

The SPECIALIST Lexicon and NLP Tools

By Dr. Chris J. Lu
NIH/NLM/LHNCBC/ACIB

April 29, 2025

Clinical Data Management and Analytic Team (CDMAT) Meeting
Lexical Systems Group: <https://lhncbc.nlm.nih.gov/LSG/index.html>



National Library of Medicine
Lister Hill National Center for Biomedical Communications

Outlines

- Natural Language Processing (NLP) Overview P:01-10
- Introduction
 - The SPECIALIST Lexicon P:11-24
 - The SPECIALIST Lexical Tools P:25-30
- NLP Tools
 - LVG P:31-37
 - Norm P:38-57
 - Derivations, Synonyms, Antonyms P:58-71
 - Multiwords P:72-74
 - CSpell P:75-78
- Questions (anytime) P:79

Natural Language Processing (NLP)

- Natural Language:
 - the ordinary language that humans use naturally.
 - may be spoken, **written** or signed.
 - communication and understanding.
- Natural Language Processing (NLP):
 - NLP in our scope is to use computer to **understand the meaning** (concept) from free **text** for further analysis and processing.
 - **The use of computer to process and analyze text for various applications.**
 - NLP includes a board range of subjects, require knowledge from linguistics, computer science, and statistics (data science), ML/DL NN and LLMs.

NLP Applications in Healthcare Research

- Information retrieval:
 - to retrieve relevant research studies on COVID-19 treatment (keyword search?)
- Information extraction:
 - to find specific information (treatments) from unstructured text (clinical notes).
- Text summarization:
 - to extract the most relevant information for quick review, such as on new medication.
- Text classification:
 - to automatically sorting content into predefined categories
 - e.g. to triage patients with emergency levels (1-5) prior to their admission for treatment at the Accident & Emergency Department (AED).
- Sentiment analysis:
 - to identify patients' emotion (feedback, review) on treatment or medication.
- Question answering:
 - to provide a specific answer to a question, e.g. consumer health.
- ...

Search by Concepts

heart disease
C00187799



heart disease
Heart disease
Heart Disease
HEART DISEASE
heart diseases
HEART DIS
Disease;heart
Disease hearts
Heart Disease, NOS

heart disorder
cardiac disease
diseases of the heart
disorder of heart
disorder cardiac
cardiopathy
morbus cordis
syndrome heart disease

- **Lexical variants:**
 - including case, inflectional variant, spelling variants, word order, abbreviation, acronyms, stopwords, etc.
 - can be handled by normalization.
- **Lexical thesaurus:**
 - including synonyms, derivational variants, antonyms (with negation).
 - Can be handled by query expansion.

Over 67 terms in UMLS found for the same concept of C00187799

Challenges in Concept Mapping

- Challenge: many to many mapping (ambiguity)

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• Normalization• Query Expansion• LLMs
<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)• Text categorization• Context dependent• LLMs

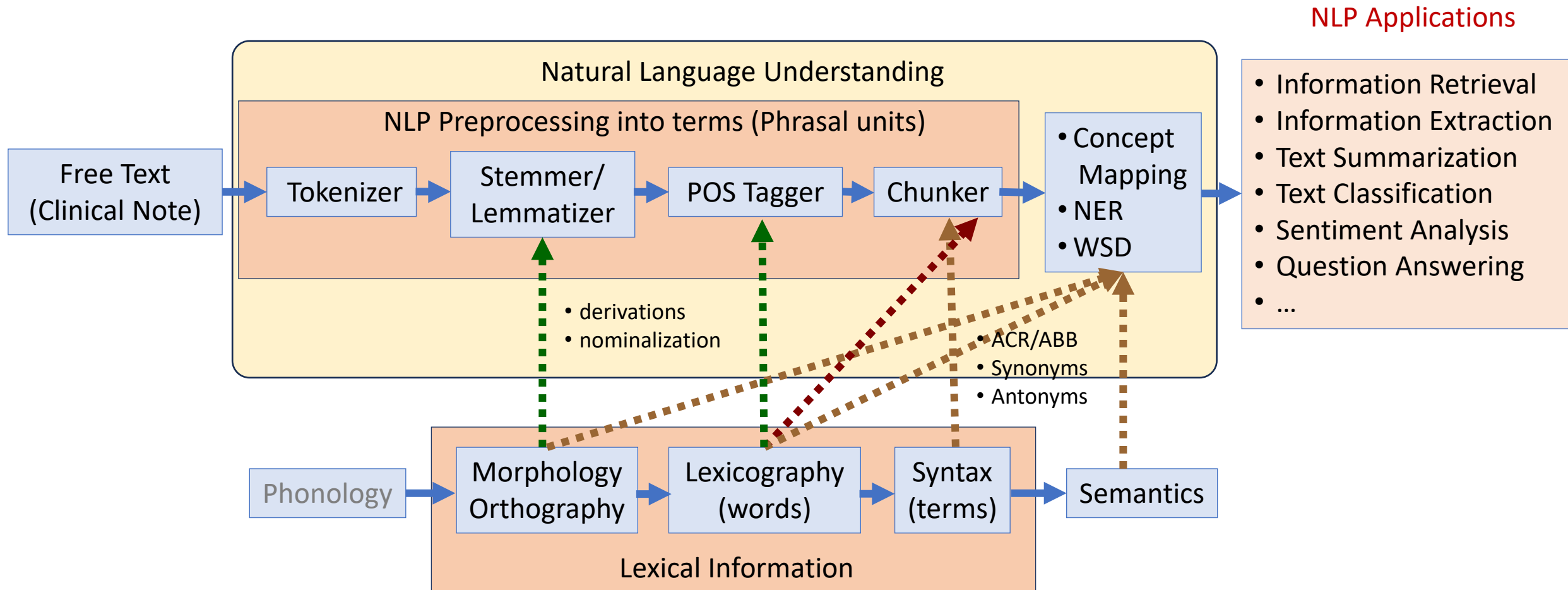


NLP Techniques

- Linguistics:
 - Tokenization
 - Multiword Expression (WME)
 - Stopword & punctuation removal
 - Stemming (uninflect) or lemmatization (inflect)
 - Normalization
- Syntax:
 - Parsers, taggers, POS tagging, chunker, etc.
- Semantics:
 - Concept mapping (meaning)
 - Query Expansion (increase recall)
 - Name Entity Recognition (NER)
- Spell checking and correction



NLP Pipeline & Lexical Information

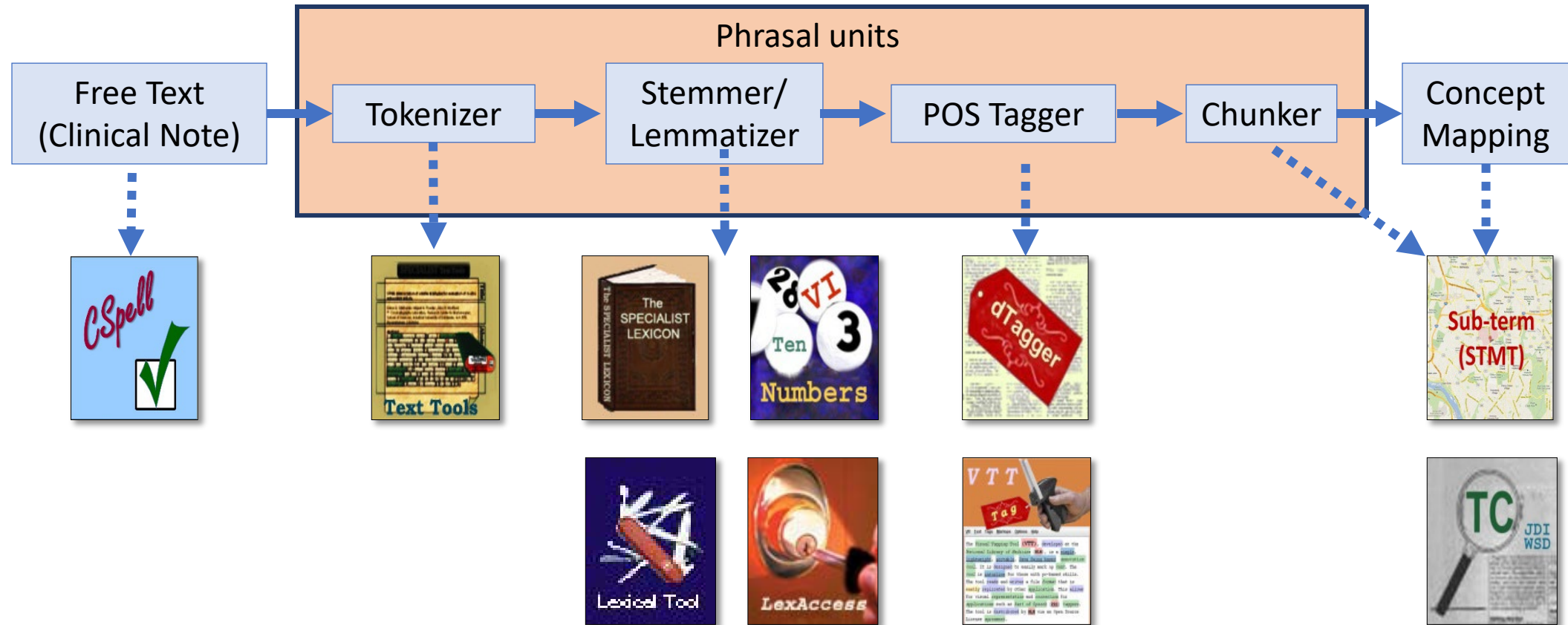


The NLP Pyramid



- **Applications:**
Information retrieval, information extraction, text summarization, text categorization, sentiment analysis, question answering, etc.
- **Semantics (concept & meaning):**
Named Entity Recognition (NER), concept mapping, relation extraction, Semantic role labelling, Word Sense Disambiguation (WSD)
- **Syntax (proper word construction):**
POS tagging, derivations, synonyms, antonyms, multiwords.
- **Morphology & Orthograph:**
prefixes/suffixes (derivation), stemming (inflect), lemmatization (uninflect), spellchecking, spelling variants, gender detection.

The SPECIALIST NLP Tools



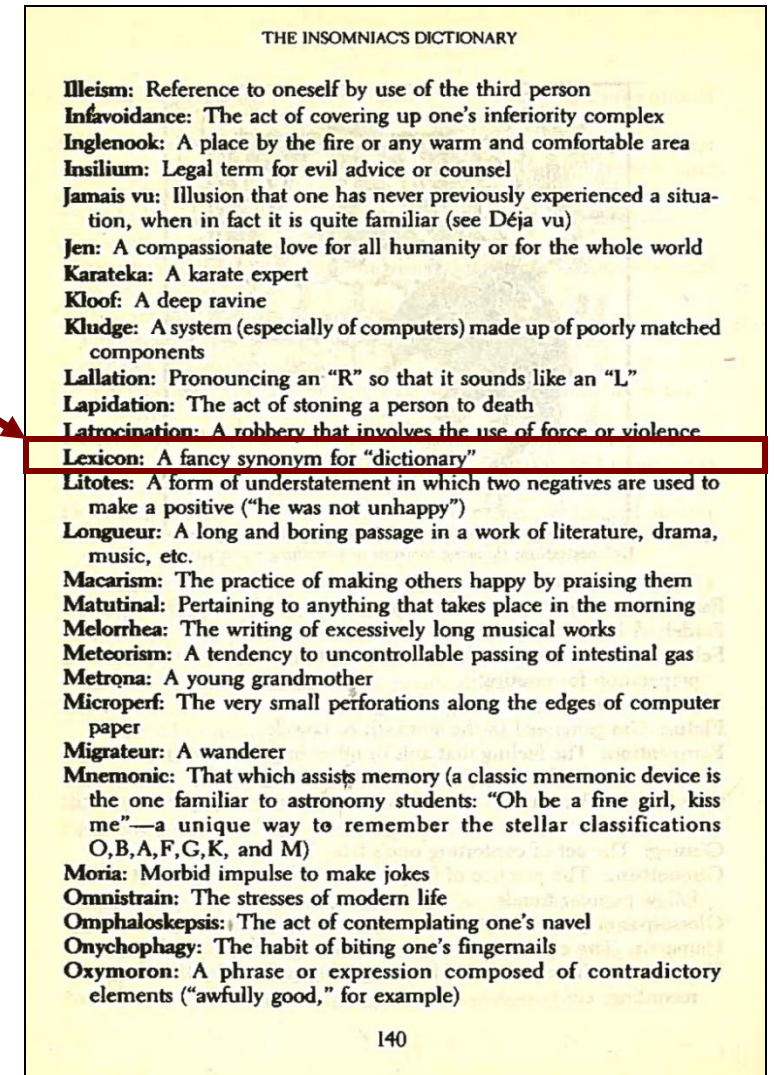
Unified Medical Language System (UMLS)

- UMLS: is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems.
- Three UMLS Knowledge Sources
 - Metathesaurus
 - Semantic Network
 - The SPECIALIST Lexicon and Lexical Tools
- Usages:
 - 5K+ UMLS users and 10K+ bibliographic records for UMLS related publications*
- Annual and semi-annual releases by the National Library of Medicine (NLM).
- Betsy HL, Fiol GD, Hua Xu H. The UMLS knowledge sources at 30: indispensable to current research and applications in biomedical informatics. JAMIA, 2020; 27(10): 1499–1501.



The SPECIALIST Lexicon

- A fancy synonym for “dictionary”
- A syntactic lexicon
- Biomedical and general English
- Over 0.53 M records, 1.2 M words (POS + forms)
- Designed/developed to provide the lexical information needed for the UMLS NLP systems



LexBuild Process (Computer-Aided)

Sources:

- Word candidates from MEDLINE
- Words from consumer data
- Words from Covid-19
- 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
- ...

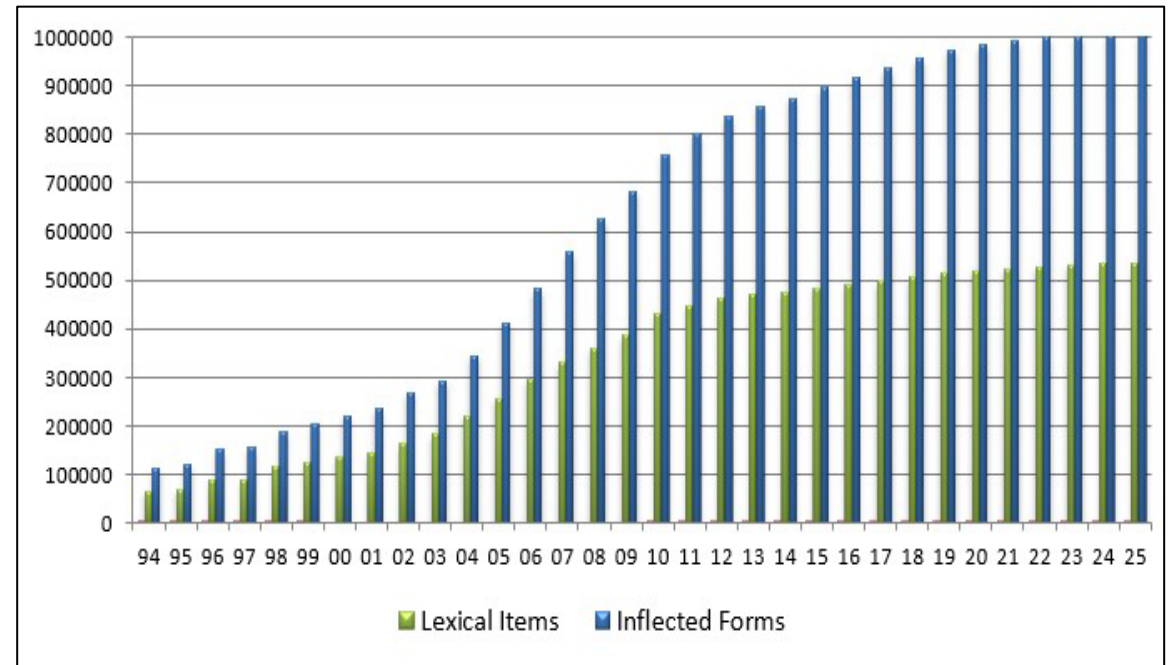
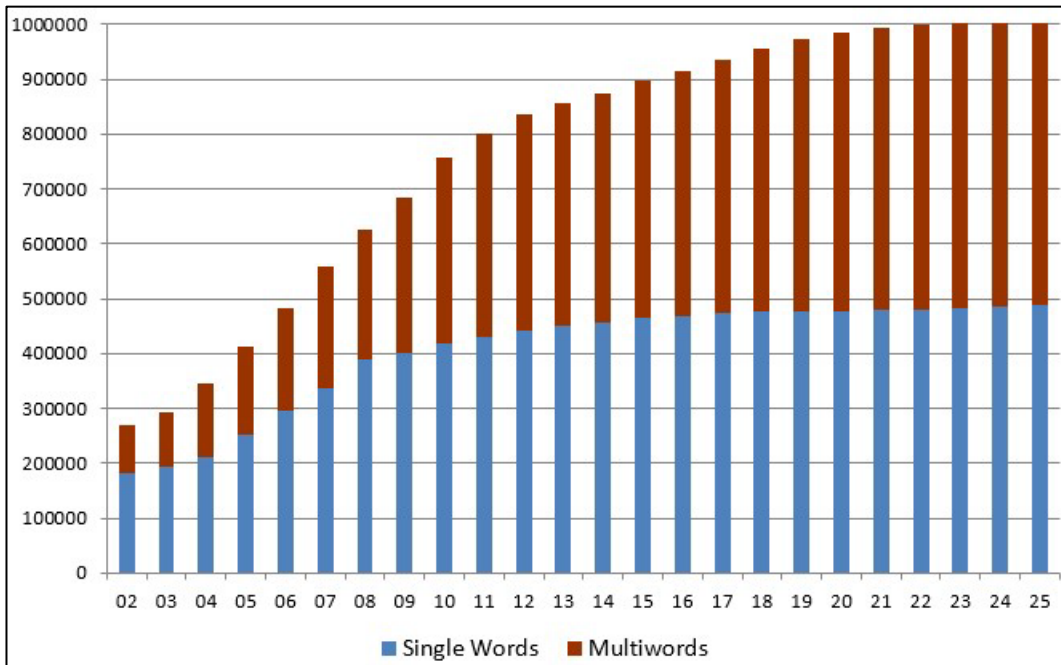


Build:

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

Lexicon Growth – 1994 to 2025

- 1,007,634 forms (orthographic words - spelling only)
 - Single words: 488,199 (48.45%); Multiwords: 519,435 (51.55%)
- 534,193 lexical records (lexical Items)
- 1,193,172 words (categories and inflections)



What Is a (Multi)-Word?

- A word is smallest unit of language that has meaning.
 - **Spelling** - Orthographic words:
 - SpVar: color vs. colour
 - Inflection (noun): dog vs. dogs
 - Inflection (verb): see vs. saw
 - Inflection (noun): saw vs. saws
 - POS: **saw** (verb) vs. **saw** (noun)
 - **Space** - Single words vs. Multi-words:
 - use spaces as word boundary
 - ice-cream vs. ice cream - space

Terms in the Lexicon

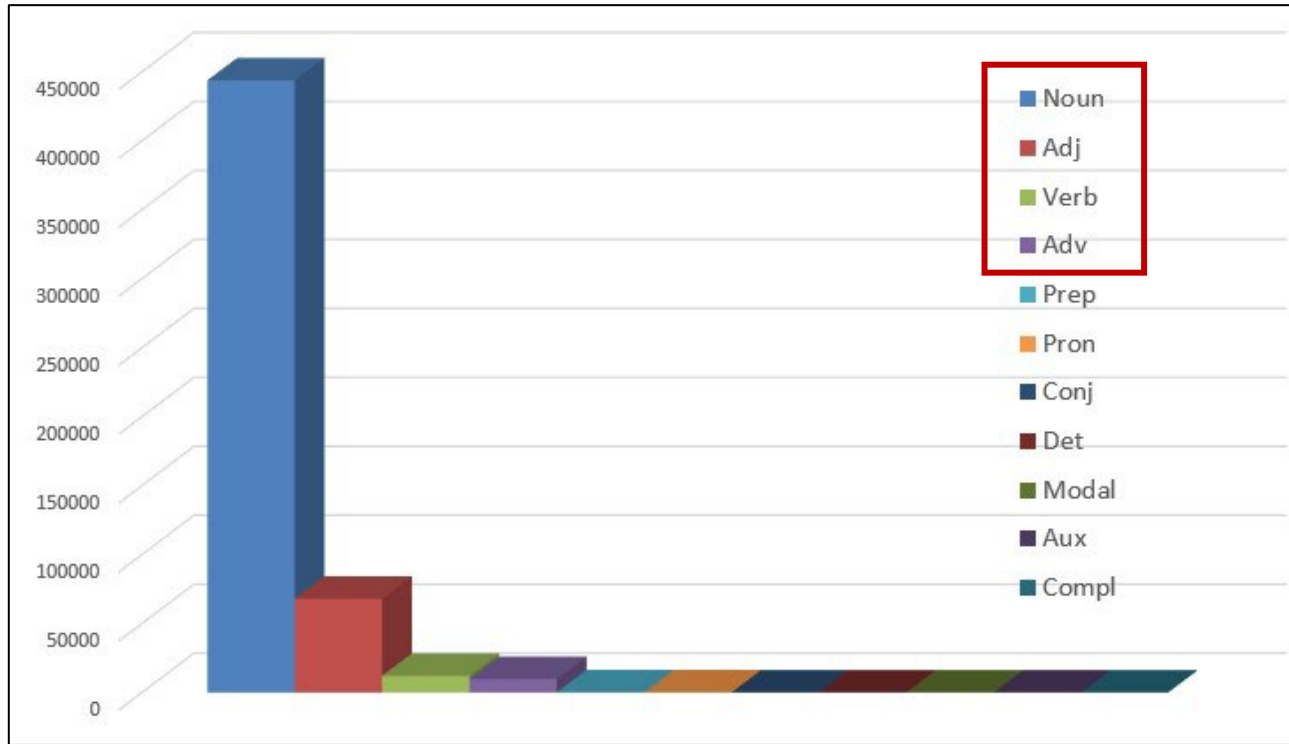
- Lexicon terms: single words and multiwords
 - Space(s): ice-cream vs. ice cream, tradeoff vs. trade-off vs. trade off.
- Four criteria for terms in the Lexicon:
 - Part of Speech (POS):
 - tear break up time, cardiac surgery, frog erythrocytic virus.
 - Inflection morphology (uninflection):
 - left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

Lexical Records (Items) – Lexical Information

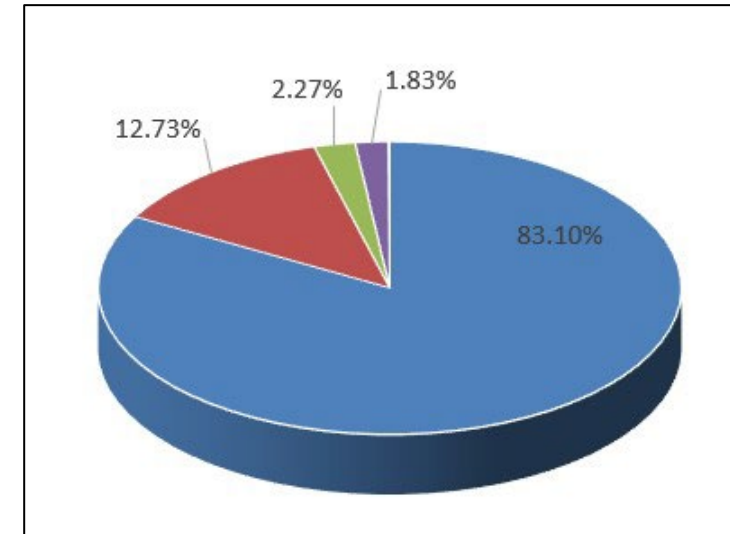
```
{base=color  
spelling_variant=colour  
entry=E0017902  
    cat=noun  
    variants=uncount  
    variants=reg  
}
```

- 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)



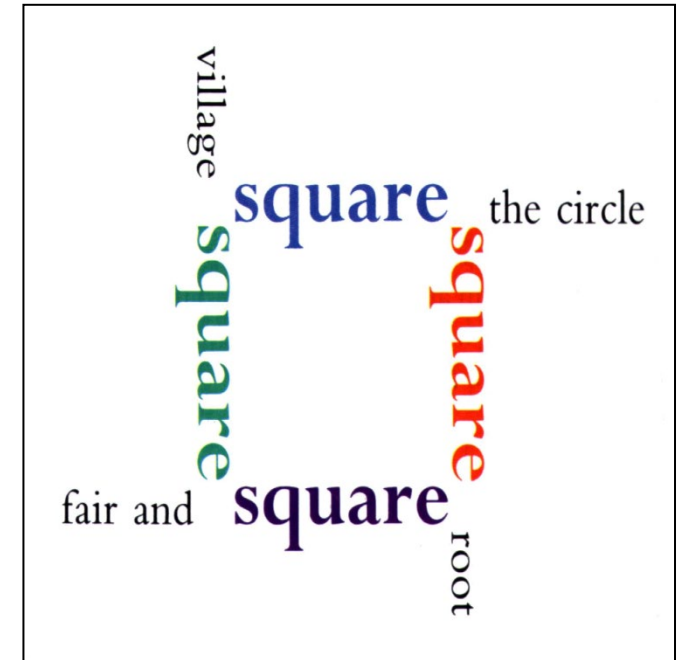
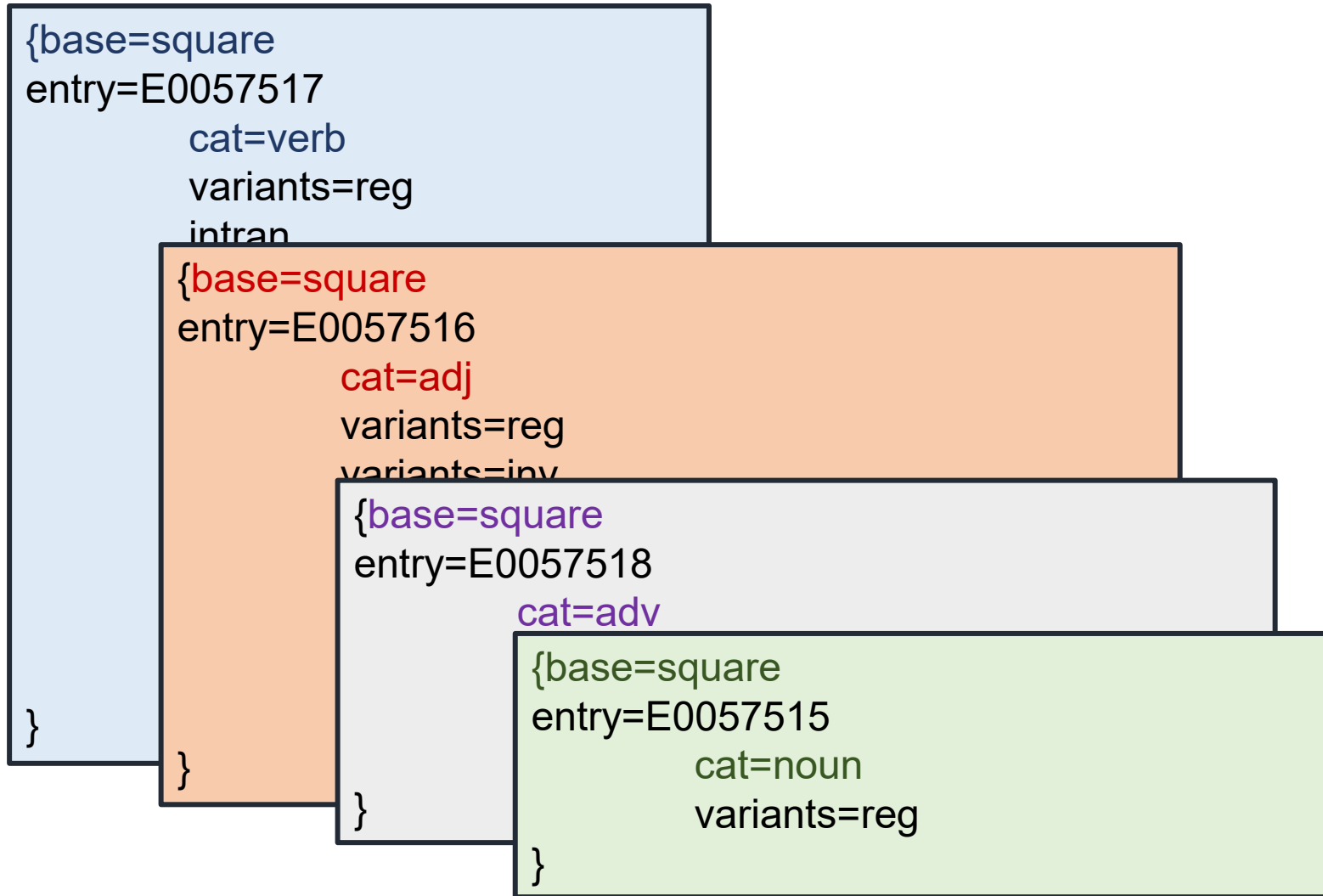
Lexicon.2025



Closed class POS (7) < 0.08% (static):

- Preposition: in, on, at
- Pronoun: it, he, they
- Conjunction: and, but, or
- Determiner: a, the, some, each
- Modal: shall, may, must, dare
- Auxiliary: be, am, is, do, does
- Complementizer: that

Lexical Records & POS



Morphology

- Inflectional Morphology
 - noun: book, books
 - verb: hide, hides, hid, hidden, hiding
 - adj: red, redder reddest
- Derivational Morphology
 - 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 | civilise
- harbor | harbour
- fetus | foetus | foetus
- centre | center
- spelt | spelled
- ice cream | ice-cream
- xray | x-ray | x ray



Syntax - Verb Complements

- Intran
 - intransitive: no direct object
 - I'll treat.
- tran=np
 - transitive: direct object
 - He treated the patient.
- ditran=np,pphr(with,np)
 - Ditransitive: indirect object + direct object
 - He treated the patient with the drug.
- ...

Lexical Information to Coded 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• ...
Synonyms	<ul style="list-style-type: none">• chromatic
Antonyms	<ul style="list-style-type: none">• black-and-white



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



Relationship

among lexRecords



```
{base=colorful
spelling_variant=colourful
entry=E0017909
    cat=adj
    variants=inv;periph
    ...
}
```

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  
}
```


The Lexicon (Data) and Lexical Tools (Software)



20+ LR DB Tables



```
{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

The SPECIALIST Lexical Tools

- Lexical Tools: Algorithm + Data (directly or derived from the Lexicon)
 - Command line tools (6)
 - 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



The Lexical Tools - Facts

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



Functions of The Lexical Tools

- 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,	...

Generated Lexical Variants

LexRecord: E0029526 | generalise | verb

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

← A LexRecord

Inflectional variants:

- generalises, generalised, generalising

← A LexRecord + Algorithm

Derivational variants:

- suffixD: generalis**ation**, generaliz**ation**, generalis**able**
- prefixD: **over**generalise, **over**-generalise

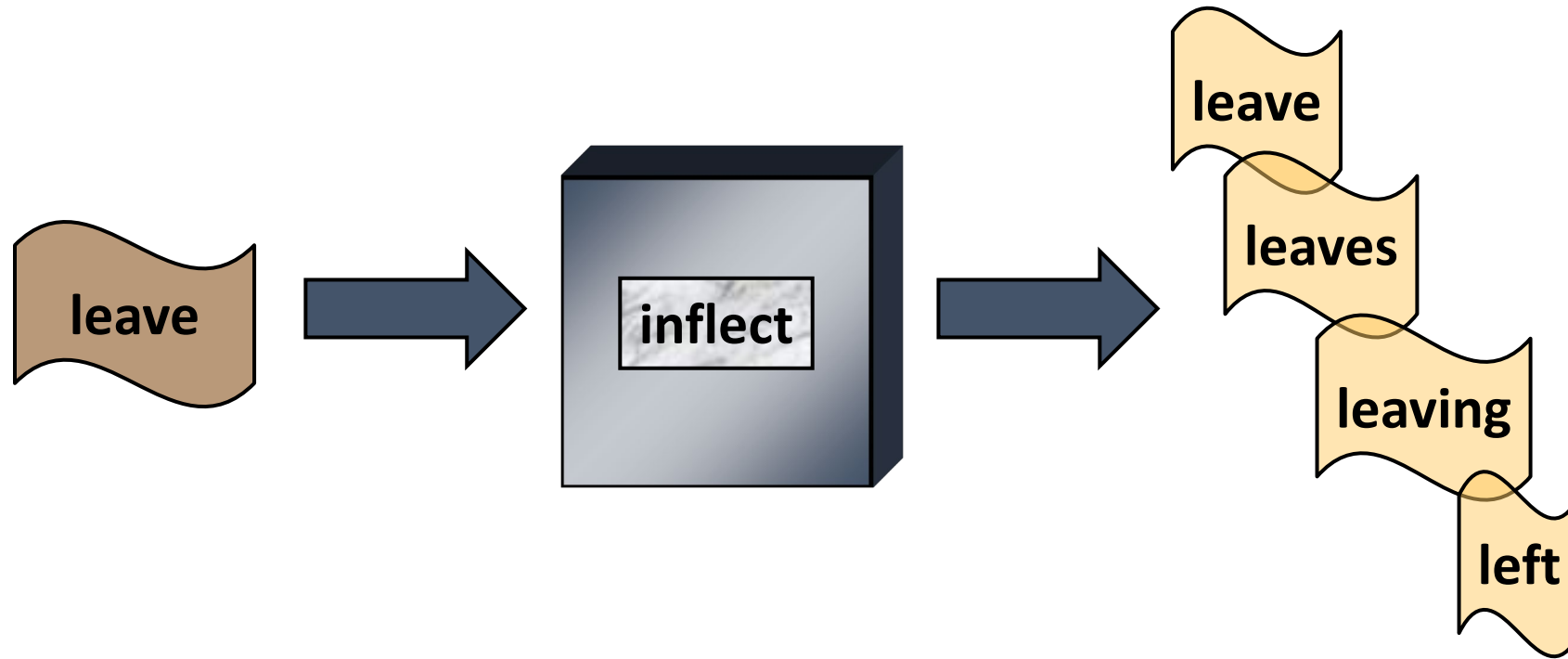
Synonyms: generalize

Antonyms: none

← Multiple LexRecords + Algorithm

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

Example - LVG Flow Component



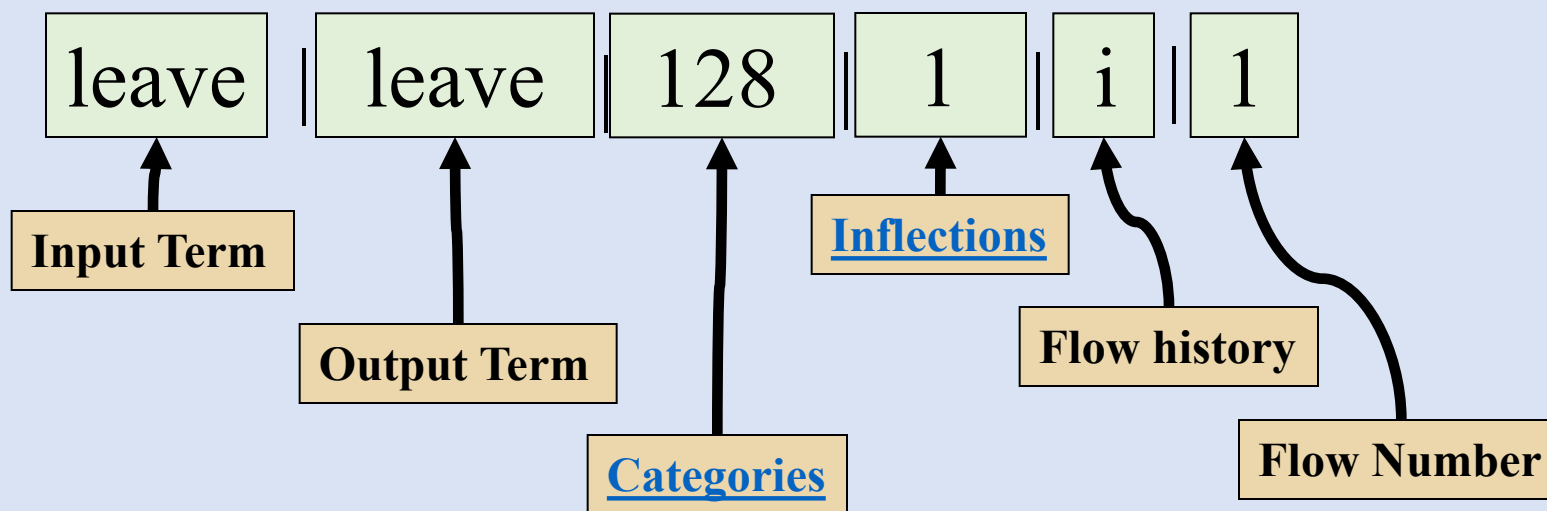
Example - LVG CmdLine

```
> 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|
```

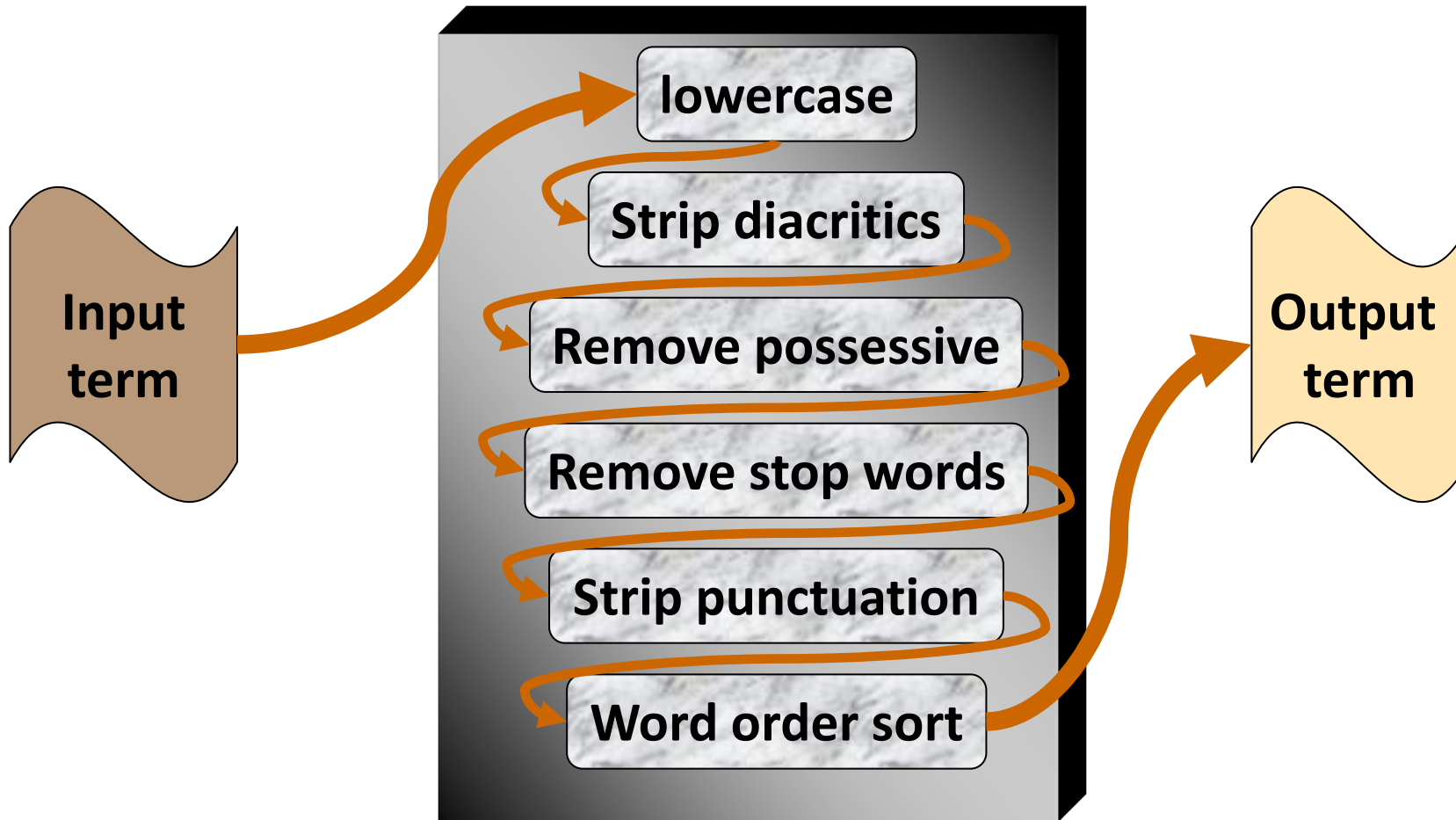


LVG Flow Component – Fielded Output

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



LVG – A Serial Flow



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

Example - A Serial Flow

```
> lvg -f:l:q:g:t:p:w
```

The Gougerot-Sjögren's Syndrome

The Gougerot-Sjögren's Syndrome |

gougerotsjogren syndrome |

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

← Input

← Output

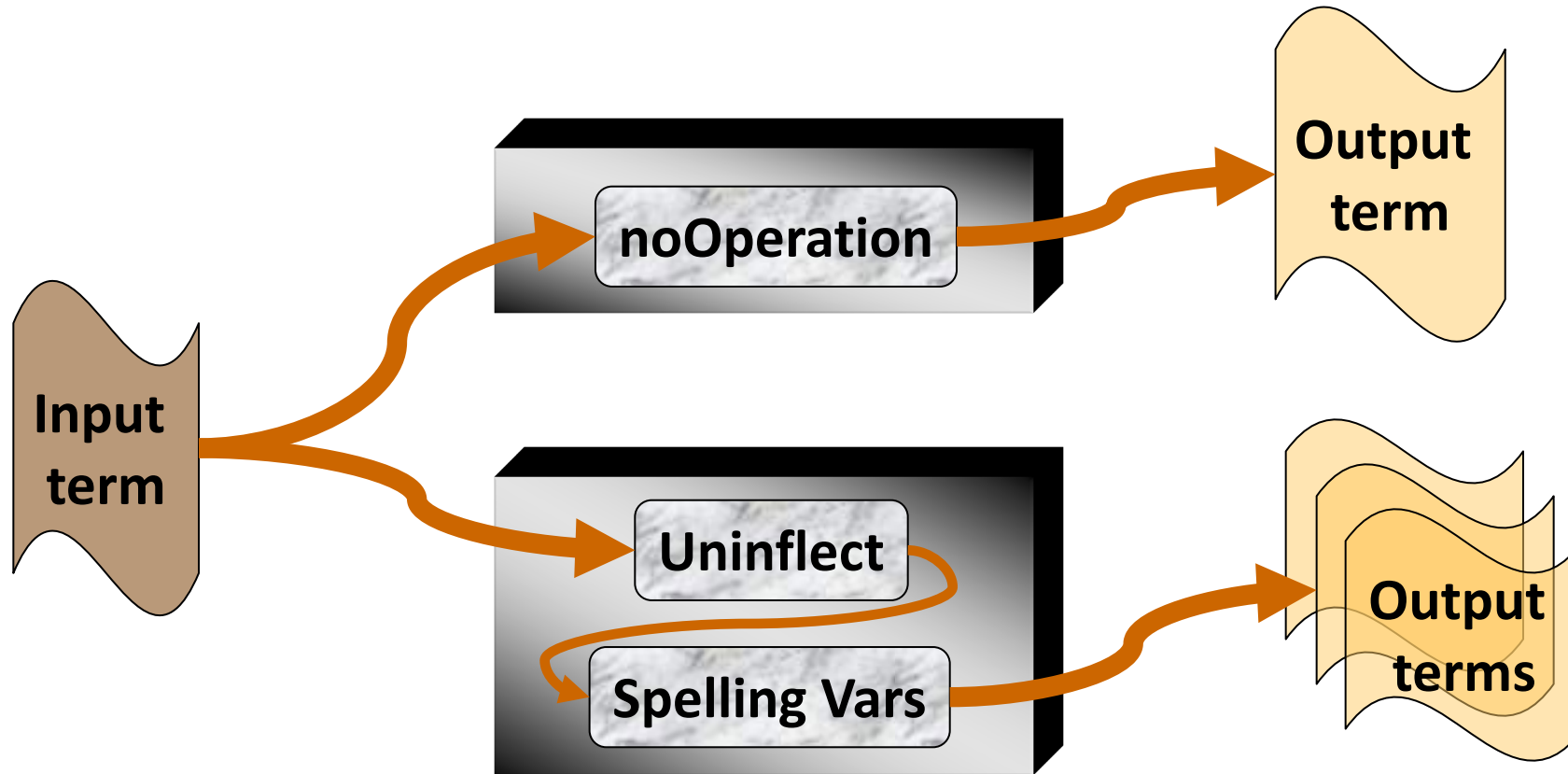
← Other information

l: low case
q: strip diacritic
g: remove genitive

t: remove stop words
p: remove punctuation
W: word order sort



LVG - Parallel Flows



- Multiple flows can be defined

Example - Parallel Flows

```
> lvg -f:n -f:B:s  
colors  
colors|colors|2047|16777215|n|1|  
colors|color|128|1|B+s|2|  
colors|color|1024|1|B+s|2|  
colors|colour|128|1|B+s|2|  
colors|colour|1024|1|B+s|2|
```

- n: no operation
- b: base form
- s: spelling variants

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



Example - Norm

“Foetoproteins α 's, NOS”

q0: map symbols to ASCII

g: remove genitives

rs: remove parenthetical plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninflect each words in a term

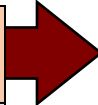
Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map Unicode to ASCII

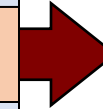
w: sort words by order

Norm (Cont.)

q0: map symbols to ASCII		“Foetoproteins α’s, NOS”
g: remove genitives		"Foetoproteins α’s, NOS"
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 (Cont.)

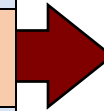
q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetical plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order



"Foetoproteins α's, NOS"
"Foetoproteins α's, NOS"
"Foetoproteins α, NOS"

Norm (Cont.)

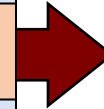
q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetical plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order



"Fœtoproteins α's, NOS"
"Fœtoproteins α's, NOS"
"Fœtoproteins α, NOS"
"Fœtoproteins α, NOS"

Norm (Cont.)

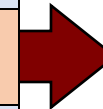
q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetical plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order



"Foetoproteins α's, NOS"
"Foetoproteins α's, NOS"
"Foetoproteins α, NOS"
"Foetoproteins α, NOS"
Foetoproteins α NOS

Norm (Cont.)

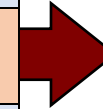
q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetical plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order



"Fœtoproteins α's, NOS"
"Fœtoproteins α's, NOS"
"Fœtoproteins α, NOS"
"Fœtoproteins α, NOS"
Fœtoproteins α NOS
Fœtoproteins α

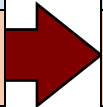
Norm (Cont.)

q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetical 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



"Foetoproteins α's, NOS"
"Foetoproteins α's, NOS"
"Foetoproteins α, NOS"
"Foetoproteins α, NOS"
Foetoproteins α NOS
Foetoproteins α
foetoproteins α

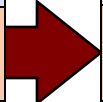
Norm (Cont.)

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

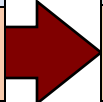
Norm (Cont.)

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

Norm (Cont.)

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

Norm (Cont.)

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

Norm (Cont.)

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

Norm (Cont.)

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

Norm – Why?

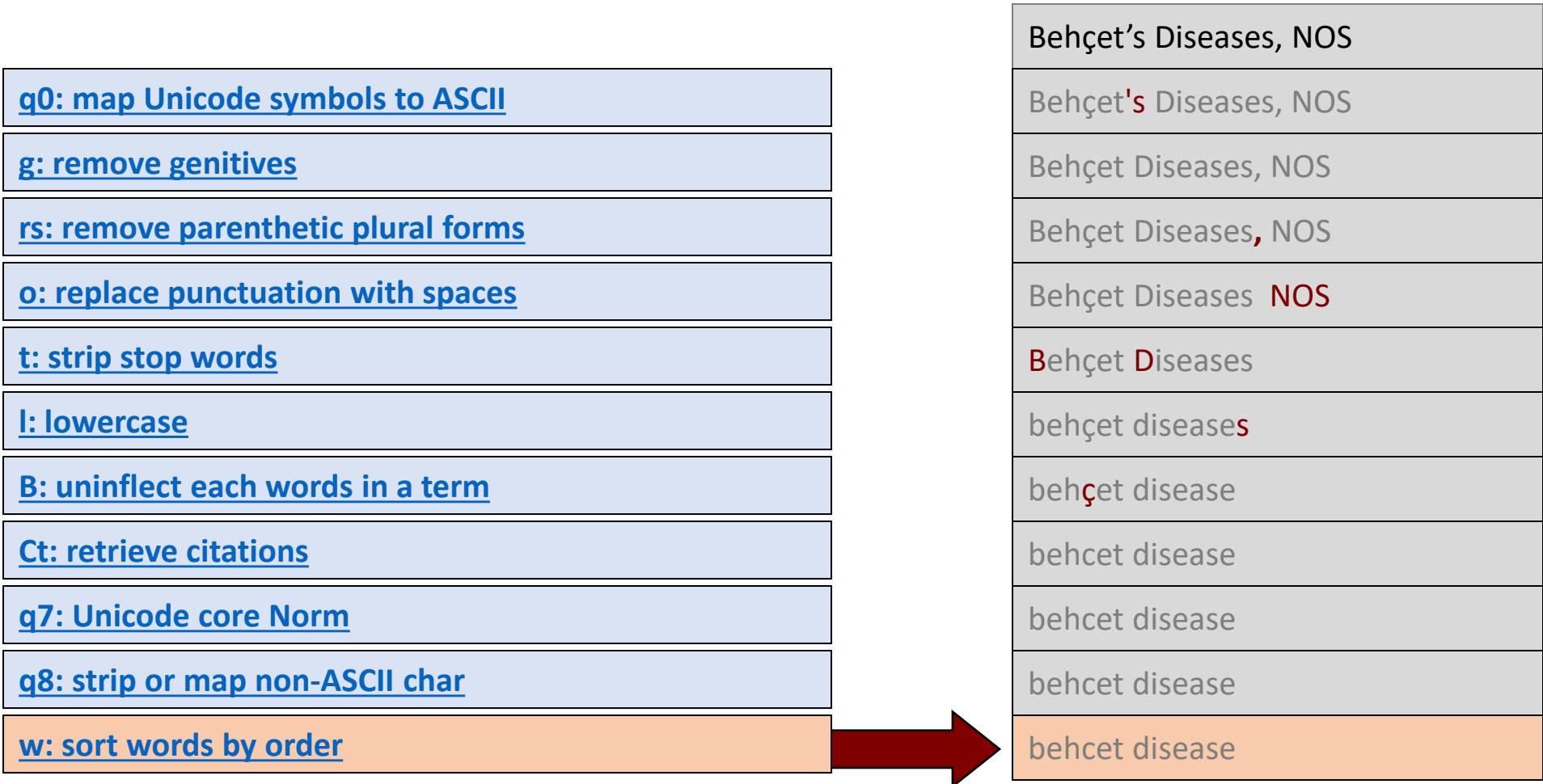
alpha Fetoprotein
alpha Fetoproteins
alpha-Fetoprotein
alpha-Fetoproteins
Alpha fetoproteins
alpha fetoprotein
alpha Foetoprotein
alpha foetoprotein
alpha fetoproteins
Alpha-fetoprotein
alpha-fetoprotein
Alpha Fetoproteins
Alpha-Fetoprotein
Alpha-fetoprotein NOS
Alpha Fetoprotein
alpha-fetoprotein
ALPHA-FETOPROTEIN
Alpha Foetoprotein
...

- In the MEDLINE articles, over 50 different forms for “alpha fetoprotein” can be normalized to the same form.

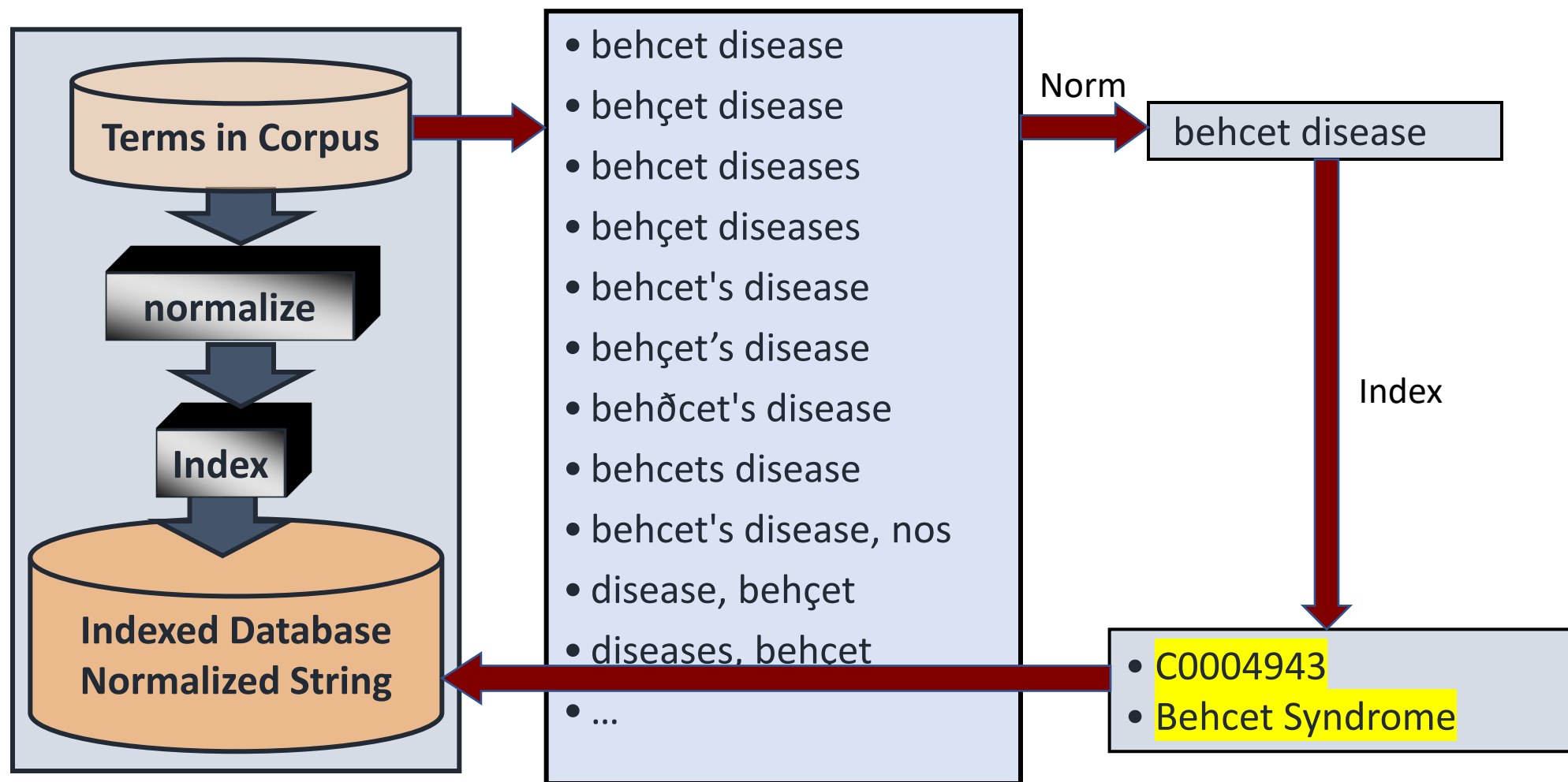


alpha fetoprotein

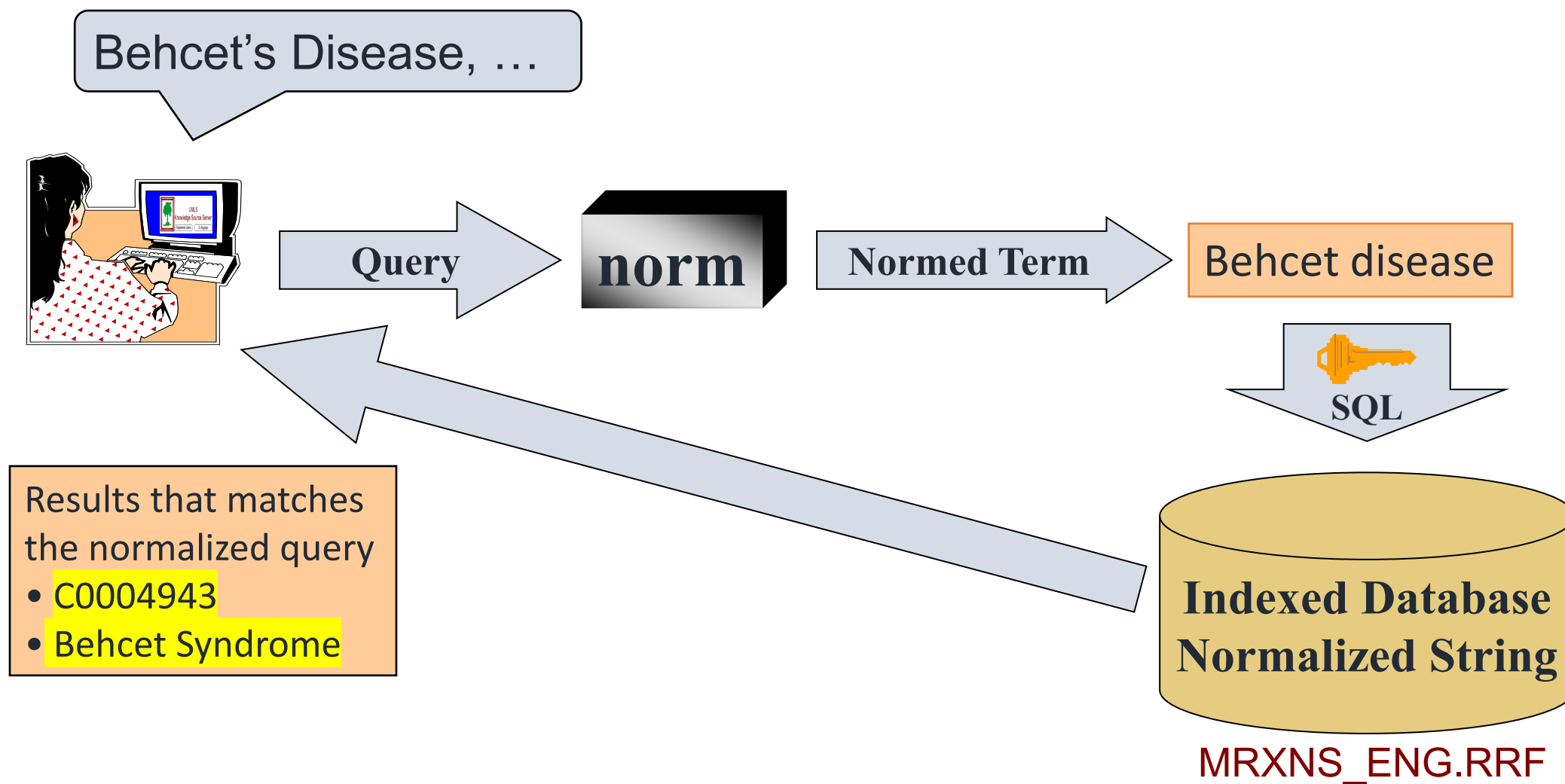
Concept Mapping (CM) - Norm



Indexing by Norm (Pre-Process Lexical Variations)



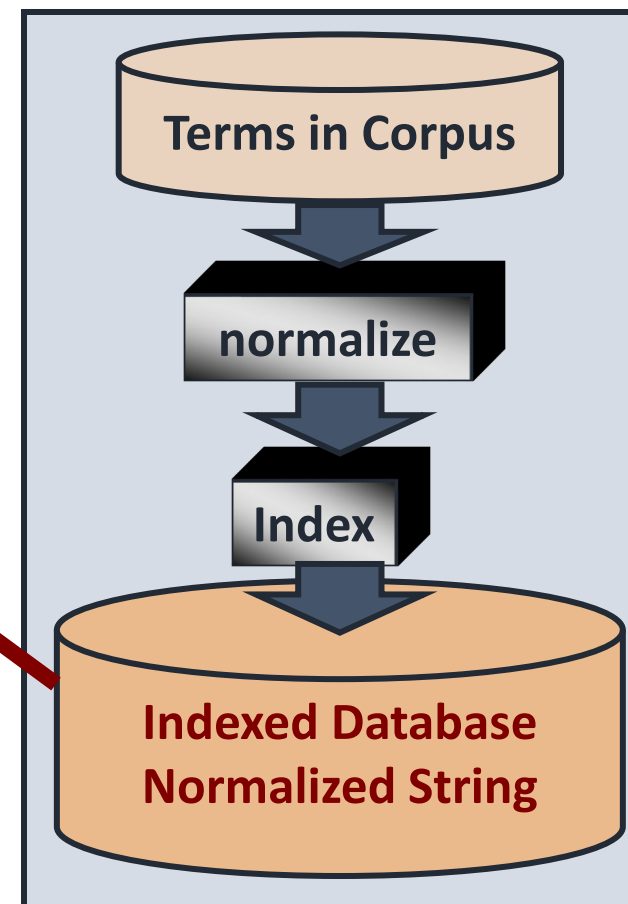
Query by Norm (Applications)



UMLS Metathesaurus

➤ UMLS Normalized Files

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

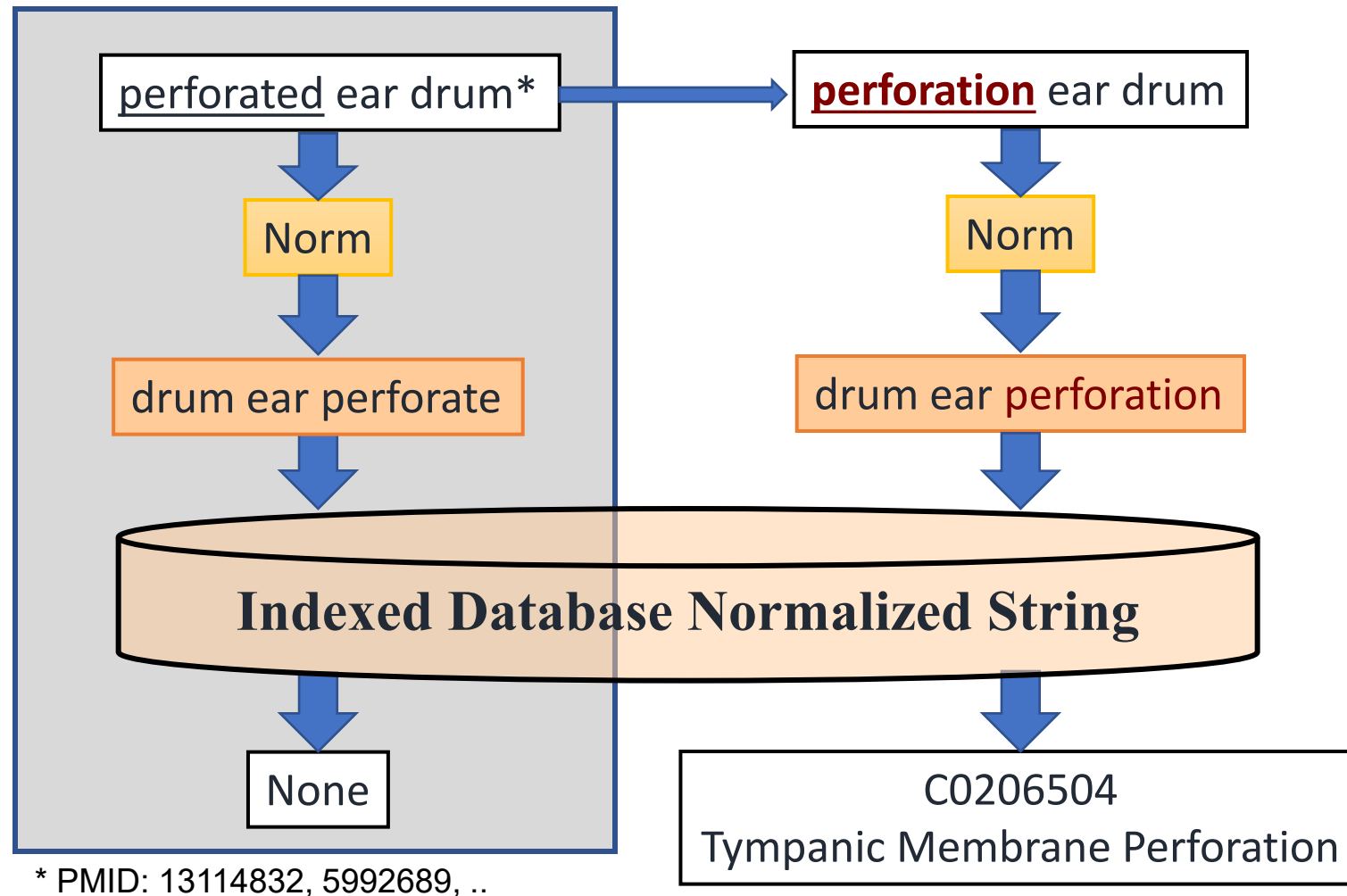


NLP Techniques in Concept Mapping

- **Normalization (same record – lexical variations):**
 - 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 with same concept – lexical thesaurus):**
 - Expand a term to its equal terms, such as subterm substitution of synonyms, derivational variants, abbreviations, etc.
 - To increase recall
- Multiword approach
- Others



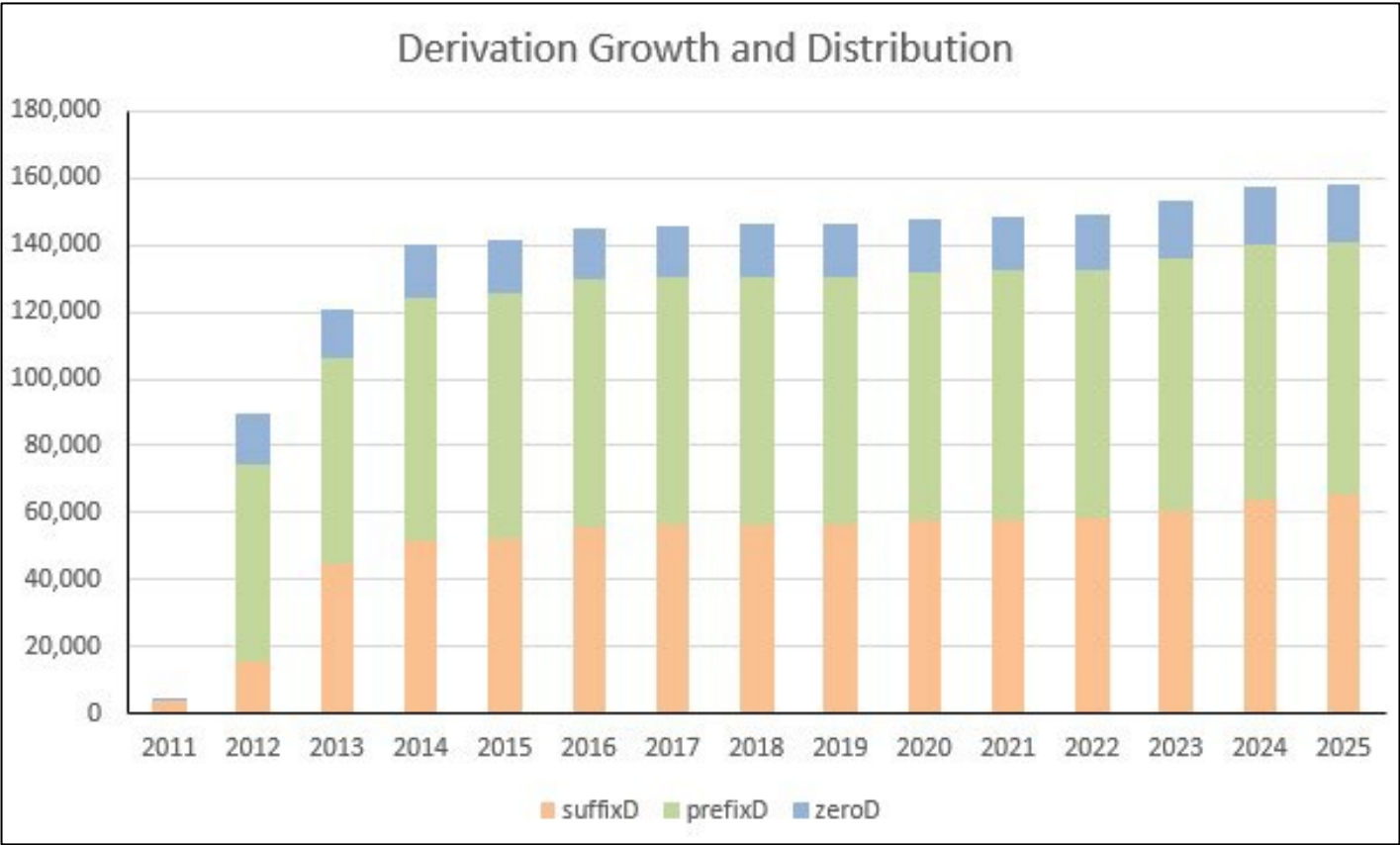
CM – Query Expansion (Derivation)



Lexicon Derivations

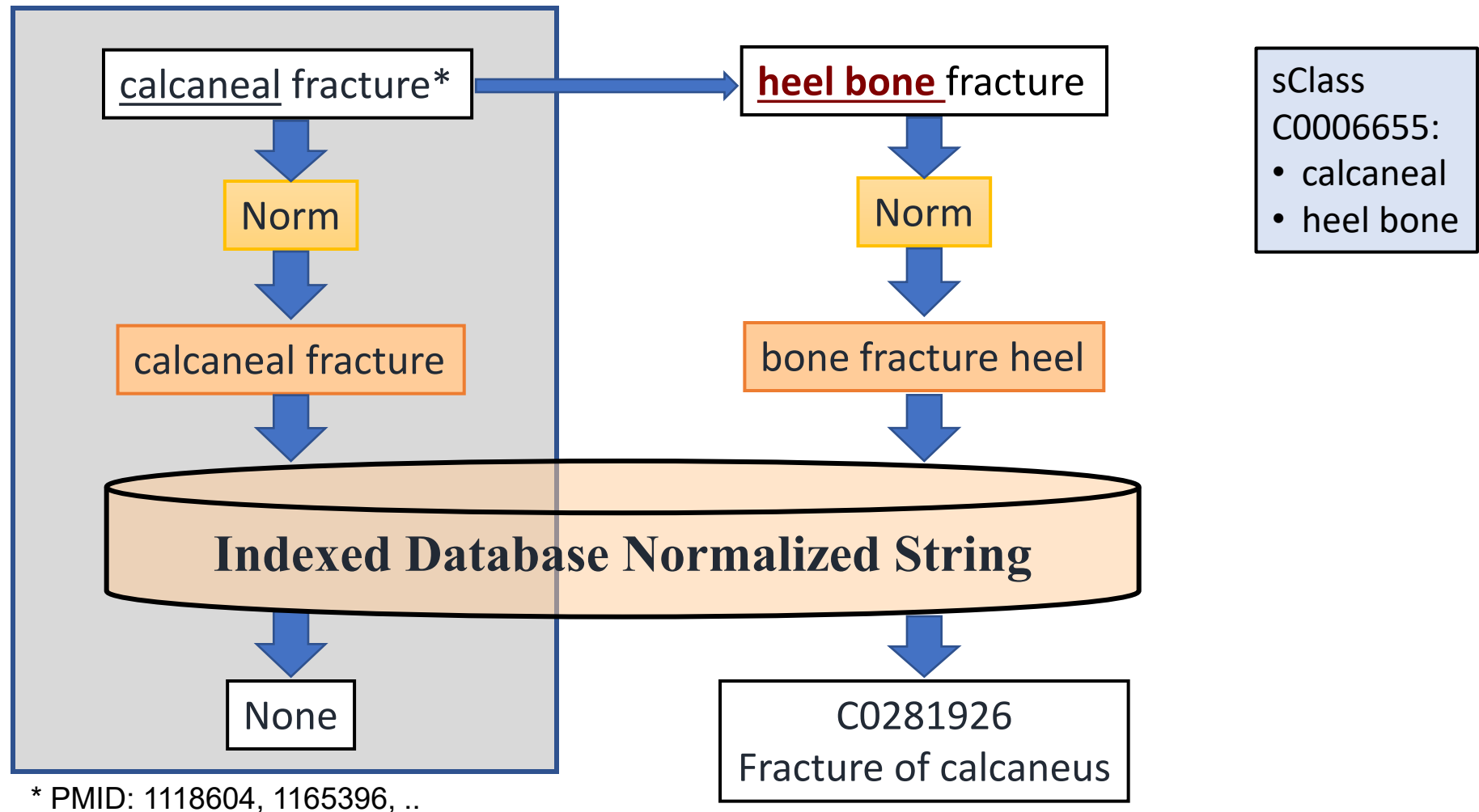
- Zero derivations (conversion)
 - transport (verb), transport (noun).
- Prefix derivations
 - autotransport, intratransport, pretransport, etc.
 - 153 prefixD rules: auto-, intra-, pre-
 - 126,168 prefixD candidates with 60.03 precision
 - negation: anti-war, contradict, disagree, dysfunction, immature, etc.
- Suffix derivations
 - transportation, transportable, transporter, .
 - 186 suffixD rules: \$|verb|ation\$|noun, \$|verb|able\$|adj, \$|noun|er\$|noun
 - 93,298 suffixD candidates with 81.55 precision
 - negation: \$|noun|less\$|adj, \$|verb|less\$|adj

Lexicon Derivation Growth

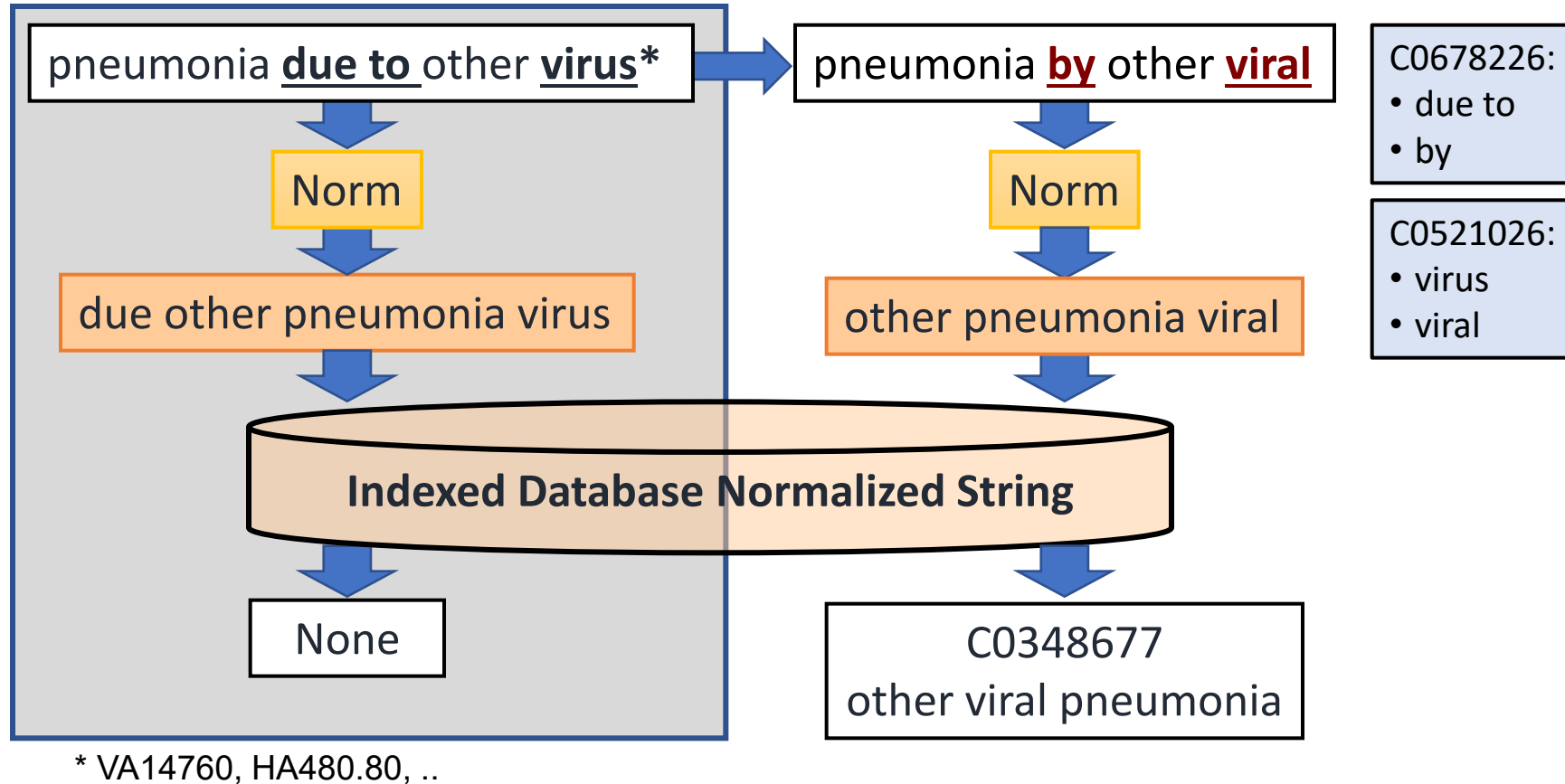


Year	Total	suffixD	prefixD	zeroD	Negation:[N O]
2025	158,409	41.25%	47.81%	10.94%	14.77% 85.23%

CM – Query Expansion (Synonym)



CM - Multiple Substitutions



Lexicon Synonyms

- Terms with the same concept but lexically dissimilar
 - Arranged in pairs, called sPairs
 - Must have EUI and CUI
 - Must be cognitive synonyms:
 - Commutativity (if $A = B$, then $B = A$)
 - Transitivity (if $A = B$ and $B = C$, then $A = C$), Ex: “happy – joy – enjoy”
- Examples:
 - “behcet syndrome” and “behcet disease”
 - “heel bone” and “calcaneal”



Lexicon Synonym Acquisition

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



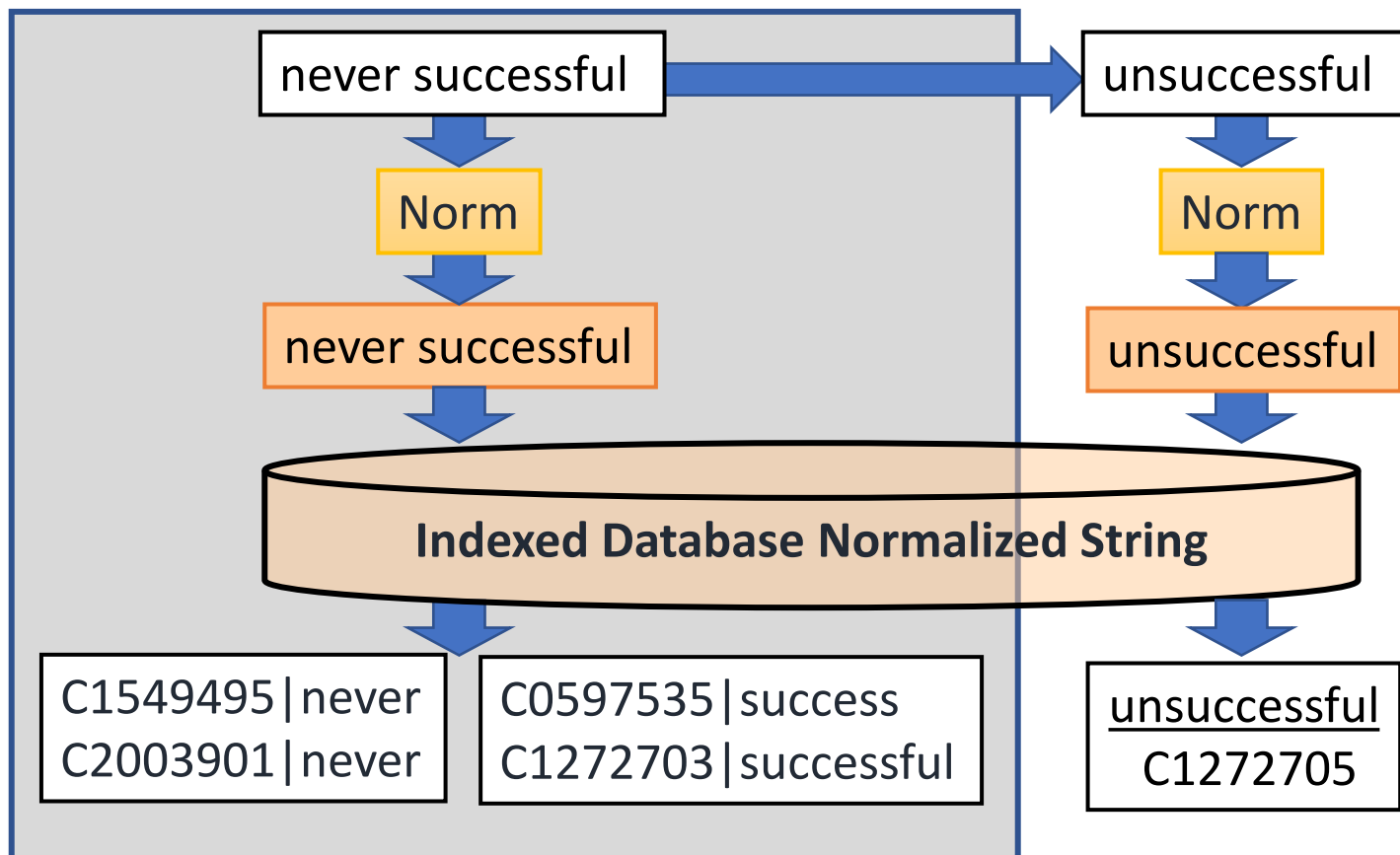
Lexicon Synonym Growth



Year	Total	CUI	EUI	NLP_LVG
2025	282,602	211,738 (74%)	66,092 (23%)	4,772 (1%)

CM – Query Expansion (Antonyms)

- PMID-2644556: Therapy was never successful for large gall stones (> 15 mm).



- Better natural language understanding: Negated antonyms can be used to substitute synonymous antonyms for better recall ("never successful" = "unsuccessful").

Lexicon Antonyms

- Lexicon Antonyms: in the Lexicon, same POS, not synonyms, canonical antonyms.
- Canonical Antonyms (13K)
 - good antonyms that have opposite or contrasting meanings in a canonical domain, a generic domain (central to human life and way of living across times and cultures)
 - Ex: black|white|adj|color vs. dark|white|adj|chocolate
 - Ex: cold|hotness|noun|temperature vs. ice|hot|adj|tea
- Source Models:
 - LEX : Negative lexical entries in the Lexicon (0.65%)
 - SD: Suffix derivations with negations in the Lexicon (3.25%)
 - PD: Prefix derivations with negations in the Lexicon (43.32%)
 - CC: Co-occurrences in a corpus (6.35%)
 - SN: Semantic network (46.42%)

Features in the Lexicon Antonyms

- **Canonical Antonyms** (13K)
 - good antonyms that have opposite or contrasting meanings in a canonical domain.
- **Canonical Domains** (11)
 - are generic and central to human life and ways of living across time and cultures.
 - existence, frequency, location, physical_property, possibility, quality, quantity, role, size, temperature, temporal.
- **Type** (4)
 - bounded type: two endpoints without middle ground (if $X = \text{not } Y$, $Y = \text{not } X$), [dead|alive]
 - unbounded type: extreme values never reach an endpoint (if $X \neq \text{not } Y$, $Y \neq \text{not } X$), [long|short]
 - asymmetric bounded (if $X = \text{not } Y$, $Y \neq \text{not } X$, where X is the negative/endpoint), [colorless|colorful]
 - NA
- **Negation** (5)
 - strict negative [always|never|adv], broadly negative [usually|rarely|adv], otherwise.

Lexicon Antonym Examples

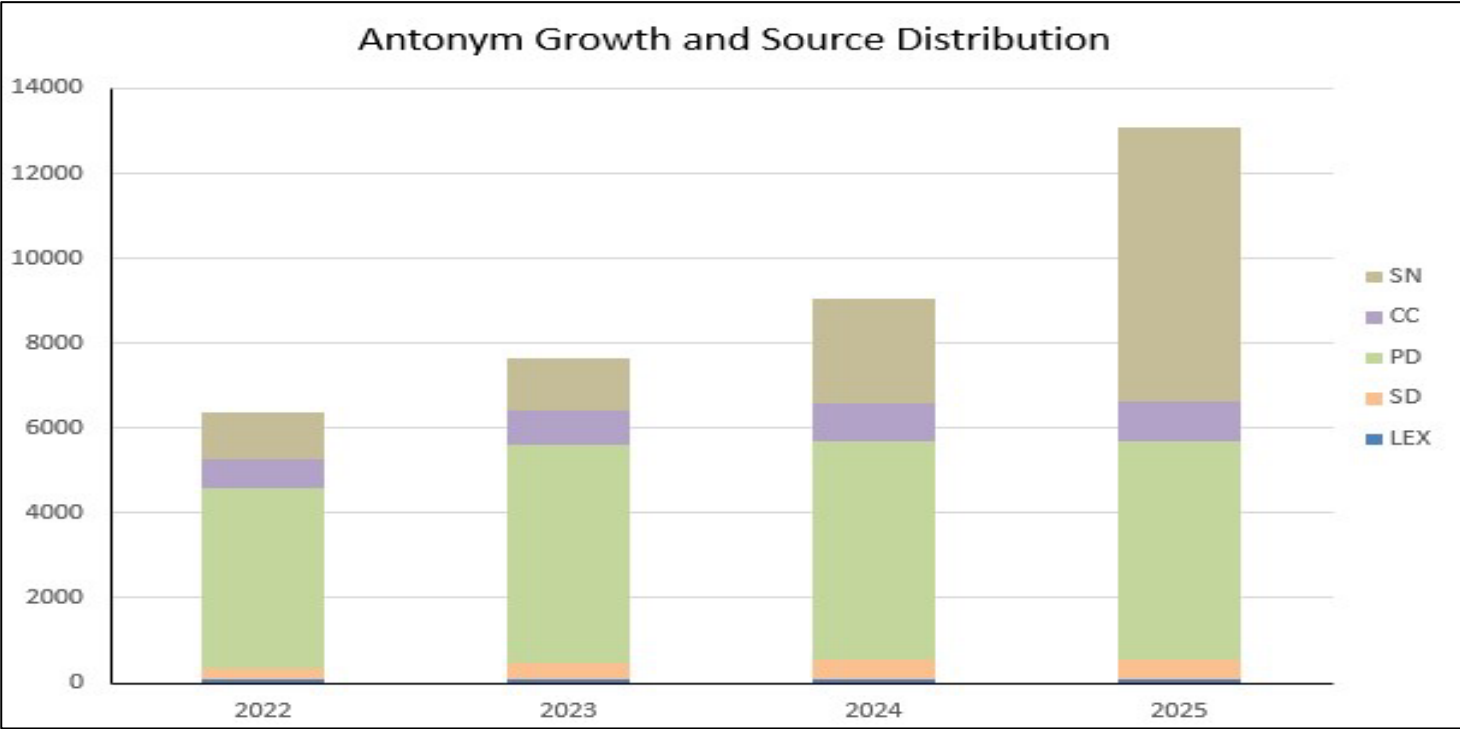
Antonym-1	Antonym-2	POS	Type*	Negation**	Model	Domain	Dist.
with	without	prep	B	N2***	LEX	existence	2.60%
never	always	adv	UB	N1***	LEX	frequency	0.84%
head	tail	noun	UB	O	SN	location	7.71%
asleep	awake	adj	B	O	CC	physical property	3.36%
believe	disbelieve	verb	AB2***	BN2***	PD	possibility	1.97%
treatable	untreatable	adj	B	O	PD	quality	74.95%
decrease	increase	verb	UB	O	CC	quantity	1.94%
amateur	pro	noun	UB	O	SN	role	0.60%
big	little	adv	UB	BN2***	SN	size	1.77%
feverous	feverless	adj	B	O	SD	temperature	0.40%
after	before	conj	UB	O	CC	temporal	3.85%

* Type: B=bounded, UB=unbounded, AB=asymmetric bounded.

** Negation: N=strict negative, BN=broad negative, O=otherwise; not negative.

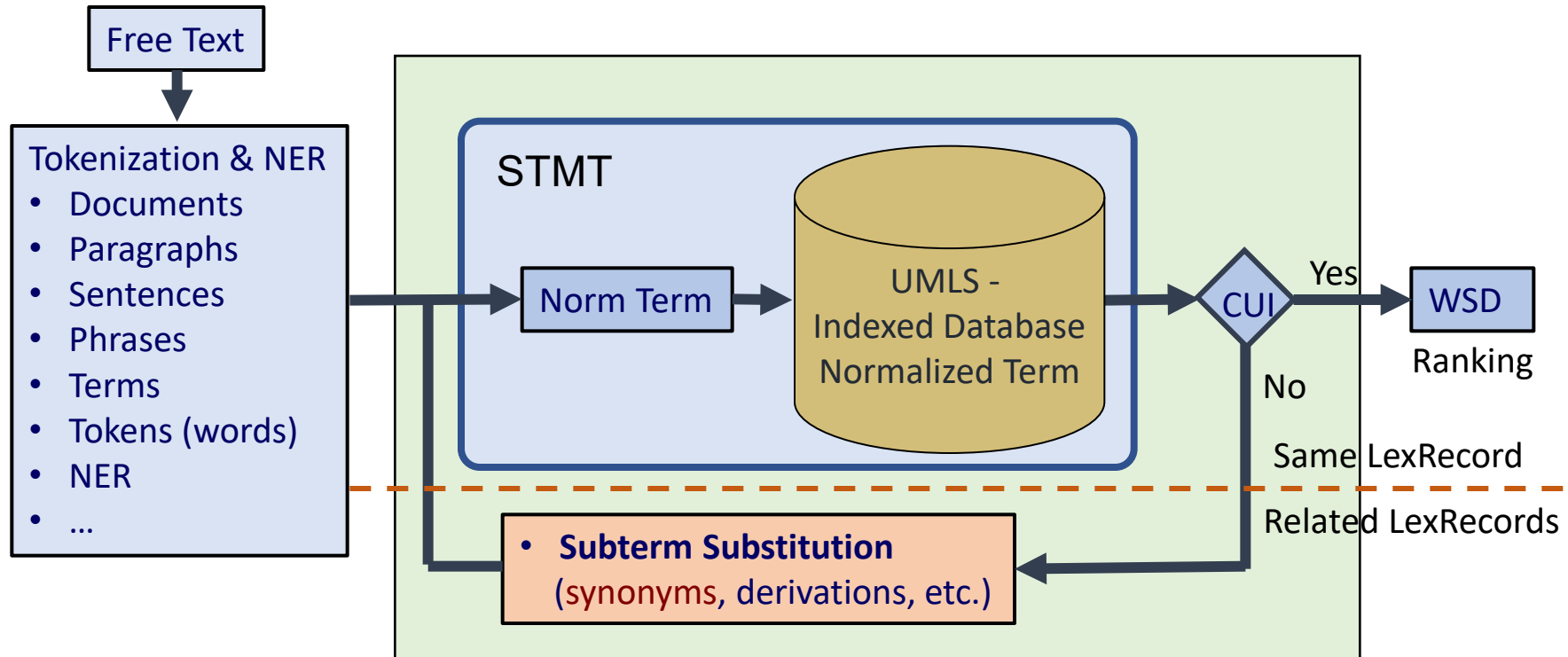
*** 1 = Antonym-1, 2 = Antonym-2.

Lexicon Antonym Growth



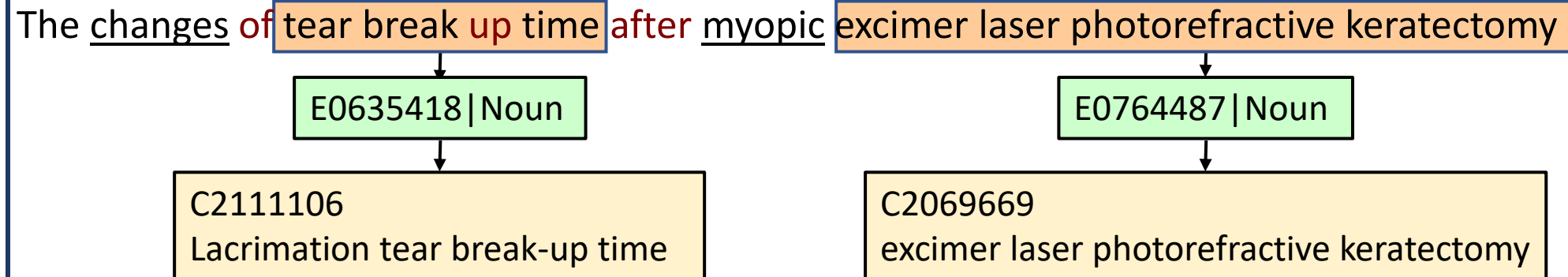
Year	Total	LEXICON	SuffixD	PrefixD	Co-occurrences	Semantic Network
2025	13,076	94 (0.72%)	450 (3.44%)	5,170 (39.54%)	932 (7.13%)	6,430 (49.17%)

Real-time Concept Mapping Model



CM – Multiword Approach

- Multiword Approach (vs. window shifting concept lookup algorithm)
 - Name Entity Recognition (NER)
 - POS tagger, parser
 - Concept mapping (the longest multiword terms)
- Example (PMID 9510650, TI):



STMT – Multiword Identification

- Run the LSF (LexItem Subterm Finder) in the STMT (Sub-Term Mapping Tools)
- Two longest Lexicon multiwords in the sentence are identified, which are used as name entity for the key concepts in that sentence.

```
shell> ls -p
- Please input a term (type "Ctl-d" to quit) >
The changes of tear break up time after myopic excimer laser photorefractive keratectomy
--- LexItem Multiword Subterms ---
break up time|E0635415
photorefractive keratectomy|E0225495
break up|E0220309
excimer laser|E0514806
excimer laser photorefractive keratectomy|E0764487
tear break up time|E0635418
changes|E0016183|E0016184
```



STMT – Concept Mapping

- Run SMT to find the concept mapping and preferred term
- The concepts and preferred terms of “tear break uptime” and “excimer laser photorefractive keratectomy” are found.

```
shell> smt -p -pt
- Please input a term (type "Ctl-d" to quit)
tear break up time
tear break up time|break tear time up|C2111106|lacrimation tear break-up time|0
- Please input a term (type "Ctl-d" to quit) >
excimer laser photorefractive keratectomy
excimer laser photorefractive keratectomy|ceratectomy excimer laser photorefractive|C2069669|excimer laser
photorefractive keratectomy|0
```

CSpell – Consumer Health

- Health information consumers
 - Patients, families, caregivers, and the general public
 - Seek health information & ask questions online every day
- Sources of consumer health questions
 - MedlinePlus, forms and emails, etc.
 - Search engine, social media, forum, etc.
- Consumer questions
 - Contain many spelling errors, informal expressions, etc.
 - Spelling errors hinder automatic question answering
 - Spelling corrections are needed (pre-processing)



Example - CSpell

My mom is 82 years old suffering from **anixity** and depression for the last 10 years was **dianosed** early **on set** **deminita** 3 years ago. Do **yall** have a office in Greensboro NC? Can you recommend someone. she has **seretona** syndrome and **nonething** helps her. ^[2]

- Corrections:



- Reference:

[2] Kilicoglu H, Fiszman M, Roberts K, et al. An Ensemble method for spelling correction in consumer health questions. AMIA Annu Symp Proc., 2015: 727–36.

Error	Correction
anixity	anxiety
dianosed	diagnosed
on set	onset
deminita	dementia
yall	y'all
seretona	serotonin
nonething	nothing

← merge

Examples - CSpell Multiple Corrections

Ex-1: Input Text	Output: different types of corrections
He was dianosed early on set deminita 3years ago.	He was <u>diagnosed</u> early <u>onset</u> <u>dementia</u> <u>3 years</u> ago.
	<div>NW Spelling</div> <div>RW Merge</div> <div>NW Spelling</div> <div>ND Split</div>

Ex-2: Input Text	Output: multiple corrections
I have a shuntfrom2007 .	I have a <u>shunt from 2007</u> .
	<div>NW Split</div> <div>ND Split</div>

Ex-3: Input Text	Output: multiple corrections
I am permanently depressed and was on 2 or 3 different anti depresants .	I am permanently depressed and was on 2 or 3 different <u>antidepressants</u> .
	<div>RW Merge</div> <div>NW Spelling</div>

Questions



- Lexical Systems Group: <https://lhncbc.nlm.nih.gov/LSG/index.html>