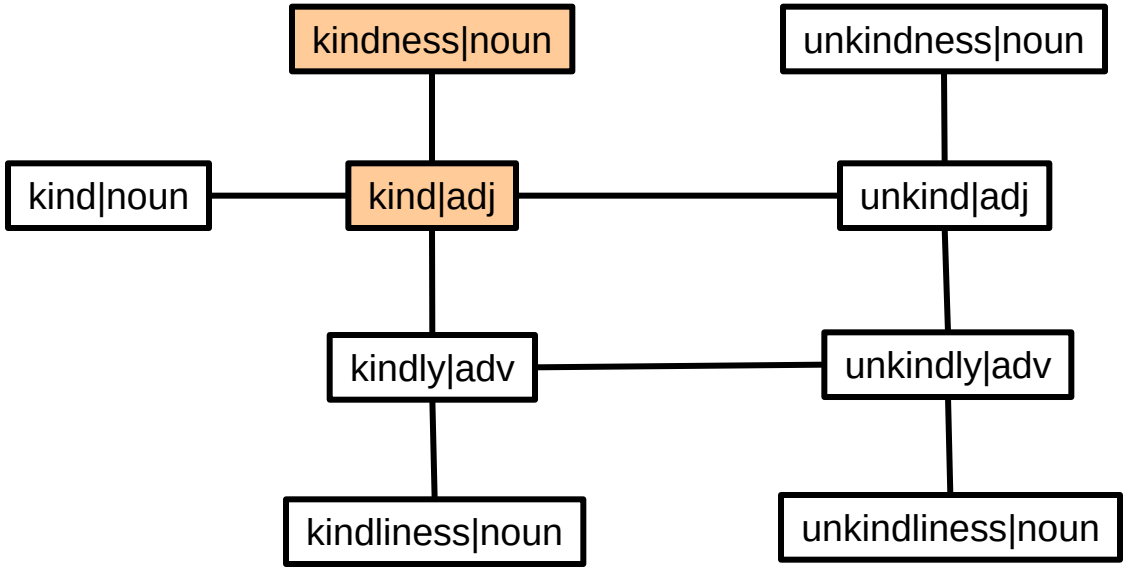
Derivational Analysis (Tagging)

- Performed by linguistic experts
- Is complicated when more than one affix involved
  - look at usage of all related words
  - peel off the derivational affixes
  - check if they are valid words
  - determine the order of derivation
  - multi-option-al, pseudo-hyper-para-thyroid-ism



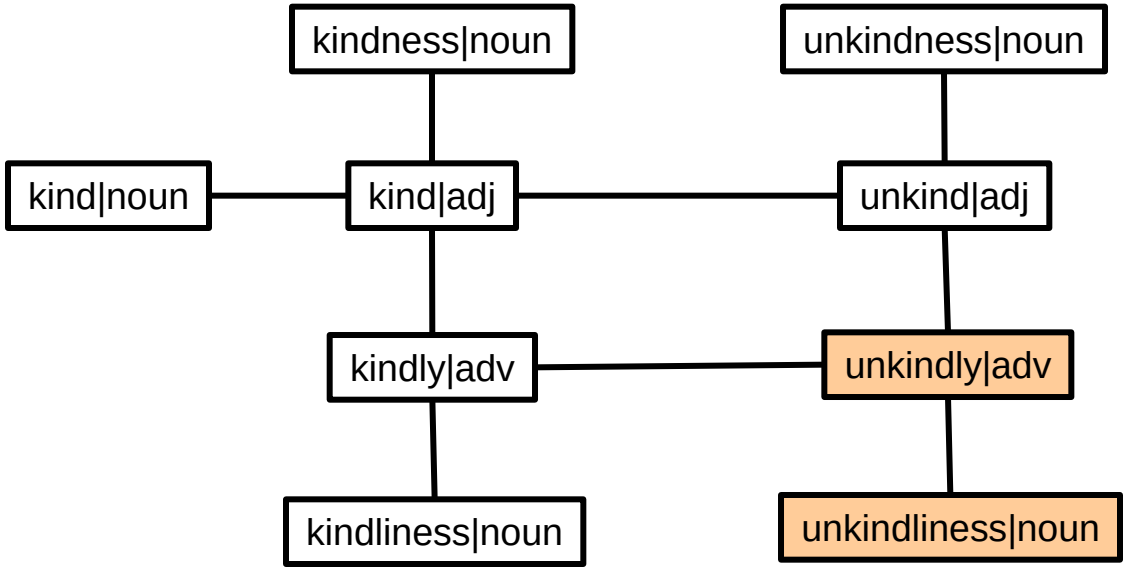
Derivational Pair & Tag

- Format: base 1|category 1|base 2|category 2|Tag
- Examples:
  - kindness|noun|kind|adj|yes
  - unkindly|adv|unkindliness|noun|yes
  - kindness|noun|kindly|adj|no
  - kindness|noun|unkindness|noun|no



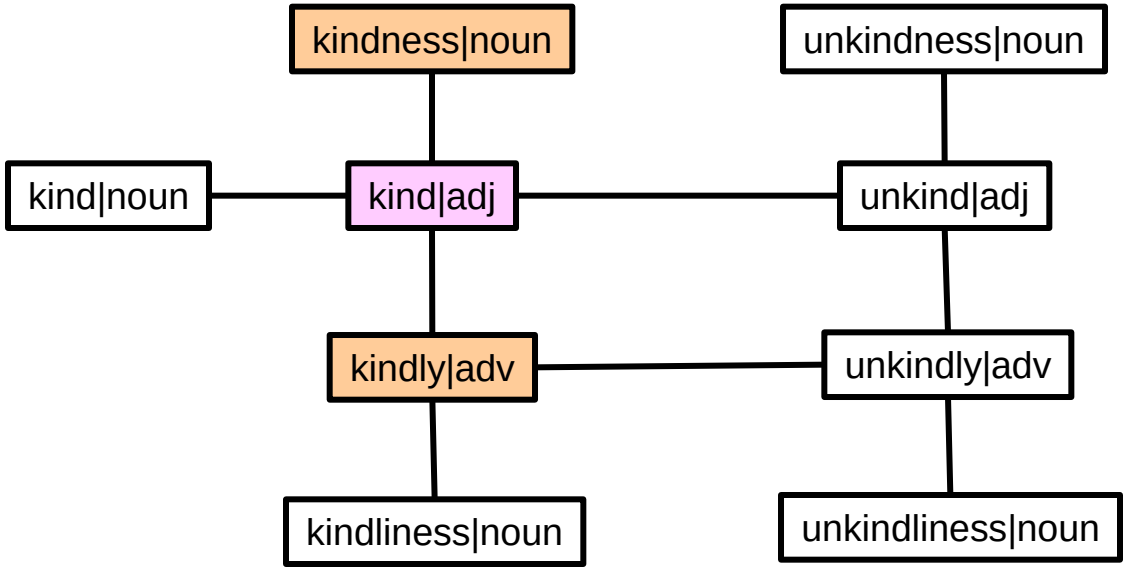
Derivational Pair & Tag

- Format: base 1|category 1|base 2|category 2|Tag
- Examples:
  - kindness|noun|kind|adj|yes
  - unkindly|adv|unkindliness|noun|yes
  - kindness|noun|kindly|adj|no
  - kindness|noun|unkindness|noun|no



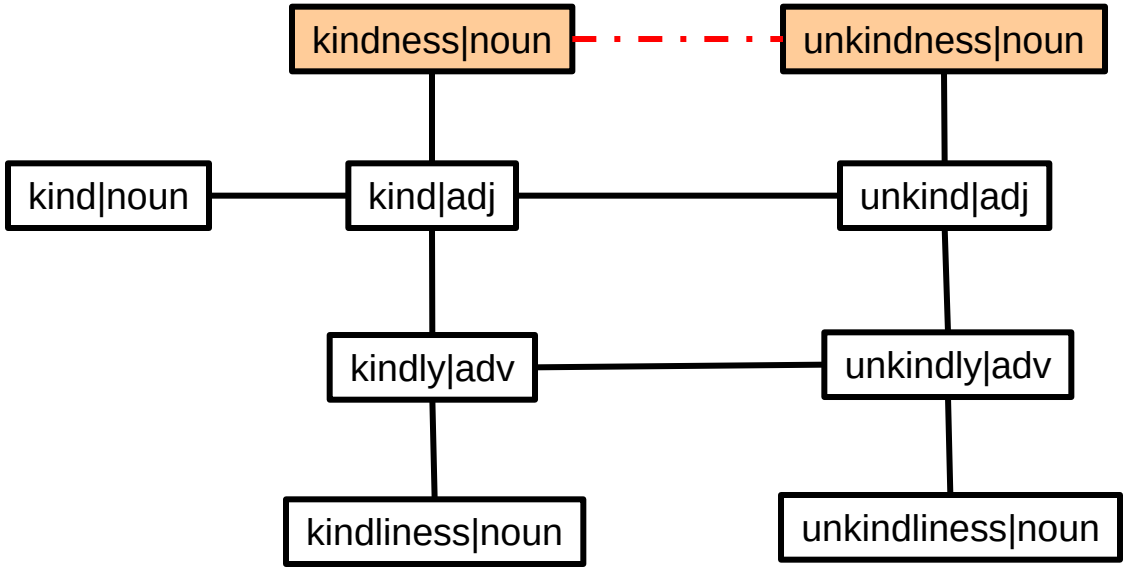
Derivational Pair & Tag

- Format: base 1|category 1|base 2|category 2|Tag
- Examples:
  - kindness|noun|kind|adj|yes
  - unkindly|adv|unkindliness|noun|yes
  - kindness|noun|kindly|adj|no
  - kindness|noun|unkindness|noun|no



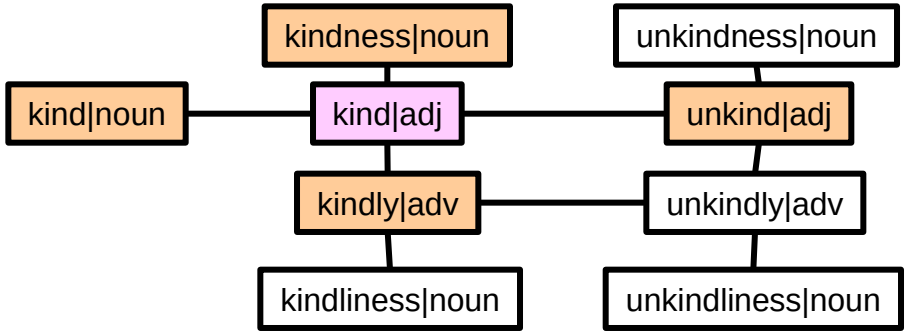
Derivational Pair & Tag

- Format: base 1|category 1|base 2|category 2|Tag
- Examples:
  - kindness|noun|kind|adj|yes
  - unkindly|adv|unkindliness|noun|yes
  - kindness|noun|kindly|adj|no
  - kindness|noun|unkindness|noun|no



Derivational Flows in LVG

- Direct derivation generation (-f:d)
  - All valid derivational pairs associated with the node
  - Example:
    4 derivational variants of kind|adj are found:
    kind|noun, kindness|noun, kindly|adv, and unkind|adj



- Recursive derivation generation (-f:R)
  - Entire derivational network
  - Also provides the distance (number of derivational pairs involved). For example, 2 for kindness|noun and kindly|adv

Derivational Flow

- Facts
  - 4,559 derivational pairs (2011)

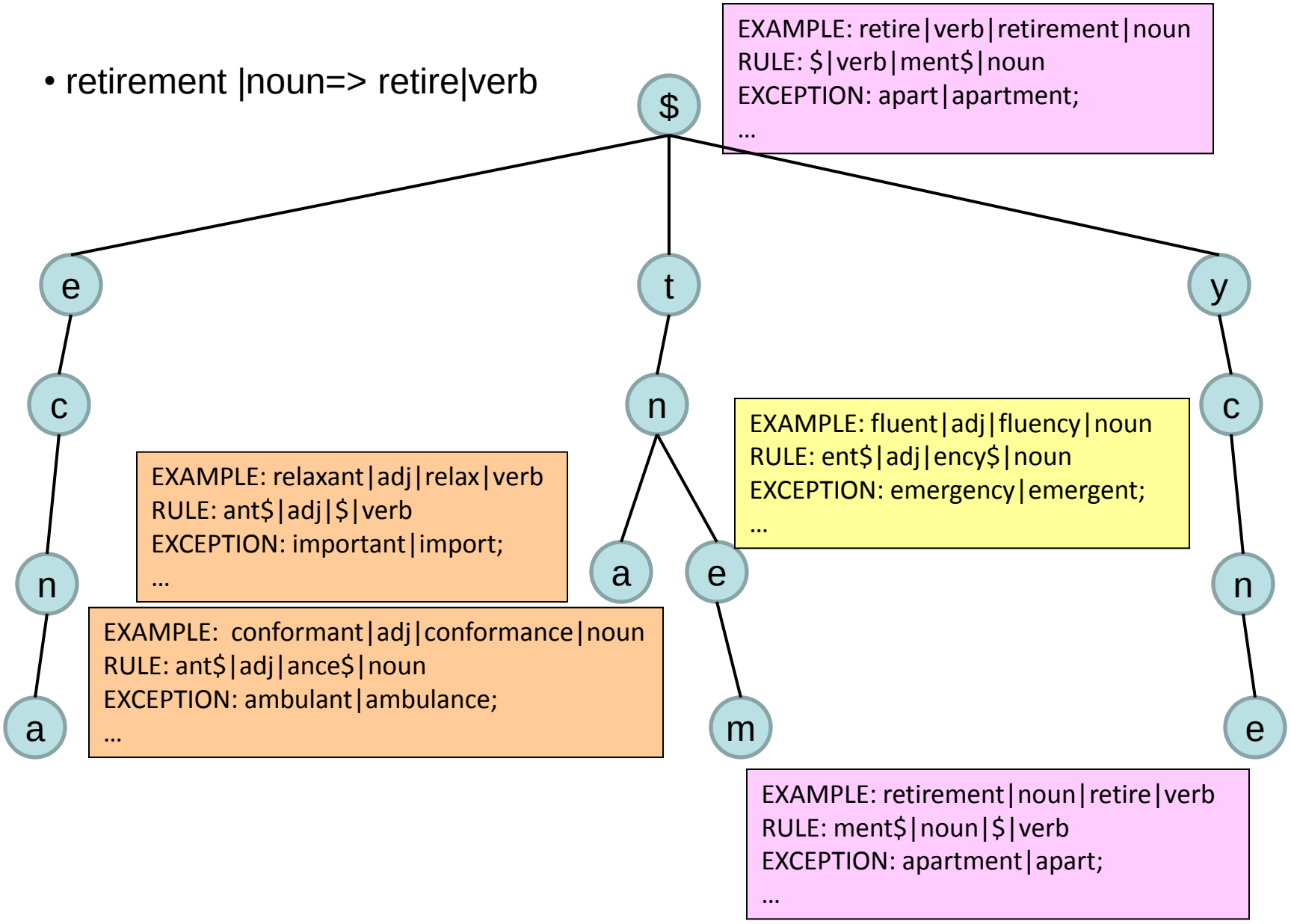
Base 1	Category 1	Base 2	Category 2
…	…	…	…
treatment	noun	treat	noun
…	…	…	…

- Rules
  - 97 SD-Rules
  - Use exceptions to increase precision

EXAMPLE: retirement|noun|retire|verb
RULE: ment$|noun|$|verb
EXCEPTION: apartment|apart;
EXCEPTION: basement|base;
EXCEPTION: department|depart;
…

SD-Rules (Trie)

- retirement |noun=> retire|verb



SD-Rules Filters

- Exception filter
  - Exclude exceptions for the rules
  - Implemented in the Trie
  - depart|verb|department|noun
- Word length filter
  - Exclude short word
  - Default (min.) value is 3
  - moment|noun|mo|verb
- Stem length filter
  - stem length = word length – suffix length
  - Default (min.) value is 3
  - lament|noun|la|verb
- Domain filter
  - Exclude words not in Lexicon
  - color|verb|colorment|noun

Derivations - Evaluation

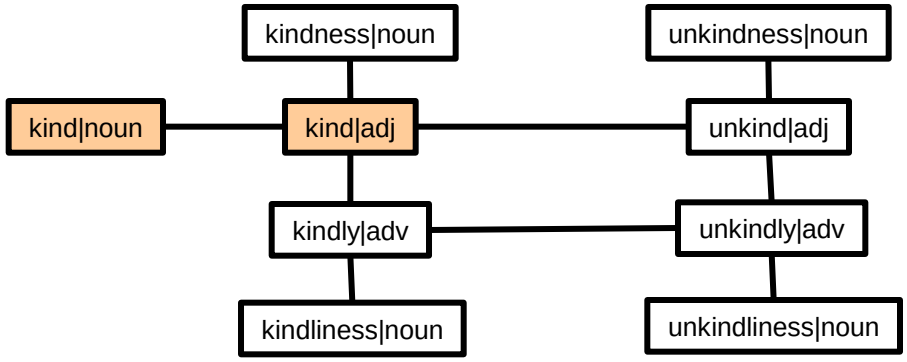
- Facts
  - 4,559 derivational pairs (2011)
  - Maintenance: collecting, validating, and tagging
  - Has not grown proportionally with Lexicon …
  - Prefix derivation & zero derivation (conversion) ?
- Rules
  - 97 SD-Rules
  - High frequency?
  - High precision?
  - Prefix derivation & zero derivation rules?

Challenges

- Facts
  - More coverage: include zeroD, prefixD, suffixD
  - Grows proportionally with Lexicon
  - Higher precision
- Rules
  - Evaluate frequency and precision for SD-Rules
  - Include PD-Rules?
  - Include ZD-Rules?

Zero Derivation (ZD)

- Also called conversion or functional shift
- assigns an existing word to a new syntactic category without any concomitant change in form
- ZD Pairs:
  - kind|noun|kind|adj|yes
  - flex|noun|flex|verb|yes
  - round|adj|round|prep|no



ZD Process

- Retrieve base forms (citation & spelling variants) and category information from Lexicon
- Raw ZD pairs: all words with multiple categories
- Filter programs:
  - Min. Word length (< 2):
    Example: a|noun|a|det|no
    - a|noun: abbreviations for 50+ nouns, such as abortion, acid, adult, …
  - Exclude abbreviations and acronyms
    Example: AAIR|noun|AAIR|adj |no
    - AAIR|noun: age-adjusted incidence rate
    - AAIR|adj: rate-adaptive atrial
- Final tagging

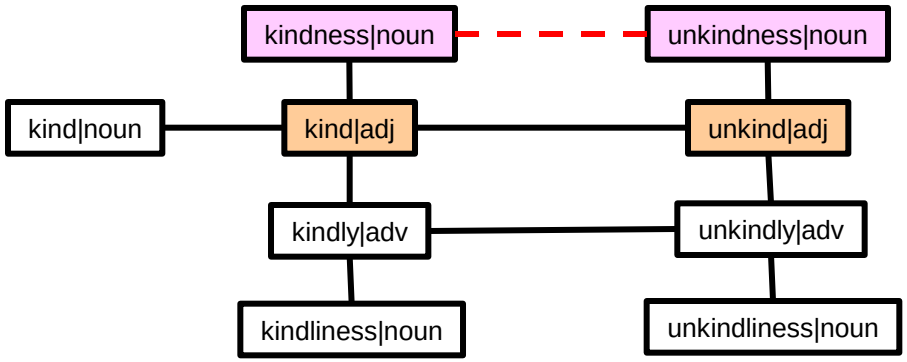
ZD Results

- Raw ZD pairs: 18,400
- Filtered: 10.52%
- Tagged: 89.48 % (recorded for future release)
- No ZD-Rules
  - Valid rate: 80.14%
  - Invalid: 1,718

	ZD Counts	Percentage
Raw	18,400	100.00%
Filtered	1,935	10.52%
Tag - Invalid	1,718	9.34%
Tag - Valid	14,747	80.14%

Prefix Derivation (PD)

- Placed at the beginning of a base word to form another word
- Three patterns:
  - **prefix:** significant|adj nonsignificant|adj
  - **prefix and a dash:** significant|adj|non-significant|adj
  - **prefix and a space:** significant|adj|non significant|adj
- PD pairs:
  - unkind|adj|kind|adj|yes
  - kindness|noun|unkindness|noun|no
  - unplug|verb|plug|noun|no
  - touchable|adj|untouchable|adj|yes
  - touchable|adj|untouchable|noun|yes



PD Process

- Collects common derivational prefixes (143)
- Retrieve all base forms from Lexicon
- Raw PD pairs: match three prefix patterns
  - prefix: nonsignificant|adj|significant|adj
  - prefix and a dash: non-significant|adj|significant|adj
  - prefix and a space: non significant|adj|significant|adj
- Final tagging
  - Tag the most frequent and user requested prefixes

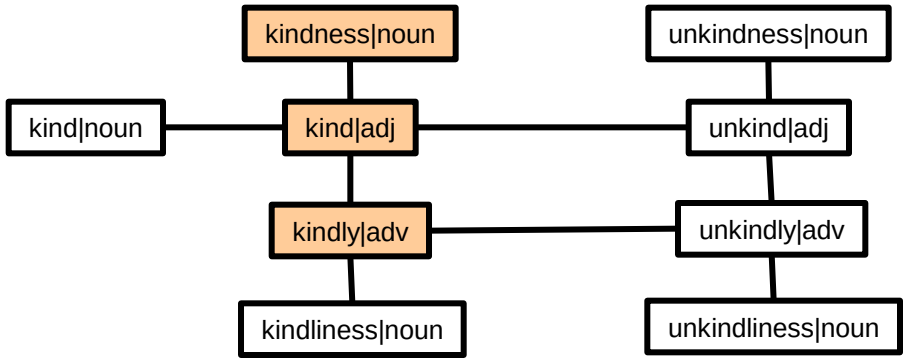
PD Results

	prefix	Raw PrefixD	Valid prefixD
1	non	16,471 (14.31%)	12,598 (76.49%)
2	pre	9,651 (8.38%)	7,224 (74.85%)
3	post	9,490 (8.24%)	7,621 (80.31%)
4	anti	6,500 (5.65%)	5,051 (77.71%)
5	sub	4,262 (3.70%)	2,698 (63.30%)
6	re	4,198 (3.65%)	1,527 (36.37%)
7	inter	4,143 (3.60%)	2,708 (65.36%)
8	multi	3,781 (3.28%)	2,169 (57.37%)
9	intra	3,575 (3.10%)	2,360 (66.01%)
10	pseudo	3,331 (2.89%)	2,082 (62.50%)
11	un	2,211 (1.92%)	1,271 (57.49%)
12	peri	1,901 (1,65%)	1,367 (71.91%)
...	...	...	...
Tag		86,333 (74.98%)	56,694 (65.67%)
Raw		115,139 (100%)	

- Raw PD pairs: 115,139
- Tagged: 74.98 %
  recorded for future release
- No PD-Rules
  - Avg. 65.67% valid rate
  - Max. 80.31% valid rate
- No Category filter:
  - 24.54% of valid PD pairs changes category
    - fog|noun|antifog|adj|yes
- No acronym or abbreviation filter:
  - 0.83% of valid PD pairs are acronyms or abbreviations
    - MDR|noun|antiMDR|adj|yes
      MDR, acronym for "multidrug resistance"

Suffix Derivation (SD)

- Also called a postfix or ending
- Placed after the stem of a word to form another word
- Several hundreds of derivational suffixes
- Collects common derivational suffixes (200)
- SD Pairs:
  - kind|adj|kindness|noun
  - kind|adj|kindnly|adv



SD Process - Nominalization

- The process of producing a noun from a verb or an adjective via the derivational suffix
- Coded in Lexicon
- A type of suffix derivation
- Bi-directional

```
{base=locate
entry=E0037939
        cat=verb
        variants=reg
        tran=np
        link=advbl
        cplxtran=np,advbl
        nominalization=location|noun|E0037940
}
```

```
{base=location
entry=E0037940
        cat=noun
        variants=reg
        variants=uncount
        compl=pphr(of,np)
        compl=pphr(by,np)
        nominalization_of=locate|verb|E0037939
}
```

SD Process - ND

- Raw ND pairs: retrieve all nominalization information from Lexicon
- Filters:
  - Pattern filter: exclude invalid SD for verb particle ND
    Pattern-1: baseParticle|noun|base|verb => backup|noun|back|verb
    Pattern-2: base-Particle|noun|base|verb => cut-through|noun|cut|verb
    Pattern-3: inflParticle|noun|base|verb => grownup|nou|grow|verb
    Pattern-4: infl-Particle|noun|base|verb => salting-in|noun|salt|verb
    Particle Exception: "per" => shopper|noun|shop|verb
  - Exception filter: exclude other known SD pairs
    Examples:
    face-saving|noun|save|verb
    decision-making|noun|make|verb
    merry-making|noun|make|verb
    lovemaking|noun|make|verb
    …

ND (SD) Results

- Raw ND pairs: 14,445
- Filtered: 0.50%
- Valid: 99.50 % ND pairs (program generated)

ND Pairs	Filtered	Valid
14,445	72	14,373
100%	0.50%	99.50%

ND to SD-Rules

- Identified SD-Rules from Valid ND pairs
  - 496 possible rules are found
  - **locat**ion|noun|**locat**e|verb => **ion$|noun|e$|verb**
- Further analysis

Derivation Suffix Rules	Example	Counts
ation$\|noun\|ate$|verb	location|noun|locate|verb	1,547
sion$\|noun\|se$|verb	tension|noun|tense|verb	77
ution$\|noun\|ute$|verb	delution|noun|delute|verb	37
etion$\|noun\|ete$|verb	completion|noun|complete|verb	22
otion$\|noun\|ote$|verb	devotion|noun|devote|verb	6
ition$\|noun\|ite$|verb	Ignition|noun|ignite|verb	4
otion$\|noun\|ote$|verb	coercion|noun|coerce|verb	1

- Map with existing SD-Rules in LVG

Identified Rules	Rules in Lexical Tools	Counts
ness$\|noun\|$|adj	ness$\|noun\|$|adj	2,481
ion$\|noun\|e$|verb	ation$\|noun\|ate$|verb	1,547
	sion$\|noun\|se$|verb	77
	Others ...	70
ity$\|noun\|$|adj	ity$\|noun\|$|adj	881
	icity$\|noun\|ic$|adj	745
ility$\|noun\|le$|adj	ability$\|noun\|able$|adj	1,036
	Others ...	253
ation$\|noun\|e$|verb	ation$\|noun\|e$|verb	1,133

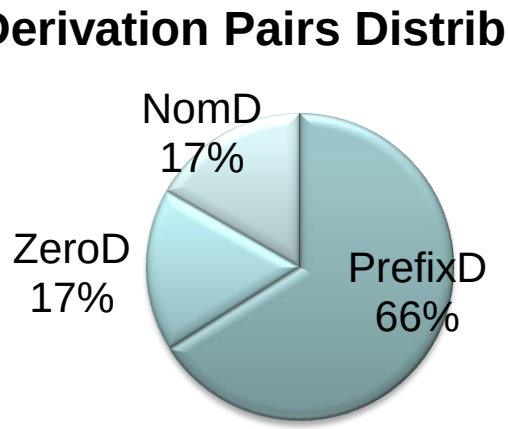
Final Compile

- Final affix validation program:
  - Affix check: check the first and last (3) characters between base forms of derivational pairs to assure only one affix is involved.
  - Exception filter:
    - able|adj|ability|noun
    - long|adj|length|noun
    - high|adj|height|noun
    - …
  - Spelling variants
    - dysmaturity|noun|dismature|adj
    - gray|adj|grey|noun
    - haemolysed|adj|hemolyzation|noun
    - …
- Combine all three lists (ZD, PD, ND)

Final Results

- More coverage (will grow with Lexicon)

2011 Lvg Facts	2012 Lvg Facts
4,559	89,950



- Virtually 100% precision rate (Facts)

Future Work

- ZeroD:
  - Rules-based filter: syntactic category and other linguistic knowledge
  - PD-Rules: syntactic category and other linguistic knowledge
- PrefixD:
  - Update prefix list annually
  - Complete tagging processes for all collected prefix
  - Rules-based filter: syntactic category and other linguistic knowledge
  - PD-Rules: syntactic category and other linguistic knowledge
- SuffixD:
  - Develop a thorough validation process for existing SD-Rules by all possible raw SD pairs in the Lexicon
  - Find all exceptions for each SD-Rules in Lexicon
  - Rules-based filter: syntactic category and other linguistic knowledge
  - SD-Rules: syntactic category and other linguistic knowledge

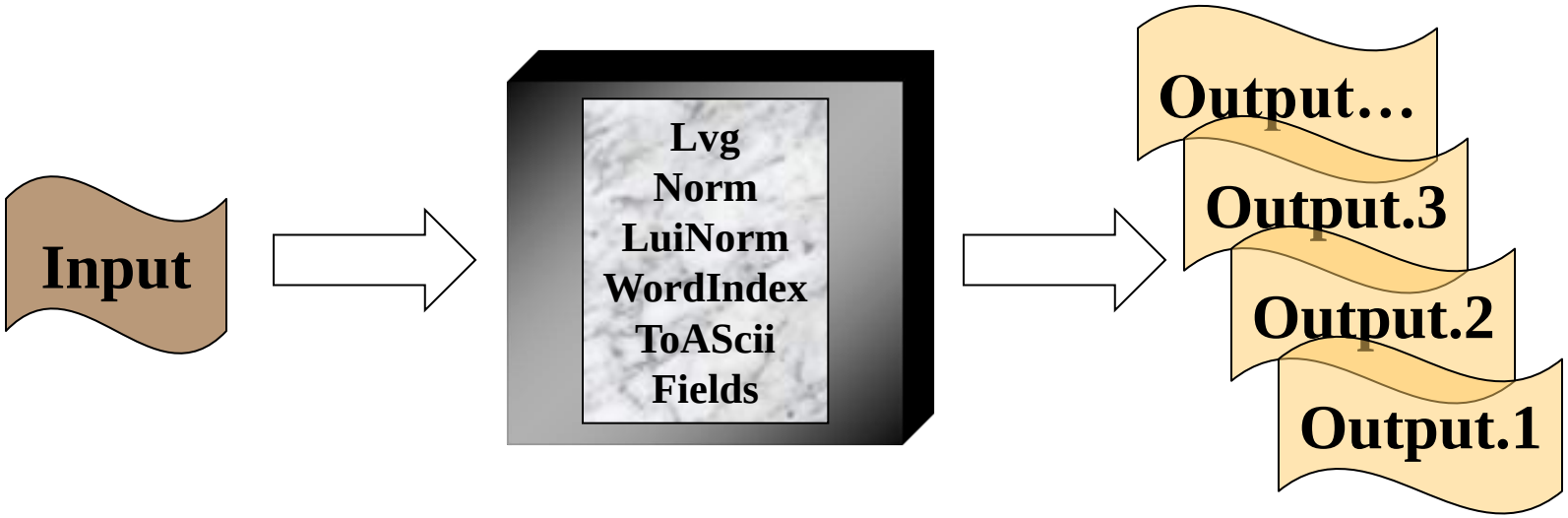
Questions

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

Lexical Tools

http://SPECIALIST.nlm.nih.gov/lvg

Lexical Tools - Six Tools



Lexical Tools - 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
  - making word, term, phrase indexes
  - matching queries with indexed entries
  - increasing recall and/or precision

Facts

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

Norm

- Composed of 11 Lvg flow components to abstract away from:
  - case
  - punctuation
  - possessive forms
  - inflections
  - spelling variants
  - stop words
  - Diacritics, ligatures & symbols (Unicode to ASCII)
  - word order

Norm

q0: map Unicode 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 non-ASCII char
w: sort words by order

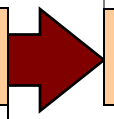
Norm

Hodgkin's Diseases, NOS

q0: map Unicode 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 non-ASCII char
w: sort words by order

Norm

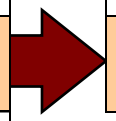
q0: map Unicode symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map non-ASCII char
w: sort words by order



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

Norm

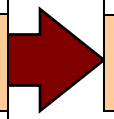
q0: map Unicode symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map non-ASCII char
w: sort words by order



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

Norm

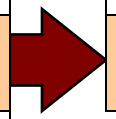
q0: map Unicode symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map non-ASCII char
w: sort words by order



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

Norm

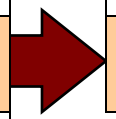
q0: map Unicode symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map non-ASCII char
w: sort words by order



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

Norm

q0: map Unicode 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 non-ASCII char
w: sort words by order



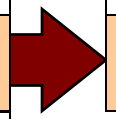
Hodgkin's Diseases, NOS
Hodgkin's Diseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases NOS
Hodgkin Diseases

Norm

	Hodgkin's Diseases, NOS
q0: map Unicode symbols to ASCII	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
l: lowercase	hodgkin diseases
B: uninflect each words in a term	
Ct: retrieve citations	
q7: Unicode core Norm	
q8: strip or map non-ASCII char	
w: sort words by order	

Norm

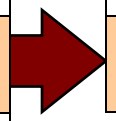
q0: map Unicode symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map non-ASCII char
w: sort words by order



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

Norm

q0: map Unicode 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 non-ASCII char
w: sort words by order



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

Norm

	Hodgkin's Diseases, NOS
q0: map Unicode symbols to ASCII	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
l: lowercase	hodgkin diseases
B: uninflect each words in a term	hodgkin disease
Ct: retrieve citations	hodgkin disease
q7: Unicode core Norm	hodgkin disease
q8: strip or map non-ASCII char	
w: sort words by order	

Norm

	Hodgkin's Diseases, NOS
q0: map Unicode symbols to ASCII	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
l: lowercase	hodgkin diseases
B: uninflect each words in a term	hodgkin disease
Ct: retrieve citations	hodgkin disease
q7: Unicode core Norm	hodgkin disease
q8: strip or map non-ASCII char	hodgkin disease
w: sort words by order	

Norm

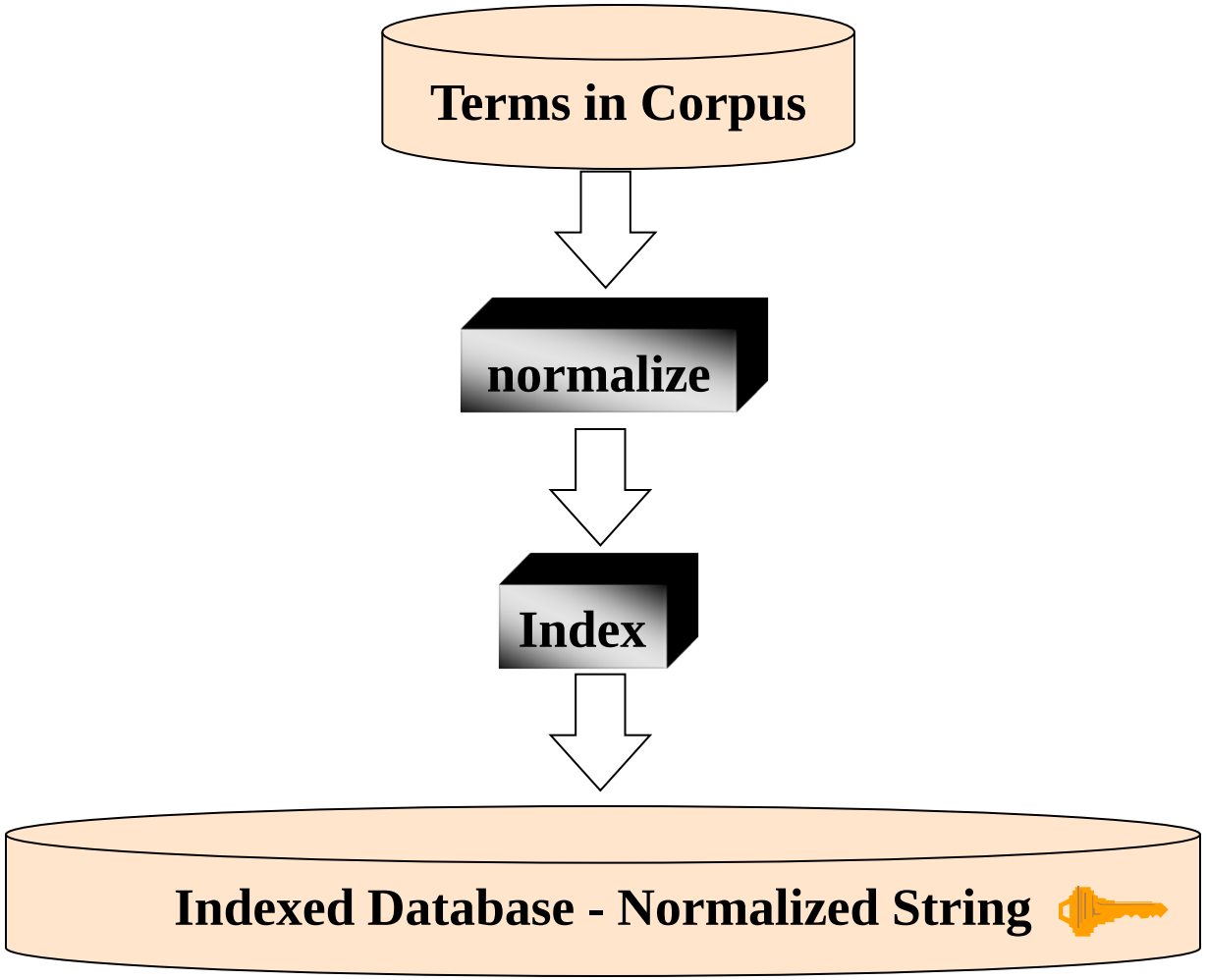
	Hodgkin's Diseases, NOS
q0: map Unicode symbols to ASCII	Hodgkin's Diseases, NOS
g: remove genitives	Hodgkin Diseases, NOS
rs: remove parenthetic plural forms	Hodgkin Diseases, NOS
o: replace punctuation with spaces	Hodgkin Diseases NOS
t: strip stop words	Hodgkin Diseases
l: lowercase	hodgkin diseases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

Norm: Example

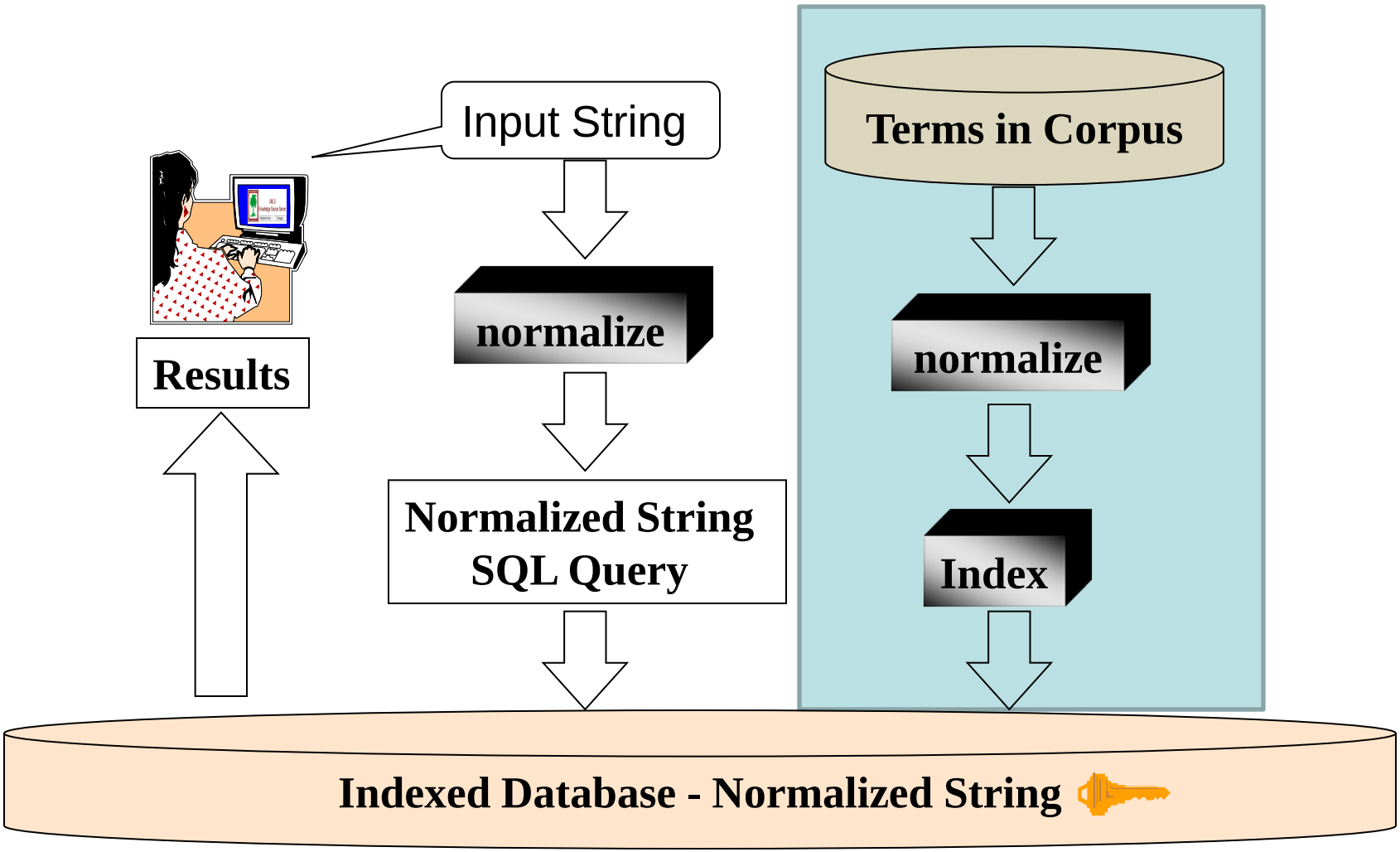
- Hodgkin Disease
- HODGKINS DISEASE
- Hodgkin's Disease
- Disease, Hodgkin's
- HODGKIN'S DISEASE
- Hodgkin's disease
- Hodgkins Disease
- Hodgkin's disease NOS
- Hodgkin's disease, NOS
- Disease, Hodgkins
- Diseases, Hodgkins
- Hodgkins Diseases
- Hodgkins disease
- hodgkin's disease
- Disease;Hodgkins
- Disease, Hodgkin
- …

disease hodgkin

Normalize and Index



Normalize and Index



Questions

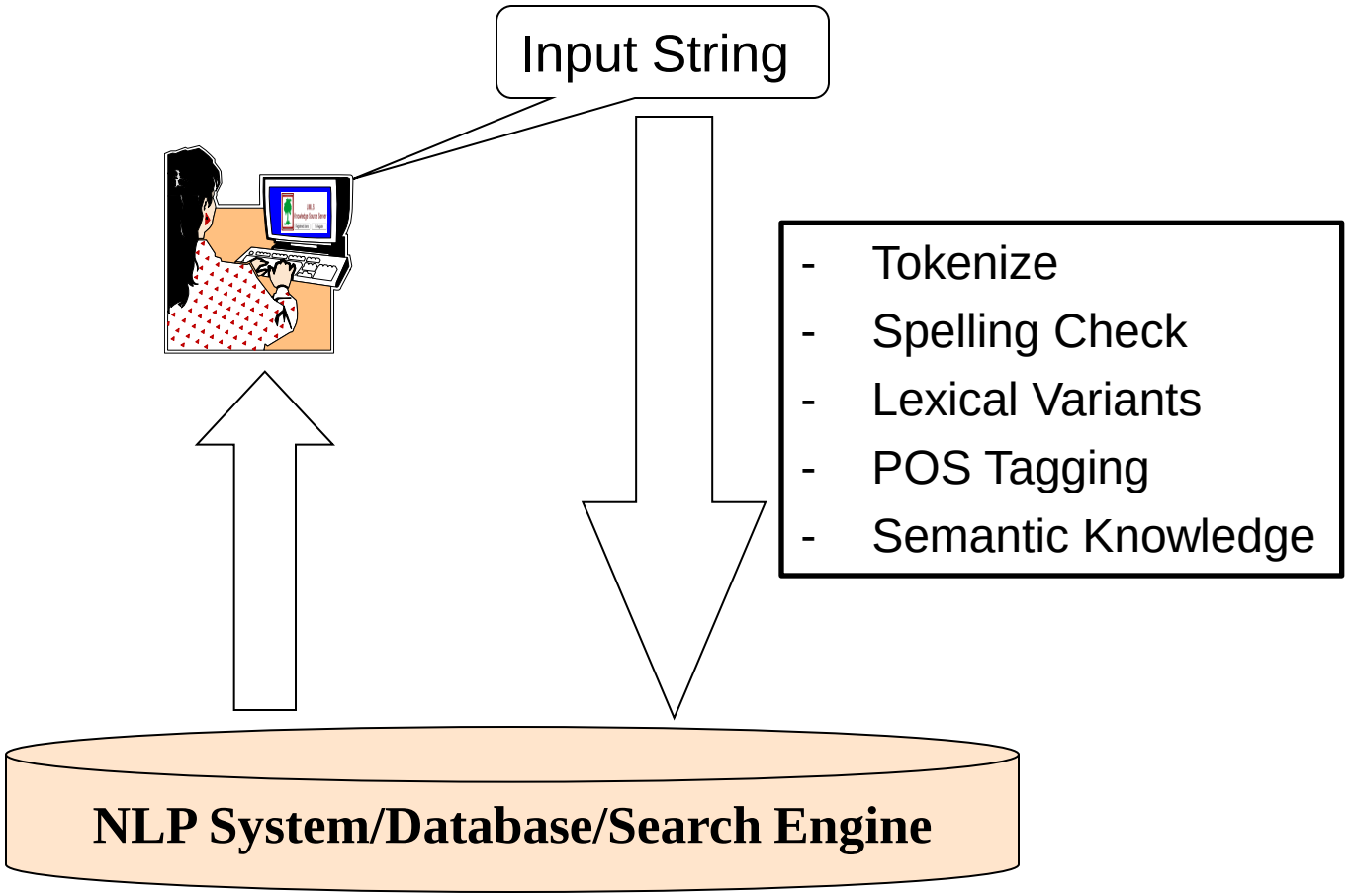
- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov

NLP & NLP Tools

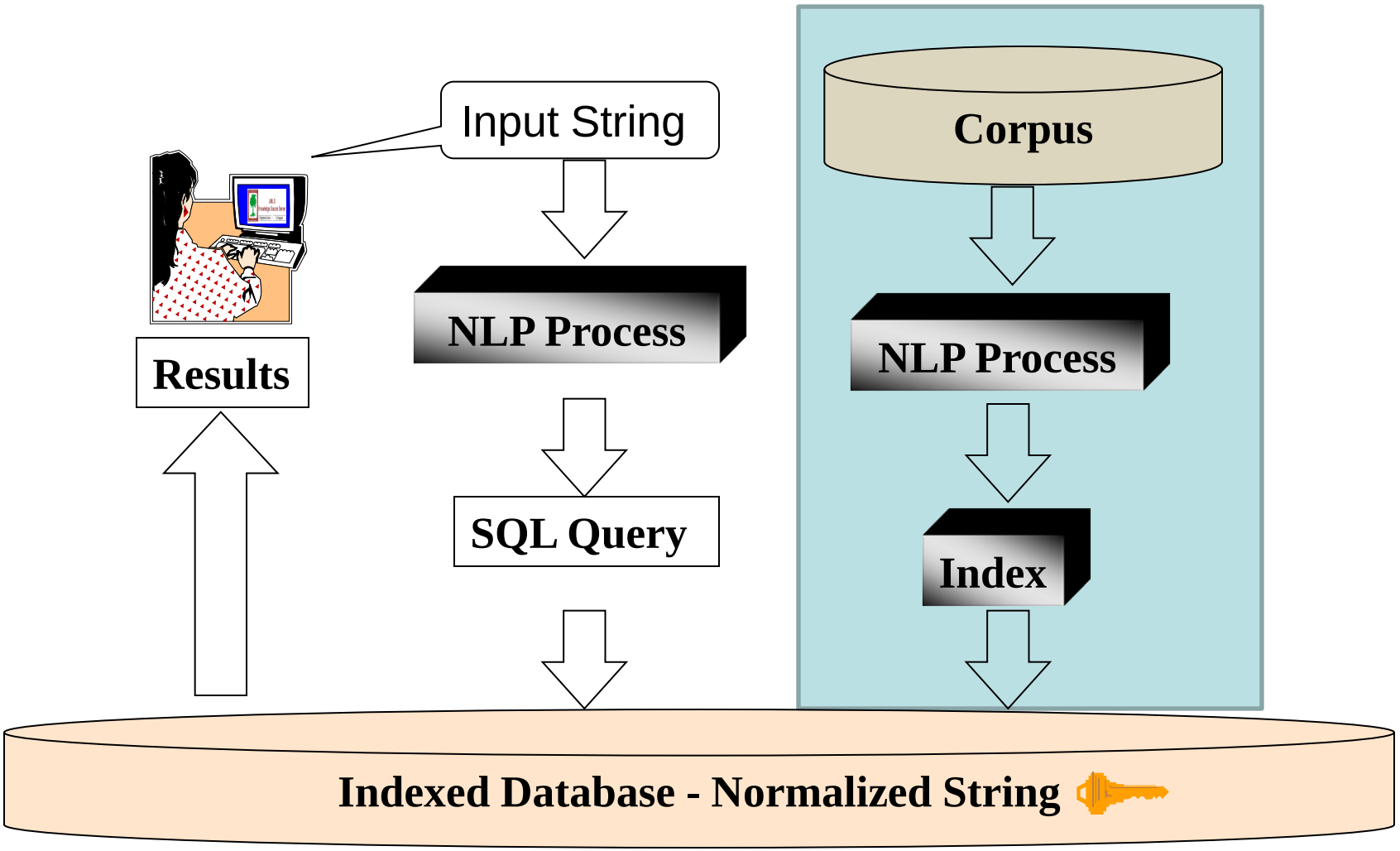
- Natural Language
  - is ordinary language that humans use naturally
  - may be spoken, signed, or written

- Natural Language Processing
  - NLP is to process human language to make their information accessible to computer applications
  - The goal is to design and build software that will analyze, understand, and generate human language
  - Most NLP applications require knowledge from linguistics, computer science, and statistics

Information Retrieval



Information Retrieval



Core NLP Tasks

- Ex: Web search engine for biomedical information
  - **Software:**
    - keyword search
      - break inputs into words
      - POS tagging
      - other annotation
    - spelling check
      - suggest correct spelling for misspelled words
    - lexical variants
      - spelling variants, inflectional/uninflectional variants, synonyms, acronyms/abbreviations, expansions, derivational variants, etc.
    - semantic knowledge
      - map text to Metathesaurus concepts
      - Word Sense Disambiguation (WSD)
  - **Data:**
    - corpus: annotation/tagging

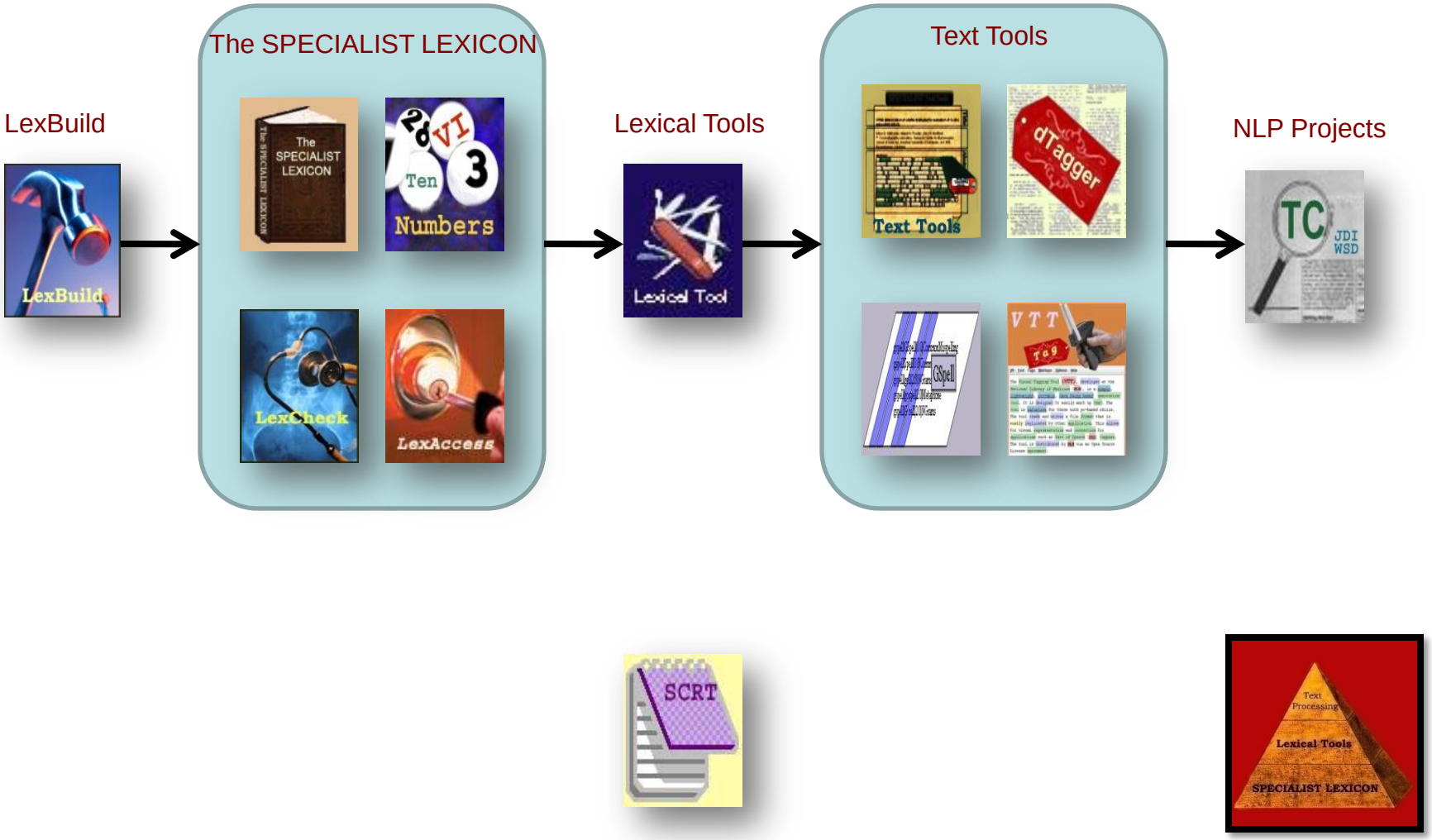
NLP Tools

- Ex: Web search engine for biomedical information
  - **Software:**
    - keyword search
      - break inputs into words (Text Tools)
      - POS tagging (dTagger)
      - Other annotation (Visual Tagging Tool, VTT)
    - spelling check
      - suggest correct spelling for misspelled words (gSpell)
    - lexical variants
      - spelling variants, inflectional/uninflectional variants, synonyms, acronyms/abbreviations, expansions, derivational variants, etc. (Lexical Tools)
    - semantic knowledge
      - map text to Metathesaurus concepts (MetaMap, MMTX)
      - Word Sense Disambiguation (TC - StWSD)
  - **Data: corpus**
    - annotation/tagging (Text Tools, dTagger, VTT, Lexical Tools)

Core NLP Tools

- Ex: Web Search Engine for biomedical information
  - Software:
    - keyword search
      - break inputs into words (Text Tools)
      - POS tagging (dTagger)
      - Other annotation (Visual Tagging Tool, VTT)
    - spelling check
      - suggest correct spelling for misspelled words (gSpell)
    - lexical variants
      - spelling variants, inflectional/uninflectional variants, synonyms, acronyms/abbreviations, expansions, derivational variants, etc. (Lexical Tools)
    - semantic knowledge
      - map text to Metathesaurus concepts (MetaMap, MMTX)
      - Word Sense Disambiguation (TC - StWSD)
  - Data: corpus
    - annotation/tagging (Text Tools, dTagger, VTT, Lexical Tools)

The SPECIALIST NLP Tools



Questions

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