The SPECIALIST Lexicon and NLP Tools

August 19, 2007

Medinfo 2007

Allen Browne, Guy Divita, Anantha Bangalore, Karen Thorn, Chris Lu, Susanne Humphrey
US National Library of Medicine
Lister Hill Nation Center for Biomedical Communication

- Allen C. Browne
- Guy Divita
- Chris Lu
- Susanne Humphrey
- Anantha Bangalore
- Karen Thorn
Medinfo Tutorial S-023
SPECIALIST NLP Tools for UMLS Developers

• Introduction – Allen Browne
• The Specialist Lexicon – Allen Browne
• Lexical Tools – Chris Lu
• Text Tools – Guy Divita
• Text Categorization – Lu & Browne
• Knowledge Source Server – Anantha Bangalore
Lexical tools

SPECIALIST LEXICON

Text processing
The SPECIALIST Lexicon

- A syntactic lexicon
- Biomedical and general English
- Over 330,000 records
Frequency Spectrum:
Alice in Wonderland
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
Derivational Morphology

I suppose you could say I'm a dictionaryologist.

Dictionary+ology+ist
Inflectional Morphology

- octopus
- octopi
- octopuses

"Fellow octopi, or octopuses ... octopi? ... Dang, it's hard to start a speech with this crowd."
Orthography

Spelling Variation

- align -- aline
- Grave’s disease -- Graves’s disease -- Graves’ disease
- anesthetize -- anesthetise
- Esophagus -- oesophagus
- foetus – fetus
- centre -- center
Orthography

HAGAR THE HORRIBLE  CHRIS BROWNE

I NEED THE CORRECT SPELLING FOR THE RECORDS...

HOW DO YOU SPELL "CROCODILE"?

DR. ZOOK

©2000 by King Features Syndicate, Inc. World Rights reserved.
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
}
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
}
Orthographic vs. Lexicographic Word:

Why, for instance, if a two-word boy scout feels chilly on his one-word campground, does he pull up a two-word camp chair in front of his one-word campfire? Anyone who seeks a strictly logical answer to such questions is chasing will-o'-the-wisps (chargeable in telegrams as a single word, because of the hyphens) in a semantic bog.

UTF-8

{base=resume
 spelling VARIANT=résumé
 spelling VARIANT=resumé
 entry=E0053099
     cat=noun
     variants=reg
}

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

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

{base=cafe
 spelling VARIANT=café
 entry=E0420690
     cat=noun
     variants=reg
}
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*

Leach – Leaches
Stomach – Stomachs ➔ irregular
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 potassiums
• much potassium

* This form does not occur
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 \textit{s}.
  – \textit{y} becomes \textit{ie} following a consonant before \textit{s}.
  – \textit{e} is inserted between \textit{z}, \textit{x}, \textit{ch}, or \textit{sh} and \textit{s}.

• The past tense suffix is \textit{ed}.
  – \textit{y} becomes \textit{ie} following a consonant before \textit{ed}.
  – Final \textit{e} is deleted before \textit{ed}.

• The past participle is the same as the past tense.

• The present participle suffix is \textit{ing}.
  - \textit{y} becomes \textit{ie} following a consonant before \textit{ing}.
  - Final \textit{e} is deleted before \textit{ing}
    unless preceded by \textit{e}, \textit{y} or \textit{o}. 
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

control: controls, controlled, controlling
Irregular Verbs

{base=bite
  entry=E0013219
  cat=verb
  variants=irreg|bite|bites|bit|bitten|biting|
  intran
  tran=np
  cplxtran=np,advbl
  }

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;
# Noun rules (glreg)

<table>
<thead>
<tr>
<th>Rule</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>us$</td>
<td>noun</td>
<td>singular</td>
</tr>
<tr>
<td>antus</td>
<td>anti</td>
<td></td>
</tr>
<tr>
<td>ma$</td>
<td>noun</td>
<td>singular</td>
</tr>
<tr>
<td>a$</td>
<td>noun</td>
<td>singular</td>
</tr>
<tr>
<td>um$</td>
<td>noun</td>
<td>singular</td>
</tr>
<tr>
<td>on$</td>
<td>noun</td>
<td>singular</td>
</tr>
<tr>
<td>sis$</td>
<td>noun</td>
<td>singular</td>
</tr>
<tr>
<td>is$</td>
<td>noun</td>
<td>singular</td>
</tr>
<tr>
<td>men$</td>
<td>noun</td>
<td>singular</td>
</tr>
<tr>
<td>ex$</td>
<td>noun</td>
<td>singular</td>
</tr>
<tr>
<td>x$</td>
<td>noun</td>
<td>singular</td>
</tr>
</tbody>
</table>
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
pneumoconiosis: a pneumoconiosis caused by the inhalation of very fine silicate or quartz dust

The Protein of a tobacco mosaic virus, Dahlemense strain

acetylseryltyrosylserylisoleucylthreonylserylprolylserylglutaminylp
henylalanylvalylphenylalanyleucylserylserylvalyltryptophylalanyla
spartylprolylisoleucylglutamylleucylleucyllasparaginyllvalylcysteiny
lthreonylserylserylleucylglycllassparaginyllglutaminylphenylalanylgl
utaminyllthreonyllglutaminylglutaminylalanynlalanynlarginynlthreonyllthreonyll
alyltryptophylllysylprolylphenylalanynlprolylglycylaspartyllvalyltyrosyllsylvalyltyr
osylarginynltyrosylasparaginyllalanynlvalylleucylaspartyllprolylleucylis
oleucylthreonyllalanynlleucylleucylglyclylthryonylphenylalanynlaspary
lthreonyllarginyllasparaginyllarginynlisoleucylisoleucylglutamyllvalylgl
utamylasparaginyllglutaminylglutaminylserylprolylthreonyllthreonyllalanynlglutamyllthreonyllleucylaspartyllalanynlthreonyllarginyllarginynlval
ylaspartyllaspartyllalanynlthreonyllvalylalanynlisoleucylarginynlserylalan
ylanysparaginyllisoleucylasparaginyllleucylvallaasparaginyllglutamylleuc
ylvalylarginylglyclylthreonyllglyclylleucultyrosylaspartyllmethionylseryllglyc
ylleucylvalyltryptophylthreonyllserylalanynlprolylalanynlserine
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
ypsiform|adj|upsiloid|adj
wolfram|noun|tungsten|noun
double vision|noun|diplopia|noun
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
The SPECIALIST Lexicon and NLP Tools

August 19, 2007

Medinfo 2007

Allen Browne, Guy Divita, Anantha Bangalore, Chris Lu, Susanne Humphrey