

The SPECIALIST Lexicon and NLP Tools (Enhanced LexSynonym Acquisition and Features)

By: Dr. Chris J. Lu

NLM – LHNCBC - CGSB

Oct., 2017

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

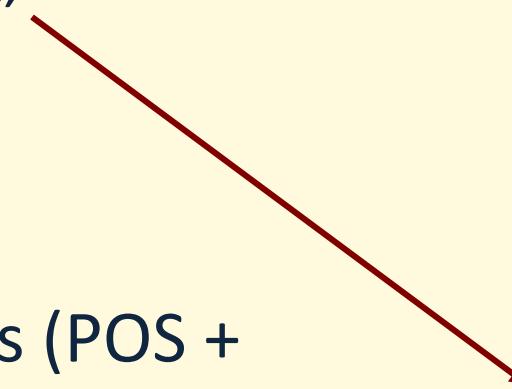
Outline

- Introduction
 - The SPECIALIST Lexicon
 - The SPECIALIST NLP Tools (Lexical Tools)
- Applications - LexSynonyms
 - Natural Language Processing (NLP)
 - LexSynonyms
- Questions (anytime)



1. The SPECIALIST Lexicon

- A fancy synonym for “dictionary”
- A syntactic lexicon
- Biomedical and general English
- Over 490,000 records, 1M words (POS + forms)
- Designed/developed to provide the lexical information needed for the NLP (Natural Language Processing) system
- Distributed in the Unified Medical Language System (UMLS) Knowledge Sources by the National Library of Medicine (NLM)



THE INSOMNIAC'S DICTIONARY

Illeism: Reference to oneself by use of the third person
Infavoidance: The act of covering up one's inferiority complex
Inglenook: A place by the fire or any warm and comfortable area
Insilium: Legal term for evil advice or counsel
Jamais vu: Illusion that one has never previously experienced a situation, when in fact it is quite familiar (see *Déjà vu*)
Jen: A compassionate love for all humanity or for the whole world
Karateka: A karate expert
Kloof: A deep ravine
Kludge: A system (especially of computers) made up of poorly matched components
Lallation: Pronouncing an “R” so that it sounds like an “L”
Lapidation: The act of stoning a person to death
Latrocination: A robbery that involves the use of force or violence

Lexicon: A fancy synonym for “dictionary”

Litotes: A form of understatement in which two negatives are used to make a positive (“he was not unhappy”)

Longueur: A long and boring passage in a work of literature, drama, music, etc.

Macarism: The practice of making others happy by praising them

Matutinal: Pertaining to anything that takes place in the morning

Melorrhea: The writing of excessively long musical works

Meteorism: A tendency to uncontrollable passing of intestinal gas

Metrona: A young grandmother

Microperf: The very small perforations along the edges of computer paper

Migrateur: A wanderer

Mnemonic: That which assists memory (a classic mnemonic device is the one familiar to astronomy students: “Oh be a fine girl, kiss me”—a unique way to remember the stellar classifications O, B, A, F, G, K, and M)

Moria: Morbid impulse to make jokes

Omnistrain: The stresses of modern life

Omphaloskepsis: The act of contemplating one's navel

Onychophagy: The habit of biting one's fingernails

Oxymoron: A phrase or expression composed of contradictory elements (“awfully good,” for example)

140



LexBuild Process (Computer-Aided)

Sources:

- Word candidates from **MEDLINE**
- Others
 - Dorland's Illustrated Medical Dictionary
 - American Heritage Word Frequency book (top 10K)
 - Longman's Dictionary of Contemporary English (Top 2K lexical items)
 - The Metathesaurus browser and retrieval system
 - The UMLS test collection
 - ...

Reviewed by lexicographers:

- Google Scholar
- Dictionaries
- Biomedical publications
- Domain-specific databases
- Nomenclature guidelines
- books
- Essie Search Engine
- ...

Build:

- **LexBuild**
- **LexAccess**
- **LexCheck**



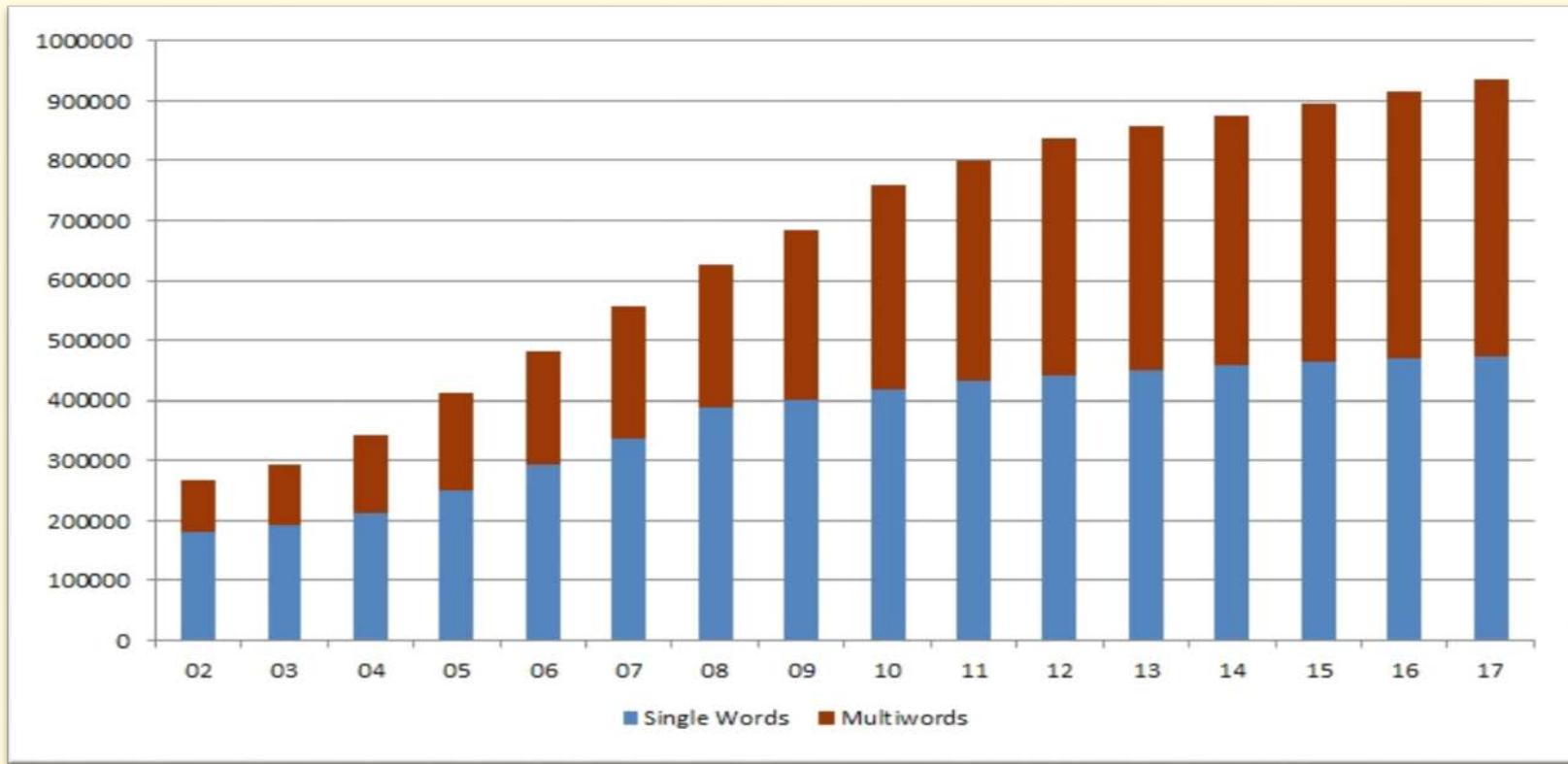
Team of Lexicon Builders

- Dr. Alexa McCray, founded in 1994 (previous LHC Director, 2005-)
- Allen Browne, father of the SPECAILIST Lexicon (retired 2017)
- Dr. Dina Demner Fushman
- Dr. Chris J. Lu
- Dr. Lynn McCreedy
- Destinee Tormey
- Francois Lang



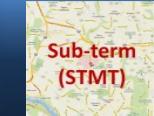
Lexicon Growth – 2002 to 2017

- 498,430 lexical records
- 1,110,321 words (categories and inflections)
- 935,276 forms (spelling only)
 - Single words: 472,608 (50.53%); Multiwords: 462,668 (49.47%)



(Multi)Words for Lexical Records

- Lexicon terms: single words and multiwords
 - Space(s): ice-cream vs. ice cream
- Four criteria for Lexicon terms:
 - Part of Speech (POS):
 - tear break up time, frog erythrocytic virus, cardiac surgery
 - Inflection morphology (uninflection):
 - left pulmonary veins (“left pulmonary vein” and “leave pulmonary vein”)
 - Specific meaning:
 - hot dog (high temperature canine?)
 - Word order:
 - trial and error, up and down (vs. food and water)
 - exercise training vs. training exercise (military)

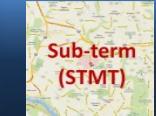
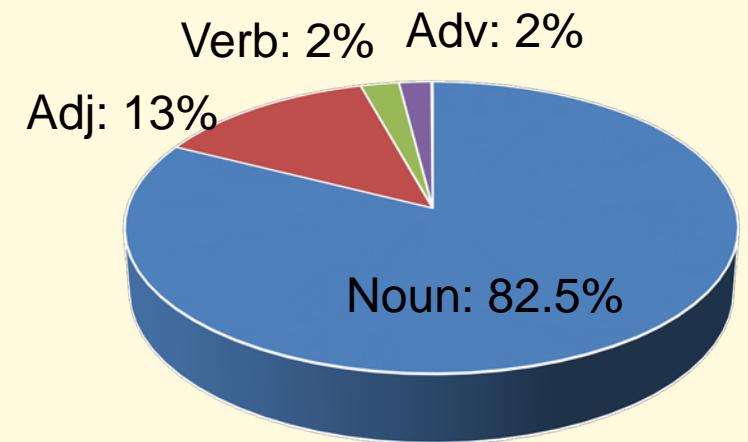
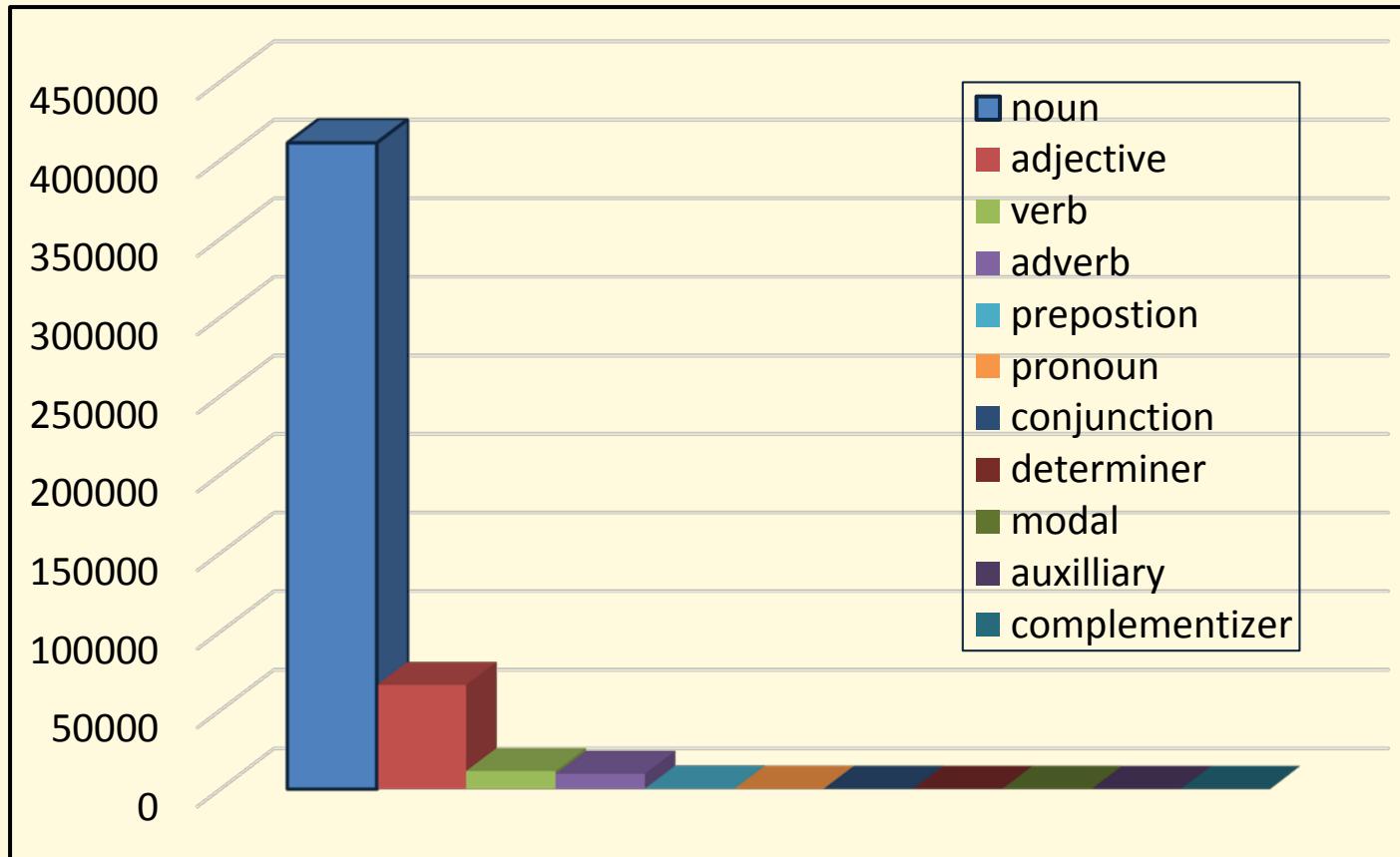


Lexical Records - Information

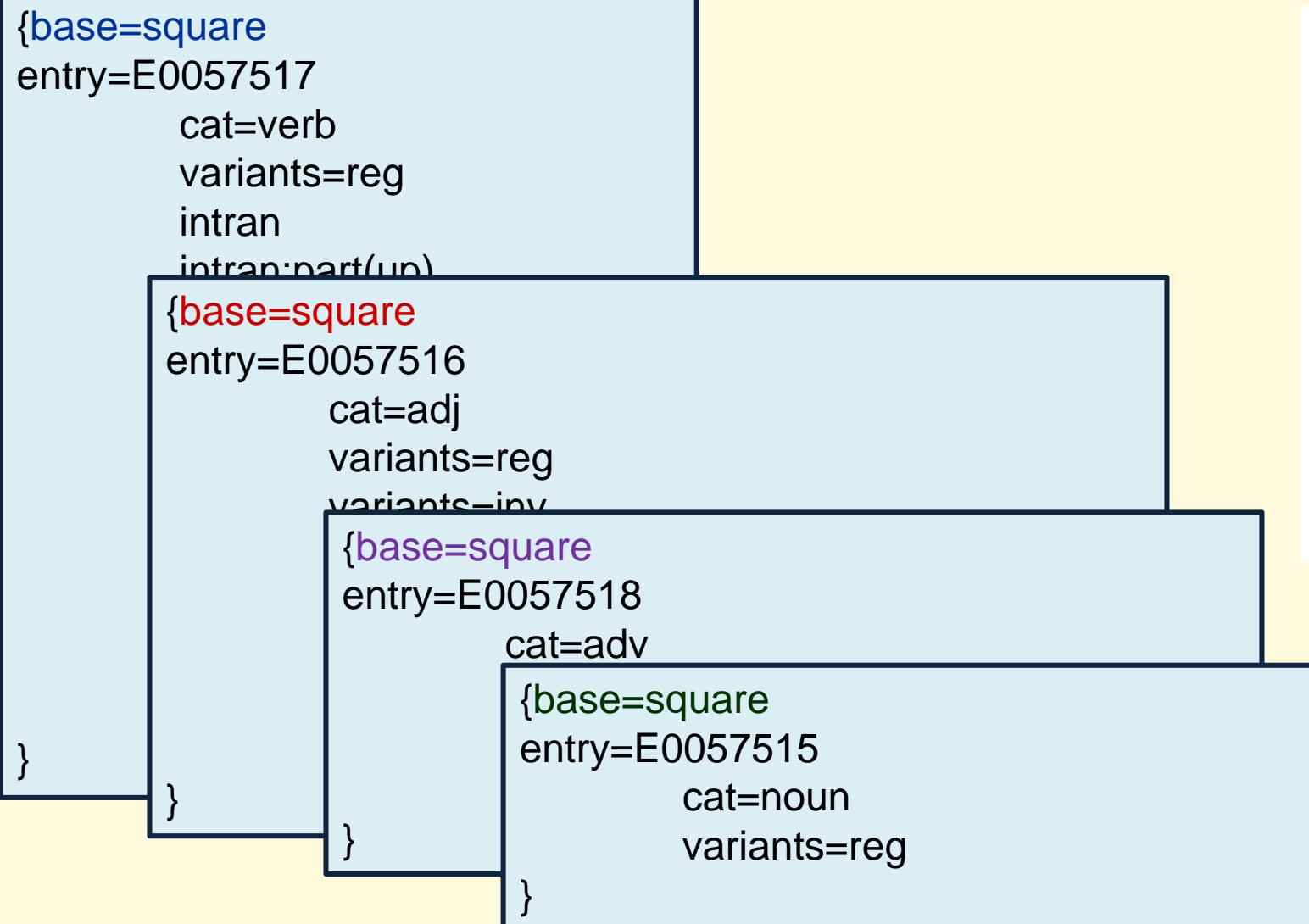
- POS (Part-of-Speech)
- Morphology
 - Inflection
 - Derivation
- Orthography
 - Spelling variants
- Syntax
 - Complementation for verbs, nouns, and adjectives
- Other
 - Expansions of abbreviations and acronyms
 - Nominalizations
 - ...



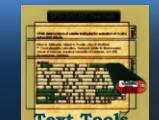
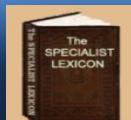
Categories – Parts of Speech (11)



Lexical Records & POS



village **square** **square** the circle
fair and **square** **square** root



Morphology

➤ Inflectional

- noun: book, books
- verb: categorize, categorizes, categorized, categorizing
- adj: red, redder reddest

➤ Derivational

- example: transport
- suffix - transportation, transportable, transporter, ...
- prefix – autotransport, intratransport, pretransport, ...
- conversion (zero) - transport (verb), transport (noun)



Orthography (Spelling Variation)

- color | colour
- grey | gray
- align | aline
- Grave's disease | Graves's disease | Graves' disease
- civilize | civilize
- harbor | harbor
- fetus | foetus | fœtus
- centre | center
- spelt | spelled
- ice cream | ice-cream
- xray | x-ray | x ray



Syntax - Verb Complements

➤intran

- I'll treat.

➤tran=np

- He treated the patient.

➤ditran=np,pphr(with,np)

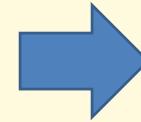
- She treated the patient with the drug.

➤ ...

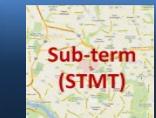


Lexical Information to Lexical Records

Lexical Information Base	color
Part of speech	<ul style="list-style-type: none"> • noun
Inflectional morphology (inflections)	<ul style="list-style-type: none"> • color • colors
Orthography	<ul style="list-style-type: none"> • colour
Abbreviation/Acronym	<ul style="list-style-type: none"> • N/A
Syntax (complementation)	<ul style="list-style-type: none"> • N/A
...	<ul style="list-style-type: none"> • ...
Derivational morphology (derivations)	<ul style="list-style-type: none"> • colorable • colorful • colorize • colorist • ...
LexSynonyms	<ul style="list-style-type: none"> • chromatic



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



UTF-8 (Since 2006)

```
{base=resume  
spelling_variant=résumé  
spelling_variant=resumé  
entry=E0053099  
    cat=noun  
    variants=reg  
}
```

```
{base=deja vu  
spelling_variant=deja-vu  
spelling_variant=déjà vu  
entry=E0021340  
    cat=noun  
    variants=uncount  
}
```

```
{base=divorcé  
entry=E0543077  
    cat=noun  
    variants=reg  
}
```

```
{base=role  
spelling_variant=rôle  
entry=E0053757  
    cat=noun  
    variants=reg  
}
```

```
{base=cafe  
spelling_variant=café  
entry=E0420690  
    cat=noun  
    variants=reg  
}
```

```
{base=Pécs  
entry=E0702889  
    cat=noun  
    variants=uncount  
    proper  
}
```



Lexicon Unigram Coverage – Without WC

- Total unique word for MEDLINE (2016): 3,619,854
- Lexicon covers 10.62 % unigrams in MEDLINE

Types	Word Count	Percentage %	Accu. %
LEXICON (S)	296,747	8.1978%	8.1978%
NUMBER	62	0.0017%	8.1995%
DIGIT	87,437	2.4155%	10.6150%
NON-WORD*	43,811	1.2103%	11.8253%
NEW	3,191,797	88.1747%	100.0000%
Total	3,619,854		

* NON-WORD: a single word only exist in multiword, such as “non”, “vitro”, “vivo”, “intra”, etc.



Lexicon Unigram Coverage – With Frequency (WC)

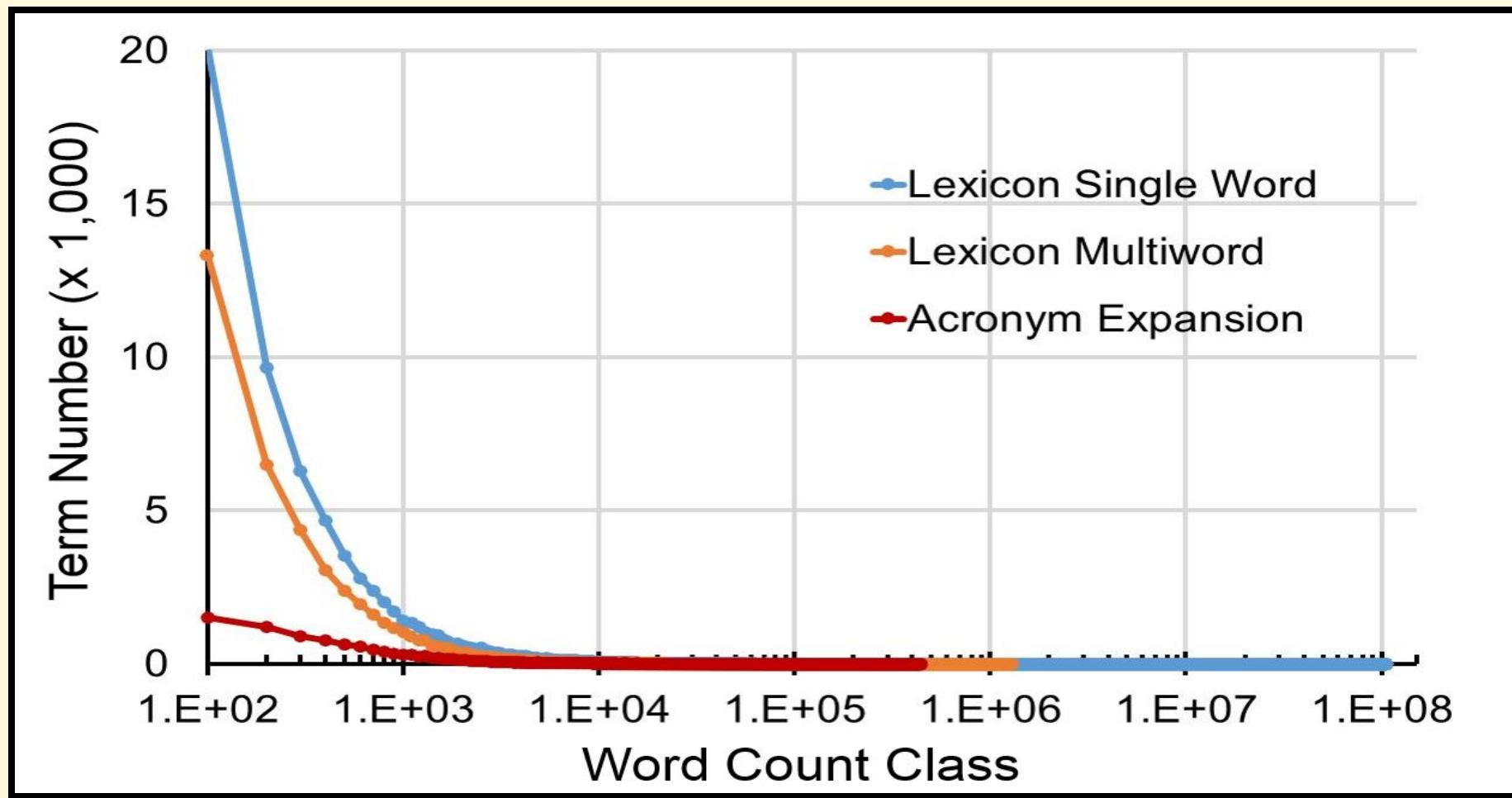
- Total word count for MEDLINE (2016): 3,114,617,940
- Lexicon covers > 98% unigrams from MEDLINE

Types	Word Count	Percentage %	Accu. %
LEXICON	2,911,156,308	93.4675%	93.4675%
NUMBER	8,753,120	0.2810%	93.7485%
DIGIT	145,548,882	4.6731%	98.4216%
NON-WORD*	19,148,557	0.6148%	99.0364%
NEW	30,011,073	0.9636%	100.0000%
Total	3,114,617,940		

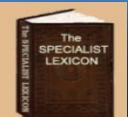
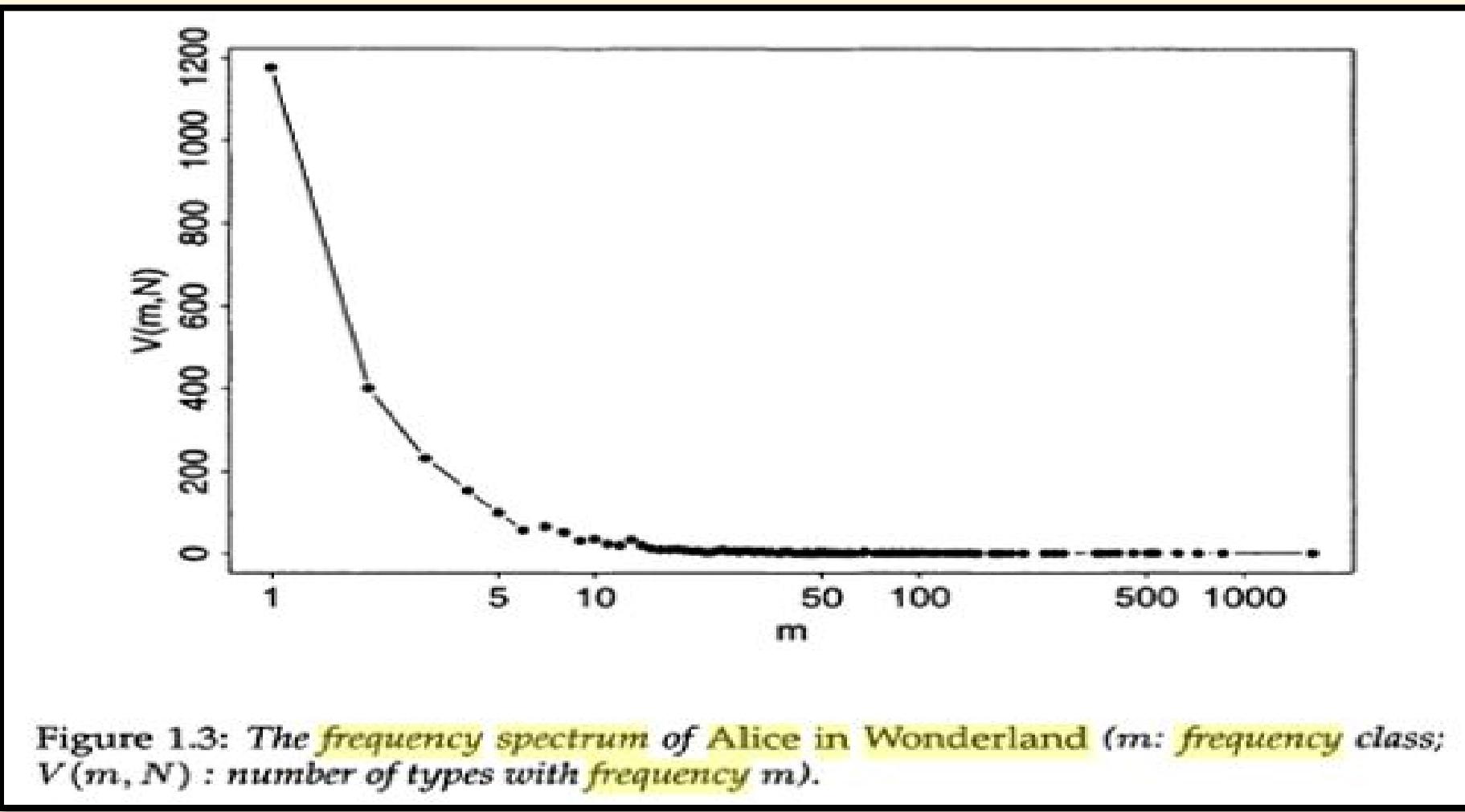
* NON-WORD: a single word only exist in multiword, such as “non”, “vitro”, “vivo”, “intra”, etc.



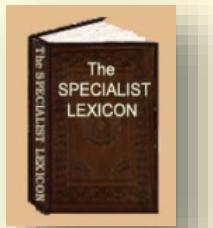
The Frequency Spectrum of Lexicon (Multi)words on MEDLINE



The Frequency Spectrum of Alice in Wonderland



Lexicon (Data) and Lexical Tools (Software)



```
{base=generalise  
spelling_variant=generalize-----> spelling variant  
entry=E0029526  
    cat=verb -----> part of speech  
    variants=reg -----> inflectional variant  
    intran  
    tran=np  
    tran=pphr(from,np) -----> chunker  
    tran=pphr(to,np)  
    nominalization=generalisation|noun|E0029525-----> derivational variant, synonym  
}
```



2. NLP - Lexical Tools

➤ Lexical Tools: Algorithm + Data (directly or derived from the Lexicon)

- Command line tools
 - lvg (Lexical Variants Generation, base of all of tools)
 - norm (UMLS - MRXNS, MRXNW)
 - luiNorm (UMLS - LUI)
 - wordInd (UMLS - MRXNW)
 - toAscii (MetaMap - BDB Tables)
 - fields (Lexicon Tables, MetaMap - BDB Tables, etc.)
- Lexical Gui Tool (lgt)
- Web Tools
- Java API's



Generated Lexical Variants

LexRecord: E0029526|generalise|verb

- POS: verb
- citation: generalise
- spVar: generalize
- inflVars: generalises, generalised, generalising
- nominalization: generalisation, generalization
- Abbreviation/acronym: n/a

Derivational variants:

- suffixD: generalisation, generalization, generalisable
- prefixD: overgeneralise, over-generalise

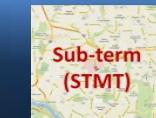
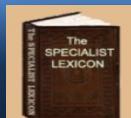
Synonyms: generalize

Fruitful Variants: generalisability, generalisable, generalisation, generalisations, generalised, generalises, generalising, generalizability, generalizable, generalization, generalizations, generalize, generalized, generalizer, generalizers, generalizes, generalizing, overgeneralize, etc.

← A LexRecord

← A LexRecord + Rules

← Multiple LexRecords + Rules



Lexical Tools - Facts

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



LVG - Lexical Variants Generation

- 62 flow components
 - base form
 - spelling variants
 - inflectional variants
 - derivational variants
 - acronyms/abbreviations
 - ...
- 34 options
 - input filter options (3)
 - global behavior options (12)
 - flow specific options (5)
 - output filter options (14)

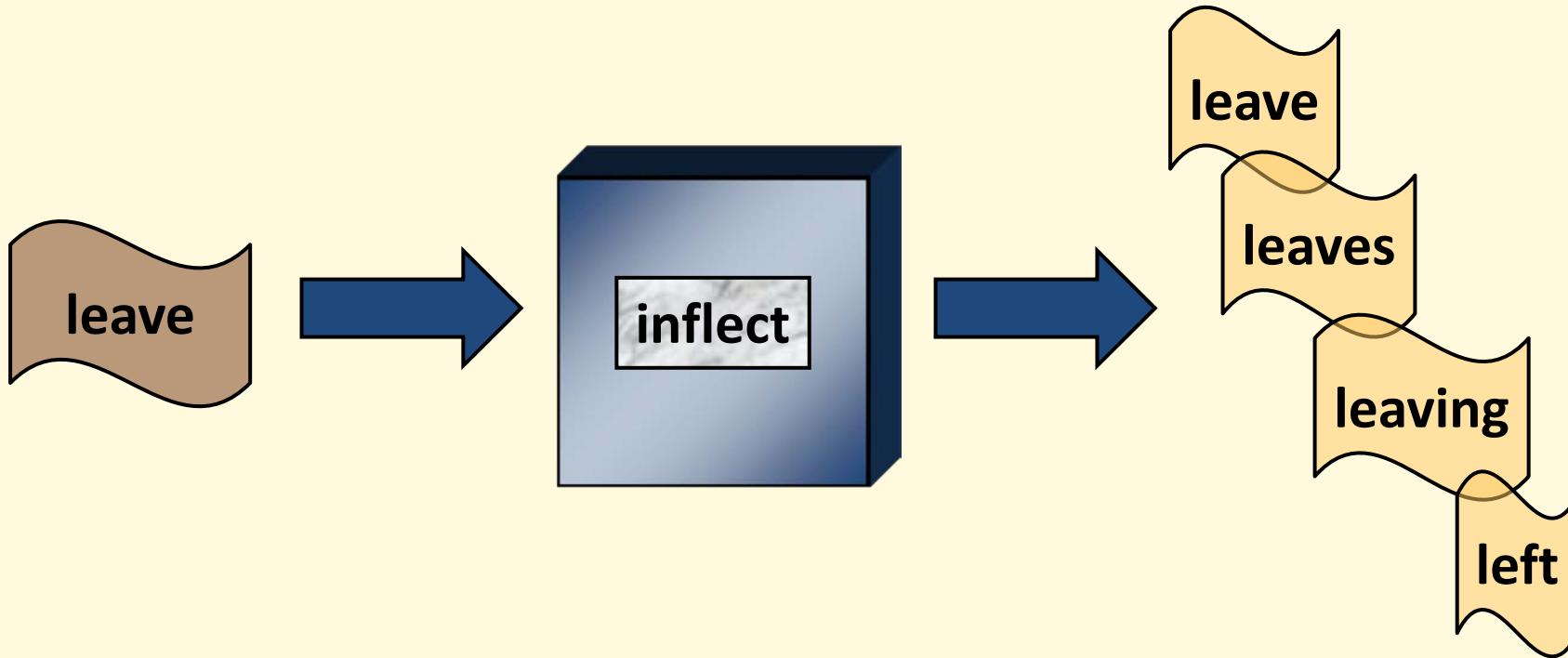


Lexical Tools – Flow Components (62)

Lexicon Related – Data (32)	Non-Lexicon related – Algorithm (30)
Inflection (10): b, B, Bn, l, ici, is, L, Ln, Lp, si,	Unicode operation (10): q, q0, q1, q2, q3, q4, q5, q6, q7, q8
Derivation (3): d, dc, R	Tokenizer (3): c, ca, ch
Acronym or abbreviation (3): a, A, fa	Punctuation operation (3): o, p, P
Spelling variant (2): e, s	Lowercase (1): l
Lexicon mapping (3): An, E, f, fp	Metaphone (1): m
Synonym (2): y, r	Remove parenthetic plural forms (1): rs
Nominalization (1): nom	Strip stop word (1): t
Citation (1): Ct	Remove genitive (1): g
Fruitful variant (4): G, Ge, Gn, V	No operation (1): n
Normalization (2): N, N3,	...



LVG Flow Component – Example

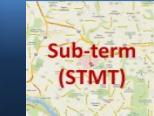


LVG Flow Component – Cmd line

```
> lvg -f:i
```

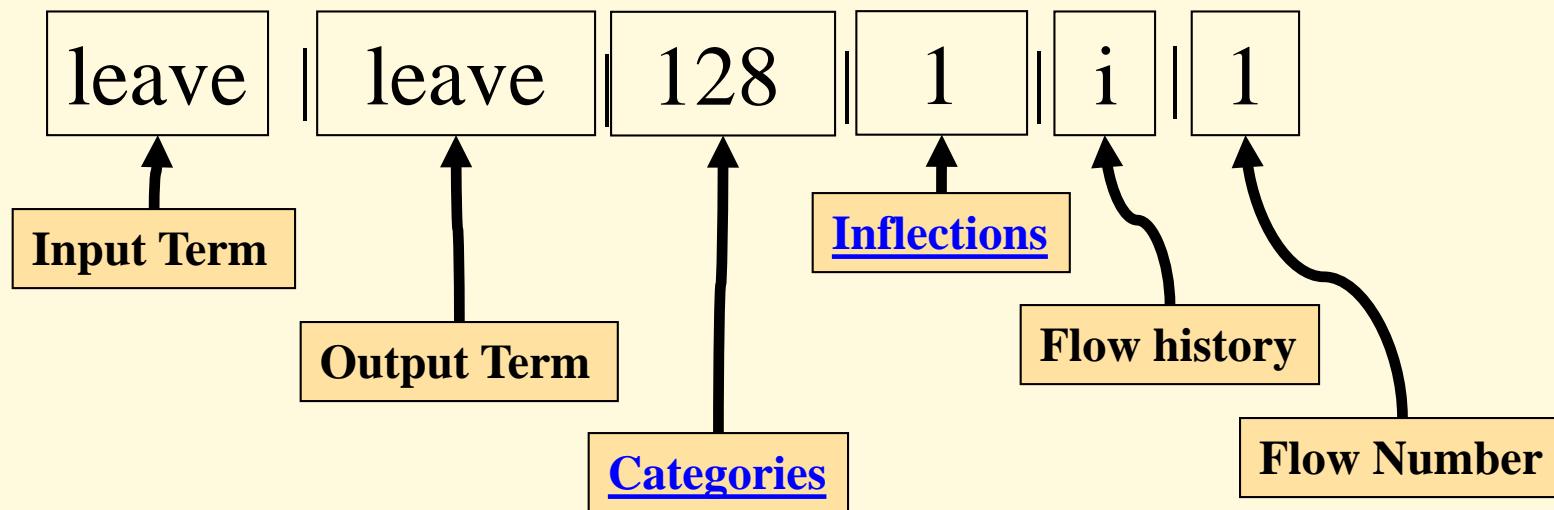
```
leave
```

```
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|
```

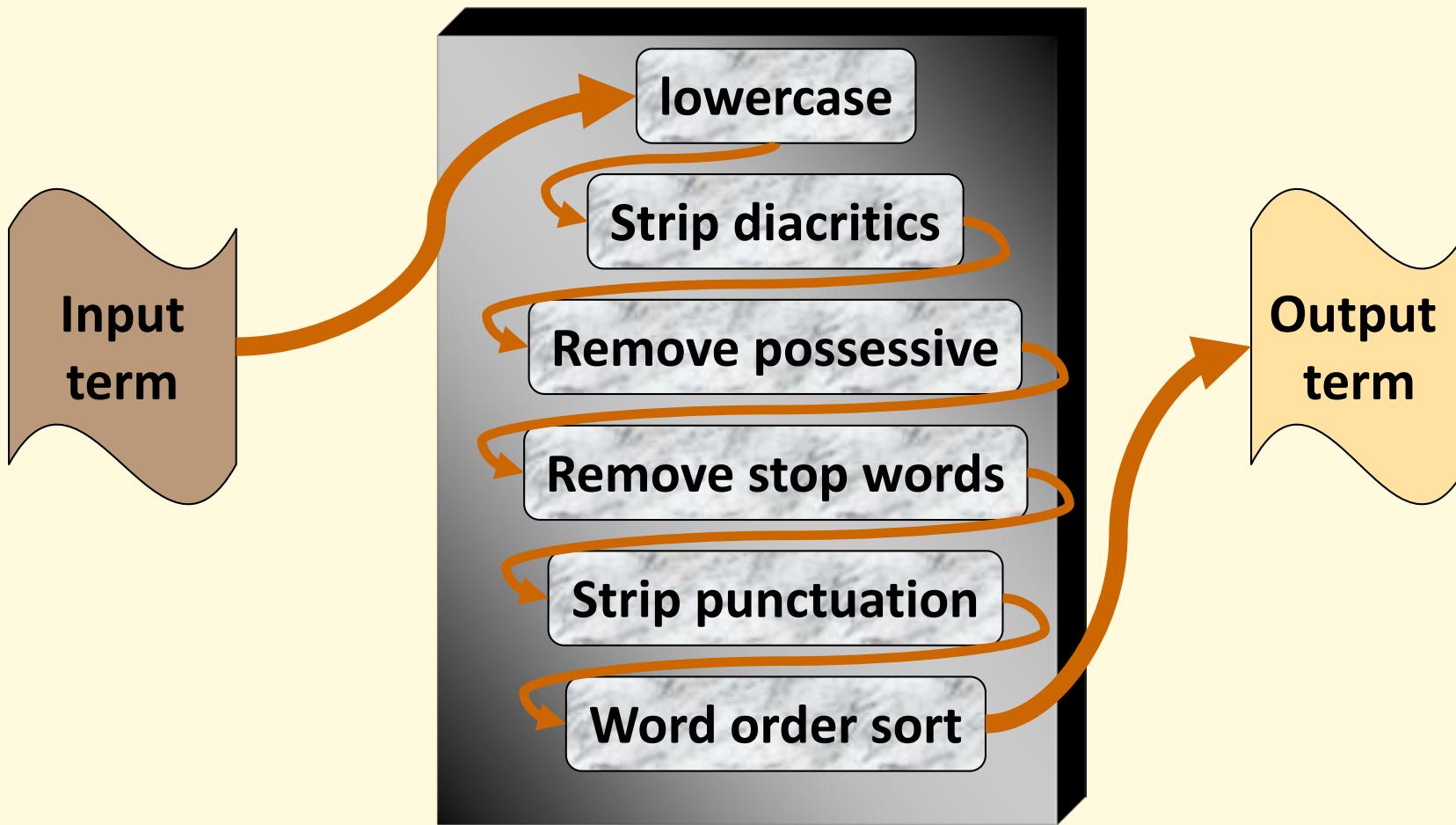


LVG Flow Component – Fielded Output

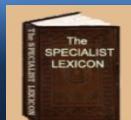
> lvg -f:i
leave



LVG – 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

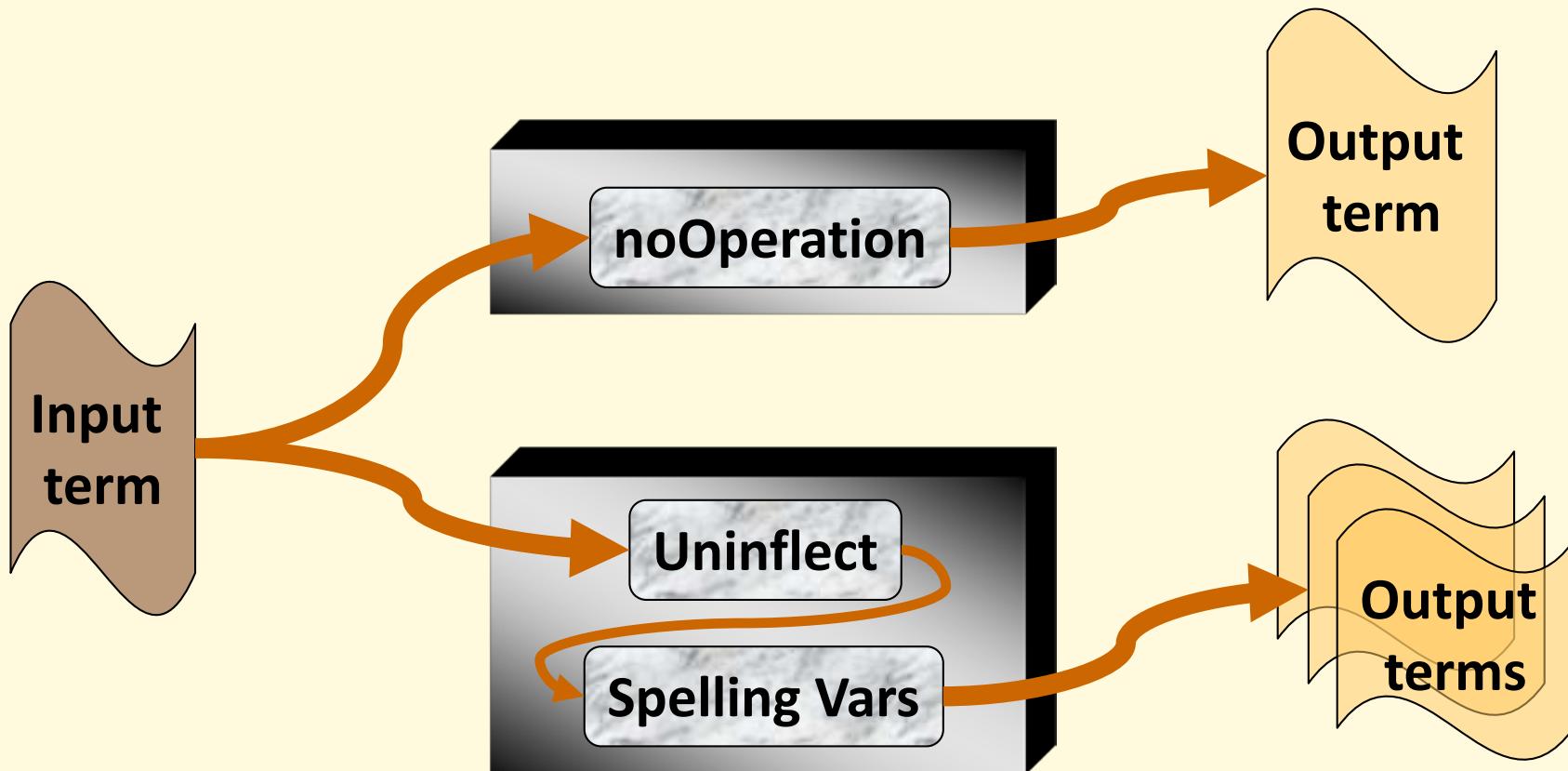
The **Gougerot-Sjögren's Syndrome**

gougerotsjogren syndrome | 2047 |

16777215 | **l+q+g+t+p+w** | 1 |



LVG - 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|

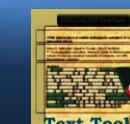
color|colour|128|1|B+s|2|

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



Norm (commonly used flow)

- Composed of 11 Lvg flow components to abstract away from (only keep meaningful words):
- case
 - punctuation
 - possessive forms
 - inflections
 - spelling variants
 - stop words
 - diacritics & ligatures (non-ASCII Unicode)
 - word order



Norm

“Fœtoproteins α's, NOS”

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: uninfect 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: uninfect each words in a term

Ct: retrieve citations

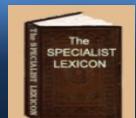
q7: Unicode core Norm

q8: strip or map Unicode to ASCII

w: sort words by order

“Fœtoproteins α’s, NOS“

"Fœtoproteins α's, NOS"



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: uninfect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

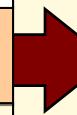
q8: strip or map Unicode to ASCII

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: uninfect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map Unicode to ASCII

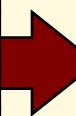
w: sort words by order

"Fœtoproteins α's, NOS"

"Fœtoproteins α's, NOS"

"Fœtoproteins α, NOS"

"Fœtoproteins α, NOS"



Norm

q0: map symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninfect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map Unicode to ASCII

w: sort words by order

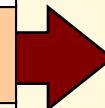
"Fœtoproteins α's, NOS"

"Fœtoproteins α's, NOS"

"Fœtoproteins α, NOS"

"Fœtoproteins α, NOS"

Fœtoproteins α NOS



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: uninfect 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"

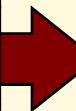
"Fœtoproteins α's, NOS"

"Fœtoproteins α, NOS"

"Fœtoproteins α, NOS"

Fœtoproteins α **NOS**

Fœtoproteins α



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: uninfect 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"

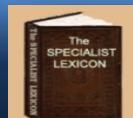
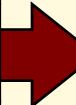
"Fœtoproteins α, NOS"

"Fœtoproteins α, NOS"

Fœtoproteins α NOS

Fœtoproteins α

fœtoproteins α



Norm

q0: map symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninfect 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"

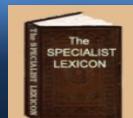
"Fœtoproteins α, NOS"

Fœtoproteins α NOS

Fœtoproteins α

fœtoproteins α

fœtoprotein α



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: uninfect 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"

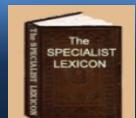
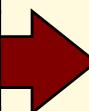
Fœtoproteins α NOS

Fœtoproteins α

fœtoproteins α

fœtoprotein α

fetoprotein α



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: uninfect 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

fœtoprotein α

fetoprotein α

fetoprotein α



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: uninfect 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

fœtoprotein α

fetoprotein α

fetoprotein α

fetoprotein alpha



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: uninfect 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

fœtoprotein α

fetoprotein α

fetoprotein α

fetoprotein alpha

alpha fetoprotein



Norm

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



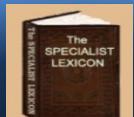
3. Natural Language Processing (NLP)

➤ 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
- NLP includes a board range of subjects, require knowledge from linguistics, computer science, and statistics.
- NLP in our scope is to use computer to understand the meaning (concept) from text for further analysis and processing.



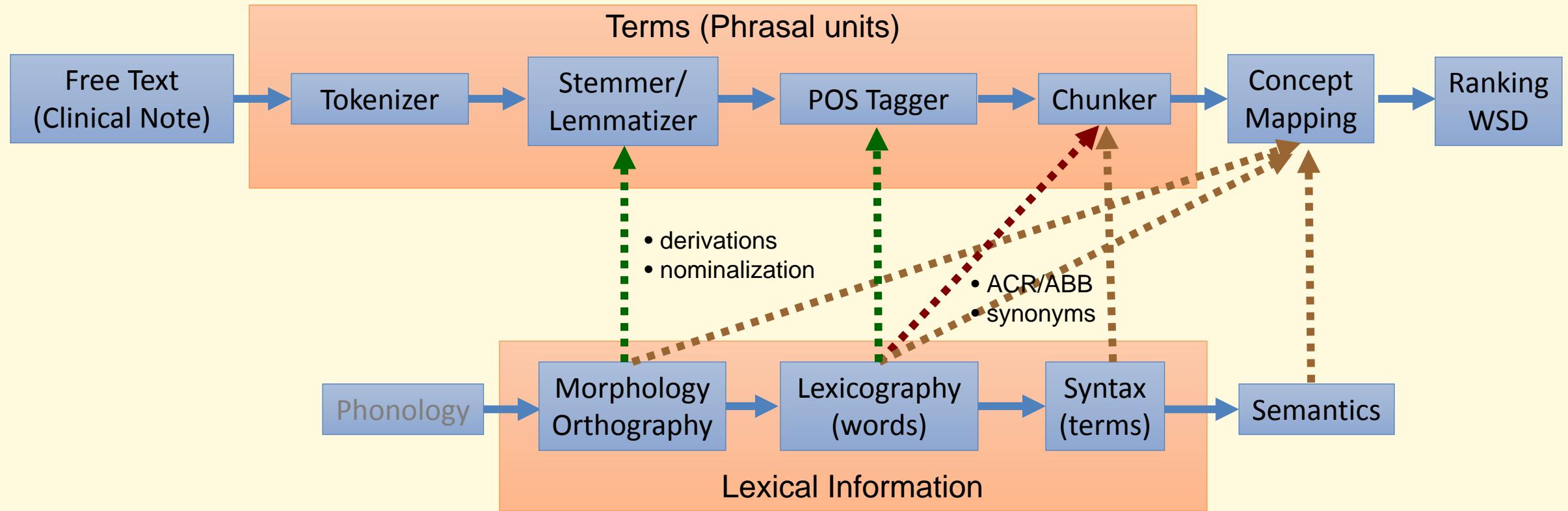
Concept Mapping Challenges

- Challenge 1: Map terms to concepts (meaning)
- Challenge 2: many to many mapping

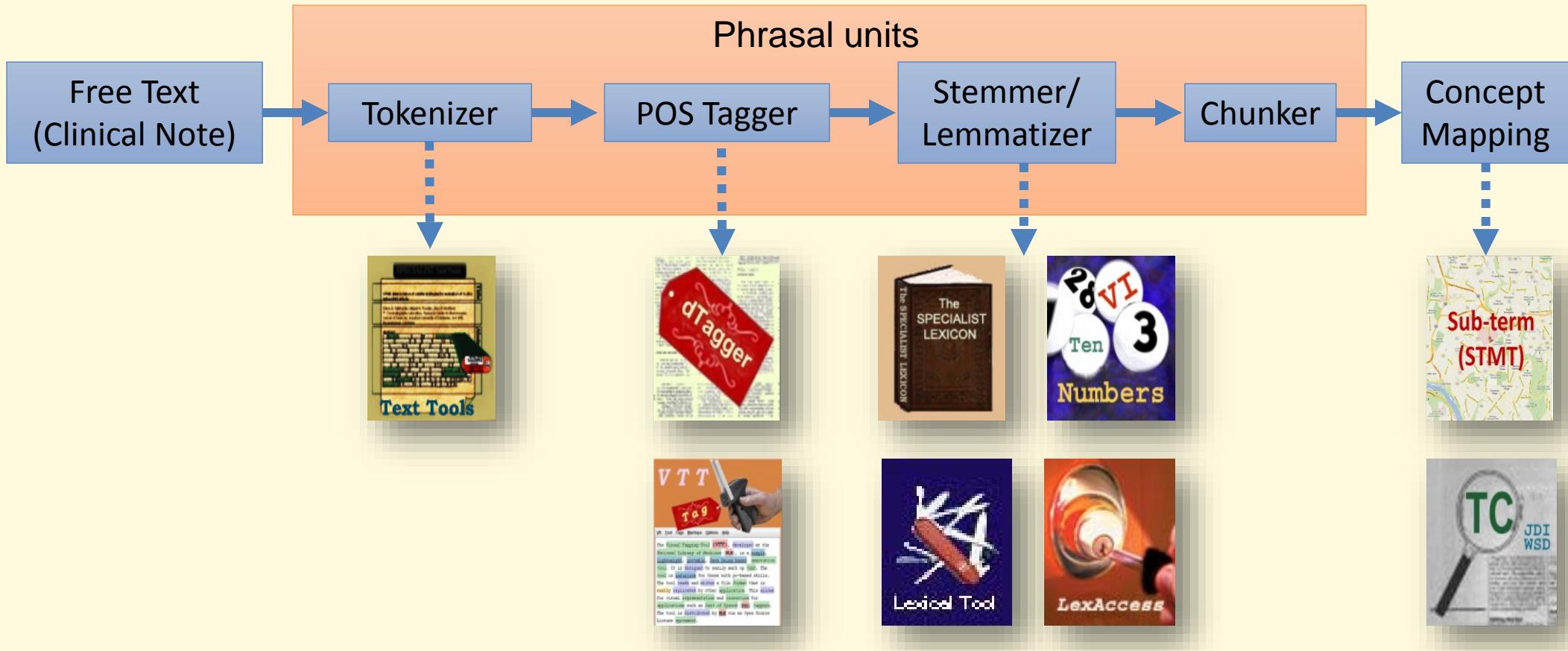
Terms	Concepts	NLP
<ul style="list-style-type: none">• cold• Cold Temperature• Cold Temperatures• Cold (Temperature)• Temperatures, Cold• Low temperature• low temperatures• ...	<ul style="list-style-type: none">• Cold Temperature C0009264	<ul style="list-style-type: none">• Concept mapping
<ul style="list-style-type: none">• cold	<ul style="list-style-type: none">• Cold Temperature C0009264• Common Cold C0009443• Cold Therapy C0010412• Cold Sensation C0234192• ...	<ul style="list-style-type: none">• WSD (Word Sense Disambiguation)



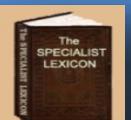
NLP Pipe Line – Lexical Information



The SPECIALIST NLP Tools



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



NLP – Concept Mapping

➤ Normalization (same record):

- A term might have a great deal of lexical variations, such as inflectional variants, spelling variants, abbreviations (expansions), cases, ASCII conversion, etc.
- Normalize different forms of a concept to a same form

➤ Query Expansion (related records):

- Expand a term to its equal terms, such as subterm substitution of synonyms, derivational variants, abbreviations, etc.
- To increase recall

➤ POS tagger:

- Assign part of speech to a single word or multiword in a text
- To increase precision

➤ Others...



Lexical Tools – 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: uninfect each words in a term](#)

[Ct: retrieve citations](#)

[q7: Unicode core Norm](#)

[q8: strip or map non-ASCII char](#)

[w: sort words by order](#)

Behçet's Diseases, NOS

Behçet's Diseases, NOS

Behçet Diseases, NOS

Behçet Diseases, NOS

Behçet Diseases NOS

Behçet Diseases

behçet diseases

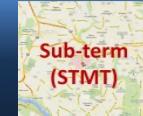
behçet disease

behcet disease

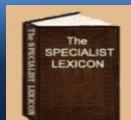
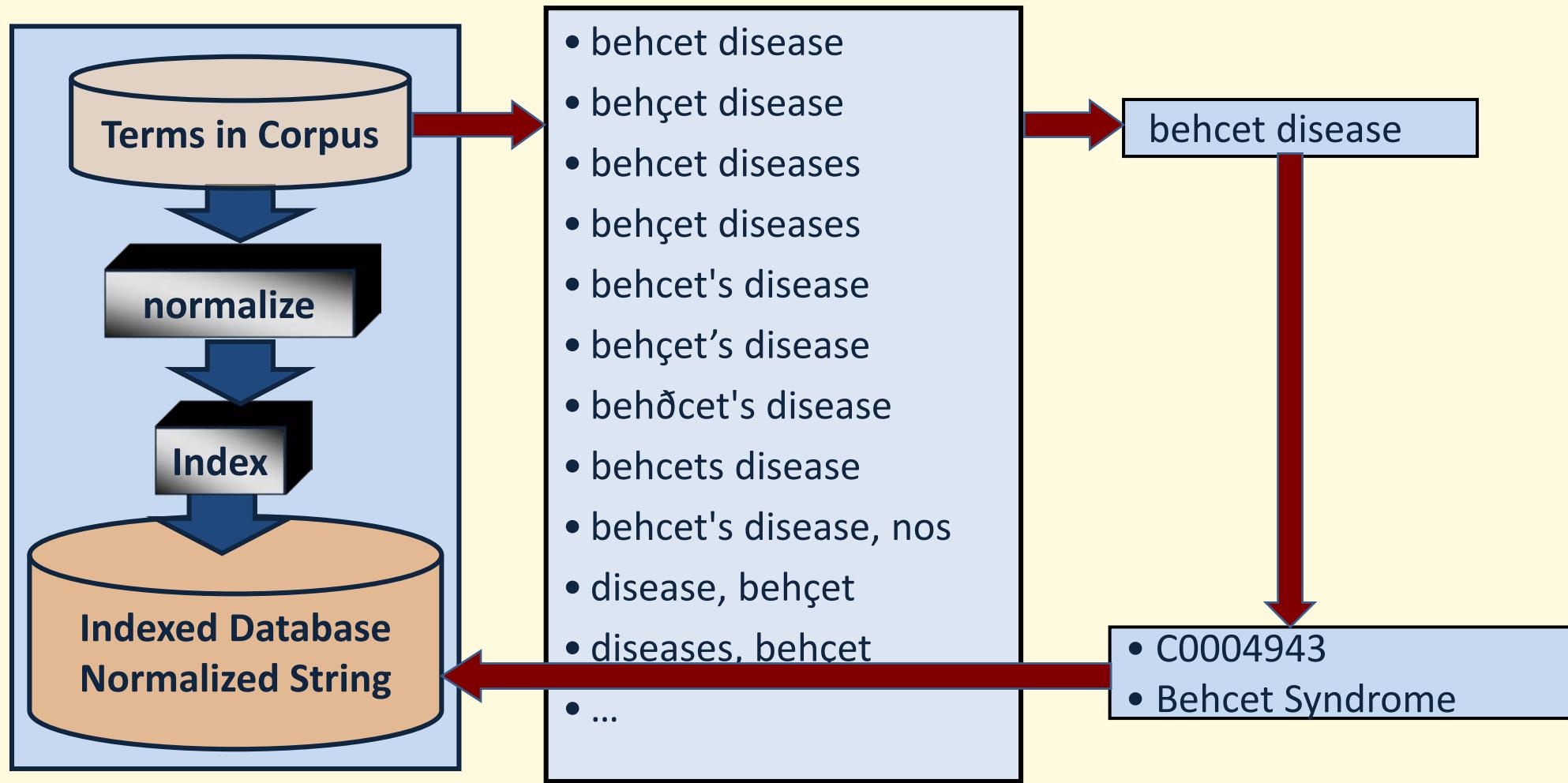
behcet disease

behcet disease

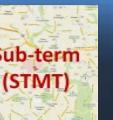
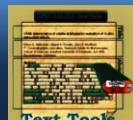
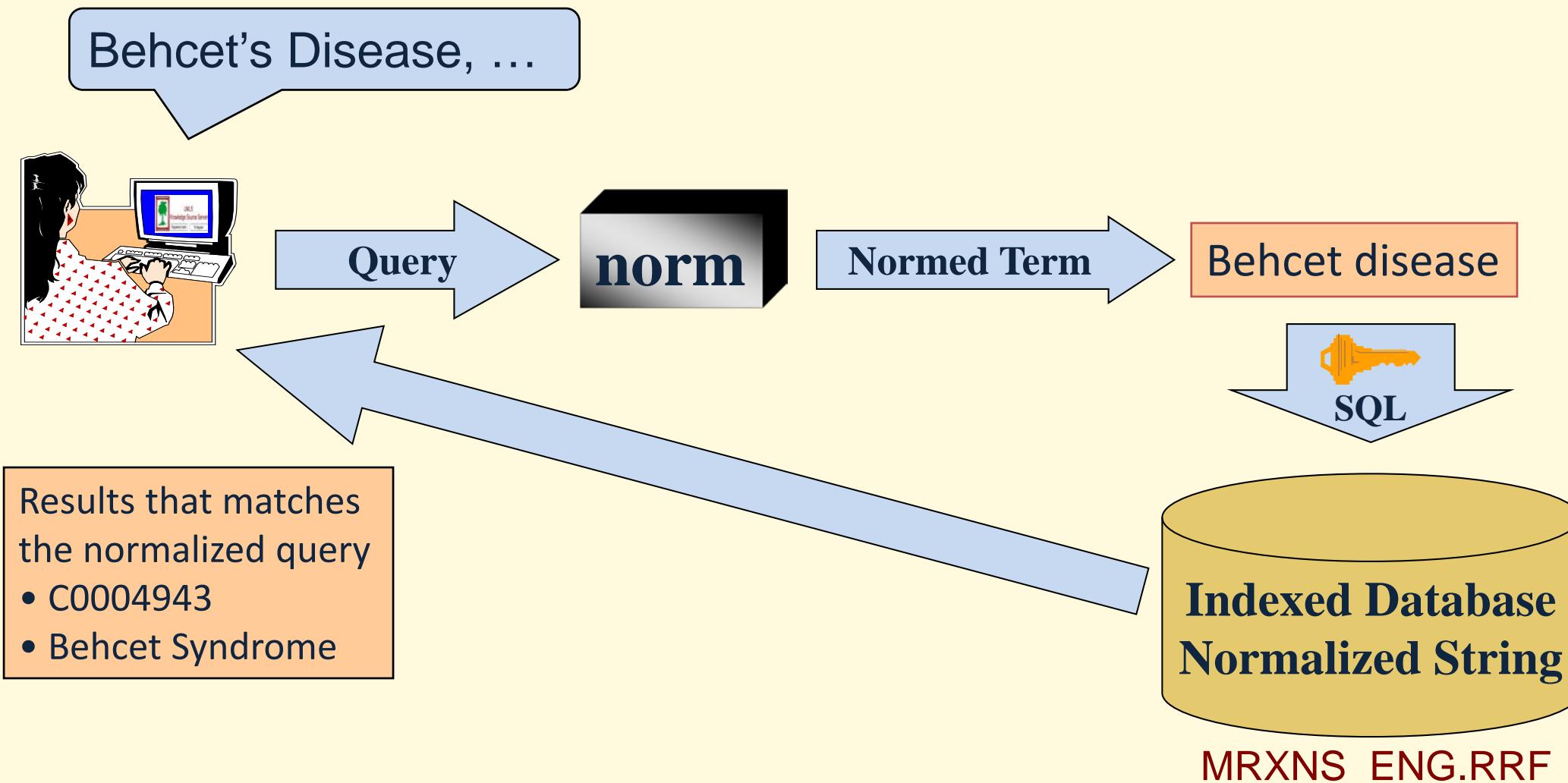
behcet disease



NLP – Norm (Pre-Process Lexical Variations)



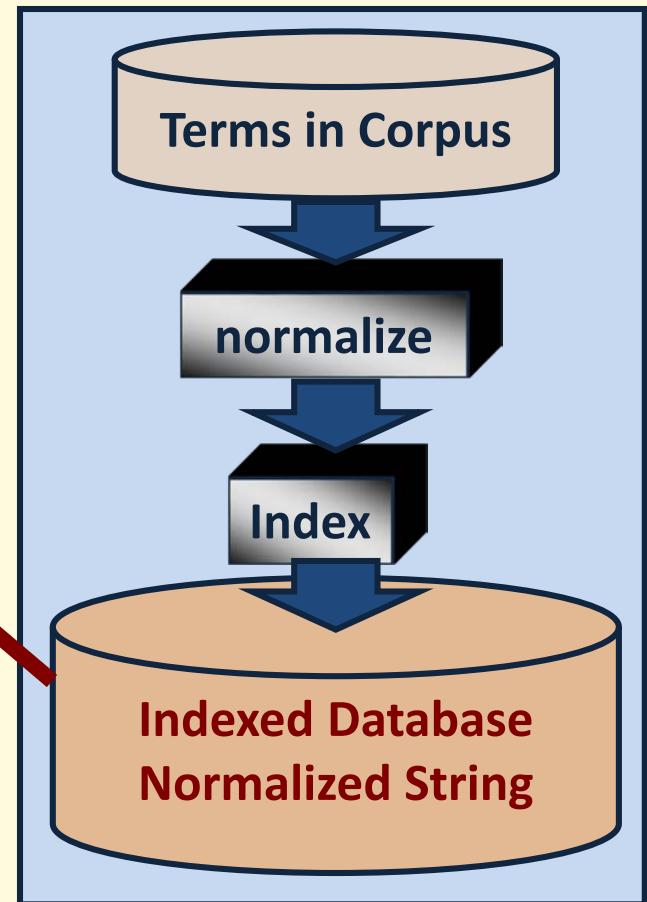
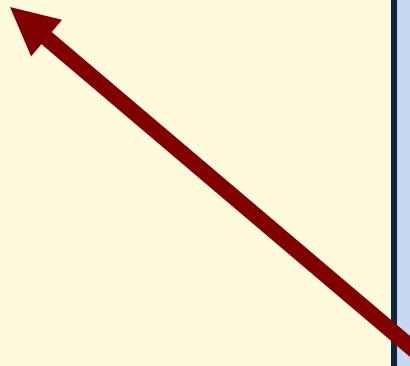
NLP – Norm (Cont.)



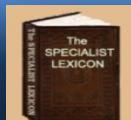
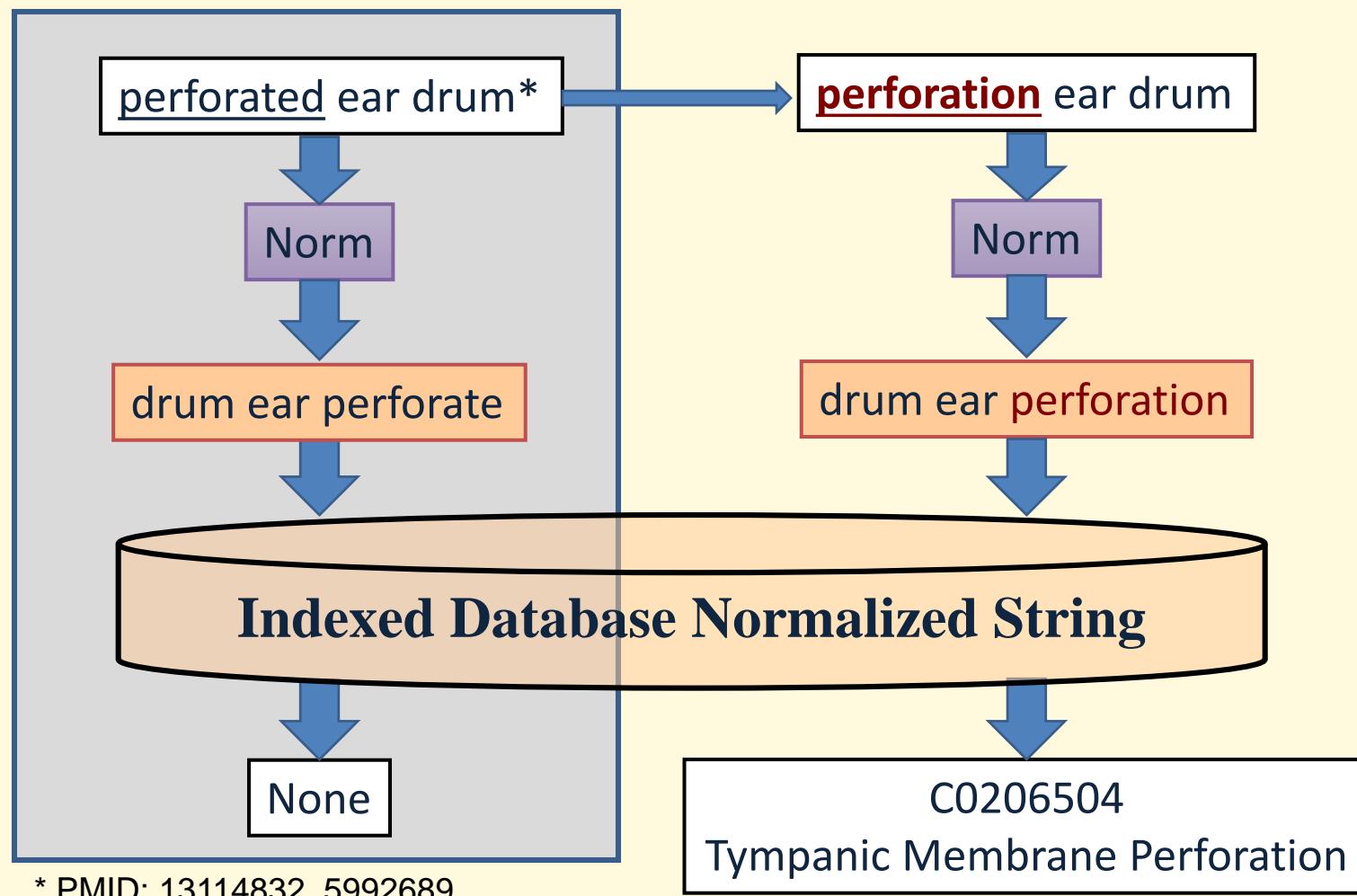
UMLS Metathesaurus

➤ UMLS Normalized Files

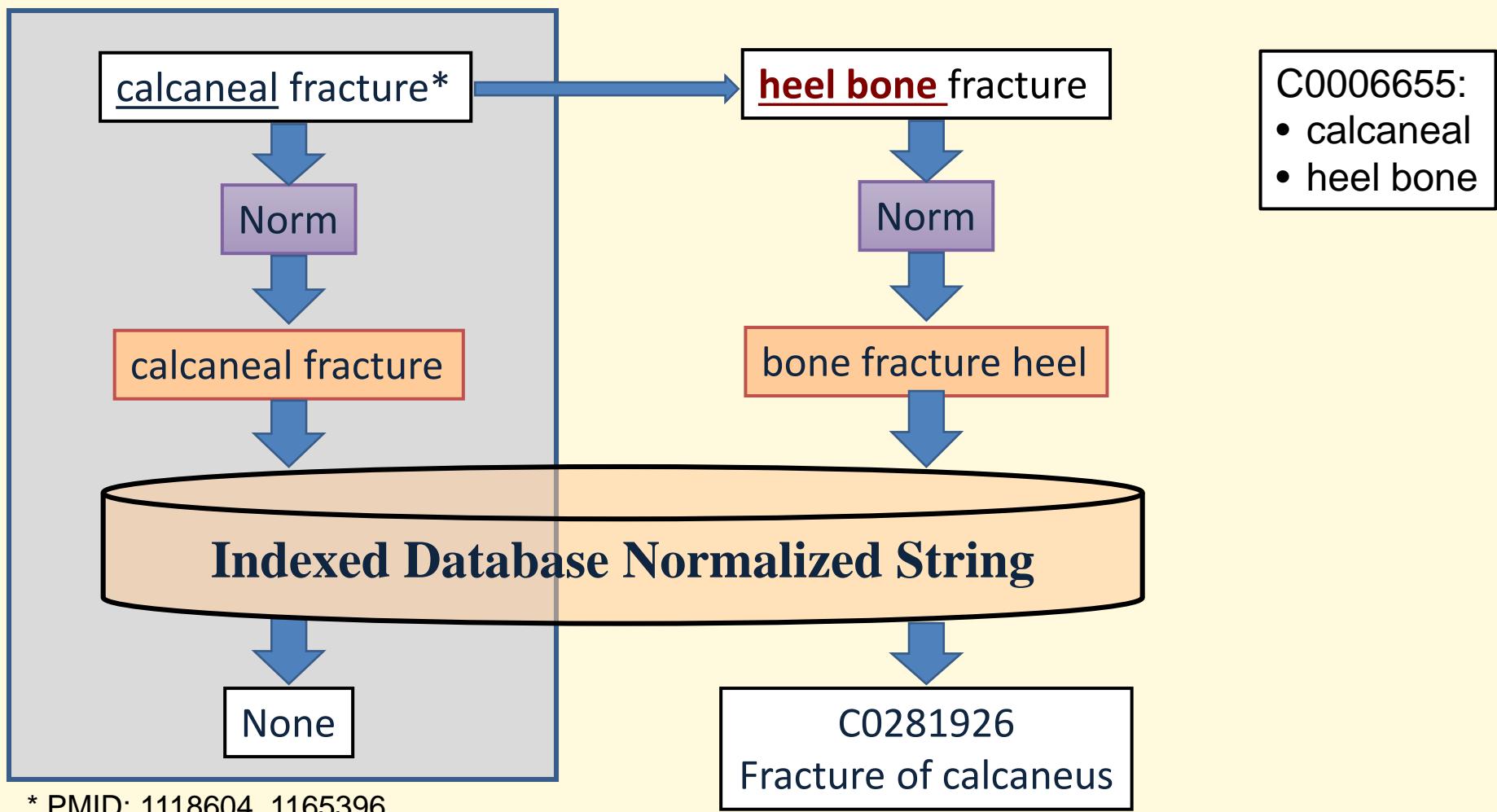
- Normalized words: MRXNW_ENG.RRF
- Normalized strings: MRXNS_ENG.RRF



NLP – Query Expansion (derivation)



NLP – Query Expansion (Synonym)

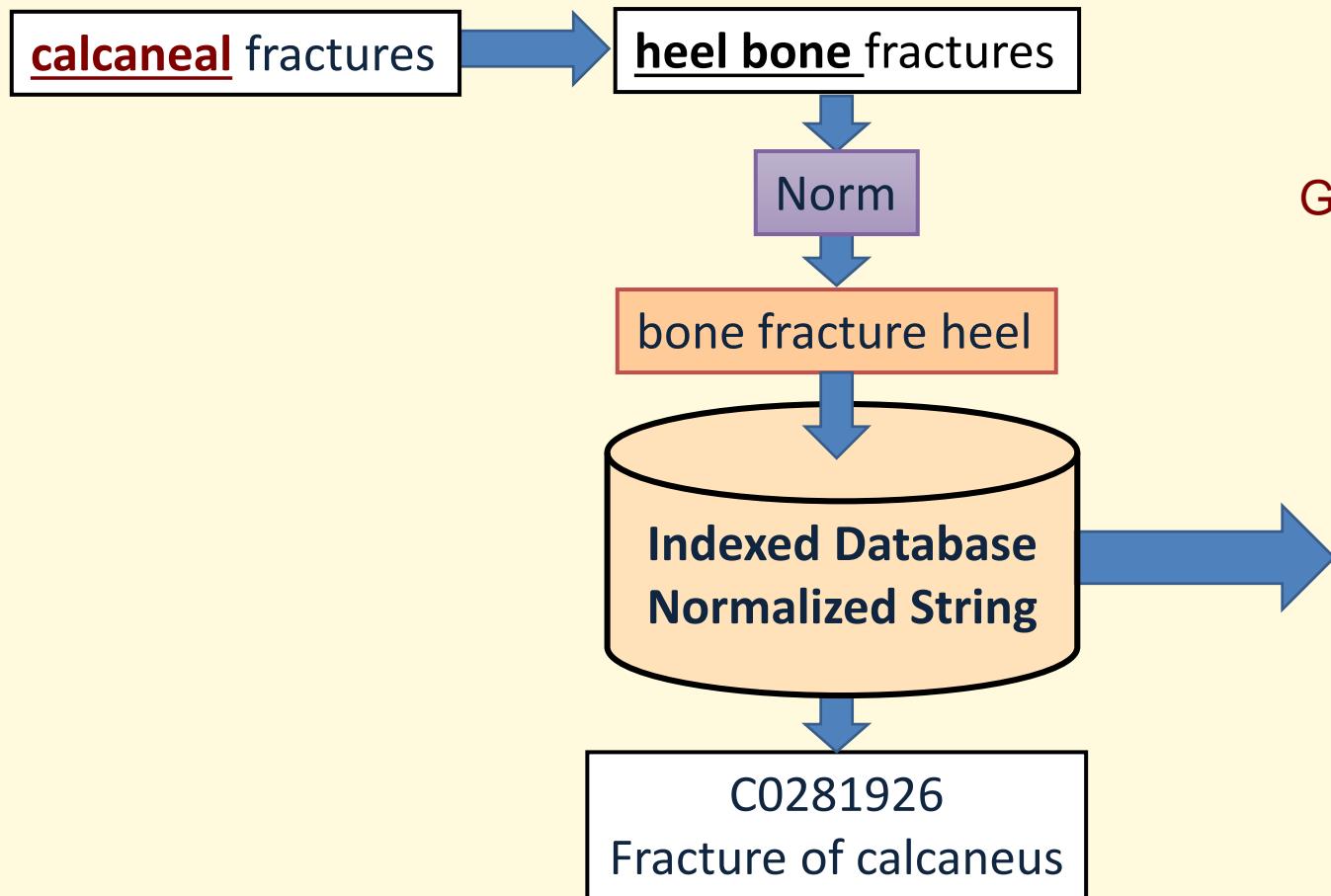


UMLS Synonymy (C0281926)

calcaneus fracture
calcaneus fractures
calcaneus; fracture
fracture calcaneus
fracture heel
fracture of calcaneus
fracture of calcaneus (diagnosis)
fracture of calcaneus (disorder)
fracture of os calcis
fracture; calcaneus
fracture; heel bone
fracture; os calcis
fracture;calcaneus
fractured calcaneus
fractured os calcis
fractures heel
heel bone
heel bone fracture
heel bone; fracture
heel fracture
of calcaneus fracture
os calcis
os calcis fracture
os calcis; fracture



UMLS Synonymy – Expanded Terms



[UMLS Synonymy]
Expanded Terms for
Concept Mapping:
Grouped by Normalization

C0281926:

- Key | calcaneus fracture
 - fractured calcaneus
 - fracture; calcaneus
 - fracture of calcaneus
 - calcaneus fracture
 - calcaneus fractures
 - calcaneus; fracture
- Key | bone fracture heel
 - heel bone fracture
 - heel bone; fracture
 - fracture; heel bone
 - ...



UMLS Synonym to Element Synonym

calcaneus fracture
calcaneus fractures
calcaneus; fracture
fracture calcaneus
fracture heel
fracture of calcaneus
fracture of calcaneus (diagnosis)
fracture of calcaneus (disorder)
fracture of os calcis
fracture; calcaneus
fracture; heel bone
fracture; os calcis
fracture;calcaneus
fractured calcaneus
fractured os calcis
fractures heel
heel bone
heel bone fracture
heel bone; fracture
heel fracture
of calcaneus fracture
os calcis
os calcis fracture
os calcis; fracture

Norm: **calcaneus fracture**

calcaneus fracture
calcaneus fractures
calcaneus; fracture
fracture calcaneus
fracture of calcaneus
fracture of calcaneus (diagnosis)
fracture of calcaneus (disorder)
fracture; calcaneus
fracture;calcaneus
fractured calcaneus
of calcaneus fracture

Norm: **bone fracture heel**

heel bone fracture
fracture; heel bone
heel bone; fracture

Norm: **calcis fracture os**

fracture of **os calcis**
fracture; os calcis
fractured os calcis
os calcis fracture
os calcis; fracture

Norm: **fracture heel**

fracture heel
fractures heel
heel fracture

heel bone
os calcis

- Other element Synonym
- **calcaneal** fracture – PMID: 1194000, 471457, ...
- **calcaneum** fracture – PMID: 13288374, 5550125, ...

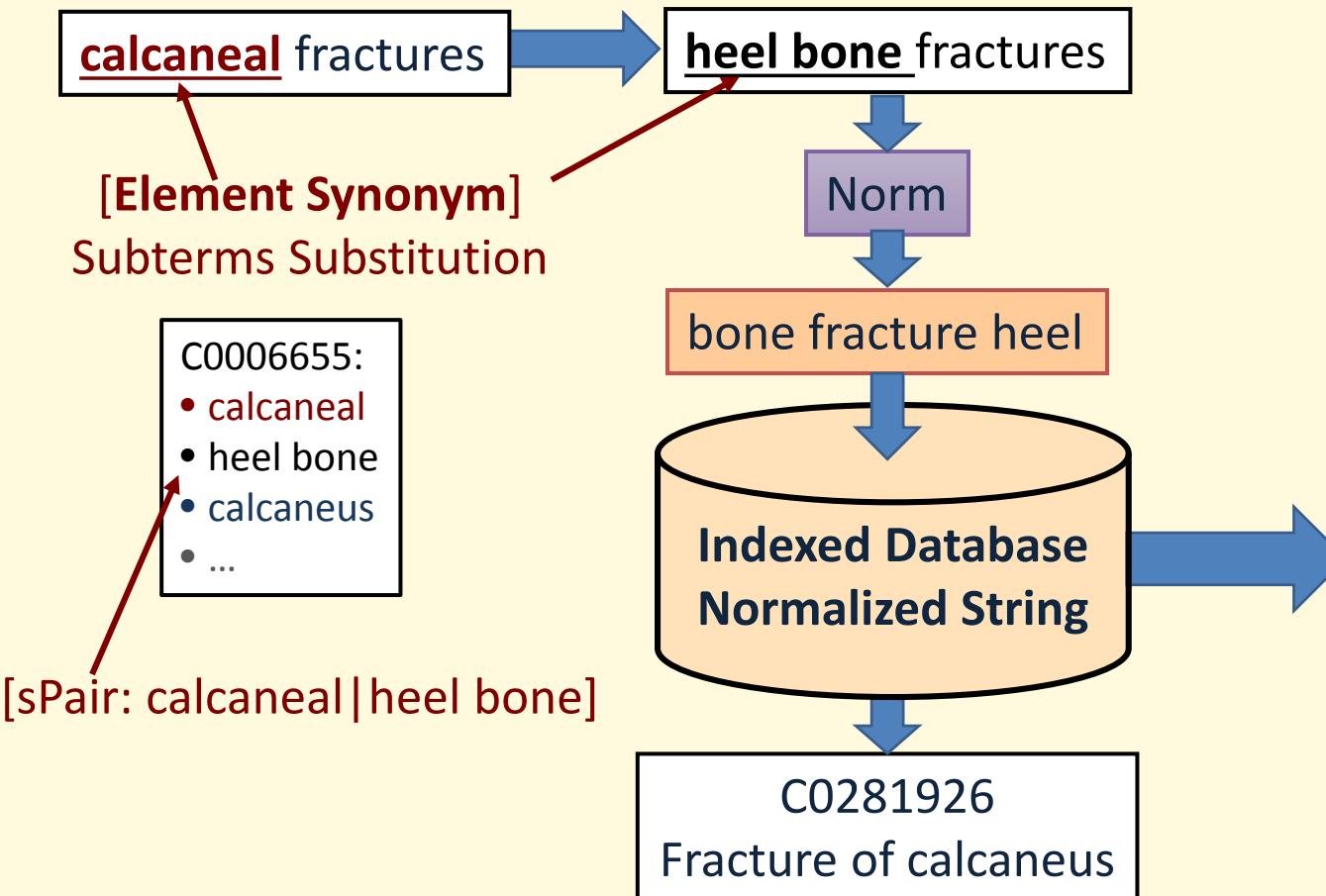


C0006655:

- calcaneal
- calcaneum
- calcaneus
- heel bone
- os calcis
- ...

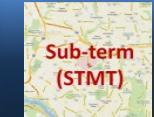
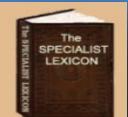


Element Synonyms

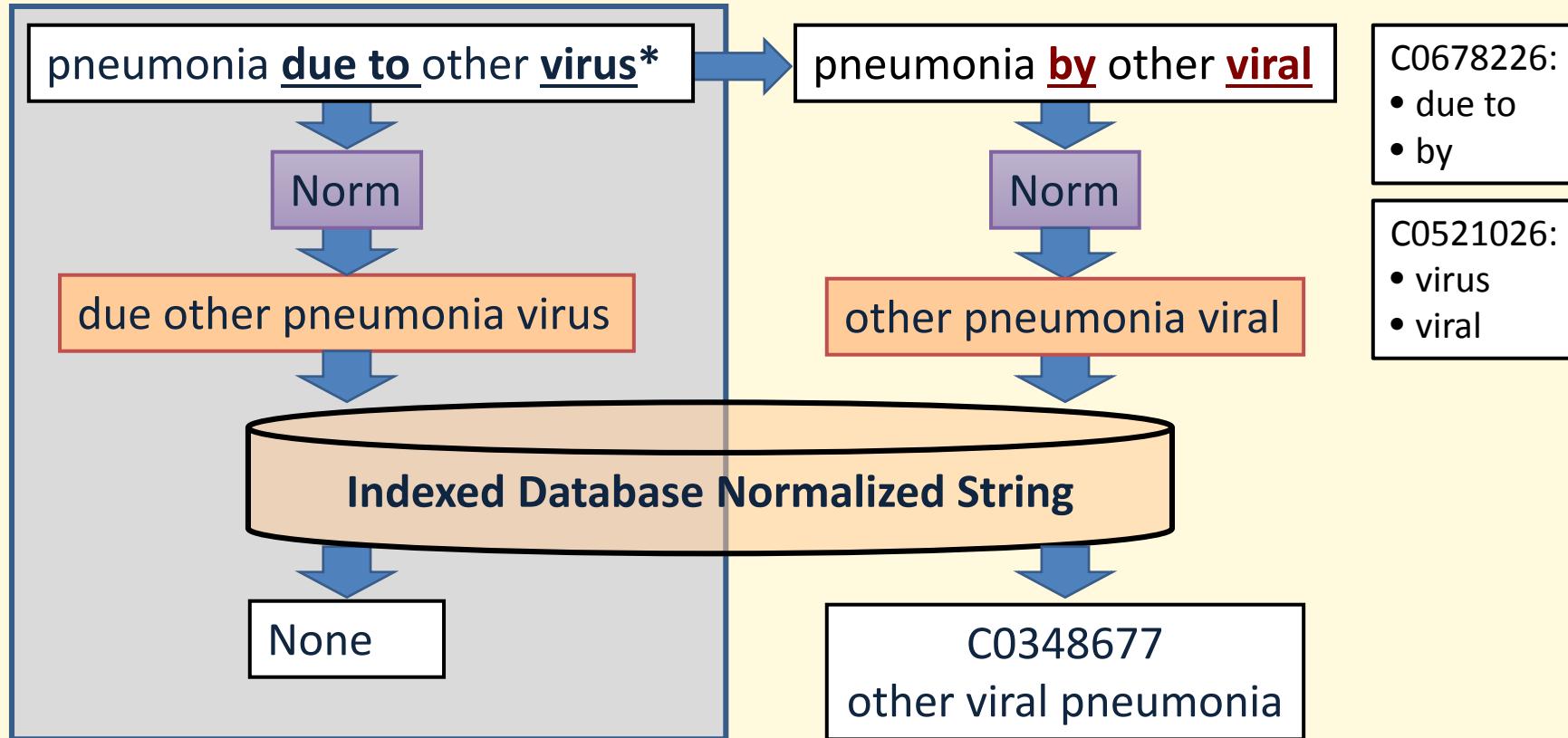


C0281926:

- Key | calcaneus fracture
 - fractured calcaneus
 - fracture; calcaneus
 - fracture of calcaneus
 - calcaneus fracture
 - calcaneus fractures
 - calcaneus; fracture
- Key | bone fracture heel
 - heel bone fracture
 - heel bone; fracture
 - fracture; heel bone
 - ...



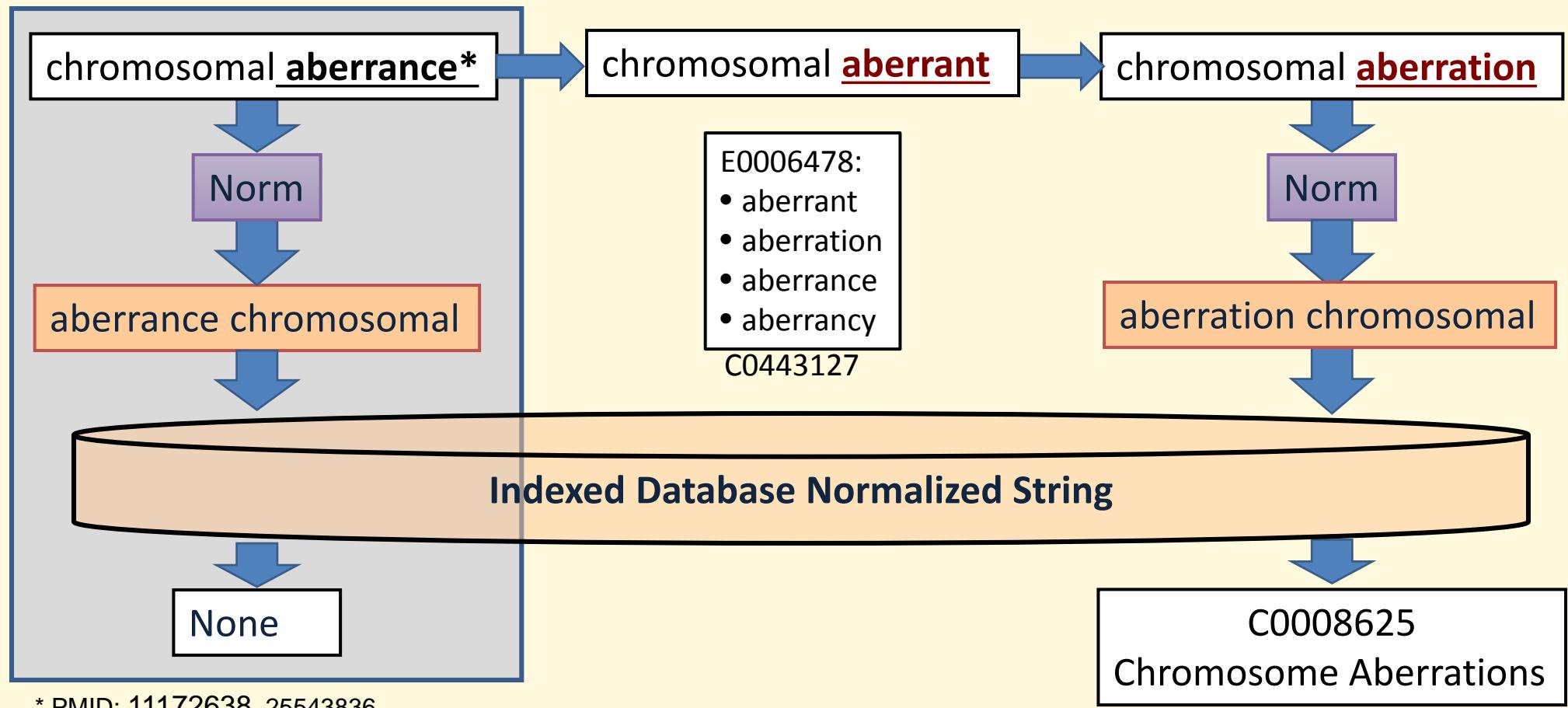
Multiple Substitutions



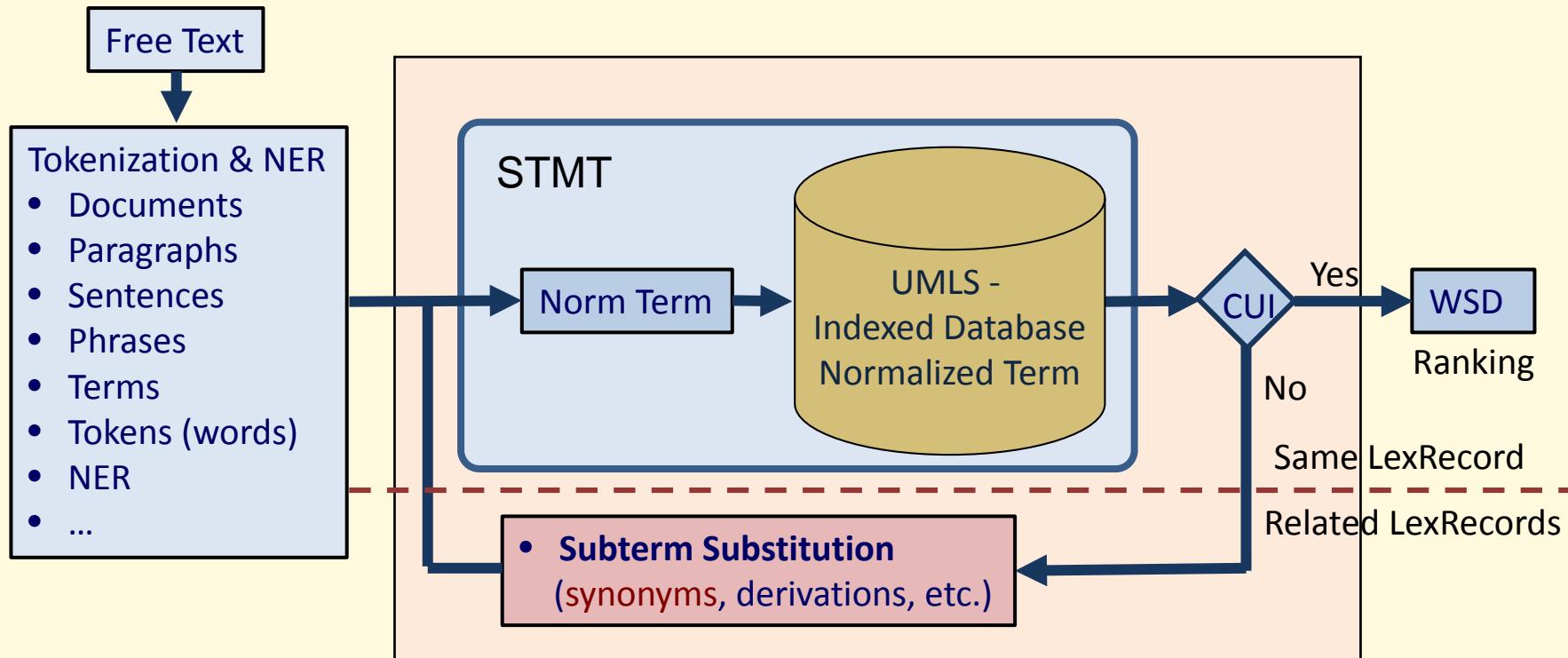
* VA14760, HA480.80, ..



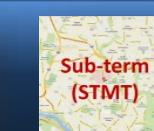
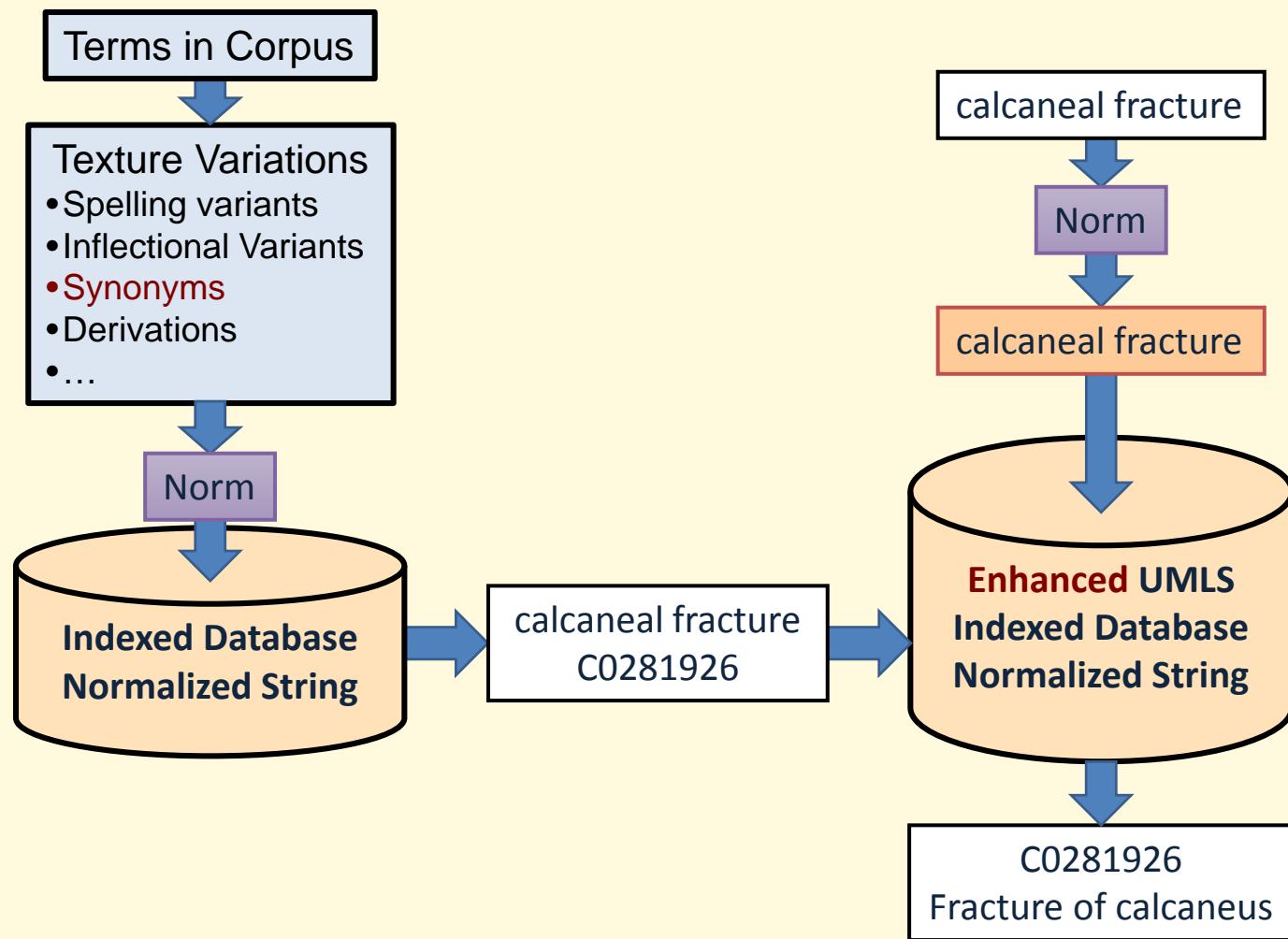
Recursive Substitutions



Real-time Model

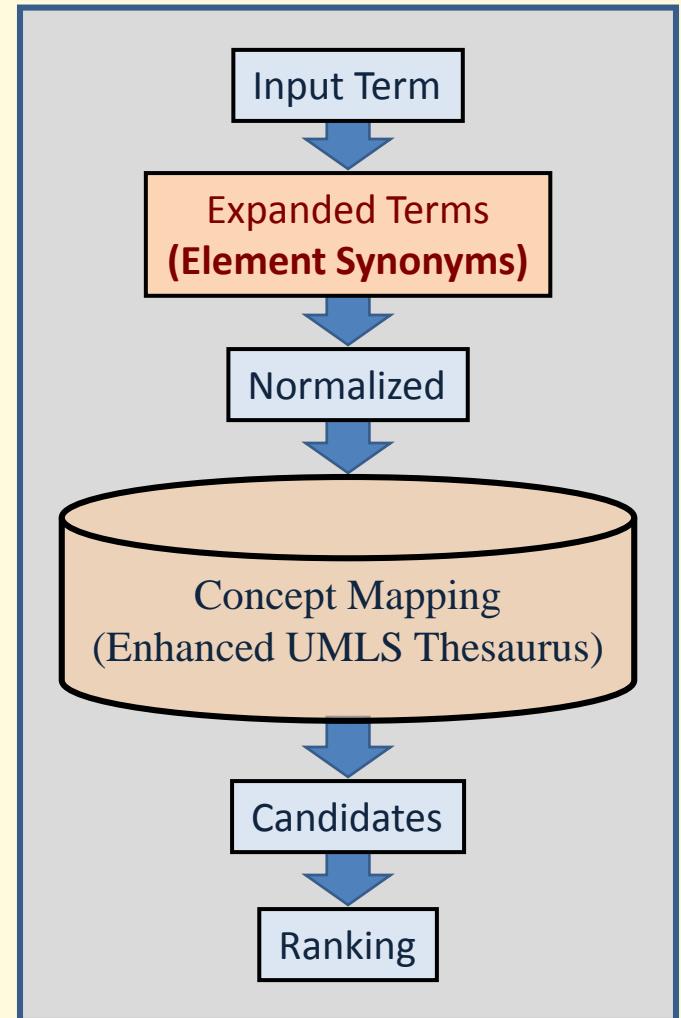


Pre-Processing Model



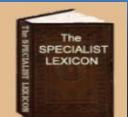
4. LexSynonym - Element Synonyms

- The key for subterm substitutions (data of synonyms) depends on the completeness and quality of both element synonyms for a given UMLS synonym thesaurus.
- Synonym Related Data:
 - Element Synonyms (for expanded terms)
 - UMLS Synonym thesaurus (for concept mapping)
- Completeness: recall
- Quality: precision



Synonym Sets

- UMLS Synonyms (13M)
- The SPECIALIST Lexicon Synonyms, 2016- (~5K)
- Others
 - UMLS-Core Projects (~12K)
 - Synonym set by Randy Miller, (~15K)
 - dictionary.com, thesaurus.com,
 - WordNet (<https://wordnet.princeton.edu>)
 - etc..



Element Synonyms - UMLS Synonyms

- Applied restrictions: source vocabulary (MeSH), term length, size of grams (1), etc..
- Issues:
 - Quantity (over-generated):
 - Example: [C0013182, Drug Allergy], “allergy drug” and “allergy medicine” (expanded terms)
 - Slow performance (if use all expanded terms for element synonyms)
 - Quality:
 - Not necessary cognitive synonyms (commutativity and transitivity)
 - Broader or narrower concept, acronyms, abbreviations, POS ambiguity, multiple CUIs, etc..
 - Single words or multiwords
 - Example: [C0281926, Fracture of calcaneus], “calcaneal fracture” and “heel bone fracture”
 - How many grams?



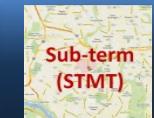
Element Synonyms – Lexicon Synonyms

- Developed in early 90's
- The original idea is to provide synonyms that are not in the UMLS Metathesaurus
 - not a complete data set
- Quantity: manually updated by user's requests (static):
 - 2004 (5,056) -> 2016 (5,198)
 - Only 142 sPairs were added since 2004
 - Need an automatic/systematic way to generate synonyms
- Quality: not necessary good sPairs
- 6 associated flow components (10%): G, Ge, Gn, r, v, y



LexSynonyms – Objectives

- To establish a system to generate a standalone set of generic element synonyms (sPairs) for effective UMLS concept mapping
 - Scope:
 - include all synonymous terms in Lexicon (LexSynonyms)
 - grow with the SPECIALIST Lexicon
 - a thorough set of element synonyms (to increase recall)
 - Feature requirements:
 - **better performance: increase recall and preserve precision**
 - resolve known issues (near-synonyms, POS ambiguity, include multiword synonyms, etc.)
 - cognitive synonyms (to preserve precision)



Enhanced Requirements

- Element synonyms for subterm substitution
 - R1: Cognitive synonyms (not near-synonyms)
 - R2: POS (meaning shift)
 - R3: Source: CUI (UMLS) and other source information
 - R4: Expansions of abbreviations and acronyms
 - R5: Word level (single POS): single words and multiwords
- ...



R1: Cognitive Synonym (Quality)

- Two properties:
 - **Commutativity:** $(x = y) \rightarrow (y = x)$
 - joy|noun|enjoy|verb -> enjoy|verb|joy|noun
 - bi-directional (sPair)
 - **Transitivity:** $((x = y) \text{ and } (y = z)) \rightarrow (x = z)$
 - enjoy|verb -> joy|noun -> happy|adj
 - multiple (recursive) substitutions
 - sClass (synonym class)
- Prevent precision issues by near-synonyms.



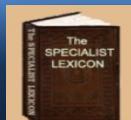
Synonym Types

- Cognitive synonym:
 - less difference
 - greater interchangeability (not context-sensitive)
 - more generic
 - can be represented as a synonym pair (sPair)
- Near-synonym:
 - greater difference
 - less interchangeability
 - specific use, can't used in generic case



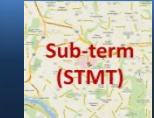
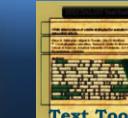
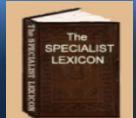
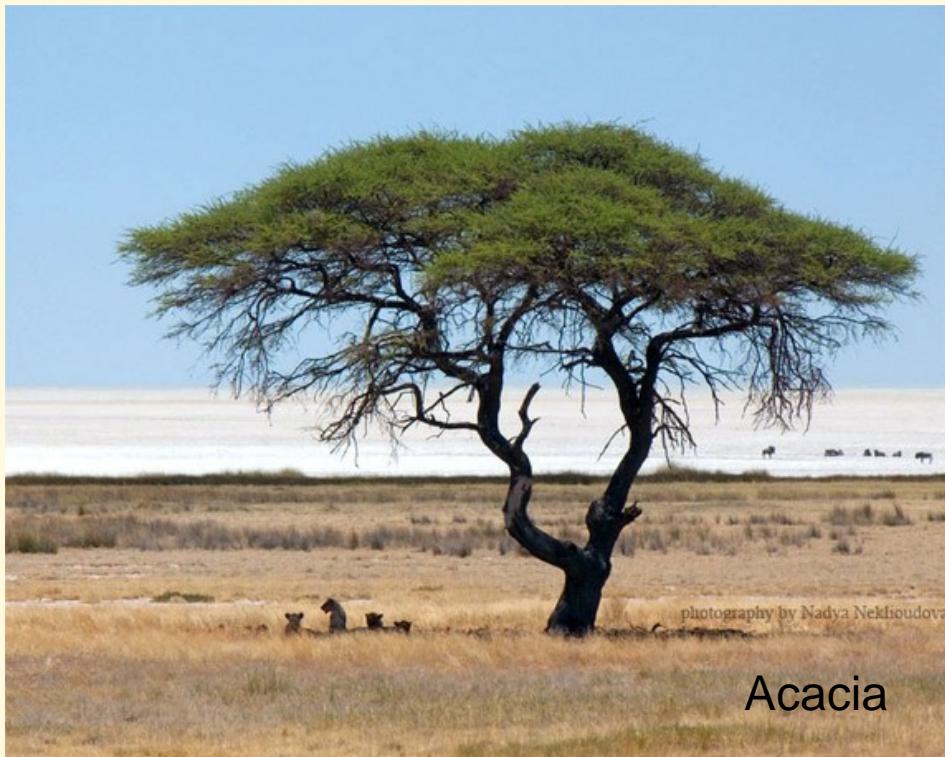
Near-Synonyms

CUI	Preferred Term	Synonym	Explanation
C0000869	Acacia	locust tree	Though both the acacia & locust tree are members of Leguminosae (pea, bean), they do seem to refer to different trees.
C0003353	Antigua	Anguilla	The islands of Antigua & Anguilla are both in the West Indies, but are not the same place.
C0032639	Pons	metencephalon	The metencephalon, per unabridged.merriam-webster.com includes the cerebellum and pons, and is different from the pons



Acacia & Locust tree

➤C0000869



Anguilla & Antigua

➤C0003353



Metencephalon & Pontine Structure (Pons)

➤C0032639

Hinbrain: Metencephalon

20

b) metencephalon

▫ pons

- Contains pneumotaxic centre which fine tunes breathing rate
- Relays information between cerebellum and cerebrum

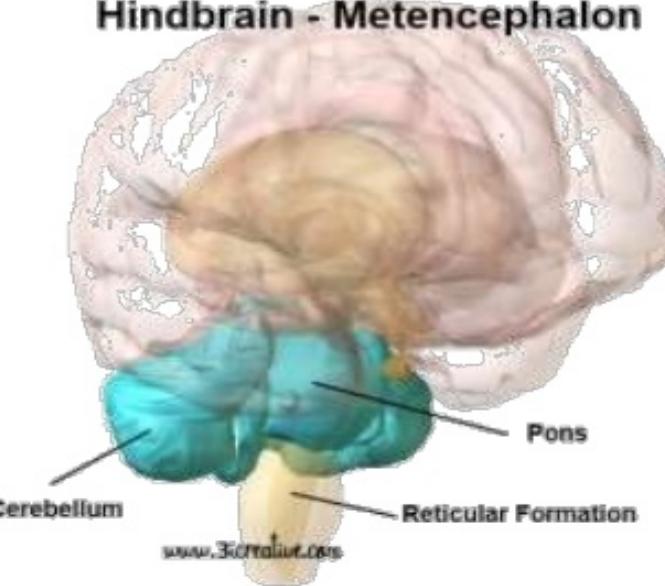
▫ cerebellum

- Feedback center for execution of motor movements
- Controls posture and balance

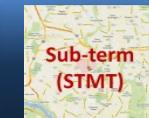
▫ reticular formation

- Nuclei diffusely located through the brainstem*
- Regulates wakefulness and muscle tone

Hindbrain - Metencephalon



*the term "brainstem" refers to the medulla oblongata, pons, and the midbrain



R2: POS Issues – Meaning Shift

CUI	Preferred Term	synonym	Explanation
C0004063	Assault	mug	The noun mug means a large cup, while the verb mug does refer to assault.
C0001774	Agaricales	Mushroom	The verb (to) mushroom means increase, spread, or develop rapidly. It does not refer to Agaricales while the noun is a synonym.
C0003459	Anura	frog	The verb (to) frog means hunt for or catch frogs. It does not refer to Anura, while the noun is a synonym.
C0003842	Arteries	arterial	The noun arterial refers to roads, not circulatory anatomy, unlike the adjective arterial.



POS: Assault & Mug



≠



mug|verb
(assault)

mug|noun
(a large cup)



R3: Source: CUI, EUI, ...

The patient expired 1 day later.

Pressure of CO₂ in expired air ...

Disposal of expired drug ...

CUI: C0011065

PT: Cessation of life

died

dead

death

deceased

...

CUI: C0231800

PT: Expiration, Function

exhaled

expiratory

expiration

...

CUI: C1704631

PT: Expiration

expire

expiration

...



R4: Acronym/Abbreviation Issues – Precision

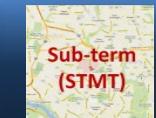
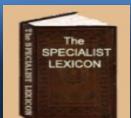
- ER (27): emergency room | efficacy ratio | ejection rate | evoked response | extended release | external resistance | eye research | energy restriction | ...

CUI	Preferred Term	synonym
C0003023	Angola	ago
C0001175	Acquired Immunodeficiency Syndrome	sida
C0001857	AIDS related complex	arc
C3714936	Non-Compliant ADaM Datasets Domain	ax

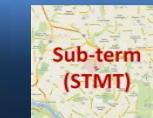
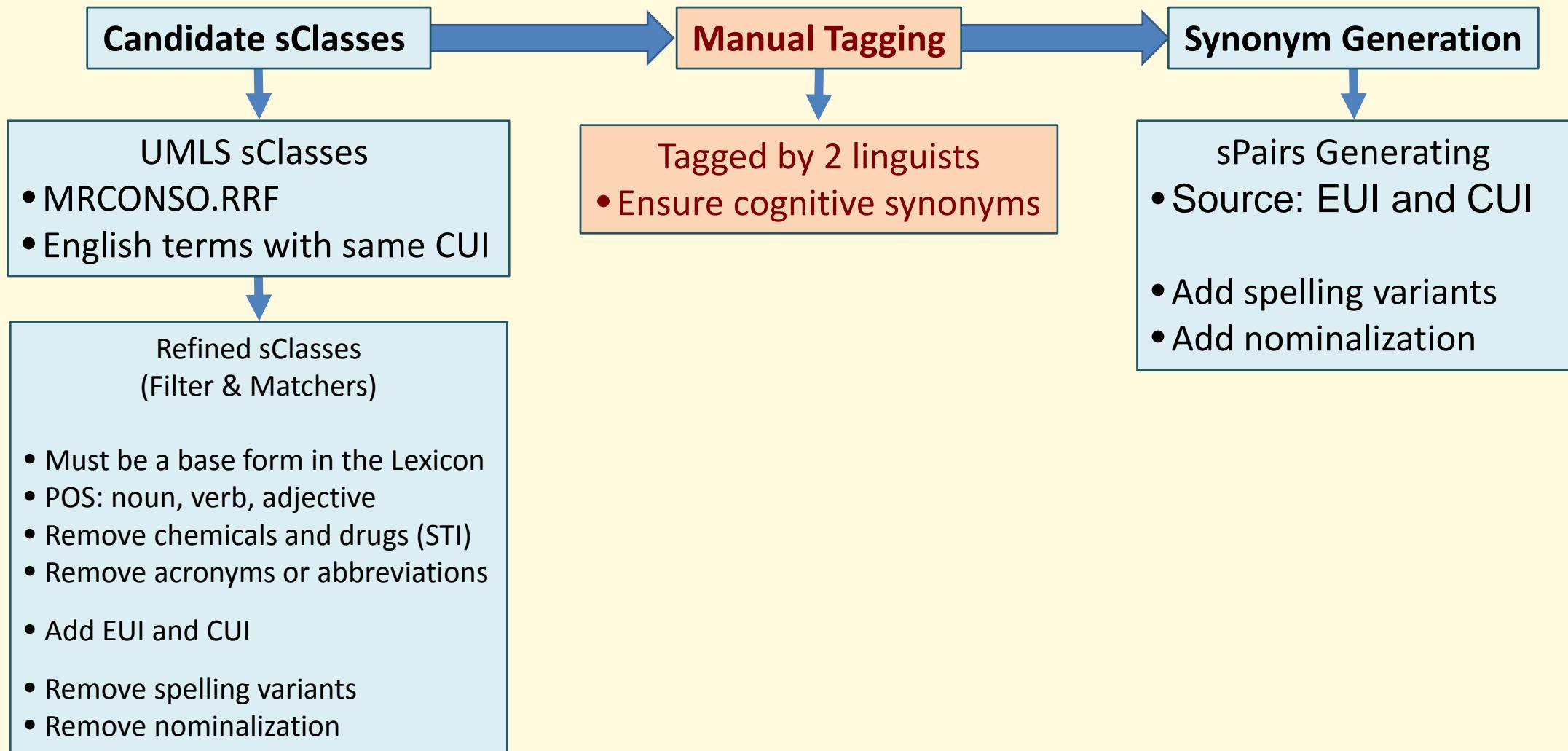


Approach - Refined sClass & Manually Tag

- English terms from MRCONSO.RRF with same CUI
- Exclude chemicals & drugs
 - use MRSTY.RRF to map CUI to STI
 - filter out disallowed STI in SemGroups.filter.txt
- In Lexicon with inflection is base and POS of adj, noun, or verb
- Remove acronyms/abbreviations => it drops precision
- Remove spVars => add them in post-process
- Remove nominalization => add them in post-process
- Remove singleton sClass (1 single candidates)
- Manually tag (for cognitive synonyms)



Computer-aided System



Example: sClass & Tags (POS)

#SYNONYM_CLASS|C0003842|Arteries

noun|E0010481|arteria|Y

noun|E0010531|artery|Y

noun|E0694191|arterial|N

adj|E0010482|arterial|Y

#SYNONYM_CLASS|C0004063|Assault

verb|E0041250|mug|Y

noun|E0010822|assault|Y

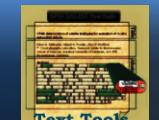
noun|E0041249|mug|N

...



Synonym Sources

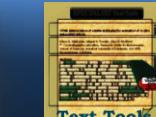
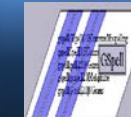
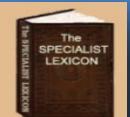
- Lexicon-Sourced Synonyms
 - Nominalizations with EUI
 - automatic retrieved from the SPECIALIST Lexicon
- UMLS-Sourced Cognitive Synonyms with CUI
- NLP Projects-Sourced Cognitive Synonyms
 - legacy data (LVG, STMT, UMLS Core, ...)
 - can be automatically retrieved
 - manually verified and add POS



Lexicon-Sourced Synonyms

- nominalizations are synonyms
- can be retrieved from the Lexicon automatically
- associated EUIs are preserved
- example:
 - sPair of [ability|noun|able|adj|E0006490]

```
{base=ability  
entry=E0006490  
cat=noun  
variants=reg  
variants=uncount  
compl=pphr(of,np)  
compl=infcomp:arbc  
nominalization_of=able|adj|E0006510  
}
```



Example: sClass & Tagging

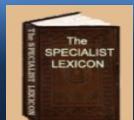
Refined sClass

```
...  
#SYNONYM_CLASS|C0011065|Cessation of life  
128|E0020918|death|Y  
1|E0020877|dead|Y  
1|E0020990|deceased|Y  
1|E0022536|die|
```

Lexical Records

```
{base=death  
entry=E0020918  
cat=noun  
variants=reg  
variants=uncount  
compl=pphr(of,np)  
compl=pphr(from,np)  
nominalization_of=die|verb|E0022536  
}
```

Removed (nominalization)



Example: sClass to sPairs

Final sClass

```
...  
#SYNONYM_CLASS|C0011065|Cessation of life  
128|E0020918|death|Y  
1|E0020877|dead|Y  
1|E0020990|deceased|Y  
1024|E0022536|die|nom  
128|E0020885|deadnes|nom  
...
```



sPairs

```
...  
deadness|128|dead|1|C0011065  
deadness|128|death|128|C0011065  
deadness|128|deceased|1|C0011065  
deadness|128|die|1024|C0011065  
dead|1|deadness|128|C0011065  
dead|1|death|128|C0011065  
dead|1|deceased|1|C0011065  
dead|1|die|1024|C0011065  
death|128|deadness|128|C0011065  
death|128|dead|1|C0011065  
death|128|deceased|1|C0011065  
death|128|die|1024|C0011065  
deceased|1|deadness|128|C0011065  
deceased|1|dead|1|C0011065  
deceased|1|death|128|C0011065  
deceased|1|die|1024|C0011065  
die|1024|deadness|128|C0011065  
die|1024|dead|1|C0011065  
die|1024|death|128|C0011065  
die|1024|deceased|1|C0011065  
...
```

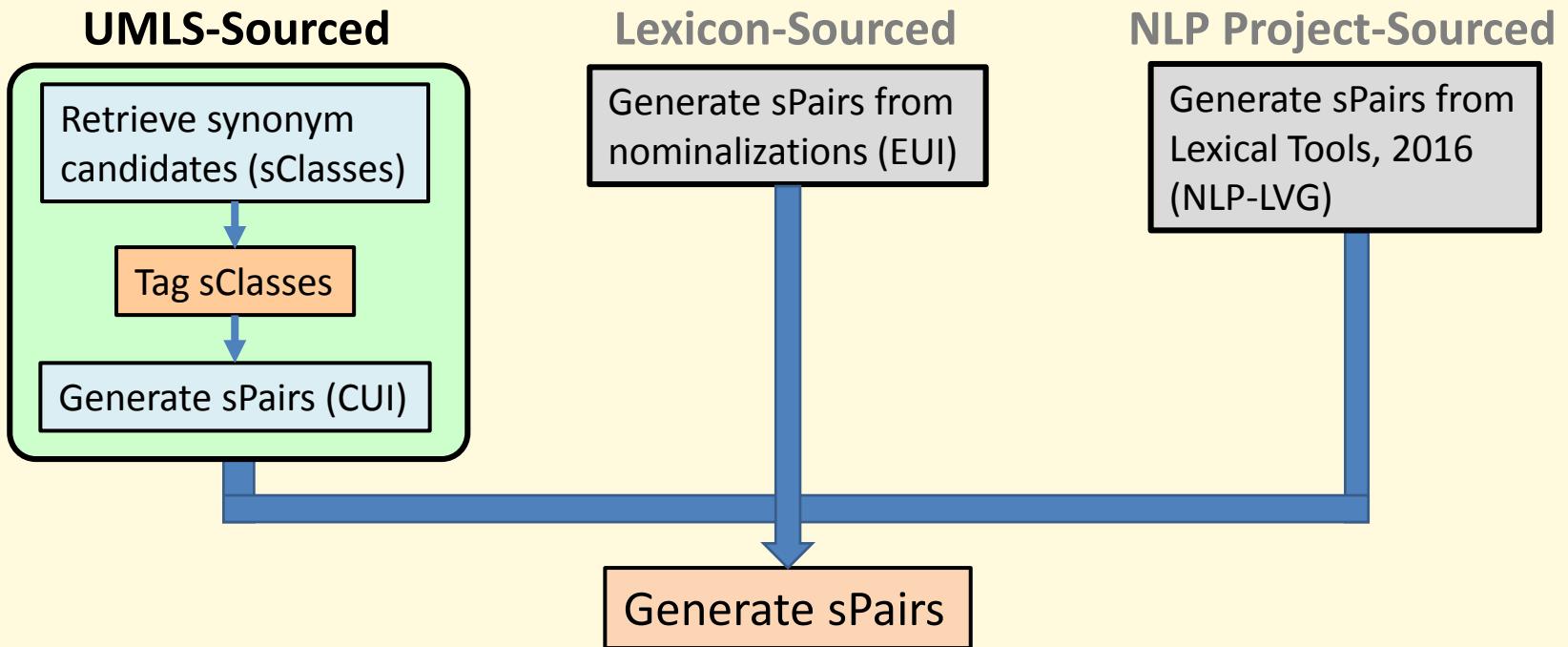
Add nominalization

```
{base=death  
entry=E0020918  
cat=noun  
variants=reg  
variants=uncount  
compl=pphr(of,np)  
compl=pphr(from,np)  
nominalization_of=die|verb|E0022536  
}
```

```
{base=dead  
entry=E0020877  
cat=adj  
variants=inv  
...  
position=pred  
stative  
nominalization=deadness|noun|E0020885  
}
```



sPairs Generation



Synonym-1	POS-1	Synonym-2	POS-2	Source
mug	verb	assault	noun	C0004063
assault	noun	mug	verb	C0004063
...



Results – 2017 Release

➤ 2017 LexSynonyms

	Candidates	Tagged	Completion (%)
sClass	22,779	7,686	33.74%
Synonyms	80,913	29,990	37.06%

➤ Synonyms (sPairs):

Year	CUI	EUI	NLP	Total
2016	0 (0%)	0 (0%)	5,198 (100%)	5,198
2017	118,468 (62%)	67,584 (35%)	4,792 (3%)	190,844

36.71 growth

➤ Format:

Synonym-1 | POS-1 | Synonym-2 | POS-2 | Source



Evaluation

➤ Model:

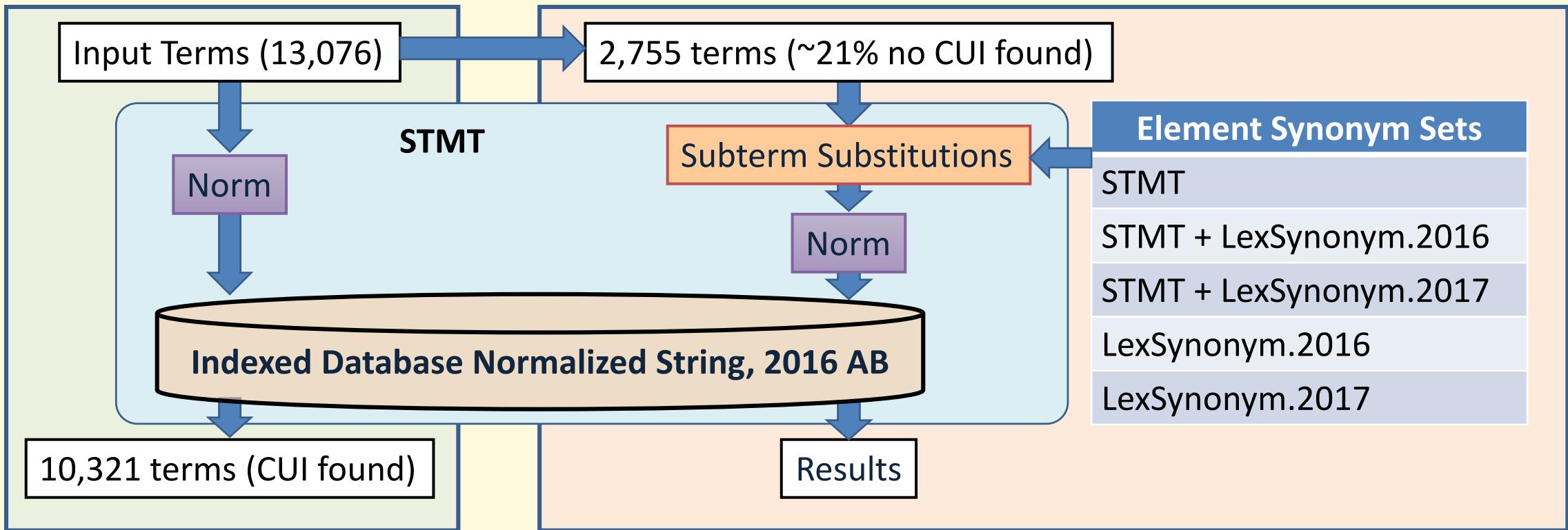
- STMT (Sub-Term Mapping Tools) [6]:
 - Real-time subterm substitution features for concept mapping
 - Easy configurable options for element synonym set

➤ Data:

- UMLS-Core Project:
 - Top 95% used terms form 8 hospitals.
 - Assigned CUI(s) to 13,076 terms
 - 2,755 terms of them do not have mapped concept through normalization in UMLS.2016AB
 - Gold Standard: 2,755 terms mapped to 2,756 CUIs



Evaluation Model



Evaluation Results

- Gold Standard: 2,755 terms mapped to 2,756 CUIs
- Element sets:
 - STMT: a validated project specific synonym set for UMLS-Core project
 - About 75% of STMT element synonyms are duplicated in LexSynonym.2017, while only ~3% are duplicated in LexSynonym.2016.

Element Synonym Set	N. Size	T.P.	F.P.	F.N.	Precision	Recall	F1	Time
STMT [6]	7,873	690	353	2,066	66.16%	25.04%	0.3633	7:57
STMT + LexSynonym.2016	12,681	691	358	2,065	65.87%	25.07%	0.3632	5:31
STMT + LexSynonym.2017	151,913	828	424	1,928	66.13%	30.04%	0.4132	9:18
Element Synonym Set	N. Size	T.P.	F.P.	F.N.	Precision	Recall	F1	Time
LexSynonym.2016	5,070	9	12	2,747	42.86%	0.33%	0.0065	0:16
LexSynonym.2017	149,912	287	117	2,469	71.04%	10.41%	0.1816	3:19



Lexical Tools – Synonym Flow

➤ Software Changes:

- Include POS and the source information in synonym database

➤ Example:

```
shell> lvg -f:y -m
```

```
die
```

```
die|dead|1|1|y|1|FACT|die|die|verb|dead|adj|C0011065|
```

```
die|deadness|128|1|y|1|FACT|die|die|verb|deadness|noun|C0011065|
```

```
die|death|128|1|y|1|FACT|die|die|verb|death|noun|C0011065|
```

```
die|deceased|1|1|y|1|FACT|die|die|verb|deceased|adj|C0011065|
```

```
die|expire|1024|1|y|1|FACT|die|die|verb|expire|verb|NLP_LVG|
```



Lexical Tools – Synonyms Flow Options

- Synonym source restriction options (-ks):
- C (CUI), E (EUI), N (NLP), CE, CN, EN, CEN.

➤ Example:

```
shell> lvg -f:y -m -ks:C
```

die

die|dead|1|1|y|1|FACT|die|die|verb|dead|adj|C0011065|

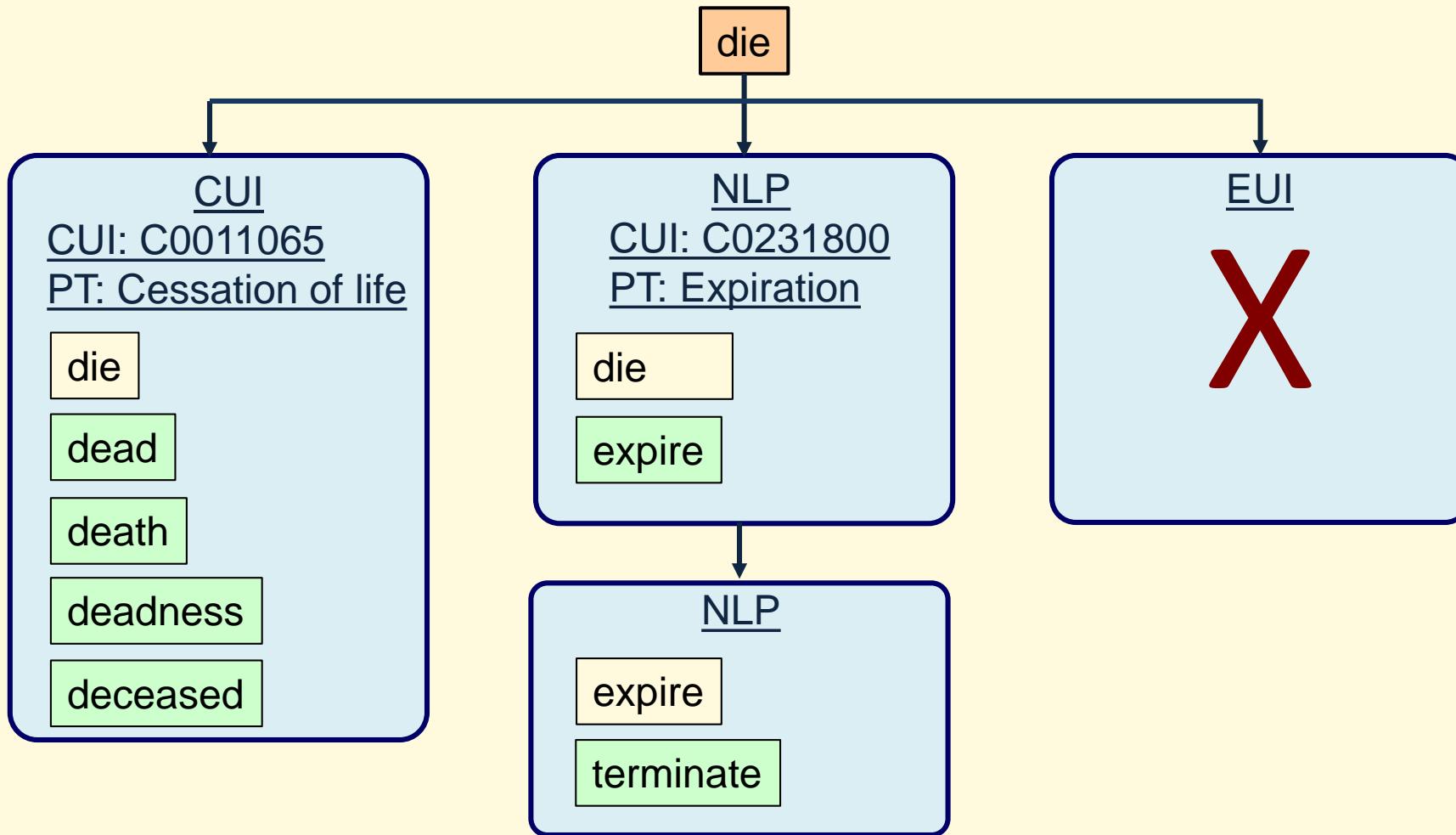
die|deadness|128|1|y|1|FACT|die|die|verb|deadness|noun|C0011065|

die|death|128|1|y|1|FACT|die|die|verb|death|noun|C0011065|

die|deceased|1|1|y|1|FACT|die|die|verb|deceased|adj|C0011065|



Lexical Tools – Recursive Synonyms



Lexical Tools – Recursive Synonym Flow

➤ Software Enhancement:

- must have the same type of source
- If the source is CUI: only synonyms from the same CUI are used (multiple CUI Issues)
- If the source is EUI: all synonyms with EUI source are used
- If the source is NLP: synonyms from same NLP source are used

➤ Example:

```
shell> lvg -f:y -m
```

```
die
```

```
die|dead|1|1|r|2|FACT|die|verb|dead|adj|C0011065|y|
```

```
die|deadness|128|1|r|2|FACT|die|verb|deadness|noun|C0011065|y|
```

```
die|death|128|1|r|2|FACT|die|verb|death|noun|C0011065|y|
```

```
die|deceased|1|1|r|2|FACT|die|verb|deceased|adj|C0011065|y|
```

```
die|expire|1024|1|r|2|FACT|die|verb|expire|verb|NLP_LVG|y|
```

```
die|terminate|1024|1|r|2|FACT|expire|verb|terminate|verb|NLP_LVG|yy|
```



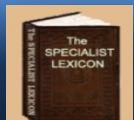
Summary

Objective & Requirements	Check	Notes
Standalone element synonym set	Yes	
All synonymous terms in the Lexicon	1/3 Yes	~ 1/3 completed
Grows with the SPECIALIST Lexicon	Yes	
Element synonyms, not expanded terms (Over-generated issues)	Yes	Must be in the Lexicon (430K, < 2% of UMLS synonyms)
R1: Cognitive Synonym	Yes	Done in tagging (cognitive synonyms)
R2: Include POS	Yes	Provide POS in sClass by Lexicon
R3: Include source (CUI, EUI, etc.)	Yes	Provide source in sClass (CUI, EUI, etc.)
R4: Exclude Acronym/abbreviation	Yes	Removed in sClass by Lexicon
R5: Include Single words and multiwords	Yes	Terms in the Lexicon include both
Improve NLP performance	Yes	Improve recall and preserve precision

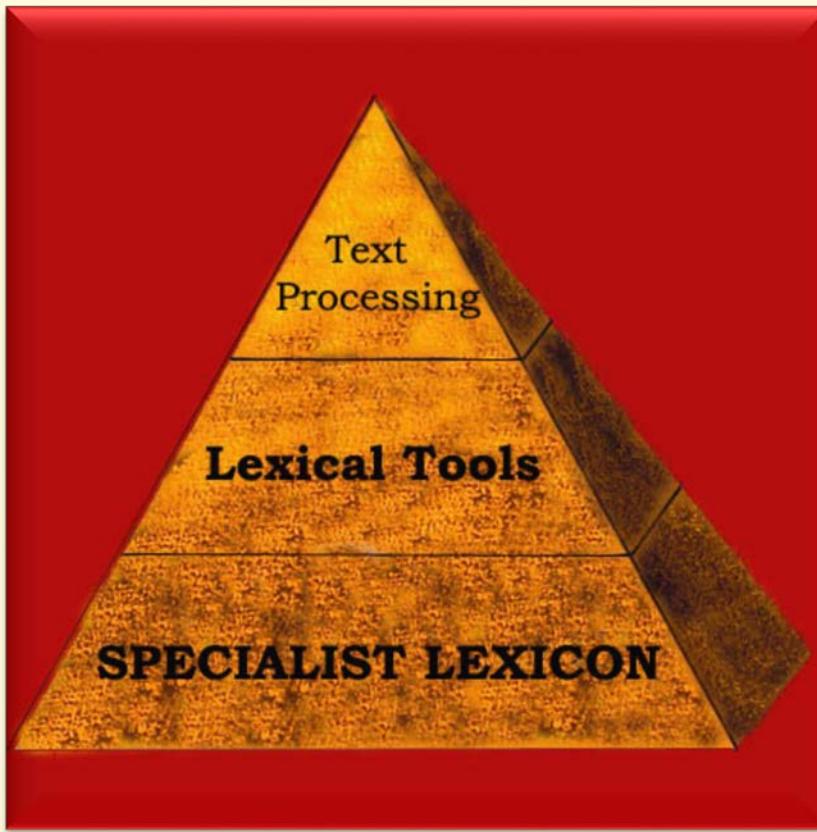


Future Work

- Complete all candidate sClasses in the future releases
- Update annually on Lexicon and Lexical Tools release with the latest Lexicon and UMLS Metathesaurus
- Include more project specific synonym set from other NLP resources (UMLS-Core, Randy Milller, etc.)
- Performance tests on NLP applications



Questions



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

