The SPECIALIST Lexicon and NLP Tools (Enhanced LexSynonym Acquisition and Features)

By: Dr. Chris J. Lu

**NLM – LHNCBC - CGSB**

Oct., 2017

Outline

- Introduction
  - The SPECIALIST Lexicon
  - The SPECIALIST NLP Tools (Lexical Tools)
- Applications - LexSynonyms
  - Natural Language Processing (NLP)
  - LexSynonyms
- Questions (anytime)
1. The SPECIALIST Lexicon

- A fancy synonym for “dictionary”
- A syntactic lexicon
- Biomedical and general English
- Over 490,000 records, 1M words (POS + forms)
- Designed/developed to provide the lexical information needed for the NLP (Natural Language Processing) system
- Distributed in the Unified Medical Language System (UMLS) Knowledge Sources by the National Library of Medicine (NLM)
LexBuild Process (Computer-Aided)

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

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

Build:
- LexBuild
- LexAccess
- LexCheck
Team of Lexicon Builders

- Dr. Alexa McCray, founded in 1994 (previous LHC Director, 2005-)
- Allen Browne, father of the SPECAILIST Lexicon (retired 2017)
- Dr. Dina Demner Fushman
- Dr. Chris J. Lu
- Dr. Lynn McCreedy
- Destinee Tormey
- Francois Lang
Lexicon Growth – 2002 to 2017

- 498,430 lexical records
- 1,110,321 words (categories and inflections)
- 935,276 forms (spelling only)
  - Single words: 472,608 (50.53%); Multiwords: 462,668 (49.47%)
Lexicon terms: single words and multiwords
- Space(s): ice-cream vs. ice cream

Four criteria for Lexicon terms:
- Part of Speech (POS):
  - tear break up time, frog erythrocytic virus, cardiac surgery
- Inflection morphology (uninflection):
  - left pulmonary veins (“left pulmonary vein” and “leave pulmonary vein”)
- Specific meaning:
  - hot dog (high temperature canine?)
- Word order:
  - trial and error, up and down (vs. food and water)
  - exercise training vs. training exercise (military)
Lexical Records - Information

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

Inflectional
- noun: book, books
- verb: categorize, categorizes, categorized, categorizing
- adj: red, redder reddest

Derivational
- example: transport
- suffix - transportation, transportable, transporter, ...
- prefix – autotransport, intratransport, pretransport, ...
- conversion (zero) - transport (verb), transport (noun)
Orthography (Spelling Variation)

- color | colour
- grey | gray
- align | aline
- Grave’s disease | Graves’s disease | Graves’ disease
- civilize | civilize
- harbor | harbor
- fetus | foetus | foetus
- 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 Lexical Records

<table>
<thead>
<tr>
<th>Lexical Information</th>
<th>Base</th>
<th>color</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part of speech</td>
<td></td>
<td><strong>noun</strong></td>
</tr>
<tr>
<td>Inflectional morphology (inflections)</td>
<td></td>
<td><strong>color</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>colors</strong></td>
</tr>
<tr>
<td>Orthography</td>
<td></td>
<td><strong>colour</strong></td>
</tr>
<tr>
<td>Abbreviation/Acronym</td>
<td></td>
<td><strong>N/A</strong></td>
</tr>
<tr>
<td>Syntax (complementation)</td>
<td></td>
<td><strong>N/A</strong></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td><strong>...</strong></td>
</tr>
<tr>
<td>Derivational morphology (derivations)</td>
<td></td>
<td><strong>colorable</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>colorful</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>colorize</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>colorist</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>...</strong></td>
</tr>
<tr>
<td>LexSynonyms</td>
<td></td>
<td><strong>chromatic</strong></td>
</tr>
</tbody>
</table>

{base=color
spelling_variant=colour
entry=E0017902
cat=noun
variants=uncount
variants=reg
}
UTF-8 (Since 2006)

{base=resume
  spelling_variant=résumé
  entry=E0053099
  cat=noun
  variants=reg
}

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

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

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

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

{base=Pécs
  entry=E0702889
  cat=noun
  variants=uncount proper
}
Lexicon Unigram Coverage – Without WC

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

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

* 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

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

* 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

![Graph showing the frequency spectrum of lexicon (multi)words on MEDLINE. The graph plots term number against word count class. The graph includes lines for Lexicon Single Word, Lexicon Multiword, and Acronym Expansion.]
The Frequency Spectrum of Alice in Wonderland

![Graph showing frequency spectrum](image)

*Figure 1.3: The frequency spectrum of Alice in Wonderland (m: frequency class; V(m, N): number of types with frequency m).*
Lexicon (Data) and Lexical Tools (Software)

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

→ spelling variant
→ part of speech
→ inflectional variant
→ chunker
→ derivational variant, synonym
2. 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
- inflVARS: generalises, generalised, generalising
- nominalization: generalisation, generalization
- Abbreviation/acronym: n/a

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

**Synonyms:** generalize

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

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

- 62 flow components
  - base form
  - spelling variants
  - inflectional variants
  - derivational variants
  - acronyms/abbreviations
  - …

- 34 options
  - input filter options (3)
  - global behavior options (12)
  - flow specific options (5)
  - output filter options (14)
## Lexical Tools – Flow Components (62)

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

leave → inflect → leaves, leaving, left
LVG Flow Component – Cmd line

```
> lvg -f:i
leave
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

<table>
<thead>
<tr>
<th>Input Term</th>
<th>Output Term</th>
<th>Inflections</th>
<th>Flow history</th>
<th>Flow Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>leave</td>
<td>leave</td>
<td>128</td>
<td>i</td>
<td>1</td>
</tr>
</tbody>
</table>
• Flow components can be arranged so that the output of one is the input to another.
A Serial Flow - Example

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

The Gougerot-Sjögren's Syndrome

The Gougerot-Sjögren's Syndrome

gougerotsjogren syndrome|2047|
16777215|l+q+g+t+p+w|1|
LVG - Parallel Flows

- Multiple flows can be defined
Parallel Flows - Example

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

color
color|color|2047|16777215|n|1|
color|color|128|1|B+s|2|
color|color|1024|1|B+s|2|
color|colour|128|1|B+s|2|
color|colour|1024|1|B+s|2|
```
Norm (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
**Norm**

<table>
<thead>
<tr>
<th>q0: map symbols to ASCII</th>
</tr>
</thead>
<tbody>
<tr>
<td>g: remove genitives</td>
</tr>
<tr>
<td>rs: remove parenthetic plural forms</td>
</tr>
<tr>
<td>o: replace punctuation with spaces</td>
</tr>
<tr>
<td>t: strip stop words</td>
</tr>
<tr>
<td>l: lowercase</td>
</tr>
<tr>
<td>B: uninflact each words in a term</td>
</tr>
<tr>
<td>Ct: retrieve citations</td>
</tr>
<tr>
<td>q7: Unicode core Norm</td>
</tr>
<tr>
<td>q8: strip or map Unicode to ASCII</td>
</tr>
<tr>
<td>w: sort words by order</td>
</tr>
</tbody>
</table>

“Fœtoproteins α’s, NOS“
**Norm**

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

"Fœtoproteins α’s, NOS“

"Fœtoproteins α’s, NOS"
<table>
<thead>
<tr>
<th>Norm</th>
<th>&quot;Fœtoproteins α’s, NOS“</th>
<th>&quot;Fœtoproteins α’s, NOS“</th>
</tr>
</thead>
<tbody>
<tr>
<td>q0: map symbols to ASCII</td>
<td>&quot;Fœtoproteins α’s, NOS“</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>g: remove genitives</td>
<td>&quot;Fœtoproteins α’s, NOS“</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>rs: remove parenthetic plural forms</td>
<td>&quot;Fœtoproteins α’s, NOS“</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>o: replace punctuation with spaces</td>
<td>&quot;Fœtoproteins α’s, NOS“</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>t: strip stop words</td>
<td>&quot;Fœtoproteins α’s, NOS“</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>l: lowercase</td>
<td>&quot;Fœtoproteins α’s, NOS“</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>B: uninflect each words in a term</td>
<td>&quot;Fœtoproteins α’s, NOS“</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>Ct: retrieve citations</td>
<td>&quot;Fœtoproteins α’s, NOS“</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>q7: Unicode core Norm</td>
<td>&quot;Fœtoproteins α’s, NOS“</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>q8: strip or map Unicode to ASCII</td>
<td>&quot;Fœtoproteins α’s, NOS“</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>w: sort words by order</td>
<td>&quot;Fœtoproteins α’s, NOS“</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
</tbody>
</table>
### Norm

<table>
<thead>
<tr>
<th>q0: map symbols to ASCII</th>
<th>&quot;Fœtoproteins α’s, NOS“</th>
</tr>
</thead>
<tbody>
<tr>
<td>g: remove genitives</td>
<td>&quot;Fœtoproteins α’s, NOS&quot;</td>
</tr>
<tr>
<td>rs: remove parenthetic plural forms</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>o: replace punctuation with spaces</td>
<td></td>
</tr>
<tr>
<td>t: strip stop words</td>
<td></td>
</tr>
<tr>
<td>l: lowercase</td>
<td></td>
</tr>
<tr>
<td>B: uninflect each words in a term</td>
<td></td>
</tr>
<tr>
<td>Ct: retrieve citations</td>
<td></td>
</tr>
<tr>
<td>q7: Unicode core Norm</td>
<td></td>
</tr>
<tr>
<td>q8: strip or map Unicode to ASCII</td>
<td></td>
</tr>
<tr>
<td>w: sort words by order</td>
<td></td>
</tr>
</tbody>
</table>

---

**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**
<table>
<thead>
<tr>
<th>q0: map symbols to ASCII</th>
<th>“Fœtoproteins α’s, NOS“</th>
</tr>
</thead>
<tbody>
<tr>
<td>g: remove genitives</td>
<td>&quot;Fœtoproteins α's, NOS&quot;</td>
</tr>
<tr>
<td>rs: remove parenthetic plural forms</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>o: replace punctuation with spaces</td>
<td>Fœtoproteins α NOS</td>
</tr>
<tr>
<td>t: strip stop words</td>
<td></td>
</tr>
<tr>
<td>l: lowercase</td>
<td></td>
</tr>
<tr>
<td>B: uninflect each words in a term</td>
<td></td>
</tr>
<tr>
<td>Ct: retrieve citations</td>
<td></td>
</tr>
<tr>
<td>q7: Unicode core Norm</td>
<td></td>
</tr>
<tr>
<td>q8: strip or map Unicode to ASCII</td>
<td></td>
</tr>
<tr>
<td>w: sort words by order</td>
<td></td>
</tr>
</tbody>
</table>
Norm

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

"Fœtoproteins α’s, NOS"
"Fœtoproteins α’s, NOS"
"Fœtoproteins α, NOS"
"Fœtoproteins α, NOS"
Fœtoproteins α NOS
Fœtoproteins α
<table>
<thead>
<tr>
<th>q0: map symbols to ASCII</th>
<th>“Fœtoproteins α’s, NOS“</th>
</tr>
</thead>
<tbody>
<tr>
<td>g: remove genitives</td>
<td>&quot;Fœtoproteins α’s, NOS&quot;</td>
</tr>
<tr>
<td>rs: remove parenthetic plural forms</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>o: replace punctuation with spaces</td>
<td>Fœtoproteins α NOS</td>
</tr>
<tr>
<td>t: strip stop words</td>
<td>Fœtoproteins α</td>
</tr>
<tr>
<td>l: lowercase</td>
<td>fœtoproteins α</td>
</tr>
<tr>
<td>B: uninflect each words in a term</td>
<td></td>
</tr>
<tr>
<td>Ct: retrieve citations</td>
<td></td>
</tr>
<tr>
<td>q7: Unicode core Norm</td>
<td></td>
</tr>
<tr>
<td>q8: strip or map Unicode to ASCII</td>
<td></td>
</tr>
<tr>
<td>w: sort words by order</td>
<td></td>
</tr>
<tr>
<td>Norm</td>
<td>“Fœtoproteins α’s, NOS“</td>
</tr>
<tr>
<td>------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>q0: map symbols to ASCII</td>
<td>&quot;Fœtoproteins α’s, NOS&quot;</td>
</tr>
<tr>
<td>g: remove genitives</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>rs: remove parenthetic plural forms</td>
<td>Fœtoproteins α NOS</td>
</tr>
<tr>
<td>o: replace punctuation with spaces</td>
<td>Fœtoproteins α</td>
</tr>
<tr>
<td>t: strip stop words</td>
<td>føtoproteins α</td>
</tr>
<tr>
<td>l: lowercase</td>
<td>føtoprotein α</td>
</tr>
<tr>
<td>B: uninflect each words in a term</td>
<td></td>
</tr>
<tr>
<td>Ct: retrieve citations</td>
<td></td>
</tr>
<tr>
<td>q7: Unicode core Norm</td>
<td></td>
</tr>
<tr>
<td>q8: strip or map Unicode to ASCII</td>
<td></td>
</tr>
<tr>
<td>w: sort words by order</td>
<td></td>
</tr>
<tr>
<td>q0: map symbols to ASCII</td>
<td>“Fœtoproteins α’s, NOS”</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>g: remove genitives</td>
<td>&quot;Fœtoproteins α’s, NOS&quot;</td>
</tr>
<tr>
<td>rs: remove parenthetic plural forms</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>o: replace punctuation with spaces</td>
<td>Fœtoproteins α NOS</td>
</tr>
<tr>
<td>t: strip stop words</td>
<td>Fœtoproteins α</td>
</tr>
<tr>
<td>l: lowercase</td>
<td>føtoproteins α</td>
</tr>
<tr>
<td>B: uninflect each words in a term</td>
<td>føetoprotein α</td>
</tr>
<tr>
<td>Ct: retrieve citations</td>
<td>fetoprotein α</td>
</tr>
<tr>
<td>q7: Unicode core Norm</td>
<td></td>
</tr>
<tr>
<td>q8: strip or map Unicode to ASCII</td>
<td></td>
</tr>
<tr>
<td>w: sort words by order</td>
<td></td>
</tr>
</tbody>
</table>
Norm

q0: map symbols to ASCII
"Fœtoproteins α’s, NOS"
"Fœtoproteins α’s, NOS"
"Fœtoproteins α, NOS"
"Fœtoproteins α, NOS"
Fœtoproteins α NOS
Fœtoproteins α
fœtoproteins α
fœtoprotein α
fetoprotein α
q7: Unicode core Norm
fetoprotein α
q8: strip or map Unicode to ASCII
w: sort words by order
<table>
<thead>
<tr>
<th>q0: map symbols to ASCII</th>
<th>“Fœtoproteins α’s, NOS“</th>
</tr>
</thead>
<tbody>
<tr>
<td>g: remove genitives</td>
<td>&quot;Fœtoproteins α's, NOS&quot;</td>
</tr>
<tr>
<td>rs: remove parenthetic plural forms</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>o: replace punctuation with spaces</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>t: strip stop words</td>
<td>Fœtoproteins α NOS</td>
</tr>
<tr>
<td>l: lowercase</td>
<td>Fœtoproteins α</td>
</tr>
<tr>
<td>B: uninflect each words in a term</td>
<td>føtoproteins α</td>
</tr>
<tr>
<td>Ct: retrieve citations</td>
<td>føtoprotein α</td>
</tr>
<tr>
<td>q7: Unicode core Norm</td>
<td>fetoprotein α</td>
</tr>
<tr>
<td>q8: strip or map Unicode to ASCII</td>
<td>fetoprotein alpha</td>
</tr>
<tr>
<td>w: sort words by order</td>
<td></td>
</tr>
<tr>
<td>Norm</td>
<td>&quot;Fœtoproteins α's, NOS“</td>
</tr>
<tr>
<td>------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>q0: map symbols to ASCII</td>
<td>&quot;Fœtoproteins α's, NOS&quot;</td>
</tr>
<tr>
<td>g: remove genitives</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>rs: remove parenthetic plural forms</td>
<td>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>o: replace punctuation with spaces</td>
<td>Fœtoproteins α NOS</td>
</tr>
<tr>
<td>t: strip stop words</td>
<td>Fœtoproteins α</td>
</tr>
<tr>
<td>l: lowercase</td>
<td>foetoproteins α</td>
</tr>
<tr>
<td>B: uninflect each words in a term</td>
<td>foetoprotein α</td>
</tr>
<tr>
<td>Ct: retrieve citations</td>
<td>fetoprotein α</td>
</tr>
<tr>
<td>q7: Unicode core Norm</td>
<td>fetoprotein α</td>
</tr>
<tr>
<td>q8: strip or map Unicode to ASCII</td>
<td>fetoprotein alpha</td>
</tr>
<tr>
<td>w: sort words by order</td>
<td>alpha fetoprotein</td>
</tr>
</tbody>
</table>
alpha fetoprotein
alpha Fetoproteins
alpha-Fetoprotein
alpha-Fetoproteins
Alpha fetoproteins
alpha fetoprotein
alpha Foetoprotein
alpha foetoprotein
alpha fetoproteins
Alpha-fetoprotein
alpha-fetoprotein
Alpha Fetoproteins
Alpha-Fetoprotein
Alpha-fetoprotein NOS
Alpha Fetoprotein
alpha-fetoprotein
ALPHA-FETOPROTEIN
Alpha Foetoprotein
…
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 broad 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 1:** Map terms to concepts (meaning)
- **Challenge 2:** many to many mapping

<table>
<thead>
<tr>
<th>Terms</th>
<th>Concepts</th>
<th>NLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>• cold</td>
<td>• Cold Temperature</td>
<td>C0009264</td>
</tr>
<tr>
<td>• Cold Temperature</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Cold Temperatures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Cold (Temperature)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Temperatures, Cold</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Low temperature</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• low temperatures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• cold</td>
<td>• Cold Temperature</td>
<td>C0009264</td>
</tr>
<tr>
<td>• Common Cold</td>
<td>C0009443</td>
<td>• Cool Temperature</td>
</tr>
<tr>
<td>• Cold Therapy</td>
<td>C0010412</td>
<td>• Cold Sensation</td>
</tr>
<tr>
<td>• Cold Sensation</td>
<td>C0234192</td>
<td>• ...</td>
</tr>
</tbody>
</table>
NLP Pipe Line – Lexical Information

Terms (Phrasal units)

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

Phonology → Morphology → Orthography → Lexicography (words) → Syntax (terms) → Semantics

Lexical Information

• derivations
• nominalization
• ACR/ABB
• synonyms
The SPECIALIST NLP Tools

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

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

POSS tagger:
• Assign part of speech to a single word or multiword in a text
• To increase precision

Others...
<table>
<thead>
<tr>
<th>Behçet’s Diseases, NOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behçet’s Diseases, NOS</td>
</tr>
<tr>
<td>Behçet Diseases, NOS</td>
</tr>
<tr>
<td>Behçet Diseases, NOS</td>
</tr>
<tr>
<td>Behçet Diseases NOS</td>
</tr>
<tr>
<td>Behçet Diseases</td>
</tr>
<tr>
<td>behçet diseases</td>
</tr>
<tr>
<td>behçet disease</td>
</tr>
<tr>
<td>behçet disease</td>
</tr>
<tr>
<td>behçet disease</td>
</tr>
<tr>
<td>behçet disease</td>
</tr>
</tbody>
</table>

| q0: map Unicode 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 non-ASCII char |

| w: sort words by order |
NLP – Norm (Pre-Process Lexical Variations)

Terms in Corpus

- behcet disease
- behçet disease
- behcet diseases
- behçet diseases
- behcet's disease
- behçet's disease
- behçet’s disease
- behçet's disease
- behcets disease
- behcets disease
- behcet's disease, nos
- disease, behçet
- diseases, behçet
- ...

Indexed Database

Normalized String

- C0004943
- Behcet Syndrome
NLP – Norm (Cont.)

Query → norm → Normed Term → Behcet disease

Results that matches the normalized query:
- C0004943
- Behcet Syndrome

Indexed Database
Normalized String

MRXNS_ENG.RRF

SQL
UMLS Metathesaurus

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

Diagram:
- Terms in Corpus
- Normalize
- Index
- Indexed Database
  - Normalized String
NLP – Query Expansion (derivation)

Indexed Database Normalized String

- perforated ear drum*
- drum ear perforate
- None

- perforation ear drum
- drum ear perforation
- C0206504
  Tympanic Membrane Perforation

* PMID: 13114832, 5992689, ..
NLP – Query Expansion (Synonym)

Indexed Database Normalized String

calcaneal fracture*
Norm
calcaneal fracture
Norm
bone fracture heel

None

C0281926
Fracture of calcaneus

C0006655:
- calcaneal
- heel bone

* PMID: 1118604, 1165396, ..
<table>
<thead>
<tr>
<th>UMLS Synonymy (C0281926)</th>
</tr>
</thead>
<tbody>
<tr>
<td>calcaneus fracture</td>
</tr>
<tr>
<td>calcaneus fractures</td>
</tr>
<tr>
<td>calcaneus; fracture</td>
</tr>
<tr>
<td>fracture calcaneus</td>
</tr>
<tr>
<td>fracture heel</td>
</tr>
<tr>
<td>fracture of calcaneus</td>
</tr>
<tr>
<td>fracture of calcaneus (diagnosis)</td>
</tr>
<tr>
<td>fracture of calcaneus (disorder)</td>
</tr>
<tr>
<td>fracture of os calcis</td>
</tr>
<tr>
<td>fracture; calcaneus</td>
</tr>
<tr>
<td>fracture; heel bone</td>
</tr>
<tr>
<td>fracture; os calcis</td>
</tr>
<tr>
<td>fracture; calcaneus</td>
</tr>
<tr>
<td>fractured calcaneus</td>
</tr>
<tr>
<td>fractured os calcis</td>
</tr>
<tr>
<td>fractures heel</td>
</tr>
<tr>
<td>heel bone</td>
</tr>
<tr>
<td>heel bone fracture</td>
</tr>
<tr>
<td>heel bone; fracture</td>
</tr>
<tr>
<td>heel fracture</td>
</tr>
<tr>
<td>of calcaneus fracture</td>
</tr>
<tr>
<td>os calcis</td>
</tr>
<tr>
<td>os calcis fracture</td>
</tr>
<tr>
<td>os calcis; fracture</td>
</tr>
</tbody>
</table>
UMLS Synonymy – Expanded Terms

calcaneal fractures ➔ heel bone fractures

Norm

bone fracture heel

Indexed Database
Normalized String

C0281926: Fracture of calcaneus

[UMLS Synonymy]
Expanded Terms for Concept Mapping:
Grouped by Normalization

C0281926:
- Key | calcaneus fracture
  - fractured calcaneus
  - fracture; calcaneus
  - fracture of calcaneus
  - calcaneus fracture
  - calcaneus fractures
  - calcaneus; fracture
- Key | bone fracture heel
  - heel bone fracture
  - heel bone; fracture
  - fracture; heel bone
  - ...
<table>
<thead>
<tr>
<th>UMLS Synonym to Element Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>calcaneus fracture</td>
</tr>
<tr>
<td>calcaneus fractures</td>
</tr>
<tr>
<td>calcaneus; fracture</td>
</tr>
<tr>
<td>fracture calcaneus</td>
</tr>
<tr>
<td>fracture heel</td>
</tr>
<tr>
<td>fracture of calcaneus</td>
</tr>
<tr>
<td>fracture of calcaneus (diagnosis)</td>
</tr>
<tr>
<td>fracture of calcaneus (disorder)</td>
</tr>
<tr>
<td>fracture of os calcis</td>
</tr>
<tr>
<td>fracture; calcaneus</td>
</tr>
<tr>
<td>fracture; heel bone</td>
</tr>
<tr>
<td>fracture; os calcis</td>
</tr>
<tr>
<td>fractured calcaneus</td>
</tr>
<tr>
<td>fractured os calcis</td>
</tr>
<tr>
<td>fractures heel</td>
</tr>
<tr>
<td>heel bone</td>
</tr>
<tr>
<td>heel bone fracture</td>
</tr>
<tr>
<td>heel bone; fracture</td>
</tr>
<tr>
<td>heel fracture</td>
</tr>
<tr>
<td>heel fracture</td>
</tr>
<tr>
<td>of calcaneus fracture</td>
</tr>
<tr>
<td>os calcis</td>
</tr>
<tr>
<td>os calcis fracture</td>
</tr>
<tr>
<td>os calcis; fracture</td>
</tr>
</tbody>
</table>

**Norm:** **calcaneus fracture**

- **calcaneus** fracture
- calcaneus fractures
- calcaneus; fracture
- fracture calcaneus
- fracture of calcaneus (diagnosis)
- fracture of calcaneus (disorder)
- fracture of os calcis
- fracture; calcaneus
- fracture; heel bone
- fracture; calcaneus
- fractured calcaneus
- fractured os calcis
- fractures heel
- heel bone
- heel bone fracture
- heel bone; fracture
- heel fracture
- of calcaneus fracture
- os calcis
- os calcis fracture
- os calcis; fracture

**Norm:** **bone fracture heel**

- **heel bone** fracture
- fracture; heel bone
- heel bone; fracture

**Norm:** **calcis fracture os**

- fracture of **os calcis**
- fracture; os calcis
- fractured os calcis
- os calcis fracture
- os calcis; fracture

**Other element Synonym**

- **calcaneal** fracture – PMID: 1194000, 471457, …
- **calcaneum** fracture – PMID: 13288374, 5550125, …

**C0006655:**
- calcaneal
- calcaneum
- calcaneus
- heel bone
- os calcis
- …
Element Synonyms

[Element Synonym] Subterms Substitution

calcaneal fractures

heel bone fractures

Norm

bone fracture heel

Indexed Database Normalized String

C0281926 Fracture of calcaneus

[sPair: calcaneal|heel bone]

C0006655:
• calcaneal
• heel bone
• calcaneus
• …

[UMLS Synonyms] Expanded Terms for Concept Mapping: Normalization

C0281926:
• Key | calcaneus fracture
  ▪ fractured calcaneus
  ▪ fracture; calcaneus
  ▪ fracture of calcaneus
  ▪ calcaneus fracture
  ▪ calcaneus fractures
  ▪ calcaneus; fracture
• Key | bone fracture heel
  ▪ heel bone fracture
  ▪ heel bone; fracture
  ▪ fracture; heel bone
• …
Multiple Substitutions

Indexed Database Normalized String

pneumonia **due to** other **virus***

due other pneumonia virus

None

pneumonia **by** other **viral**

other pneumonia viral

C0348677
other viral pneumonia

* VA14760, HA480.80, ..
Recursive Substitutions

chromosomal aberrance* → chromosomal aberrant → chromosomal aberration

E0006478:
- aberrant
- aberration
- aberrance
- aberrancy

C0443127

Indexed Database Normalized String

C0008625
Chromosome Aberrations

* PMID: 11172638, 25543836, ..
Real-time Model

Tokenization & NER
- Documents
- Paragraphs
- Sentences
- Phrases
- Terms
- Tokens (words)
- NER
- ...

Free Text

STMT

Norm Term

UMLS - Indexed Database Normalized Term

CUI

WSD

Yes

No

Ranking

Same LexRecord

Related LexRecords

• Subterm Substitution (synonyms, derivations, etc.)
Pre-Processing Model

Terms in Corpus

Texture Variations
- Spelling variants
- Inflectional Variants
- Synonyms
- Derivations
...

Norm

Indexed Database
Normalized String

calcaneal fracture

Norm

calcaneal fracture

Enhanced UMLS
Indexed Database
Normalized String

C0281926
Fracture of calcaneus
4. LexSynonym - Element Synonyms

- The key for subterm substitutions (data of synonyms) depends on the completeness and quality of both element synonyms for a given UMLS synonym thesaurus.

- Synonym Related Data:
  - Element Synonyms (for expanded terms)
  - UMLS Synonym thesaurus (for concept mapping)

- Completeness: recall
- Quality: precision
Synonym Sets

- UMLS Synonyms (13M)
- The SPECIALIST Lexicon Synonyms, 2016- (~5K)
- Others
  - UMLS-Core Projects (~12K)
  - Synonym set by Randy Miller, (~15K)
  - dictionary.com, thesaurus.com,
  - WordNet (https://wordnet.princeton.edu)
  - etc..
Element Synonyms - UMLS Synonyms

- Applied restrictions: source vocabulary (MeSH), term length, size of grams (1), etc..

- Issues:
  - Quantity (over-generated):
    - Example: [C0013182, Drug Allergy], “allergy drug” and “allergy medicine” (expanded terms)
    - Slow performance (if use all expanded terms for element synonyms)
  - Quality:
    - Not necessary cognitive synonyms (commutativity and transitivity)
    - Broader or narrower concept, acronyms, abbreviations, POS ambiguity, multiple CUIs, etc..
  - Single words or multiwords
    - Example: [C0281926, Fracture of calcaneus], “calcaneal fracture” and “heel bone fracture”
    - How many grams?
Element Synonyms – Lexicon Synonyms

- Developed in early 90's
- The original idea is to provide synonyms that are not in the UMLS Metathesaurus
  - not a complete data set
- Quantity: manually updated by user’s requests (static):
  - 2004 (5,056) -> 2016 (5,198)
  - Only 142 sPairs were added since 2004
  - Need an automatic/systematic way to generate synonyms
- Quality: not necessary good sPairs
- 6 associated flow components (10%): G, Ge, Gn, r, v, y
LexSynonyms – Objectives

➢ To establish a system to generate a standalone set of generic element synonyms (sPairs) for effective UMLS concept mapping

• Scope:
  o include all synonymous terms in Lexicon (LexSynonyms)
  o grow with the SPECIALIST Lexicon
  o a thorough set of element synonyms (to increase recall)

• Feature requirements:
  o **better performance: increase recall and preserve precision**
  o resolve known issues (near-synonyms, POS ambiguity, include multiword synonyms, etc.)
  o cognitive synonyms (to preserve precision)
Enhanced Requirements

- Element synonyms for subterm substitution
- R1: Cognitive synonyms (not near-synonyms)
- R2: POS (meaning shift)
- R3: Source: CUI (UMLS) and other source information
- R4: Expansions of abbreviations and acronyms
- R5: Word level (single POS): single words and multiwords
- ...
R1: Cognitive Synonym (Quality)

- Two properties:
  - **Commutativity**: \((x = y) \rightarrow (y = x)\)
    - joy|noun|enjoy|verb -> enjoy|verb|joy|noun
    - bi-directional (sPair)
  - **Transitivity**: \(((x = y) \text{ and } (y = z)) \rightarrow (x = z)\)
    - enjoy|verb -> joy|noun -> happy|adj
    - multiple (recursive) substitutions
    - sClass (synonym class)
- Prevent precision issues by near-synonyms.
Synonym Types

- Cognitive synonym:
  - less difference
  - greater interchangeability (not context-sensitive)
  - more generic
  - can be represented as a synonym pair (sPair)

- Near-synonym:
  - greater difference
  - less interchangeability
  - specific use, can’t used in generic case
### Near-Synonyms

<table>
<thead>
<tr>
<th>CUI</th>
<th>Preferred Term</th>
<th>Synonym</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0000869</td>
<td>Acacia</td>
<td>locust tree</td>
<td>Though both the acacia &amp; locust tree are members of Leguminosae (pea, bean), they do seem to refer to different trees.</td>
</tr>
<tr>
<td>C0003353</td>
<td>Antigua</td>
<td>Anguilla</td>
<td>The islands of Antigua &amp; Anguilla are both in the West Indies, but are not the same place.</td>
</tr>
<tr>
<td>C0032639</td>
<td>Pons</td>
<td>metencephalon</td>
<td>The metencephalon, per unabridged.merriam-webster.com includes the cerebellum and pons, and is different from the pons</td>
</tr>
</tbody>
</table>
Acacia & Locust tree

- C0000869

[Image of Acacia tree]

[Image of Locust Tree]
Anguilla & Antigua

C0003353
**Metencephalon & Pontine Structure (Pons)**

- C0032639

---

**Hinbrain: Metencephalon**

- **b) metencephalon**
  - **pons**
    - Contains pneumotaxic centre which fine tunes breathing rate
    - Relays information between cerebellum and cerebrum
  - **cerebellum**
    - Feedback center for execution of motor movements
    - Controls posture and balance
  - **reticular formation**
    - Nuclei diffusely located through the brainstem*
    - Regulates wakefulness and muscle tone

*the term “brainstem” refers to the medulla oblongata, pons, and the midbrain*
# R2: POS Issues – Meaning Shift

<table>
<thead>
<tr>
<th>CUI</th>
<th>Preferred Term</th>
<th>synonym</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0004063</td>
<td>Assault</td>
<td>mug</td>
<td>The noun mug means a large cup, while the verb mug does refer to assault.</td>
</tr>
<tr>
<td>C0001774</td>
<td>Agaricales</td>
<td>Mushroom</td>
<td>The verb (to) mushroom means increase, spread, or develop rapidly. It does not refer to Agaricales while the noun is a synonym.</td>
</tr>
<tr>
<td>C0003459</td>
<td>Anura</td>
<td>frog</td>
<td>The verb (to) frog means hunt for or catch frogs. It does not refer to Anura, while the noun is a synonym.</td>
</tr>
<tr>
<td>C0003842</td>
<td>Arteries</td>
<td>arterial</td>
<td>The noun arterial refers to roads, not circulatory anatomy, unlike the adjective arterial.</td>
</tr>
</tbody>
</table>
POS: Assault & Mug

mug|verb
(assault)

≠

mug|noun
(a large cup)
The patient **expired** 1 day later.

Pressure of CO2 in **expired** air ...

Disposal of **expired** drug ...

---

**CUI: C0011065**

PT: Cessation of life

died
dead
death
deceased
...

**CUI: C0231800**

PT: Expiration, Function

exhaled

expiratory

expiration

...

**CUI: C1704631**

PT: Expiration

expire

expiration

...
R4: Acronym/Abbreviation Issues – Precision

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

<table>
<thead>
<tr>
<th>CUI</th>
<th>Preferred Term</th>
<th>synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0003023</td>
<td>Angola</td>
<td>ago</td>
</tr>
<tr>
<td>C0001175</td>
<td>Acquired Immunodeficiency Syndrome</td>
<td>sida</td>
</tr>
<tr>
<td>C0001857</td>
<td>AIDS related complex</td>
<td>arc</td>
</tr>
<tr>
<td>C3714936</td>
<td>Non-Compliant ADaM Datasets Domain</td>
<td>ax</td>
</tr>
</tbody>
</table>
Approach - Refined sClass & Manually Tag

- English terms from MRCONSO.RRF with same CUI
- Exclude chemicals & drugs
  - use MRSTY.RRF to map CUI to STI
  - filter out disallowed STI in SemGroups.filter.txt
- In Lexicon with inflection is base and POS of adj, noun, or verb
- Remove acronyms/abbreviations => it drops precision
- Remove spVars => add them in post-process
- Remove nominalization => add them in post-process
- Remove singleton sClass (1 single candidates)
- Manually tag (for cognitive synonyms)
Computer-aided System

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

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

**Manual Tagging**
- Tagged by 2 linguists
  - Ensure cognitive synonyms

**Synonym Generation**
- sPairs Generating
  - Source: EUI and CUI
  - Add spelling variants
  - Add nominalization
Example: sClass & Tags (POS)

#SYNONYM_CLASS|C0003842|Arteries
noun|E0010481|arteria|Y
noun|E0010531|artery|Y
noun|E0694191|arterial|N
adj|E0010482|arterial|Y
#SYNONYM_CLASS|C0004063|Assault
verb|E0041250|mug|Y
noun|E0010822|assault|Y
noun|E0041249|mug|N
...

...
Synonym Sources

- Lexicon-Sourced Synonyms
  - Nominalizations with EUI
  - automatic retrieved from the SPECIALIST Lexicon

- UMLS-Sourced Cognitive Synonyms with CUI

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

- nominalizations are synonyms
- can be retrieved from the Lexicon automatically
- associated EUIs are preserved
- example:
  
  - sPair of [ability|noun|able|adj|E0006490]

```{base=ability
  entry=E0006490
  cat=noun
  variants=reg
  variants=uncount
  compl=pphr(of,np)
  compl=infcomp:arbc
  nominalization_of=able|adj|E0006510
}```
Example: sClass & Tagging

Refined sClass

#SYNONYM_CLASS|C0011065|Cessation of life
128|E0020918|death|Y
1|E0020877|dead|Y
1|E0020990|deceased|Y
1|E0022536|die|

Lexical Records

{base=death
tentry=E0020918
cat=noun
variants=reg
variants=uncount
compl=pphr(of,np)
compl=pphr(from,np)
nominalization_of=die|verb|E0022536
}
Example: sClass to sPairs

Final sClass

...  
#SYNONYM_CLASS|C0011065|Cessation of life
128|E0020918|death|Y
1|E0020877|dead|Y
1|E0020990|deceased|Y
1024|E0022536|die|nom
128|E0020885|deadness|nom
...

Add nominalization

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

sPairs

...  
deadness|128|dead|1|C0011065
deadness|128|death|128|C0011065
deadness|128|deceased|1|C0011065
deadness|128|die|1024|C0011065
dead|1|deadness|128|C0011065
dead|1|death|128|C0011065
dead|1|deceased|1|C0011065
dead|1|die|1024|C0011065
dead|128|deadness|128|C0011065
dead|128|death|1|C0011065
dead|128|deceased|1|C0011065
dead|128|die|1024|C0011065
dead|128|deadness|128|C0011065
dead|128|death|1|C0011065
dead|128|deceased|1|C0011065
dead|128|die|1024|C0011065
dead|128|deadness|128|C0011065
dead|128|death|1|C0011065
dead|128|deceased|1|C0011065
dead|128|die|1024|C0011065
dead|128|deadness|128|C0011065
dead|128|death|1|C0011065
dead|128|deceased|1|C0011065
dead|128|die|1024|C0011065
dead|128|deadness|128|C0011065
dead|128|death|1|C0011065
dead|128|deceased|1|C0011065
dead|128|die|1024|C0011065
dead|128|deadness|128|C0011065
dead|128|death|1|C0011065
dead|128|deceased|1|C0011065
dead|128|die|1024|C0011065
dead|128|deadness|128|C0011065
dead|128|death|1|C0011065
dead|128|deceased|1|C0011065
dead|128|die|1024|C0011065
dead|128|deadness|128|C0011065
dead|128|death|1|C0011065
dead|128|deceased|1|C0011065
dead|128|die|1024|C0011065
dead|128|deadness|128|C0011065
dead|128|death|1|C0011065
dead|128|deceased|1|C0011065
dead|128|die|1024|C0011065
...

{base=dead
 entry=E0020877
 cat=adj
 variants=inv
 ...
 position=pred
 stative
 nominalization=deadness|noun|E0020885
 }
sPairs Generation

UMLS-Sourced

Retrieve synonym candidates (sClasses)

Tag sClasses

Generate sPairs (CUI)

Lexicon-Sourced

Generate sPairs from nominalizations (EUI)

NLP Project-Sourced

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

Generate sPairs

<table>
<thead>
<tr>
<th>Synonym-1</th>
<th>POS-1</th>
<th>Synonym-2</th>
<th>POS-2</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>mug</td>
<td>verb</td>
<td>assault</td>
<td>noun</td>
<td>C0004063</td>
</tr>
<tr>
<td>assault</td>
<td>noun</td>
<td>mug</td>
<td>verb</td>
<td>C0004063</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Results – 2017 Release

2017 LexSynonyms

<table>
<thead>
<tr>
<th></th>
<th>Candidates</th>
<th>Tagged</th>
<th>Completion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sClass</td>
<td>22,779</td>
<td>7,686</td>
<td>33.74%</td>
</tr>
<tr>
<td>Synonyms</td>
<td>80,913</td>
<td>29,990</td>
<td>37.06%</td>
</tr>
</tbody>
</table>

Synonyms (sPairs):

<table>
<thead>
<tr>
<th>Year</th>
<th>CUI</th>
<th>EUI</th>
<th>NLP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>5,198 (100%)</td>
<td>5,198</td>
</tr>
<tr>
<td>2017</td>
<td>118,468 (62%)</td>
<td>67,584 (35%)</td>
<td>4,792 (3%)</td>
<td>190,844</td>
</tr>
</tbody>
</table>

Format:

<table>
<thead>
<tr>
<th>Synonym-1</th>
<th>POS-1</th>
<th>Synonym-2</th>
<th>POS-2</th>
<th>Source</th>
</tr>
</thead>
</table>

36.71 growth
Evaluation

Model:
- STMT (Sub-Term Mapping Tools) [6]:
  - Real-time subterm substitution features for concept mapping
  - Easy configurable options for element synonym set

Data:
- UMLS-Core Project:
  - Top 95% used terms form 8 hospitals.
  - Assigned CUI(s) to 13,076 terms
  - 2,755 terms of them do not have mapped concept through normalization in UMLS.2016AB
  - Gold Standard: 2,755 terms mapped to 2,756 CUIs
Evaluation Model

Input Terms (13,076) → 2,755 terms (~21% no CUI found)

Indexed Database Normalized String, 2016 AB

STMT

Subterm Substitutions

Norm

Element Synonym Sets
- STMT
- STMT + LexSynonym.2016
- STMT + LexSynonym.2017
- LexSynonym.2016
- LexSynonym.2017

10,321 terms (CUI found)
### Evaluation Results

- **Gold Standard:** 2,755 terms mapped to 2,756 CUIs
- **Element sets:**
  - STMT: a validated project specific synonym set for UMLS-Core project
  - About 75% of STMT element synonyms are duplicated in LexSynonym.2017, while only ~3% are duplicated in LexSynonym.2016.

<table>
<thead>
<tr>
<th>Element Synonym Set</th>
<th>N. Size</th>
<th>T.P.</th>
<th>F.P.</th>
<th>F.N.</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>STMT [6]</td>
<td>7,873</td>
<td>690</td>
<td>353</td>
<td>2,066</td>
<td>66.16%</td>
<td>25.04%</td>
<td>0.3633</td>
<td>7:57</td>
</tr>
<tr>
<td>STMT + LexSynonym.2016</td>
<td>12,681</td>
<td>691</td>
<td>358</td>
<td>2,065</td>
<td>65.87%</td>
<td>25.07%</td>
<td>0.3632</td>
<td>5:31</td>
</tr>
<tr>
<td>STMT + LexSynonym.2017</td>
<td>151,913</td>
<td>828</td>
<td>424</td>
<td>1,928</td>
<td>66.13%</td>
<td>30.04%</td>
<td><strong>0.4132</strong></td>
<td>9:18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Element Synonym Set</th>
<th>N. Size</th>
<th>T.P.</th>
<th>F.P.</th>
<th>F.N.</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>LexSynonym.2016</td>
<td>5,070</td>
<td>9</td>
<td>12</td>
<td>2,747</td>
<td>42.86%</td>
<td>0.33%</td>
<td>0.0065</td>
<td>0:16</td>
</tr>
<tr>
<td>LexSynonym.2017</td>
<td>149,912</td>
<td>287</td>
<td>117</td>
<td>2,469</td>
<td>71.04%</td>
<td><strong>10.41%</strong></td>
<td>0.1816</td>
<td>3:19</td>
</tr>
</tbody>
</table>
Lexical Tools – Synonym Flow

Software Changes:
• Include POS and the source information in synonym database

Example:
shell> lvg -f:y -m
die
die|dead|1|1|y|1|FACT|die|die|verb|dead|adj|C0011065|
die|deadness|128|1|y|1|FACT|die|die|verb|deadness|noun|C0011065|
die|death|128|1|y|1|FACT|die|die|verb|death|noun|C0011065|
die|deceased|1|1|y|1|FACT|die|die|verb|deceased|adj|C0011065|
die|expire|1024|1|y|1|FACT|die|die|verb|expire|verb|NLP_LVG|
Lexical Tools – Synonyms Flow Options

Synonym source restriction options (-ks):
• C (CUI), E (EUI), N (NLP), CE, CN, EN, CEN.

Example:
shell> lvg –f:y –m –ks:C
die
die|dead|1|1|y|1|FACT|die|die|verb|dead|adj|C0011065|
die|deadness|128|1|y|1|FACT|die|die|verb|deadness|noun|C0011065|
die|death|128|1|y|1|FACT|die|die|verb|death|noun|C0011065|
die|deceased|1|1|y|1|FACT|die|die|verb|deceased|adj|C0011065|
Lexical Tools – Recursive Synonym Flow

Software Enhancement:
- must have the same type of source
- If the source is CUI: only synonyms from the same CUI are used (multiple CUI Issues)
- If the source is EUI: all synonyms with EUI source are used
- If the source is NLP: synonyms from same NLP source are used

Example:
```
shell> lvg -f:y -m
die
die|dead|1|1|r|2|FACT|die|verb|dead|adj|C0011065|y|
die|deadness|128|1|r|2|FACT|die|verb|deadness|noun|C0011065|y|
die|death|128|1|r|2|FACT|die|verb|death|noun|C0011065|y|
die|deceased|1|1|r|2|FACT|die|verb|deceased|adj|C0011065|y|
die|expire|1024|1|r|2|FACT|die|verb|expire|verb|NLP_LVG|y|
die|terminate|1024|1|r|2|FACT|expire|verb|terminate|verb|NLP_LVG|yy|
```
### Summary

<table>
<thead>
<tr>
<th>Objective &amp; Requirements</th>
<th>Check</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standalone element synonym set</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>All synonymous terms in the Lexicon</td>
<td>1/3 Yes</td>
<td>~ 1/3 completed</td>
</tr>
<tr>
<td>Grows with the SPECIALIST Lexicon</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Element synonyms, not expanded terms (Over-generated issues)</td>
<td>Yes</td>
<td>Must be in the Lexicon (430K, &lt; 2% of UMLS synonyms)</td>
</tr>
<tr>
<td>R1: Cognitive Synonym</td>
<td>Yes</td>
<td>Done in tagging (cognitive synonyms)</td>
</tr>
<tr>
<td>R2: Include POS</td>
<td>Yes</td>
<td>Provide POS in sClass by Lexicon</td>
</tr>
<tr>
<td>R3: Include source (CUI, EUI, etc.)</td>
<td>Yes</td>
<td>Provide source in sClass (CUI, EUI, etc.)</td>
</tr>
<tr>
<td>R4: Exclude Acronym/abbreviation</td>
<td>Yes</td>
<td>Removed in sClass by Lexicon</td>
</tr>
<tr>
<td>R5: Include Single words and multiwords</td>
<td>Yes</td>
<td>Terms in the Lexicon include both</td>
</tr>
<tr>
<td>Improve NLP performance</td>
<td>Yes</td>
<td>Improve recall and preserve precision</td>
</tr>
</tbody>
</table>
Future Work

- Complete all candidate sClasses in the future releases
- Update annually on Lexicon and Lexical Tools release with the latest Lexicon and UMLS Metathesaurus
- Include more project specific synonym set from other NLP resources (UMLS-Core, Randy Milller, etc.)
- Performance tests on NLP applications
Questions