The SPECIALIST Lexicon and NLP Tools

By: Dr. Chris J. Lu

NLM – LHNCBC - CGSB

Jun., 2017

• The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov
Table of Contents

- 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. Lynn McCready
- Destinee Tormey
- Francois Lang
- Dr. Chris J. Lu
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%)
(Multi)Words for Lexical Records

- 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
  - …
Categories – Parts of Speech (11)

Lexicon.2017

Noun: 82.5%

Adj: 13%

Verb: 2%

Adv: 2%
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
position=attrib(1)
position=attrib(3)
position=pred
stative
nominalization=squareness|noun|E0057519
}

{base=square
entry=E0057518
cat=adv
modification_type=intensifier
modification_type=verb_modifier;manner
}

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

village square the circle

fair and square root
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 Records to Lexical Information

<table>
<thead>
<tr>
<th>Lexical Information</th>
<th>Base</th>
<th>color</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part of speech</td>
<td>• noun</td>
<td></td>
</tr>
<tr>
<td>Inflectional morphology (inflections)</td>
<td>• color</td>
<td>• colors</td>
</tr>
<tr>
<td>Orthography</td>
<td>• colour</td>
<td></td>
</tr>
<tr>
<td>Abbreviation/Acronym</td>
<td>• N/A</td>
<td></td>
</tr>
<tr>
<td>Syntax (complementation)</td>
<td>• N/A</td>
<td>• ...</td>
</tr>
<tr>
<td>Derivational morphology (derivations)</td>
<td>• colorable</td>
<td>• colorful</td>
</tr>
<tr>
<td>LexSynonyms</td>
<td>• chromatic</td>
<td></td>
</tr>
</tbody>
</table>
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=role
  spelling_variant=rôle
  entry=E0053757
  cat=noun
  variants=reg
}

{base=cave
  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>Total</td>
<td>3,619,854</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 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).
The Frequency Spectrum of Lexicon (Multi)words on MEDLINE
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.
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. 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)
<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, l, 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 \rightarrow \text{inflect} \rightarrow \text{leaves, leaving, left}
```
> 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 –f:i
leave

```plaintext
leave
---
128
---
i
---
1
```
• 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

Input term → noOperation → Output term

Input term → Uninflect → Spelling Vars → Output terms
Parallel Flows - Example

```plaintext
> 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

- Composed of 11 Lvg flow components to abstract away from:
  - case
  - punctuation
  - possessive forms
  - inflections
  - spelling variants
  - stop words
  - diacritics & ligatures (non-ASCII Unicode)
  - word order
<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>
<tr>
<td>q0: map symbols to ASCII</td>
</tr>
<tr>
<td>--------------------------</td>
</tr>
<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>
<tr>
<td>q0: map symbols to ASCII</td>
</tr>
<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>
Norm

q0: map symbols to ASCII
"Fœtoproteins α’s, NOS"
"Fœtoproteins α’s, NOS"
"Fœtoproteins α, NOS"
"Fœtoproteins α, NOS"

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

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

**Norm**
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

g: remove genitives
"Fœtoproteins α's, NOS"
"Fœtoproteins α, NOS"
Fœtoproteins α
fœtoproteins α

rs: remove parenthetic plural forms
Fœtoproteins α
fœtoproteins α

o: replace punctuation with spaces
Fœtoproteins α
fœtoproteins α

l: lowercase
Fœtoproteins α
fœtoproteins α

B: uninflect each words in a term
Fœtoproteins α
fœtoproteins α

Ct: retrieve citations

q7: Unicode core Norm
Fœtoproteins α
fœtoproteins α

q8: strip or map Unicode to ASCII
Fœtoproteins α
fœtoproteins α

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

<p>| &quot;Fœtoproteins α's, NOS&quot; | &quot;Fœtoproteins α's, NOS&quot; |
| &quot;Fœtoproteins α, NOS&quot;  | &quot;Fœtoproteins α, NOS&quot;  |
| Fœtoproteins α NOS     | Fœtoproteins α          |
| fœtoproteins α         | fœtoprotein α           |</p>
<table>
<thead>
<tr>
<th>Norm</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>q0: map symbols to ASCII</td>
<td>&quot;Fœtoproteins α’s, NOS“</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: uninfllect 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>
<tr>
<td>q0: map symbols to ASCII</td>
<td>&quot;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>føtoprotein 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></td>
</tr>
<tr>
<td>w: sort words by order</td>
<td></td>
</tr>
</tbody>
</table>
q0: map symbols to ASCII

"Fœtoproteins α’s, NOS"

"Fœtoproteins α, NOS"

"Fœtoproteins α, NOS"

Fœtoproteins α NOS

Fœtoproteins α

foetoproteins α

foetoprotein α

fetoprotein α

fetoprotein α

fetoprotein alpha

q8: strip or map Unicode to ASCII
<table>
<thead>
<tr>
<th>Operation</th>
<th>Text</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's, 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>&quot;Fœtoproteins α, NOS&quot;</td>
</tr>
<tr>
<td>l: lowercase</td>
<td>Fœtoproteins α NOS</td>
</tr>
<tr>
<td>B: uninflect each words in a term</td>
<td>Fœtoproteins α</td>
</tr>
<tr>
<td>Ct: retrieve citations</td>
<td>foœtoproteins α</td>
</tr>
<tr>
<td>q7: Unicode core Norm</td>
<td>foœtoprotein α</td>
</tr>
<tr>
<td>q8: strip or map Unicode to ASCII</td>
<td>fetoprotein α</td>
</tr>
<tr>
<td>w: sort words by order</td>
<td>fetoprotein α</td>
</tr>
<tr>
<td></td>
<td>fetoprotein alpha</td>
</tr>
<tr>
<td></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 Fœtoprotein
…
3. Natural Language Processing (NLP)

• Natural language is ordinary language that humans use naturally, may be spoken, written, or sign.

• The main purpose of language is communication, for us to understand the meaning.

• NLP includes a board range of subjects.

• NLP in our scope is to use computer to understand the meaning (concept) from text for further analysis and processing.
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.
# NLP 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>
</tbody>
</table>

<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>• Common Cold</td>
<td>C0009443</td>
<td></td>
</tr>
<tr>
<td>• Cold Therapy</td>
<td>C0010412</td>
<td></td>
</tr>
<tr>
<td>• Cold Sensation</td>
<td>C0234192</td>
<td></td>
</tr>
<tr>
<td>• ...</td>
<td></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

Lexical Information

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

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

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

q0: map Unicode symbols to ASCII

Behçet’s Diseases, NOS
Behçet Diseases, NOS
Behçet Diseases NOS
Behçet Diseases

Behçet disease
behçet disease
behçet disease
behçet disease
behçet disease
behçet disease
behçet disease

Behçet’s Diseases, NOS
Behçet Diseases, NOS
Behçet Diseases NOS
Behçet Diseases

Behçet disease
behçet disease
behçet disease
behçet disease
behçet disease
behçet disease
behçet disease

w: sort words by order
NLP – Norm (Lexical Variations)

- Behcet Disease
- Behçet’s Disease
- Behcet Diseases
- Behçet Diseases
- Behçet's Disease
- Behçet's Disease
- Behcets Disease
- Behçets Disease
- Behçet’s Disease, NOS
- Behçet's Disease, NOS
- behcet disease
- behcet diseases
- behcet’s disease
- behcet’s disease, nos
disease, Behçet
diseases, behçet
...

[Indexed Database]

[Normalized String]

[Terms in Corpus]

normalize
Index

behcet disease

• C0004943
• Behcet Syndrome

[UMLS Synonyms]
Behcet’s Disease

Query → norm → Normed Term → Behcet disease

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

Indexed Database
Normalized String

SQL
NLP – Query Expansion (derivation)

Indexed Database Normalized String

- perforated ear drum
  - Norm
  - drum ear perforate
  - None

- perforation ear drum
  - Norm
  - drum ear perforation
  - C0206504
    - Tympanic Membrane Perforation
NLP – Query Expansion (Synonym)

Indexed Database Normalized String

calcaneal fracture

Norm

calcaneal fracture

None

heel bone fracture

Norm

bone fracture heel

C0281926 Fracture of calcaneus

C0006655: • calcaneal • heel bone
NLP – Query Expansion (Synonym)

[Input term]
calcaneal fracture

[Element Synonym]
Norm
calcaneal fracture

C0006655:
- calcaneal
- heel bone

[Indexed Database Normalized String]

[Expanded Term]
heel bone fracture

Norm
bone fracture heel

None

C0281926
Fracture of calcaneus

[sPair: calcaneal|heel bone]

[UMLS Synonyms]
NLP – Concept Mapping Model

- Free Text

Tokenization (Segmentation)
- Documents
- Paragraphs
- Sentences
- Phrases
- Terms (Lexical Lookup)
- Tokens (words)
- …

Norm Term

Ontology (UMLS) - Indexed Database Normalized Term

CUI

Yes
- WSD (STI)
- Ranking

No
- Same Records
- Related Records

Terms
- Query Expansion (STMT)

(derivations, synonyms, co-occurrences, or fruitful variants, etc.)
Generates expanded terms of the input term
- derivational variants
- synonyms (recursive)
- fruitful variants (combination of above)

Normalization (lexical variants from the same record)

Enhanced UMLS thesaurus
- Pre-generated expanded term pool
- Add new expanded terms (synonyms) to UMLS thesaurus

Find candidates (mapped concepts)

Ranking & Filters (keyword match, frequency, semantic types, concept distance, longest terms, etc.)
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
Element Synonyms Review

- **UMLS Synonyms**
  - Semantically equivalent terms that have the same or very similar meaning (concept, CUI).
  - 2016AA UMLS Metathesaurus containing over 3.25M concepts and nearly 13 M unique concept names from over 190 source vocabularies.

- The SPECIALIST Lexicon and Lexical Tools Synonyms, 2016- (~5K)
- UMLS-Core Projects (~12K)
- Synonym set by Randy Miller, (~15K)
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)
    o include all synonymous terms in Lexicon (LexSynonyms)
    o grow with the SPECIALIST Lexicon
Synonym Types

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

- **Near-synonym:**
  - greater difference
  - less interchangeableability
  - specific use, can’t used in generic case
Properties of Cognitive Synonyms (sPairs)

- **Commutativity**: 
  - $(x = y) \rightarrow (y = x)$
  - bi-directional
  - `joy | noun | enjoy | verb` -> `enjoy | verb | joy | noun`

- **Transitivity**: 
  - $((x = y) \text{ and } (y = z)) \rightarrow (x = z)$
  - `enjoy | verb | joy | noun`
  - `joy | noun | happy | adj`
  - recursive
  - `enjoy | verb` -> `joy | noun` -> `happy | adj`

- Suitable for sPairs (element synonyms)
- Resolve many issues in element synonyms.
## Broader/Narrower Issues – Near Synonyms

<table>
<thead>
<tr>
<th>CUI</th>
<th>Preferred Term</th>
<th>synonym</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0001613</td>
<td>Adrenal Cortex</td>
<td>cortical</td>
<td>The adjective cortical can refer to any of several types of cortex &amp; so does not have synonymy with “adrenal cortex”</td>
</tr>
<tr>
<td>C0032639</td>
<td>Pontine structure</td>
<td>metencephalon</td>
<td>The metencephalon, per m-w.com includes the cerebellum and pons, and is different from the pons</td>
</tr>
<tr>
<td>C0001575</td>
<td>Uterine adnexae structure</td>
<td>adnexa</td>
<td>There are several types of adnexa, such as eye adnexa, adnexa of skin, etc.</td>
</tr>
<tr>
<td>C0000936</td>
<td>Visual Accommodation</td>
<td>accommodation</td>
<td>There are other accommodations.</td>
</tr>
</tbody>
</table>
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
## Distinct Issues – Similar but Different

<table>
<thead>
<tr>
<th>CUI</th>
<th>Preferred Term</th>
<th>synonym</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0000741</td>
<td>Abducens nerve structure</td>
<td>abductor digiti minimi</td>
<td>The abductor digiti minimi is a muscle, not a nerve.</td>
</tr>
<tr>
<td>C0003864</td>
<td>Arthritis</td>
<td>arthritide</td>
<td>Per Dorland’s an arthride is “any skin eruption of arthritic or gouty origin.”</td>
</tr>
<tr>
<td>C0005400</td>
<td>Bile duct structure</td>
<td>choledochitis</td>
<td>Choledochitis is a condition of the common bile duct, not structure.</td>
</tr>
<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>
</tbody>
</table>
Acacia & Locust tree
Anguilla & Antigua
## Acronym/Abbreviation Issues – Precision

<table>
<thead>
<tr>
<th>CUI</th>
<th>Preferred Term</th>
<th>synonym</th>
</tr>
</thead>
<tbody>
<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><strong>C0003023</strong></td>
<td>Angola</td>
<td>ago</td>
</tr>
<tr>
<td>C3714936</td>
<td>Non-Compliant ADaM Datasets Domain</td>
<td>ax</td>
</tr>
</tbody>
</table>

- ER (27): emergency room | efficacy ratio | ejection rate | evoked response | extended release | external resistance | eye research | energy restriction | …
### 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>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>
<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>
</tbody>
</table>
Recursive Issues – Multiple Concepts

- Multiple CUIs (transitivity?)
- Example (cold):

<table>
<thead>
<tr>
<th>CUIs</th>
<th>Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0009443, common cold</td>
<td>• cold</td>
</tr>
<tr>
<td></td>
<td>• coryza</td>
</tr>
<tr>
<td></td>
<td>• acute coryza</td>
</tr>
<tr>
<td></td>
<td>• common cold</td>
</tr>
<tr>
<td>C0009264, cold temp</td>
<td>• cold</td>
</tr>
<tr>
<td></td>
<td>• low temperature</td>
</tr>
<tr>
<td></td>
<td>• low-temperature</td>
</tr>
<tr>
<td></td>
<td>• low-temperature</td>
</tr>
<tr>
<td>C0234192, cold sens</td>
<td>• cold</td>
</tr>
<tr>
<td></td>
<td>• psychoesthesia</td>
</tr>
<tr>
<td>...</td>
<td>• ...</td>
</tr>
</tbody>
</table>

- common cold|cold
- cold|cold temperature
- cold|cold sensation

=> common cold|cold|cold temperature ?
=> common cold|cold|cold sensation?
Recursive Issue 2 – Endless loop

Example – cold blooded animal

<table>
<thead>
<tr>
<th>Synonyms</th>
<th>cold temperature</th>
<th>cold therapy</th>
<th>common cold</th>
<th>cold sensation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-G substitution</td>
<td>cold temperature</td>
<td>cold therapy</td>
<td>common cold</td>
<td>cold sensation</td>
</tr>
<tr>
<td>2-G substitution</td>
<td>cold temperature</td>
<td>cold therapy</td>
<td>common cold</td>
<td>common cold sensation</td>
</tr>
<tr>
<td></td>
<td>temperature</td>
<td>cold therapy</td>
<td>common cold</td>
<td>common cold sensation</td>
</tr>
<tr>
<td></td>
<td>cold therapy</td>
<td>therapy</td>
<td>common cold</td>
<td>common cold sensation</td>
</tr>
<tr>
<td></td>
<td>common cold</td>
<td>therapy</td>
<td>common cold</td>
<td>common cold sensation</td>
</tr>
<tr>
<td></td>
<td>cold sensation</td>
<td>therapy</td>
<td>common cold</td>
<td>common cold sensation</td>
</tr>
<tr>
<td></td>
<td>temperature</td>
<td></td>
<td></td>
<td>[...]</td>
</tr>
</tbody>
</table>

...
LexSynonyms – Objectives

➢ To establish a system to:
  • generate a standalone set of generic element synonyms (sPairs)
    o include all synonymous terms in Lexicon (LexSynonyms)
    o grow with the SPECIALIST Lexicon
  • use for effective UMLS concept mapping
    o a **thorough set** of element synonyms (to increase recall)
    o **cognitive synonyms** (to preserve precision)
LexSynonyms – Requirements

- Requirements (sClass):
  - All synonymous terms (cognitive synonyms) in the Lexicon
  - Bi-directional (commutativity) - interchangeable sPair in NLP
  - Recursive (transitivity) - use in NLP to improve Recall, yet preserve precision

- Resolve all above observed issues
  - Broader issues
  - Distinct issues
  - Acronym/abbreviation issues
  - POS issues
  - Recursive issues
Approach - Refined sClass

- 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)
sClass Example

#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
}
```
UMLS-Sourced Cognitive Synonyms

UMLS sClasses
(English terms with same CUI)

Filter & Matchers:
• remove chemicals and drugs
• must be a base form in the Lexicon
• POS: noun, verb, adjective
• remove acronyms or abbreviations

Auto-Processing sClasses:
• Spelling variants
• Nominalization
• EUI/CUI

Manual tagging on refined sClasses:
• To ensure cognitive synonyms

Auto-Generating sPairs:
• Spelling variants
• Nominalization
• EUI/CUI
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
entry=E0020918
cat=noun
variants=reg
variants=uncount
compl=pphr(of,np)
compl=pphr(from,np)
nominalization_of=die|verb|E0022536
}

Removed (nominalization)
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=dead
tentry=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|128|C0011065
dead|128|deceased|1|C0011065
dead|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|128|C0011065
dead|128|deceased|1|C0011065
dead|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|128|C0011065
dead|128|deceased|1|C0011065
dead|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|128|C0011065
dead|128|deceased|1|C0011065
dead|128|die|1024|C0011065
dead|1|deadness|128|C0011065
dead|1|death|128|C0011065
dead|1|deceased|1|C0011065
dead|1|die|1024|C0011065
...

{base=death
tentry=E0020877
cat=adj
variants=inv
...
position=pred
stative
nominalization=deadness|noun|E0020885
}
LexSynonym Generation

1. Retrieve synonym candidates (sClasses)
2. Tag sClasses
3. Generate sPairs (CUI)
4. Generate sPairs from nominalizations (EUI)
5. Generate sPairs from Lexical Tools, 2016 (NLP-LVG)
6. Combine sPairs
Results

➢ 2017 release:
  • 2016AB Metathesaurs, 2016 Lexicon

<table>
<thead>
<tr>
<th></th>
<th>Total</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>

• Synonym stats:

<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>

36.71 growth
Tests

➢ **Model:**
  - STMT (Sub-Term Mapping Tools):
    o Real-time subterm substitution tools for concept mapping
    o Easy configurable options for element synonym set

➢ **Data:**
  - UMLS-Core Project:
    o Assigned CUI(s) to 13,076 terms
    o 2,755 terms of them do not have mapped concept through normalization in UMLS.2016AB
    o Gold Standard: 2,755 terms mapped to 2,756 CUIs
### Test Results

- **Gold Standard:** 2,755 terms mapped to 2,756 CUIs
- **Element sets:**
  - **STMT:** include a validated cognitive synonym set
  - **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</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>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>10.41%</td>
<td>0.1816</td>
<td>3:19</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>0.4132</td>
<td>9:18</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 Synonyms

- **CUI**
  - CUI: C0011065
  - PT: Cessation of life
    - die
    - dead
    - death
    - deadness
    - deceased

- **NLP**
  - CUI: C0231800
  - PT: Expiration
    - die
    - expire
    - terminate

- **EUI**
  - X
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>Goals</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>Yes</td>
<td>~ 1/3 completed</td>
</tr>
<tr>
<td>Grows with the SPECIALIST Lexicon</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Over-generated issues</td>
<td>Resolved</td>
<td>Must be in the Lexicon (430K, ~2% of UMLS synonyms)</td>
</tr>
<tr>
<td>Single words and multiwords</td>
<td>Resolved</td>
<td>Bases in Lexicon include both</td>
</tr>
<tr>
<td>Broader issues</td>
<td>Resolved</td>
<td>Done in tagging (cognitive synonyms)</td>
</tr>
<tr>
<td>Distinct issues</td>
<td>Resolved</td>
<td>Done in tagging (cognitive synonyms)</td>
</tr>
<tr>
<td>Acronym/abbreviation issues</td>
<td>Resolved</td>
<td>Removed in sClass</td>
</tr>
<tr>
<td>POS issues</td>
<td>Resolved</td>
<td>Provide POS in sClass</td>
</tr>
<tr>
<td>Recursive issues</td>
<td>Resolved</td>
<td>Provide source in sClass (CUI, EUI, etc.)</td>
</tr>
<tr>
<td>Improve NLP performance</td>
<td>Yes</td>
<td>Improve recall and preserve precision</td>
</tr>
</tbody>
</table>
Future Work

- Complete tagging on all sClasses

- Updated annually on Lexicon and Lexical Tools release
  - with the latest Lexicon and Metathesaurus
- Computer-aided (automatic) process on the sClass tagging
- Add more synonyms from other NLP projects (UMLS-Core, Randy Milller, etc.)
- Performance tests on NLP applications

<table>
<thead>
<tr>
<th></th>
<th>Total</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>
Questions