The SPECIALIST Lexicon and NLP Tools

By: Dr. Chris J. Lu

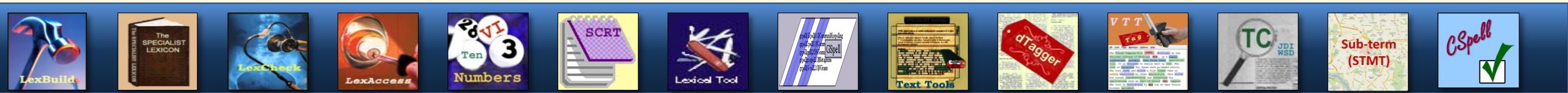
NLM – LHNCBC - CGSB

Oct., 2020

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

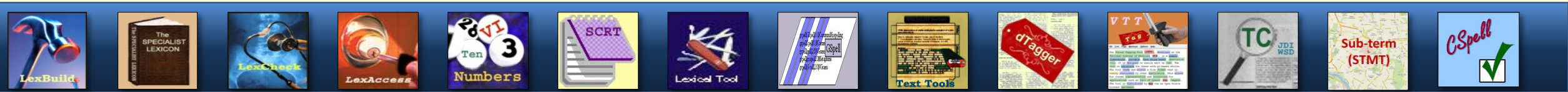
Disclaimer

- The views and options expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.



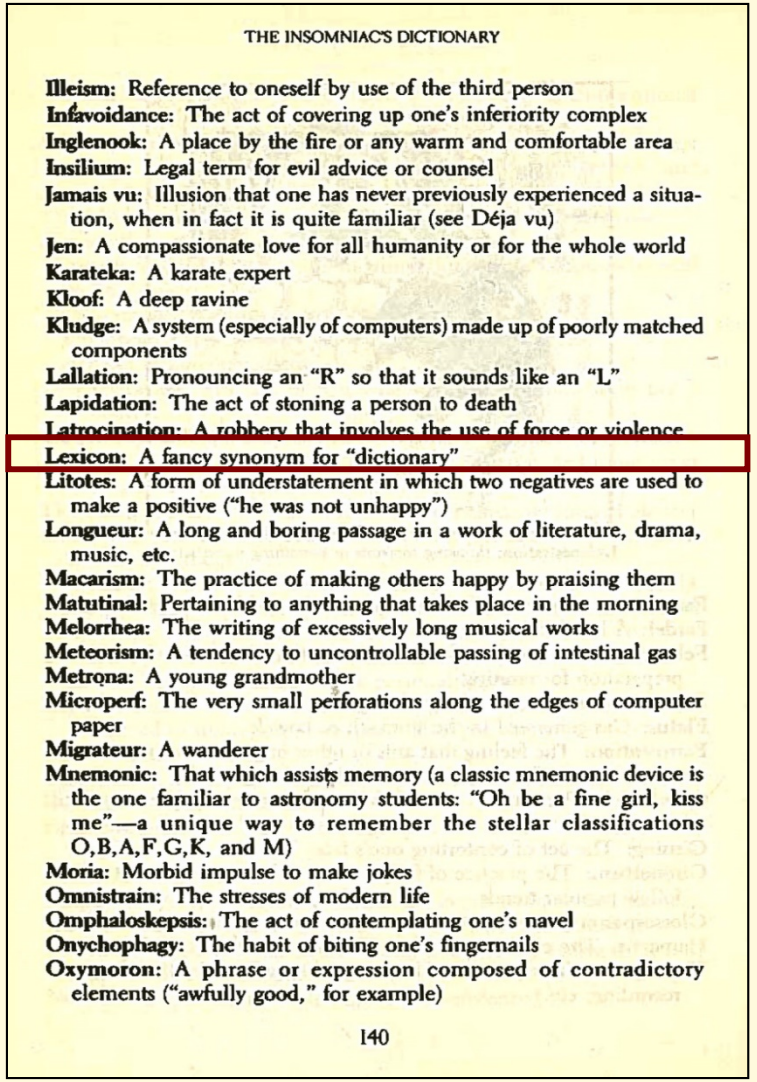
Outline

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



1. The SPECIALIST Lexicon

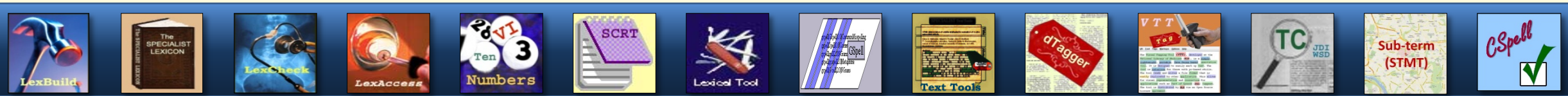
- A fancy synonym for "dictionary"
- A syntactic lexicon
- Biomedical and general English
- Over 0.5M records, 1M words (POS + forms)
- Designed/developed to provide the lexical information needed for NLP (Natural Language Processing) systems
- Distributed in the Unified Medical Language System (UMLS) 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

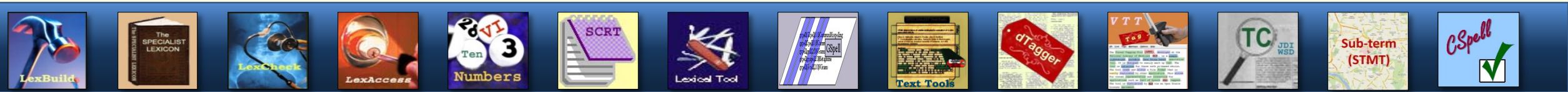


Team of Lexicon Builders

- Dr. Alexa McCray, founded in 1994 (previous LHNCBC Director, 2005-)
- Allen Browne, father of the SPECIALIST Lexicon (retired 2017)

- James Mork (Branch Chief)
- Dr. Chris J. Lu
- Dr. Amanda Payne

☺



LexBuild Process (Computer-Aided)

Sources:

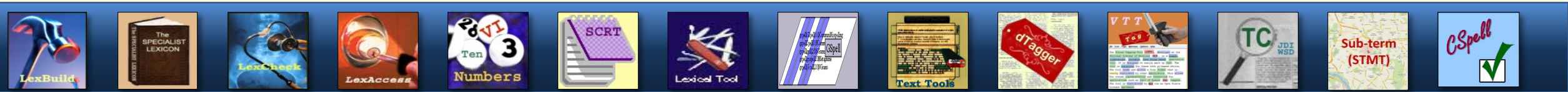
- Word candidates from MEDLINE
- Words from consumer data
- Others
  - Dorland's Illustrated Medical Dictionary
  - American Heritage Word Frequency book (top 10K)
  - Longman's Dictionary of Contemporary English (Top 2K lexical items)
  - The Metathesaurus browser and retrieval system
  - The UMLS test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

Reviewed by lexicographers:

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

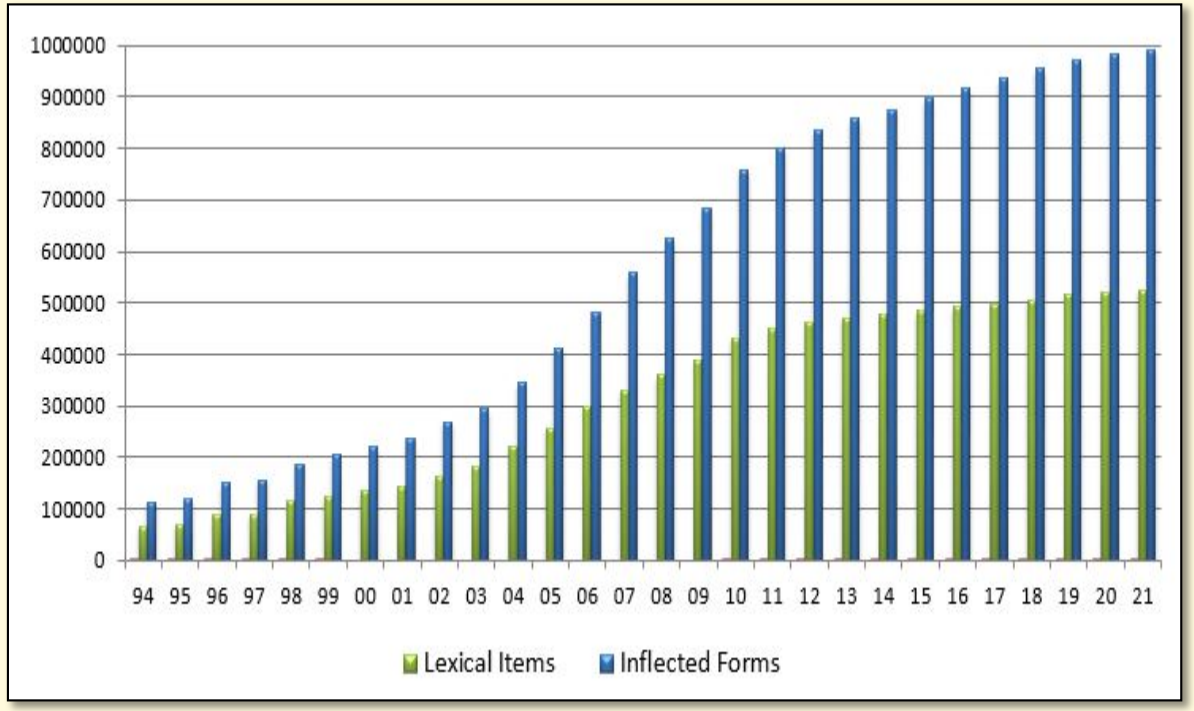
Build:

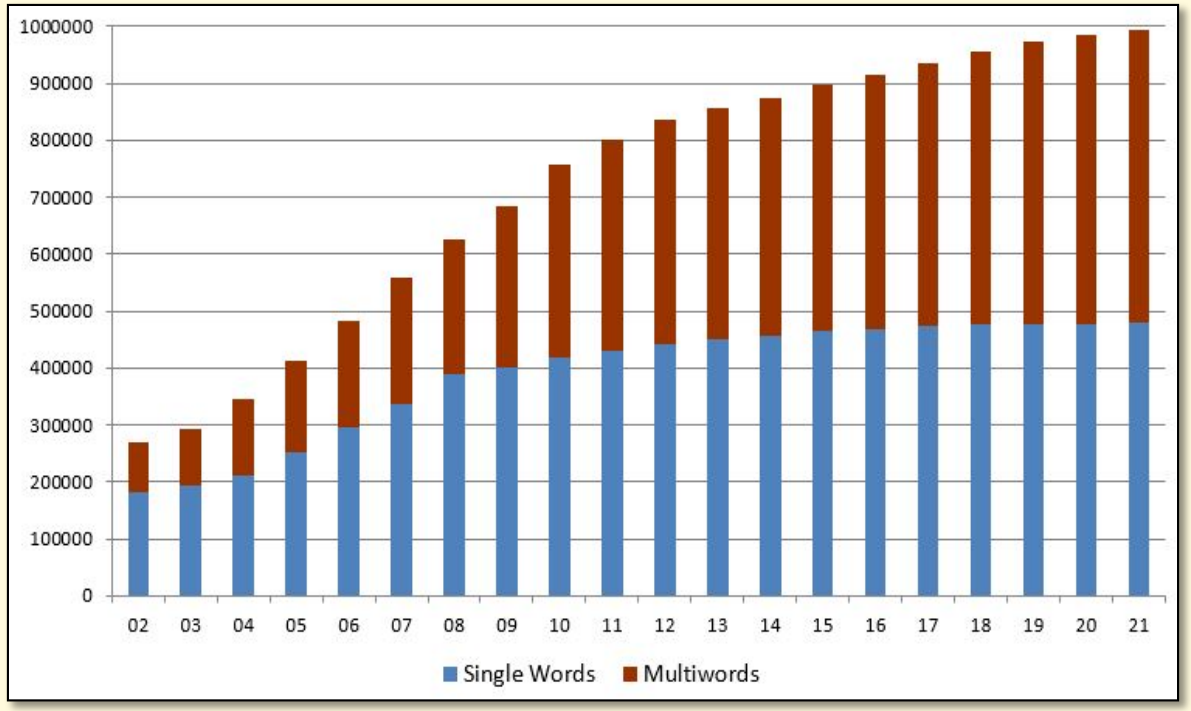
- **LexBuild**
- **LexAccess**
- **LexCheck**

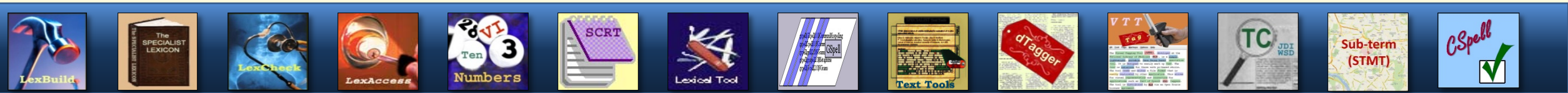


Lexicon Growth – 1994 to 2021

- 524,327 lexical records
- 1,174,195 words (categories and inflections)
- 992,545 forms (spelling only)
  - Single words: 478,585 (48.22%); Multiwords: 513,960 (**51.78%**)

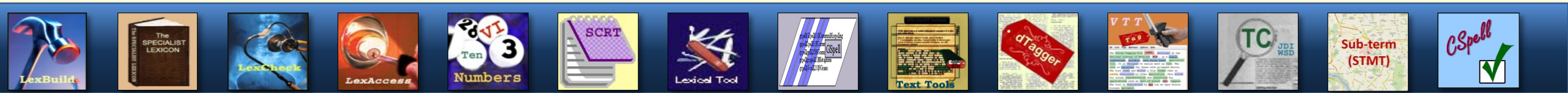






What Is a Word?

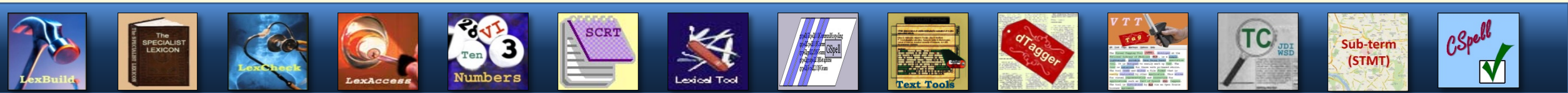
- Orthographic words:
  - Spelling
    - spVar: color vs. colour
    - Inflection (noun): dog vs. dogs
    - Inflection (verb): see vs. saw
    - Inflection (noun): saw vs. saws
    - POS: square, see vs. saws
    - meaning: cold
  - use spaces as word boundary
    - ice-cream vs. ice cream - space



(Multi)Words for Lexical Records

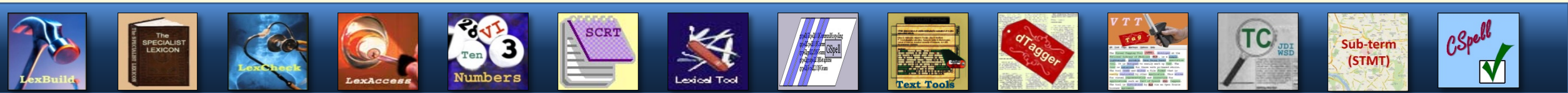
- Lexicon terms: single words and multiwords
  - Space(s): ice-cream vs. ice cream, tradeoff vs. trade-off vs. trade off
- Four criteria for terms in the Lexicon:
  - Part of Speech (POS):
    - tear break up time, cardiac surgery, frog erythrocytic virus,
  - Inflection morphology (uninflection):
 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

- Generating A Distilled N-Gram Set: Effective Lexical Multiword Building in the SPECIALIST Lexicon, HealthInf 2017
- Multiword Frequency Analysis Based on the MEDLINE N-Gram Set, AMIA 2016
- Generating the MEDLINE N-Gram Set, AMIA 2015
- Using Element Words to Generate (Multi)Words for the SPECIALIST Lexicon, AMIA 2014

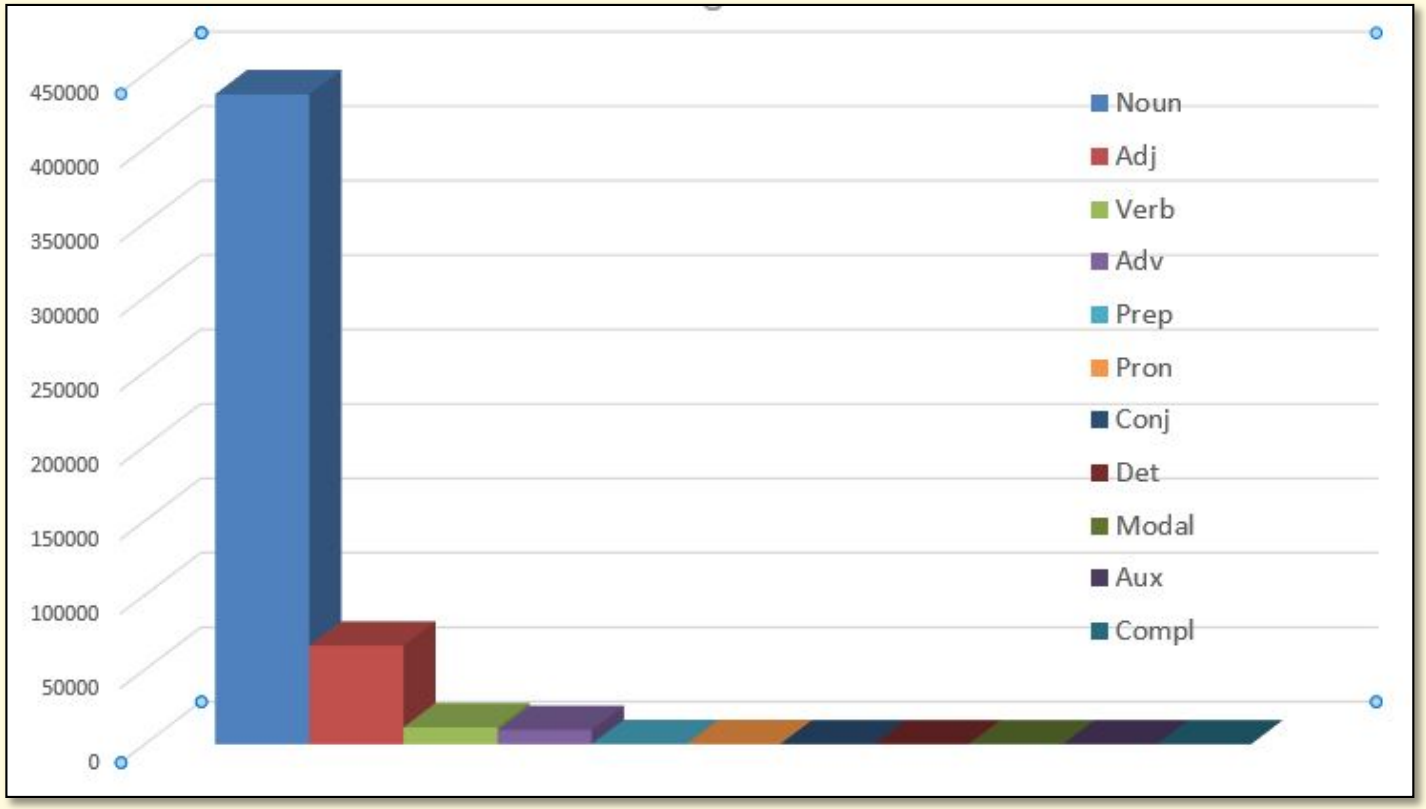


Lexical Records - Information

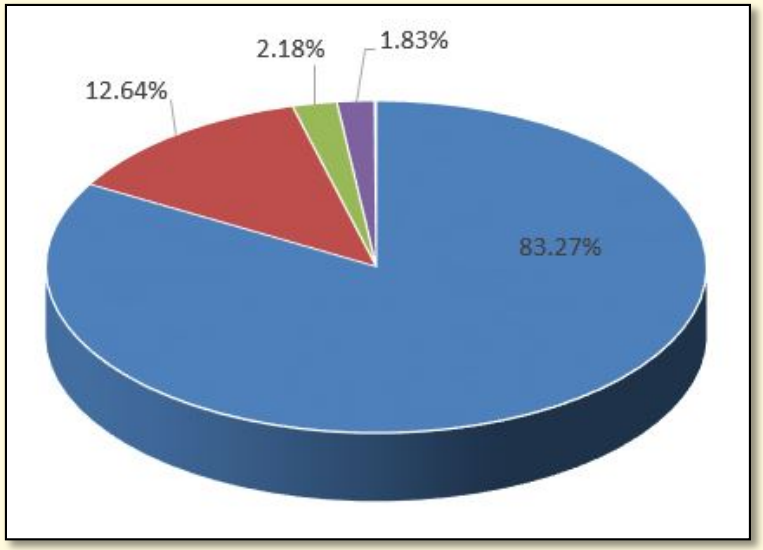
- POS (Part-of-Speech)
- Morphology
  - Inflection
  - Derivation
- Orthography
  - Spelling variants
- Syntax
  - Complementation for verbs, nouns, and adjectives
- Other
  - Expansions of abbreviations and acronyms
  - Nominalizations
  - …



Categories – Parts of Speech (11)

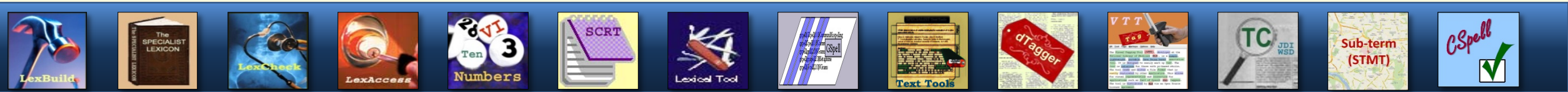


Lexicon.2021



The rest of POS < 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
- Compl: that



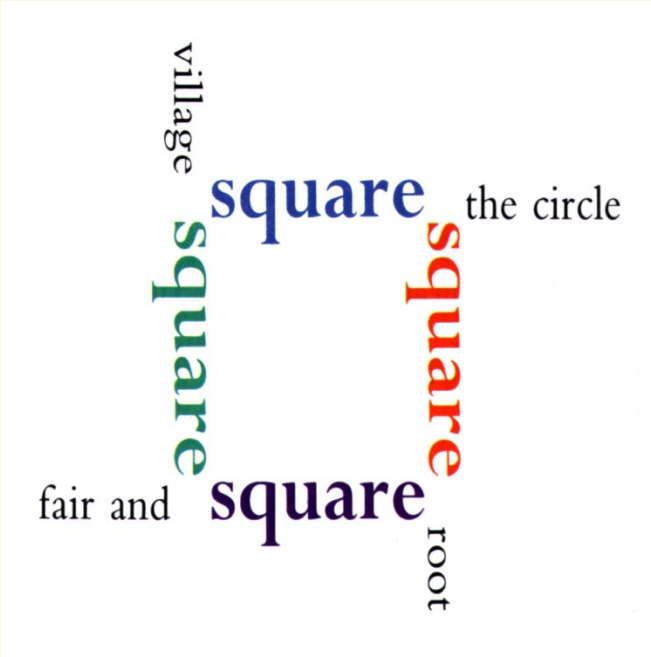
Lexical Records & POS

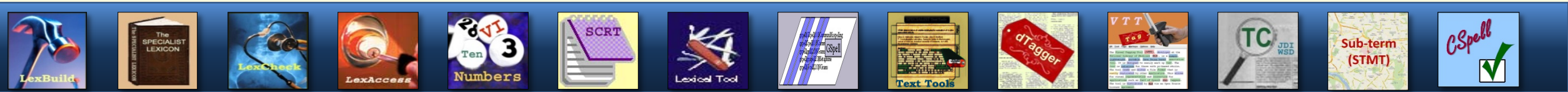
```
{base=square
entry=E0057517
        cat=verb
        variants=reg
        intran
        intran:part(up)
}
```

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

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

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

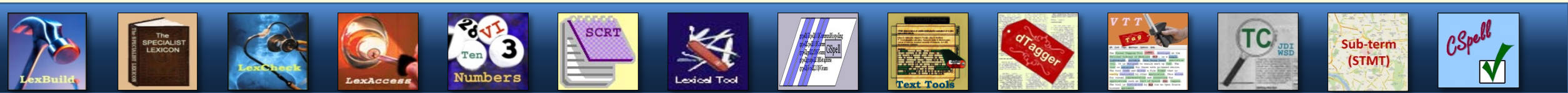




Morphology

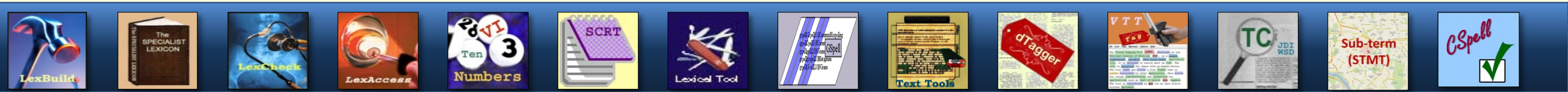
- Inflectional
  - noun: book, books
  - verb: categorize, categorizes, categorized, categorizing
  - adj: red, redder reddest
- Derivational
  - example: transport
    - suffix - transportation, transportable, transporter, …
    - prefix – autotransport, intratransport, pretransport, …
    - conversion (zero) - transport (verb), transport (noun)

- Generating SD-Rules in the SPECISLIST Lexical Tools - Optimization for Suffix Derivation Rule Set, HealthInf 2016
- A Systematic Approach for Automatically Generating Derivational Variants in Lexical Tools Based on the SPECIALIST Lexicon, IEEE IT Professional Magazine 2012
- Implementing Comprehensive Derivational Features in Lexical Tools Using a Systematical Approach, AMIA 2013



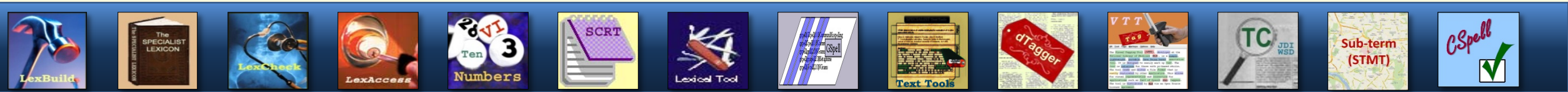
Orthography (Spelling Variation)

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



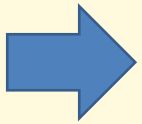
Syntax - Verb Complements

- intran
  - I'll treat.
- tran=np
  - He treated the patient.
- ditran=np,pphr(with,np)
  - She treated the patient with the drug.
- …

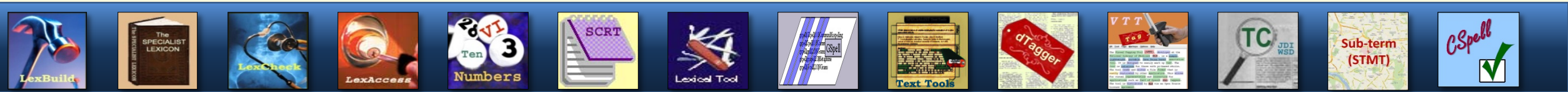


Lexical Information to Coded Lexical Records

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



```
{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
}
```

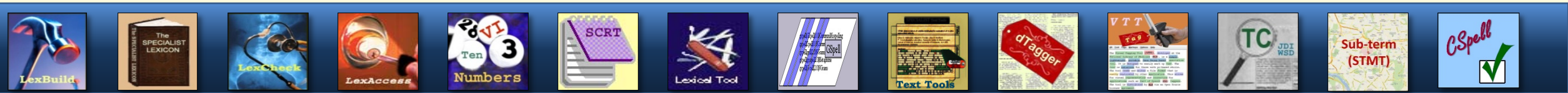
```
{base=divorcé
entry=E0543077
        cat=noun
        variants=reg
}
```

```
{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
}
```

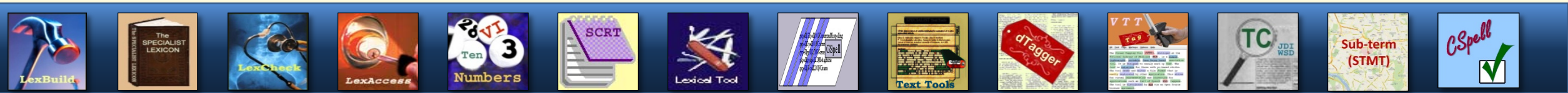
- Converting Unicode Lexicon and Lexical Tools for ASCII NLP Applications, AMIA 2011
- Using lexical tools to convert Unicode characters to ASCII, AMIA 2008



Classification Type (2020+)

Code	Definitions	Examples
class_type=archaic|arg1	indicates that the specified base form is no longer in common use	class_type=archaic|colde
class_type=taxonomic|arg1	indicates that a variant is a term from biological taxonomy (genus, species, etc)	class_type=taxonomic|Bacterium
class_type=source|arg1|arg2	indicates the language or dialect where the specified base form originated	class_type=source|colour|british class_type=source|bonafide|latin
class_type=informal|arg1|arg2|arg3	indicates that the specified base form is used primarily in colloquial contexts	class_type=informal|nite|night|E0042638 class_type=informal|nite|evening|E0026437
class_type=other	indicates some other type of classification information (gene, protein, etc.)	class_type=other

- Classification Types: A new Feature in the SPECIALIST Lexicon, AMIA 2019

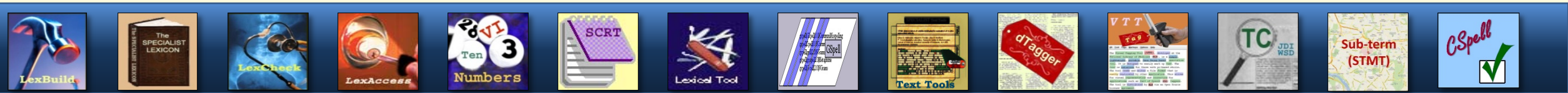


Lexicon Unigram Coverage – Without WC

- Total unique word for MEDLINE (2016): 3,619,854
- Lexicon covers 10.62-11.83 % 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%**
NW-EW*	43,811	1.2103%	**11.8253%**
NEW	3,191,797	88.1747%	100.0000%
Total	3,619,854		

* NW-EW: an element word only exist in multiword, such as “non”, “vitro”, “vivo”, “intra”, etc.

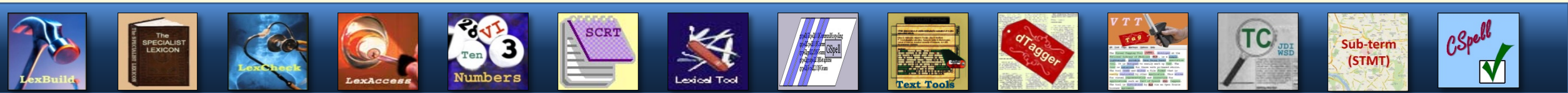


Lexicon Unigram Coverage – With Frequency (WC)

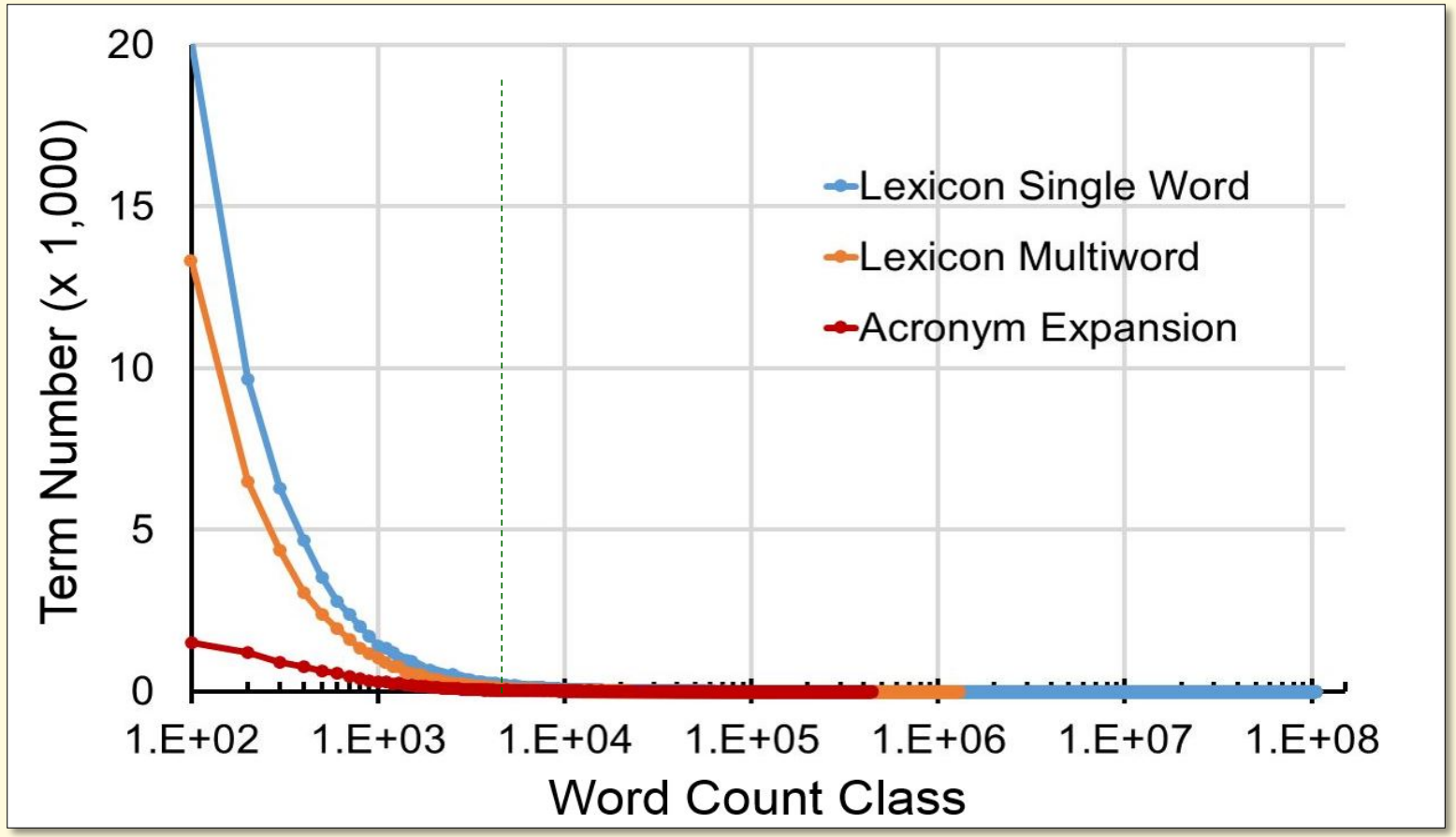
- Total word count for MEDLINE (2016): 3,114,617,940
- Lexicon covers > 98-99% unigrams from MEDLINE

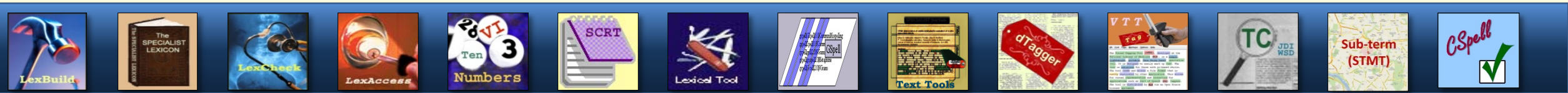
Types	Word Count	Percentage %	Accu. %
LEXICON	2,911,156,308	93.4675%	93.4675%
NUMBER	8,753,120	0.2810%	93.7485%
DIGIT	145,548,882	4.6731%	**98.4216%**
NW-EW*	19,148,557	0.6148%	**99.0364%**
NEW	30,011,073	0.9636%	100.0000%
Total	3,114,617,940		

* NW-EW: an element word only exist in multiword, such as “non”, “vitro”, “vivo”, “intra”, etc.



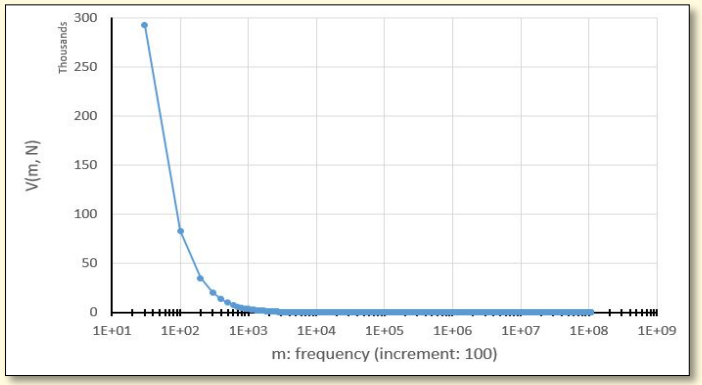
The Frequency Spectrum
Lexicon (Multi)words on MEDLINE



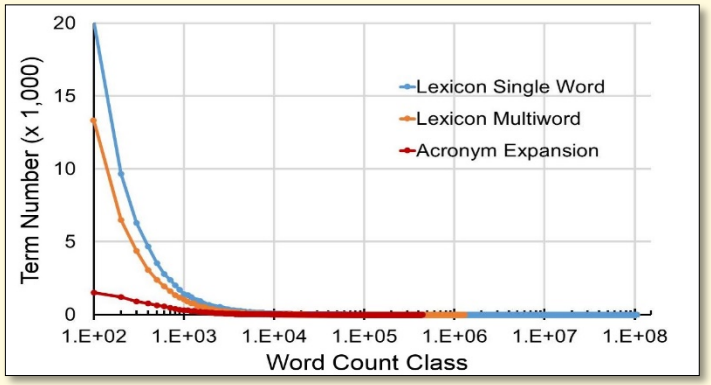


The Frequency Spectrum and Corpus Size

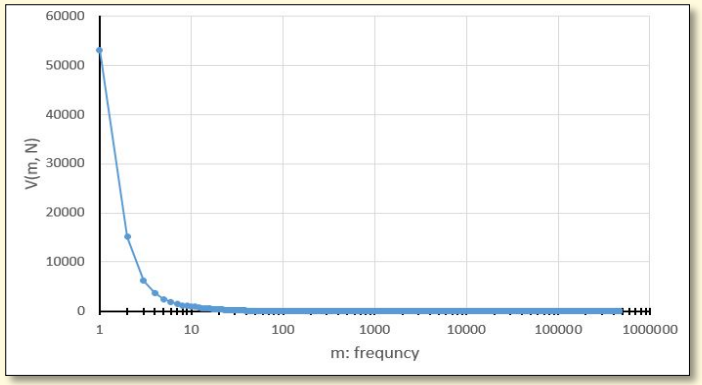
- MEDLINE 2019 (3,824,268,997 words)



- Lexicon on MEDLINE 2015 (464,781 words)



- Consumer Health Corpus (10,197,915 words)



- Alice in Wonderland (26,432 words)

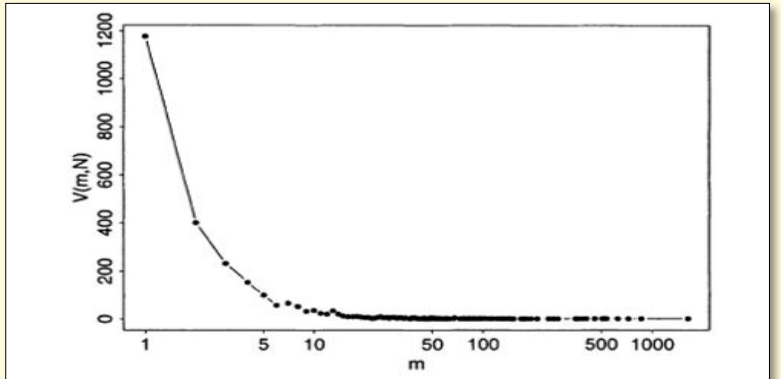
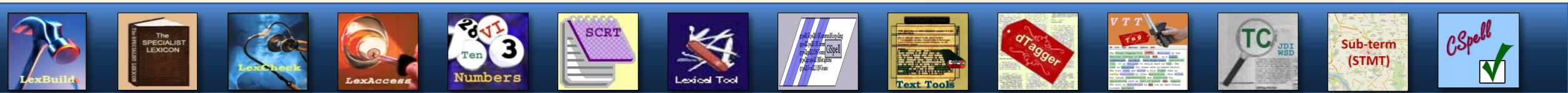


Figure 1.3: *The frequency spectrum of Alice in Wonderland (m: frequency class; V(m, N) : number of types with frequency m).*

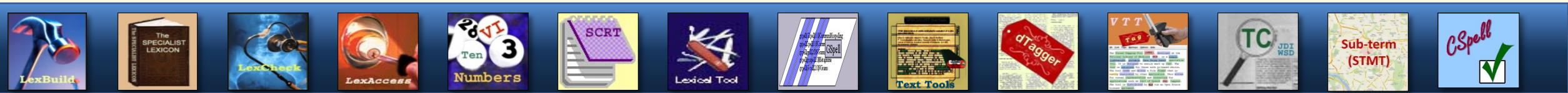


Lexicon (Data) and Lexical Tools (Software)

DB - LR 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
}
```

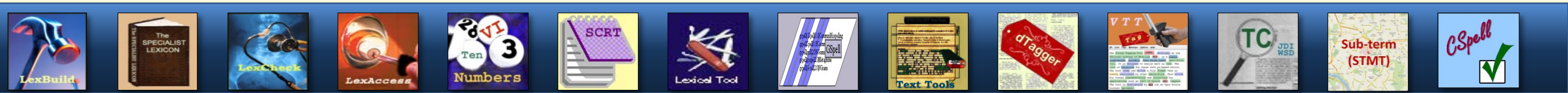
- spelling_variant → spelling variant
- cat=verb → part of speech
- variants=reg → inflectional variant
- tran=pphr(from,np) → chunker
- nominalization=generalisation|noun|E0029525 → derivational variant, synonym



2. NLP - Lexical Tools

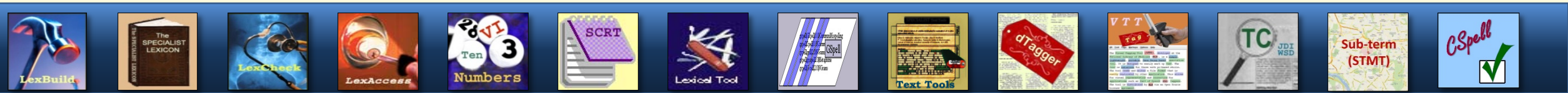
- Lexical Tools: Algorithm + Data (directly or derived from the Lexicon)
  - Command line tools
    - lvg (Lexical Variants Generation, base of all of tools)
    - norm (UMLS - MRXNS, MRXNW)
    - luiNorm (UMLS - LUI)
    - wordInd (UMLS - MRXNW)
    - toAscii (MetaMap - BDB Tables)
    - fields (Lexicon Tables, MetaMap - BDB Tables, etc.)

  - Lexical Gui Tool (lgt)
  - Web Tools
  - Java API's



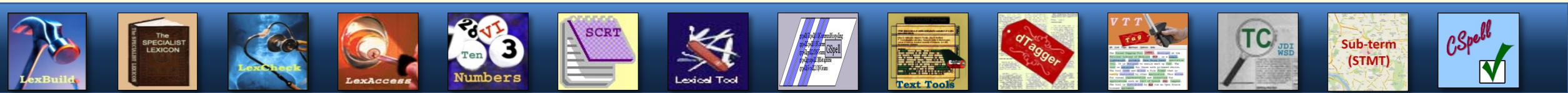
Lexical Tools - Facts

- 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)



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

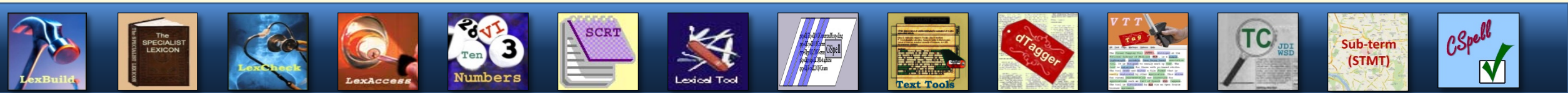
Derivational variants:
- suffixD: generalisation, generalization, generalisable
- prefixD: overgeneralise, over-generalise

Synonyms: generalize

← Multiple LexRecords + Algorithm

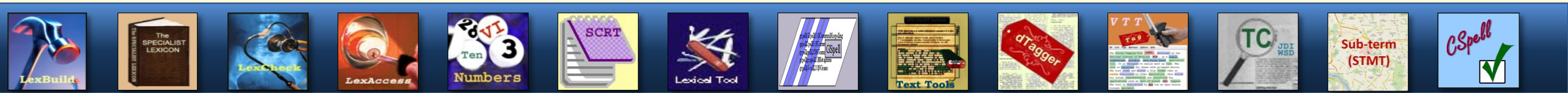
Fruitful Variants: generalisability, generalisable, generalisation, generalisations, generalised, generalises, generalising, generalizability, generalizable, generalization, generalizations, generalize, generalized, generalizer, generalizers, generalizes, generalizing, overgeneralize, etc.

- Enhancing LexSynonym Features in the Lexical Tools, AMIA 2017
- Enhanced LexSynonym Acquisition for Effective UMLS Concept Mapping, MedInfo 2017

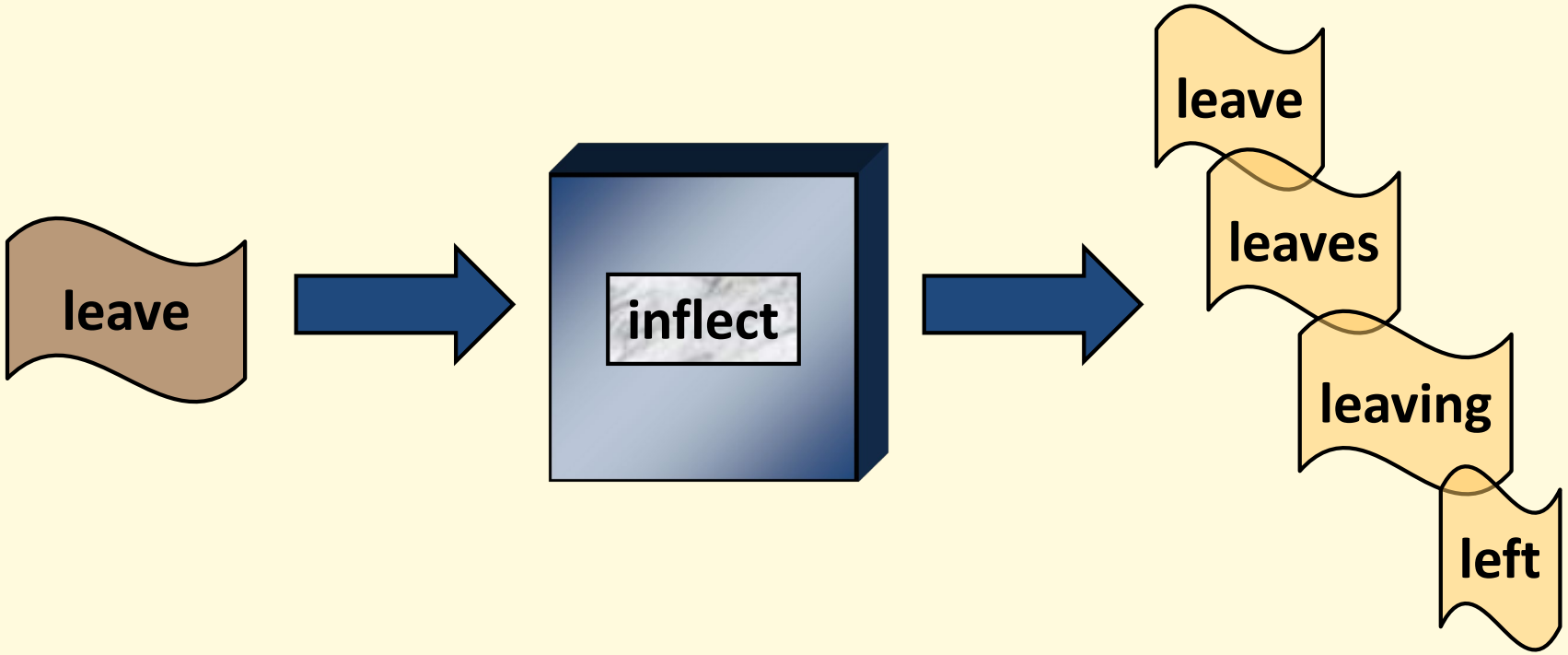


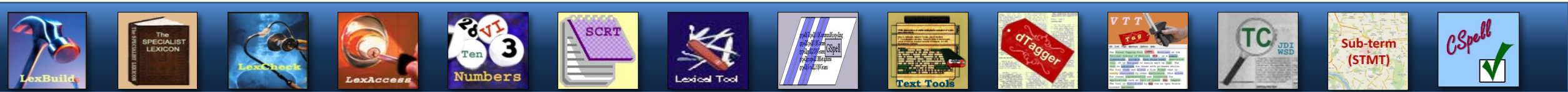
Lexical Tools – Flow Components (62)

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,	…



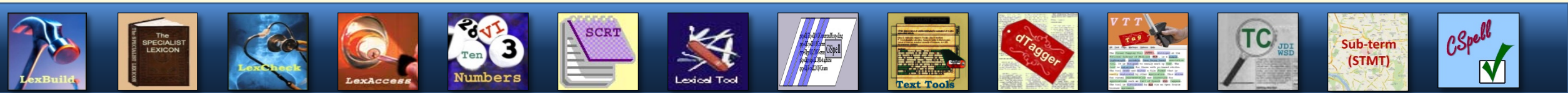
LVG Flow Component – Example





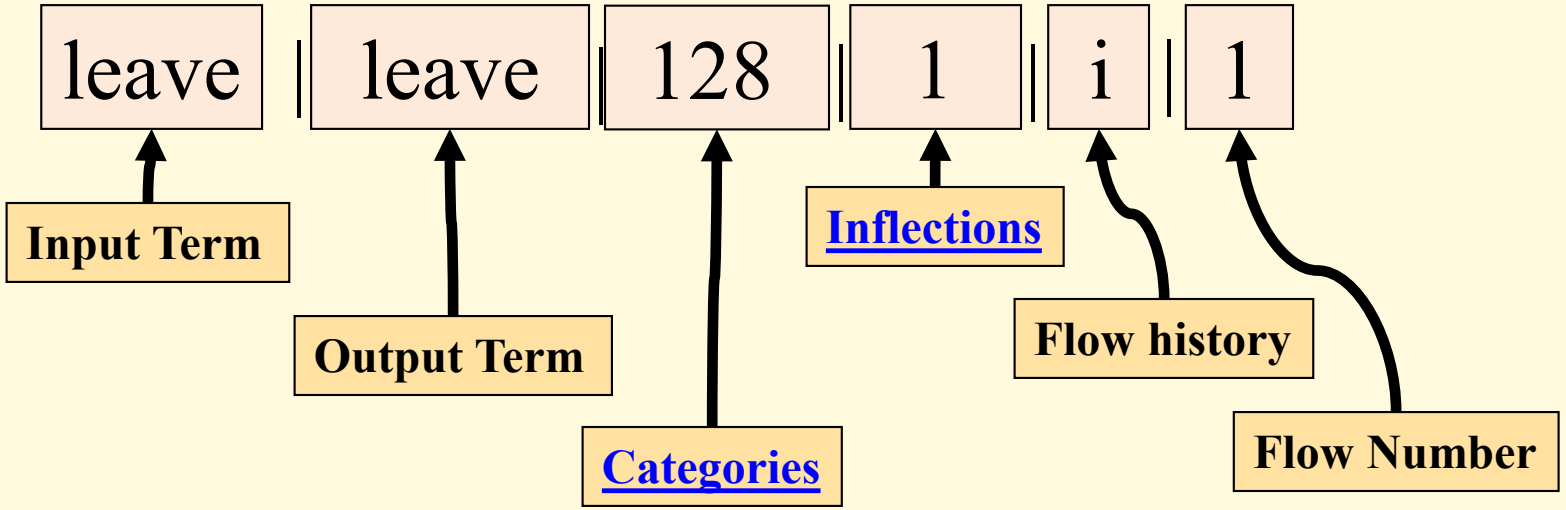
LVG Flow Component – CmdLine

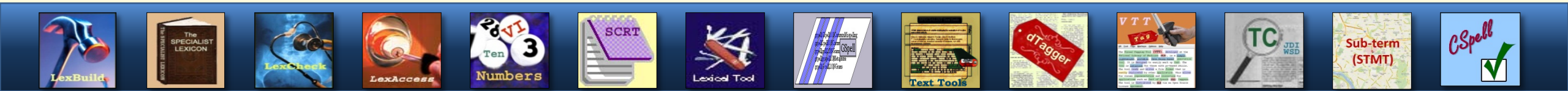
```
> lvg -f:i
leave
leave|leave|128|1|i|1|
leave|leave|128|512|i|1|
leave|leaves|128|8|i|1|
leave|left|1024|64|i|1|
leave|left|1024|32|i|1|
leave|leave|1024|1|i|1|
leave|leave|1024|262144|i|1|
leave|leave|1024|1024|i|1|
leave|leaves|1024|128|i|1|
leave|leaving|1024|16|i|1|
```



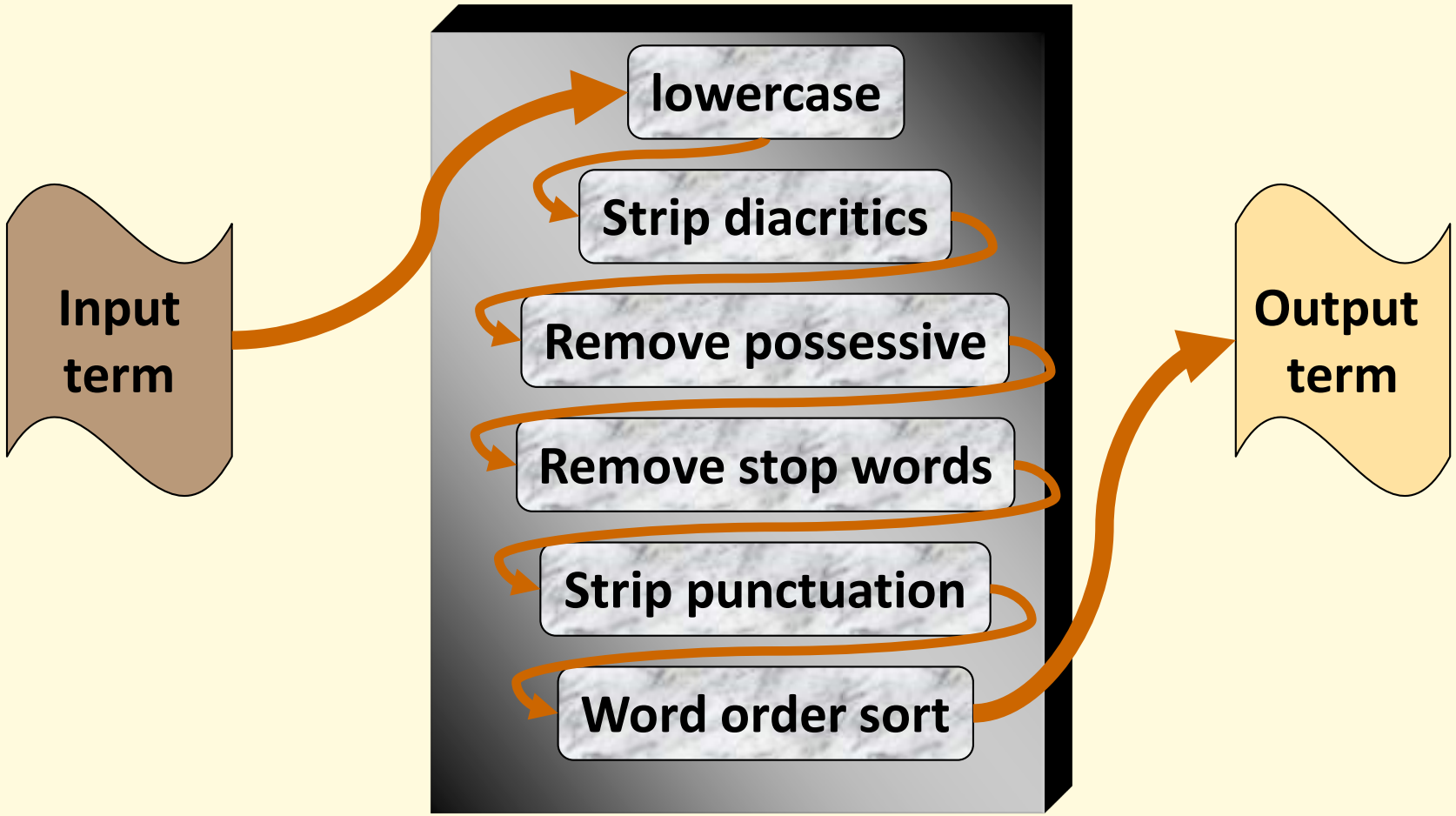
LVG Flow Component – Fielded Output

```
> lvg –f:i
  leave
```

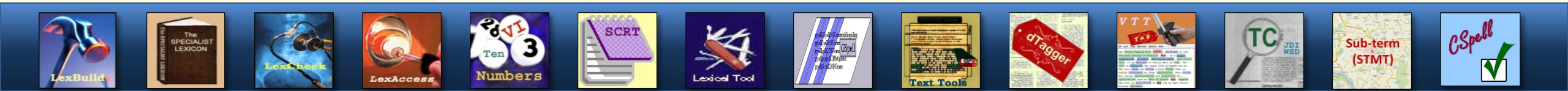




LVG – A Serial Flow



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



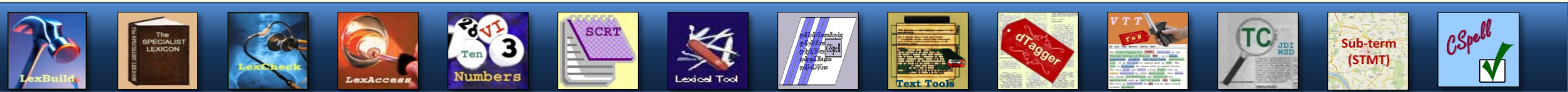
A Serial Flow - Example

```
➢ 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|
```

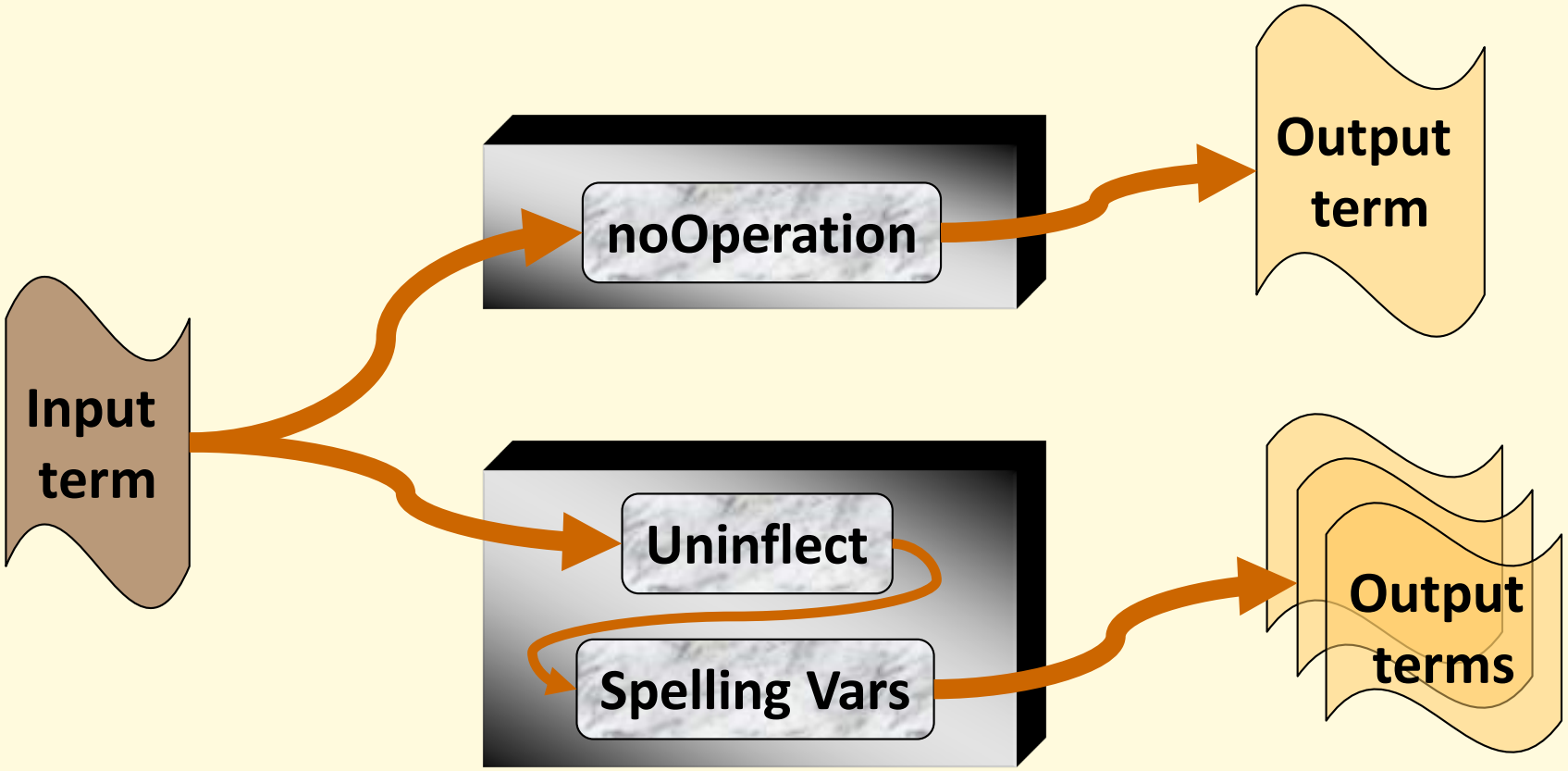
Input

Output

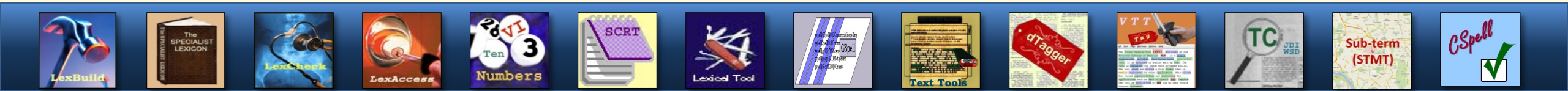
Other information



LVG - Parallel Flows

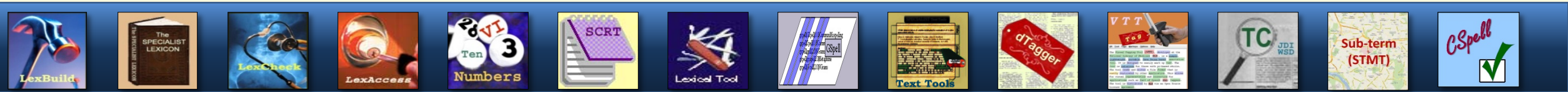


- Multiple flows can be defined



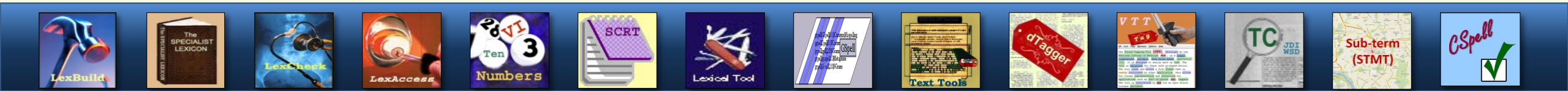
Parallel Flows - Example

```
> 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|
```



Norm (commonly used flow)

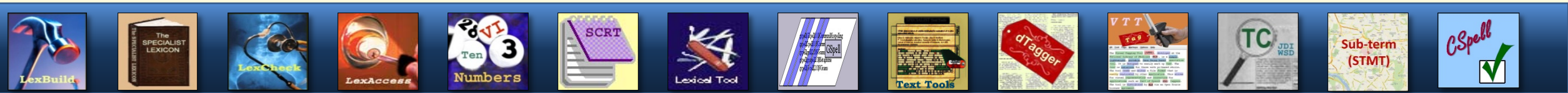
- Composed of 11 Lvg flow components to abstract away from (only keep meaningful words):
  - case
  - punctuation
  - possessive forms
  - inflections
  - spelling variants
  - stop words
  - diacritics & ligatures (non-ASCII Unicode)
  - word order



Example - Norm

"Fœtoproteins α's, NOS"

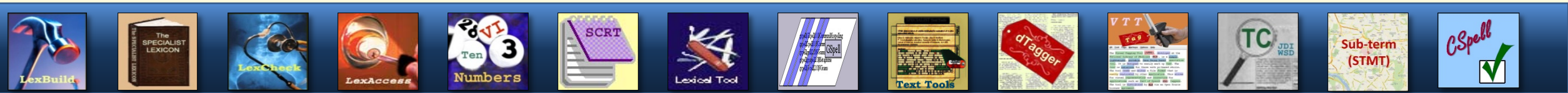
q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order



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

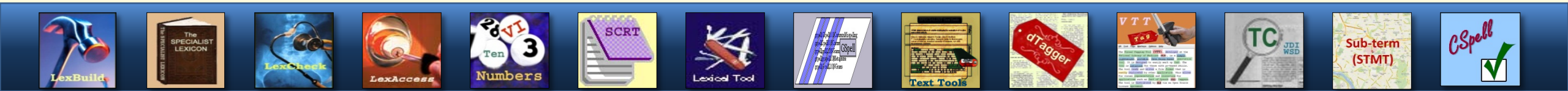
“Fœtoproteins α’s, NOS“
"Fœtoproteins α’s, NOS"



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

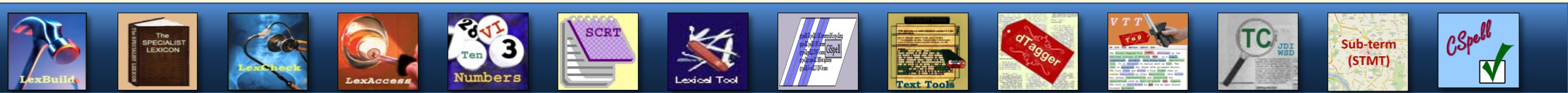
“Fœtoproteins α’s, NOS“
"Fœtoproteins α**’s**, NOS"
"Fœtoproteins α, NOS"



Norm

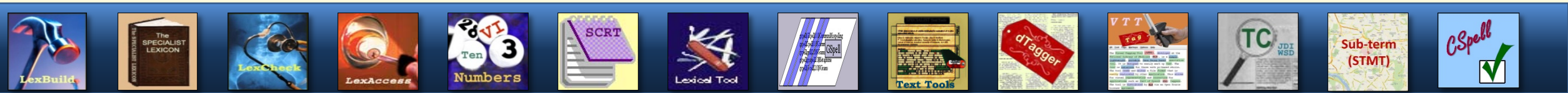
q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order

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



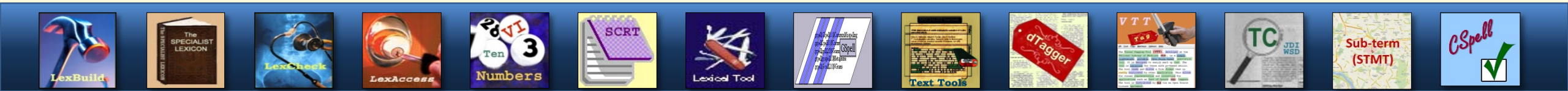
Norm

q0: map symbols to ASCII	“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	
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	



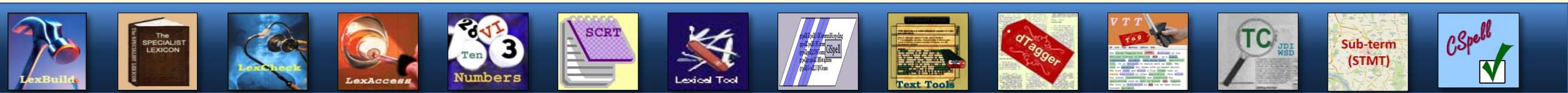
Norm

q0: map symbols to ASCII	“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	
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	“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	
q7: Unicode core Norm	
q8: strip or map Unicode to ASCII	
w: sort words by order	

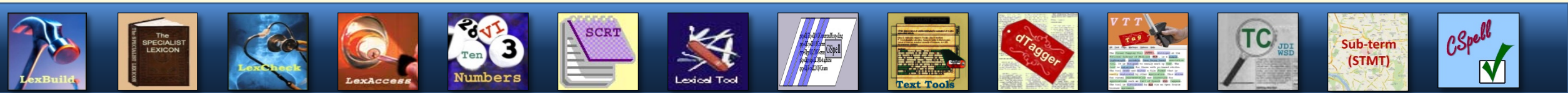


Norm

	“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	

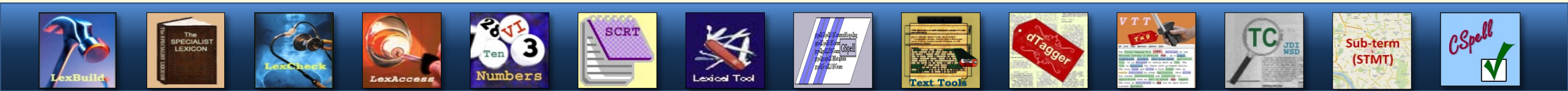
Norm

	“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	



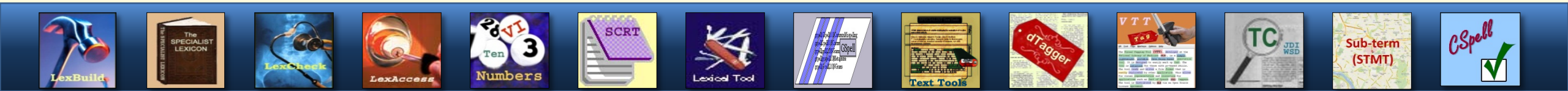
Norm

	“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	



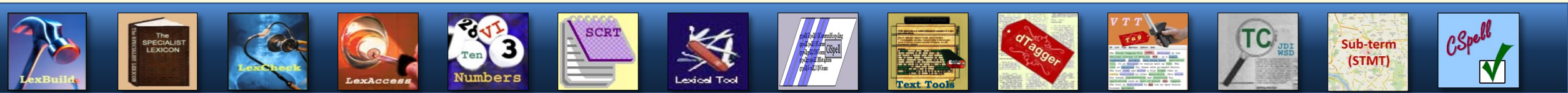
Norm

	“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

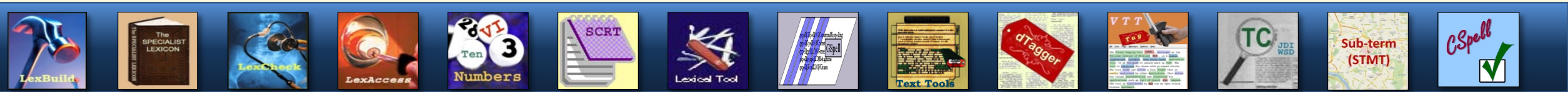
	“Fœtoproteins α’s, NOS“
q0: map symbols to ASCII	"Fœtoproteins α’s, NOS"
g: remove genitives	"Fœtoproteins α, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	Fœtoproteins α NOS
t: strip stop words	Fœtoproteins α
l: lowercase	fœtoproteins α
B: uninflect 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 (Why?)

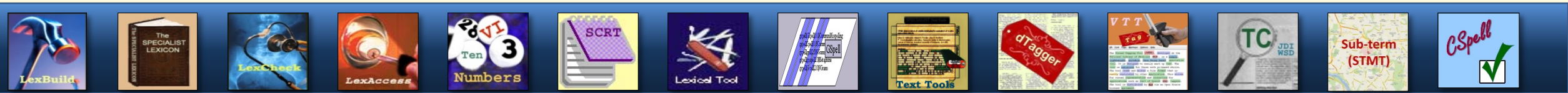
alpha Fetoprotein
alpha Fetoproteins
alpha-Fetoprotein
alpha-Fetoproteins
Alpha fetoproteins
alpha fetoprotein
alpha Foetoprotein
alpha foetoprotein
alpha fetoproteins
Alpha-fetoprotein
alpha-fetoprotein
Alpha Fetoproteins
Alpha-Fetoprotein
Alpha-fetoprotein NOS
Alpha Fetoprotein
alpha-fetoprotein
ALPHA-FETOPROTEIN
Alpha 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 (data science).
  - NLP in our scope is to use computer to understand the meaning (concept) from text for further analysis and processing.



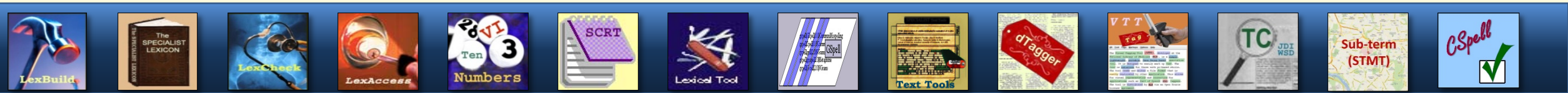
Concept Mapping Challenges

➢Challenge: many to many mapping (ambiguity)

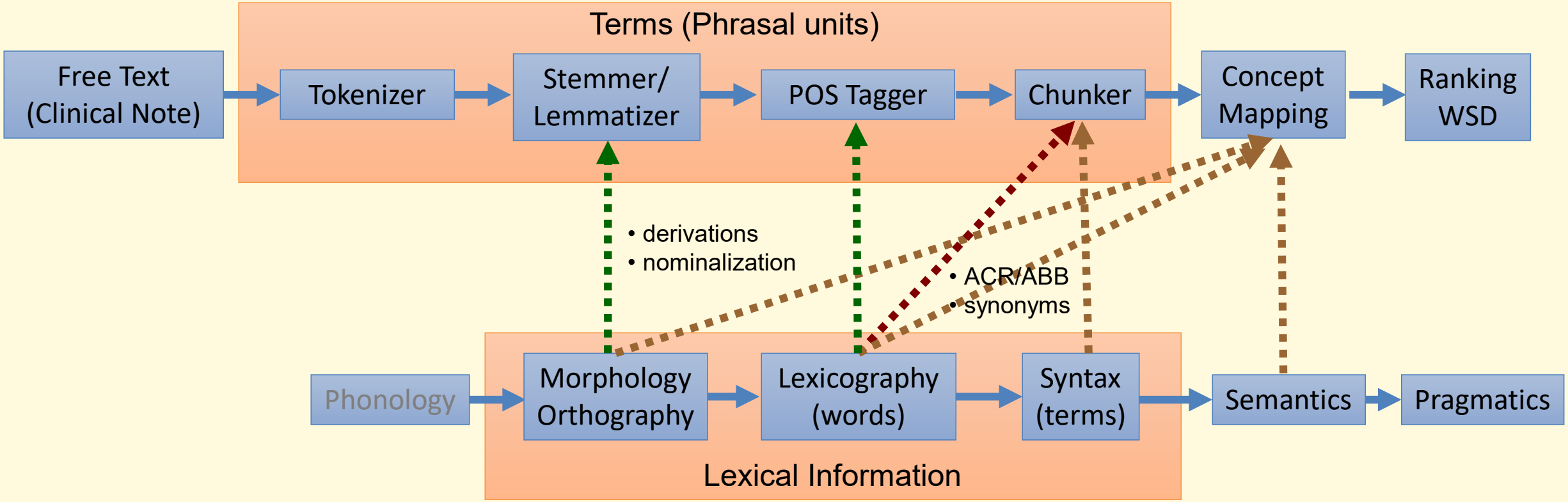
Terms	Concepts	NLP
• cold • Cold Temperature • Cold Temperatures • Cold (Temperature) • Temperatures, Cold • Low temperature • low temperatures • …	• Cold Temperature|C0009264	• Concept mapping • Normalization
• cold	• Cold Temperature|C0009264 • Common Cold|C0009443 • Cold Therapy|C0010412 • Cold Sensation|C0234192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

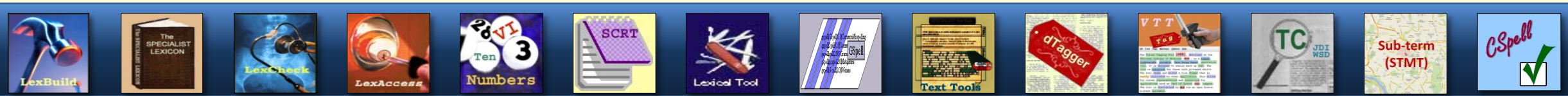
☺

☺



NLP Pipe Line – Lexical Information



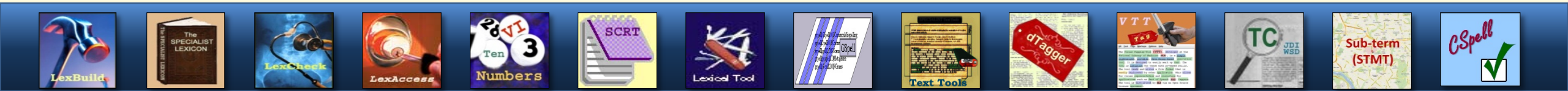


The NLP Pyramid

- Pragmatics (analyze whole text):
  Question Answering, summarization, topic segmentation, sentiment analysis, classification, machine translation
- Semantics (concept & meaning):
  Named Entity Recognition (NER), relation extraction, Semantic role labelling, Word Sense Disambiguation (WSD)
- Syntax (proper word construction):
  POS tagging, syntax tree, dependency tree
- Morphology:
  prefixes/suffixes (derivation), word inflection, lemmatization (the base form of a word), spellchecking, singularization/pluralization, gender detection.

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

☺

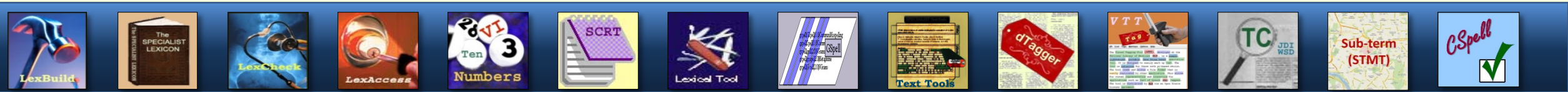


NLP Applications

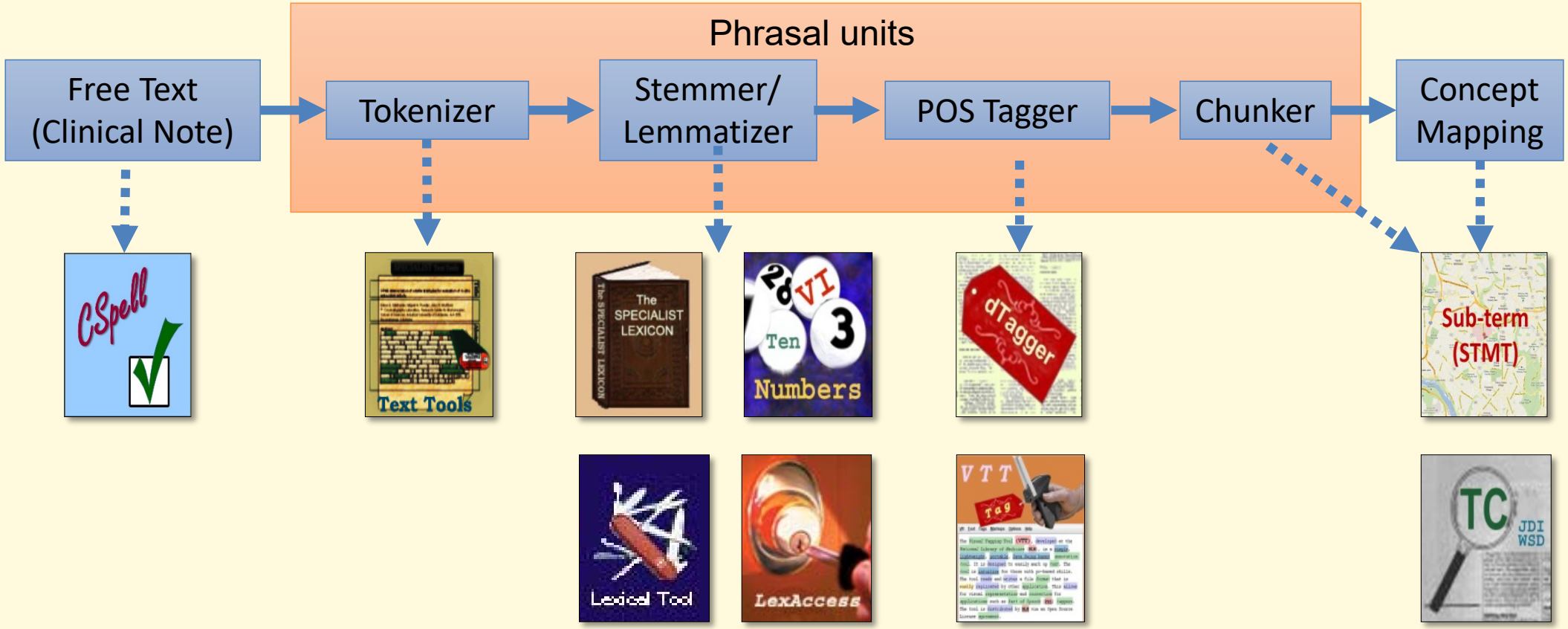
- Syntax:
  - parsers, taggers, POS tagging, etc.
- Semantics:
  - name entity recognition, **concept mapping**, etc.

- Pragmatic
  - Classification
  - Machine Translation
  - Summarization:
    - sentiment analysis and figure out the topics of a page
  - Question answering
    - find answers for queries

- Spell checking, keyword search, finding synonyms
- Knowledge (information) extraction:
  - learn relations between entities, recognize events, etc.

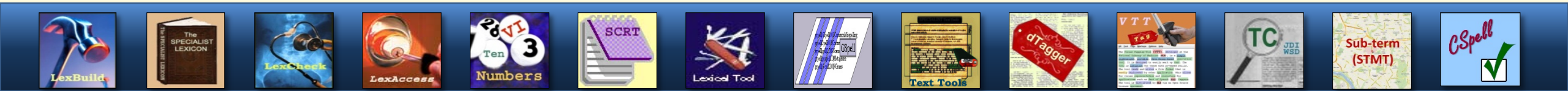


The SPECIALIST NLP Tools



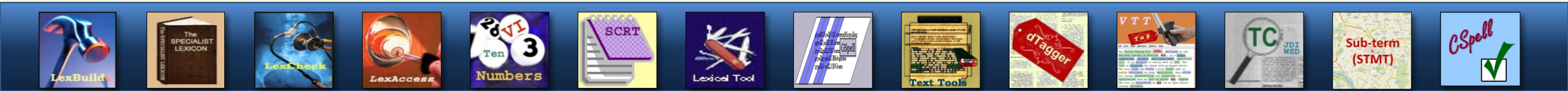
- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov

☺



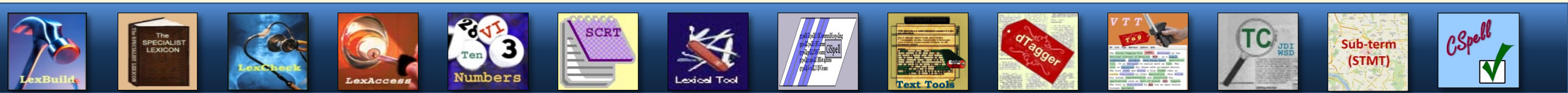
NLP – Concept Mapping

- Normalization (same record – 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 – same concept):
  - 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…

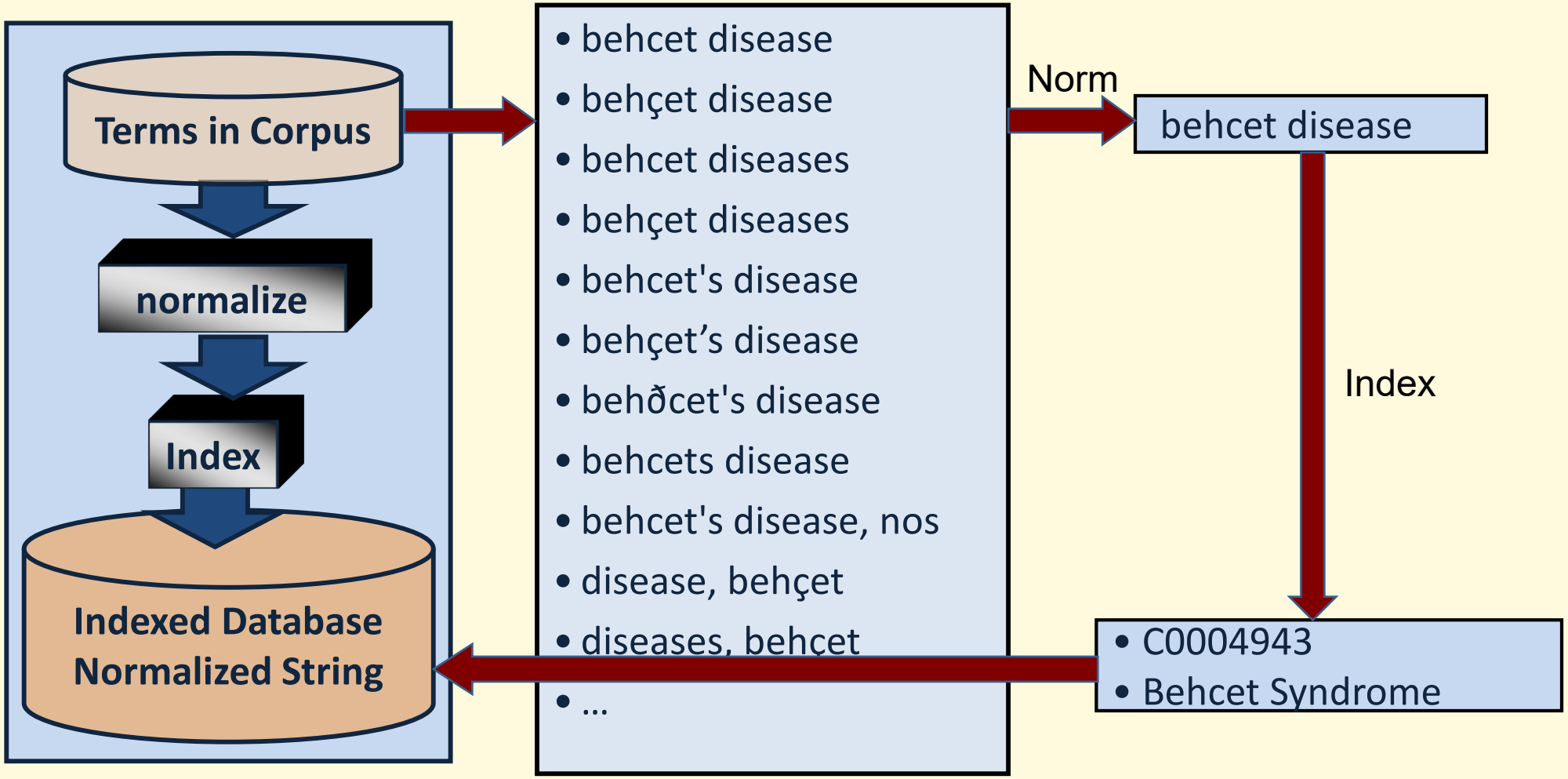


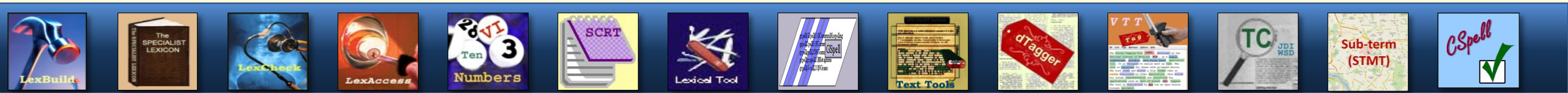
Lexical Tools – Norm

	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 (Pre-Process Lexical Variations)





NLP – Norm (Application)

Behcet's Disease, …

Query

norm

Normed Term

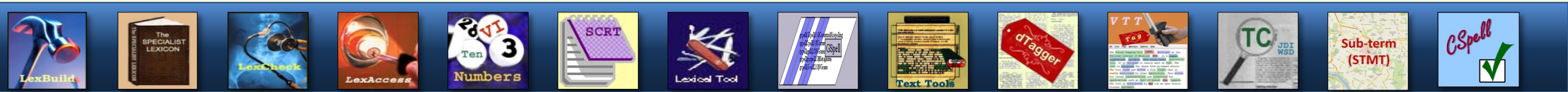
Behcet disease

SQL

Results that matches the normalized query
- C0004943
- Behcet Syndrome

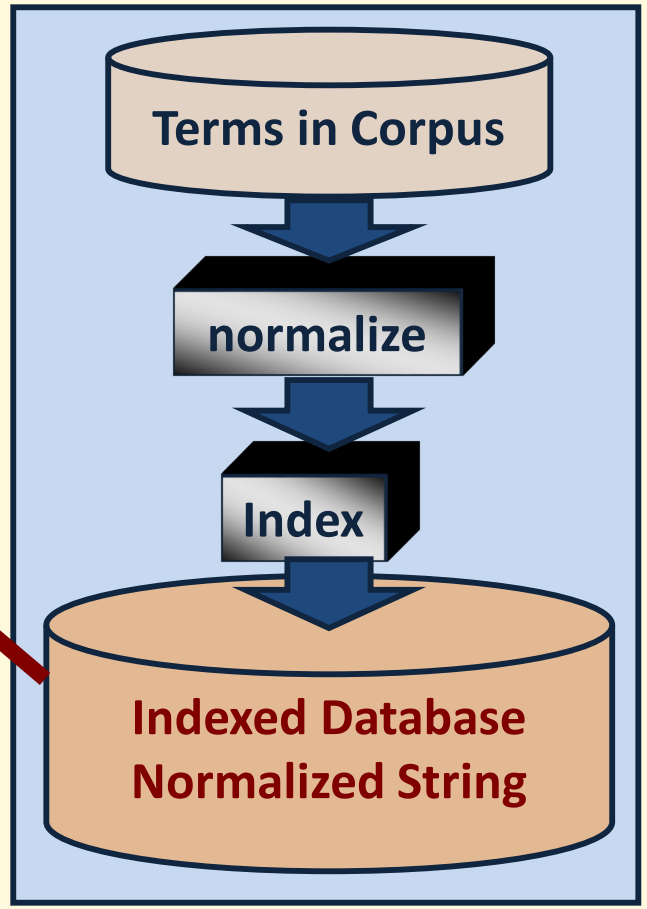
Indexed Database
Normalized String

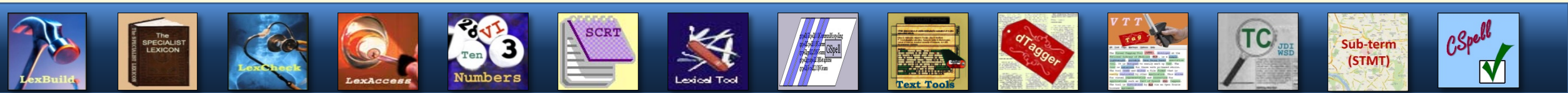
MRXNS_ENG.RRF



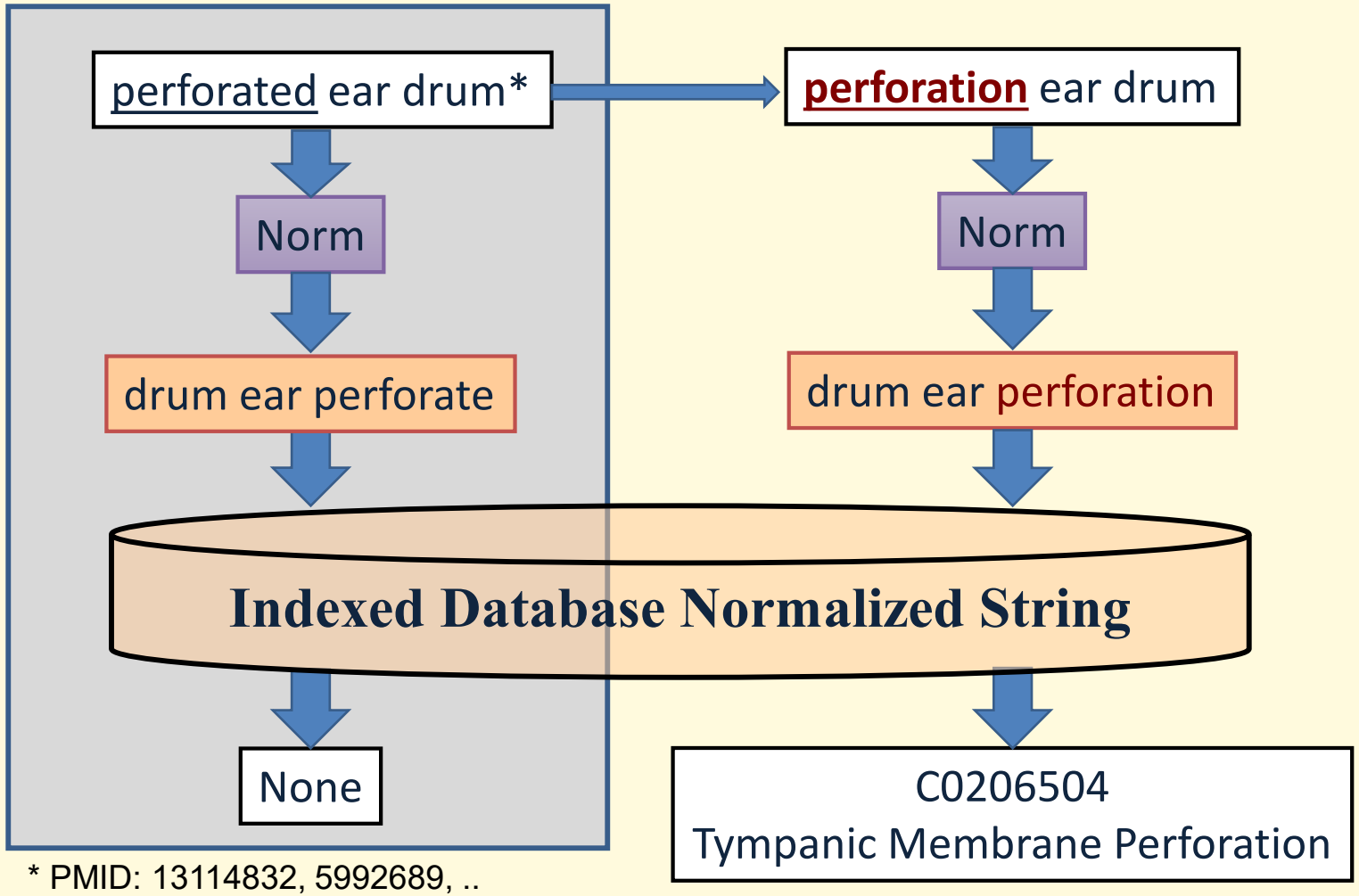
UMLS Metathesaurus

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

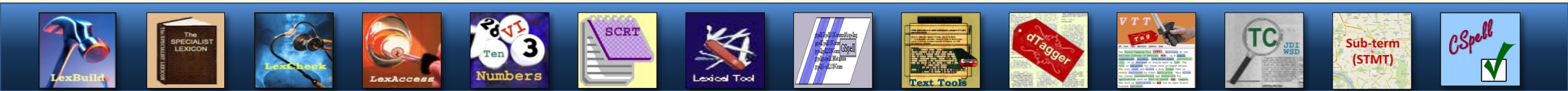




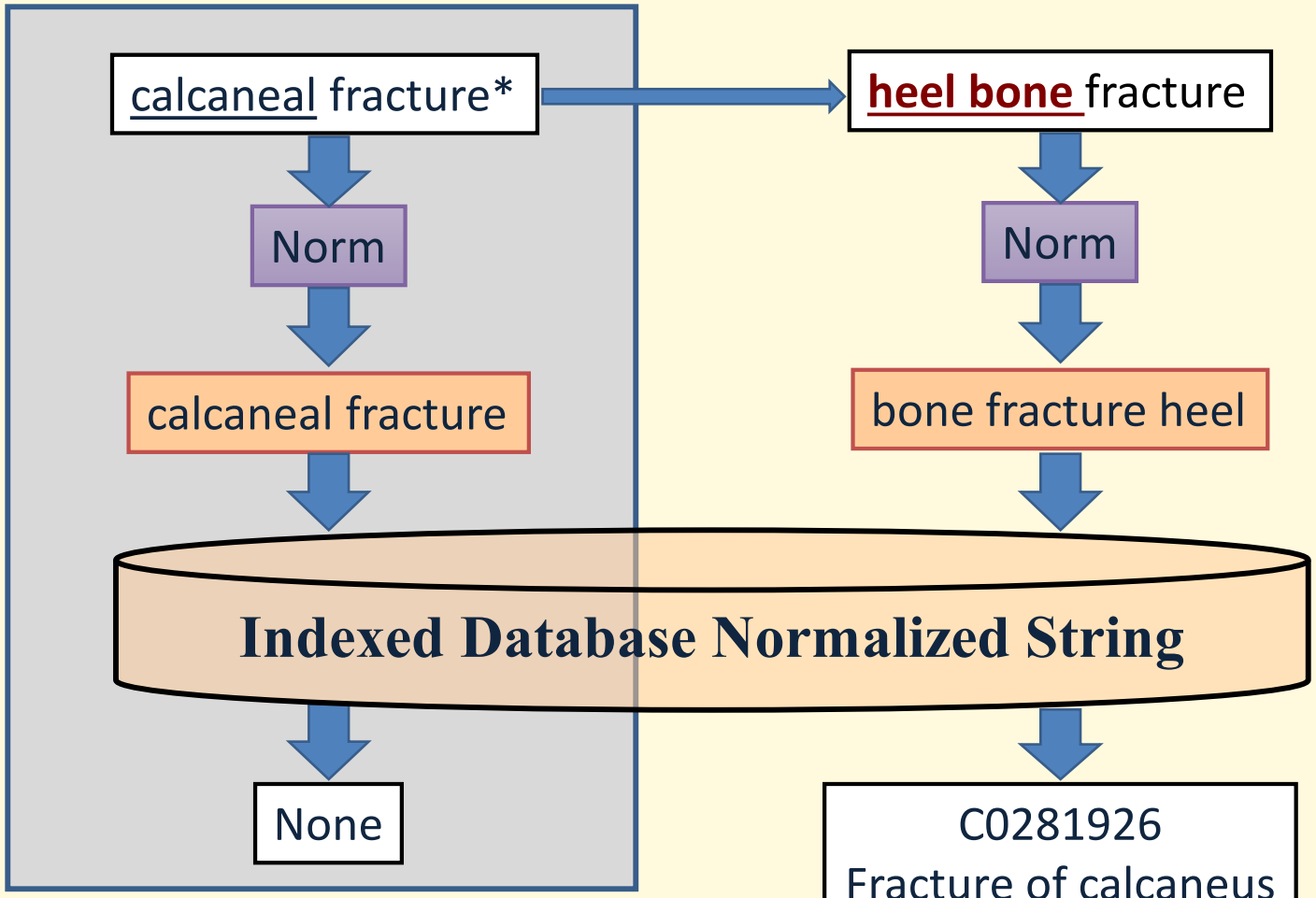
NLP – Query Expansion (Derivation)



* PMID: 13114832, 5992689, ..



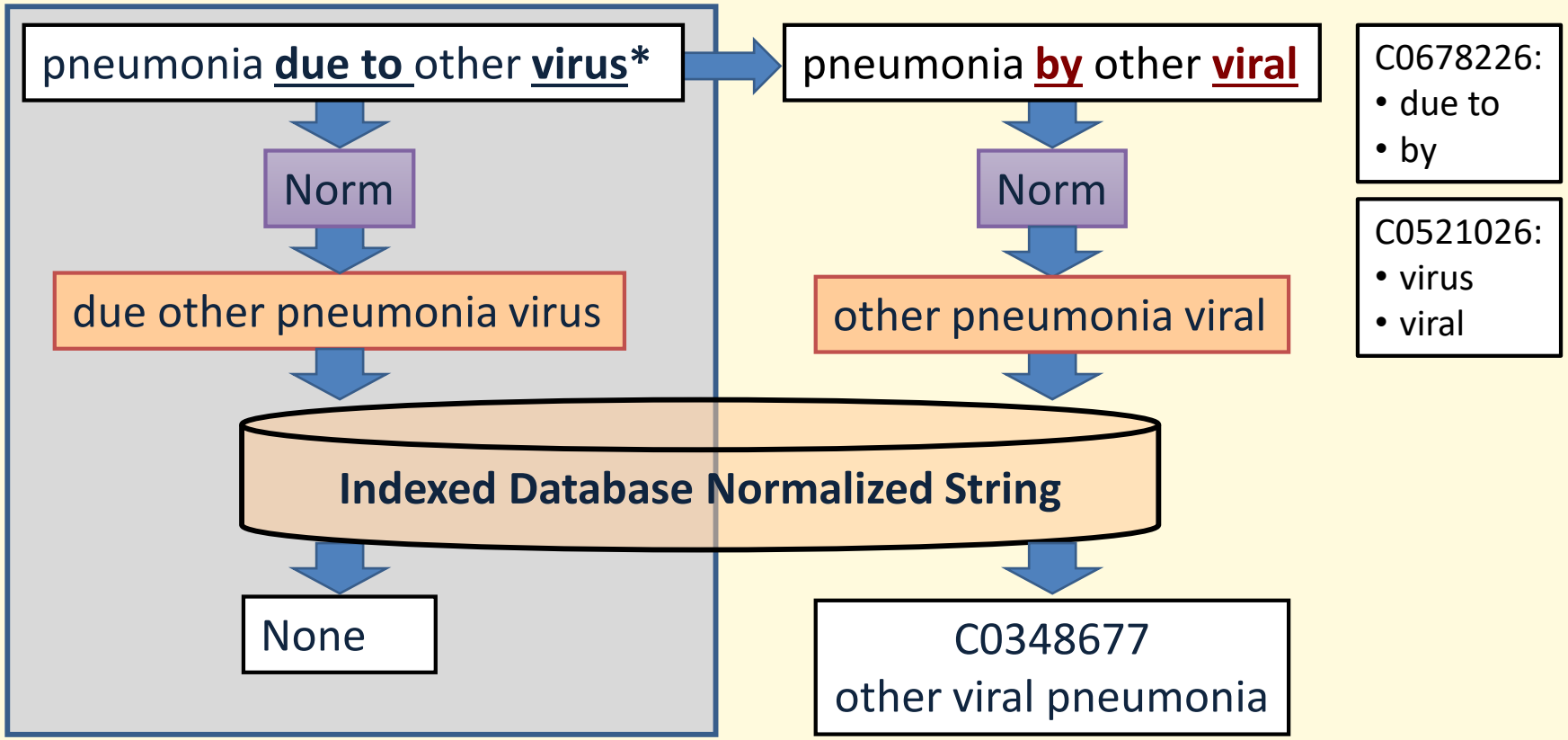
NLP – Query Expansion (Synonym)



C0006655:
- calcaneal
- heel bone

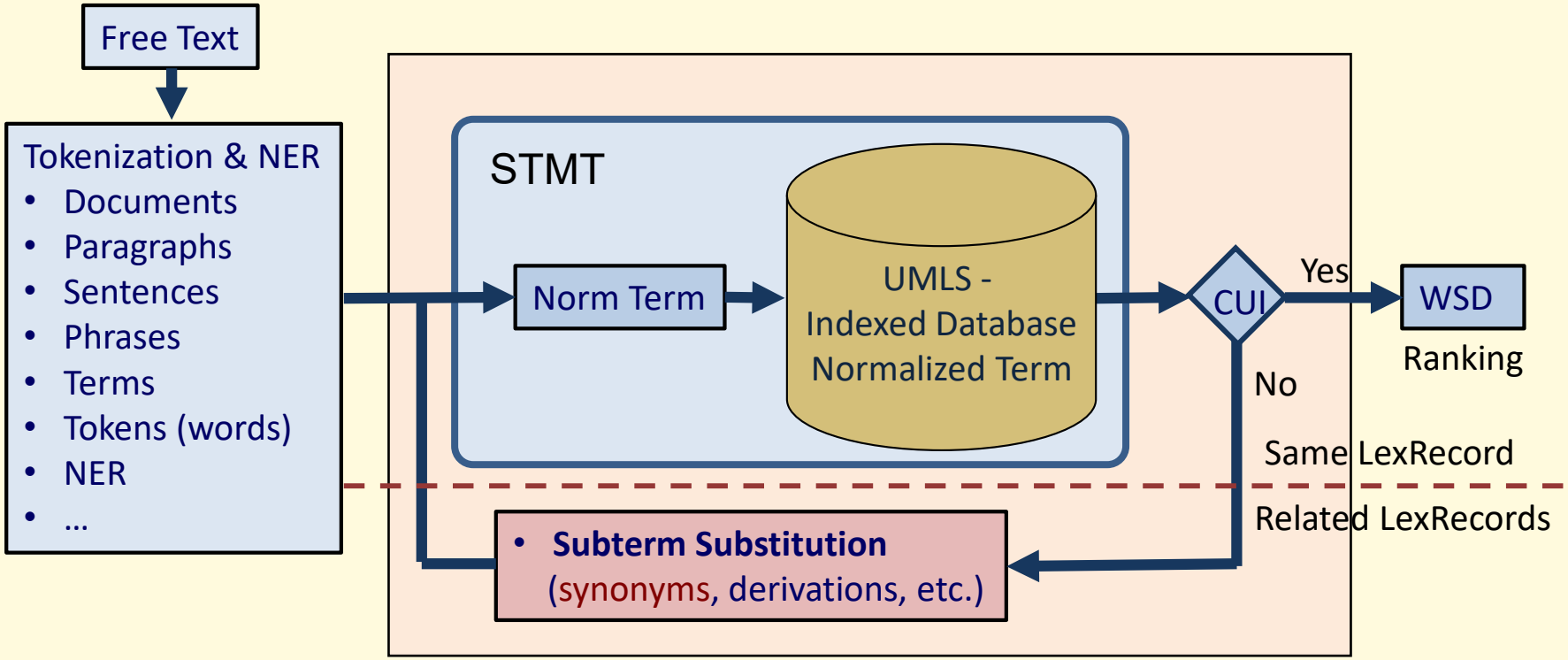
* PMID: 1118604, 1165396, ..

Multiple Substitutions

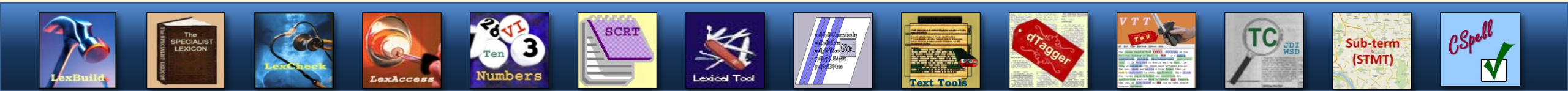


* VA14760, HA480.80, ..

Real-time Concept Mapping Model

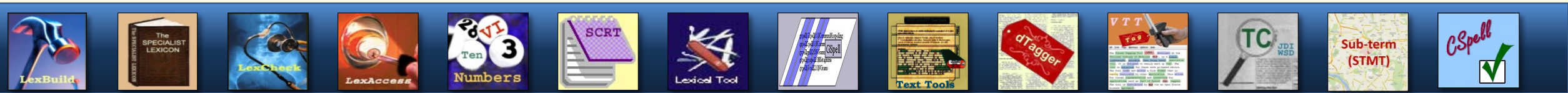


- Enhanced LexSynonym Acquisition for Effective UMLS Concept Mapping, MedInfo 2017
- Development of Sub-Term Mapping Tools (STMT), AMIA 2012



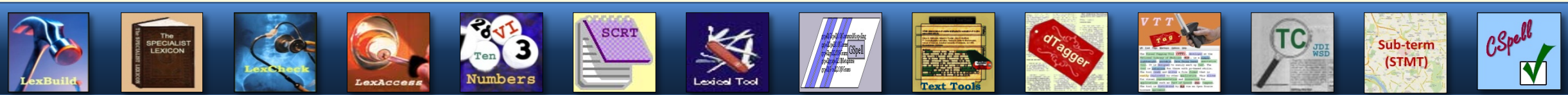
4. Current Research

- CSpell
- Multiwords
- Derivations
- Synonyms
- Antonyms



CSpell

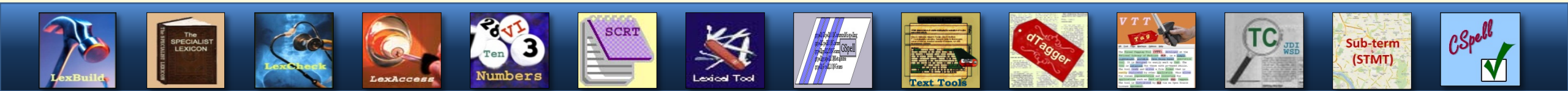
- JAMIA Journal Club Webinar, 04/11/2019
- "Spell checker for consumer language (CSpell)", **Editor's Choice paper in JAMIA**, Volume 26, Issue 3, 1 March 2019, Pages 211-218
- "Improving Spelling Correction with Consumer Health Terminology", AMIA 2018 Annual Symposium



CSpell - Background

- 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)

☺



CSpell Application Example

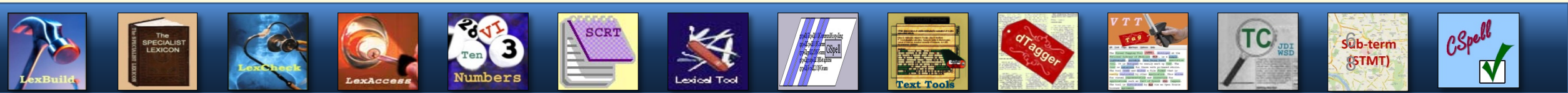
My mom is 82 years old suffering from anixity and depression for the last 10 years was dianosed early on set deminita 3 years ago. Do yall have a office in Greensboro NC? Can you recommend someone. she has seretona syndrome and nonething helps her. [2]

- Corrections:

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

- 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 Types Analysis from Training Set

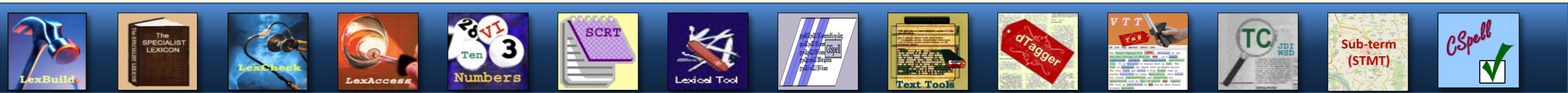
- 471 Consumer health questions
- 24,837 Tokens

Distribution of errors in the training set

Correction needed	Nonwords	Real words	ND	Multiple[a]	Total by type
Spelling	348	153	113	N/A	614
Merge	10	38	0	N/A	48
Split	24	10	281	N/A	315
Multiple	N/A	N/A	N/A	31	31
Total	382	201	394	31	1008
Percentage	37.90	19.94	39.09	3.08	100.00

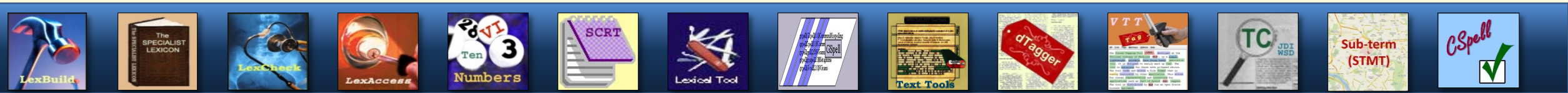
Note: ND: not dictionary based.

[a]Errors that combine several types and require multiple corrections.



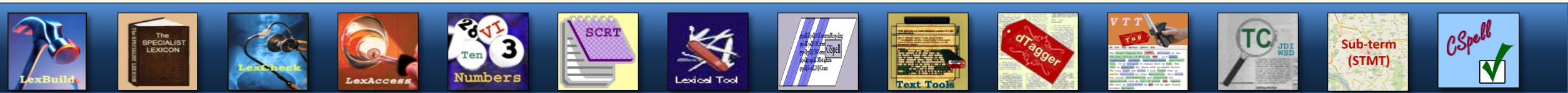
Dictionary-based Examples

Nonword (38%)	Real-word (20%)
➢ Spelling: • dianosed => diagnosed	➢ Spelling: • bowl movement => bowel movement
➢ Split: • knowabout => know about	➢ Split: • for along time => for a long time
➢ Merge: • dur ing => during	➢ Merge: • diagnosed on set => diagnosed onset



Non-dictionary-based Examples

Handler (11%)	Splitter (28%)
➢ XML/HTML: • "germs" => “germs” • & => & ➢ Informal expression: • pls => please • whos => who’s	➢ Leading digit: • 1.5years => 1.5 years • 42nd ➢ Ending digit: • from2007=> from 2007 • Co-Q10 ➢ Leading punctuation &([{: • volunteers(healthy) => volunteers (healthy) • finger(s) ➢ Ending punctuation : .?!,:;&)]}: • (..)why=> (..) why • NAD(P)H

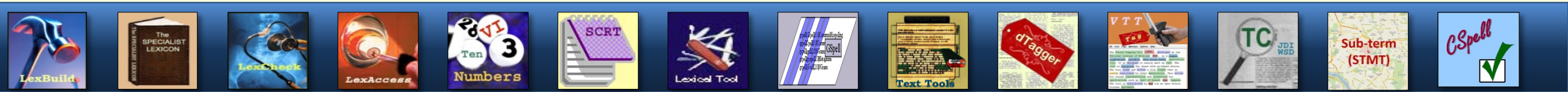


Dictionary-based Model

Detector → Candidates → Ranker → Corrector

- Example:

Input	Detector	Candidates	Ranker	Corrector
diagnost	nonword	• diagnose • diagnosed • diagnostic • diagnosis • diagnoses • diagnoser • …	1) diagnosis 2) diagnosed 3) …	diagnosis



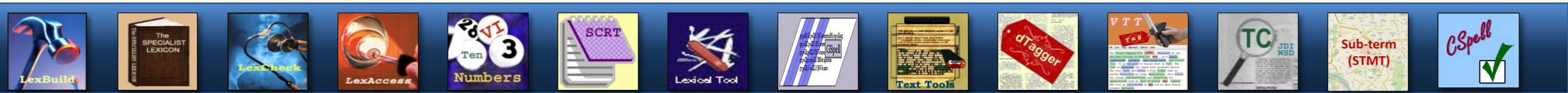
Toward Context-dependent Corrections

- Isolated-word corrections

Input	Technique	Correction
diagnost	Orthographic	diagnose
diagnost	Word frequency	diagnosis
diagnost	Noisy Channel	diagnosis

- Context-dependent corrections
  - Word2vec – CBOW Dual Embedding

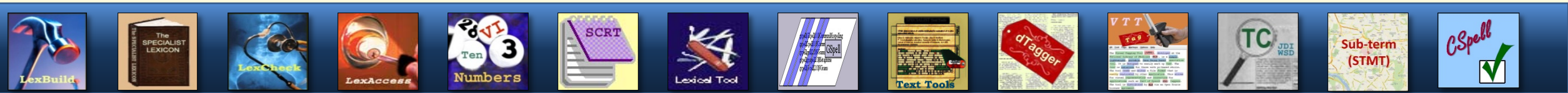
Input	Correction
the diagnost	the diagnosis
was diagnost	was diagnosed



2-stage Ranking Example

- Input: **havy**
- Candidates: 441
  - Stage 1: find top candidates by orthographic similarity score
  - Stage 2: use context score, then Noisy Channel score

Input Text	Correction	Stage-1 Scores	N.C. Scores	Stage 2 - Context Scores			
				heavy	have	hay	wavy
havy	**have***			0.0000	0.0000	0.0000	0.0000
havy duty	**heavy** duty	2.25	0.00198	**0.0597**	-0.0302	-0.0053	0.0074
havy diabetes	**have** diabetes	2.20	**0.14933**	-0.0067	**0.0586**	-0.0518	-0.0813
havy fever	**hay** fever	2.13	0.00032	-0.1331	**0.2280**	**0.2292**	-0.0391
havy lines	**wavy** lines	2.13	0.00004	-0.0170	-0.0410	-0.0702	**0.1495**



CSpell Multiple Correction Examples

Ex-1: Input Text	Output: different types of corrections
He was dianosed early on set deminita 3years ago.	He was diagnosed early onset dementia 3 years ago.

NW Spelling

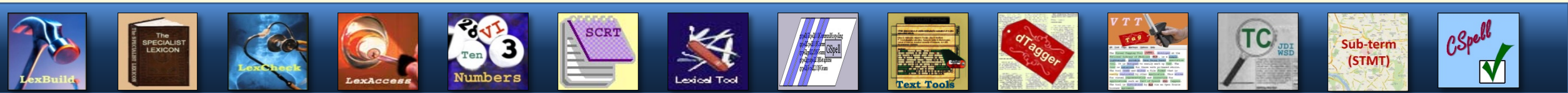
RW Merge | NW Spelling | ND Split

Ex-2: Input Text	Output: multiple corrections
I have a shuntfrom2007.	I have a shunt from 2007.

NW Split | ND Split

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 antidepressants.

RW Merge | NW Spelling



Results

➢ Training Set Detection:

Method	Precision	Recall	F1	δ F1		
ESpell	0.3475	0.4253	0.3825	-----	-0.11	-0.31
Jazzy	0.8499	0.3465	0.4923	0.11	-----	-0.20
Ensemble	0.8078	0.6017	0.6897	0.31	0.20	-----
CSpell	0.9289	0.7178	**0.8098**	**0.42**	**0.32**	**0.12**

➢ Test Set Detection:

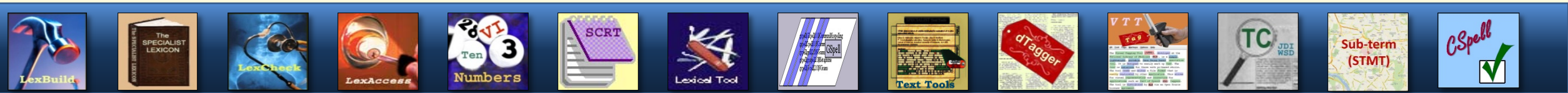
Method	Precision	Recall	F1	δ F1
Ensemble	0.8210	0.5645	0.6690	-----
CSpell	0.8900	0.7419	0.8093	**0.14**

➢ Training Set Correction:

Method	Precision	Recall	F1	δ F1		
ESpell	0.2076	0.2541	0.2285	-----	-0.05	-0.38
Jazzy	0.4860	0.1981	0.2815	0.05	-----	-0.33
Ensemble	0.7201	0.5363	**0.6147**	0.38	**0.33**	-----
CSpell	0.8416	0.6504	**0.7338**	**0.50**	**0.45**	**0.12**

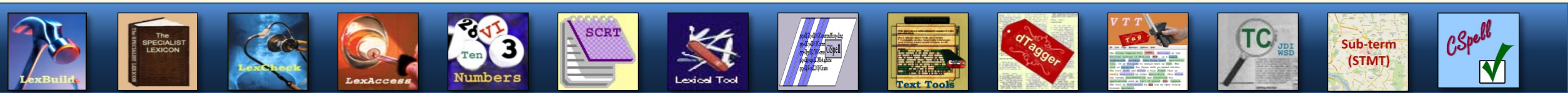
➢ Test Set Correction:

Method	Precision	Recall	F1	δ F1
Ensemble	0.6975	0.4796	0.5684	-----
CSpell	0.7607	0.6341	0.6917	**0.12**



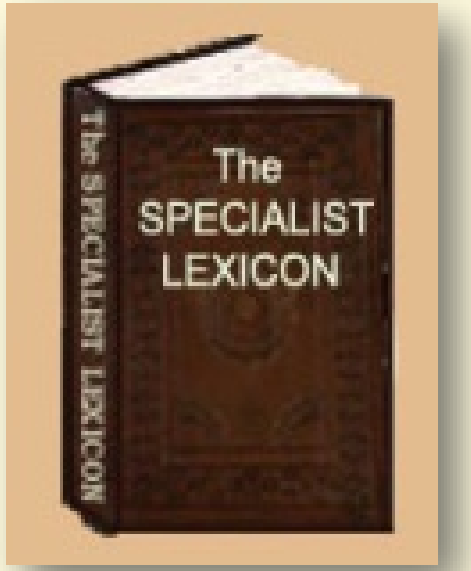
CSpell Summary

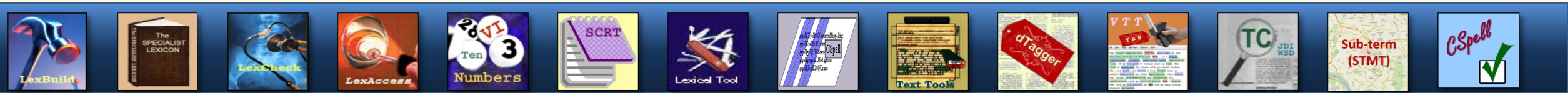
- A deployable context-dependent correction tool
- Correct all types of errors in consumer language
- Open source with public supports
- Provides many configurable options
  - Dictionary
  - Corpus
  - Types of corrections
- Configuration file
  - Default with empirical best values of thresholds and other variables
- Command line tool and Java APIs



Lexicon – Multiwords (MWE)

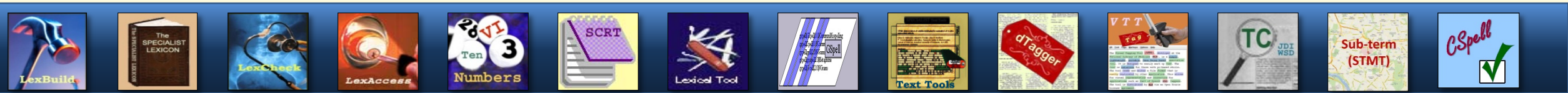
- "The Unified Medical Language System SPECIALIST Lexicon and Lexical Tools: Development and applications", Journal of the American Medical Informatics Association, 29 May 2020
- "Generating A Distilled N-Gram Set: Effective Lexical Multiword Building in the SPECIALIST Lexicon", The 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017)
- "Multiword Frequency Analysis Based on the MEDLINE N-Gram Set", AMIA 2016 Annual Symposium
- "Generating the MEDLINE N-Gram Set", AMIA 2015 Annual Symposium
- "Using Element Words to Generate (Multi)Words for the SPECIALIST Lexicon", AMIA 2014 Annual Symposium





Lexicon Multiwords - Introduction

- 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:
    - hot dog (high temperature canine?)
  - Word order:
    - trial and error, up and down (vs. food and water)
    - exercise training vs. training exercise (military)



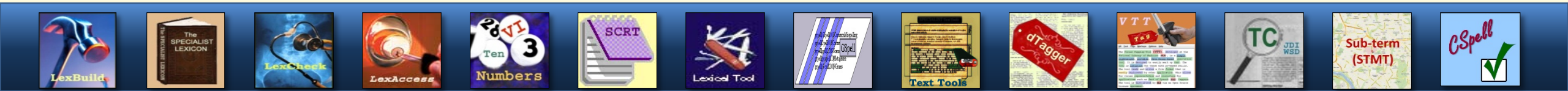
Multiwords - Application Examples

- Example (PMID 9510650, TI):
- Multiword Approach:
  - Name Entity Recognition (NER)
  - POS tagger, parser
  - Concept mapping (the longest word)

- The changes of tear break up time after myopic excimer laser photorefractive keratectomy

E0635418|Noun → C2111106 Lacrimation tear break-up time

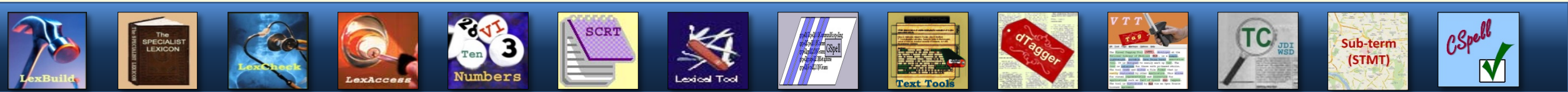
E0764487|Noun → C2069669 excimer laser photorefractive keratectomy



Multiwords Examples (continued)

- Install the STMT (Sub-Term Mapping Tools), https://umlslex.nlm.nih.gov/stmt
- Run the LSF (LexItem Subterm Finder)
- Two longest Lexicon multiwords in the sentence are identified, which are used as name entity for the key concepts in that sentence.
- This feature is implemented in MetaMap Lite.

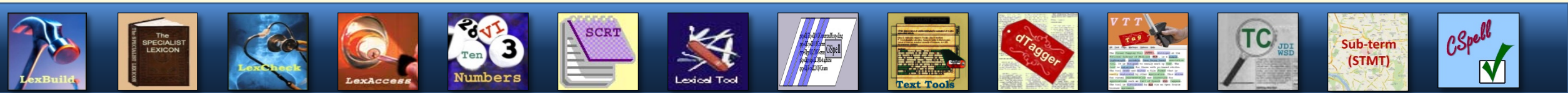
```
shell> lsf –p
- Please input a term (type "Ctl-d" to quit) >
The changes of tear break up time after myopic excimer laser photorefractive keratectomy
--- LexItem Multiword Subterms ---
break up time|E0635415
photorefractive keratectomy|E0225495
break up|E0220309
excimer laser|E0514806
excimer laser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



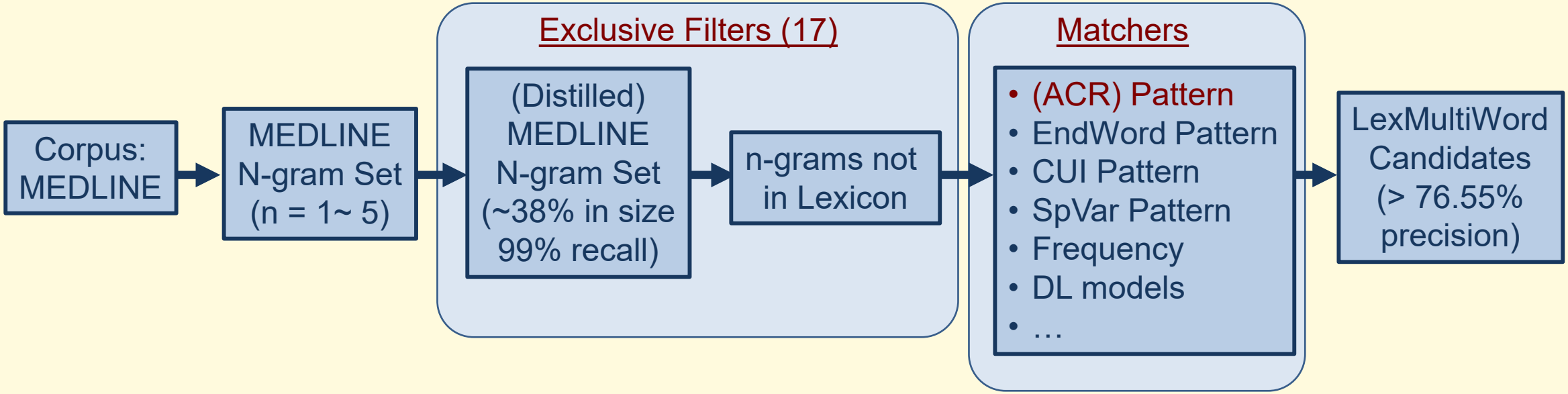
Multiwords Examples (continued)

- Install the STMT (Sub-Term Mapping Tools), https://umlslex.nlm.nih.gov/stmt
- 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

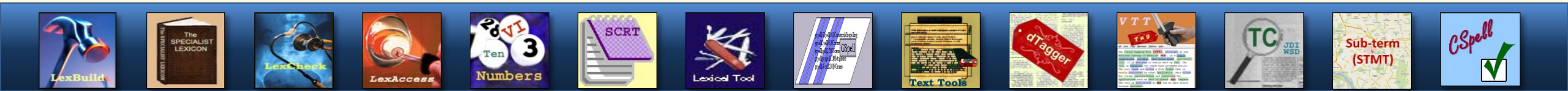
```
shell> smt –p -pt
-   Please input a term (type "Ctl-d" to quit)
tear break up time
tear break up time|break tear time up|C2111106|lacrimation tear break-up time|0
- Please input a term (type "Ctl-d" to quit) >
excimer laser photorefractive keratectomy
excimer laser photorefractive keratectomy|ceratectomy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



Multiword Generation Models

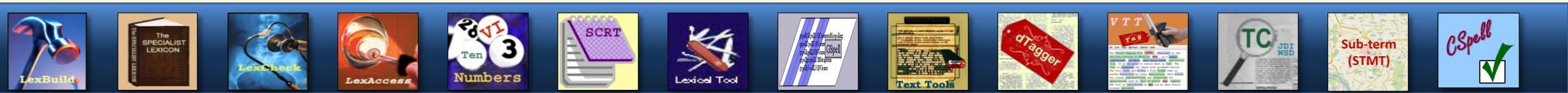


- A valid multiwords must meet 4 requirements:
  - POS
  - Inflection morphology
  - a specific meaning
  - Word order



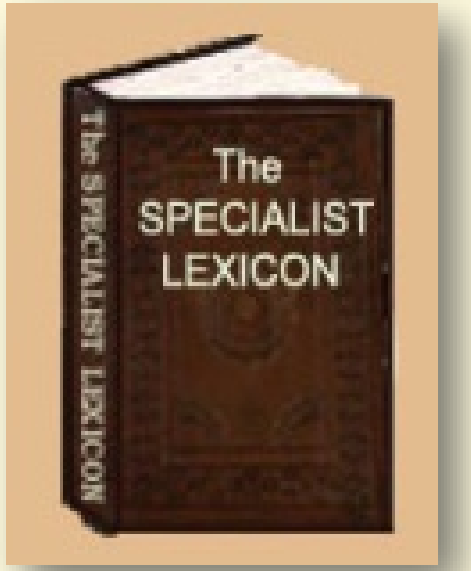
Multiwords Summary

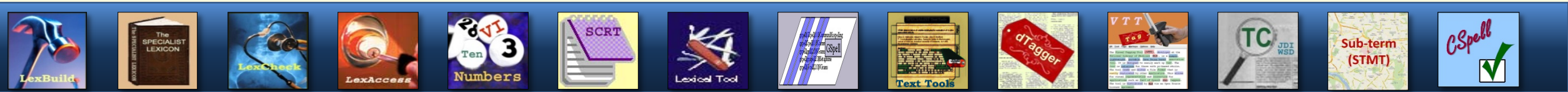
- Multiwords are pervasive, challenging and vital to NLP
- The SPECAILIST Lexicon include over 51.78% (513,960) multiwords
- Improve lexBuilding on multiwords (> 40% efficiency improvement)
- Distribute the MEDLINE n-gram set (2014+) to public
- Build a biomedical Lexicon to support NLP applications
- Multiword Approach (embedded information)



Lexicon - Derivations

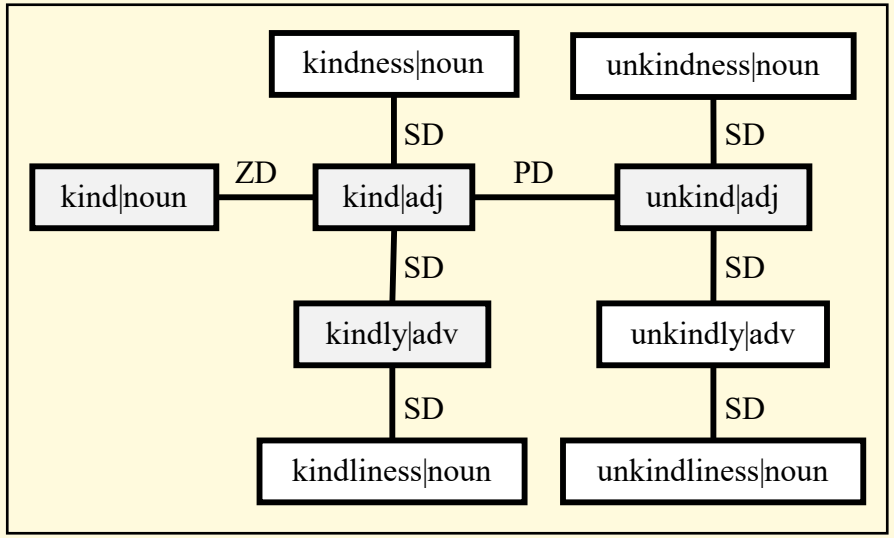
- "The Unified Medical Language System SPECIALIST Lexicon and Lexical Tools: Development and applications", Journal of the American Medical Informatics Association, 29 May 2020
- "Generating SD-Rules in the SPECISLIST Lexical Tools - Optimization for Suffix Derivation Rule Set", The 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016)
- "Implementing Comprehensive Derivational Features in Lexical Tools Using a Systematical Approach", AMIA 2013 Annual Symposium
- "A Systematic Approach for Automatically Generating Derivational Variants in Lexical Tools Based on the SPECIALIST Lexicon", IEEE IT Professional Magazine, May/June, 2012

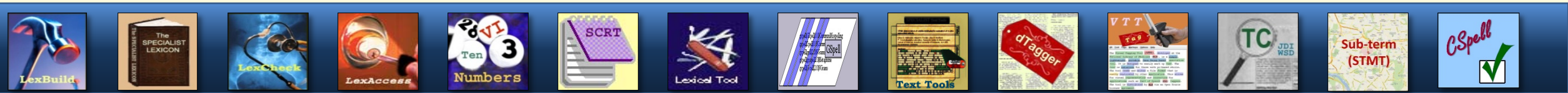




Lexicon Derivations

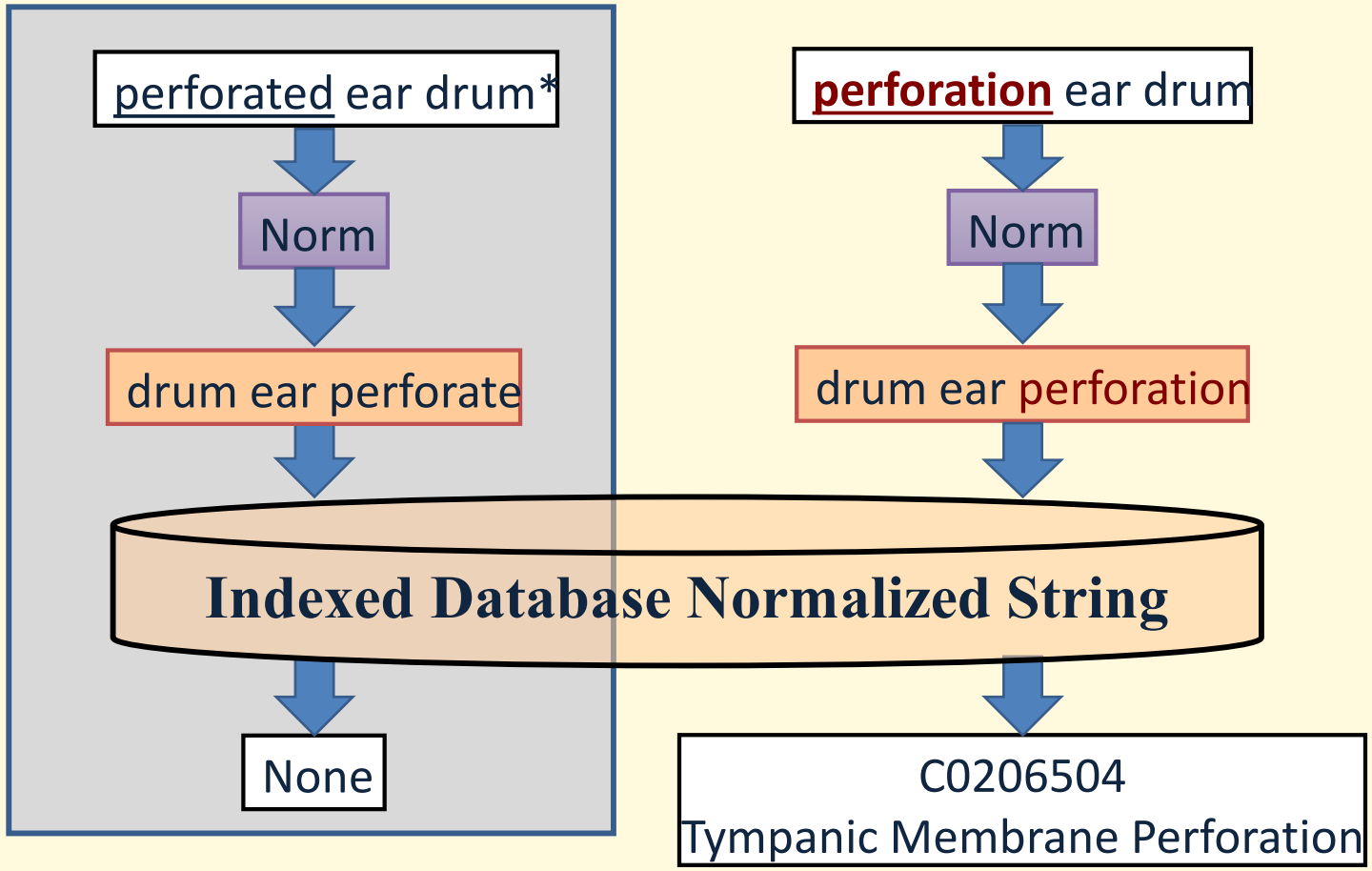
- Based on derivational morphology
- Arranged in pairs, called dPairs
- Types of derivation - example: transport
  - suffix - transport<u>ation</u>, transport<u>able</u>, transport<u>er</u>, …
  - prefix – <u>auto</u>transport, <u>intra</u>transport, <u>pre</u>transport, …
  - conversion (zero) - transport (verb), transport (noun)
- Derivation network:

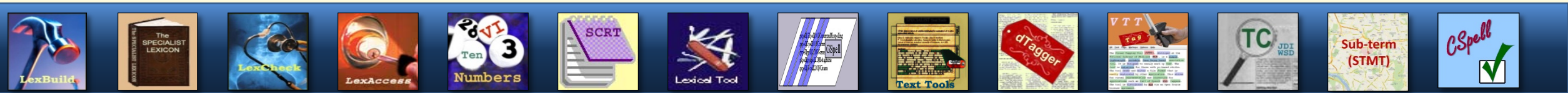




Derivation - Application Examples

- Used in Query Expansion for Concept Mapping





Derivation Generation Models

Lexicon

Automatic generation:
- zeroD
- prefixD (150 PD-Rules)
- suffixD (170 SD-Rules)

Raw dPairs

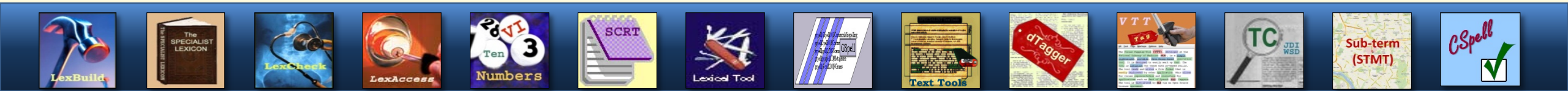
Automatic tagging

Heuristic Rules:
- Nominalization, etc.

Experts' tagging
- type
- negation
- etc.

Derivation table

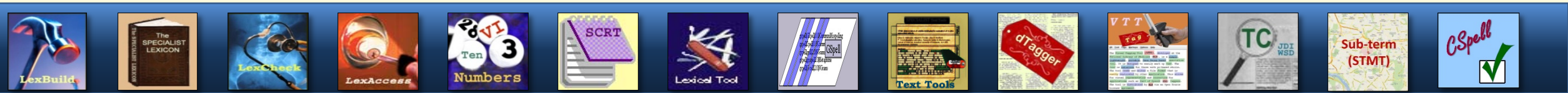
Example:
PD-Rule: anti-, dis-, etc.
SD-Rule:
- $|adj|ness$|noun
- $|verb|less$|adj



Derivation Summary

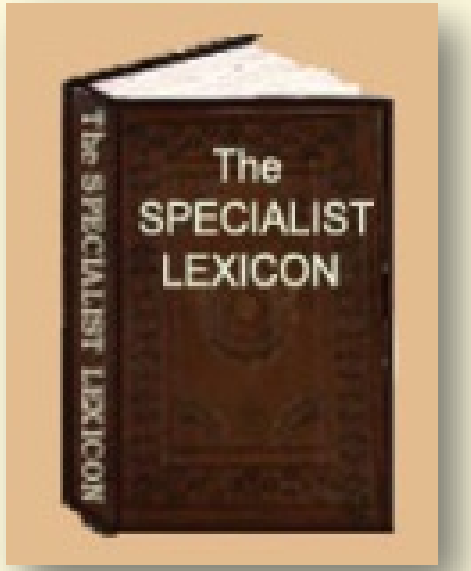
Year	Total	suffixD	prefixD	zeroD	Negation:[N|O]
2021	148,315	39.15%	50.15%	10.69%	15.26%|84.74%
2020	147,792	38.96%	50.32%	10.72%	15.31%|84.69%
2019	146,380	38.50%	50.68%	10.82%	15.44%|84.56%
2018	146,323	38.50%	50.68%	10.82%	15.43%|84.57%
2017	146,203	38.49%	50.58%	10.83%	15.43%|84.57%
2016	145,339	38.21%	50.92%	10.88%	15.49%|84.51%
2015	141,623	36.90%	51.98%	11.12%	15.85%|84.15%
2014	140,203	36.62%	52.19%	11.20%	15.97%|84.03%
2013	121,078	37.0%	50.6%	12.4%	17.29%|82.71%
2012	89,950	17%	66%	17%	N/A
2011-	4,559	92%	0.1%	7.9%	N/A

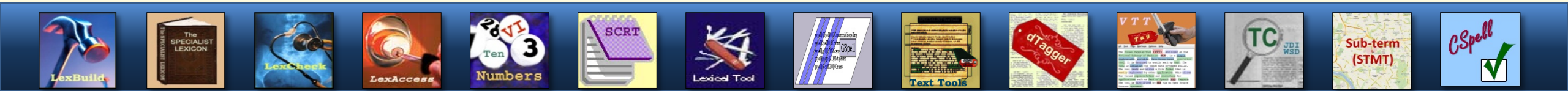
- Stable growth over 32 times since this systematic approach was implemented in 2012
- New derivational features, such as negation, derivation types, and recursive derivations with options, are implemented in the Lexical Tools for better performance in NLP applications.



Lexicon - Synonyms

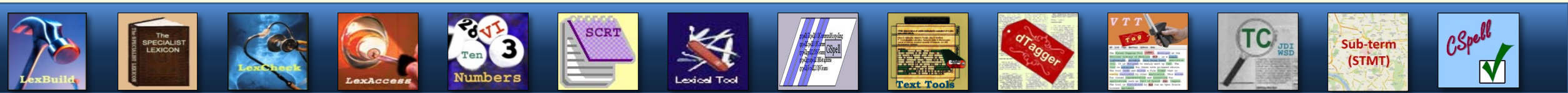
- "The Unified Medical Language System SPECIALIST Lexicon and Lexical Tools: Development and applications", Journal of the American Medical Informatics Association, 29 May 2020
- "Enhancing LexSynonym Features in the Lexical Tools", AMIA 2017 Annual Symposium
- "Enhanced LexSynonym Acquisition for Effective UMLS Concept Mapping", MedInfo 2017: Precision Healthcare through Informatics





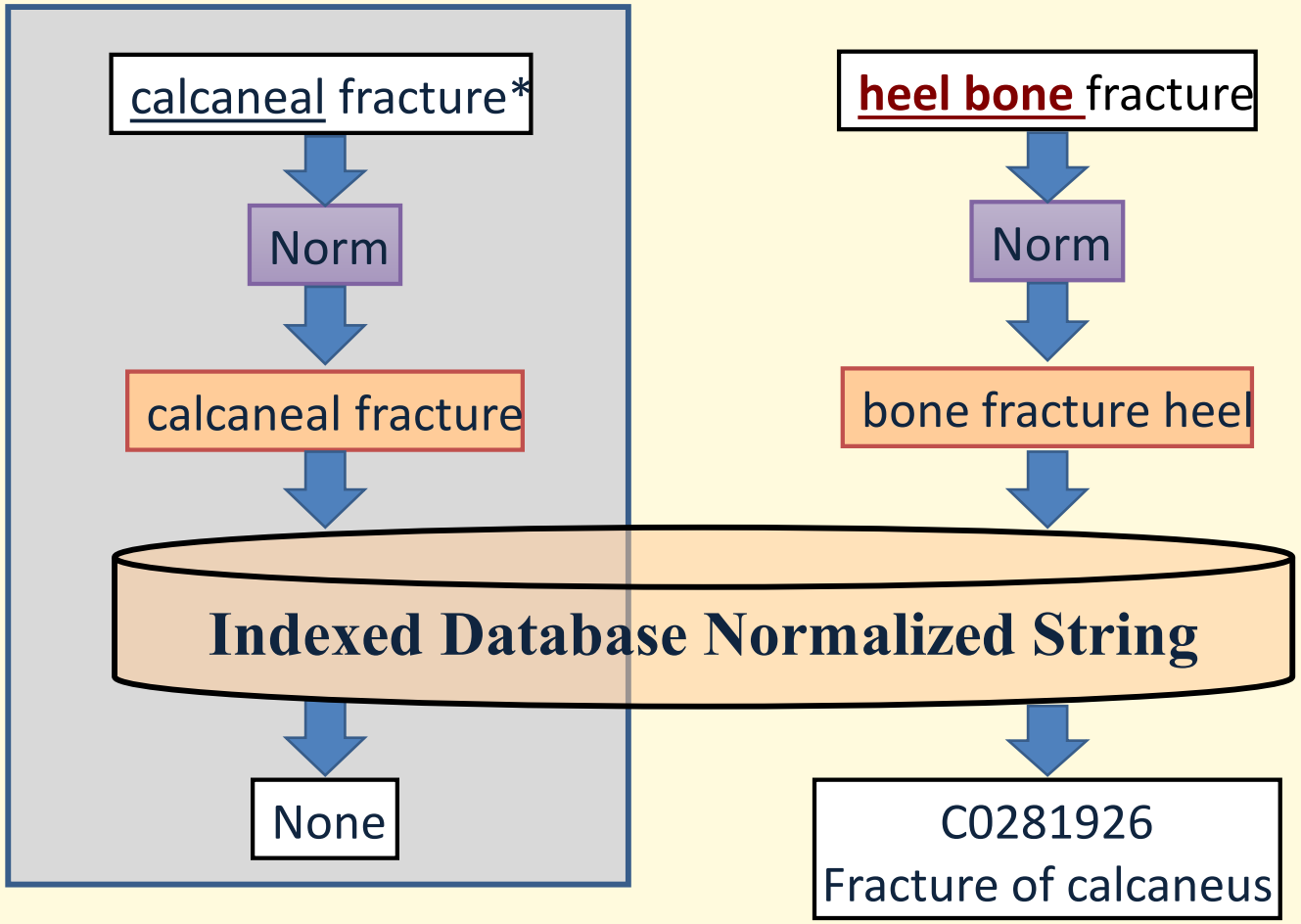
Lexicon Synonyms

- 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”

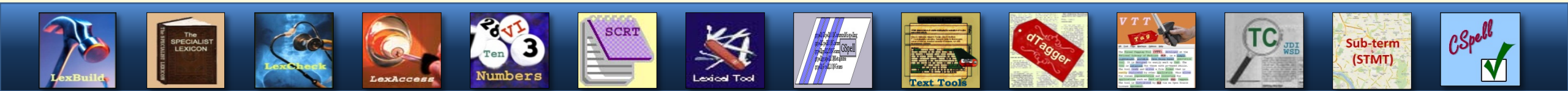


Synonym Application Examples

- **Used in Query Expansion for Concept Mapping**



C0006655:
- calcaneal
- heel bone



Synonym Generation Models

Candidate sClasses → **Manual Tagging** → **Synonym Generation**

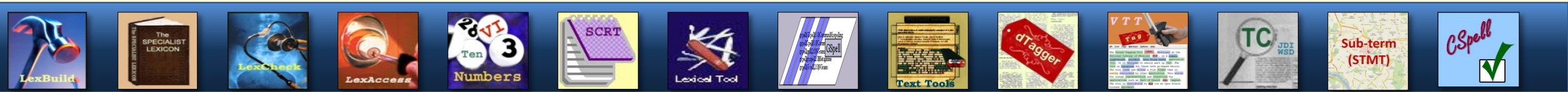
UMLS sClasses
- MRCONSO.RRF
- English terms with same CUI

Refined sClasses
(Heuristic Rules)

- Must be a base form in the Lexicon
- POS: noun, verb, adjective
- Remove chemicals and drugs (STI)
- Remove acronyms or abbreviations
- Add EUI and CUI
- Remove spelling variants
- Remove nominalization

Tagged by linguists
- Ensure cognitive synonyms
  - Commutativity:
    (x = y) => (y = x)
  - Transitivity:
    ((x = y) and (y = z)) -> (x = z)

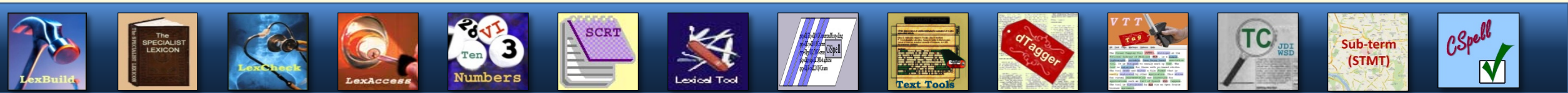
sPairs Generating
- Source: EUI and CUI
- Add spelling variants
- Add nominalization



Synonym Summary

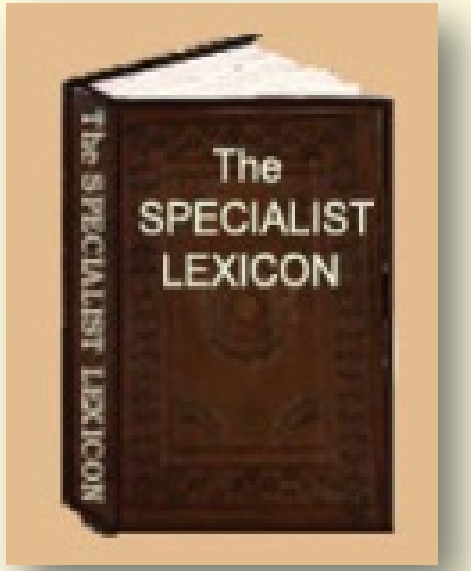
Year	Total	Source		
		CUI	EUI	NLP_LVG
2021	233,994	161,690 (69%)	67,532 (29%)	4,772 (2%)
2020	227,692	155,488 (68%)	67,432 (30%)	4,772 (2%)
2019	213,164	140,726 (66%)	67,664 (32%)	4,774 (2%)
2018	197,116	124,508 (63%)	67,830 (34%)	4,778 (3%)
2017	190,844	118,468 (62%)	67,584 (35%)	4,792 (3%)
2016-	5,198	0 (0%)	0 (0%)	5,198 (100%)

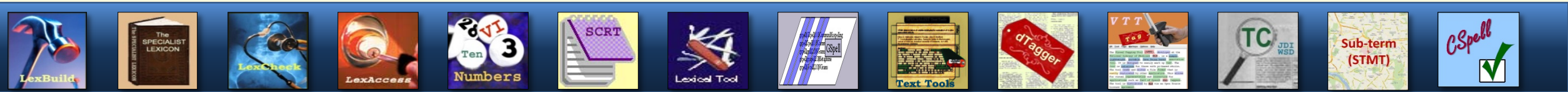
- Stable growth over 45 times since this systematic approach was implemented in 2017
- New synonym features, such as POS, source information, and recursive synonyms with source options, were implemented in the Lexical Tools to provide better performance for downstream NLP applications.



Lexicon - Antonyms

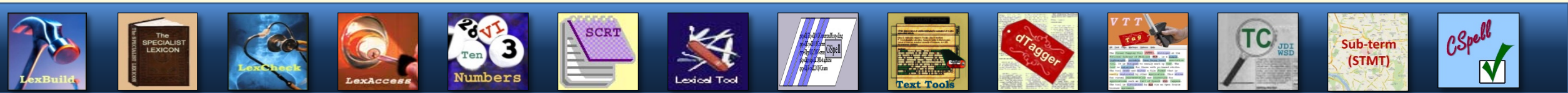
- "The Unified Medical Language System SPECIALIST Lexicon and Lexical Tools: Development and applications", Journal of the American Medical Informatics Association, 29 May 2020
- "Enhanced Features in the SPECIALIST Lexicon - Antonyms", AMIA 2020 Annual Symposium





Lexicon Antonyms

- Antonyms are words that have opposite or contrasting meanings in a specific domain.
- Lexicon Antonyms:
  - arranged in pairs, called aPairs
  - must be single words in the Lexicon with the same part-of-speech
  - must be canonical antonyms
    - members of a canonical antonym pair express opposite properties on generic domains, that are central to human life and ways of living across times and cultures, such as color, temperature, length, etc.
    - Examples: “black-white-color” and “dark-white-chocolate”



Antonym - Application Examples

- Example: In acute suppuration of the knee, excision is never successful.

  ↓

  unsuccessful
  C1272705

To capture the correct meaning:

- Negated antonyms can be used to substitute synonymous antonyms for better recall (“never successful” = “unsuccessful”).
- Negation detection cue words can be retrieved from negative antonyms with strict negation (N), such as “unsuccessful”, “useless”, “never”, “not”, “without”, etc.

Antonym Generation Models

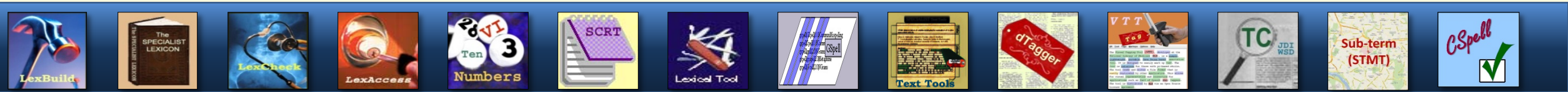
Antonym Candidates → **Manual Tagging** → **Antonym Generation**

Sources
- Lexical records with negative tags
- Suffix derivations with negation
- Prefix derivations with negation
- Collocates in a corpus

Heuristic Rules
- Lexicon single words
- With the same POS
- Co-occurring
- Not synonyms
- Pattern Matches

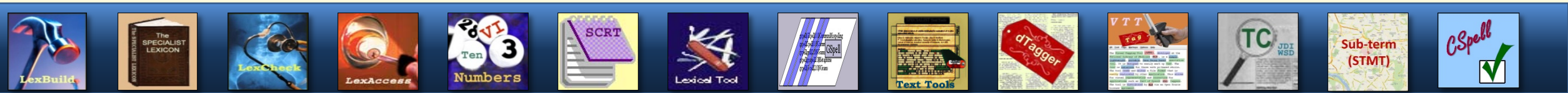
Tagged by linguists
- Ensure canonical antonyms
  - In a generic domain

- Negation:
  Negation Detection Cue Words
- POS
- Type: B|UB|AB
- Domain: quality, length, etc.
- Sources: PD, SD, LEX, CC

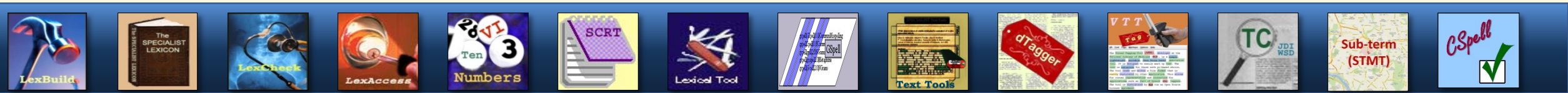


Antonym Summary

- Provide generic and comprehensive antonyms for NLP applications
  - canonical antonyms
  - types
  - negations
  - domains
  - sources
- Plan to include antonyms in the future release of the Lexicon
- Plan to integrate antonym features in the future release of Lexical Tools



Questions

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

☺