The SPECIALIST Lexicon and NLP Tools

By: Dr. Chris J. Lu

NLM – LHNCBC - CGSB

Jun., 2017

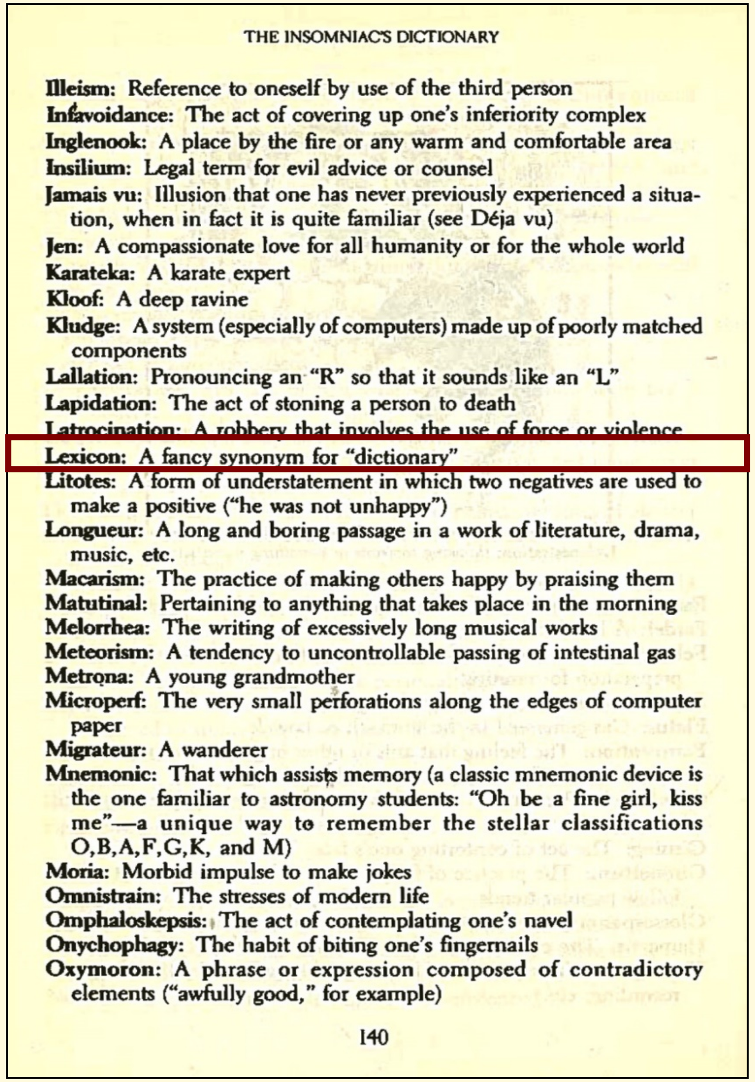
- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov

Table of Contents

- 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)

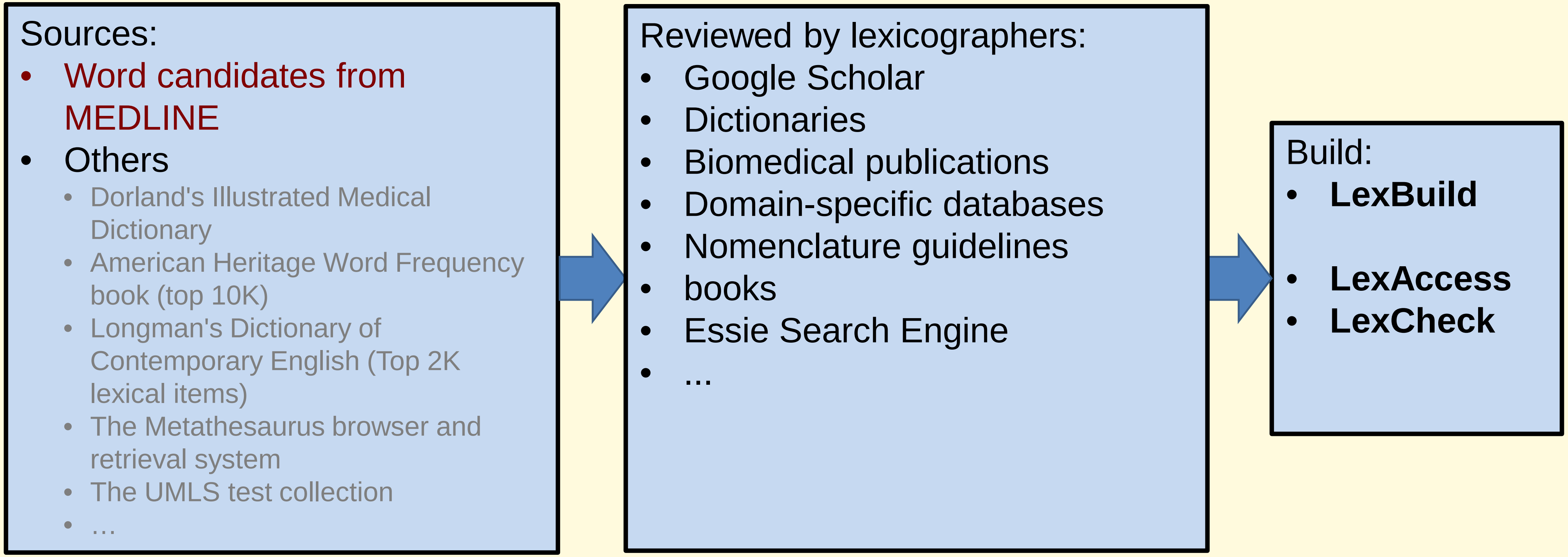


THE INSOMNIAC'S DICTIONARY

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 Déja 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)

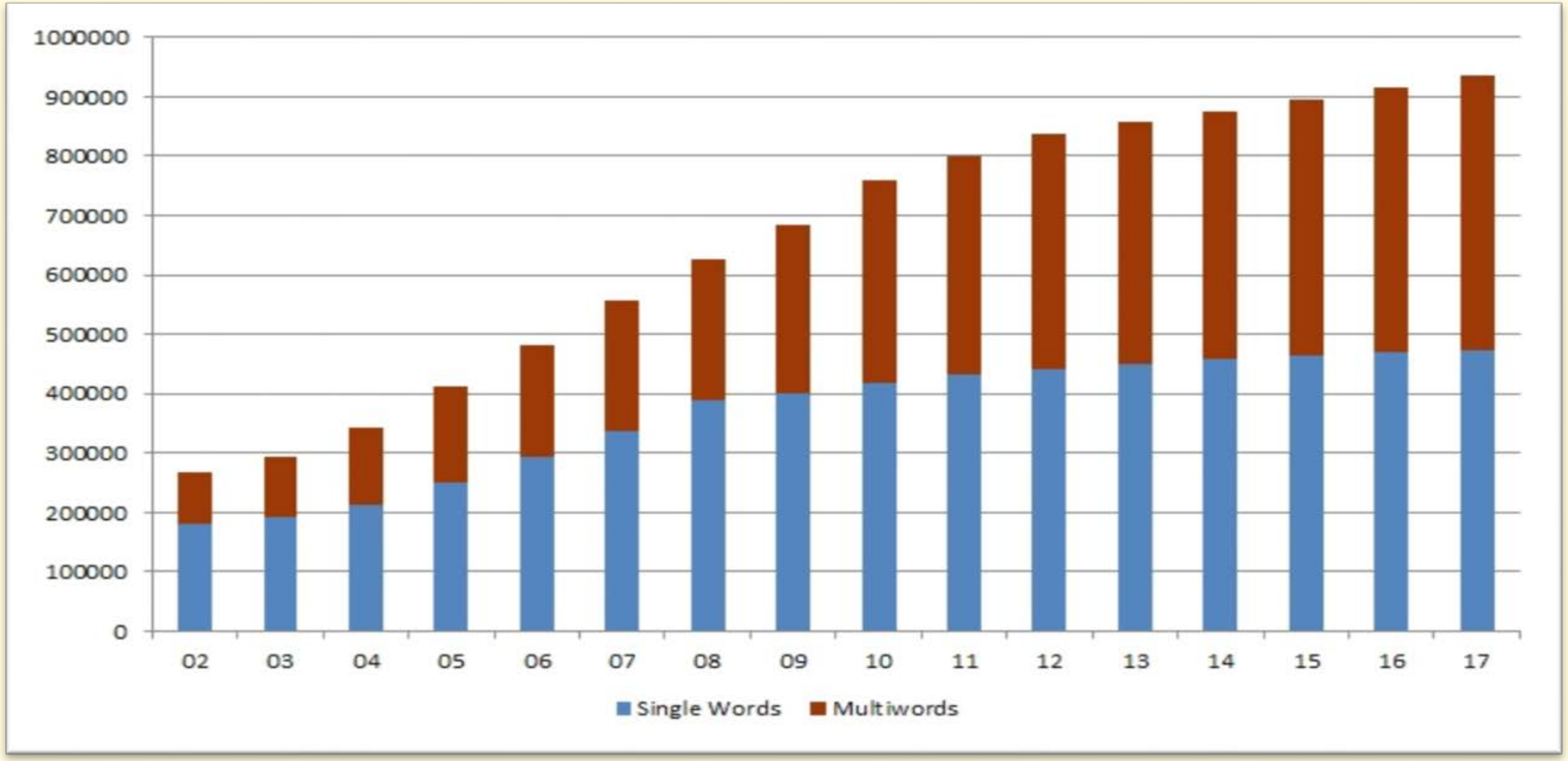
- Dr. Dina Demner Fushman

- Dr. Lynn McCreedy
- Destinee Tormey
- Francois Lang

- Dr. Chris J. Lu

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%)



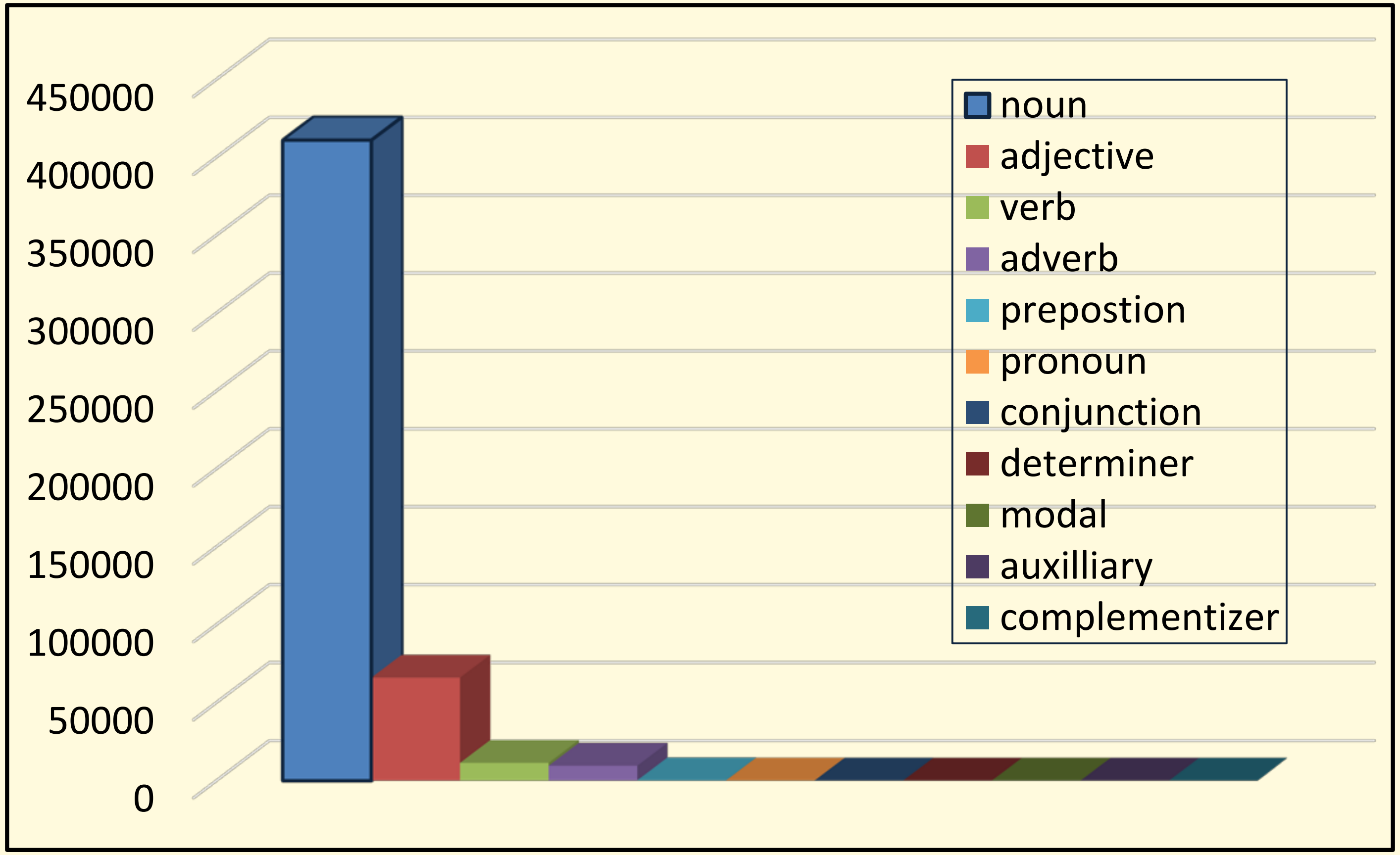
(Multi)Words for Lexical Records

- 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)

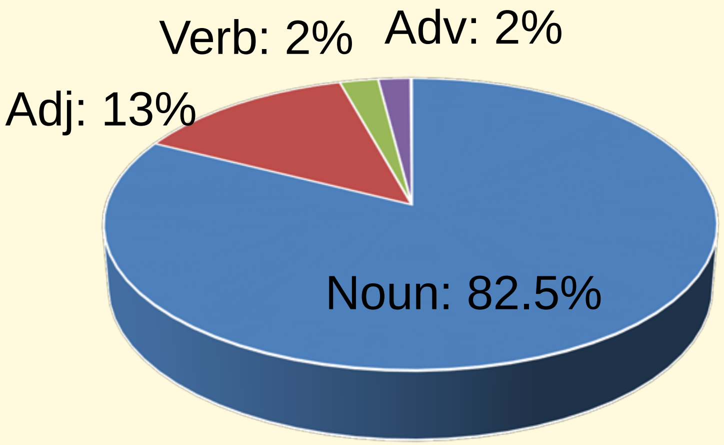
Lexical Records - Information

- 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)



Lexicon.2017



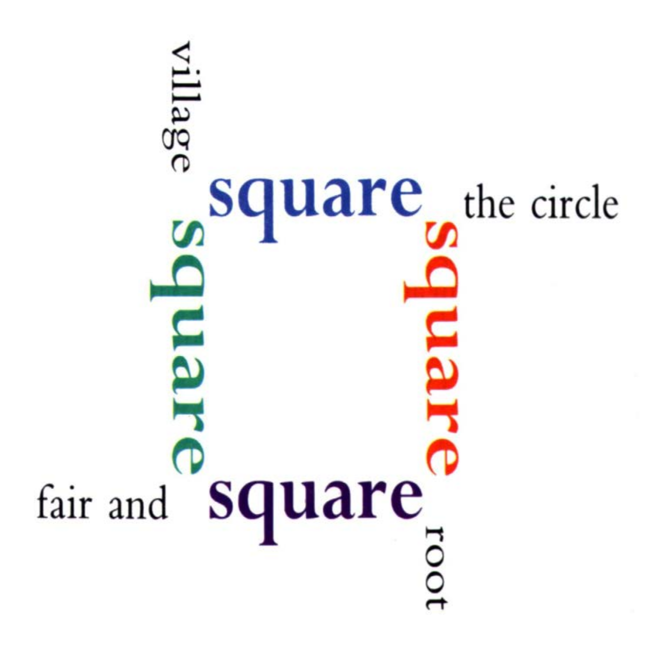
Lexical Records & POS

```
{base=square
entry=E0057517
        cat=verb
        variants=reg
        intran
        intran;part(up)
}
```

```
{base=square
entry=E0057516
        cat=adj
        variants=reg
        variants=inv
}
```

```
{base=square
entry=E0057518
        cat=adv
}
```

```
{base=square
entry=E0057515
        cat=noun
        variants=reg
}
```



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 Records to Lexical Information

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

Lexical Information | Base	color
Part of speech	• noun
Inflectional morphology (inflections)	• color • colors
Orthography	• colour
Abbreviation/Acronym	• N/A
Syntax (complementation)	• N/A
…	• …
Derivational morphology (derivations)	• colorable • colorful • colorize • colorist • …
LexSynonyms	• chromatic

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.

The Frequency Spectrum of Alice in Wonderland

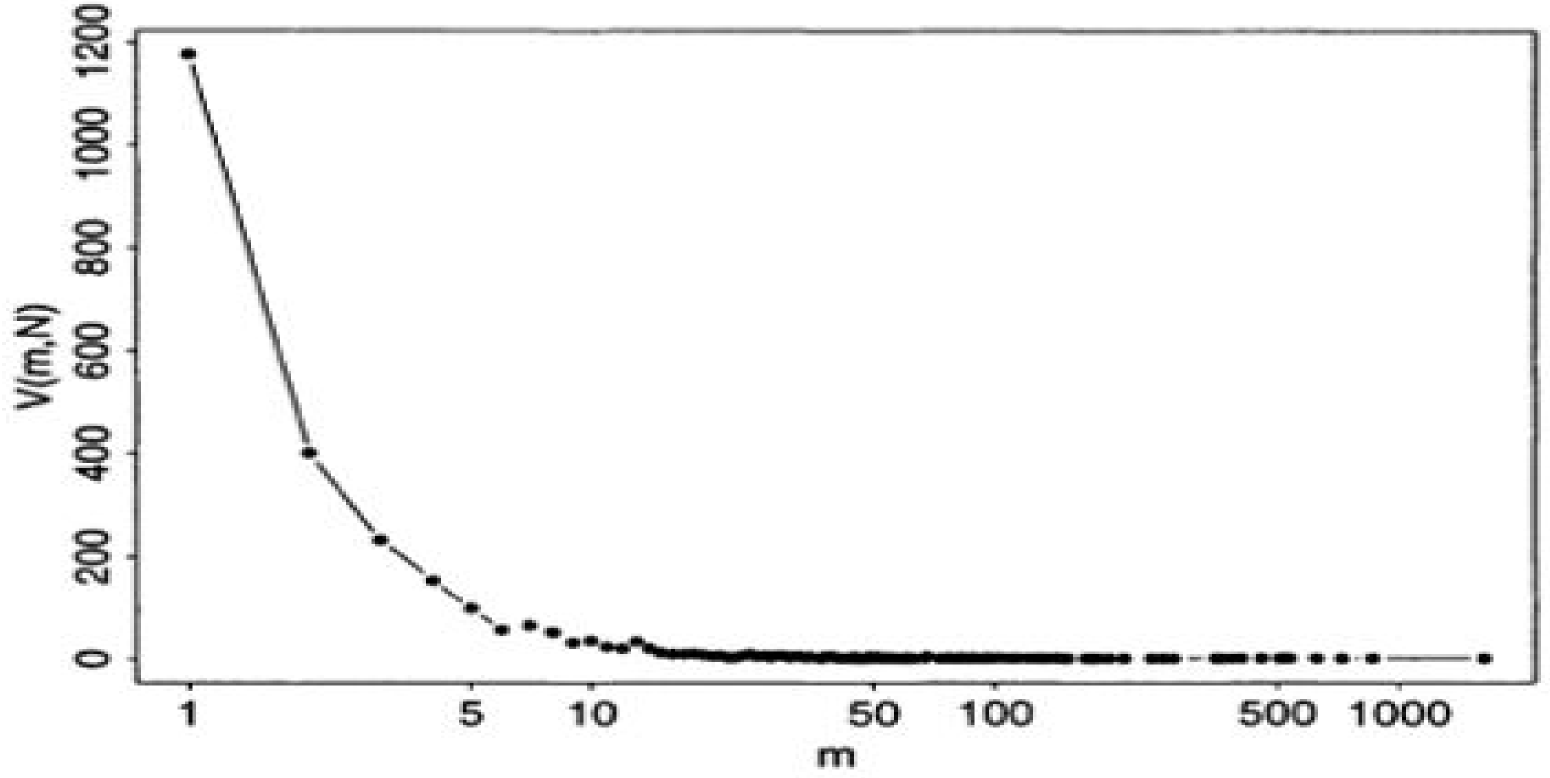
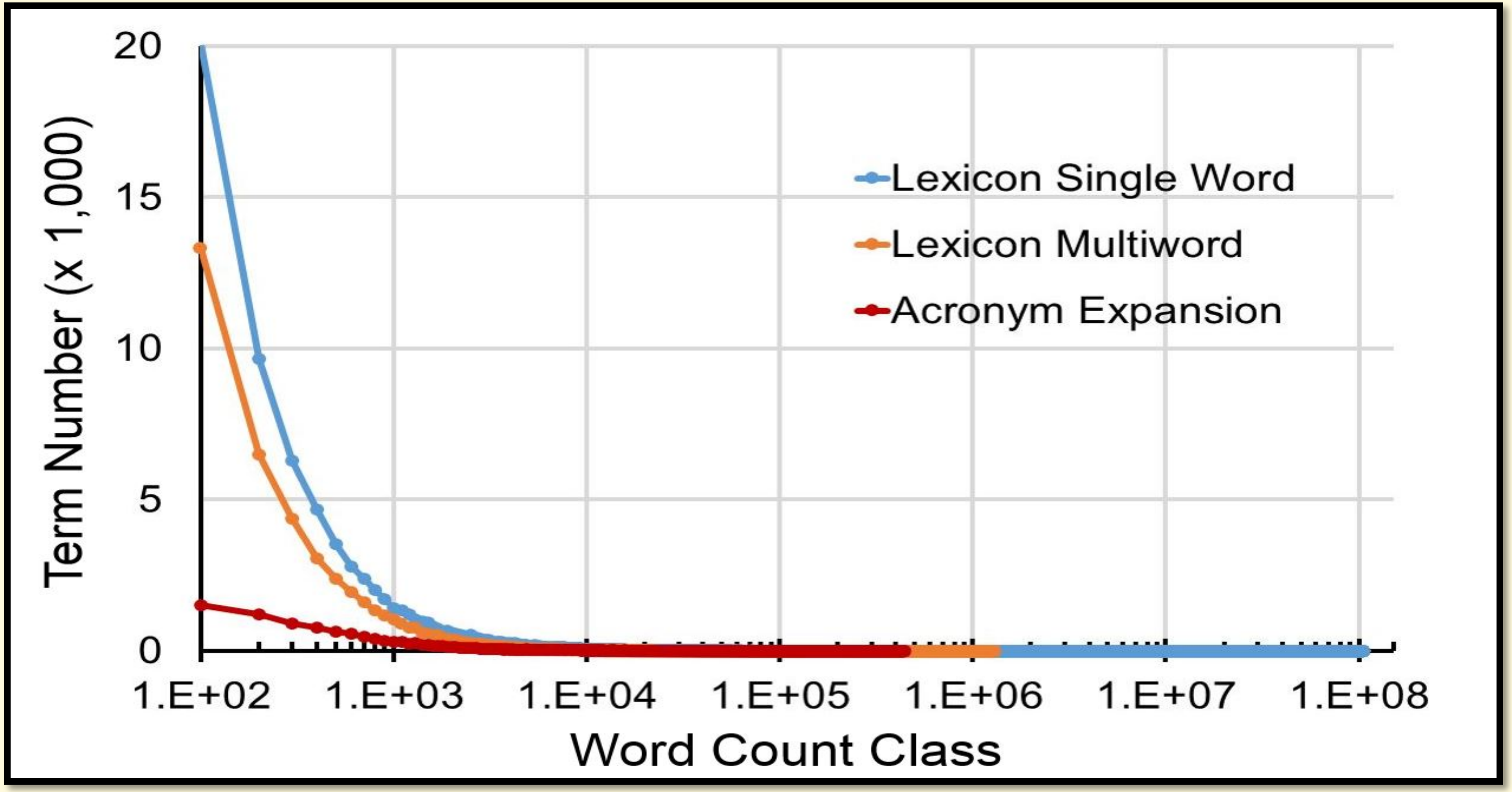


Figure 1.3: *The frequency spectrum of Alice in Wonderland (m: frequency class; V(m, N) : number of types with frequency m).*

The Frequency Spectrum of Lexicon (Multi)words on MEDLINE



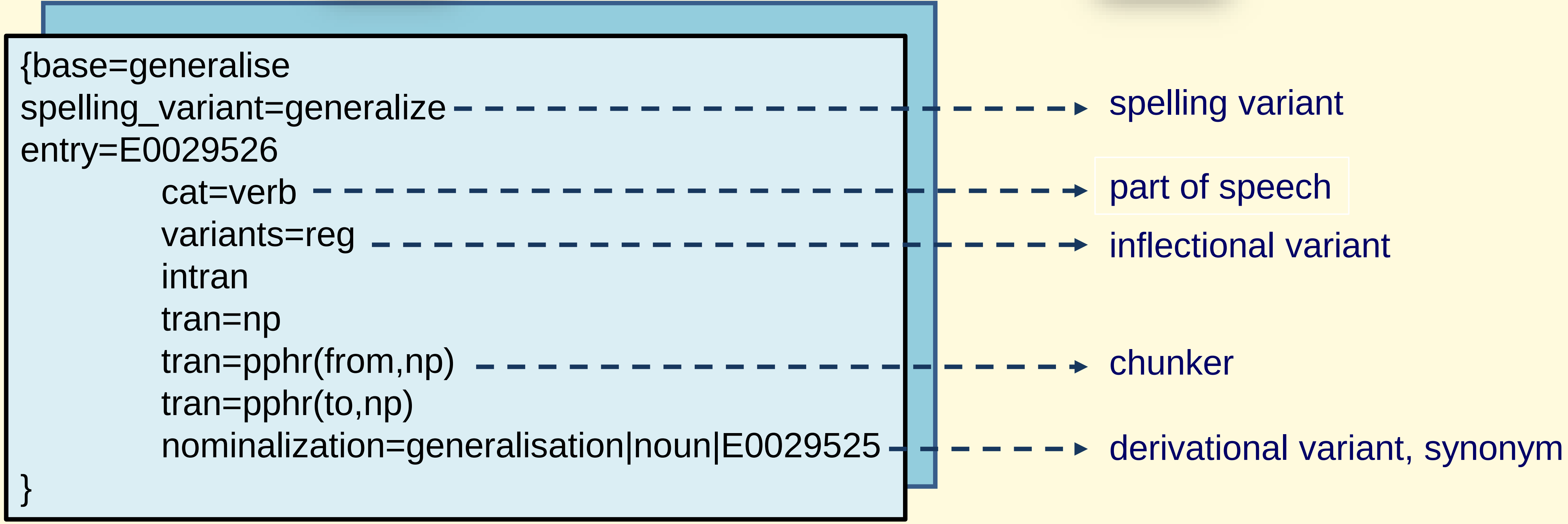
Lexicon Unigram Coverage – With Frequency (WC)

- 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.

Lexicon (Data) and Lexical Tools (Software)



```
{base=generalise
spelling_variant=generalize
entry=E0029526
        cat=verb
        variants=reg
        intran
        tran=np
        tran=pphr(from,np)
        tran=pphr(to,np)
        nominalization=generalisation|noun|E0029525
}
```

spelling variant

part of speech

inflectional variant

chunker

derivational variant, synonym

2. 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.

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

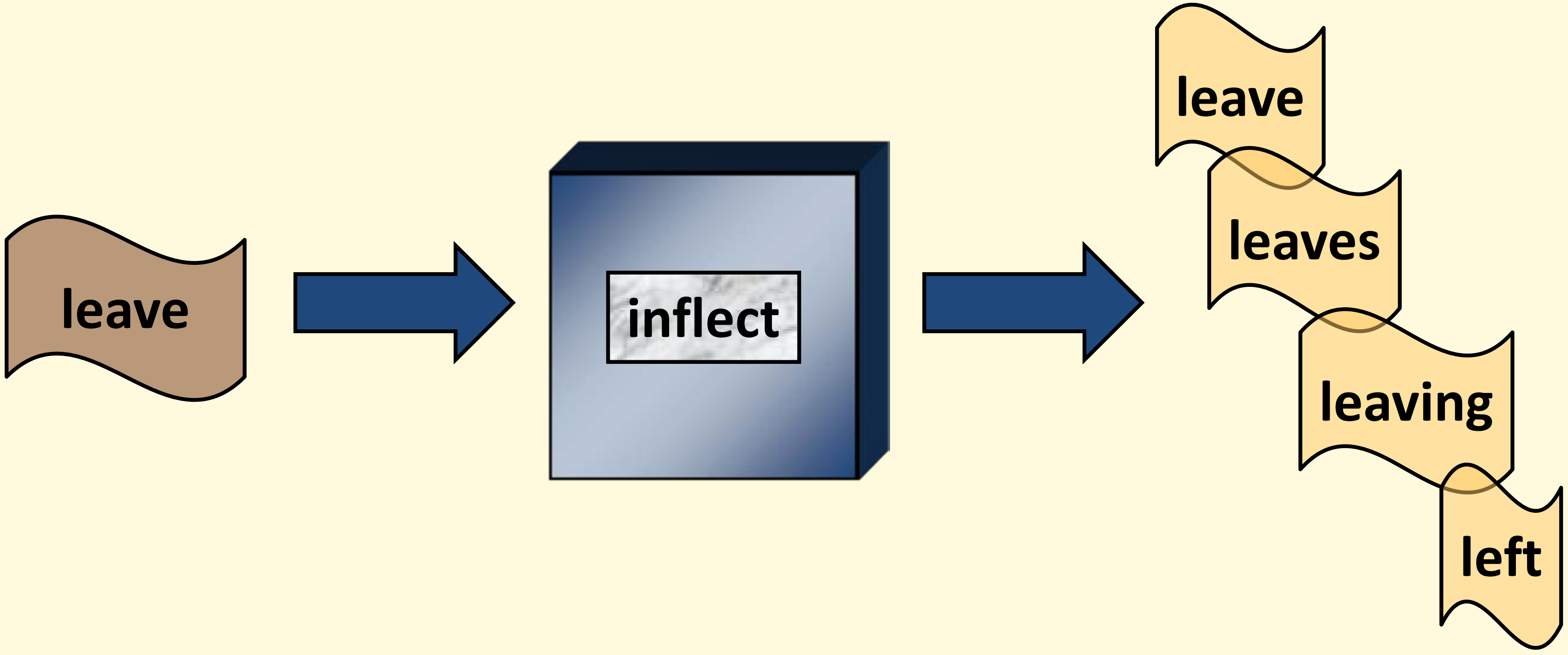
LVG - Lexical Variants Generation

- 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, I, 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

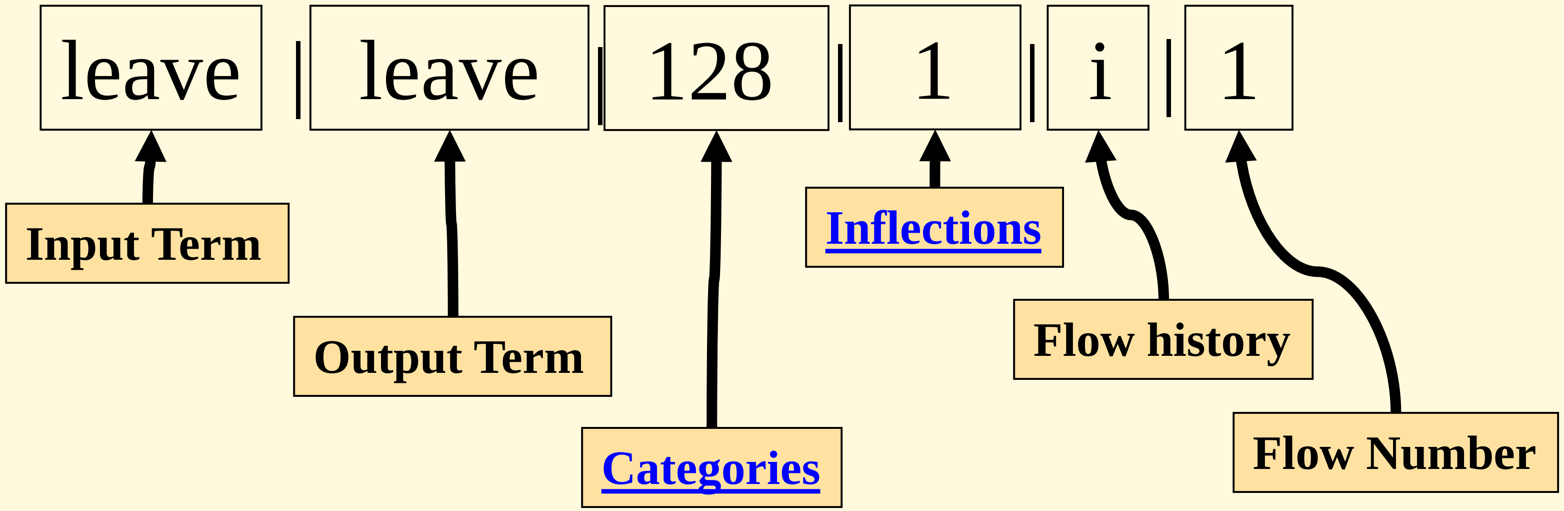


LVG Flow Component – Cmd line

```
> 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
```



```
leave|leave|128|1|i|1
```

Input Term

Output Term

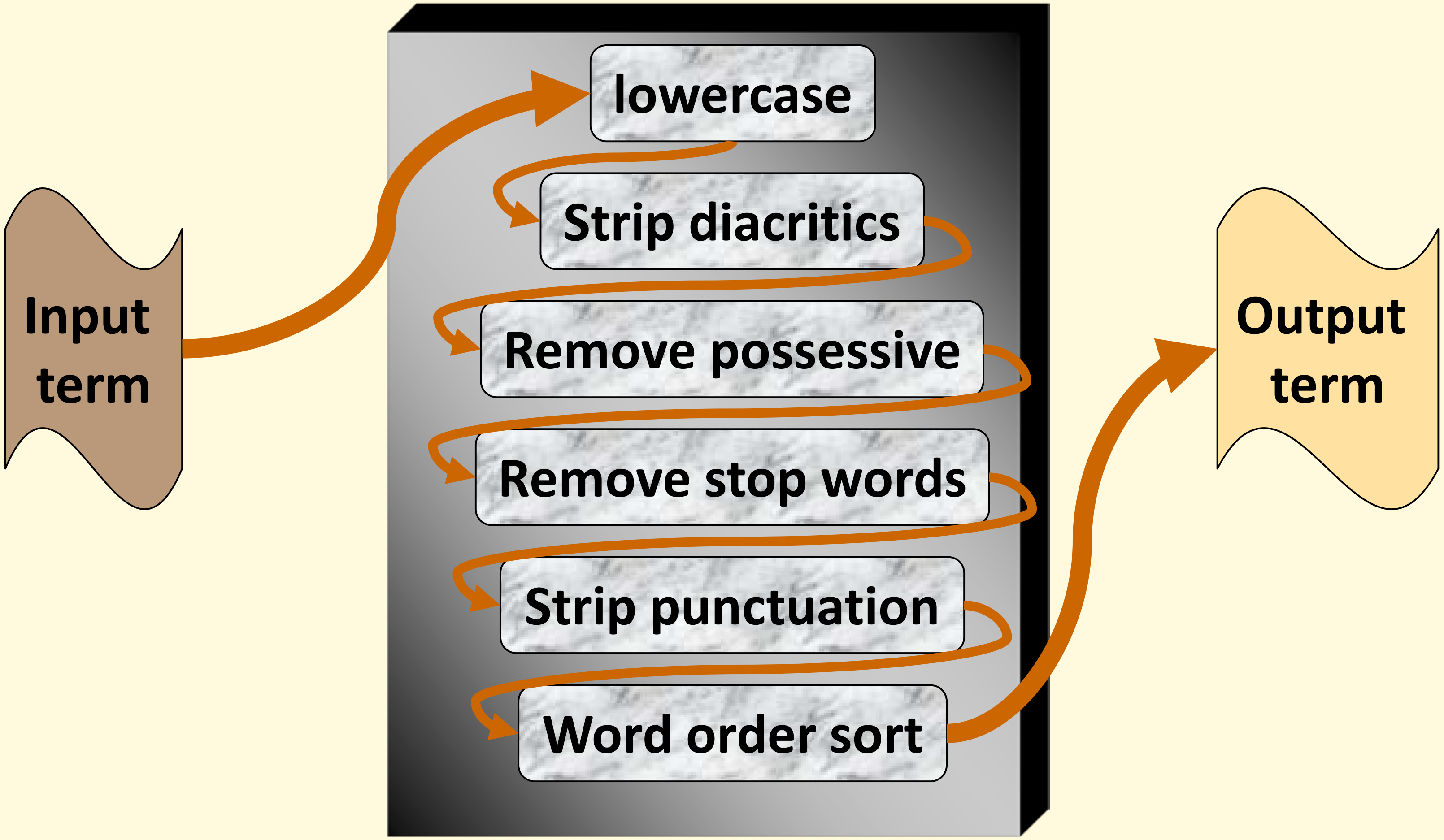
Categories

Inflections

Flow history

Flow Number

LVG – A Serial Flow



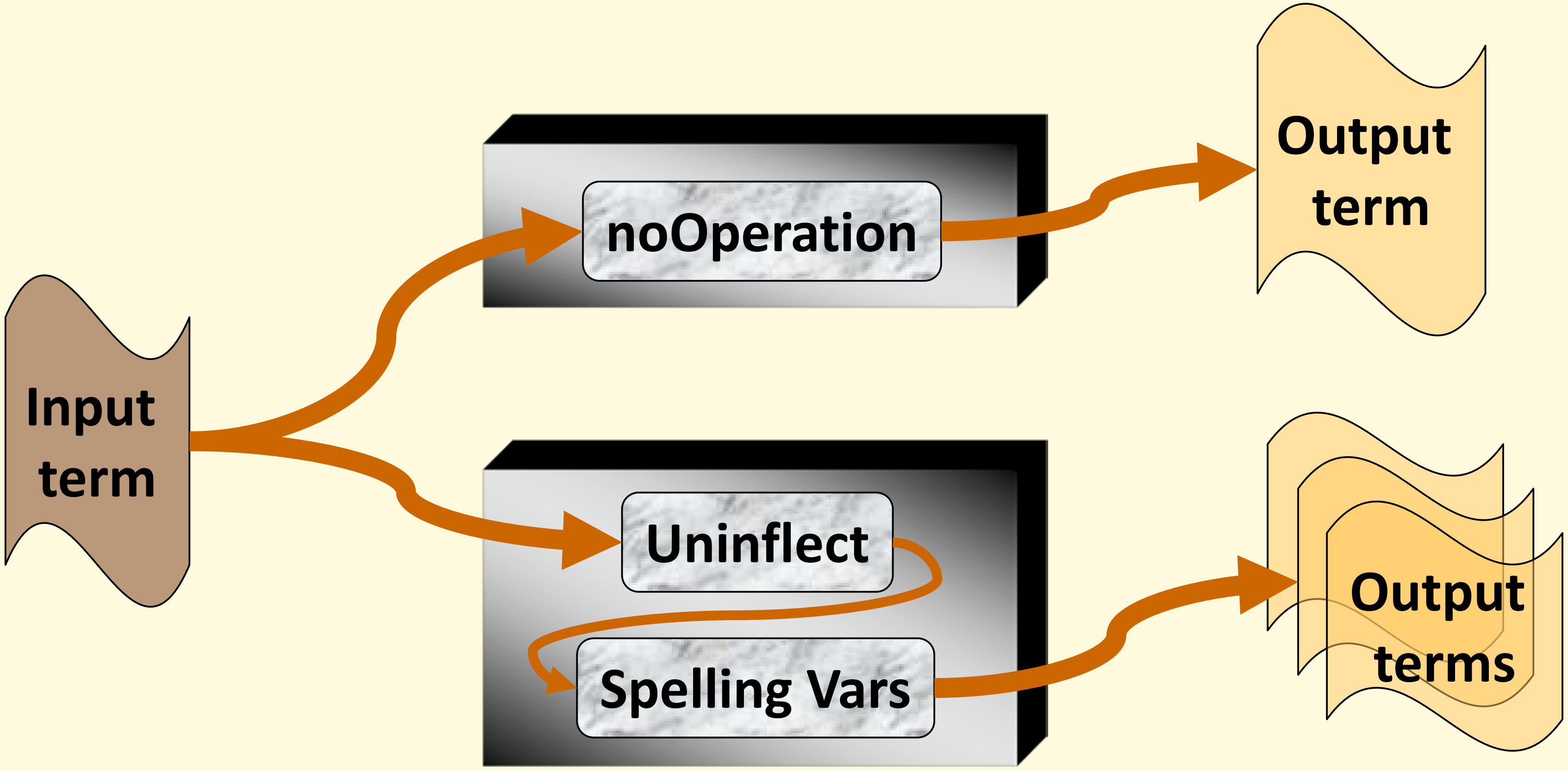
- Flow components can be arranged so that the output of one is the input to another.

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

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

Norm

“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: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order

Norm

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

“Fœtoproteins α’s, NOS“

"Fœtoproteins α’s, NOS"

Norm

q0: map symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map Unicode to ASCII

w: sort words by order

“Fœtoproteins α’s, NOS“

"Fœtoproteins α**’s**, NOS"

"Fœtoproteins α, NOS"

Norm

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

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

Norm

q0: map symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map Unicode to ASCII

w: sort words by order

“Fœtoproteins α’s, NOS“

"Fœtoproteins α’s, NOS"

"Fœtoproteins α, NOS"

"Fœtoproteins α, NOS"

Fœtoproteins α NOS

Norm

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

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

Norm

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

Norm

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

Norm

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

Norm

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

Norm

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

Norm

	“Fœtoproteins α’s, NOS“
q0: map symbols to ASCII	"Fœtoproteins α’s, NOS"
g: remove genitives	"Fœtoproteins α, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	Fœtoproteins α NOS
t: strip stop words	Fœtoproteins α
l: lowercase	fœtoproteins α
B: uninflect each words in a term	fœtoprotein α
Ct: retrieve citations	fetoprotein α
q7: Unicode core Norm	fetoprotein α
q8: strip or map Unicode to ASCII	fetoprotein alpha
w: sort words by order	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, written, or sign.
- The main purpose of language is communication, for us to understand the meaning.
- NLP includes a board range of subjects.
- NLP in our scope is to use computer to understand the meaning (concept) from text for further analysis and processing.

3. Natural Language Processing (NLP)

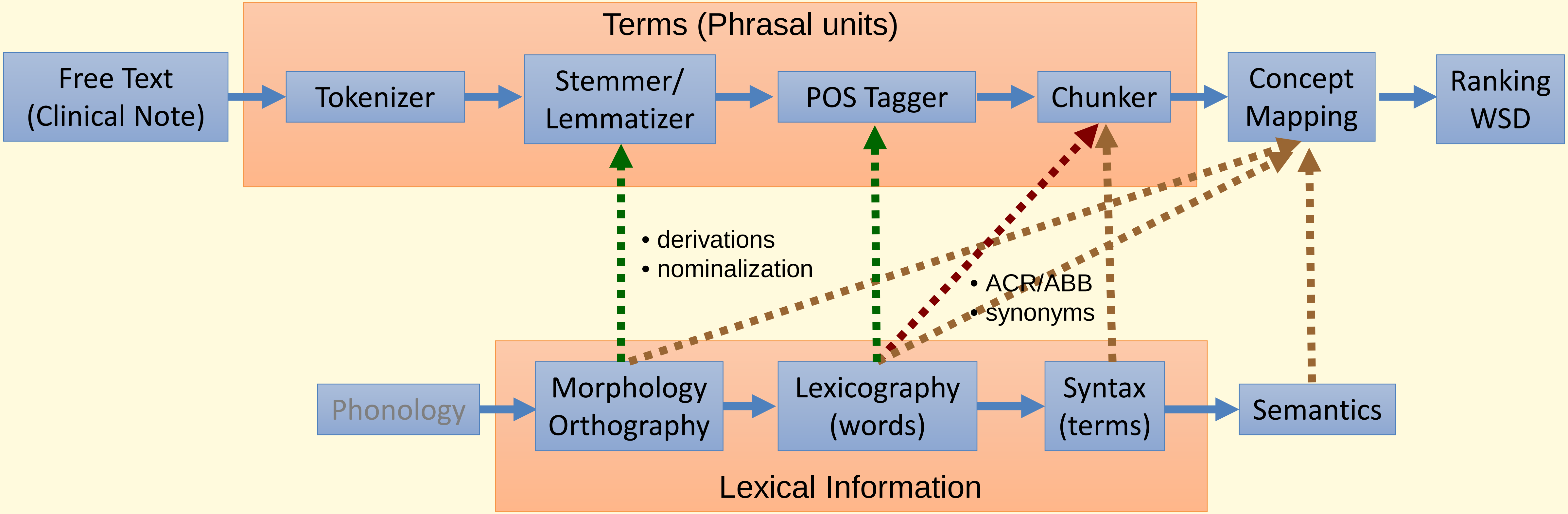
- Natural Language
  - is ordinary language that humans use naturally
  - may be spoken, signed, or written
- Natural Language Processing
  - NLP is to process human language to make their information accessible to computer applications
  - The goal is to design and build software that will analyze, 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.

NLP Challenges

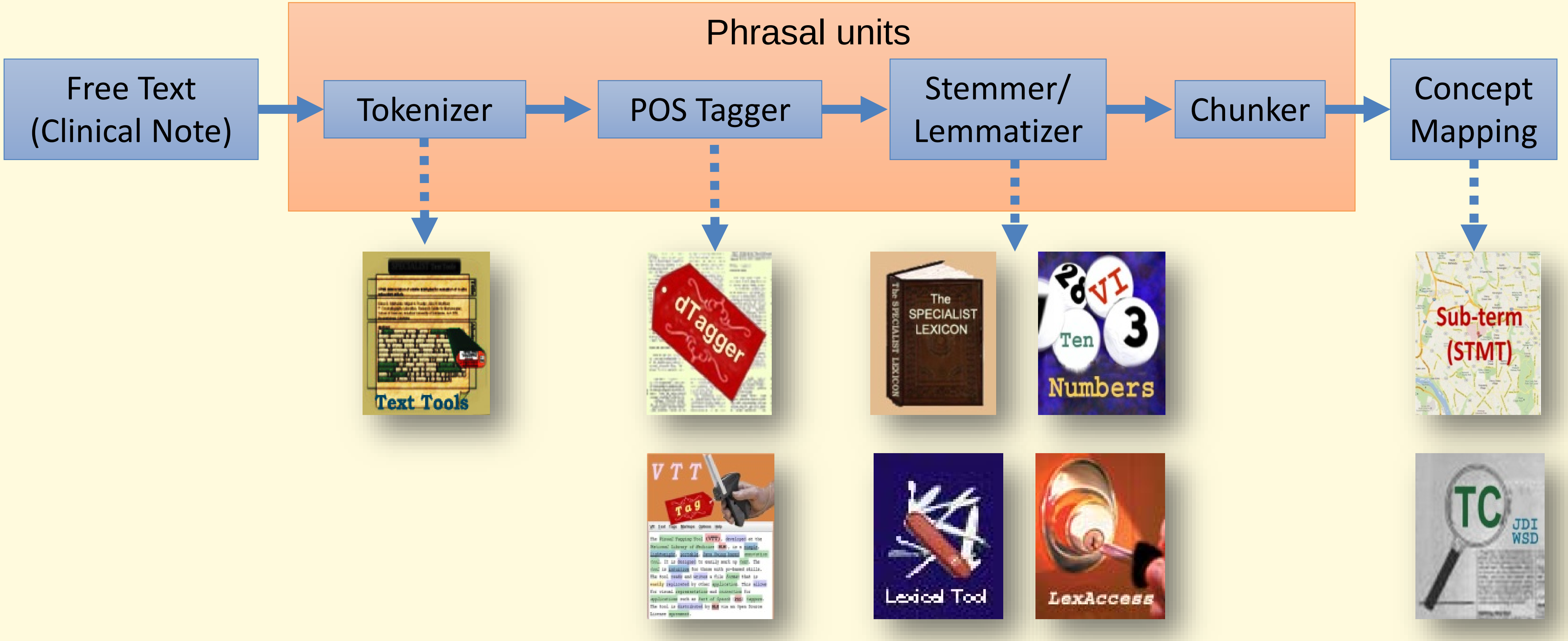
- Challenge 1: Map terms to concepts (meaning)
- Challenge 2: many to many mapping

Terms	Concepts	NLP
• cold • Cold Temperature • Cold Temperatures • Cold (Temperature) • Temperatures, Cold • Low temperature • low temperatures • …	• Cold Temperature|C0009264	• Concept mapping
• cold	• Cold Temperature|C0009264 • Common Cold|C0009443 • Cold Therapy|C0010412 • Cold Sensation|C0234192 • …	• WSD (Word Sense Disambiguation)

NLP Pipe Line – Lexical Information



The SPECIALIST NLP Tools



Phrasal units

Free Text (Clinical Note) → Tokenizer → POS Tagger → Stemmer/ Lemmatizer → Chunker → Concept Mapping

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

NLP – Concept Mapping

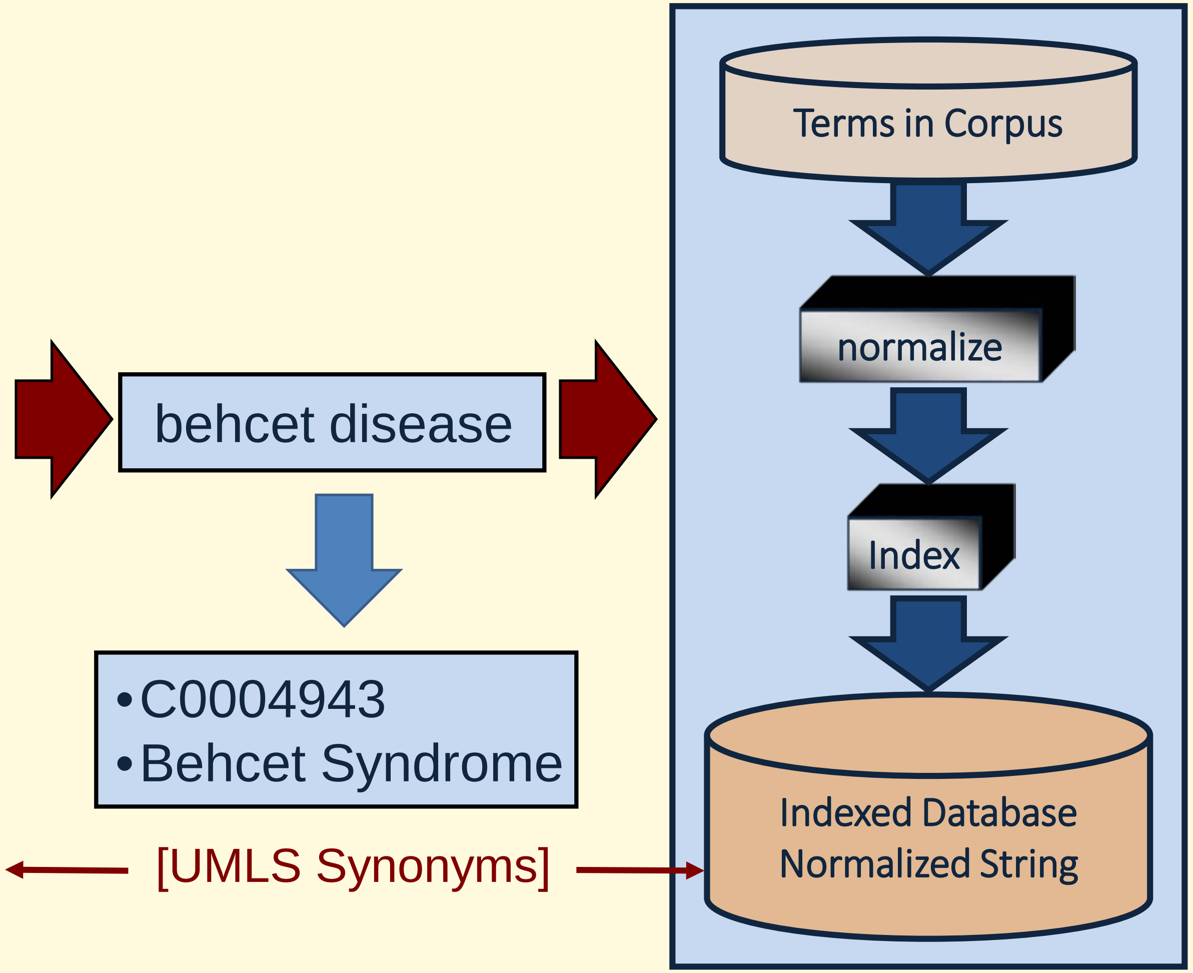
- Normalization (same record):
  - A term might have a great deal of lexical variations, such as inflectional variants, spelling variants, abbreviations (expansions), cases, ASCII conversion, etc.
  - Normalize different forms of a concept to a same form
- Query Expansion (related records):
  - Expand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

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

NLP – Norm (Lexical Variations)

- Behcet Disease
- Behçet’s Disease
- Behcet Diseases
- Behçet Diseases
- Behcet's Disease
- Behçet's Disease
- Behcets Disease
- Behçets Disease
- Behcet’s Disease, NOS
- Behçet’s Disease, NOS
- behcet disease
- behcet diseases
- behcet’s disease
- behcet’s disease, nos
- disease, Behçet
- diseases, behçet
- …



behcet disease

- C0004943
- Behcet Syndrome

[UMLS Synonyms]

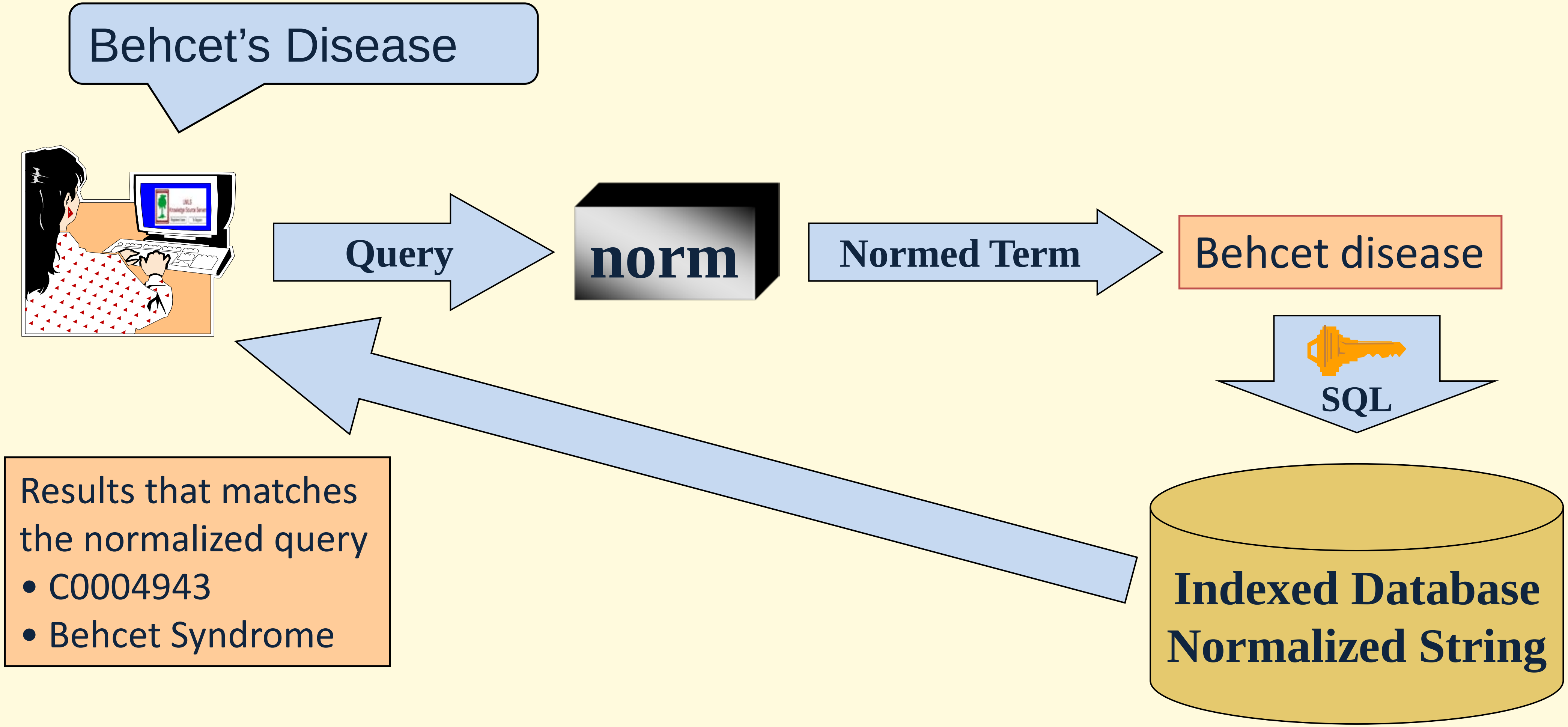
Terms in Corpus

normalize

Index

Indexed Database
Normalized String

NLP – Norm (Cont.)



Behcet’s Disease

Query

norm

Normed Term

Behcet disease

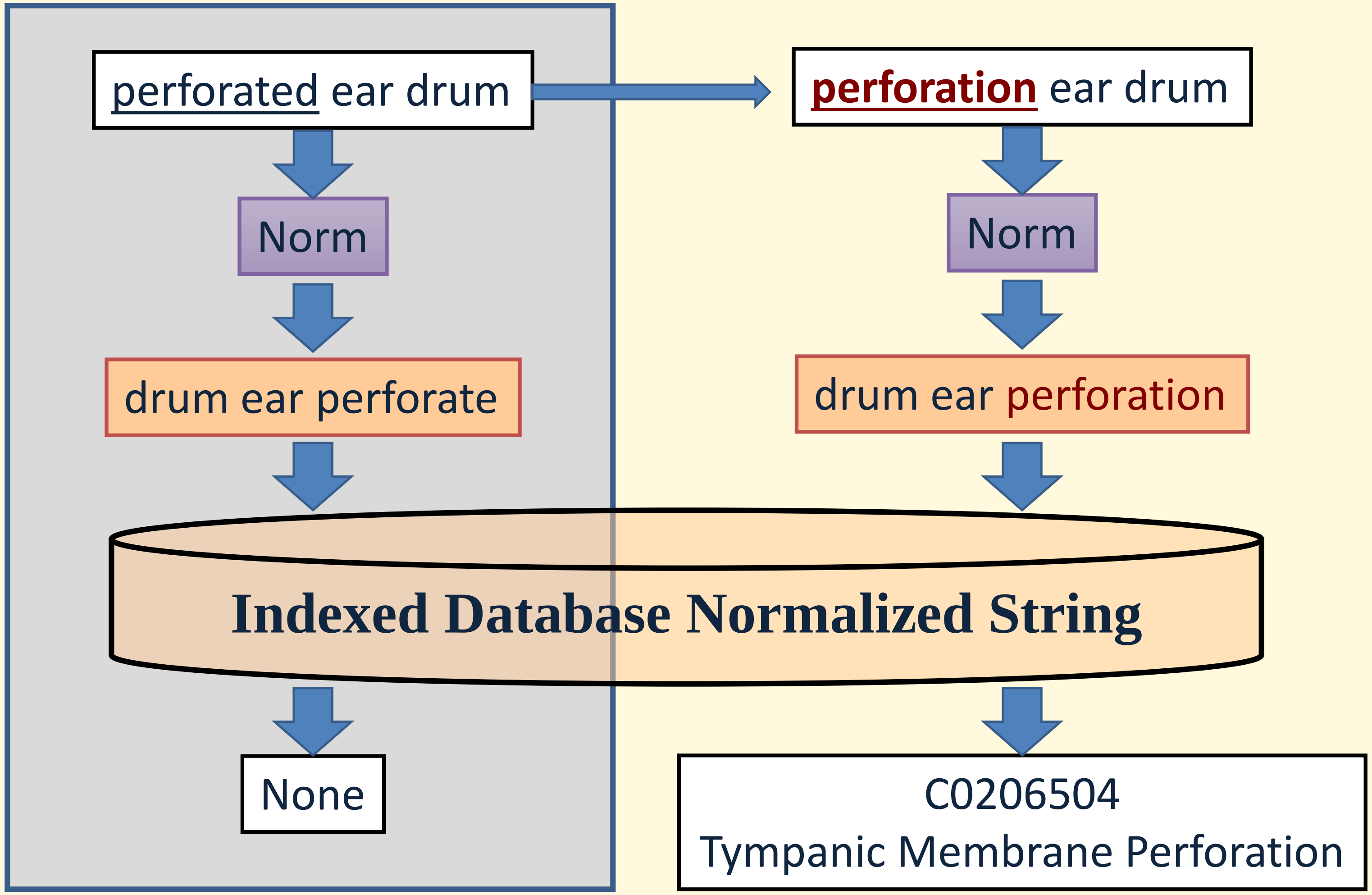
SQL

Indexed Database Normalized String

Results that matches the normalized query

- C0004943
- Behcet Syndrome

NLP – Query Expansion (derivation)



perforated ear drum

Norm

drum ear perforate

None

perforation ear drum

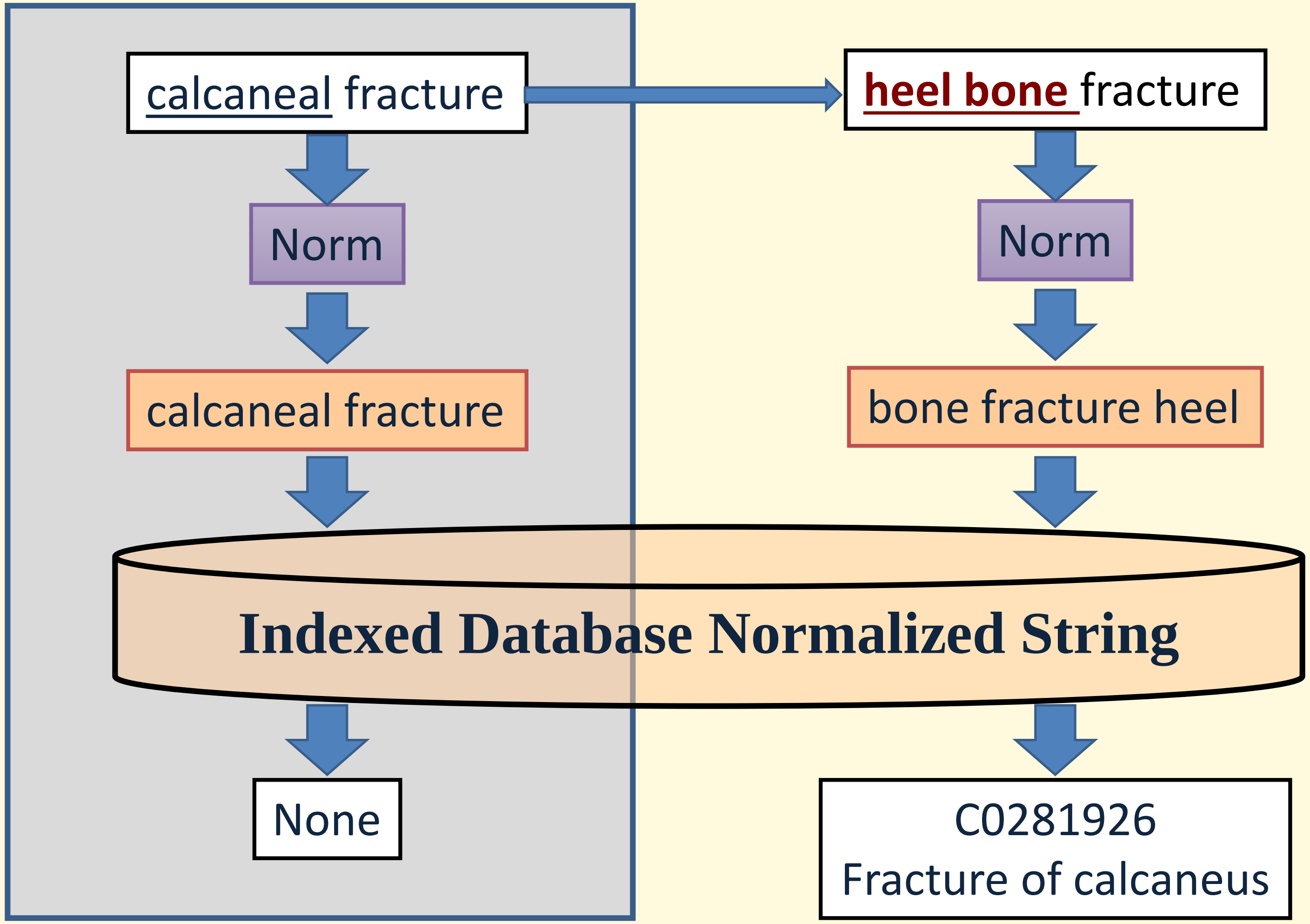
Norm

drum ear perforation

Indexed Database Normalized String

C0206504
Tympanic Membrane Perforation

NLP – Query Expansion (Synonym)



calcaneal fracture

Norm

calcaneal fracture

None

heel bone fracture

Norm

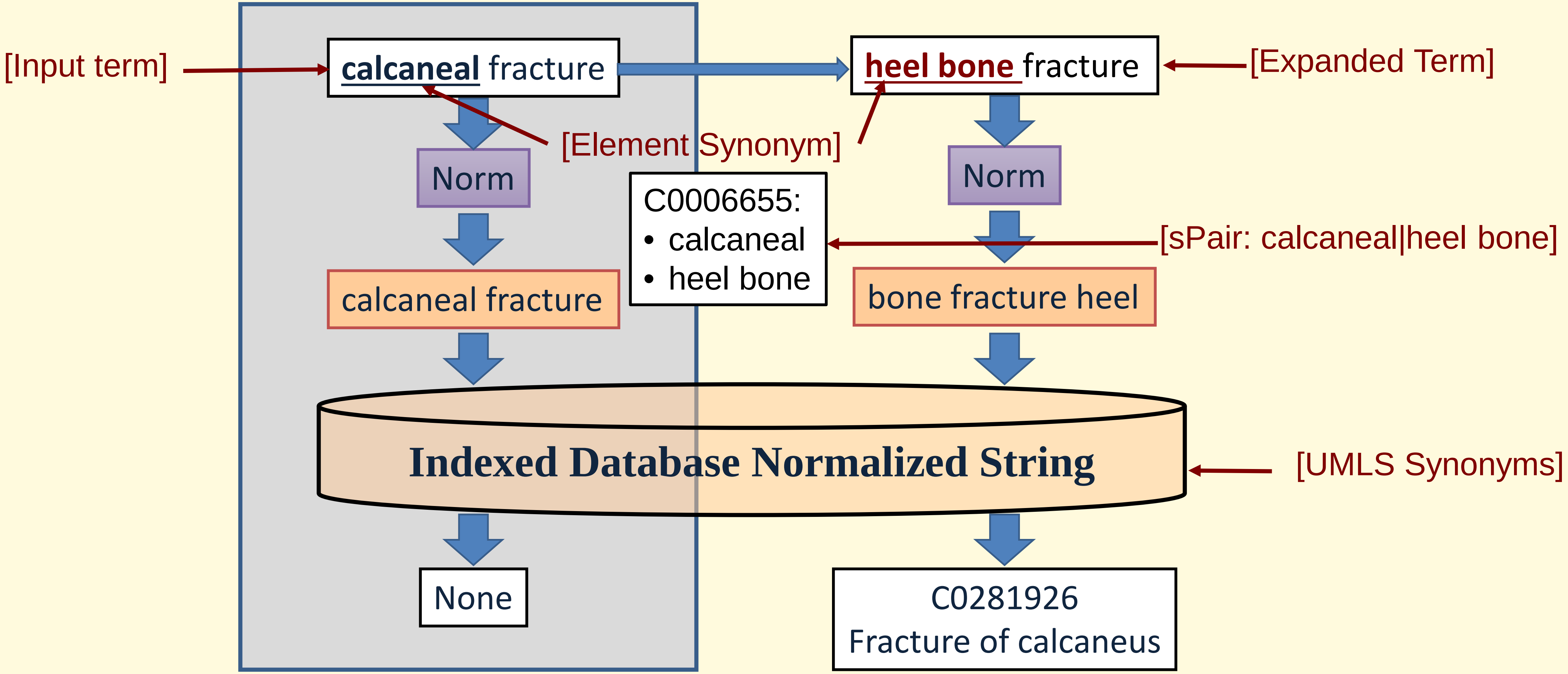
bone fracture heel

Indexed Database Normalized String

C0281926
Fracture of calcaneus

C0006655:
- calcaneal
- heel bone

NLP – Query Expansion (Synonym)



[Input term]

calcaneal fracture

[Element Synonym]

Norm

calcaneal fracture

None

C0006655:
- calcaneal
- heel bone

heel bone fracture

[Expanded Term]

Norm

[sPair: calcaneal|heel bone]

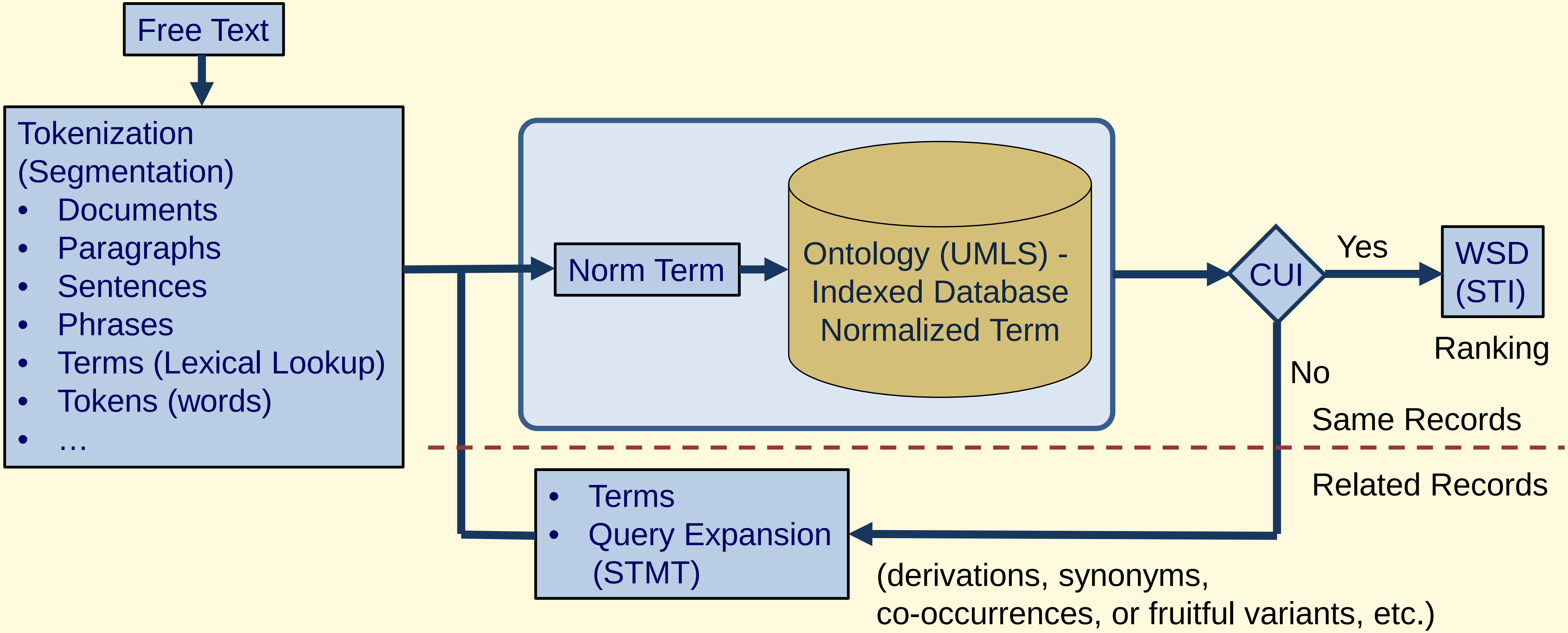
bone fracture heel

Indexed Database Normalized String

[UMLS Synonyms]

C0281926
Fracture of calcaneus

NLP – Concept Mapping Model



Free Text

Tokenization (Segmentation)

- Documents
- Paragraphs
- Sentences
- Phrases
- Terms (Lexical Lookup)
- Tokens (words)
- …

Norm Term

Ontology (UMLS) - Indexed Database Normalized Term

CUI

Yes

WSD (STI)

Ranking

No

Same Records

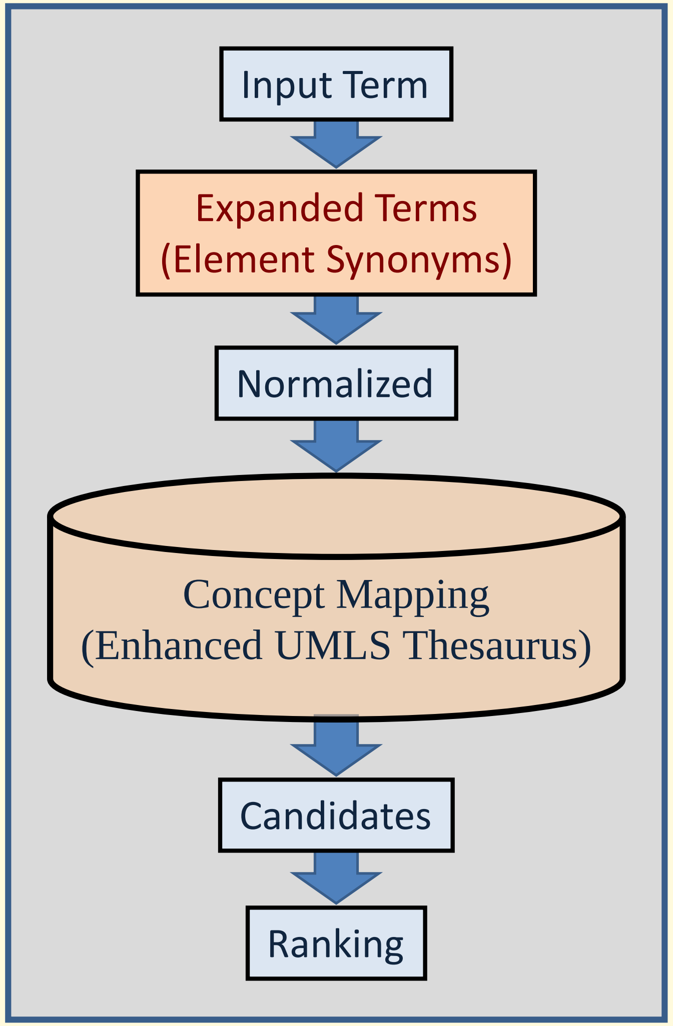
Related Records

- Terms
- Query Expansion (STMT)

(derivations, synonyms, co-occurrences, or fruitful variants, etc.)

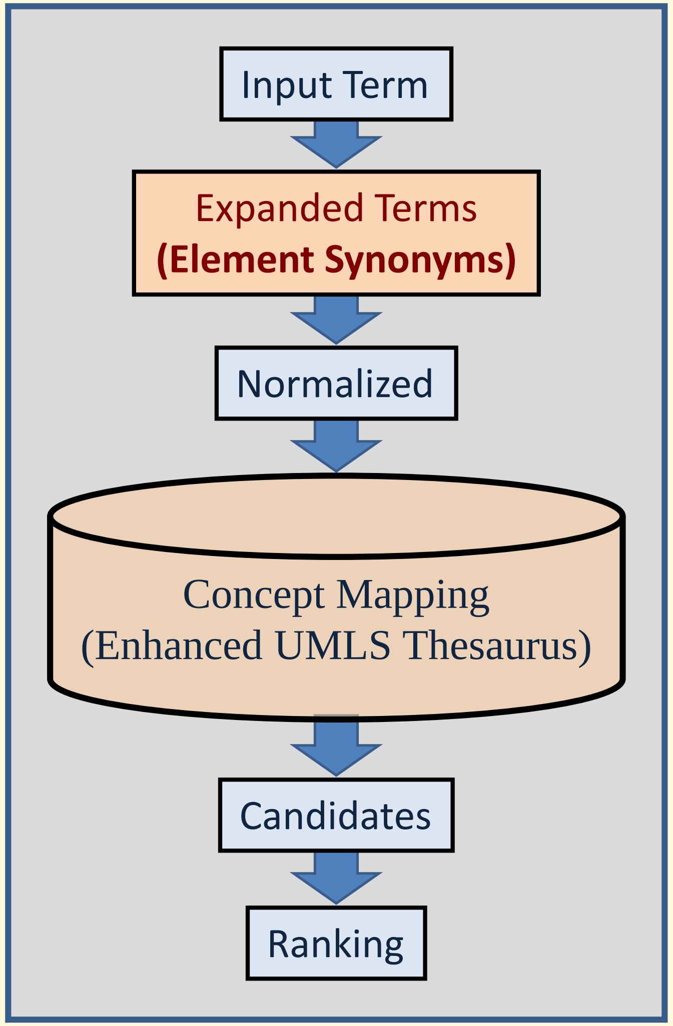
Generic Implementation

- Generates expanded terms of the input term
  - derivational variants
  - synonyms (recursive)
  - fruitful variants (combination of above)
- Normalization (lexical variants from the same record)
- Enhanced UMLS thesaurus
  - Pre-generated expanded term pool
  - Add new expanded terms (synonyms) to UMLS thesaurus
- Find candidates (mapped concepts)
- Ranking & Filters (keyword match, frequency, semantic types, concept distance, longest terms, etc.)



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



Element Synonyms Review

- ➢ UMLS Synonyms
- Semantically equivalent terms that have the same or very similar meaning (concept, CUI).
- 2016AA UMLS Metathesaurus containing over 3.25M concepts and nearly 13 M unique concept names from over 190 source vocabularies.

- ➢ The SPECIALIST Lexicon and Lexical Tools Synonyms, 2016- (~5K)
- ➢ UMLS-Core Projects (~12K)
- ➢ Synonym set by Randy Miller, (~15K)

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)
    - include all synonymous terms in Lexicon (LexSynonyms)
    - grow with the SPECIALIST Lexicon

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

Properties of Cognitive Synonyms (sPairs)

- Commutativity: (x = y) -> (y = x)
  - bi-directional
  - joy|noun|enjoy|verb -> enjoy|verb|joy|noun
- Transitivity: ((x = y) and (y = z)) -> (x = z)
  - enjoy|verb|joy|noun
  - joy|noun|happy|adj

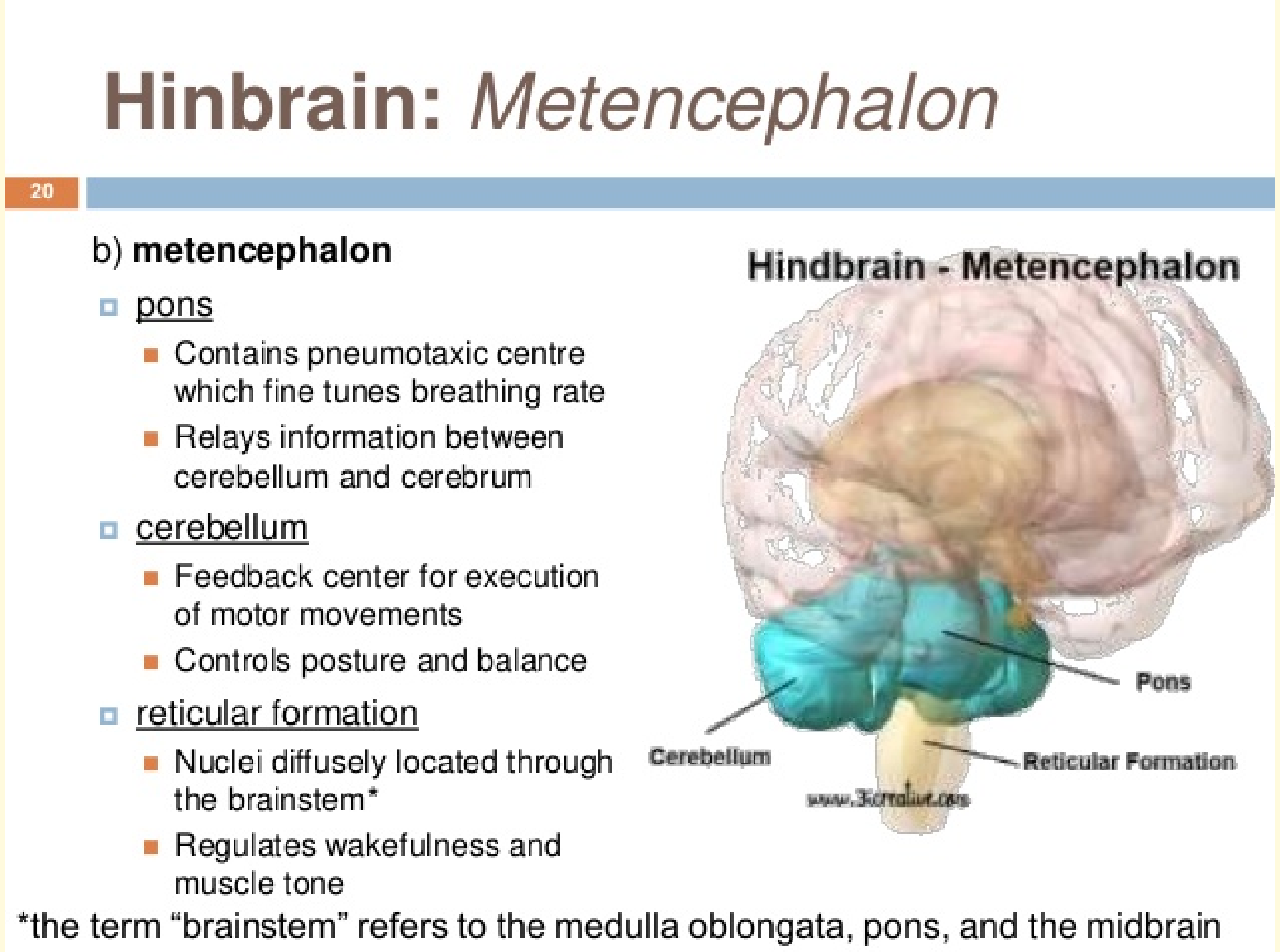
  ->

  - recursive
  - enjoy|verb -> joy|noun -> happy|adj
- Suitable for sPairs (element synonyms)
- Resolve many issues in element synonyms.

Broader/Narrower Issues – Near Synonyms

CUI	Preferred Term	synonym	Explanation
C0001613	Adrenal Cortex	cortical	The adjective cortical can refer to any of several types of cortex & so does not have synonymy with “adrenal cortex”
C0032639	Pontine structure	metencephalon	The metencephalon, per m-w.com includes the cerebellum and pons, and is different from the pons
C0001575	Uterine adnexae structure	adnexa	There are several types of adnexa, such as eye adnexa, adnexa of skin, etc.
C0000936	Visual Accommodation	accommodation	There are other accommodations.

Metencephalon & Pontine Structure (Pons)



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

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

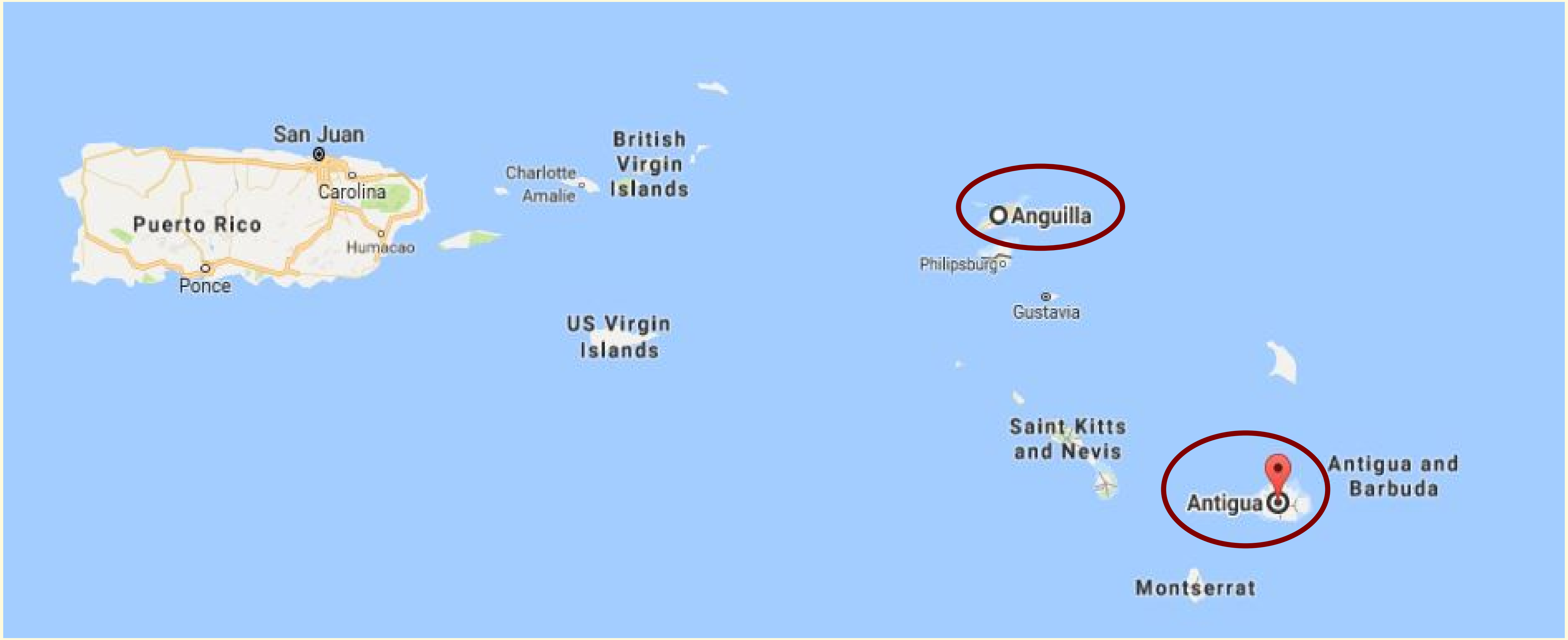
Distinct Issues – Similar but Different

CUI	Preferred Term	synonym	Explanation
C0000741	Abducens nerve structure	abductor digiti minimi	The abductor digiti minimi is a muscle, not a nerve.
C0003864	Arthritis	arthritide	Per Dorland's an arthride is "any skin eruption of arthritic or gouty origin.
C0005400	Bile duct structure	choledochitis	Choledochitis is a condition of the common bile duct, not structure.
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.

Acacia & Locust tree

Locust Tree

Anguilla & Antigua



Acronym/Abbreviation Issues – Precision

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

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

POS Issues – Meaning Shift

CUI	Preferred Term	synonym	Explanation
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.
C0004063	Assault	mug	The noun mug means a large cup, while the verb mug does refer to assault.

Recursive Issues – Multiple Concepts

- Multiple CUIs (transitivity?)
- Example (cold):

CUIs	Synonym
C0009443 common cold	• **cold** • coryza • acute coryza • common cold
C0009264 cold temperature	• **cold** • low temperature • low-temperature • lowtemperature
C0234192 cold sensation	• **cold** • psychroesthesia
…	• …

- common cold|cold
- cold|cold temperature
- cold|cold sensation

=> common cold|cold|cold temperature ?
=> common cold|cold|cold sensation?

Recursive Issue 2 – Endless loop

- Example – cold blooded animal

Synonyms	cold temperature	cold therapy	common cold	cold sensation
1-G substitution	cold temperature	cold therapy	common cold	cold sensation
2-G substitution	cold temperature temperature cold therapy temperature common cold temperature cold sensation temperature	cold temperature therapy cold therapy therapy common cold therapy cold sensation therapy	common cold temperature common cold therapy common common cold common cold sensation	…
…				

LexSynonyms – Objectives

- To establish a system to:
  - generate a standalone set of generic element synonyms (sPairs)
    - include all synonymous terms in Lexicon (LexSynonyms)
    - grow with the SPECIALIST Lexicon
  - use for effective UMLS concept mapping
    - a **thorough set** of element synonyms (to increase recall)
    - **cognitive synonyms** (to preserve precision)

LexSynonyms – Requirements

- Requirements (sClass):
  - All synonymous terms (cognitive synonyms) in the Lexicon
  - Bi-directional (commutativity) - interchangeable sPair in NLP
  - Recursive (transitivity) - use in NLP to improve Recall, yet preserve precision

- Resolve all above observed issues
  - Broader issues
  - Distinct issues
  - Acronym/abbreviation issues
  - POS issues
  - Recursive issues

Approach - Refined sClass

- 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)

sClass Example

```
#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

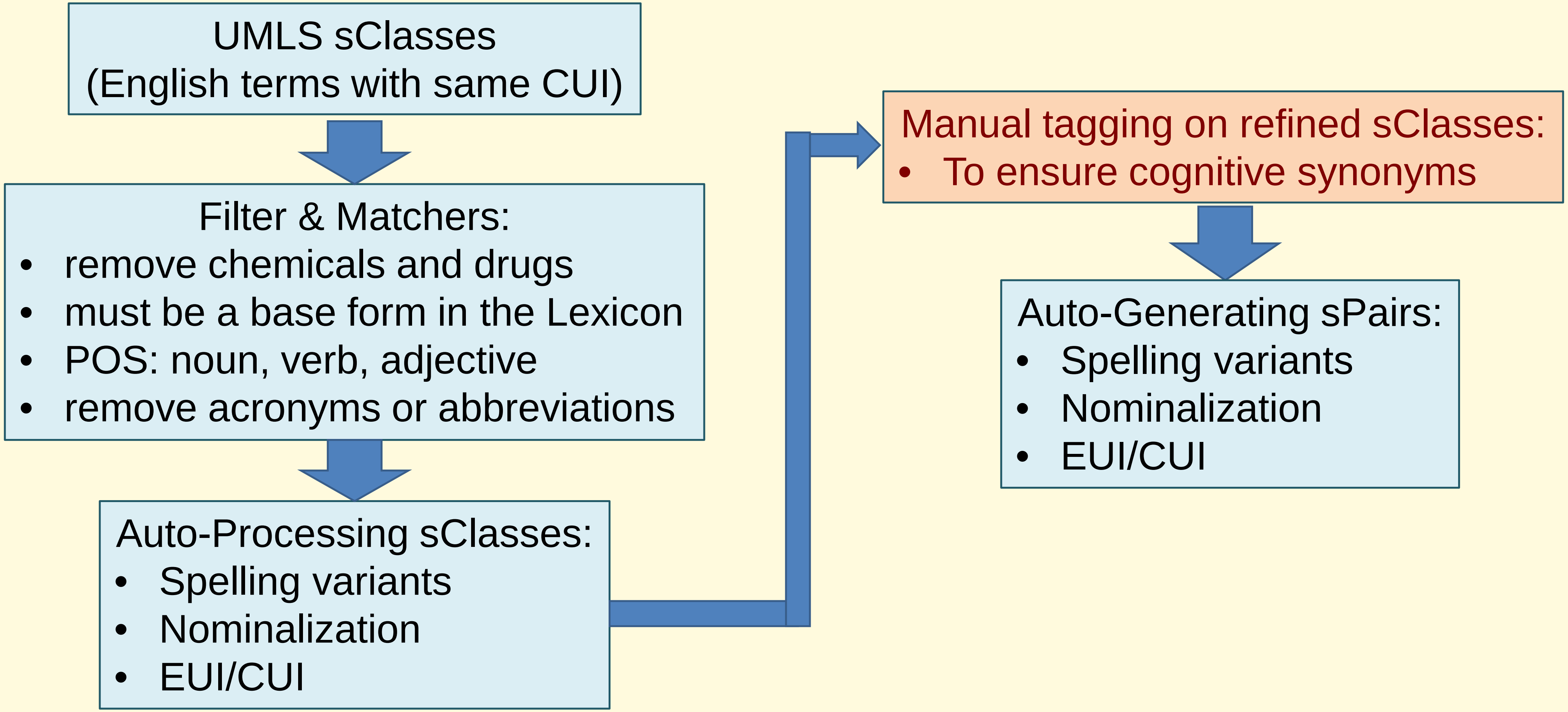
- ➢ NLP Projects-Sourced Cognitive Synonyms
  - legacy data (LVG, STMT, UMLS Core, …)
  - can be automatically retrieved
  - manually verified and add POS

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
}
```

UMLS-Sourced Cognitive Synonyms



UMLS sClasses
(English terms with same CUI)

Filter & Matchers:

- remove chemicals and drugs
- must be a base form in the Lexicon
- POS: noun, verb, adjective
- remove acronyms or abbreviations

Auto-Processing sClasses:

- Spelling variants
- Nominalization
- EUI/CUI

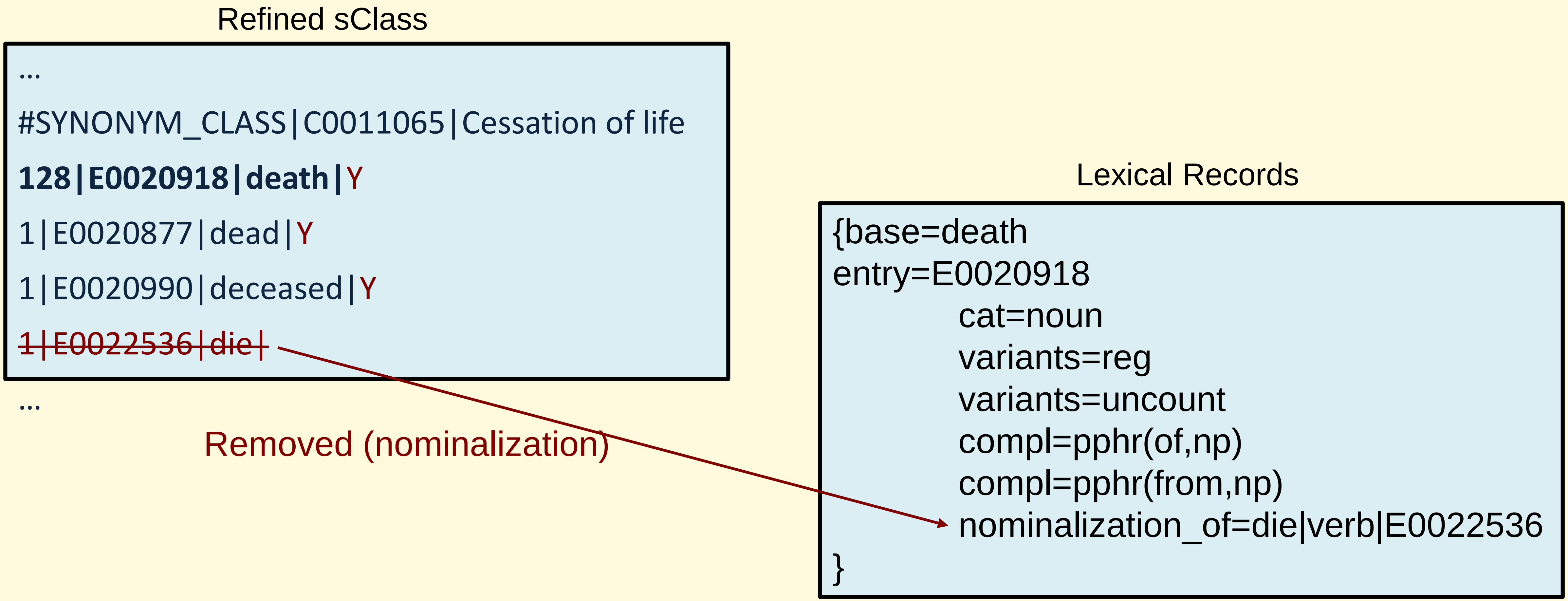
Manual tagging on refined sClasses:

- To ensure cognitive synonyms

Auto-Generating sPairs:

- Spelling variants
- Nominalization
- EUI/CUI

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|~~

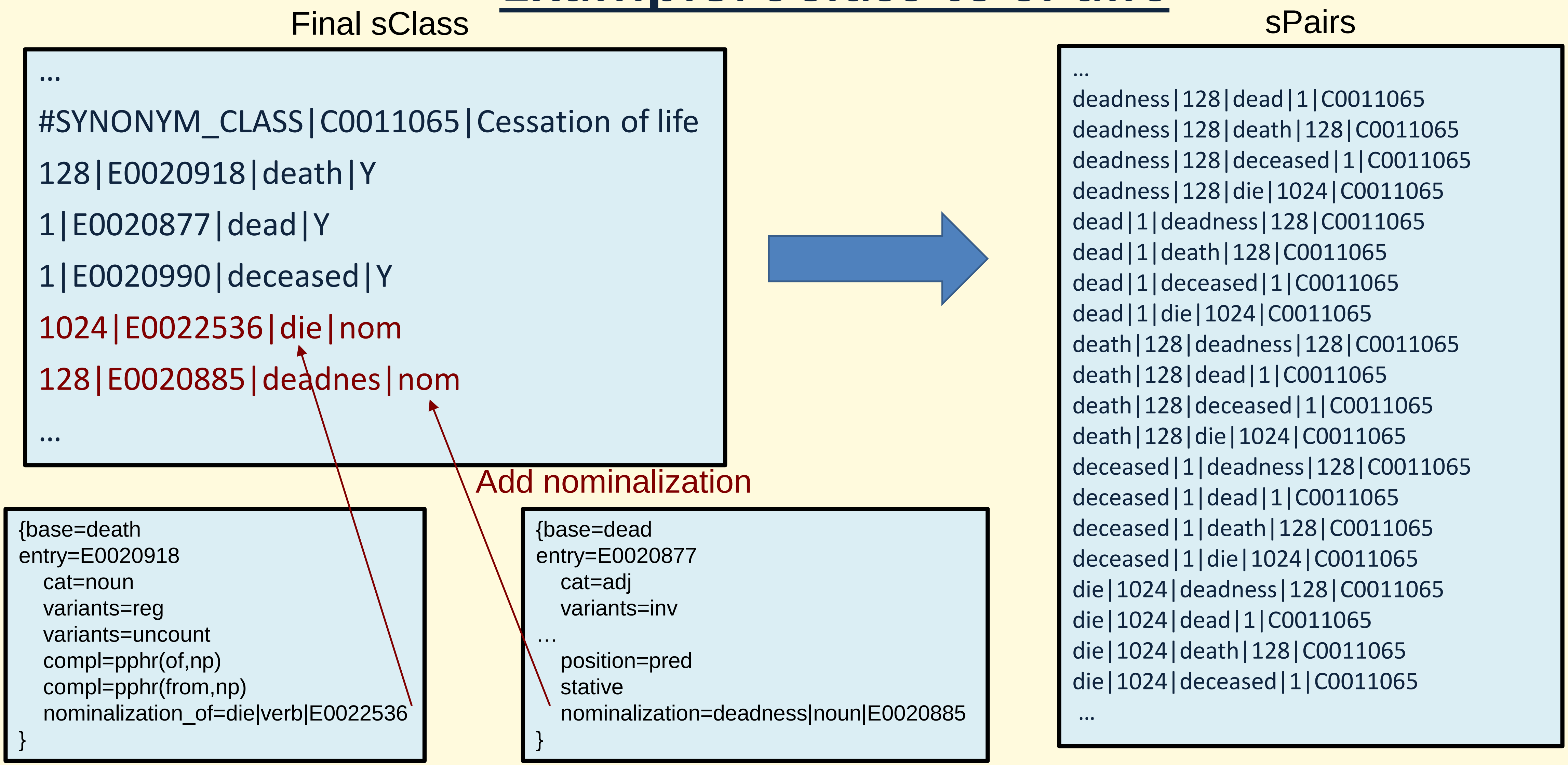
…

Removed (nominalization)

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
}
```

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
…
```

Add nominalization

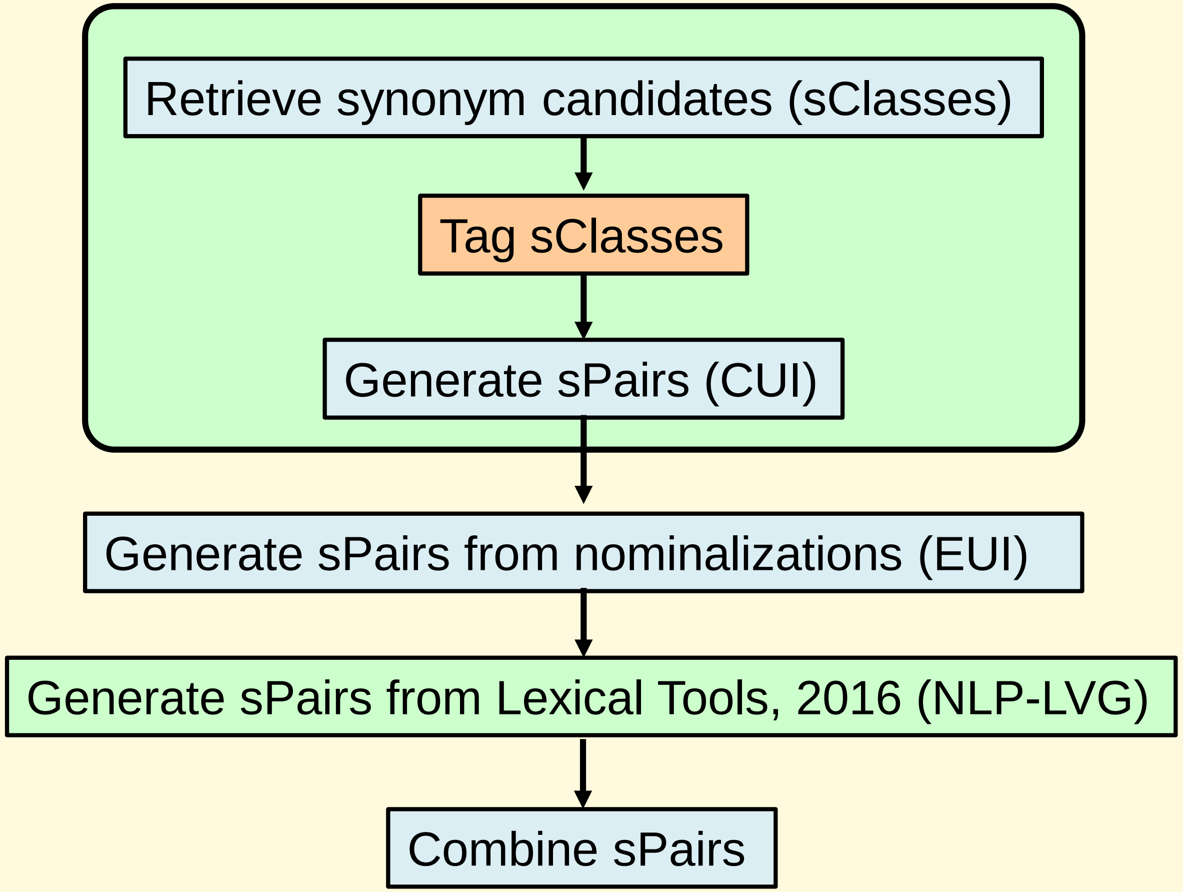
```
{base=death
entry=E0020918
    cat=noun
    variants=reg
    variants=uncount
    compl=pphr(of,np)
    compl=pphr(from,np)
    nominalization_of=die|verb|E0022536
}
```

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

sPairs

```
…
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
…
```

LexSynonym Generation



Retrieve synonym candidates (sClasses)

Tag sClasses

Generate sPairs (CUI)

Generate sPairs from nominalizations (EUI)

Generate sPairs from Lexical Tools, 2016 (NLP-LVG)

Combine sPairs

Results

- 2017 release:
  - 2016AB Metathesaurs, 2016 Lexicon

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

- Synonym stats:

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

Tests

- Model:
  - STMT (Sub-Term Mapping Tools):
    - Real-time subterm substitution tools for concept mapping
    - Easy configurable options for element synonym set
- Data:
  - UMLS-Core Project:
    - 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

Test Results

- Gold Standard: 2,755 terms mapped to 2,756 CUIs
- Element sets:
  - STMT: include a validated cognitive synonym set
  - 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	7,873	690	353	2,066	66.16%	25.04%	0.3633	7:57
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
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

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|
```

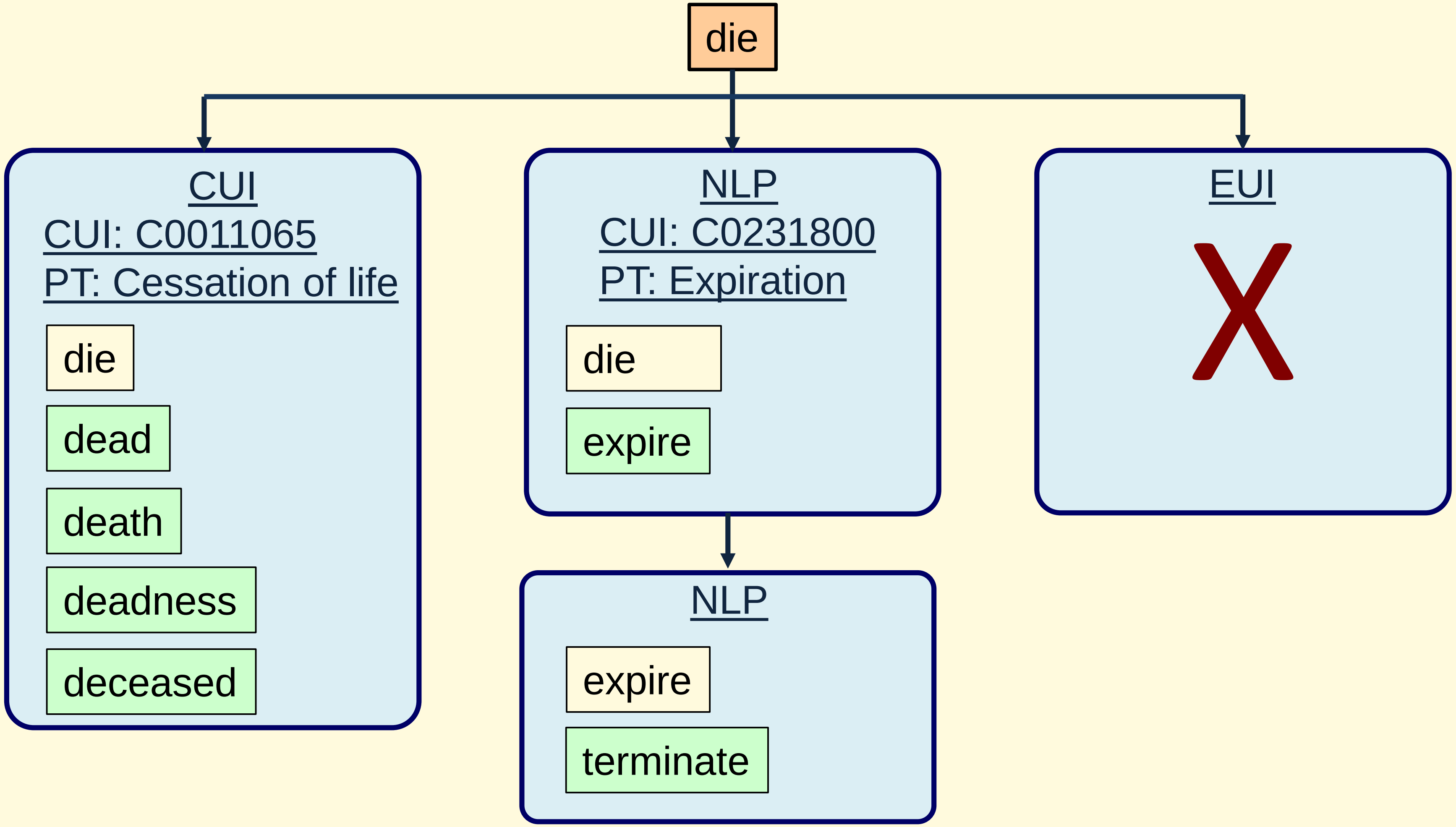
Lexical Tools – Synonyms Flow Options

- 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



die

CUI
CUI: C0011065
PT: Cessation of life

die

dead

death

deadness

deceased

NLP
CUI: C0231800
PT: Expiration

die

expire

NLP

expire

terminate

EUI

X

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

Goals	Check	Notes
Standalone element synonym set	Yes	
All synonymous terms in the Lexicon	Yes	~ 1/3 completed
Grows with the SPECIALIST Lexicon	Yes	
Over-generated issues	Resolved	Must be in the Lexicon (430K, ~2% of UMLS synonyms)
Single words and multiwords	Resolved	Bases in Lexicon include both
Broader issues	Resolved	Done in tagging (cognitive synonyms)
Distinct issues	Resolved	Done in tagging (cognitive synonyms)
Acronym/abbreviation issues	Resolved	Removed in sClass
POS issues	Resolved	Provide POS in sClass
Recursive issues	Resolved	Provide source in sClass (CUI, EUI, etc.)
Improve NLP performance	Yes	Improve recall and preserve precision

Future Work

- Complete tagging on all sClasses
  - 2017 release: 2016AB Metathesaurs, 2016 Lexicon

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

- Updated annually on Lexicon and Lexical Tools release
  - with the latest Lexicon and Metathesaurus
- Computer-aided (automatic) process on the sClass tagging
- Add more synonyms from other NLP projects (UMLS-Core, Randy Milller, etc.)
- Performance tests on NLP applications

Questions

Text Processing
Lexical Tools
SPECIALIST LEXICON

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