The SPECIALIST Lexicon and NLP Tools (Spell Checker for Consumer Language - CSpell)

By: Dr. Chris J. Lu

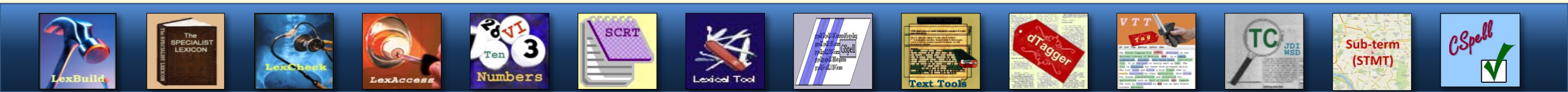
NLM – LHNCBC - CGSB

June, 2019

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

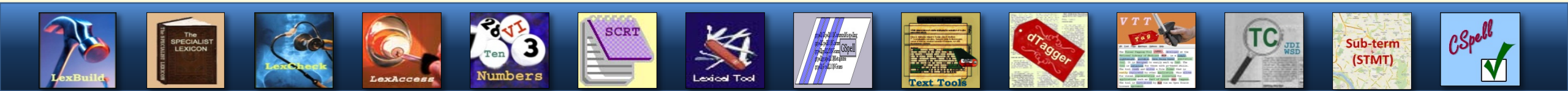
Disclaimer

- 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.



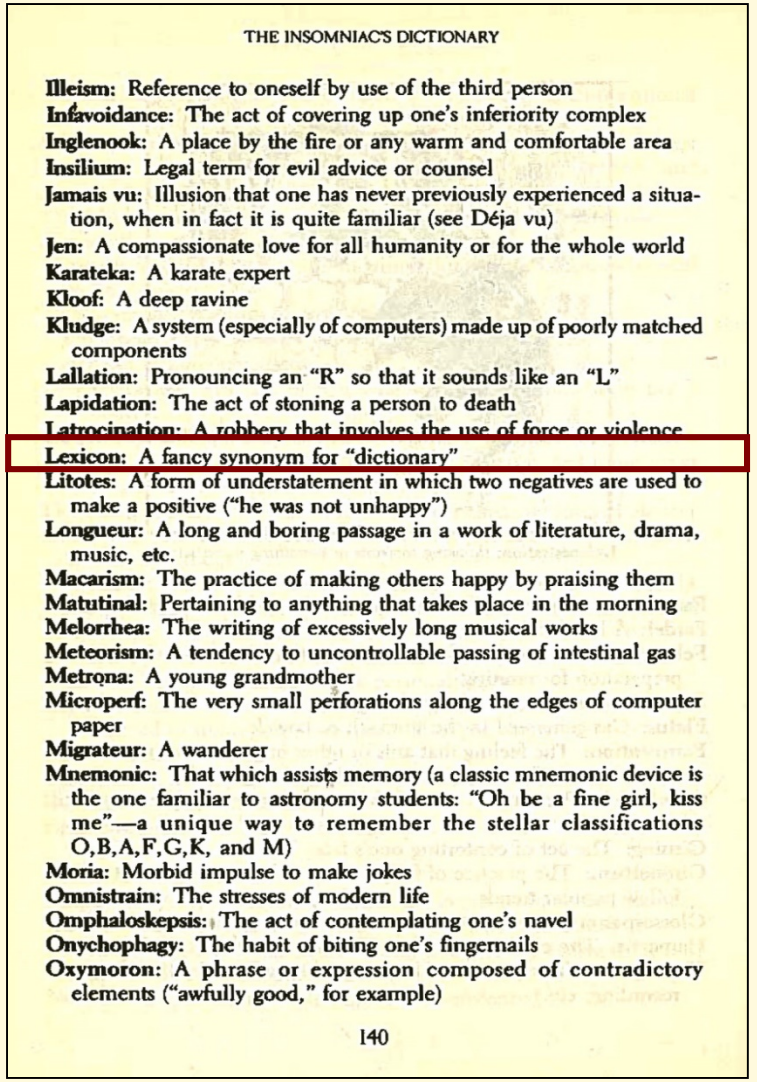
Outline

- Introduction
  - The SPECIALIST Lexicon
  - The SPECIALIST NLP Tools (Lexical Tools)
- Applications – Concept mapping & CSpell
  - Natural Language Processing (NLP)
  - CSpell (Spell Checker for Consumer Langauage)
- Questions (anytime)



1. The SPECIALIST Lexicon

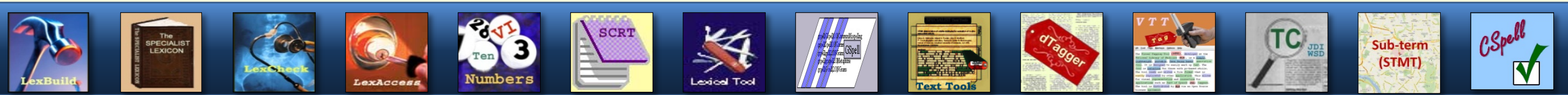
- A fancy synonym for “dictionary”
- A syntactic lexicon
- Biomedical and general English
- Over 0.5M records, 1M words (POS + forms)
- Designed/developed to provide the lexical information needed for the NLP (Natural Language Processing) system
- Distributed in the Unified Medical Language System (UMLS) Knowledge Sources by the National Library of Medicine (NLM)



THE INSOMNIAC'S DICTIONARY

Illeism: Reference to oneself by use of the third person
Infavoidance: The act of covering up one's inferiority complex
Inglenook: A place by the fire or any warm and comfortable area
Insilium: Legal term for evil advice or counsel
Jamais vu: Illusion that one has never previously experienced a situation, when in fact it is quite familiar (see Déja vu)
Jen: A compassionate love for all humanity or for the whole world
Karateka: A karate expert
Kloof: A deep ravine
Kludge: A system (especially of computers) made up of poorly matched components
Lallation: Pronouncing an "R" so that it sounds like an "L"
Lapidation: The act of stoning a person to death
Latrocination: A robbery that involves the use of force or violence
Lexicon: A fancy synonym for "dictionary"
Litotes: A form of understatement in which two negatives are used to make a positive ("he was not unhappy")
Longueur: A long and boring passage in a work of literature, drama, music, etc.
Macarism: The practice of making others happy by praising them
Matutinal: Pertaining to anything that takes place in the morning
Melorrhea: The writing of excessively long musical works
Meteorism: A tendency to uncontrollable passing of intestinal gas
Metrona: A young grandmother
Microperf: The very small perforations along the edges of computer paper
Migrateur: A wanderer
Mnemonic: That which assists memory (a classic mnemonic device is the one familiar to astronomy students: "Oh be a fine girl, kiss me"—a unique way to remember the stellar classifications O,B,A,F,G,K, and M)
Moria: Morbid impulse to make jokes
Omnistrain: The stresses of modern life
Omphaloskepsis: The act of contemplating one's navel
Onychophagy: The habit of biting one's fingernails
Oxymoron: A phrase or expression composed of contradictory elements ("awfully good," for example)

140

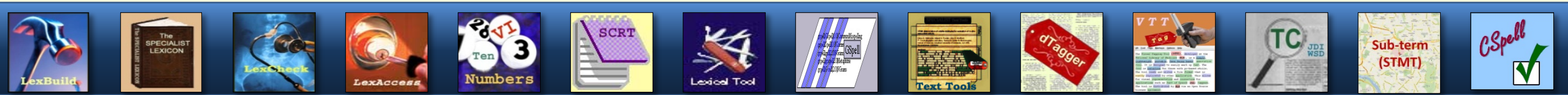


LexBuild Process (Computer-Aided)

Sources:
- Word candidates from MEDLINE
- 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**

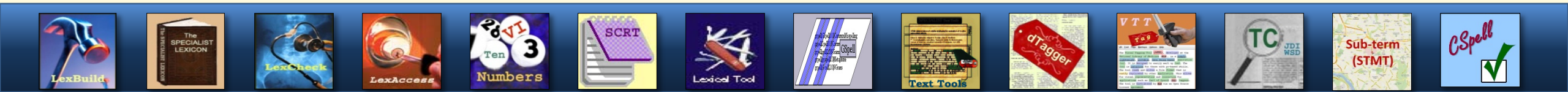


Team of Lexicon Builders

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

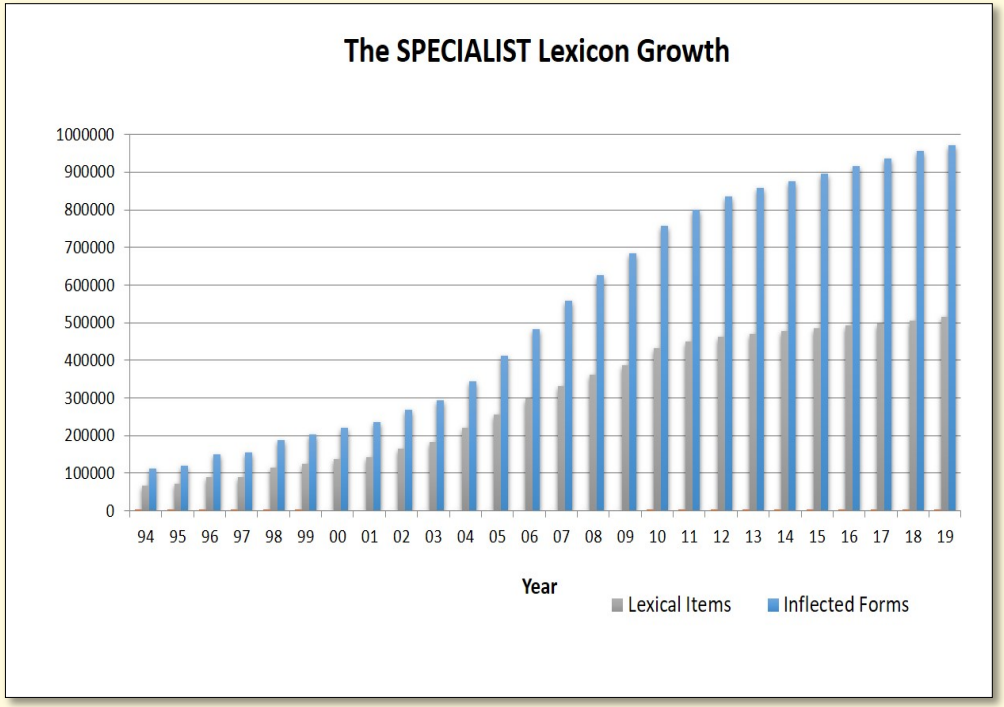
- Dr. Dina Demner Fushman (Sr. PI)
- Dr. Chris J. Lu
- Dr. Amanda Payne

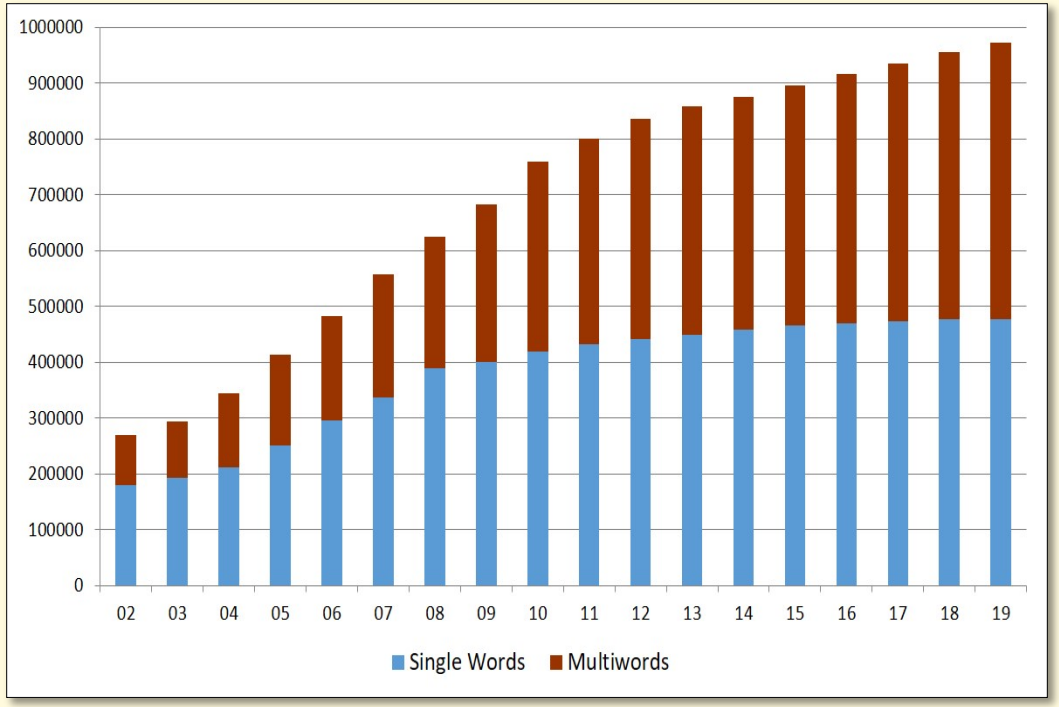
☺



Lexicon Growth – 1994 to 2019

- 515,012 lexical records
- 1,154,224 words (categories and inflections)
- 972,721 forms (spelling only)
  - Single words: 477,618 (49.10%); Multiwords: 495,103 (**50.90%**)

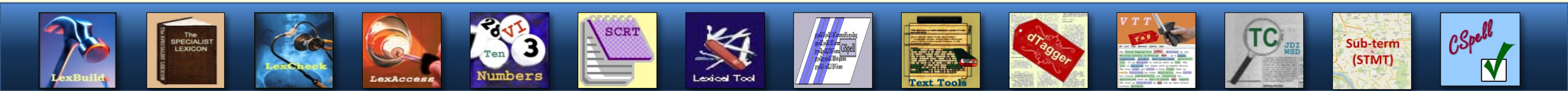




What Is a Word?

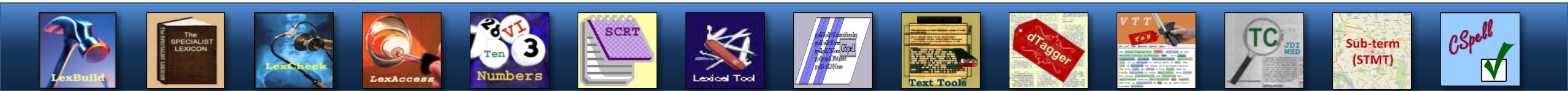
- Orthographic words:
  - spelling
  - use space as word boundary

- How about:
  - dog vs. dogs?
  - color vs. colour?
  - ice-cream vs. ice cream?
  - see vs. saw? (verb)
  - saw vs. saws (noun)?
  - see vs. saws ??



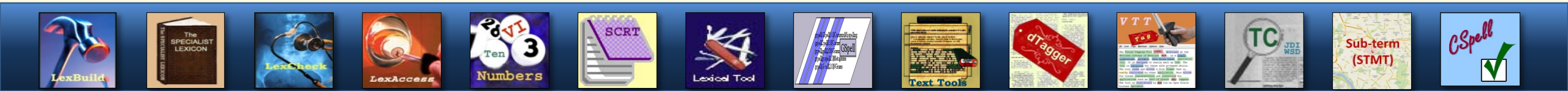
(Multi)Words for Lexical Records

- 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, frog erythrocytic virus, cardiac surgery
  - Inflection morphology (uninflection):
    - left pulmonary veins (“left pulmonary vein” and “leave pulmonary vein”)
  - Specific meaning:
    - hot dog (high temperature canine?)
  - Word order:
    - trial and error, up and down (vs. food and water)
    - exercise training vs. training exercise (military)

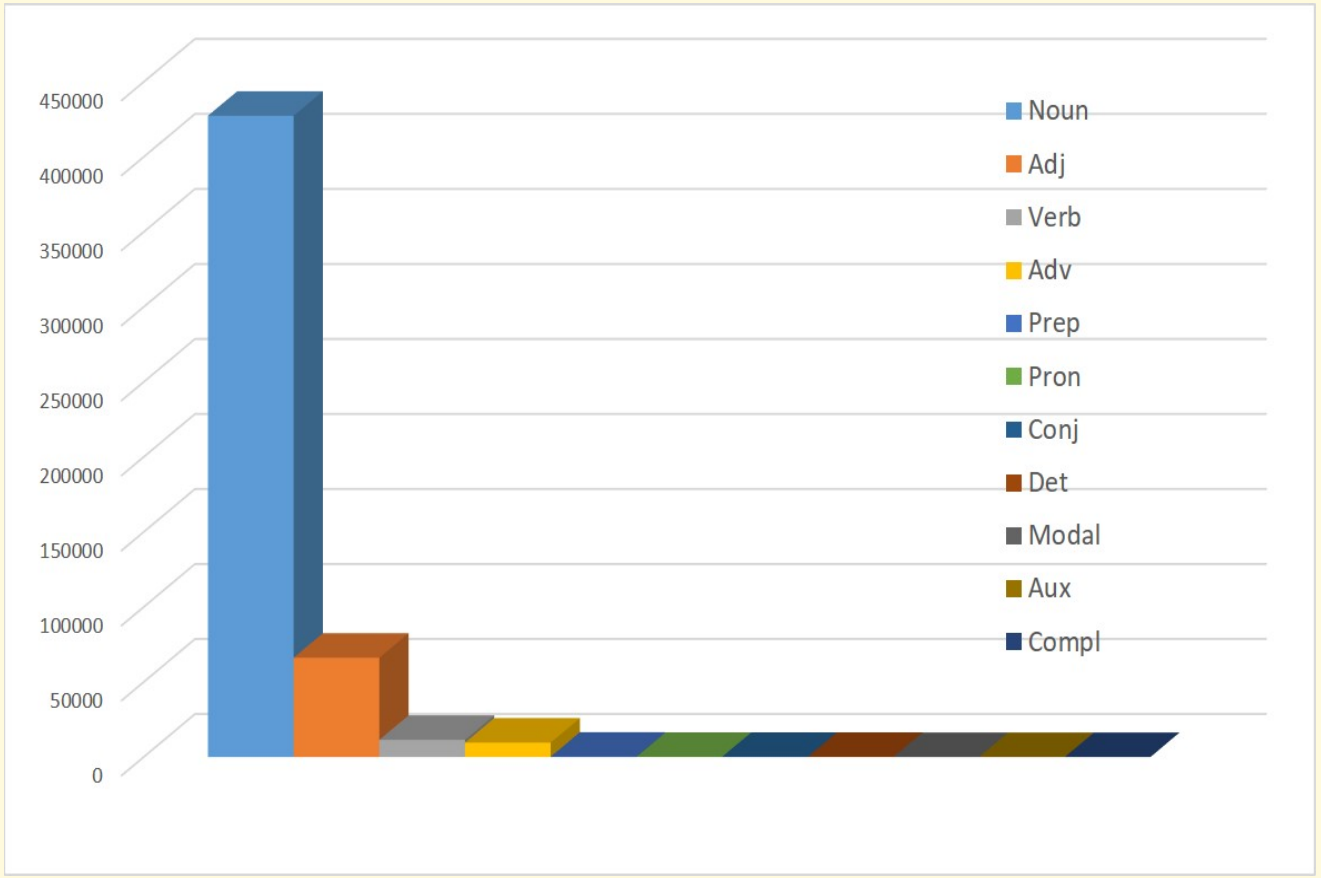


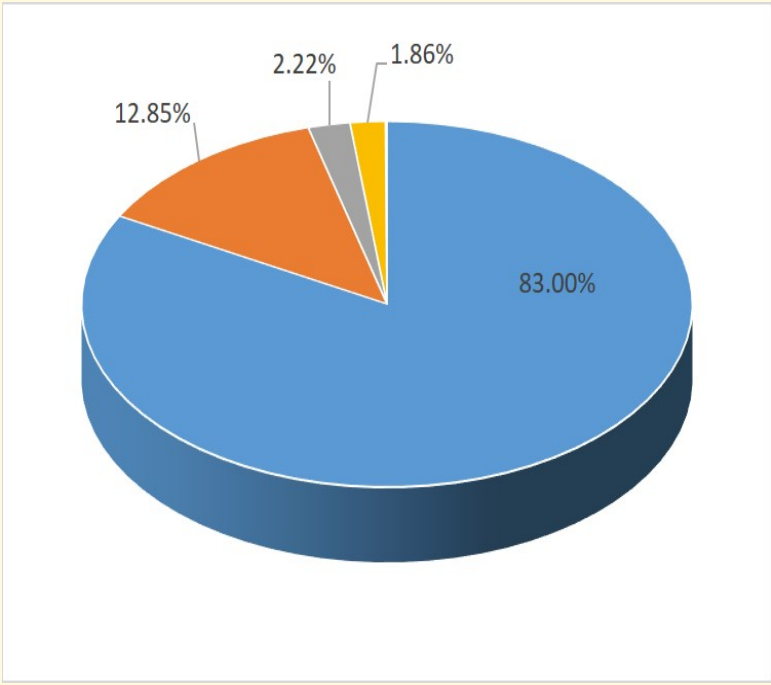
Lexical Records - Information

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



Categories – Parts of Speech (11)





Lexicon.2019

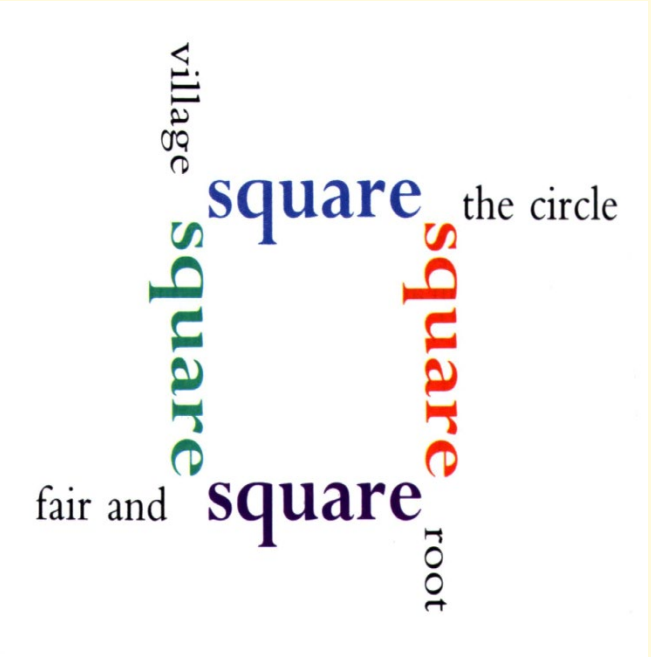
Lexical Records & POS

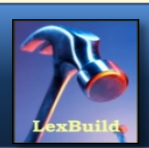
```
{base=square
entry=E0057517
        cat=verb
        variants=reg
        intran
        intran:part(up)
}
```

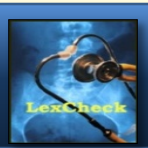
```
{base=square
entry=E0057516
        cat=adj
        variants=reg
        variants=inv
}
```

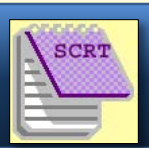
```
{base=square
entry=E0057518
        cat=adv
}
```

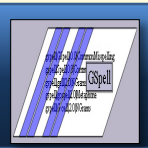
```
{base=square
entry=E0057515
        cat=noun
        variants=reg
}
```

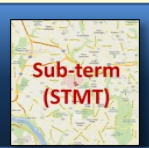






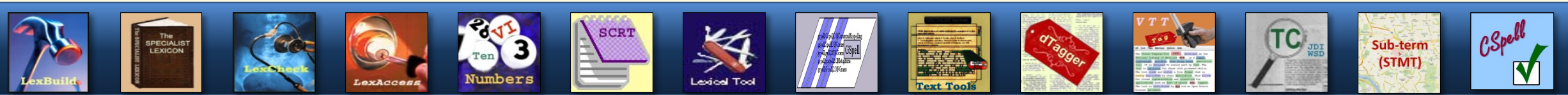






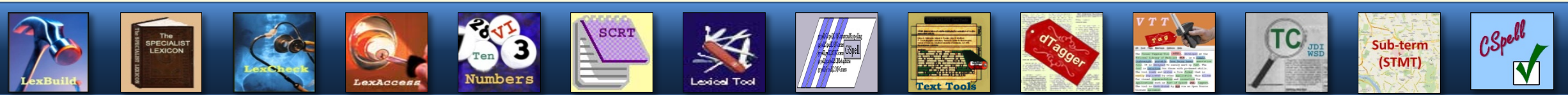
Morphology

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



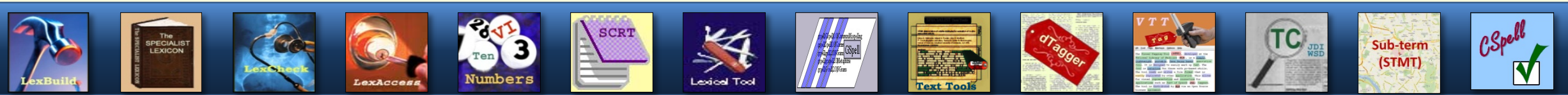
Orthography (Spelling Variation)

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



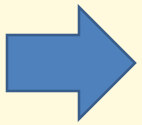
Syntax - Verb Complements

- 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
}
```

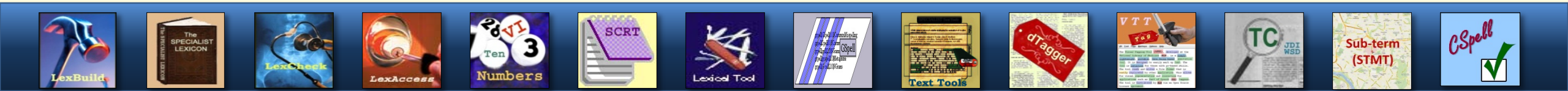
```
{base=deja vu
spelling_variant=deja-vu
spelling_variant=déjà vu
entry=E0021340
        cat=noun
        variants=uncount
}
```

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

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

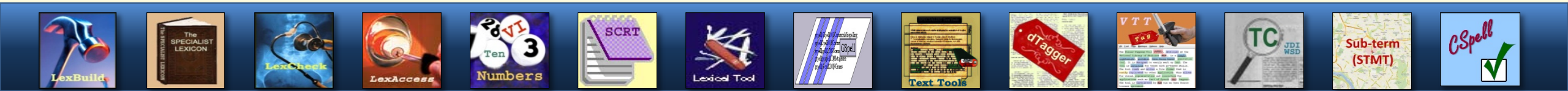
```
{base=cafe
spelling_variant=café
entry=E0420690
        cat=noun
        variants=reg
}
```

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



Classification Type (2021+)

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

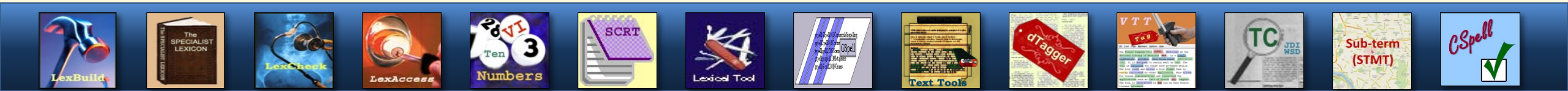


Lexicon Unigram Coverage – Without WC

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

Types	Word Count	Percentage %	Accu. %
LEXICON (S)	296,747	8.1978%	8.1978%
NUMBER	62	0.0017%	8.1995%
DIGIT	87,437	2.4155%	**10.6150%**
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.

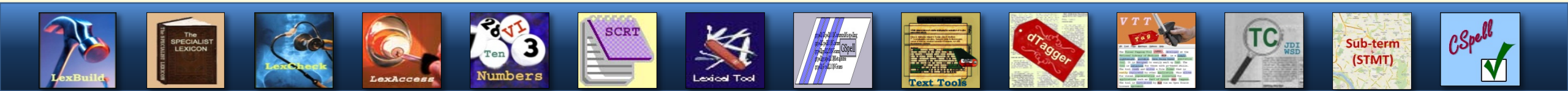


Lexicon Unigram Coverage – With Frequency (WC)

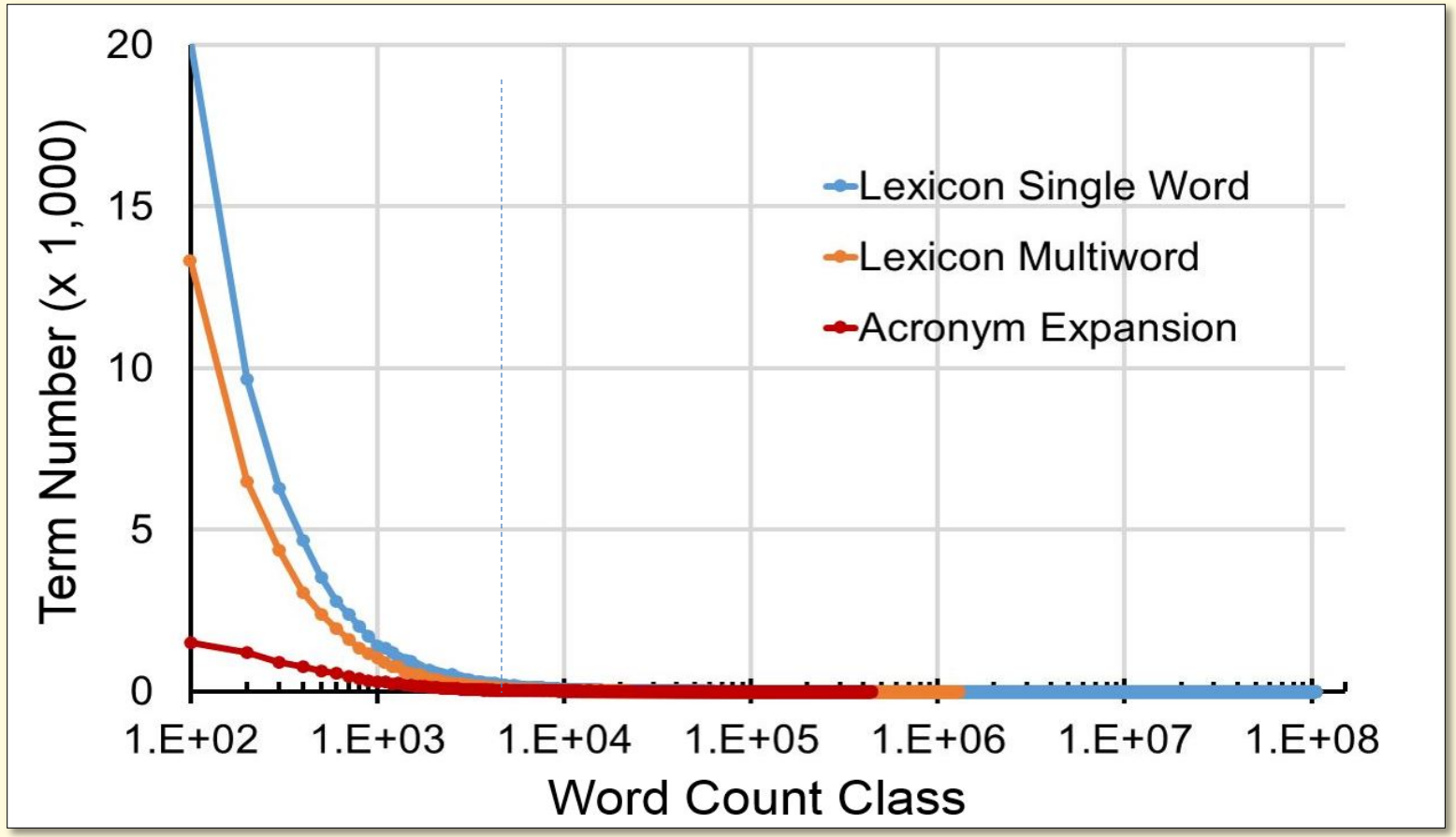
- Total word count for MEDLINE (2016): 3,114,617,940
- Lexicon covers > 98-99% unigrams from MEDLINE

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		

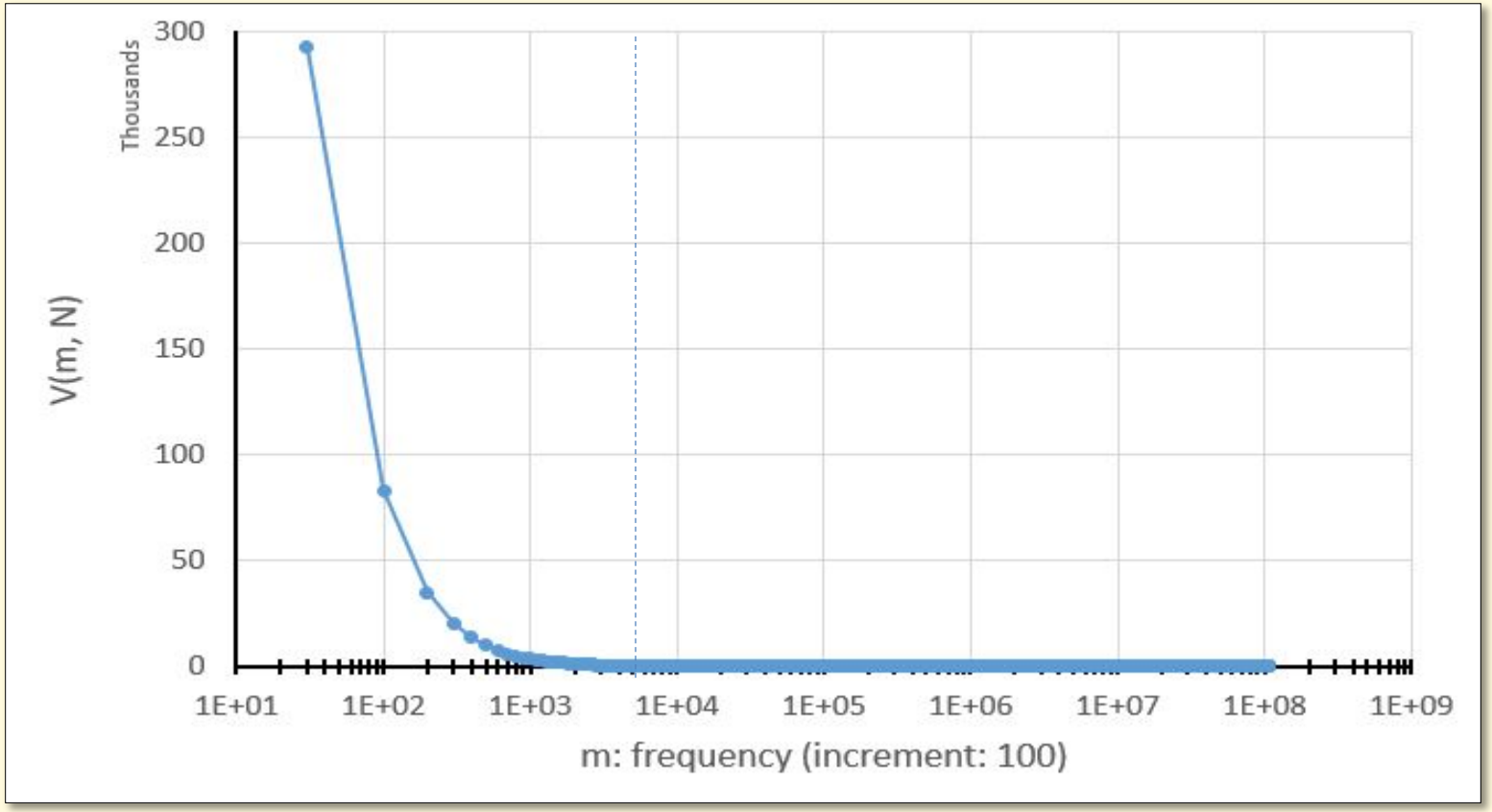
* NW-EW: an element word only exist in multiword, such as “non”, “vitro”, “vivo”, “intra”, etc.



The Frequency Spectrum of Lexicon (Multi)words on MEDLINE

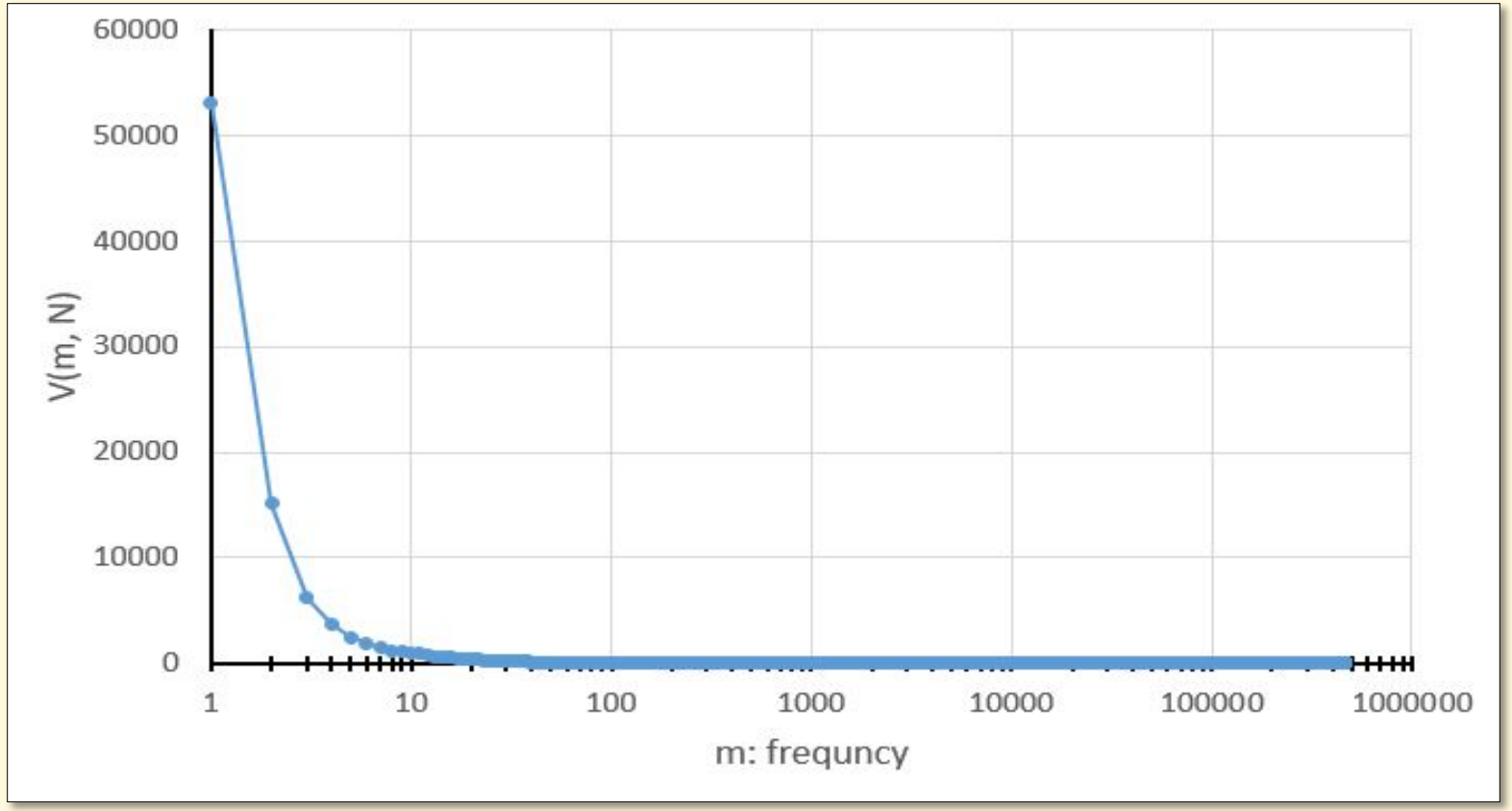


The Frequency Spectrum of MEDLINE (2019)



- 3,824,268,997 unigrams from MNS.2019, min. WC = 30

The Frequency Spectrum of Consumer Health Corpus



- 10,197,915 words

The Frequency Spectrum of Alice in Wonderland

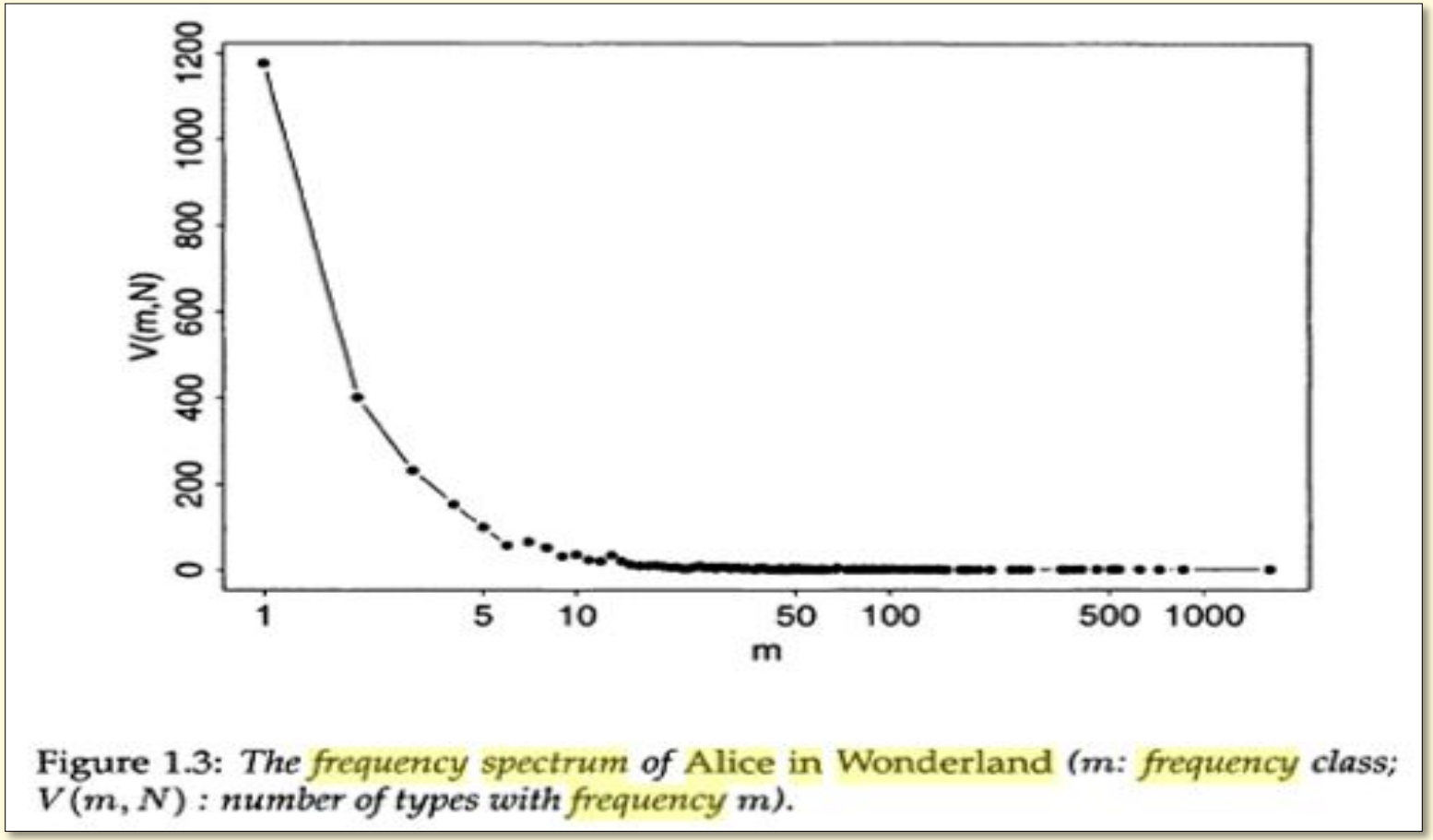


Figure 1.3: *The frequency spectrum of* Alice in Wonderland *(m: frequency class; V (m, N) : number of types with frequency m).*

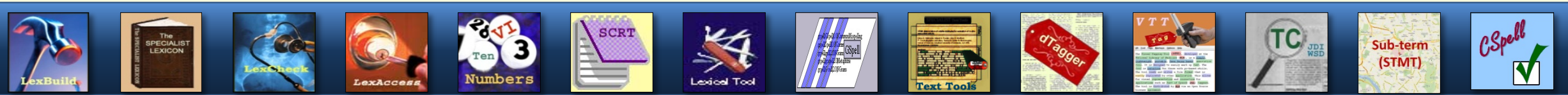
- 26,432 words

Lexicon (Data) and Lexical Tools (Software)

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
}
```

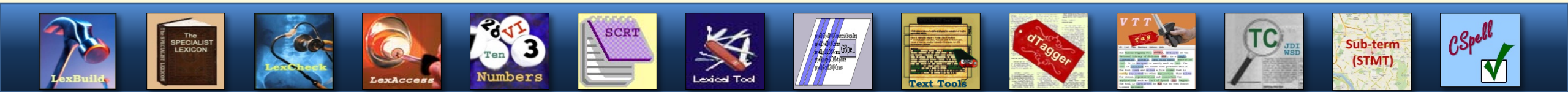
- spelling_variant → spelling variant
- cat=verb → part of speech
- variants=reg → inflectional variant
- tran=pphr(from,np) → chunker
- nominalization=generalisation|noun|E0029525 → derivational variant, synonym



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



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

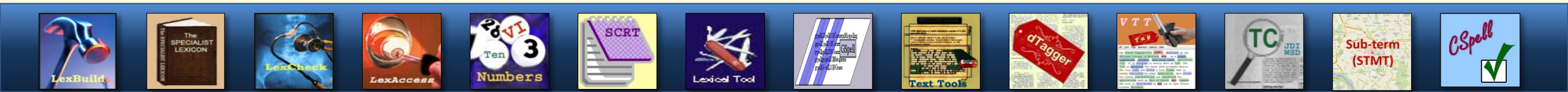
Derivational variants:
- suffixD: generalisation, generalization, generalisable
- prefixD: overgeneralise, over-generalise

← A LexRecord + Rules

Synonyms: generalize

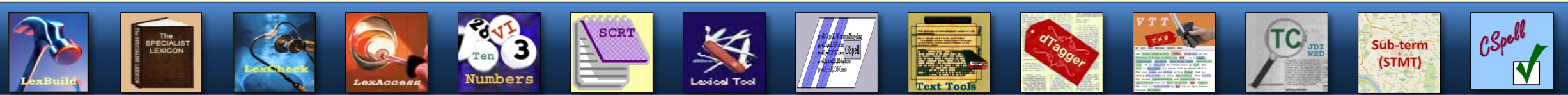
← Multiple LexRecords + Rules

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



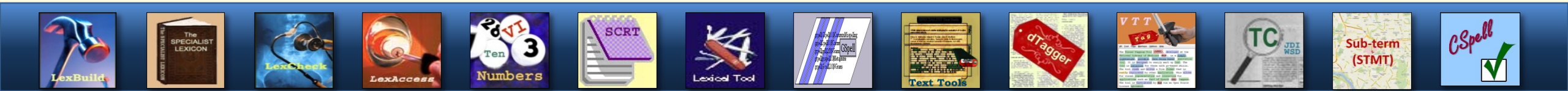
Lexical Tools - Facts

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



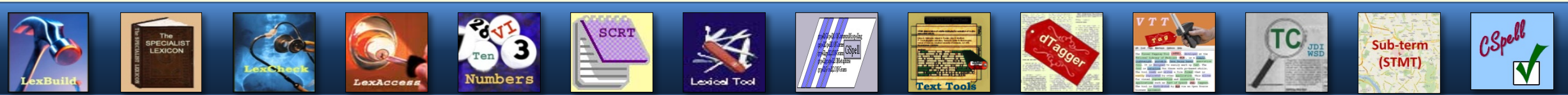
LVG - Lexical Variants Generation

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

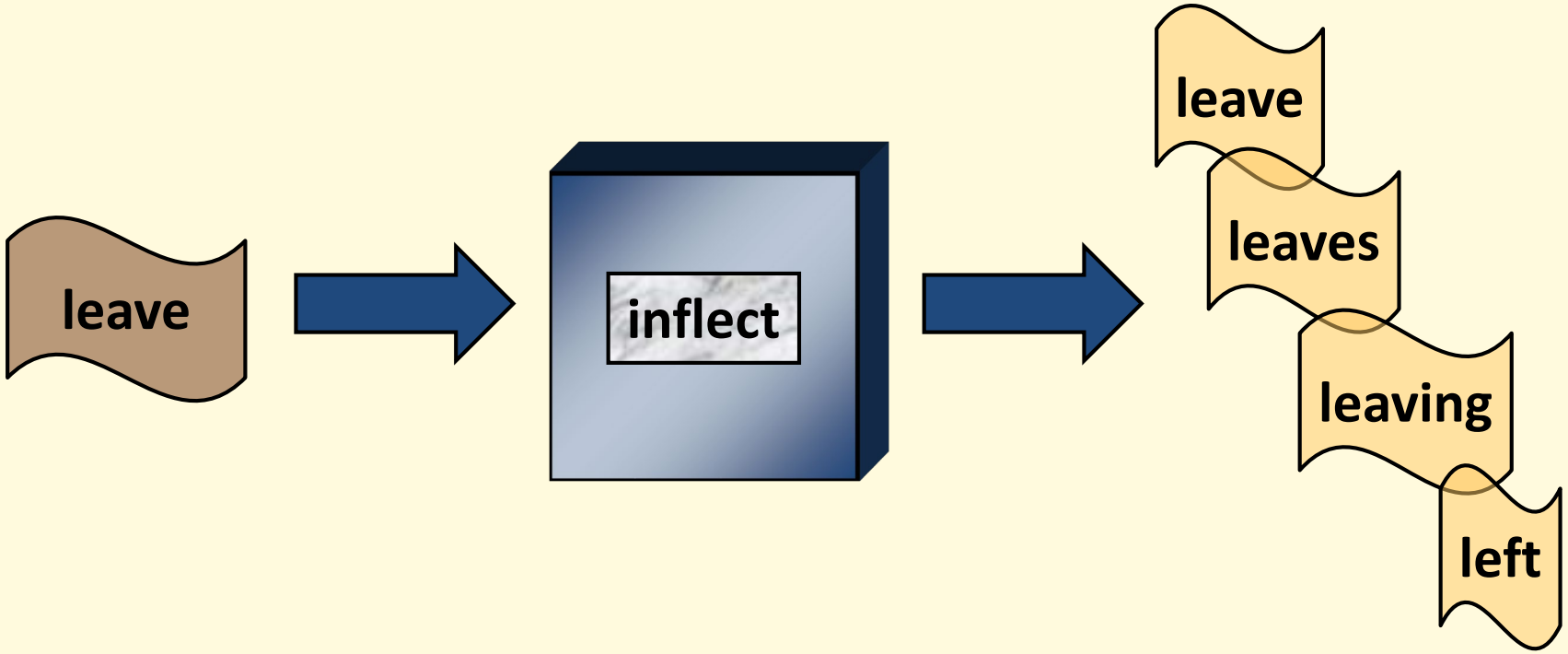


Lexical Tools – Flow Components (62)

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

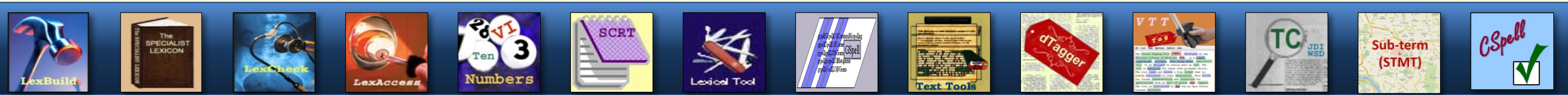


LVG Flow Component – Example



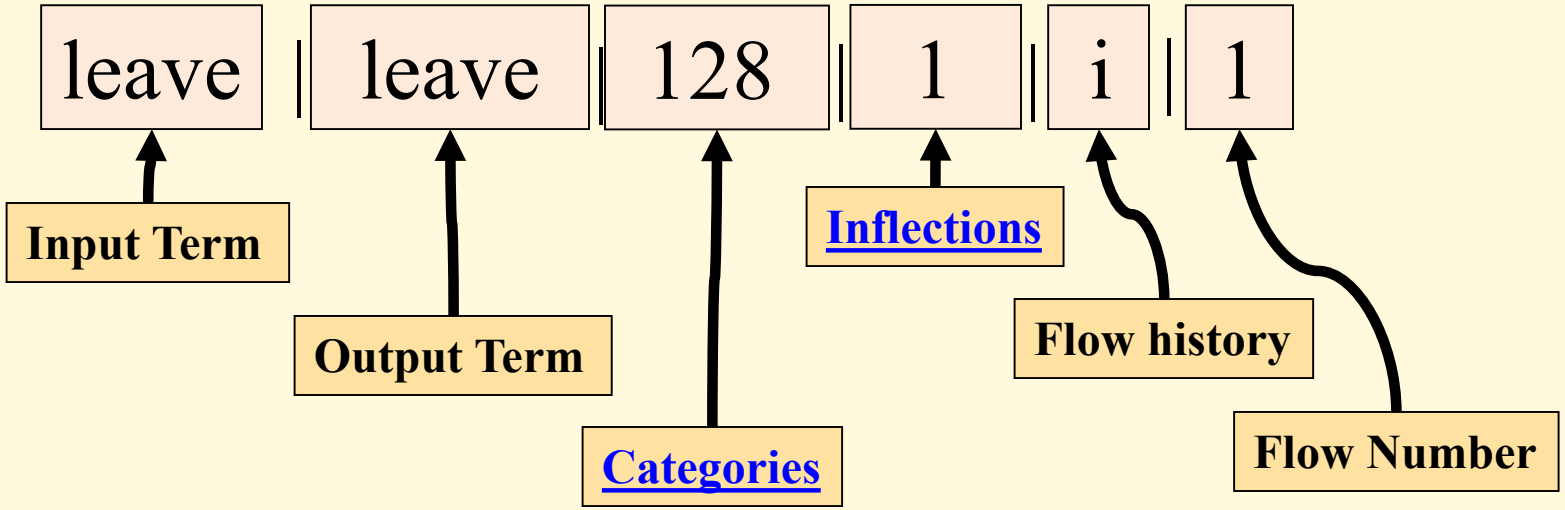
LVG Flow Component – CmdLine

```
> 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|
```

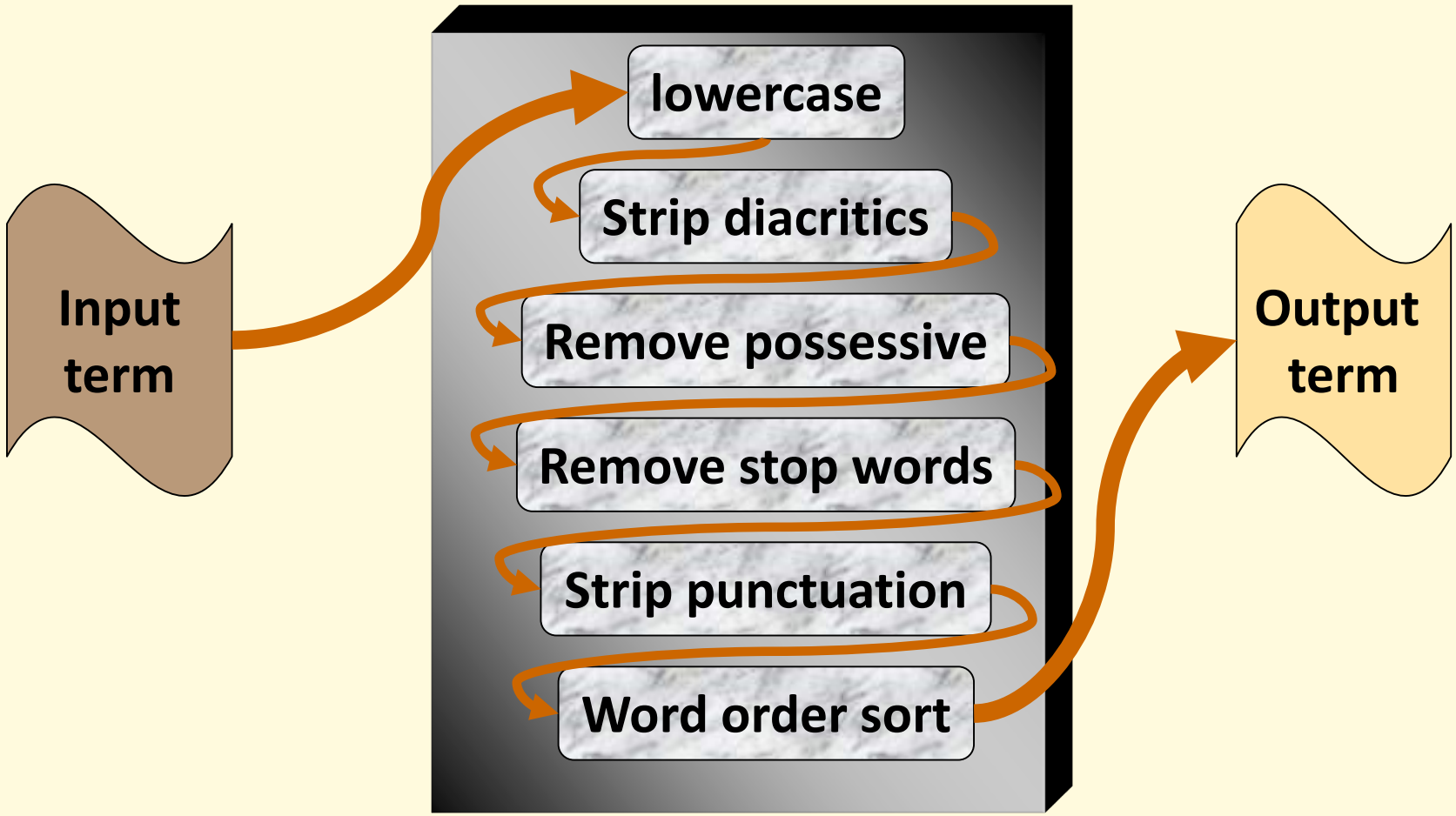


LVG Flow Component – Fielded Output

> lvg –f:i
leave



LVG – A Serial Flow

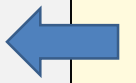


- Flow components can be arranged so that the output of one is the input to another.

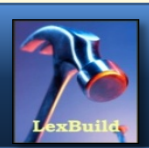
A Serial Flow - Example

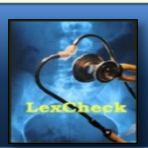
```
➢ lvg -f:l:q:g:t:p:w

The Gougerot-Sjögren's Syndrome
The Gougerot-Sjögren's Syndrome|
    gougerotsjogren syndrome|
    2047|16777215|l+q+g+t+p+w|1|
```

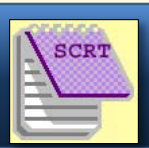

Input

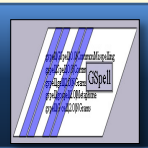

Output

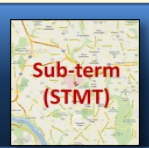

Other information

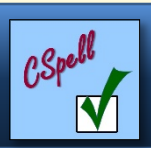




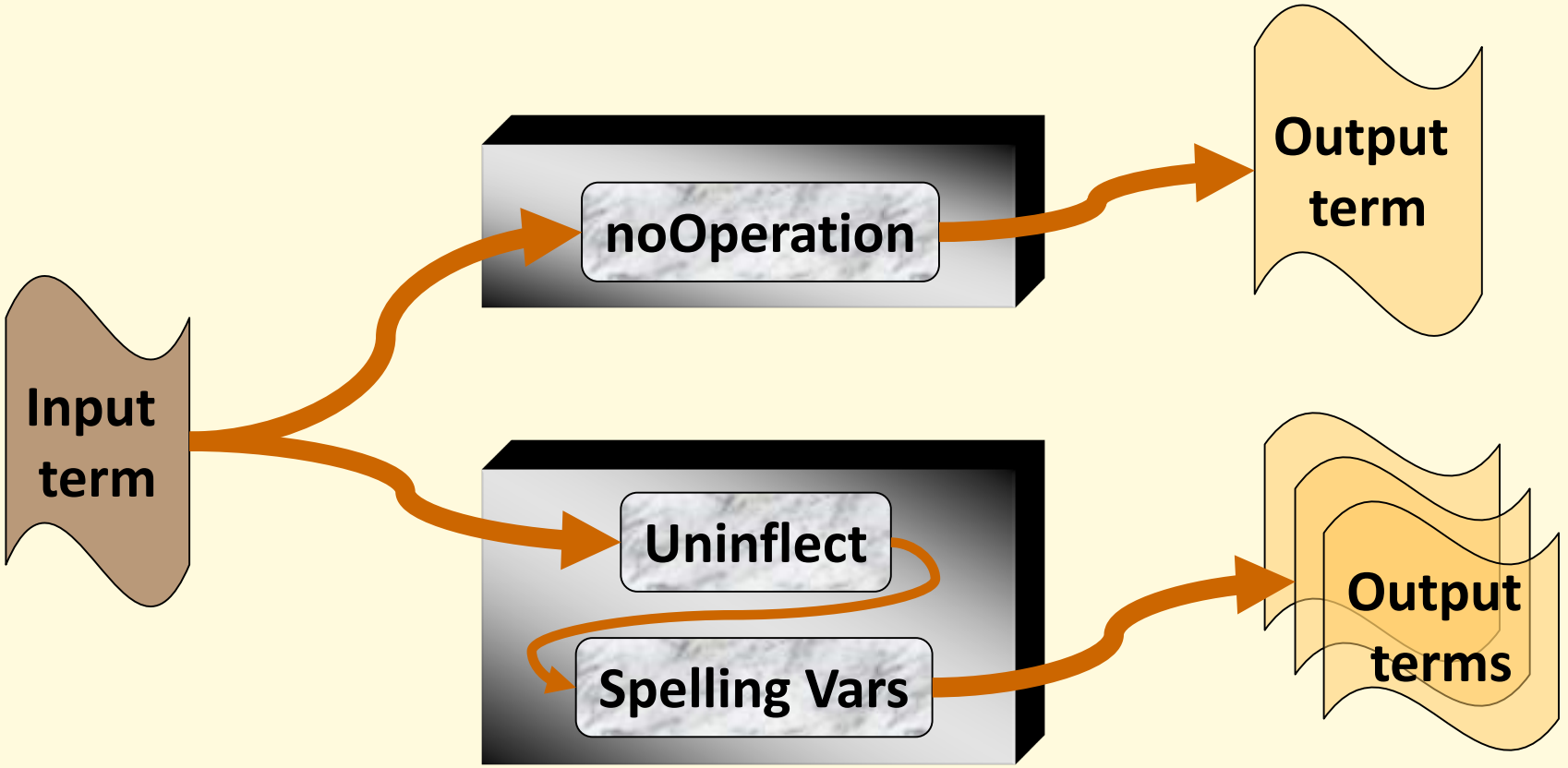








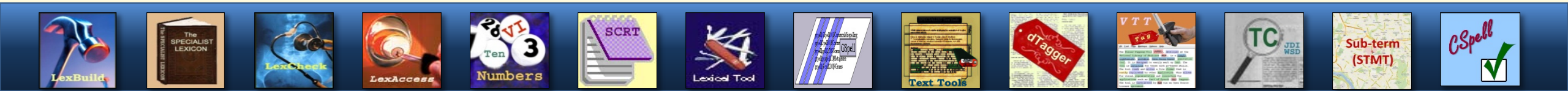
LVG - Parallel Flows



- Multiple flows can be defined

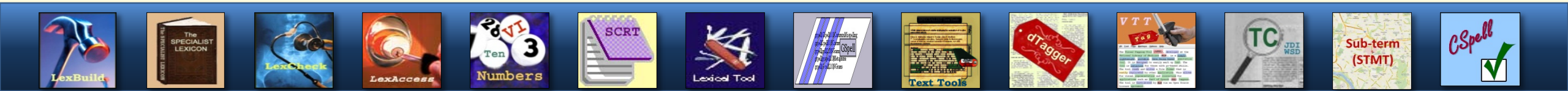
Parallel Flows - Example

```
> lvg -f:n -f:B:s

color
color|color|2047|16777215|n|1|

color|color|128|1|B+s|2|
color|color|1024|1|B+s|2|
color|colour|128|1|B+s|2|
color|colour|1024|1|B+s|2|
```



Norm (commonly used flow)

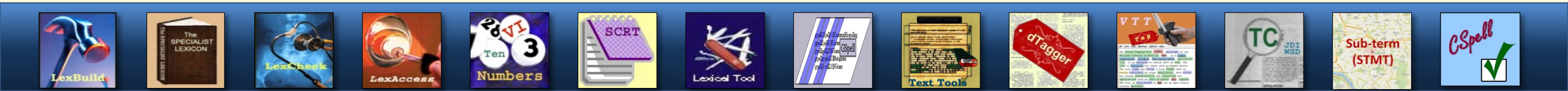
- Composed of 11 Lvg flow components to abstract away from (only keep meaningful words):
  - case
  - punctuation
  - possessive forms
  - inflections
  - spelling variants
  - stop words
  - diacritics & ligatures (non-ASCII Unicode)
  - word order



Example - Norm

"Fœtoproteins α's, NOS"

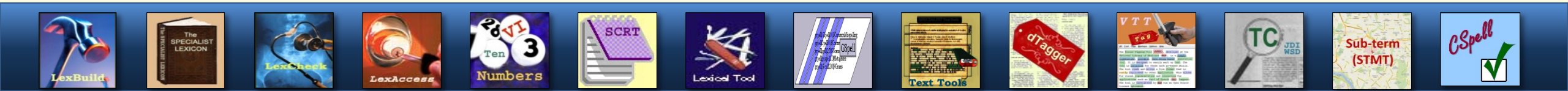
q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order



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

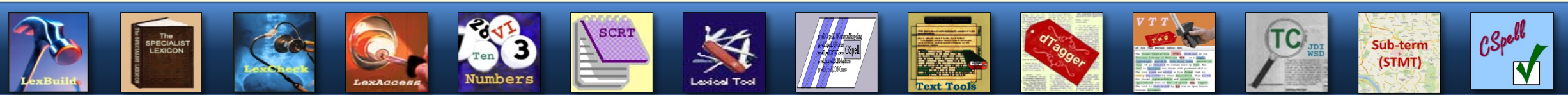
“Fœtoproteins α’s, NOS“
"Fœtoproteins α’s, NOS"



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

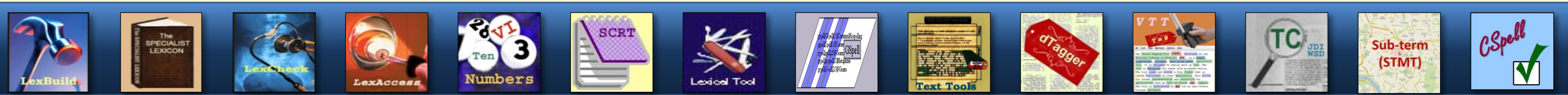
“Fœtoproteins α’s, NOS“
"Fœtoproteins α**’s**, NOS"
"Fœtoproteins α, NOS"



Norm

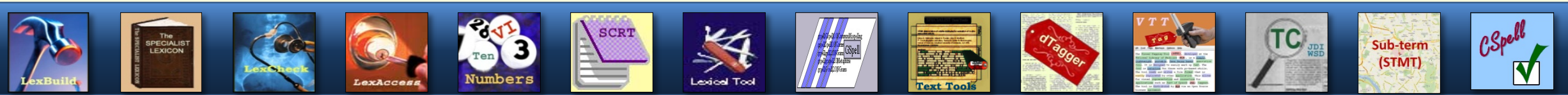
q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order

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



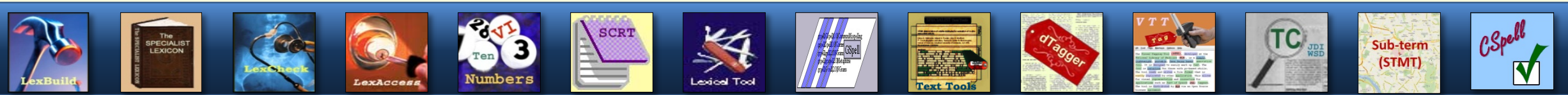
Norm

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	



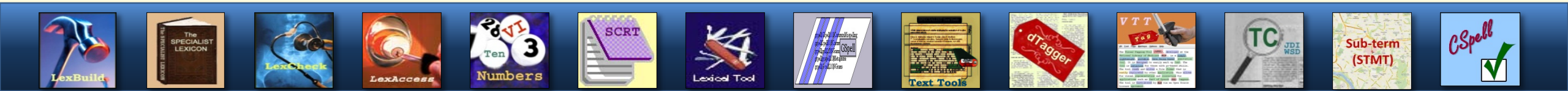
Norm

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	



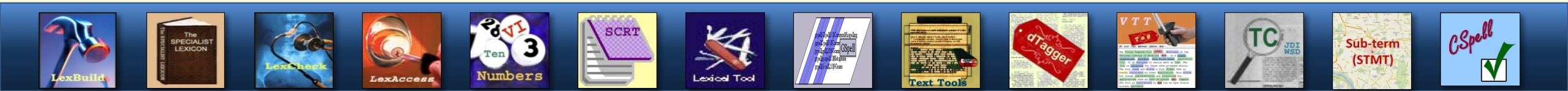
Norm

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	



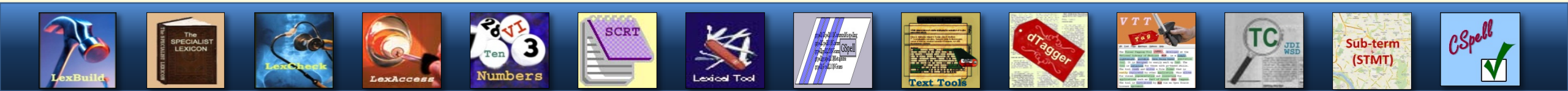
Norm

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	
q8: strip or map Unicode to ASCII	
w: sort words by order	



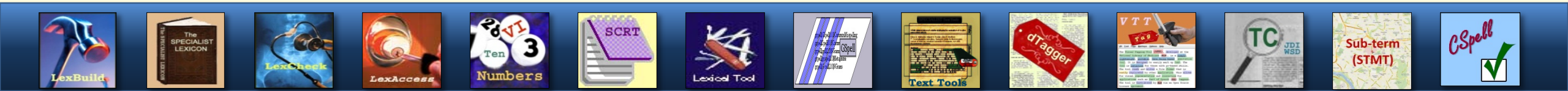
Norm

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	fœtoprotein α
q7: Unicode core Norm	fetoprotein α
q8: strip or map Unicode to ASCII	
w: sort words by order	



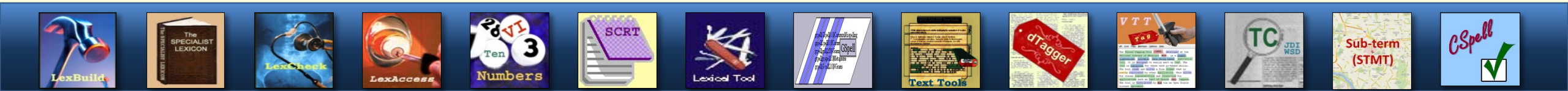
Norm

	“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	
w: sort words by order	



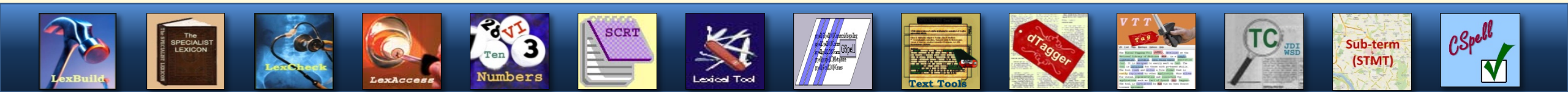
Norm

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



Norm

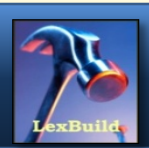
	“Fœtoproteins α’s, NOS“
q0: map symbols to ASCII	"Fœtoproteins α’s, NOS"
g: remove genitives	"Fœtoproteins α, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	Fœtoproteins α NOS
t: strip stop words	Fœtoproteins α
l: lowercase	fœtoproteins α
B: uninflect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

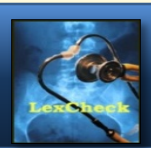


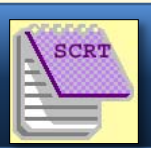
Norm

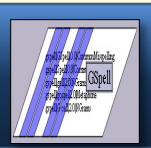
alpha Fetoprotein
alpha Fetoproteins
alpha-Fetoprotein
alpha-Fetoproteins
Alpha fetoproteins
alpha fetoprotein
alpha Foetoprotein
alpha foetoprotein
alpha fetoproteins
Alpha-fetoprotein
alpha-fetoprotein
Alpha Fetoproteins
Alpha-Fetoprotein
Alpha-fetoprotein NOS
Alpha Fetoprotein
alpha-fetoprotein
ALPHA-FETOPROTEIN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

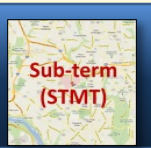
alpha fetoprotein





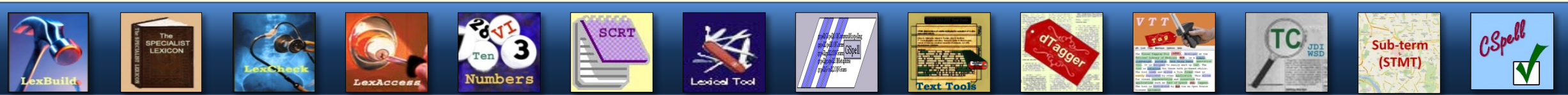






3. Natural Language Processing (NLP)

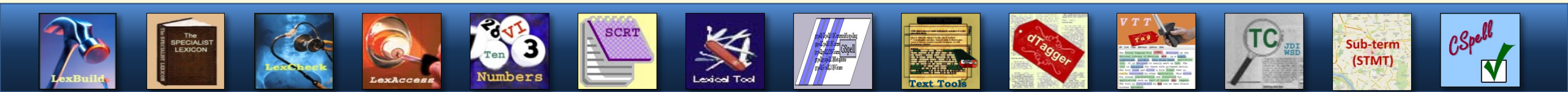
- Natural Language
  - is ordinary language that humans use naturally
  - may be spoken, signed, or written
- Natural Language Processing
  - NLP is to process human language to make their information accessible to computer applications
  - The goal is to design and build software that will analyze, understand, and generate human language
  - NLP includes a board range of subjects, require knowledge from linguistics, computer science, and statistics.
  - NLP in our scope is to use computer to understand the meaning (concept) from text for further analysis and processing.



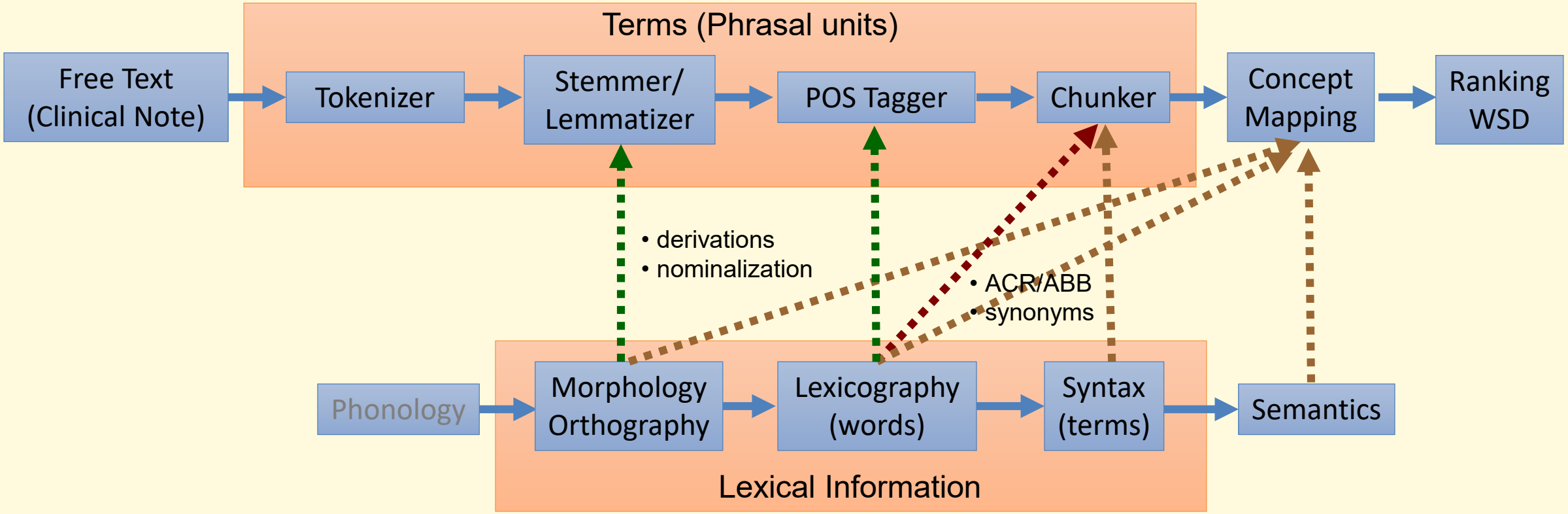
Concept Mapping Challenges

- Challenge: many to many mapping

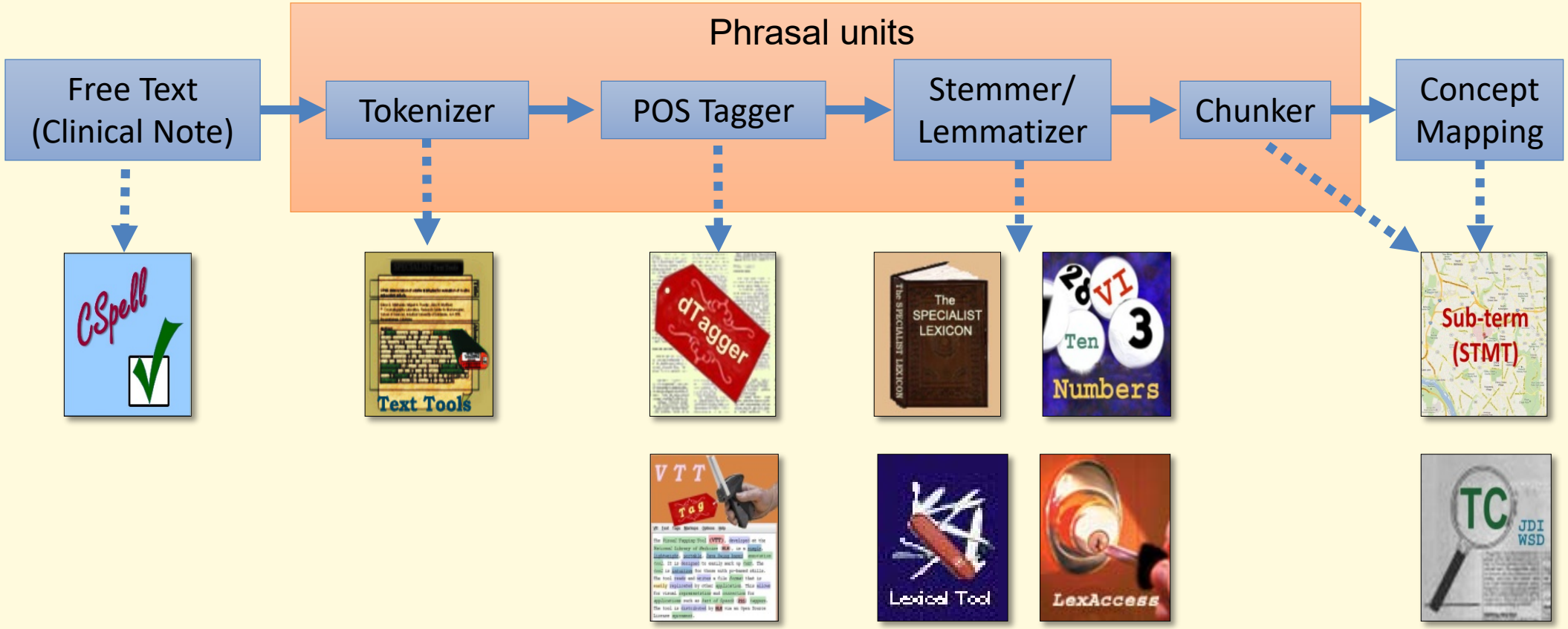
Terms	Concepts	NLP
• cold • Cold Temperature • Cold Temperatures • Cold (Temperature) • Temperatures, Cold • Low temperature • low temperatures • …	• Cold Temperature|C0009264	• Concept mapping
• cold	• Cold Temperature|C0009264 • Common Cold|C0009443 • Cold Therapy|C0010412 • Cold Sensation|C0234192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



NLP Pipe Line – Lexical Information

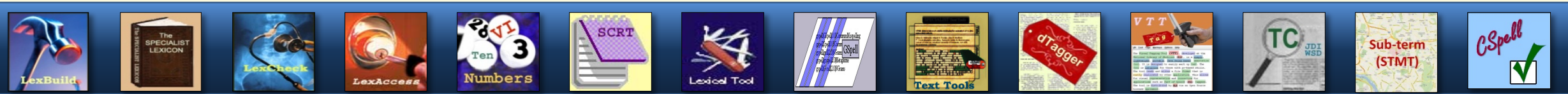


The SPECIALIST NLP Tools



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

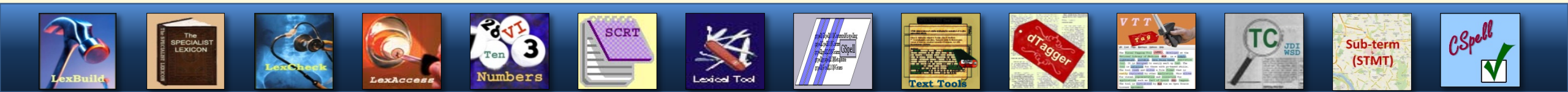
☺



NLP Applications

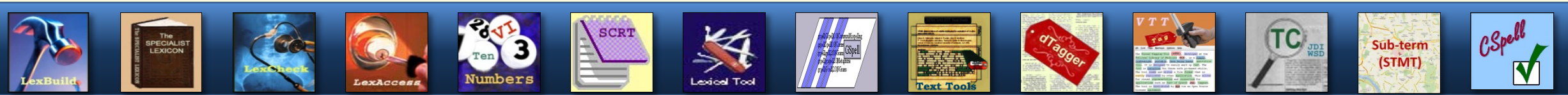
- Syntax:
  - parsers, taggers, POS tagging, etc.
- Semantics:
  - name entity recognition, **concept mapping**, etc.

- Knowledge extraction:
  - learn relations between entities, recognize events, etc.
- Summarization:
  - sentiment analysis and figure out the topics of a page
- Question answering
  - find answers for queries



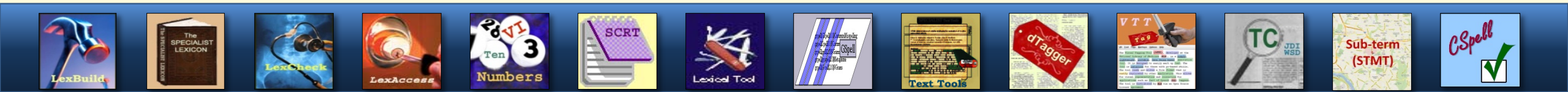
NLP – Concept Mapping

- Normalization (same record):
  - A term might have a great deal of lexical variations, such as inflectional variants, spelling variants, abbreviations (expansions), cases, ASCII conversion, etc.
  - Normalize different forms of a concept to a same form
- Query Expansion (related records):
  - Expand a term to its equal terms, such as subterm substitution of synonyms, derivational variants, abbreviations, etc.
  - To increase recall
- POS tagger:
  - Assign part of speech to a single word or multiword in a text
  - To increase precision
- Others…

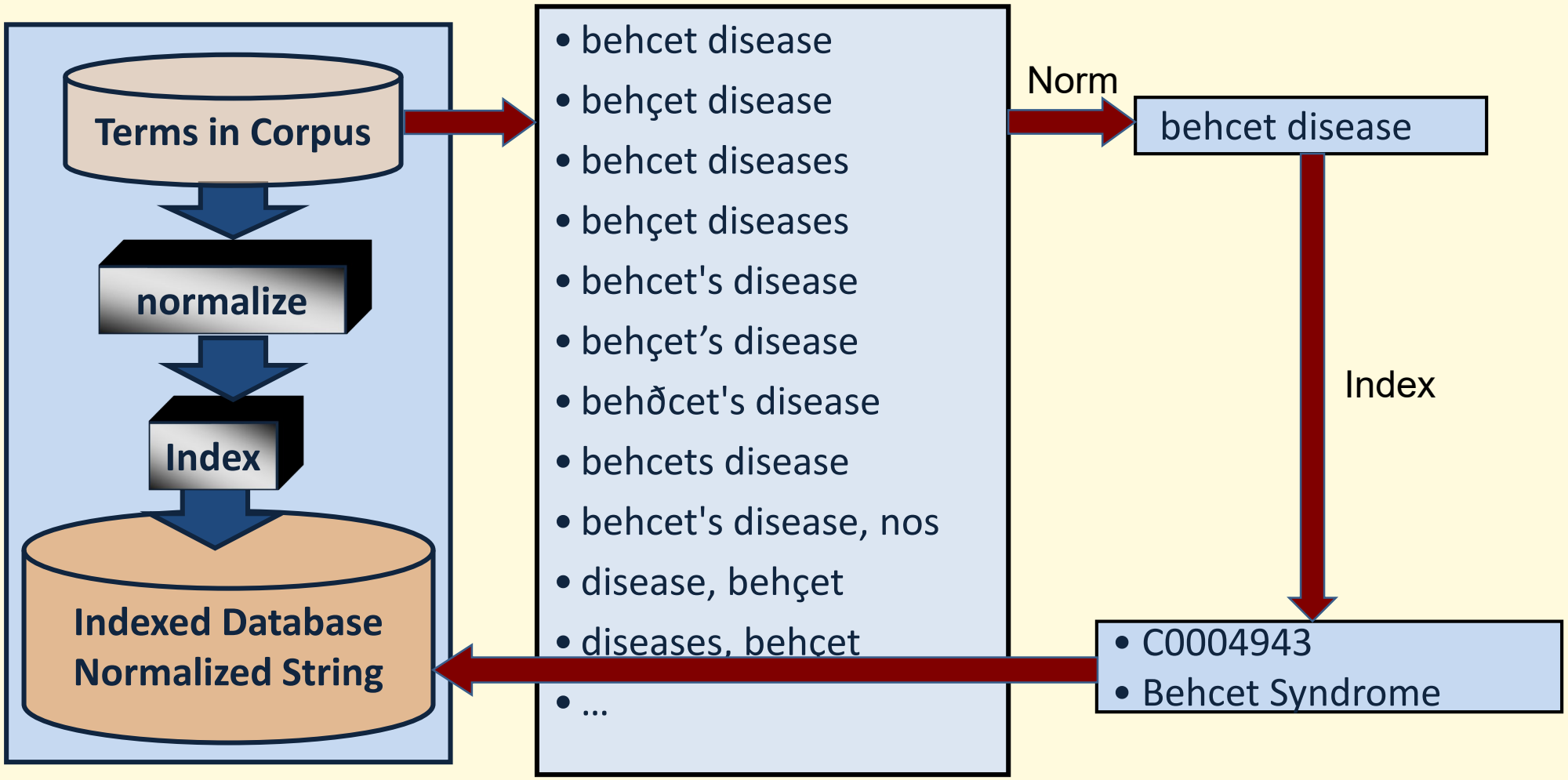


Lexical Tools – Norm

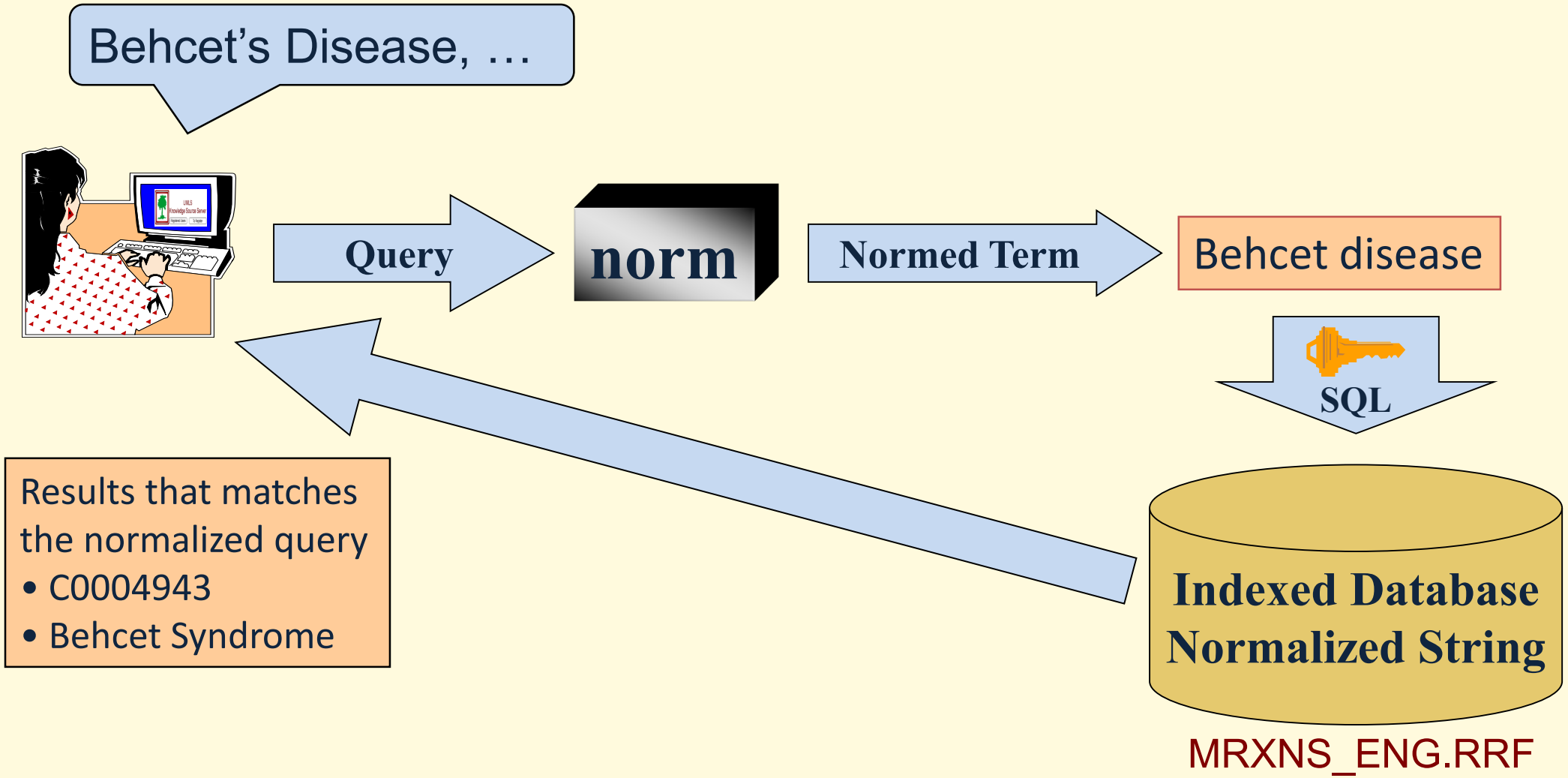
	Behçet’s Diseases, NOS
q0: map Unicode symbols to ASCII	Behçet's Diseases, NOS
g: remove genitives	Behçet Diseases, NOS
rs: remove parenthetic plural forms	Behçet Diseases, NOS
o: replace punctuation with spaces	Behçet Diseases NOS
t: strip stop words	Behçet Diseases
l: lowercase	behçet diseases
B: uninflect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



NLP – Norm (Pre-Process Lexical Variations)

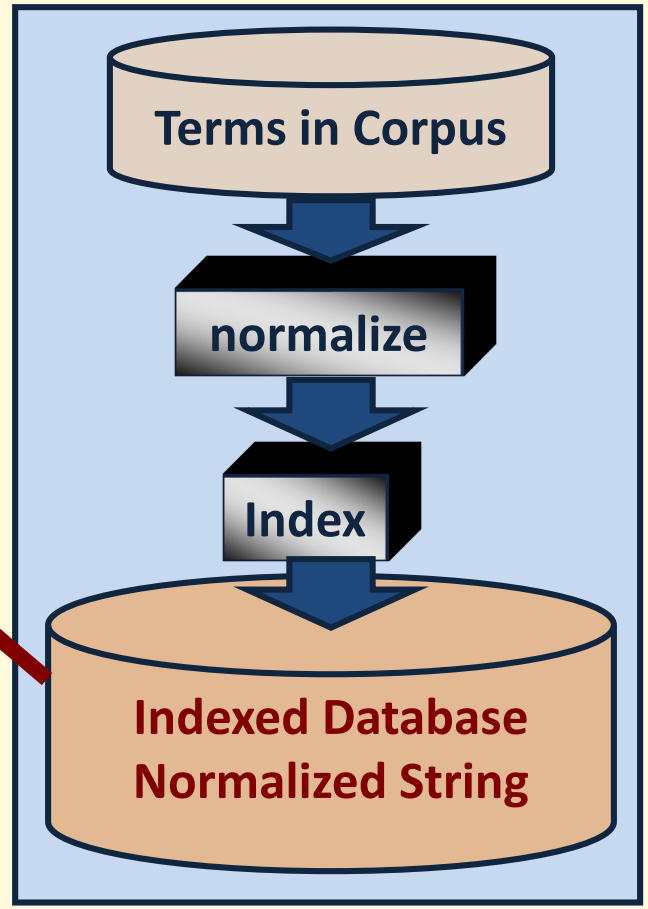


NLP – Norm (Application)

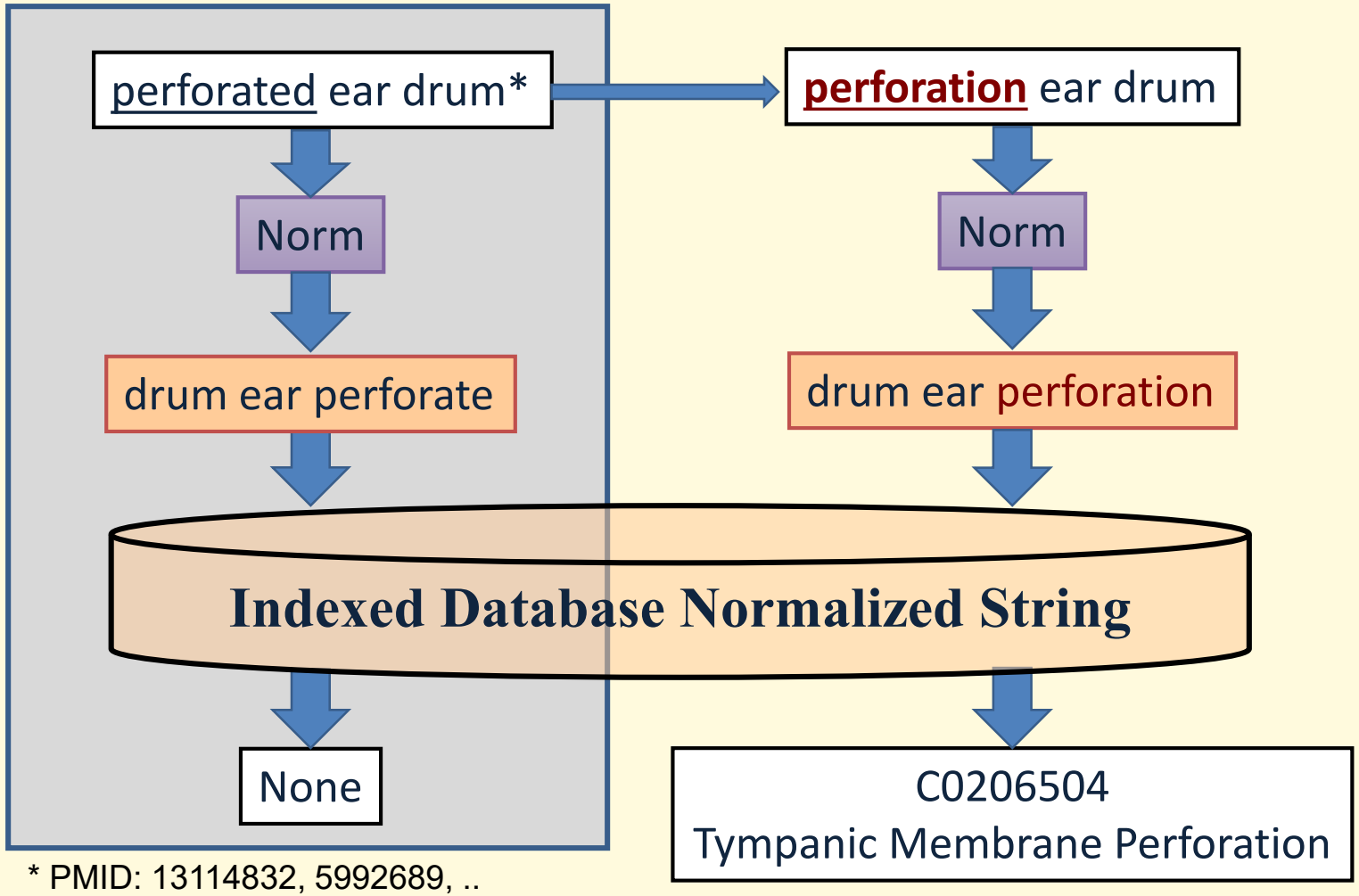


UMLS Metathesaurus

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

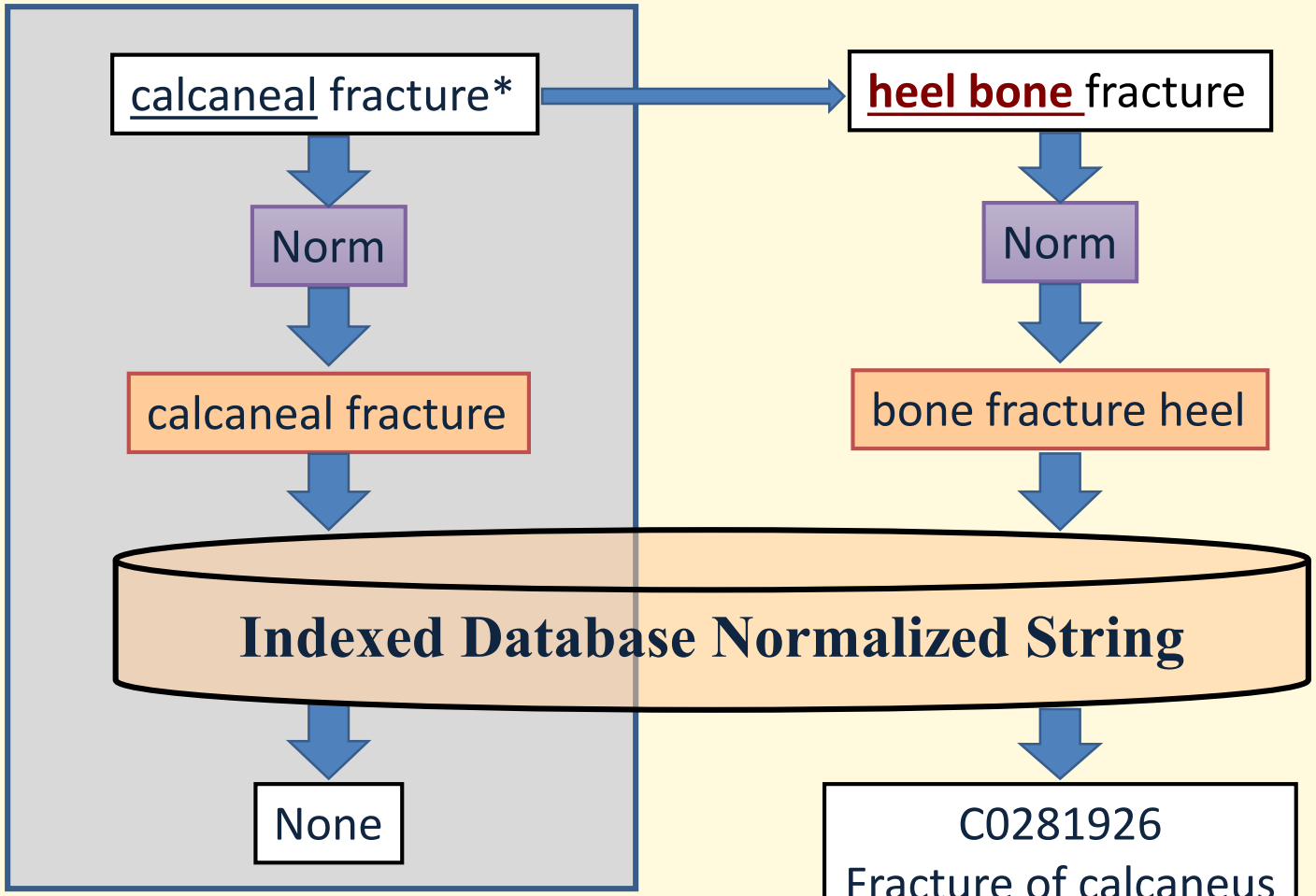


NLP – Query Expansion (derivation)



* PMID: 13114832, 5992689, ..

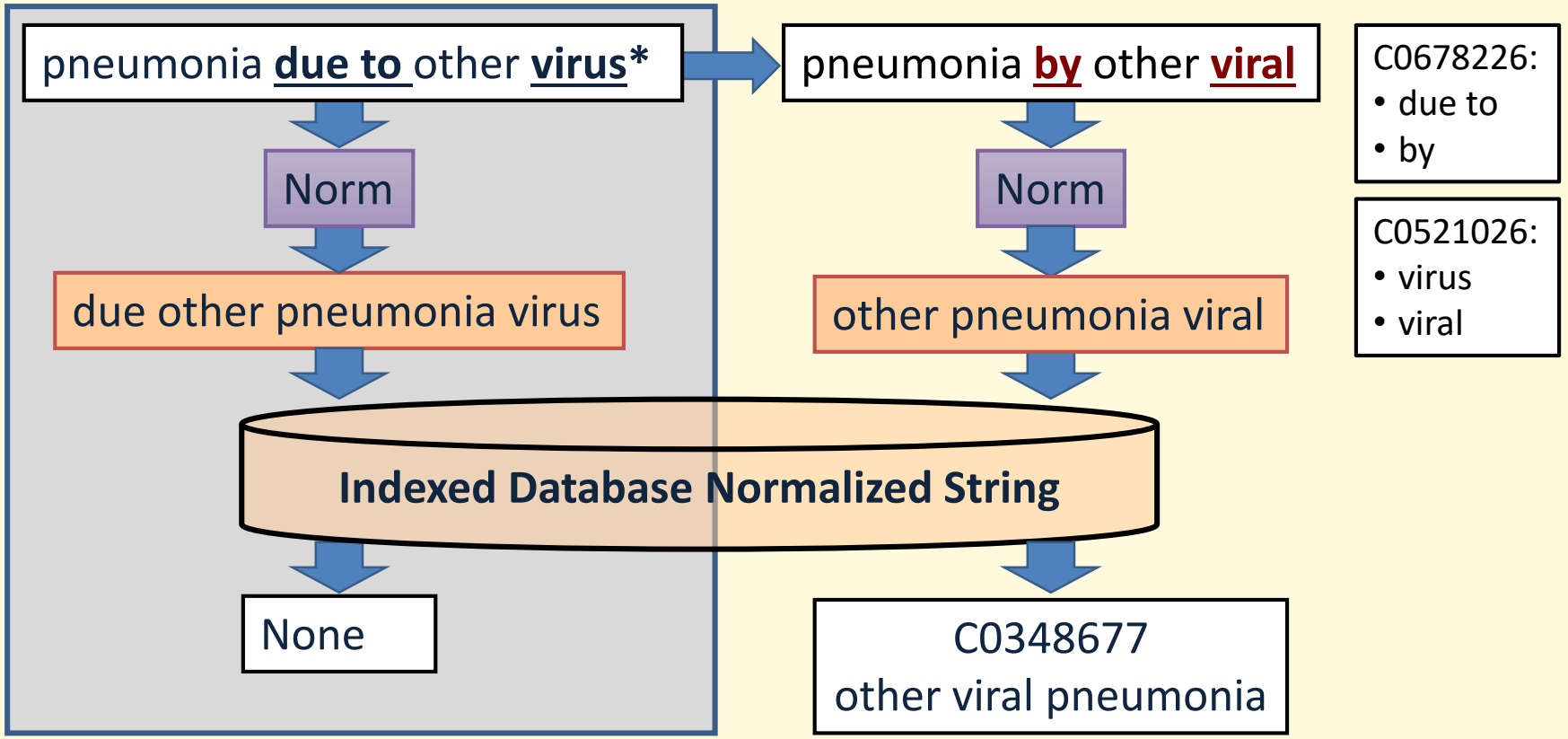
NLP – Query Expansion (Synonym)



C0006655:
- calcaneal
- heel bone

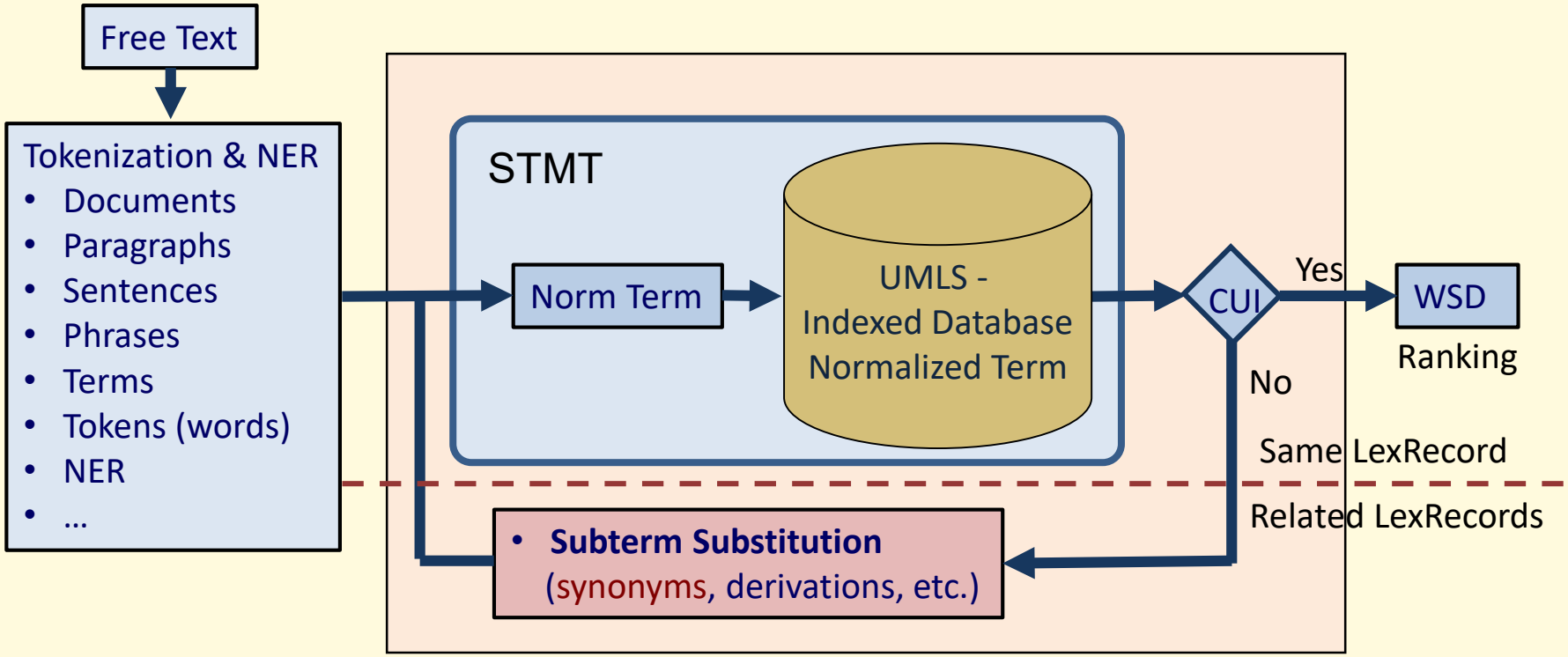
* PMID: 1118604, 1165396, ..

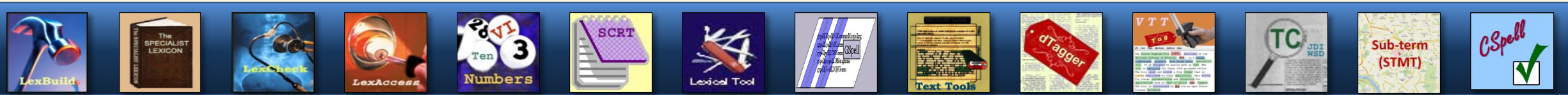
Multiple Substitutions



* VA14760, HA480.80, ..

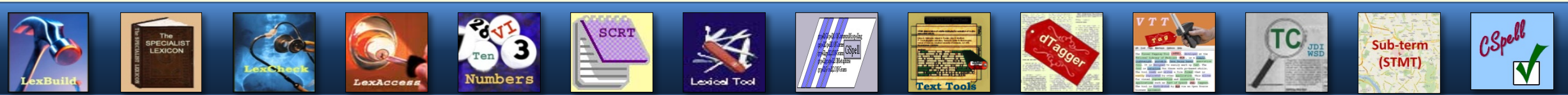
Real-time Concept Mapping Model





4. Applications - CSpell

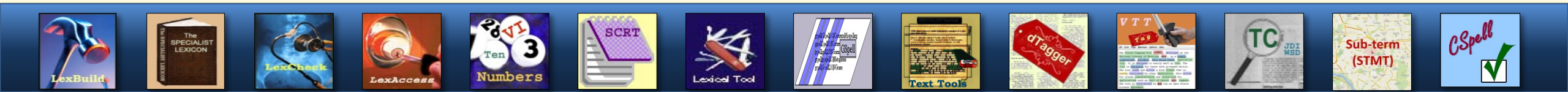
- CSpell - Spell checker for consumer language
- Editor's Choice paper in JAMIA, Volume 26, Issue 3, 1 March 2019, Pages 211-218
- JAMIA Journal Club Webinar, 04/11/2019
- http://umlslex.nlm.nih.gov/cSpell
- http://SPECIALIST.nlm.nih.gov/cSpell
- Email: umlslex@nlm.nih.gov



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)

☺



Consumer Questions Example

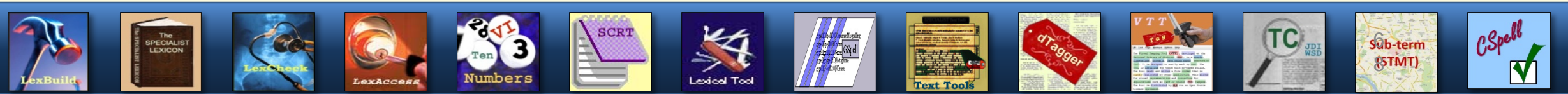
My mom is 82 years old suffering from anixity and depression for the last 10 years was dianosed early on set deminita 3 years ago. Do yall have a office in Greensboro NC? Can you recommend someone. she has seretona syndrome and nonething helps her. [2]

- 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.



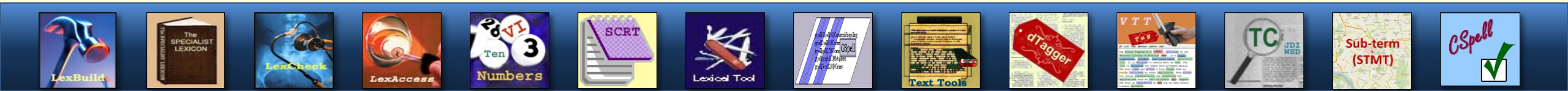
Ensemble Method (2015)

- Nonword

Method	Precision	Recall	F1	δ F1
ESpell	0.53	0.20	0.29	-----
Ensemble	0.64	0.58	0.61	0.32

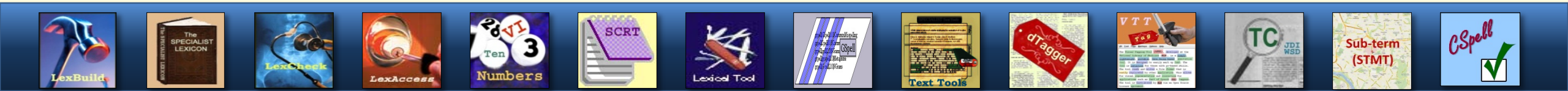
- Real-word

Method	Precision	Recall	F1	δ F1
ESpell	0.23	0.26	0.25	------
Ensemble	0.57	0.59	0.58	0.33



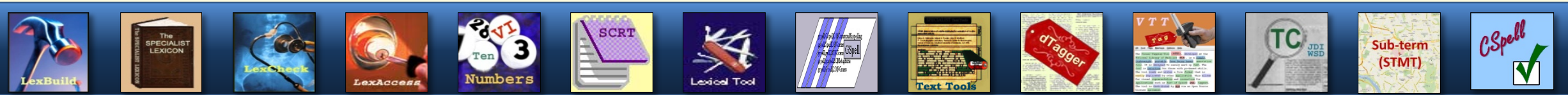
Project Objectives

- To develop a spelling tool to detect and correct all types of spelling errors in consumer language
  - Performance
  - Speed
  - Distributable
  - Generic
  - Open-source
  - Configurable



Types of Spelling Errors & Corrections

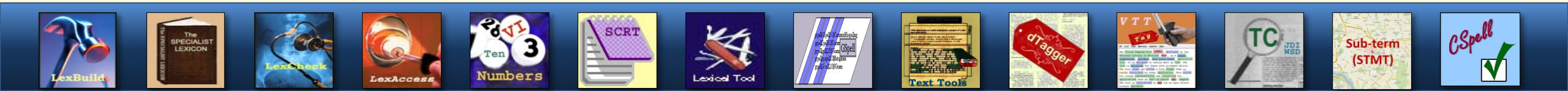
- Nonword vs. real-word errors
- Spelling errors
- Word boundary infraction errors: split and merge
- Punctuation errors
- Dictionary-based vs. non-dictionary-based corrections
- Isolated-word vs. context-dependent corrections
- Single token vs. multi-token corrections
- Others: informal expression, HTML/XML tag, …
- Multiple corrections of combination errors



Training Data Set

- Used both the training set and test set from Ensemble method for development

Training Set Summary Statistics	
Consumer health questions	471
Tokens	24,837
Annotation tags	1,008
Nonword corrections	774
Real-word corrections	964
Word count/question	5-328
Average word count/question	52.49
Errors per/question	0-27
Average error/question	2.14
Error rate (error/token)	0.04



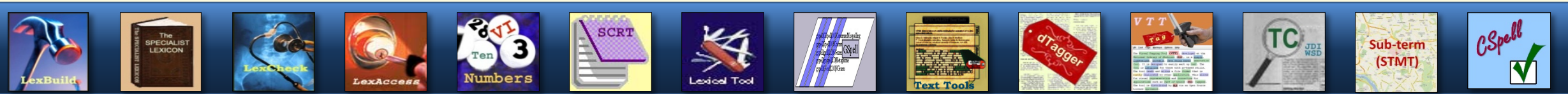
Error Types Analysis

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

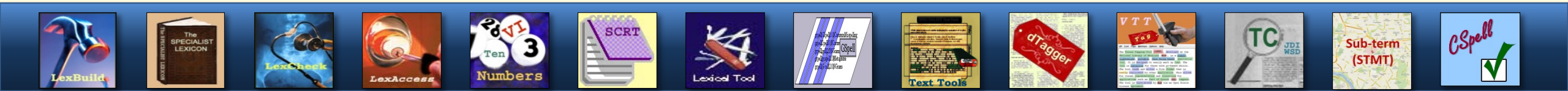
Note: ND: not dictionary based.

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



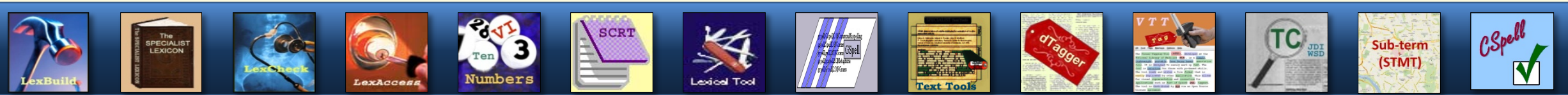
Dictionary-based Examples

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



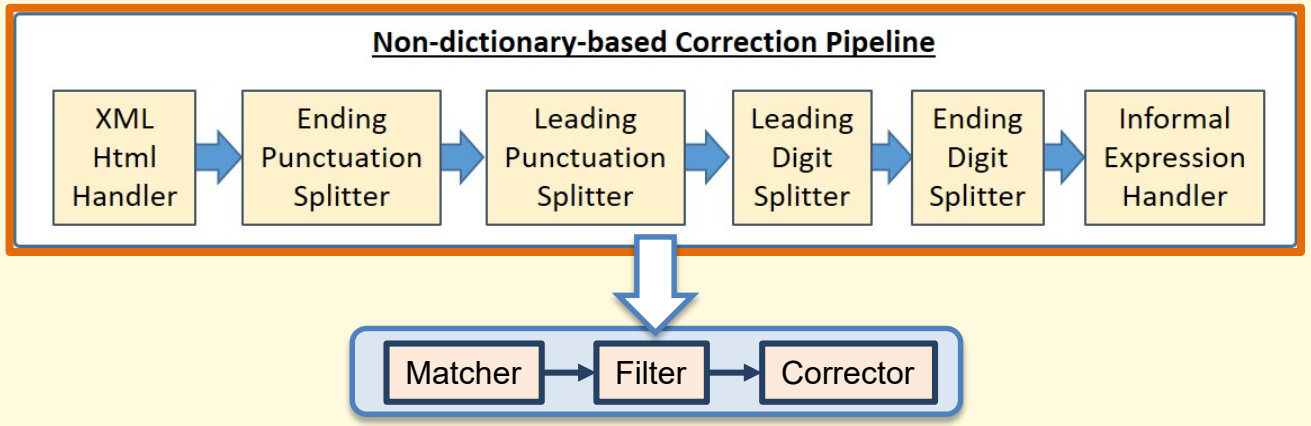
Non-dictionary-based Examples

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

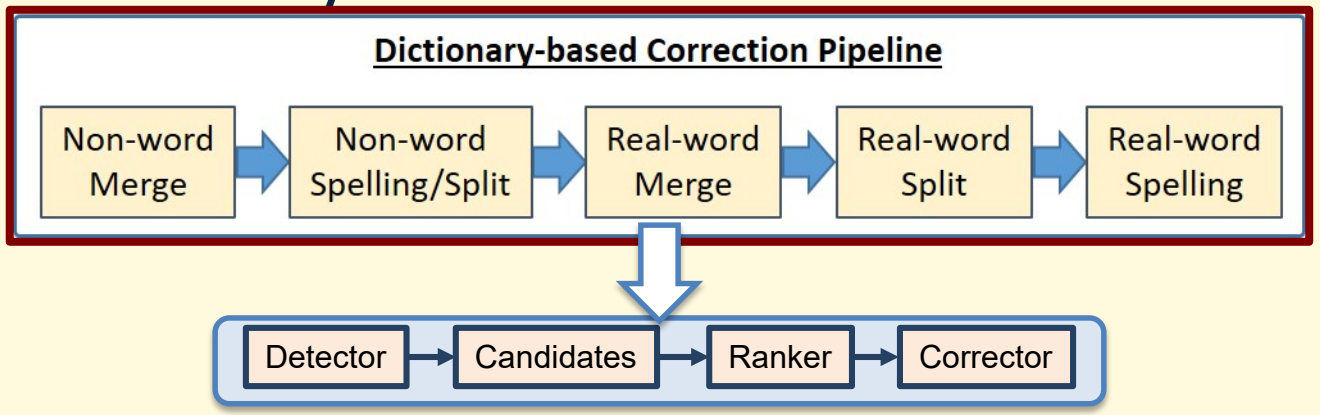


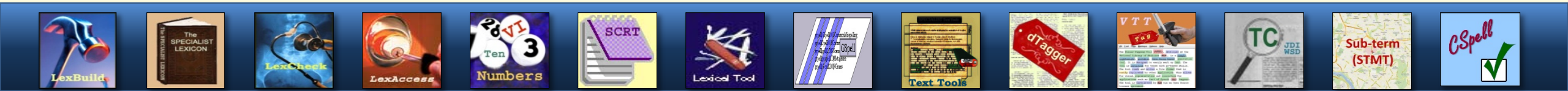
Error & Correction Types

- Non-dictionary-based correction model:

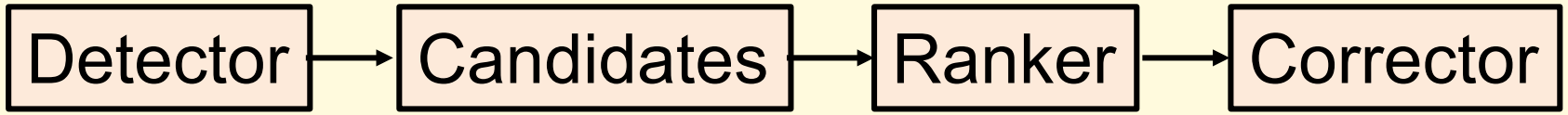


- Dictionary-based correction model:





Dictionary-based Model

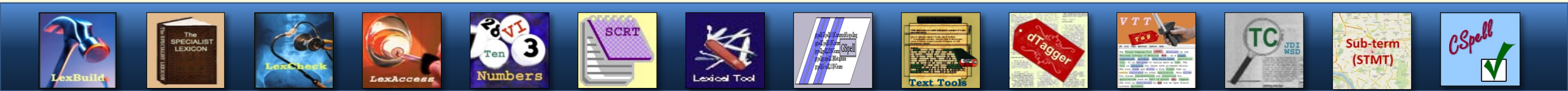


- Example:

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

Isolated-word Correction Techniques

- Edit Distance Similarity
- Phonetic Similarity
- Overlap Similarity
- Orthographic Similarity
- Word Frequency

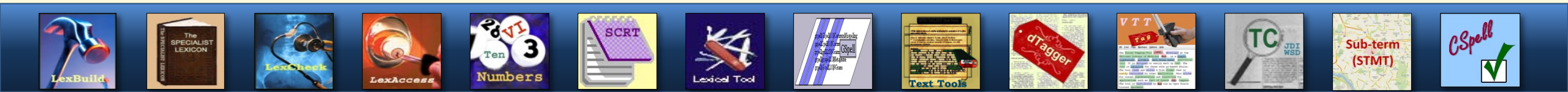


Edit Distance Similarity

- Edit **Dist**ance:
  The minimum number of operations required to transform one string into the other
- Example of “truly”:

Input	Output	Operation	Dist.	Cost	EDSS*
truely	truly	Deletion	1	0.096	0.904
trly	truly	Insertion	1	0.090	0.910
truli	truly	Substitution	1	0.100	0.900
turly	truly	Transposition	1	0.094	0.906

* EDSS: Edit Distance Similarity Score

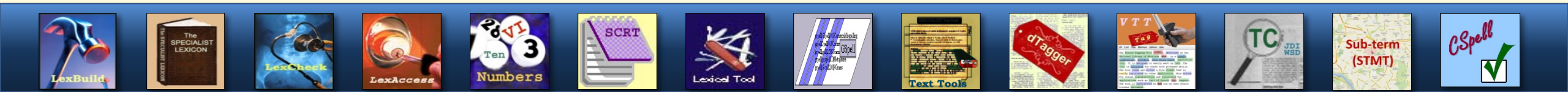


Phonetic Similarity

- Phonetic algorithm:
  - Algorithm converts strings to codes for indexing by pronunciation
- Phonetic similarity:
  - Phonetic code + edit distance similarity
- Example: “diagnost” vs. “diagnosis”

Phonetic Algorithm	diagnost	diagnosis	E.D.	PSS*
Refined Soundex	D6048036	D60480303	2	0.80
Metaphone	TNST	TNSS	1	0.90
Caverphone 2	TKNST11111	TKNSS11111	1	0.90
Double Metaphone	TKNST	TKNSS	1	0.90

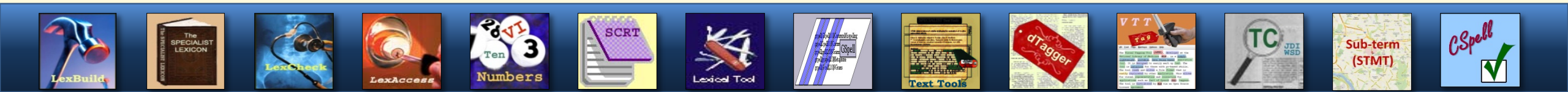
* PSS: Phonetic Similarity Score



Overlap Similarity

- Overlap similarity:
  - Calculates the overlap of matching characters at the beginning and the end of 2 terms, divided by the length of the longer term
  - Similarity score is between 0.00 and 1.00
- Example: “truely” and “truly”

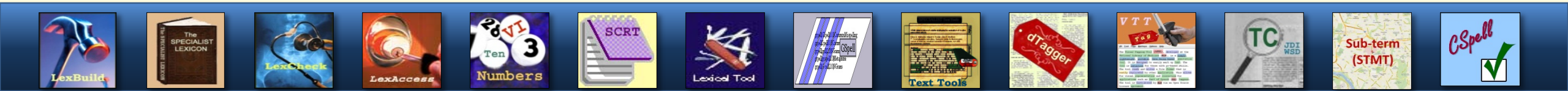
	Values	Notes
Lead Overlap Character	3	tru-
Trail Overlap Character	2	-ly
Max Length	6	truley:6, truly: 5
Overlap Similarity Score	0.832	= (3+2)/6 = 5/6



Orthographic Similarity

- Orthographic (token) similarity
  - Looks and sounds alike (error model)
  - Enhanced similarity score (weighted sum)
    = 1.0*Edit Distance + 0.7*Phonetic similarity + 0.8*Overlap similarity
- Example: “truely” and “truly”
  - Orthographic similarity score
    = 1.0*0.904 + 0.7*1.0 + 0.8*0.83
    = 2.27

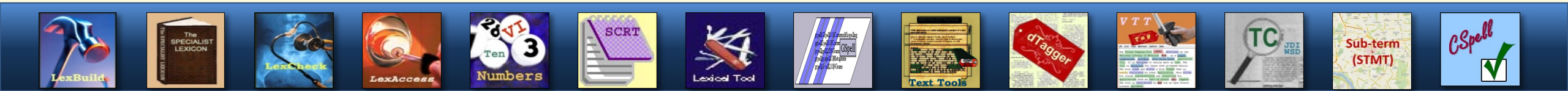
	Edit Distance Similarity	Phonetic Similarity	Overlap Similarity
Operation	deletion	[TRL], [TRL]	tru-, -ly
Score	0.904 = 1 – 0.096	1.0	0.83 = (3+2)/6
Weights	1.0	0.7	0.8



Word Frequency

- Word frequency score
  - Word count of occurrence within a given text corpus (language model)
  - Word: unigram (lowercase)
  - Normalized score: 0.0 ~ 1.0, normalized by the max. WC (the, 467,713)
- Example:

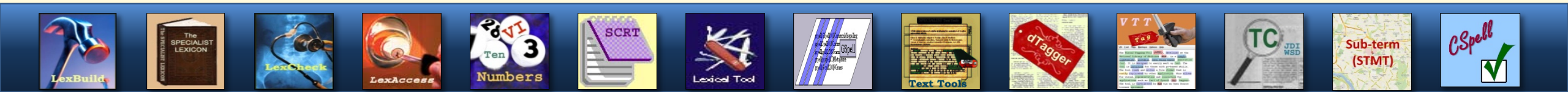
Word	Word Count	Norm. Word Frequency
diagnosis	9,083	0.019420029
diagnosed	1,948	0.004164947
diagnose	2,115	0.004522004
diagnostic	769	0.001644171
diagnoses	203	0.000434027
diagnoser	0	0.000000000



Consumer Health Corpus

- Collected articles from 16 consumer-facing NIH websites
  - 17,139 articles
  - 10,228,699 tokens
  - 192,818 unique words

Resources	Articles
Genetic and Rare Diseases - diseases	6484
Genetics Home Reference - conditions	1215
Genetics Home Reference - genes	1439
MedlinePlus - drugs	1383
MedlinePlus - medical encyclopedia	4425
MedlinePlus - all health topics	1013
MedlinePlus - herbs and supplements	153
National Eye Institute	36
National Heart, Lung, and Blood Institute	141
National Institute of Allergy and Infectious Diseases	53
National Institute of Arthritis and Musculoskeletal and Skin Diseases	55
National Institute of Child Health and Human Development	81
National Institute on Deafness and Other Communication Disorders	15
National Institute of Diabetes and Digestive and Kidney Disease	181
National Institute of Mental Health	26
National Institute of Neurological Disorders and Stroke	439



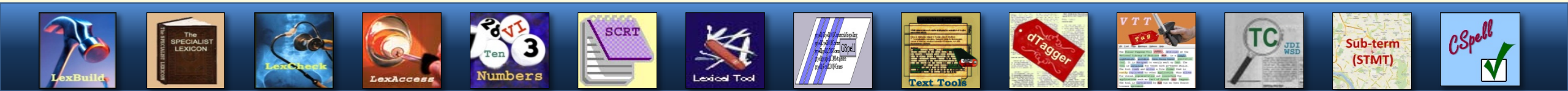
Toward Context-dependent Corrections

- Isolated-word corrections

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

- Context-dependent corrections

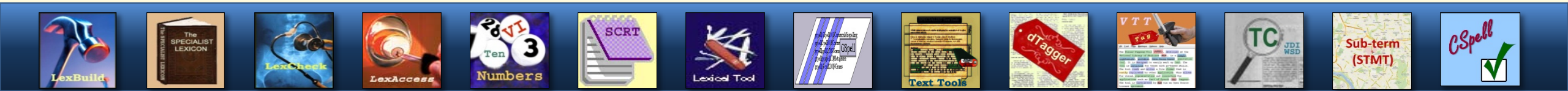
Input	Correction
the diagnost	the diagnosis
was diagnost	was diagnosed



Word2vec: Word Embedding

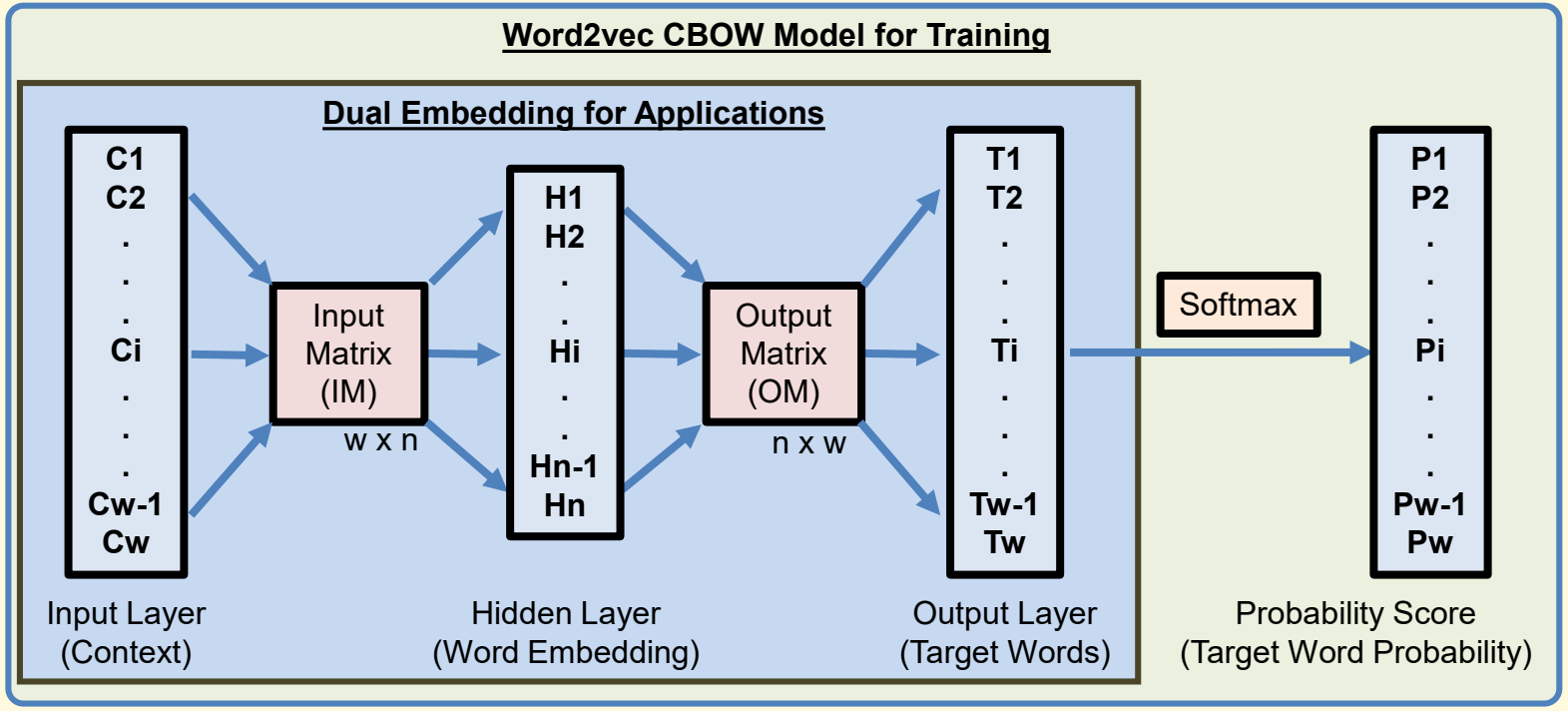
- Word vectors (word2vec, 2013 [22-23]):
  - Each word has an associated vector
  - Represent the meaning of a word in some abstract way
  - Capture meaningful syntactic and semantic regularities
- 2 Models:
  - Continuous bag-of-words (CBOW): predict a word from context
  - Continuous skip-gram: predict context from a word
- Examples (prediction by similarity):
  - Man -> Kings, Woman -> Queens

  => Kings – Man + Woman = Queens (vector operations)



CBOW - Dual Embedding

- Dual embedding: use both the IM and OM matrices to compute context scores of the predicted target word with given contexts
- Context score = [T] = [C] X [IM] X [OM]



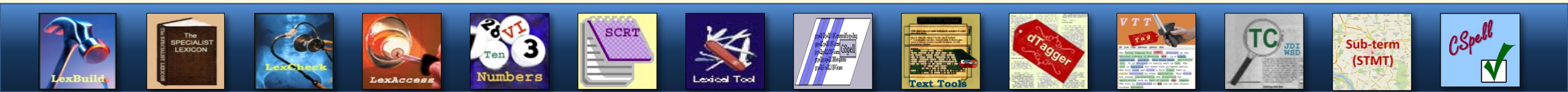
Single vs. Dual Embedding

- Single embedding:
  - Use the [IM] as word vectors
  - Use the cosine similarity between context and candidates' word vectors for context score [2,15]
- Dual embedding:
  - Use both [IM] and [OM] to feed the context into the original CBOW model for context score of candidates

Embedding	Matrixes	Precision	Recall	F1 *
Single	IM	0.5887	0.5917	0.5902
Dual	IM & OM	0.8035	0.5917	**0.6815**

=> Dual embedding has better precision and F1 (+9.13%)

* Training set, nonword test by context scores



Context-dependent Corrections

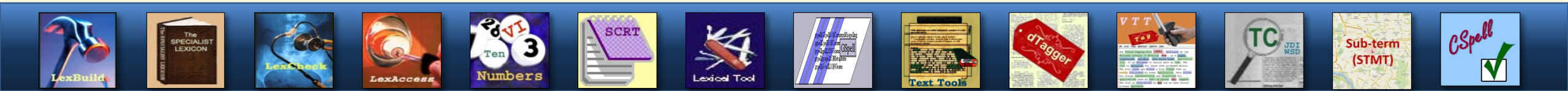
- Nonword corrections

Input	Correction
the diagnost	the diagnosis
was diagnost	was diagnosed

- Real-word corrections

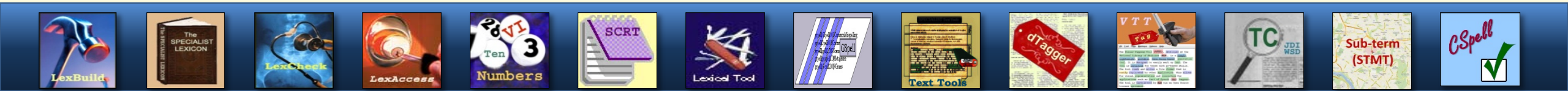
Input Text	Correction
smell size	small size
smell amount	small amount
smell intestine	small intestine

Input Text	Correction
foul small	foul smell
small an order	smell an order
taste and small	taste and smell



2-stage Ranking System

- Efficiently utilize the knowledge sources, similar to regular season/play-offs in sports championship
- Stage-1 (regular season):
  - Orthographic similarity scores > threshold
- Stage-2 (playoffs):
  - Chain comparators by the context score, then the noisy channel score in a sequential order
  - Rank in stage-1 is ignored

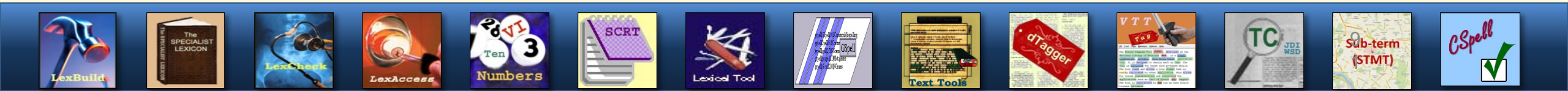


1-stage vs. 2-stage Ranking

➢ 2-stage is better than 1-stage (Nonword, training set)

Stage-1	Stage-2	Precision	Recall	F1 *
Orthographic	N/A	0.7606	0.7636	**0.7621**
Word Frequency	N/A	0.6970	0.6925	0.6948
Noisy Channel	N/A	0.7134	**0.7171**	0.7152
Context Score	N/A	**0.8035**	0.5917	0.6815
Ensemble	N/A	0.7516	0.7545	0.7531
Orthographic	Word Frequency	0.8241	0.7687	0.7955
Orthographic	Noisy Channel	0.8255	**0.7700**	0.7968
Orthographic	Context Score	**0.8996**	0.5672	0.6957
Orthographic	Context Score, Noisy Channel	0.8047	0.7842	**0.8115**

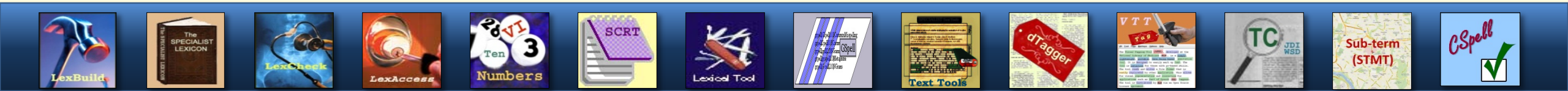
=> Best F1



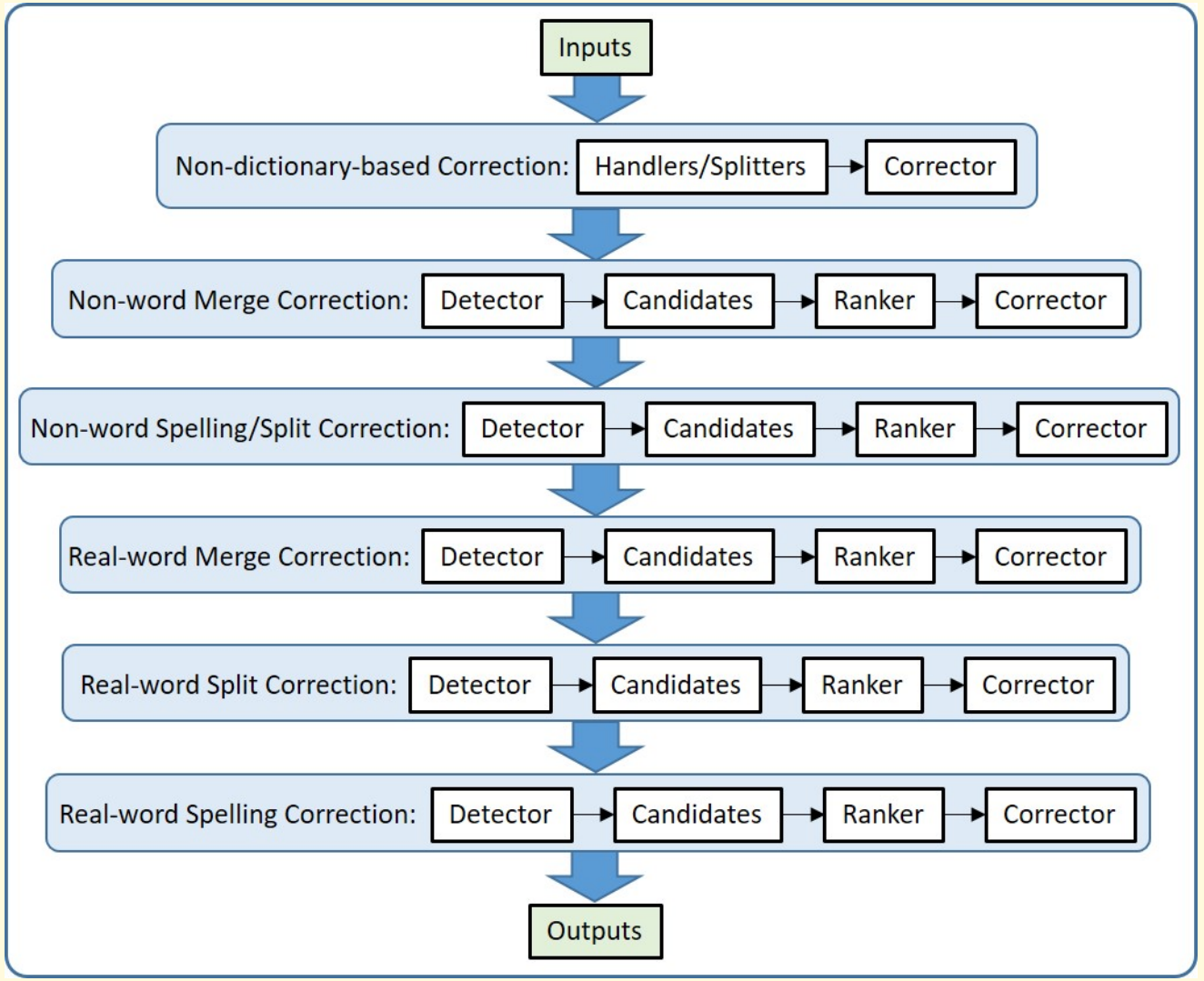
2-stage Ranking Example

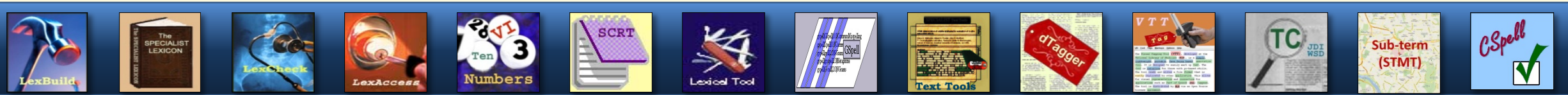
- Input: havy
- Candidates: 441
  - Stage 1: find top candidates by orthographic similarity score
  - Stage 2: use context score, then Noisy Channel score

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



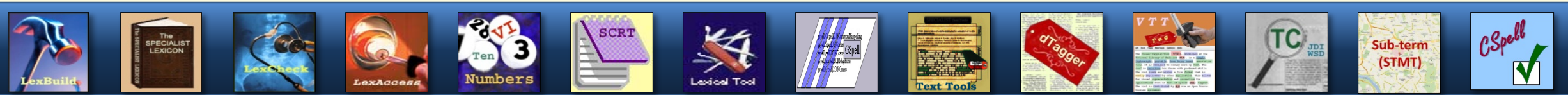
Multilayer Design





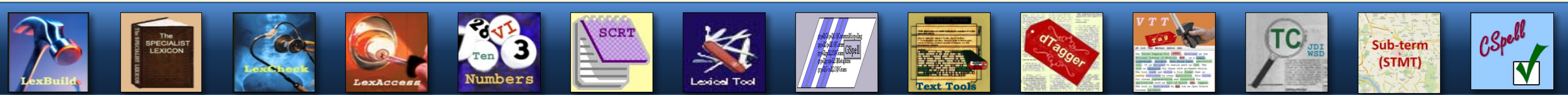
CSpell Algorithm Summary-1

Process	Detector	Candidates	Ranker	Corrector	Features
1. Non-Dic, PreCorrector (Java 8 Stream)	• Pattern Match • HashMap	• Algorithm • HashMap	• None	• Replace • FlatMap • UpdateHist	• XmlHtmlHandler • EndPuncSplitter • LeadingPuncSplitter • LeadingDigitSplitter • EndingDigitSplitter • InformalExpHandler
2. Dic, Non-Word, Merge	• Not Valid Word: splitDic (SwNoAaLc.dic) • Not Exceptions: Digit, Punc, Url, Email, EmptyStr, UpperCase, 1Char, Measurement	• <= MergeNo • Merge " " and "-" • context not an exception • orgWord not in mwDic, • cand in suggDic, not in aaDic	• Word2Vec: topScore != 0 • Frequency	• Reconstruct the whole text • UpdateHist	
3. Dic, Non-Word, 1-to-1 & Split	• Not Valid Word: CheckDic (Ew, Num) • Not Exceptions: Digit, Punc, Url, Email, EmptyStr, Measurement	• Dist <= 2 • In suggDic • < SplitNo • In mwDic • In splitDic • Not digit, unit	• Sorted by Orthographic • Refined candidates to a range (within 0.08 of top Oscore) • Use top context score • Use top NoisyChannel score	• Replace • FlatMap • UpdateHist	



CSpell Algorithm Summary-2

Process	Detector	Candidates	Ranker	Corrector
4. Dic, Real-Word, Merge	• Not corrected yet • Valid Word (splitDic) • Not Exceptions	• <= MergeNo • Merge " ", not "-" • context not an exception • orgWord not in mwDic, • cand in suggDic, not in aaDic • Cand has word2Vec • Cand WC >= 10 (Conf) • No short word merge	• Use Word2Vec to find top rank: • Valid top rank - orgS > 0 & topS*Fac > orgS - orgS < 0 & topS > 0 - orgS < 0 & topS < 0 & topS > orgS*Fac • cFac = 0.60 (Conf)	• Reconstruct the whole text • UpdateHist
5. Dic, Real-Word, Split	• Not corrected yet • Valid Word (checkDic) • Not Exceptions • Length > 3 (Conf) • Has word2Vec • WC > 200 (Conf)	• <= SplitNo • The split term is LMW • Short split word - total no of short split word • split word - Valid split words (splitDic) - Has word2Vec - WC >= 200 (Conf) - Not unit - Not proper noun	• Use Word2Vec to find top rank: • Valid top rank: - Same score ranking rules as merge • cFac = 0.00 (Conf)	• Add the FlatMap of the split term to the in token list • UpdateHist
6. Dic, Real-Word, 1-to-1	• Not corrected yet • Valid Word (checkDic) • Not Exceptions (aa, pn) • Length >= 2 (Conf) • Has word2Vec • WC > 65 (Conf)	• Valid word (suggDic) • Cand has word2Vec • Cand WC >= 0 (Conf) • Cand Length >= 2 (Conf) • Not inflVar of inWord • Heuristics Rules - pmDist=0 & prDist =0 - tDist1 < 3 & tDist2 < 4 & pmDist == 0	• Find top rank cand (sorted): - top Orthographic Score - top freqScore - top EditScore - top phoneticScore - top overlapScore • Validate top rank cand: - cScore (top > 0, org < 0) * top > -0.1*org * top-org > 0.085 * org > -0.085 * topF > 0.0006 * topF/orgF > 0.035 * orgF – topF < 0.02 or topF > orgF	



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

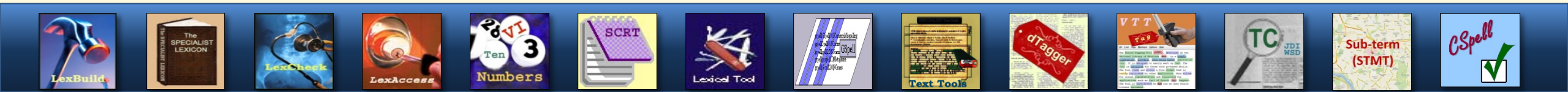
RW Merge | NW Spelling | ND Split

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



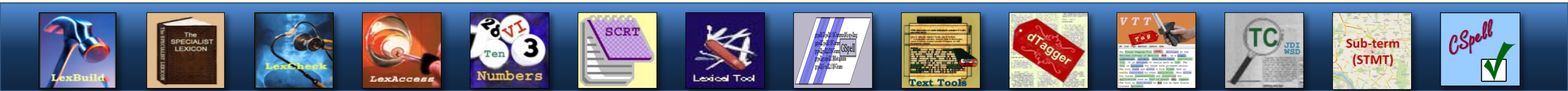
Results From Training Set

- 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**

- Correction:

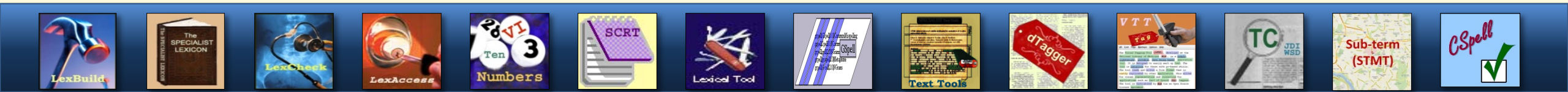
Method	Precision	Recall	F1	δ F1		
ESpell	0.2076	0.2541	0.2285	-----	-0.05	-0.38
Jazzy	0.4860	0.1981	0.2815	0.05	-----	-0.33
Ensemble	0.7201	0.5363	**0.6147**	0.38	**0.33**	-----
CSpell	0.8416	0.6504	**0.7338**	**0.50**	**0.45**	**0.12**



Test set

- Consumer health questions with most OOV (out of vocabulary) terms
  - 2 annotators, 1 arbitrator
  - Discrepancies: reconciled, then arbitrated

	Test	Training
Consumer health questions	224	471
Tokens	16,707	24,837
Annotation tags	1,946	1,008
Nonword corrections	974	774
Real-word corrections	1178	964
Word count/question	3-337	5-328
Average word count/question	72.36	52.49
Errors/question	0-22	0-27
Average error/question	4.90	2.14
Error rate (error/token)	0.07	0.04



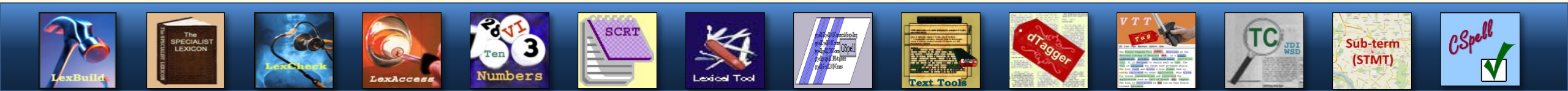
Results From Test Set

- Detection:

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

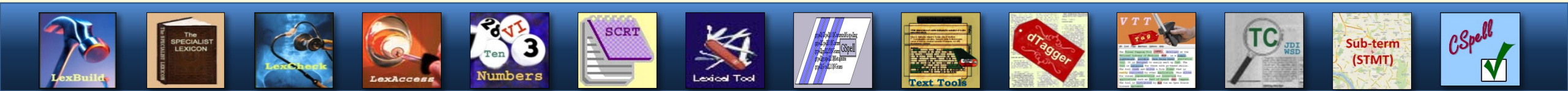
- Correction:

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



Speed Performance Test

- Real-time corrections
  - **11.38** faster than Ensemble (= 2064/181.28)
  - 430 words with 12.6 nonword corrections per second
  - 120 words with 3.6 real-word corrections per second
  - Tested on Red Hat Enterprise Workstation 7.3 Maipo (Red Hat Inc, Raleigh, NC), Intel Xeon(R) CPU E5506 @2.13GHz (Dell)



Dictionary Selection

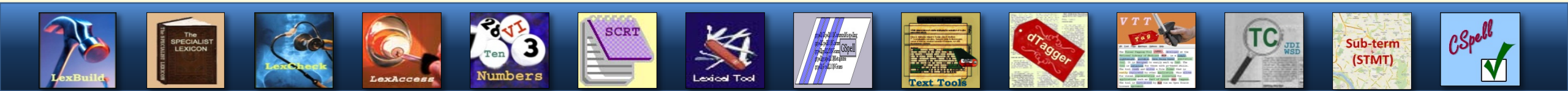
- Relevant, supervised and comprehensive dictionary is better

Dictionary	Size	B*	S*	C*	Precision	Recall	F1**
Jazzy	159,345	No	No	No	0.2085	0.6835	0.3195
Ensemble	459,038	Yes	No	No	0.7139	0.7610	0.7367
MEDLINE	496,387	Yes	No	No	0.7506	0.7468	0.7487
Lexicon	558,353	Yes	Yes	Yes	0.8407	0.7842	0.8115

*

- B: Biomedical terms
- S: Supervised (manually verified)
- C: Comprehensive lexical information (abbreviations, acronyms, proper nouns, multiwords, etc.)

** Training set, nonword test

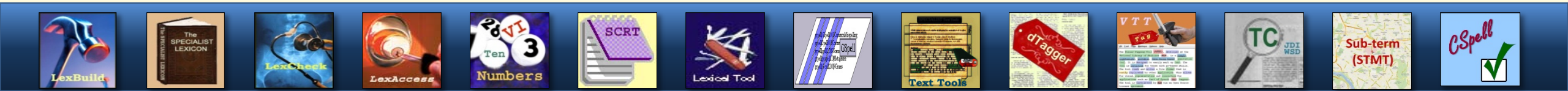


Corpus Selection

- Relevant (not size) corpus is better

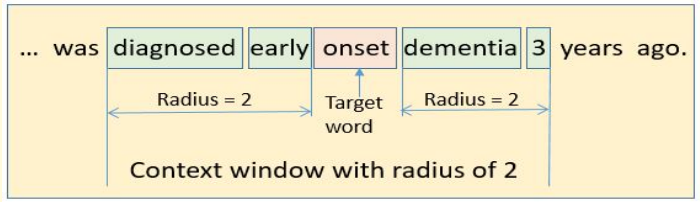
	Consumer Health Corpus	Medline N-gram Set
Resources	16 NIH public web sites	MEDLINE
Articles	17,139	26,759,399
Sentences	550,193	163,021,640
Tokens	10,228,699	3,386,661,350
Unique Word (LC)	109,818	496,388
Dictionary words	8.5886%	37.8507%
Precision	0.8407	0.8085
Recall	0.7842	0.7907
F1*	**0.8115**	**0.7995**

* Training set, nonword test



Context Window Size Selection

- Training model (CBOW):
  - window size = 5
- Application:
  - context radius = 2

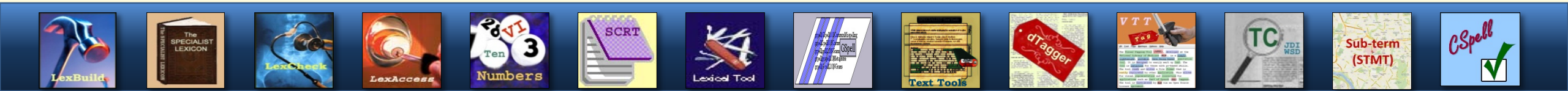


- Local context is more important than global context in spelling correction application

Context Radius	Precision	Recall	F1
1	0.8380	0.7817	0.8088
2	**0.8407**	**0.7842**	**0.8115**
3	0.8366	0.7804	0.8075
4	0.8352	0.7791	0.8061
5	0.8352	0.7791	0.8061
6	0.8296	0.7739	0.8008
…	…	…	…
10	0.8296	0.7739	0.8008
25	0.8283	0.7726	0.7995
50	0.8283	0.7726	0.7995
100	0.8283	0.7726	0.7995

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



Questions

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

☺