The SPECIALIST Lexicon and Lexical Tools

• Allen Browne
• Guy Divita
• Chris Lu
Metathesaurus Focused Search:

1) Select UMLS Release:

2) Enter a term or a concept unique identifier (CUI):

3) **Restrict source vocabulary to:**
   - All Source Vocabularies
   - Restrict to selected sources:

4) **String Matching Criteria**
The SPECIALIST Lexicon

- A syntactic lexicon
- Biomedical and general English
- Over 180,000 records
The SPECIALIST Lexicon

• General English:
• 10,000 most frequent words from the American Heritage word frequency list
• 2,000 words used by Longman’s Dictionary of Contemporary English
• Verbs and adjectives identified by heuristics
The SPECIALIST Lexicon

• Morphology
  – Inflection
  – Derivation

• Orthography
  – Spelling variants

• Syntax
  – Complementation for verbs, nouns, and adjectives
Morphology

• Inflectional
  – nucleus -- nuclei
  – cauterize, cauterizes, cauterized, cauterizing
  – red, redder reddest

• Derivational
  – laryngeal -- larynx
  – transport -- transportation
Inflectional Morphology

The pitcher wound up and he flang the ball at the batter. The batter swang and missed. The pitcher flang the ball again and this time the batter connected. He hit a high fly right to the center fielder. The center fielder was all set to catch the ball, but at the last minute his eyes were blound by the sun and he dropped it.

--J. H. "Dizzy" Dean
Derivational Morphology

I suppose you could say I'm a dictionaryologist.

Dictionary + ology + ist
Orthography

Spelling Variation

• align -- aline
• Grave’s disease -- Graves’s disease -- Graves’ disease
• anesthetize -- anaesthetise
• esophagus -- oesophagus
British and American Spelling

• Criticise -- criticize
• naturalise -- naturalize
• centre -- center
• foetus -- fetus
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.
Syntax -- Verb Complements

{base=treat
 entry=E0061964
   cat=verb
   variants=reg
   intran
   tran=np
   tran=pphr(with,np)
   tran=pphr(of,np)
   ditran=np,pphr(to,np)
   ditran=np,pphr(with,np)
   ditran=np,pphr(for,np)
   cplxtran=np,advbl
   nominalization=treatment|noun|E0061968
}

The 2003 SPECIALIST Lexicon

Number of lexical items

- noun
- adjective
- verb
- adverb
- preposition
- pronoun
- conjunction
- determiner
- modal
- modal
- auxilliary
- compl
village square

square

fair and square

square root

square

the circle
Lexicon Unit Records

{base=Kaposi's sarcoma
 spelling_variant=Kaposi sarcoma
 entry=E0003576
   cat=noun
   variants=uncount
   variants=reg
   variants=glreg
}

{base=chronic
 entry=E0016869
   cat=adj
   variants=inv
   position=attrib(1)
   position=pred
   stative
}

{base=aspirate
 entry=E0010803
   cat=verb
   variants=reg
   tran=np
   nominalization=aspiration|noun|E0010804
}

{base=in
 entry=E0033870
   cat=prep
}

{base=chronic
 entry=E0016869
   cat=adj
   variants=inv
   position=attrib(1)
   position=pred
   stative
}

{base=aspirate
 entry=E0010803
   cat=verb
   variants=reg
   tran=np
   nominalization=aspiration|noun|E0010804
}

{base=in
 entry=E0033870
   cat=prep
}
Noun Variants

{base=Kaposi's sarcoma
 spelling_variant=Kaposi sarcoma
 entry=E0003576
   cat=noun
   variants=uncount
   variants=reg
   variants=glreg
 }

- Kaposi’s sarcoma
- Kaposi’s sarcomas
- Kaposi’s sarcomata
- Kaposi sarcoma
- Kaposi sarcomas
- Kaposi sarcomata
Regular Nouns

The plural suffix is *s*. y becomes *ie* following a consonant before *s*. e is inserted before *s* if the base ends in *s, z, x, ch, or s*.
## Regular Nouns

<table>
<thead>
<tr>
<th>Base ends with</th>
<th>Plural ends with</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cy</td>
<td>Cies</td>
<td>fly: flies</td>
</tr>
<tr>
<td>-s</td>
<td>-ses</td>
<td>illness: illnesses</td>
</tr>
<tr>
<td>-z</td>
<td>-zes</td>
<td>waltz: waltzes</td>
</tr>
<tr>
<td>-x</td>
<td>-xes</td>
<td>box: boxes</td>
</tr>
<tr>
<td>-ch</td>
<td>-ches</td>
<td>match: matches</td>
</tr>
<tr>
<td>-sh</td>
<td>-shes</td>
<td>splash: splashes</td>
</tr>
<tr>
<td>X</td>
<td>Xs</td>
<td>book: books</td>
</tr>
</tbody>
</table>
### Greco-latin Regular Nouns

<table>
<thead>
<tr>
<th>Singular ends with:</th>
<th>Plural ends with:</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>-us</td>
<td>-i</td>
<td>focus/foci</td>
</tr>
<tr>
<td>-ma</td>
<td>-mata</td>
<td>trauma/traumata</td>
</tr>
<tr>
<td>-a</td>
<td>-ae</td>
<td>larva/larvae</td>
</tr>
<tr>
<td>-um</td>
<td>-a</td>
<td>ilium/ilia</td>
</tr>
<tr>
<td>-on</td>
<td>-a</td>
<td>taxon/taxa</td>
</tr>
<tr>
<td>-sis</td>
<td>-ses</td>
<td>analysis/analyses</td>
</tr>
<tr>
<td>-is</td>
<td>-ides</td>
<td>cystis/cystides</td>
</tr>
<tr>
<td>-men</td>
<td>-mina</td>
<td>foramen/foramina</td>
</tr>
<tr>
<td>-ex</td>
<td>-ices</td>
<td>index/indices</td>
</tr>
<tr>
<td>-x</td>
<td>-ces</td>
<td>matrix/matrices</td>
</tr>
</tbody>
</table>
Uncount Nouns
(abstract or mass)

{base=smallpox
 entry=E0056359
   cat=noun
   variants=uncount
}

{base=potassium
 entry=E0049387
   cat=noun
   variants=uncount
}

• * a smallpox
• * two smallpoxes
• much smallpox
• * a potassium
• * two potassiuims
• much potassium
Fixed Plural Nouns

{base=police
 entry=E0048616
  cat=noun
  variants=plur
}

{base=scissors
 entry=E0054633
  cat=noun
  variants=plur
}
Irregular Nouns

{base=corpus
 entry=E0019113
   cat=noun
   variants=irreg|corpora|
   variants=reg
 }

{base=larynx
 entry=E0036919
   cat=noun
   variants=irreg|larynges|
   variants=reg
 }
Regular Verbs

• The third person present tense suffix is *s*.  
  – *y* becomes *ie* following a consonant before *s*.  
  – *e* is inserted between *z, x, ch, or sh* and *s*.

• The past tense suffix is *ed*.  
  – *y* becomes *ie* following a consonant before *ed*.  
  – Final *e* is deleted before *ed*. 
Regular Verbs

- dismiss: dismisses, dismissed, dismissing
- agree: agrees; agreed; agreeing
- dry: dries, dried, drying
Regular Doubling Verbs

• End in a CVC pattern
• Double the final consonant before *ed* and *ing*.  
• Are otherwise regular
• variants=regd
• e.g. control: controls, controlled, controlling
Irregular Verbs

{base=dive
cat=verb
    variants=reg
    variants=irreg|dives|dove|dove|diving|
intran
intran;part(in)
...
}

Dive vs. Dove
Regular Adjectives and Adverbs

- The comparative suffix is *er*.
- The superlative suffix is *est*.
  - *y* become *ie* after a consonant before *er* or *est*.
  - Final *e* is deleted before *er* or *est*.
- e.g. green: greener, greenest
Regular Doubling Adjectives and Adverbs

- CVC final pattern
- Final consonant is doubled before ed or est.
- Otherwise regular
- e.g. red: redder, reddest
Ancillary Data Bases

- Synonymy
  - sm.db
- Derivation
  - dm.db, dm.rules
- Inflection
  - im.rules
- Neoclassical compounds
  - nc.db
Derivational Facts and Rules

dm.facts

treatment|noun|treat|verb
prohibition|noun|prohibitive|adj
cell lineage|noun|cell line|noun
photochemotherapeutic|adj|photochemotherapy|noun
pharmacotherapeutic|adj|pharmacotherapy|noun
Derivational Facts and Rules

dm.rules

# e.g. alienation|alienate
ation$|noun|ate|verb
 ration|rate; station|state;
Inflectional Facts and Rules

im.rules

# Noun rules (glreg)
us$|noun|singular|i$|noun|plural
   antus|anti;
ma$|noun|singular|mata$|noun|plural
a$|noun|singular|ae$|noun|plural
um$|noun|singular|a$|noun|plural
on$|noun|singular|a$|noun|plural
sis$|noun|singular|ses$|noun|plural
is$|noun|singular|ides$|noun|plural
men$|noun|singular|mina$|noun|plural
ex$|noun|singular|ices$|noun|plural
x$|noun|singular|ces$|noun|plural
Neoclassical compounds

nc.db

abdomin(o)|abdomen|root
ab|away from|prefix
acanth(o)|prickle|root
acar(o)|mite|root
acetabul(o)|acetabulum|root
ad|towards|prefix
agogue|inducing|terminal
albumin(o)|albumin|root
sis|condition|terminal
stomy|surgical opening|terminal
Synonyms

sm.db

alar|adj|wing|noun
amygdaline|adj|tonsil|noun
articular|adj|joint|noun
bulbar|adj|medulla oblongata|noun
fununcular|adj|boil|noun
genicular|adj|knee|noun
hepatocellular|adj|liver cells|noun
lazar|adj|leprosy|noun
lenticular|adj|crystalline lens|noun
ypsiliform|adj|upsiloid|adj
wolfram|noun|tungsten|noun
double vision|noun|diplopia|noun
Relational Tables

• One line records
• Pipe separated Fields -- “|”
• Keyed to EUI
• LRAGR matches forms to EUIs
• Word index: LRWD
Relational Tables

- LRAGR - Agreement
- LRCMP - Complements
- LRFIL - Files
- LRFLD - Fields
- LRMOD - Modification
- LRNOM - Nominalization
- LRPRN - Pronouns
- LRPRP - Properties
- LRSPL - Spelling
- LRTRM - Trademarks
- LRWD - Word index
LRAGR

Agreement and Inflection

- EUI - Entry ID
- STR - Inflected form
- SCA - Syntactic category
- AGR - agreement information
- BAS - Base form (morphological)
- CIT - Citation form (base=)
LRAGR

E0003576| Kaposi sarcomas| noun| count(thr_plur)| Kaposi sarcoma| Kaposi's sarcoma|
E0003576| Kaposi sarcomata| noun| count(thr_plur)| Kaposi sarcoma| Kaposi's sarcoma|
E0003576| Kaposi sarcoma| noun| count(thr_sing)| Kaposi sarcoma| Kaposi's sarcoma|
E0003576| Kaposi sarcoma| noun| uncount(thr_sing)| Kaposi sarcoma| Kaposi's sarcoma|
E0003576| Kaposi's sarcomas| noun| count(thr_plur)| Kaposi's sarcoma| Kaposi's sarcoma|
E0003576| Kaposi's sarcomata| noun| count(thr_plur)| Kaposi's sarcoma| Kaposi's sarcoma|
E0003576| Kaposi's sarcoma| noun| count(thr_sing)| Kaposi's sarcoma| Kaposi's sarcoma|
E0003576| Kaposi's sarcoma| noun| uncount(thr_sing)| Kaposi's sarcoma| Kaposi's sarcoma|
Number Words

- Not in the lexicon.
  - No part of speech
  - Used to construct number expressions:
    “Three thousand eight hundred and five”
- To be released in the 2003 lexicon.
- Accompanying number tools.
{base=two
cat=number_word
entry=N0000003
variant=second;ordinal
variant=second;denominator,singular;part_denominator
variant=second;denominator,plural;part_denominator
variant=half;denominator,singular;full_denominator
variant=halves;denominator,plural;full_denominator
number_type=unit
value=2
digit=2
}
{base=twenty
  cat=number_word
  entry=N00000021
  variants=reg
  number_type=decade
  value=20
  digit=2
}

{base=billion
  cat=number_word
  entry=N00000032
  variants=reg
  number_type=magnitude
  power=3
}

{base=sexdecillion
  cat=number_word
  entry=N00000046
  variants=reg
  number_type=magnitude
  power=17
}

{base=twelve
  cat=number_word
  entry=N0000013
  variants=reg
  number_type=teen
  value=12
}
sixty four million four hundred thousand

<table>
<thead>
<tr>
<th>Multiplier</th>
<th>Head</th>
<th>Addition</th>
</tr>
</thead>
<tbody>
<tr>
<td>sixty Four</td>
<td>million</td>
<td>four hundred thousand</td>
</tr>
</tbody>
</table>

64 * 1,000,000 + 400,000 = 64,400,000
Lexical tools

SPECIALIST LEXICON

Text processing
Lexical Tools

• **Wordind** -- breaks strings into words
  – Produces the Metathesaurus word indexes (MRXW)
• **LVG** -- performs various lexical transformations
• **NORM** -- a selection of LVG transformations,
  – Used for Metathesaurus indexing
  – Produces the Metathesaurus Normalized word and string indexes (MRXNW & MRXNS)
  – Used to access those indexes
Normalization

- Hodgkin Disease
- HODGKINS DISEASE
- Hodgkin's Disease
- Disease, Hodgkin's
- HODGKIN'S DISEASE
- Hodgkin's disease
- Hodgkins Disease
- Hodgkin's disease NOS
- Hodgkin's disease, NOS
- Disease, Hodgkins
- Diseases, Hodgkins
- Hodgkins Diseases
- Hodgkins disease
- hodgkin's disease
- Disease;Hodgkins
- Disease, Hodgkin
- disease hodgkin