

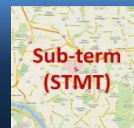
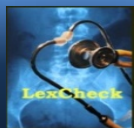
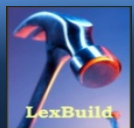
# NLP Group Meeting

By: Dr. Chris J. Lu  
2017.01.18

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

# Table of Contents

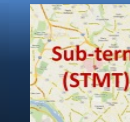
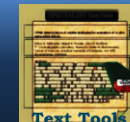
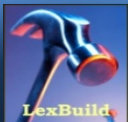
1. Overview
2. NLP Usage



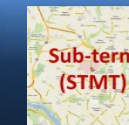
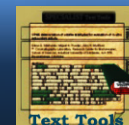
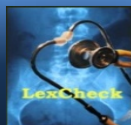
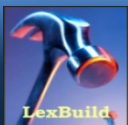
# 1. Overview: The SPECIALIST Lexicon and NLP Tools



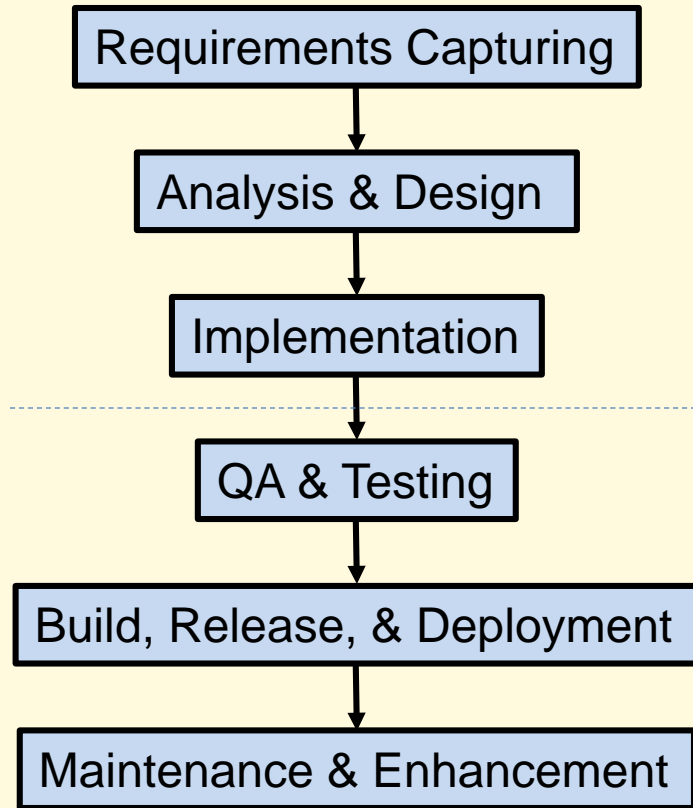
- Lexical Systems Group: <http://umlslex.nlm.nih.gov>
- SPECIALIST NLP Tools: <http://specialist.nlm.nih.gov>
- Intranet: <https://lexdev.nlm.nih.gov>



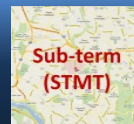
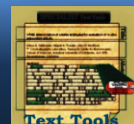
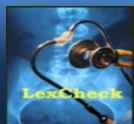
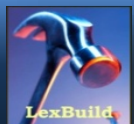
# The SPECIALIST NLP Tools



# Life Cycle Development & Task Components

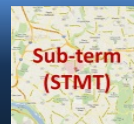
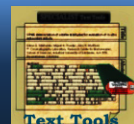
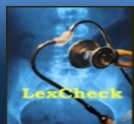
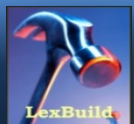


Task Components	Notes
SCR (Software Change Request)	Feature enhancements, bug fixes, library upgrades
Data updates & integration	Lexical Tools, LexAccess, STMT, MNS
Tests	All projects
Documentation	Design doc, user doc, API doc
Release package & backup	Annually & Semi-annually
Web-site & Web tools	All projects
Technical supports	All projects
Presentations	NLM summer Lectures, library associate talks, conferences
New research & development	Paper publications



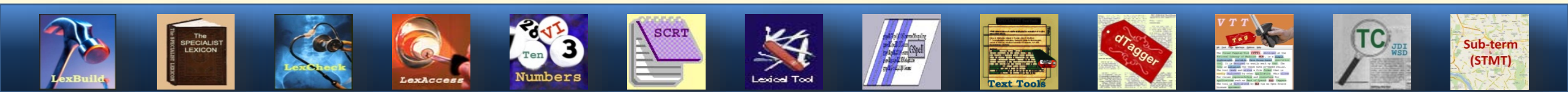
# Projects Maintenance Plan

Category	Descriptions	Projects (13)	Notes
1	Legacy Projects	gSpell, Text Tools, dTagger	Host project website
2	Completed Projects	Numbers, SCRT, VTT, TC	+Limited supports (approval required)
3	Supporting Projects	STMT	+Data updates (with UMLS release)
4	Core Supporting Projects	LB, LC	+Limited SCRs (approval required)
5	Core Projects	Lexicon, Lexical Tools, LexAccess	+Release & supports





# The SPECIALIST NLP Tools



# LexBuild Process (Computer-Aided)

## Sources:

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



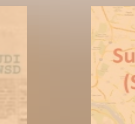
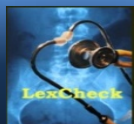
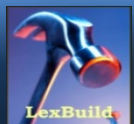
## Reviewed by lexicographers:

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



## Build:

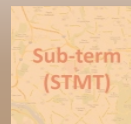
- **LexBuild**
- **LexAccess**
- **LexCheck**





# Lexical Records

```
{base=make
entry=E0038623
    cat=verb
    variants=irreg|make|makes|made|made|making|
    intran;part(off)
    intran;part(out)
    intran;part(up)
    tran=np
    tran=np|headway|
    tran=np|love|
    tran=np|mischief|
    tran=np|merry|
    tran=np|believe|
    tran=np|decision|
    tran=np|it|
    tran=np;part(out)
    tran=np;part(up)
    tran=np;part(over)
    tran=pphr(after,np)
    tran=pphr(with,np);part(away)
    ...
    nominalization=decisionmaking|noun|E0021045
    nominalization=lovemaking|noun|E0502721
    nominalization=makeup|noun|E0038625
}
```



# Lexical Records

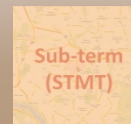
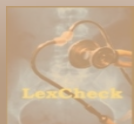
```
{base=herpes zoster
```

```
{base=zoster
```

```
{base=herpes  
entry=E0031440  
cat=noun  
variants=uncount  
}
```

```
herpes-zoster
```

```
count eg  
ncount
```



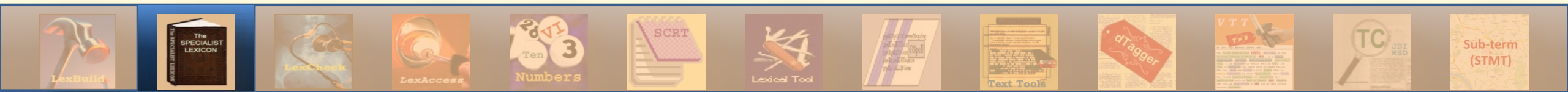
# Lexical Information from the Lexical Records

```
{base=herpes
entry=E0031440
  cat=noun
  variants=uncount
}
```

```
{base=zoster
entry=E0066013
  cat=noun
  variants=uncount
}
```

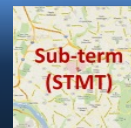
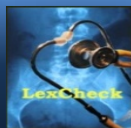
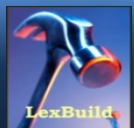
```
{base=herpes zoster
spelling_variant=herpes-zoster
entry=E0201295
  cat=noun
  variants=reg
  variants=uncount
}
```

Lexical Information   Base	herpes	zoster	herpes zoster
Part of speech	• noun	• noun	• noun
<b>Inflectional morphology</b> (inflections)	• herpes   1	• zoster   1	• herpes zoster   1 • <b>herpes zoster</b>   8
Orthography	• N/A	• N/A	• herpes-zoster
Abbreviation/Acronym	• N/A	• N/A	• N/A
Syntax (complementation)	• N/A	• N/A	• N/A
...	• ...	• ...	• ...
<b>Derivational morphology</b> (derivations)	• <b>herpetic</b>	• postzoster • post-zoster	• N/A
<b>LexSynonyms</b> (Element Synonymous words for Query Expansion)	• N/A	• singles • zona • ..	• singles • <b>zona</b> • ...



# What are Lexicon (Multi)Words?

	Descriptions	Examples
Lexicon words and LexMultiwords (LMWs)	<ul style="list-style-type: none"> <li>• Inflection morphology</li> <li>• POS</li> <li>• Specific meaning</li> <li>• Word order</li> <li>• <b>Space(s)</b></li> </ul>	<ul style="list-style-type: none"> <li>• herpes zoster</li> <li>• frog erythrocytic virus, cardiac surgery</li> <li>• hot dog</li> <li>• trial and error, exercise training, training exercise, up and down</li> <li>• ice cream (ice-cream)</li> </ul>
Non-Words (not in Lexicon)	<ul style="list-style-type: none"> <li>• (Single) Words does not exist by itself, only exist in multiwords</li> </ul>	<ul style="list-style-type: none"> <li>• non: <u>non</u> drug coated, persona <u>non</u> grata</li> <li>• vitro: in <u>vitro</u>, in <u>vitro</u> diagnostic</li> <li>• vivo: ex <u>vivo</u>, in <u>vivo</u> grown</li> <li>• intra: <u>intra</u> vires, <u>intra</u> articular route</li> <li>• etc.</li> </ul>

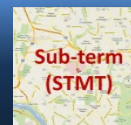
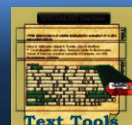
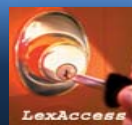
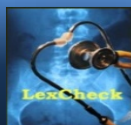
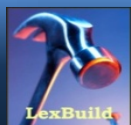


# LMW, MWE, UMLS String

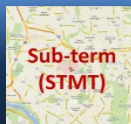
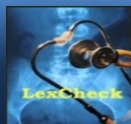
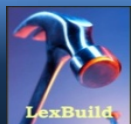
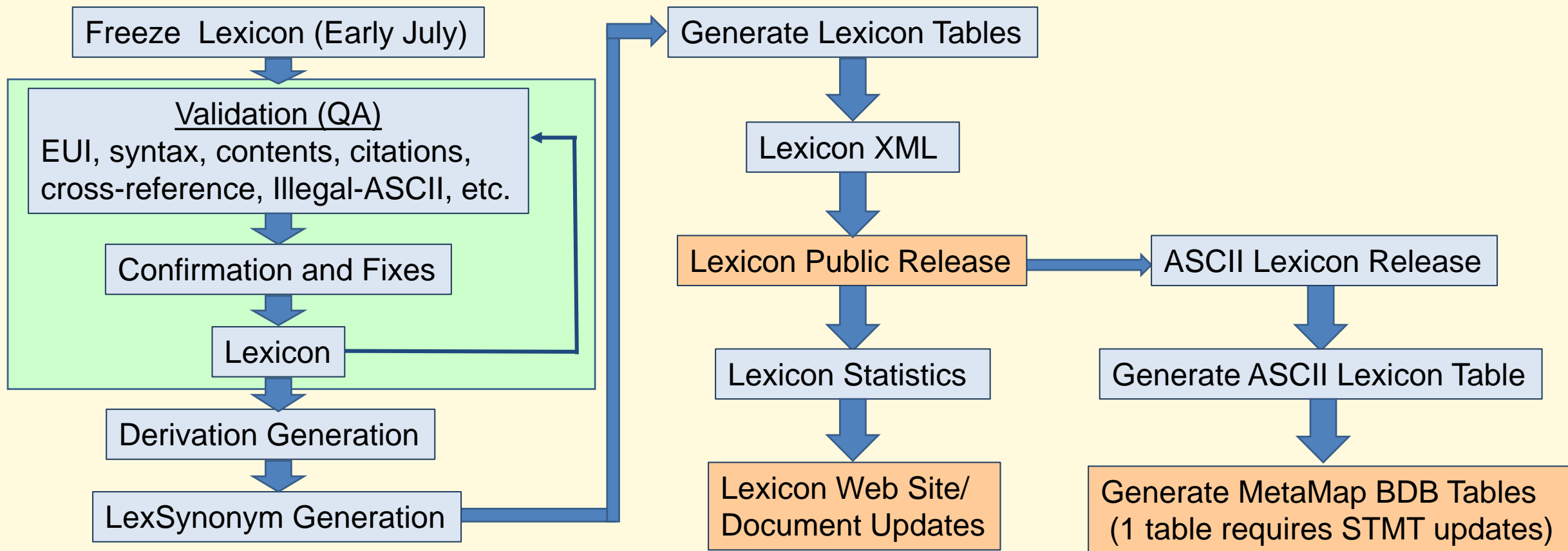
	Descriptions	Examples
LMWs and MWE (Overlap)	<ul style="list-style-type: none"> <li>Phrasal position</li> <li>Adverb, adjective</li> <li>Fixed phrases (non-decomposable)</li> </ul>	<ul style="list-style-type: none"> <li>because of, due to</li> <li>face down, in house</li> <li>kingdom come, by and large</li> </ul>
LMWs and MWE (Difference)	<ul style="list-style-type: none"> <li>Collocation</li> <li>Specific meaning</li> <li>Complementation</li> <li>Idioms</li> <li>...</li> </ul>	<ul style="list-style-type: none"> <li>undergoing <u>cardiac surgery</u></li> <li><u>in the house</u></li> <li><u>beat</u> someone <u>up</u>, <u>give birth</u></li> <li>kick the bucket, shoot the breeze</li> </ul>
UMLS String:	<ul style="list-style-type: none"> <li>Term<sup>1</sup> (conventionalized terminology)</li> <li>Phrase<sup>2</sup></li> </ul>	<ul style="list-style-type: none"> <li>Food and water, pain and fever, [disease, Hodgkin's]</li> <li>group 2, very low, may be a, not available</li> <li>right heart failure due to pulmonary hypertension</li> </ul>

1. A word or group of words with a specific meaning, especially in a particular field.

2. A phrase is a group of words that expressing a thought, but lacking a subject or a verb or both.



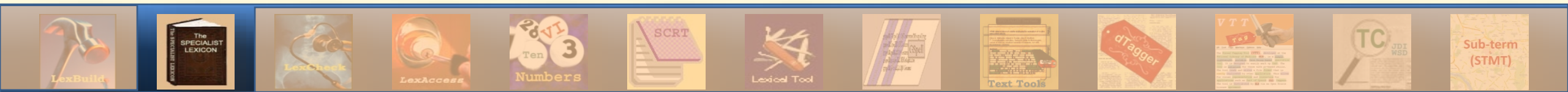
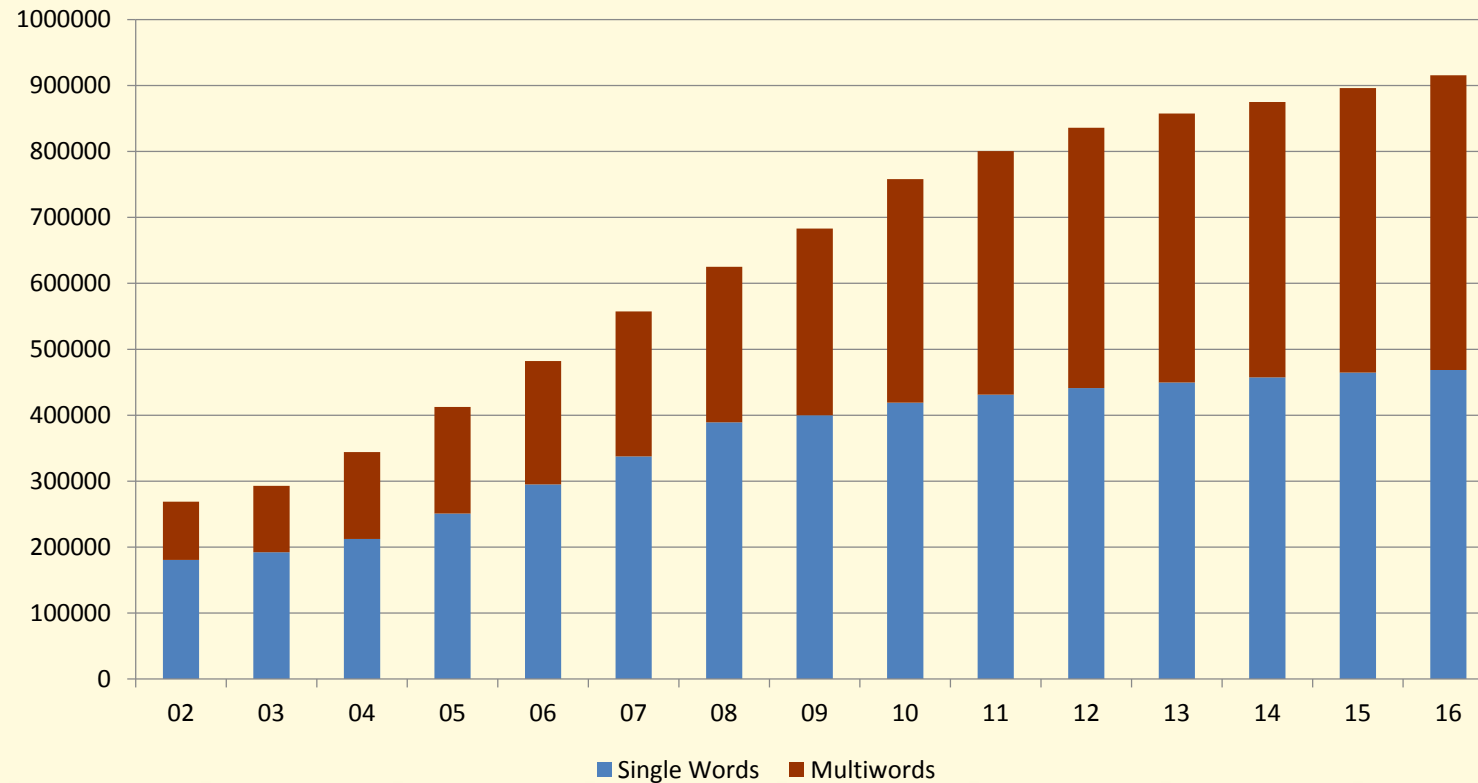
# Annual Lexicon Release Procedures





# Lexicon Growth – 2002 to 2016

- 491,639 lexical records
- 1,090,050 words (categories and inflections)
- 915,583 forms (spelling only)
  - Single words: 468,655 (51.19%); Multiwords: 446,928 (48.81%)

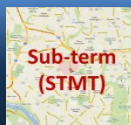
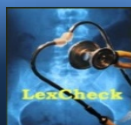


# Lexicon Unigram Coverage – Without WC

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

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

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

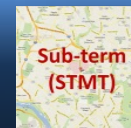
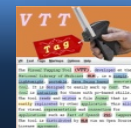
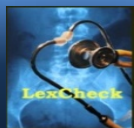
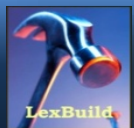


# Lexicon Unigram Coverage – With Frequency (WC)

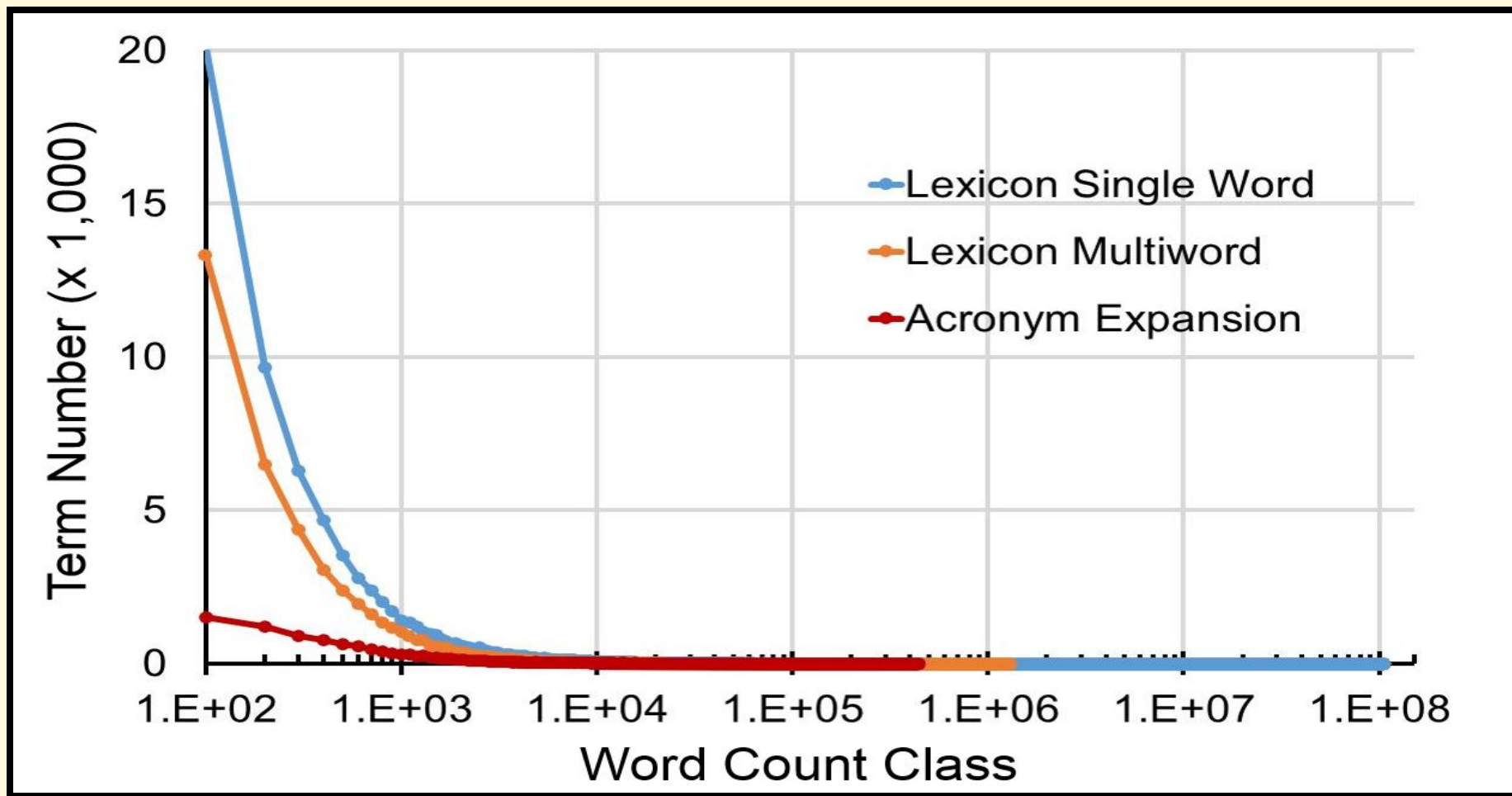
- Total word count for MEDLINE (2016): 3,114,617,940
- Lexicon covers > 98% unigrams from MEDLINE

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

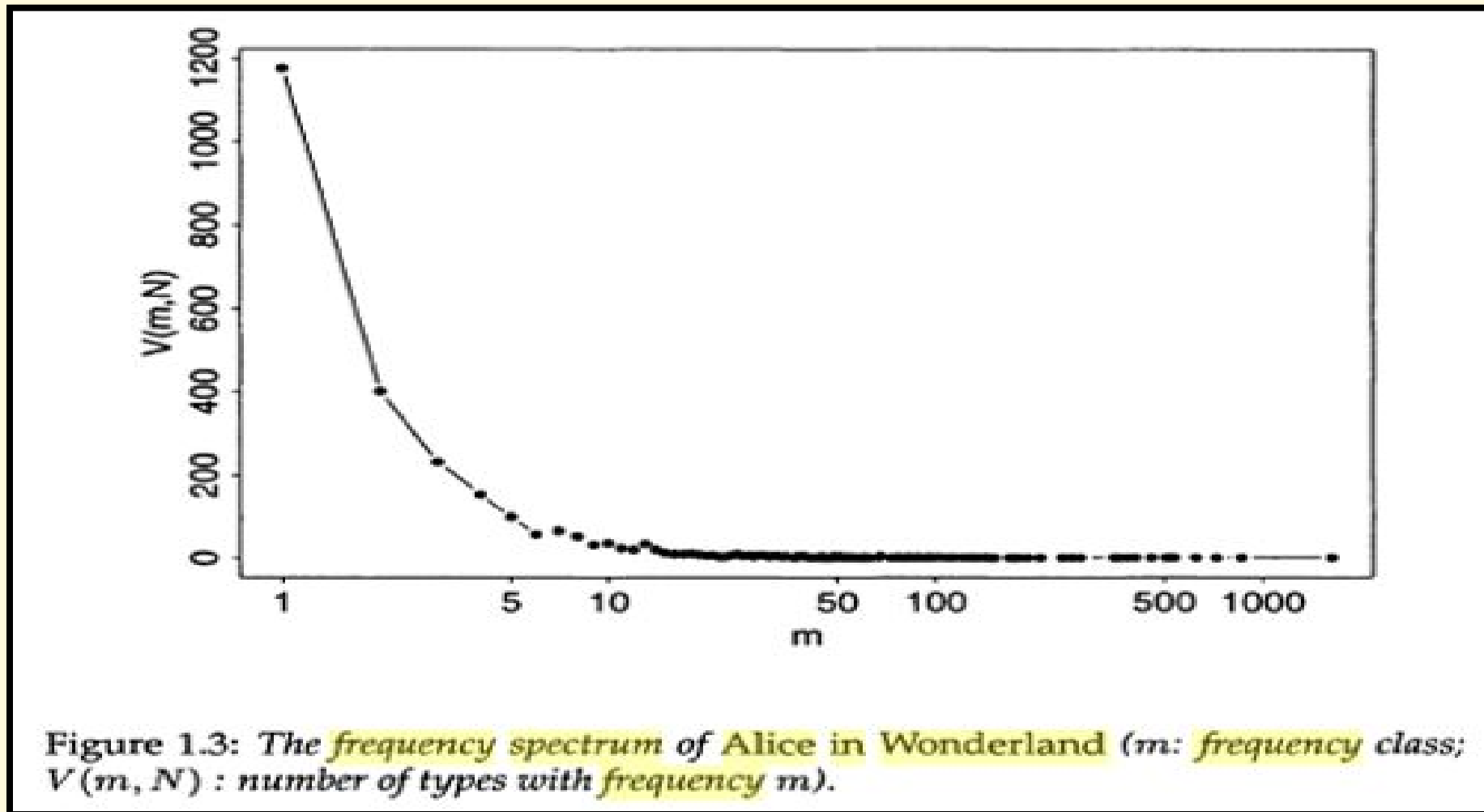
\* NON-WORD: a single word only exist in multiword, such as “non”, “vitro”, “vivo”, “intra”, etc.



# The Frequency Spectrum of Lexicon (Multi)words on MEDLINE

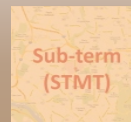
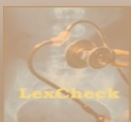


# The Frequency Spectrum of Alice in Wonderland



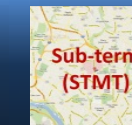
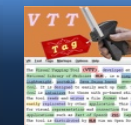
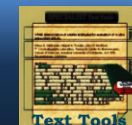
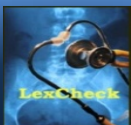
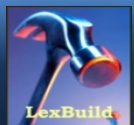
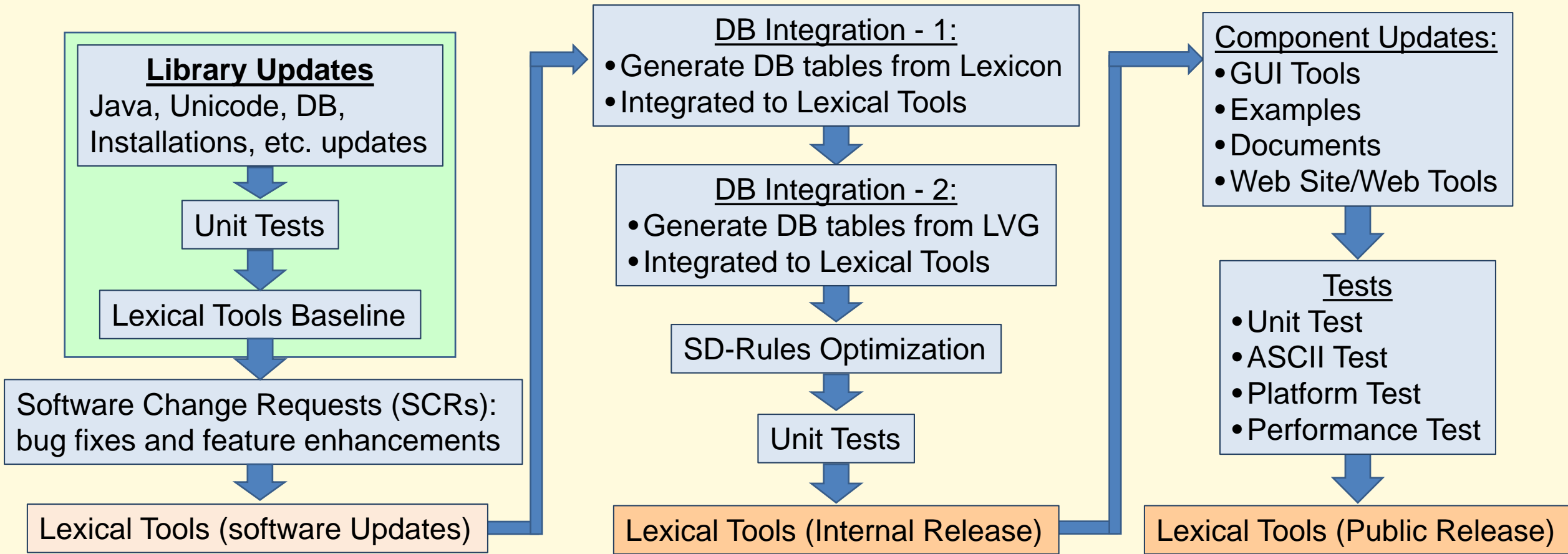
# Lexical Tools

- 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

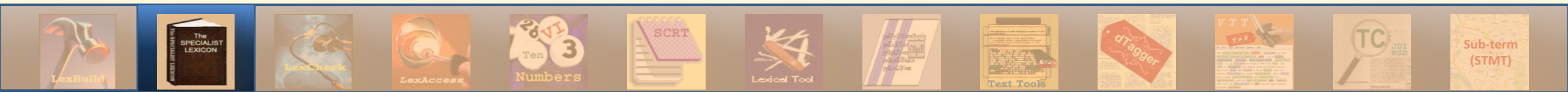
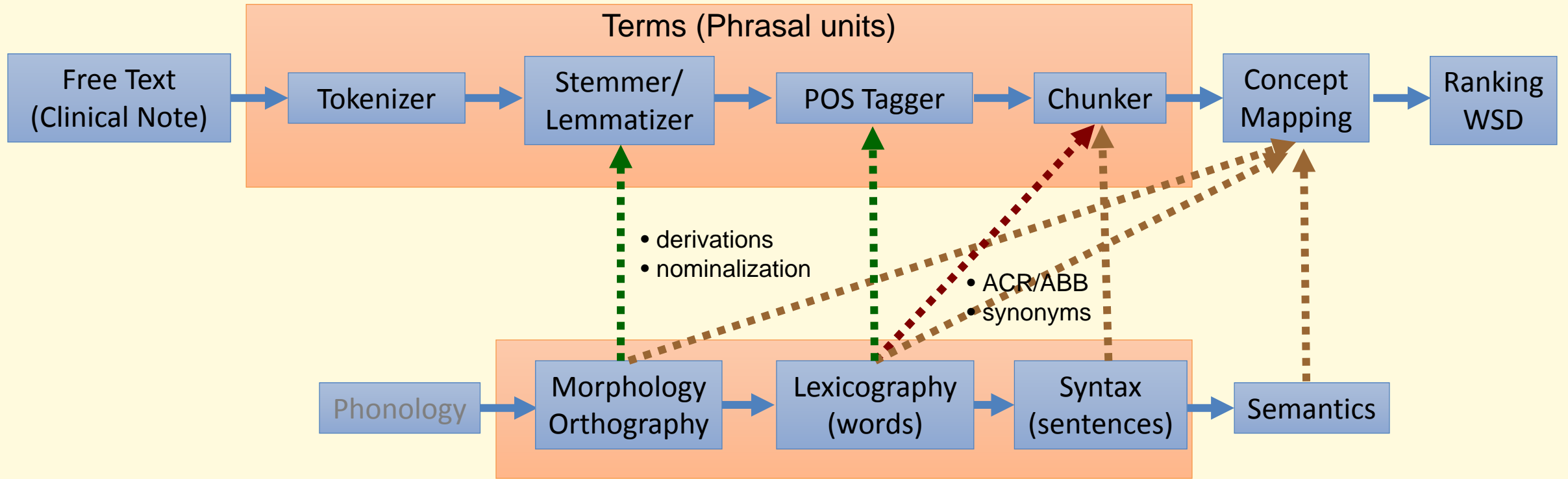




# Lexical Tools Release Procedures



# 2. NLP Usage



# Phrasal Level: Name Entity Recognition (The Longest Lead-Term)

➤ Example (PMID 23477346, TI):

• Follicular variant of papillary thyroid carcinoma is a unique clinical entity.

=> papillary thyroid carcinoma is a unique clinical

=> papillary thyroid carcinoma is a unique

=> papillary thyroid carcinoma is a

=> papillary thyroid carcinoma is

=> papillary thyroid carcinoma -> Match

=> is a unique clinical entity

=> is a unique clinical

=> is a unique

=> is a

=> is

=> a unique clinical entity

=>

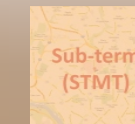
...  
=> papillary thyroid carcinoma is a unique clinical entity.

C0238463

C1710548

C0205210

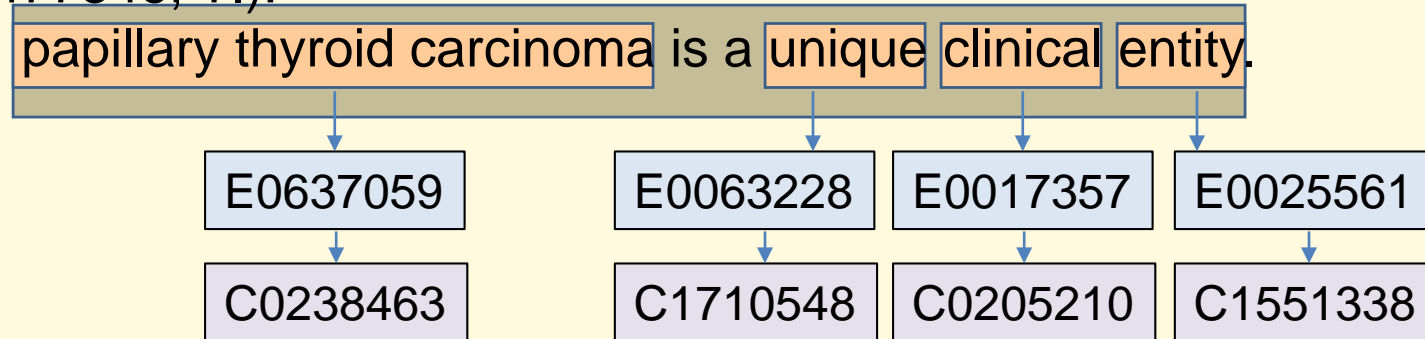
C1551338



# Phrasal Level - Name Entity Recognition

➤ Example (PMID 23477346, TI):

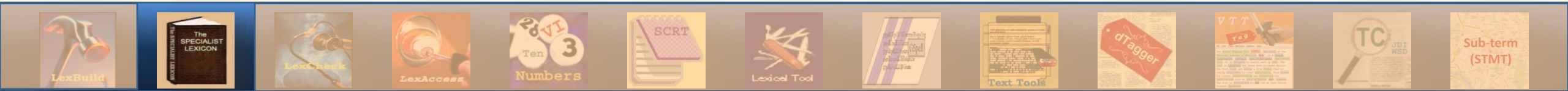
- Follicular variant of papillary thyroid carcinoma is a unique clinical entity.



➤ Lexicon Subterm Finder (LSF):

Find all subterms that have mapped CUI

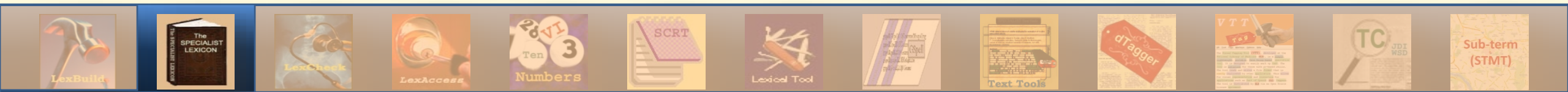
- Load Lexicon (multi)words to Trie (normalized, have mapped CUI).
- Retrieve subterms from Trie



# Processed Thesaurus: Lexicon (Multi)words - LSF

Normalized* InflVars	EUI	CUI	STI
...	...	...	...
carcinoma	E0015213	C0007097	T191   Neoplastic Process
clinical	E0017357	C0205210	T080   Qualitative Concept
entity	E0025561	C1551338	T071   Entity
papillary	E0045335	C0205312	T080   Qualitative Concept
papillary carcinoma	E0045337	C0007133	T191   Neoplastic Process
<b>papillary carcinomas</b>	E0045337	C0007133	T191   Neoplastic Process
<b>papillary plasma flow</b>	E0700280	No CUI	N/A
papillary thyroid carcinoma	E0637059	C0238463	T191   Neoplastic Process
thyroid	E0060948	C0040132	T023   Body Part, Organ, or Organ Component
unique	E0063228	C1710548	T080   Qualitative Concept
...			...

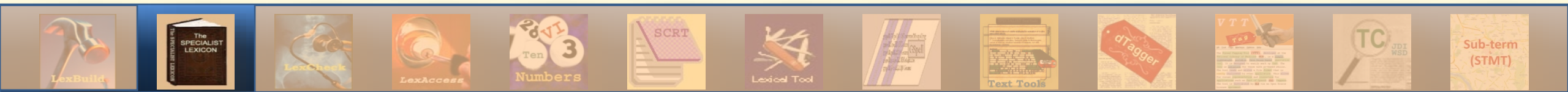
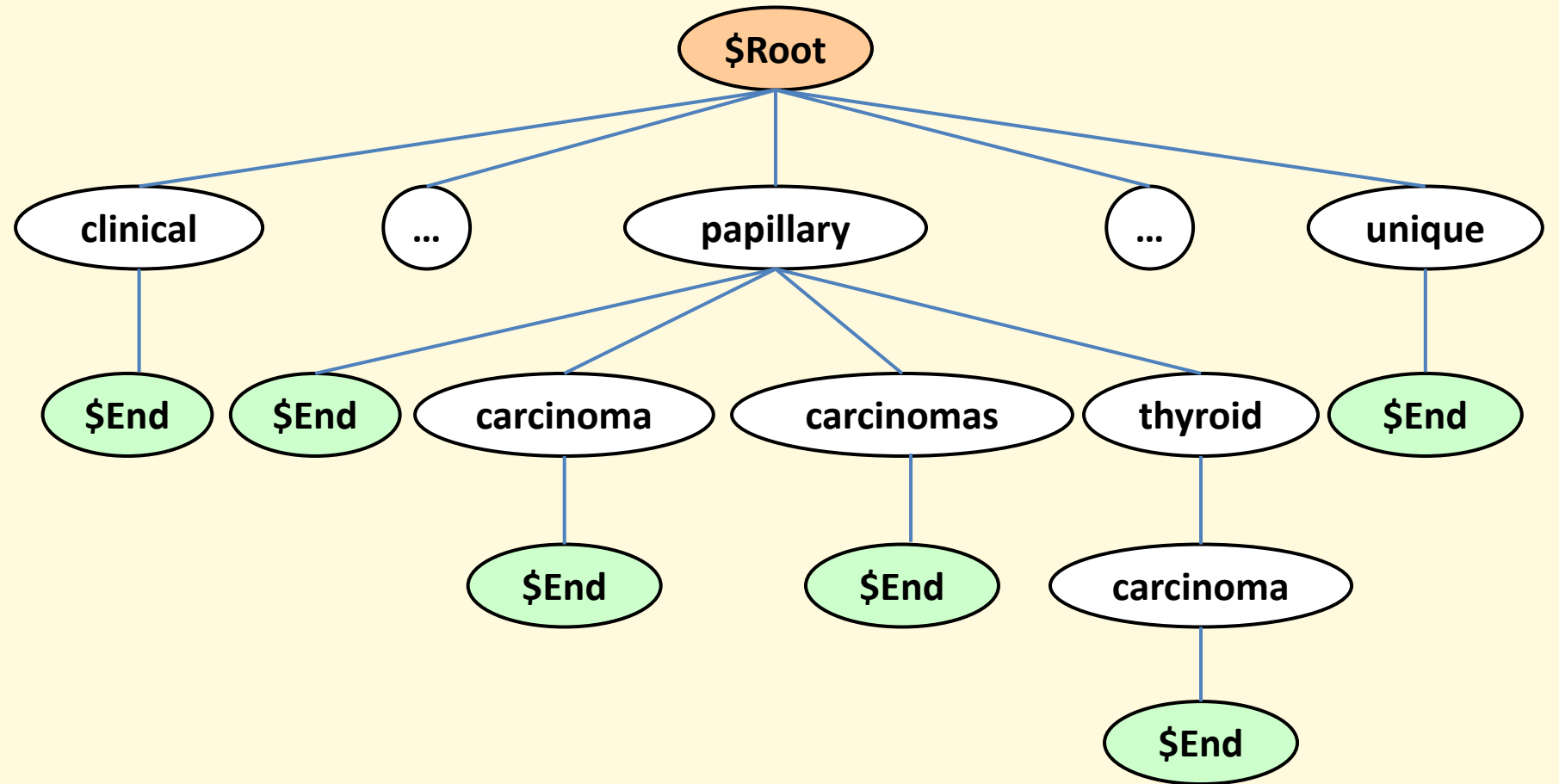
\* Norm: Remove genitive, remove parenthetical plural forms (s), (es), (ies), remove punctuation, lower cases, trim, remove duplicated results



# LSF – Load Lexicon (Multi)words to Trie

## Normalized InfiVars

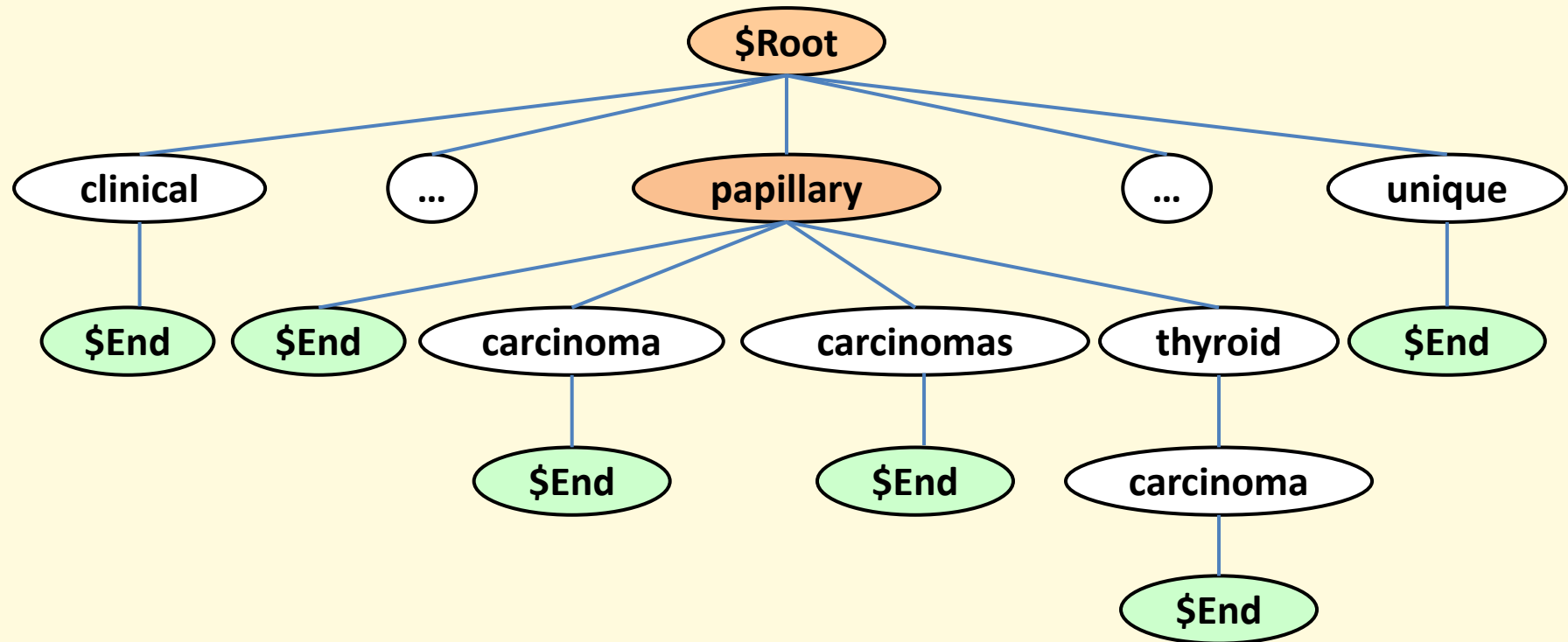
...
clinical
entity
papillary
papillary carcinoma
<b>papillary carcinomas</b>
<del>papillary plasma-flow</del>
papillary thyroid carcinoma
thyroid
unique
...





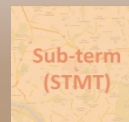
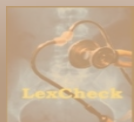
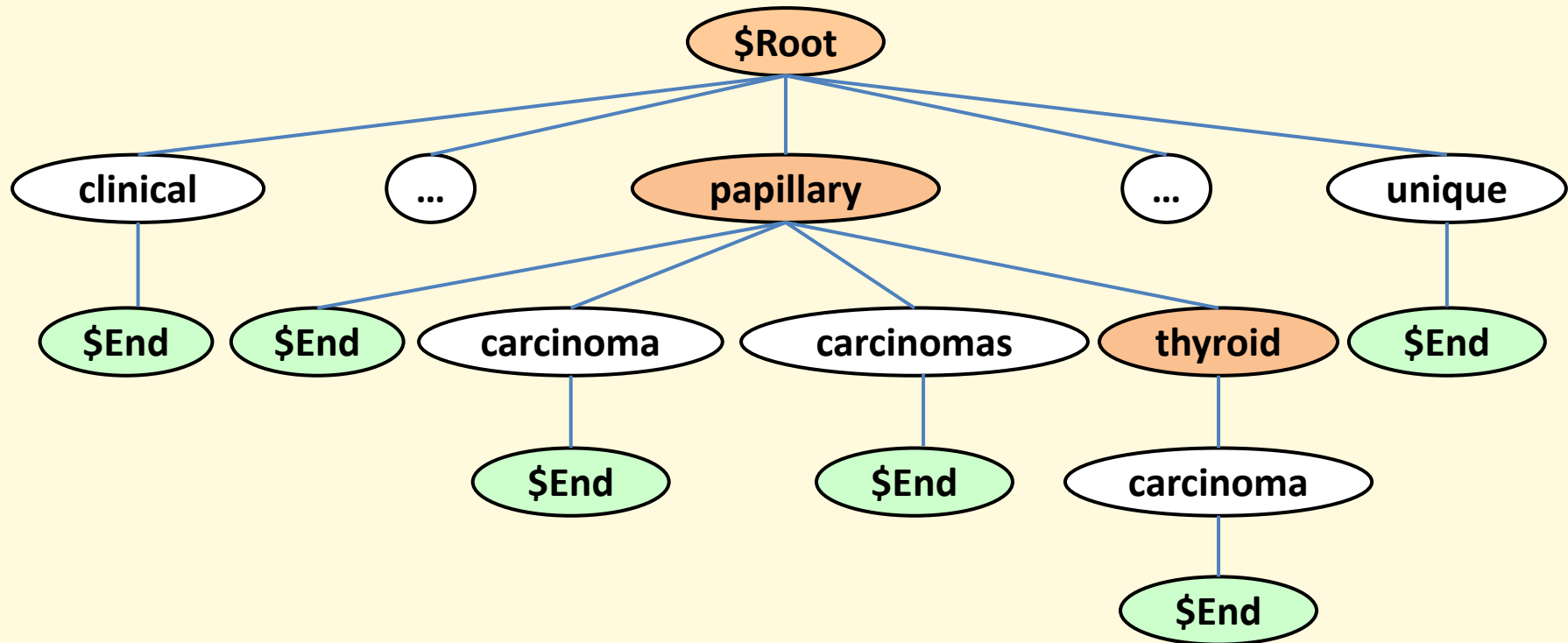
# LSF – Load Lexicon (Multi)words to Trie

- **Example:** papillary thyroid carcinoma is a unique clinical entity



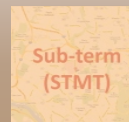
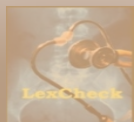
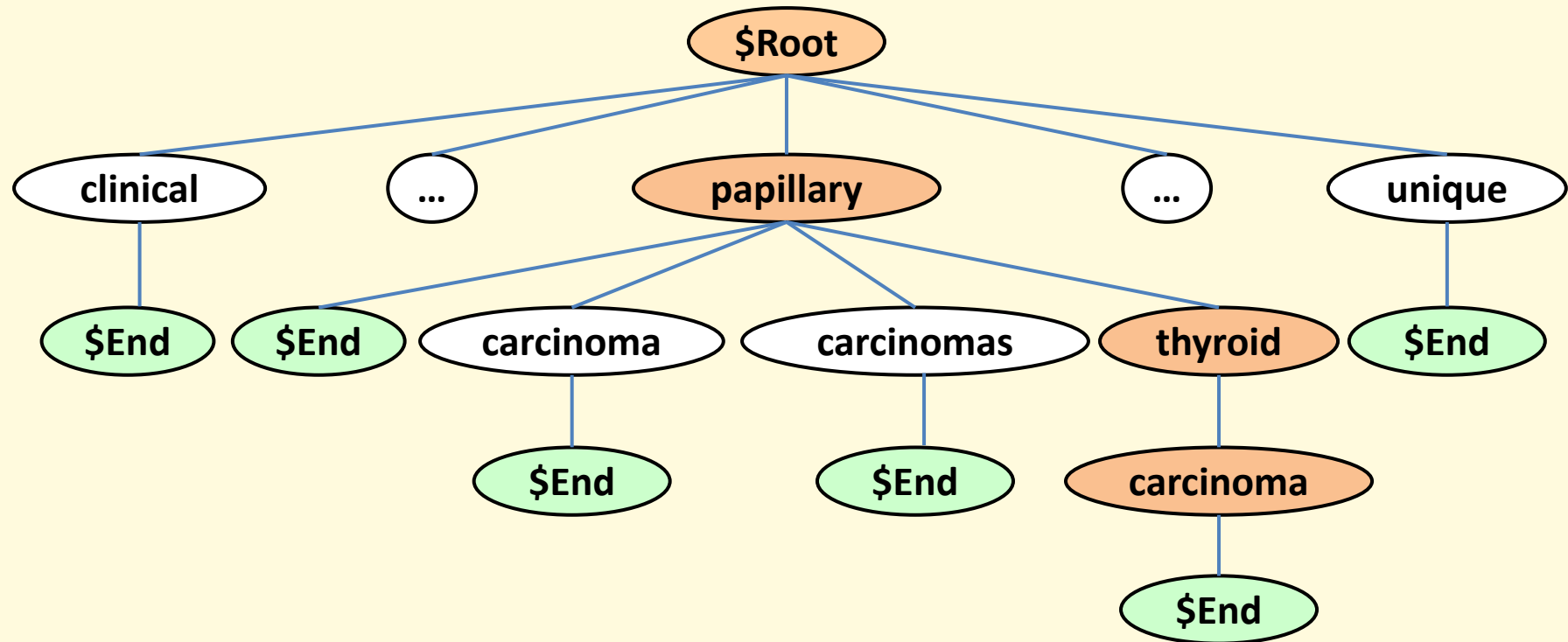
# LSF – Load Lexicon (Multi)words to Trie

- **Example:** papillary thyroid carcinoma is a unique clinical entity



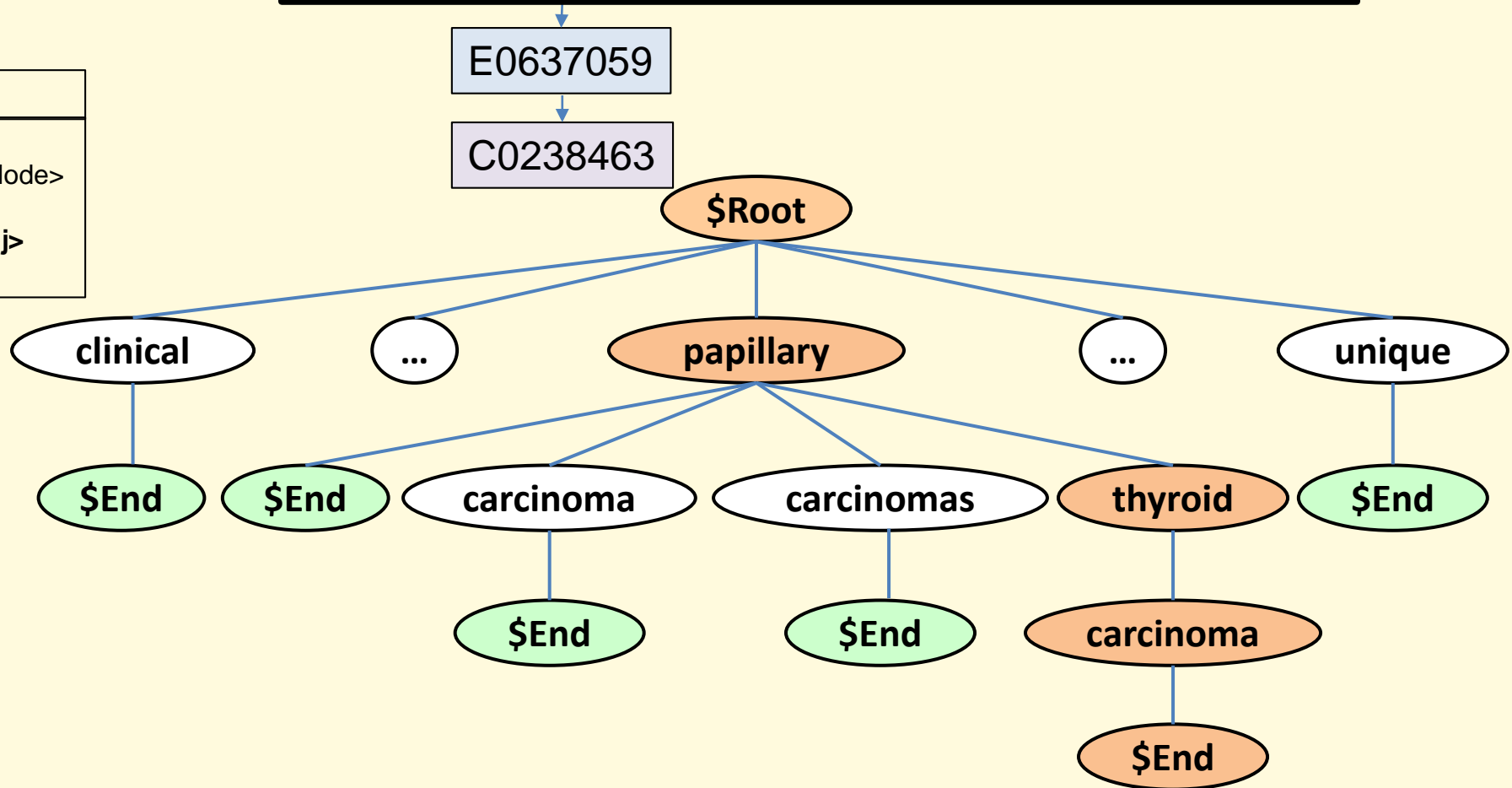
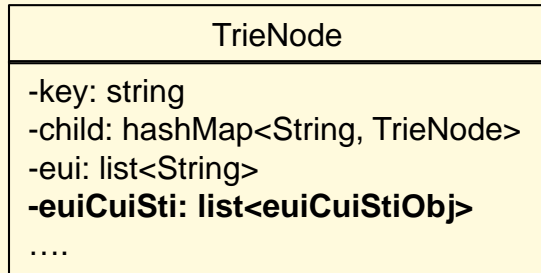
# LSF – Load Lexicon (Multi)words to Trie

- **Example:** papillary thyroid carcinoma is a unique clinical entity



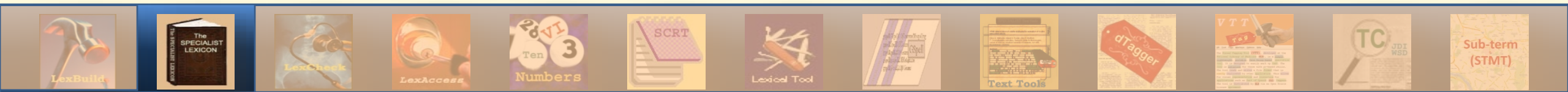
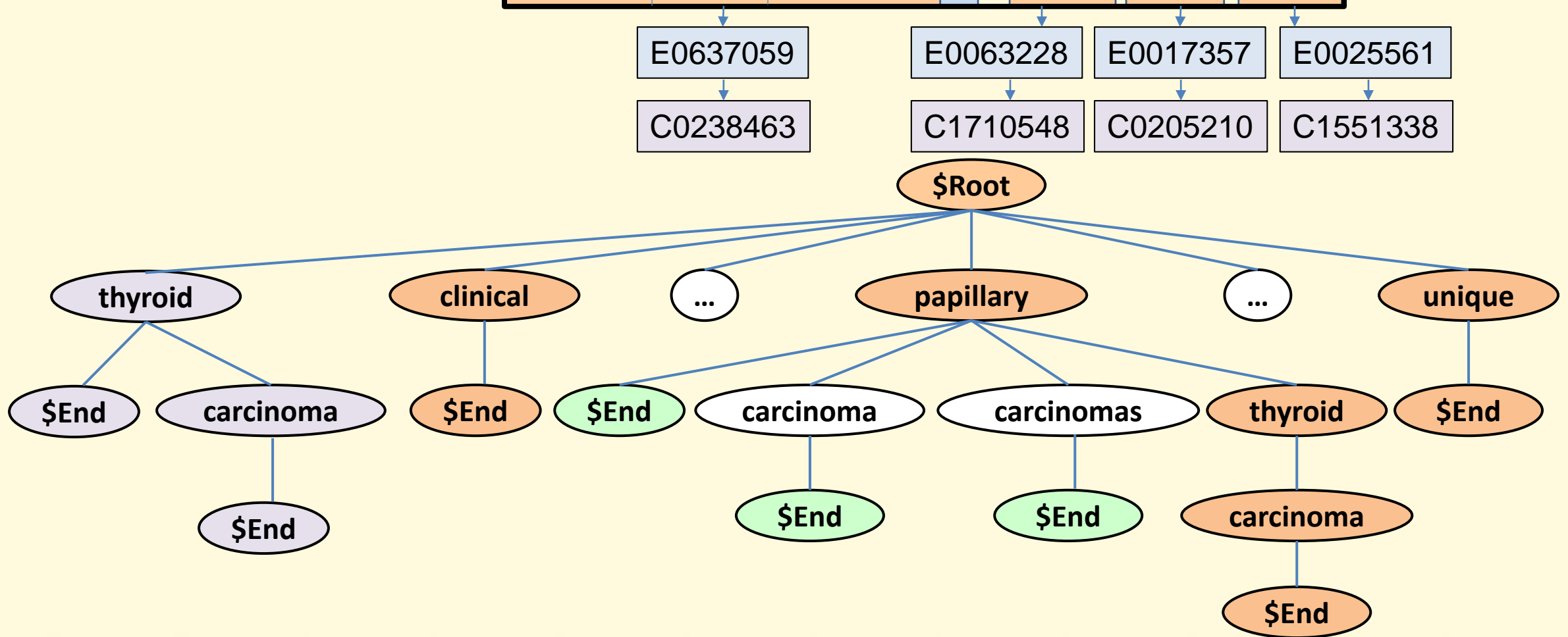
# LSF – Load Lexicon (Multi)words to Trie

- **Example:** papillary thyroid carcinoma is a unique clinical entity



# LSF – Load Lexicon (Multi)words to Trie

- **Example:** papillary thyroid carcinoma is a unique clinical entity



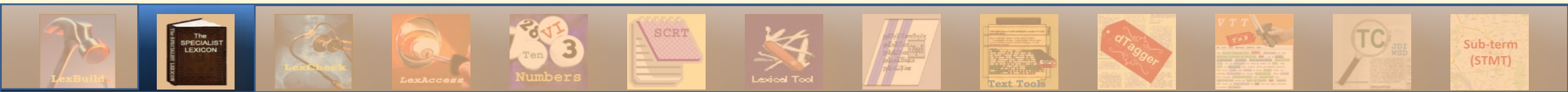
# Overlap Issues: Lead-Term, End-Term, Mid-Term?

➤ Example (PMID 12792778, TI):

- Inhibition of metastatic brain tumor growth by intramuscular administration of the endostatin gene.
  - Lead-Term: metastatic brain tumor growth: C0220650
  - End-Term: metastatic brain tumor growth: C0598934
  - LMWs: metastatic brain tumor|C0220650, brain tumor|C0006118, tumor growth|C0598934, ..

➤ Example (PMID 20162874, AB):

- In the present patient right pulmonary agenesis is co-occurring with VACTERL syndrome.
  - Lead-Term: the present patient right pulmonary agenesis: C0030706
  - End-Term: the present patient right pulmonary agenesis: C0265784
  - LMWs: patient right|C0030706, right pulmonary agenesis|C0265784, pulmonary agenesis|...





# POS Parser Issues → Sentence Level

➤ Example (PMID 9510650, TI):

- **Shallow Parser's (U of Illinois):**

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

NP PP NP VP PRT NP PP NP

- **Parser (Stanford):**

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

NP PP NP VP PRT NP SBAR NP

- Multiword Approach:

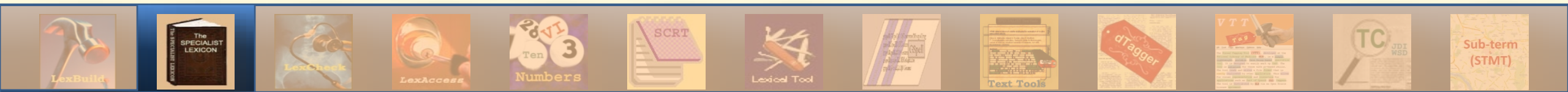
• The changes of tear break up time after myopic excimer laser photorefractive keratectomy

E0635418|Noun

E0764487|Noun

C2111106  
Lacrimation tear break-up time

C2069669  
excimer laser photorefractive keratectomy



# Order Issues (Norm)

➤ Example (PMID 5820369, TI):

- Cardiac arrest during **exercise training.**

E0566972|noun

C4279936|Exercise Training

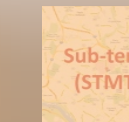
➤ Example (PMID 14719633, AB):

- Military **training exercises** are conducted routinely in the Mojave Desert.

E0764715|Noun

C4279936|Exercise Training

?



# Concept Ranking (CR) – The Longest Word?

➤ Example (PMID 9510650, TI):

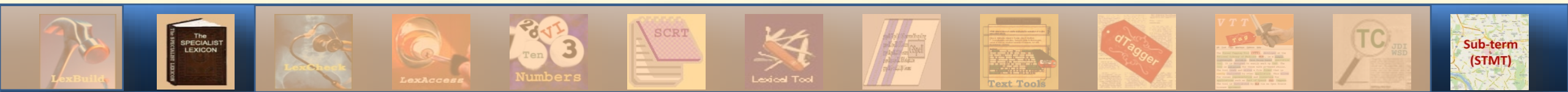
The changes of **tear break up time** after myopic excimer laser photorefractive keratectomy

Noun

sub-term	Lexicon - EUI
tear	E0060021, E0060022
break	E0013997, E0013998
up	E0063423, E0063424, ...
time	E0061086, E0061087
break up	E0220309
break up time	E0635415
tear break up time	E0635418

Noun

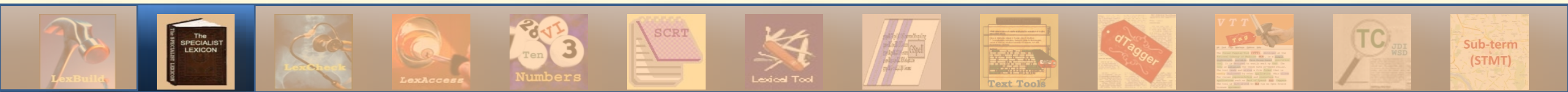
sub-term	Lexicon - EUI
myopic	E0041717, E0041718
excimer	E0319304
laser	E0036924, E0336817
photorefractive	E0418725
keratectomy	E0036428
excimer laser	E0514806
photorefractive keratectomy	E0225495
excimer laser photorefractive keratectomy	E0764487



# CR on Overlap case – The Longest Word?

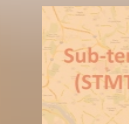
- Example (PMID 4771012, TI) – Overlap:
  - Early diagnosis and management of infected artificial heart valve.
  - Lead-Term: infected artificial heart valve.
  - End-Term: infected artificial heart valve.

sub-term	EUI	CUI
infected	E0034360, ..	C0439663
artificial	E0010589	C2004457
heart	E0030957	C0018787, ..
valve	E0063958	C0184252, ..
artificial heart	E0010602	C0018829
heart valve	E0030978	C0018826, ...
<b>artificial heart valve</b>	E0584205	C0018825, ...



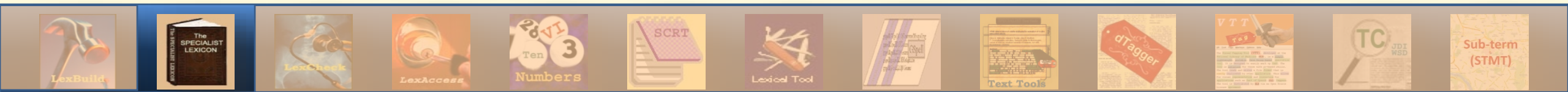
# CR: The Longest Word in the Sentence Level?

- Example (PMID 23477346, TI) – Beginning:
  - Follicular variant of papillary thyroid carcinoma is a unique clinical entity.
- Example (PMID 581461, TI) – Ending:
  - Nucleolar abnormalities in human papillary thyroid carcinoma.
- Example (PMID 6143549, AB) – Middle:
  - Coexisting papillary thyroid carcinoma occurred in three patients with HCA.



# Strings, MWEs, or Words

- Example of UMLS String (PMID 15528223, AB):
  - Right heart failure due to pulmonary hypertension causes significant morbidity and mortality.
    - UMLS String: right heart failure due to pulmonary hypertension|C1960038
    - ⇒ Search: “Right heart failure because of pulmonary hypertension”
    - LMW: right heart failure|C0235527, pulmonary hypertension|C0020542
  
- Example of MWE (PMID 23477346, TI):
  - Thirty patients undergoing cardiac surgery and 7 patients undergoing thoracic surgery not involving the heart were studied.
    - MWE: Undergoing cardiac surgery (MWE), no meaning, no POS, no morphology
    - LMW: cardiac surgery (MWE): C0018821
  
- Strings and MWEs do not have all 4 criteria of LMWs: POS, morphology, order, specific meaning

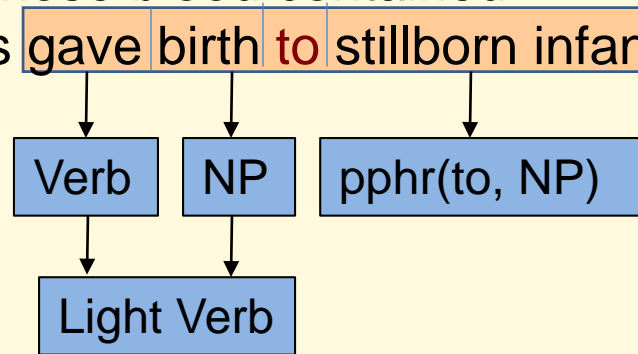




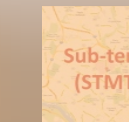
# Pattern: Complementation - Verb

➤ Example (PMID: 47945, AB):

Two **of** ten women whose blood contained  
Mycoplasma hominis **gave** birth **to** stillborn infants.



```
{base=give
entry=E0029785
  cat=verb
  variants=irreg|give|gives|gave|given|giving|
  intran
  intran;part(out)
  ...
  ditran=np|birth|,pphr(to,np)
  cplxtran=np,infcomp:objr
  cplxtran=np,infcomp:objr;part(out)
  cplxtran=np,ingcomp:arbc;part(over)
  nominalization=gift|noun|E0029737
}
```



# Multiwords, Patterns, StopWords

- Example (PMID: 47945, AB):

~~Two~~ ~~of~~ ~~ten~~ women ~~whose~~ blood contained Mycoplasma hominis gave birth to stillborn infants.

Number Pattern

Pronoun

Noun|E0341750

Light Verb

pphr(to, NP)

C0317838  
Mycoplasma hominis

C0005615  
Birth

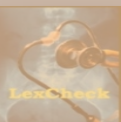
C0595939  
Stillbirth

- Multiword: “Mycoplasma hominis”

- Patterns:

- complementation: “gave birth”
- Measurement: (0.1-2.3 mg/day)
- title: Mr. Song, Dr. Coma
- ...

- StopWords: preposition, auxiliary, modal, determiner, conjunction, complementizer, pronoun, numbers, ...



# Summary

- Coverage:
  - How many multiwords do we need?
  - Query Expansion
  - Hybrid: use multiword as default, single word as supplement
- Other Resources (MWE? UMLS String?)
  - Linguistics concerns: POS, morphology, order, specific meaning
  - Technical concerns (size – performance):

	Size	Notes
<b>The Lexicon - InflVars</b>	0.9 M	Biomedical and general English
<b>UMLS Strings</b>	9.4 M	Terms and phrases, no POS, no morphology, ...
<b>UMLS Norm Terms</b>	11 M	+Word order is not preserved after Norm, morphology issues, ...
<b>WordNet</b>	0.15 M	General English, only 4 POS, no morphology, ...



# Conclusion

- Single word approach vs. multiword approach:
  - Data driven – embedded knowledge (Facts instead of Rules)
- The concept of approach (identify words), not the algorithm
  
- Other Components:
  - Morphology: inflections, uninflections, derivations
  - POS tagger (frog erythrocytic virus)
  - Norm (left pulmonary veins)
  - Query Expansion (zona vaccine)
  - Element Synonyms (zona -> herpes zoster)
  - Standard data set
  - etc.

