Table of Contents

1. Overview
2. NLP Usage
1. Overview: The SPECIALIST Lexicon and NLP Tools

The SPECIALIST NLP Tools

The SPECIALIST LEXICON

LexBuild

Lexical Tools

LexCheck

LexAccess

LVG

Text Tools

Lexical Tool

NLP Applications

Sub-term (STMT)

SCRT
Life Cycle Development & Task Components

**Task Components**

<table>
<thead>
<tr>
<th>Task Components</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCR (Software Change Request)</td>
<td>Feature enhancements, bug fixes, library upgrades</td>
</tr>
<tr>
<td>Data updates &amp; integration</td>
<td>Lexical Tools, LexAccess, STMT, MNS</td>
</tr>
<tr>
<td>Tests</td>
<td>All projects</td>
</tr>
<tr>
<td>Documentation</td>
<td>Design doc, user doc, API doc</td>
</tr>
<tr>
<td>Release package &amp; backup</td>
<td>Annually &amp; Semi-annually</td>
</tr>
<tr>
<td>Web-site &amp; Web tools</td>
<td>All projects</td>
</tr>
<tr>
<td>Technical supports</td>
<td>All projects</td>
</tr>
<tr>
<td>Presentations</td>
<td>NLM summer Lectures, library associate talks, conferences</td>
</tr>
<tr>
<td>New research &amp; development</td>
<td>Paper publications</td>
</tr>
</tbody>
</table>
# Projects Maintenance Plan

<table>
<thead>
<tr>
<th>Category</th>
<th>Descriptions</th>
<th>Projects (13)</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Legacy Projects</td>
<td>gSpell, Text Tools, dTagger</td>
<td>Host project website</td>
</tr>
<tr>
<td>2</td>
<td>Completed Projects</td>
<td>Numbers, SCRT, VTT, TC</td>
<td>+Limited supports (approval required)</td>
</tr>
<tr>
<td>3</td>
<td>Supporting Projects</td>
<td>STMT</td>
<td>+Data updates (with UMLS release)</td>
</tr>
<tr>
<td>4</td>
<td>Core Supporting Projects</td>
<td>LB, LC</td>
<td>+Limited SCRs (approval required)</td>
</tr>
<tr>
<td>5</td>
<td>Core Projects</td>
<td>Lexicon, Lexical Tools, LexAccess</td>
<td>+Release &amp; supports</td>
</tr>
</tbody>
</table>
The SPECIALIST NLP Tools

The SPECIALIST LEXICON

LexBuild

Numbers

Lexical Tools

LexAccess

LVG

Text Tools

1

Sub-term

STMT

NLP Applications

2

3
LexBuild Process (Computer-Aided)

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

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

Build:
- LexBuild
- LexAccess
- LexCheck
Lexical Records

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

Lexical Records

- **{base=herpes zoster**
  - spelling_variant=herpes-zoster
  - entry=E0201295
    - cat=noun
    - variants=reg
    - variants=uncount
- **{base=zoster**
  - entry=E0066013
    - cat=noun
    - variants=uncount
- **{base=herpes**
  - entry=E0031440
    - cat=noun
    - variants=uncount

<table>
<thead>
<tr>
<th>Lexical Information</th>
<th>herpes</th>
<th>zoster</th>
<th>herpes zoster</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Part of speech</strong></td>
<td>noun</td>
<td>noun</td>
<td>noun</td>
</tr>
<tr>
<td><strong>Inflectional morphology</strong> (inflections)</td>
<td>herpes</td>
<td>1</td>
<td>zoster</td>
</tr>
<tr>
<td><strong>Orthography</strong></td>
<td>N/A</td>
<td>N/A</td>
<td>herpes-zoster</td>
</tr>
<tr>
<td><strong>Abbreviation/Acronym</strong></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Syntax (complementation)</strong></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Derivational morphology</strong> (derivations)</td>
<td>herpetic</td>
<td>postzoster</td>
<td>post-zoster</td>
</tr>
<tr>
<td><strong>LexSynonyms</strong> (Element Synonymous words for Query Expansion)</td>
<td>N/A</td>
<td>singles</td>
<td>zona</td>
</tr>
</tbody>
</table>

**Lexical Information**

- Base = herpes
  - Entry: E0031440
  - Category: noun
  - Variants: uncount

- Base = zoster
  - Entry: E0066013
  - Category: noun
  - Variants: uncount

- Base = herpes zoster
  - Spelling variant: herpes-zoster
  - Entry: E0201295
  - Category: noun
  - Variants: reg, uncount

**Abbreviation/Acronym**

- N/A

**Syntax (complementation)**

- N/A

**Derivational morphology** (derivations)

- herpetic
- postzoster
- post-zoster

**LexSynonyms** (Element Synonymous words for Query Expansion)

- N/A
- singles
- zona
- N/A
## What are Lexicon (Multi)Words?

<table>
<thead>
<tr>
<th>Descriptions</th>
<th>Examples</th>
</tr>
</thead>
</table>
| **Lexicon words** and **LexMultiwords (LMWs)** | • Inflection morphology  
• POS  
• Specific meaning  
• Word order  
• Space(s)  
• herpes zosters  
• frog erythrocytic virus, cardiac surgery  
• hot dog  
• trial and error, exercise training, training exercise, up and down  
• ice cream (ice-cream) |
| **Non-Words** (not in Lexicon) | • (Single) Words does not exist by itself, only exist in multiwords  
• non: non drug coated, persona non grata  
• vitro: in vitro, in vitro diagnostic  
• vivo: ex vivo, in vivo grown  
• intra: intra vires, intra articular route  
• etc. |
### LMW, MWE, UMLS String

<table>
<thead>
<tr>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMWs and MWE (Overlap)</td>
<td>• Phrasal position</td>
</tr>
<tr>
<td></td>
<td>• Adverb, adjective</td>
</tr>
<tr>
<td></td>
<td>• Fixed phrases (non-decomposable)</td>
</tr>
<tr>
<td></td>
<td>• because of, due to</td>
</tr>
<tr>
<td></td>
<td>• face down, in house</td>
</tr>
<tr>
<td></td>
<td>• kingdom come, by and large</td>
</tr>
<tr>
<td>LMWs and MWE (Difference)</td>
<td>• Collocation</td>
</tr>
<tr>
<td></td>
<td>• Specific meaning</td>
</tr>
<tr>
<td></td>
<td>• Complementation</td>
</tr>
<tr>
<td></td>
<td>• Idioms</td>
</tr>
<tr>
<td></td>
<td>• ing the house</td>
</tr>
<tr>
<td></td>
<td>• beat someone up, give birth</td>
</tr>
<tr>
<td></td>
<td>• kick the bucket, shoot the breeze</td>
</tr>
<tr>
<td>UMLS String:</td>
<td>• Term(^1) (conventionalized terminology)</td>
</tr>
<tr>
<td></td>
<td>• Phrase(^2)</td>
</tr>
<tr>
<td></td>
<td>• Food and water, pain and fever, [disease, Hodgkin’s]</td>
</tr>
<tr>
<td></td>
<td>• group 2, very low, may be a, not available</td>
</tr>
<tr>
<td></td>
<td>• right heart failure due to pulmonary hypertension</td>
</tr>
</tbody>
</table>

1. A word or group of words with a specific meaning, especially in a particular field.
2. A phrase is a group of words that expressing a thought, but lacking a subject or a verb or both.
Annual Lexicon Release Procedures

Freeze Lexicon (Early July)
- Validation (QA)
  - EUI, syntax, contents, citations, cross-reference, Illegal-ASCII, etc.
- Confirmation and Fixes
- Lexicon
- Derivation Generation
- LexSynonym Generation

Generate Lexicon Tables
- Lexicon XML
  - Lexicon Public Release
  - Lexicon Statistics
  - Lexicon Web Site/Document Updates

ASCII Lexicon Release
- Generate ASCII Lexicon Table
  - Generate MetaMap BDB Tables (1 table requires STMT updates)
Lexicon Growth – 2002 to 2016

- 491,639 lexical records
- 1,090,050 words (categories and inflections)
- 915,583 forms (spelling only)
  - Single words: 468,655 (51.19%); Multiwords: 446,928 (48.81%)
**Lexicon Unigram Coverage – Without WC**

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

<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>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.
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 x-axis represents word count class (1.E+02 to 1.E+08), and the y-axis represents term number (x 1,000). The graph compares Lexicon Single Word, Lexicon Multiword, and Acronym Expansion.]
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 $m$).
Lexical Tools

- Lexical Tools: Algorithm + Data (directly or derived from the Lexicon)
  - Command line tools
    - lvg (Lexical Variants Generation, base of all of tools)
    - norm (UMLS - MRXNS, MRXNW)
    - luiNorm (UMLS - LUI)
    - wordInd (UMLS - MRXNW)
    - toAscii (MetaMap - BDB Tables)
    - fields (Lexicon Tables, MetaMap - BDB Tables, etc.)
  - Lexical Gui Tool (lgt)
  - Web Tools
  - Java API’s
Lexical Tools Release Procedures

**Library Updates**
Java, Unicode, DB, Installations, etc. updates

**Component Updates:**
- GUI Tools
- Examples
- Documents
- Web Site/Web Tools

**Tests**
- Unit Test
- ASCII Test
- Platform Test
- Performance Test

**Software Change Requests (SCRs):**
bug fixes and feature enhancements

**Unit Tests**

**Lexical Tools Baseline**

**DB Integration - 1:**
- Generate DB tables from Lexicon
- Integrated to Lexical Tools

**DB Integration - 2:**
- Generate DB tables from LVG
- Integrated to Lexical Tools

**SD-Rules Optimization**

**Lexical Tools (Internal Release)**

**Lexical Tools (Public Release)**
2. NLP Usage

Terms (Phrasal units)

- Free Text (Clinical Note)
- Tokenizer
- Stemmer/Lemmatizer
- POS Tagger
- Chunker
- Concept Mapping
- Ranking WSD
- Phonology
- Morphology
- Orthography
- Lexicography (words)
- Syntax (sentences)
- Semantics

- derivations
- nominalization
- ACR/ABB
- synonyms
Example (PMID 23477346, TI):

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

  \[ \text{papillary thyroid carcinoma is a unique clinical entity.} \]
  \[ \text{papillary thyroid carcinoma is a unique} \]
  \[ \text{papillary thyroid carcinoma is a} \]
  \[ \text{papillary thyroid carcinoma is} \]
  \[ \text{papillary thyroid carcinoma -> Match} \]
  \[ \text{is a unique clinical entity} \]
  \[ \text{is a unique clinical} \]
  \[ \text{is a unique} \]
  \[ \text{is a} \]
  \[ \text{is} \]
  \[ \text{a unique clinical entity} \]
  \[ \text{...} \]
  \[ \text{papillary thyroid carcinoma is a unique clinical entity.} \]

C0238463 C1710548 C0205210 C1551338
Phrasal Level - Name Entity Recognition

- Example (PMID 23477346, TI):
  - Follicular variant of papillary thyroid carcinoma is a unique clinical entity.

- Lexicon Subterm Finder (LSF):
  Find all subterms that have mapped CUI
  - Load Lexicon (multi)words to Trie (normalized, have mapped CUI).
  - Retrieve subterms from Trie
## Processed Thesaurus: Lexicon (Multi)words - LSF

<table>
<thead>
<tr>
<th>Normalized* InflVars</th>
<th>EUI</th>
<th>CUI</th>
<th>STI</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>carcinoma</td>
<td>E0015213</td>
<td>C0007097</td>
<td>T191</td>
</tr>
<tr>
<td>clinical</td>
<td>E0017357</td>
<td>C0205210</td>
<td>T080</td>
</tr>
<tr>
<td>entity</td>
<td>E0025561</td>
<td>C1551338</td>
<td>T071</td>
</tr>
<tr>
<td>papillary</td>
<td>E0045335</td>
<td>C0205312</td>
<td>T080</td>
</tr>
<tr>
<td>papillary carcinoma</td>
<td>E0045337</td>
<td>C0007133</td>
<td>T191</td>
</tr>
<tr>
<td>papillary carcinomas</td>
<td>E0045337</td>
<td>C0007133</td>
<td>T191</td>
</tr>
<tr>
<td>papillary plasma flow</td>
<td>E0700280</td>
<td>No CUI</td>
<td>N/A</td>
</tr>
<tr>
<td>papillary thyroid carcinoma</td>
<td>E0637059</td>
<td>C0238463</td>
<td>T191</td>
</tr>
<tr>
<td>thyroid</td>
<td>E0060948</td>
<td>C0040132</td>
<td>T023</td>
</tr>
<tr>
<td>unique</td>
<td>E0063228</td>
<td>C1710548</td>
<td>T080</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

* Norm: Remove genitive, remove parenthetical plural forms (s), (es), (ies), remove punctuation, lower cases, trim, remove duplicated results
LSF – Load Lexicon (Multi)words to Trie

Normalized InflVars

- clinical
- entity
- papillary
- papillary carcinoma
- papillary carcinomas
- papillary plasma flow
- papillary thyroid carcinoma
- thyroid
- unique
- ...

Diagram:

- $Root$
  - $End$
  - clinical
  - papillary
  - unique
  - ...
  - carcinoma
  - carcinomas
  - thyroid
  - $End$
  - $End$
  - $End$
  - $End$
  - $End$
  - $End$

- $End$
  - $End$
  - $End$
  - $End$
  - $End$
papillary thyroid carcinoma is a unique clinical entity
papillary thyroid carcinoma is a unique clinical entity
papillary thyroid carcinoma is a unique clinical entity
LSF – Load Lexicon (Multi)words to Trie

- Example: papillary thyroid carcinoma is a unique clinical entity

TrieNode
- key: string
- child: hashMap<String, TrieNode>
- eui: list<String>
- euiCuiSti: list<euiCuiStiObj>
papillary thyroid carcinoma is a unique clinical entity.

Example: papillary thyroid carcinoma is a unique clinical entity

- thyroid
- clinical
- papillary
- unique
- carcinoma
- carcinomas
- thyroid
- carcinoma

E0637059
C0238463
E0063228
C1710548
E0017357
C0205210
E0025561
C1551338
Overlap Issues: Lead-Term, End-Term, Mid-Term?

Example (PMID 12792778, TI):
- Inhibition of metastatic brain tumor growth by intramuscular administration of the endostatin gene.
  - Lead-Term: metastatic brain tumor growth: C0220650
  - End-Term: metastatic brain tumor growth: C0598934
  - LMWs: metastatic brain tumor|C0220650, brain tumor|C0006118, tumor growth|C0598934, ..

Example (PMID 20162874, AB):
- In the present patient right pulmonary agenesis is co-occurring with VACTERL syndrome.
  - Lead-Term: the present patient right pulmonary agenesis: C0030706
  - End-Term: the present patient right pulmonary agenesis: C0265784
  - LMWs: patient right|C0030706, right pulmonary agenesis|C0265784, pulmonary agenesis|...
POS Parser Issues –> Sentence Level

- Example (PMID 9510650, TI):
  - **Shallow Parser’s (U of Illinois):**
    The changes of tear break up time after myopic excimer laser photorefractive keratectomy
  - **Parser (Stanford):**
    The changes of tear break up time after myopic excimer laser photorefractive keratectomy
  - **Multiword Approach:**
    The changes of tear break up time after myopic excimer laser photorefractive keratectomy

- **POS Parser Issues –> Sentence Level**
  - C2111106 Lacrimation tear break-up time
  - C2069669 excimer laser photorefractive keratectomy
Order Issues (Norm)

- Example (PMID 5820369, TI):
  - Cardiac arrest during exercise training.

- Example (PMID 14719633, AB):
  - Military training exercises are conducted routinely in the Mojave Desert.
Concept Ranking (CR) – The Longest Word?

Example (PMID 9510650, TI):
The changes of tear break up time after myopic excimer laser photorefractive keratectomy

<table>
<thead>
<tr>
<th>sub-term</th>
<th>Lexicon - EUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>tear</td>
<td>E0060021, E0060022</td>
</tr>
<tr>
<td>break</td>
<td>E0013997, E0013998</td>
</tr>
<tr>
<td>up</td>
<td>E0063423, E0063424, ...</td>
</tr>
<tr>
<td>time</td>
<td>E0061086, E0061087</td>
</tr>
<tr>
<td>break up</td>
<td>E0220309</td>
</tr>
<tr>
<td>break up time</td>
<td>E0635415</td>
</tr>
<tr>
<td>tear break up time</td>
<td>E0635418</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sub-term</th>
<th>Lexicon - EUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>myopic</td>
<td>E0041717, E0041718</td>
</tr>
<tr>
<td>excimer</td>
<td>E0319304</td>
</tr>
<tr>
<td>laser</td>
<td>E0036924, E0336817</td>
</tr>
<tr>
<td>photorefractive</td>
<td>E0418725</td>
</tr>
<tr>
<td>keratectomy</td>
<td>E0036428</td>
</tr>
<tr>
<td>excimer laser</td>
<td>E0514806</td>
</tr>
<tr>
<td>photorefractive keratectomy</td>
<td>E0225495</td>
</tr>
<tr>
<td>excimer laser photorefractive keratectomy</td>
<td>E0764487</td>
</tr>
</tbody>
</table>
CR on Overlap case – The Longest Word?

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

<table>
<thead>
<tr>
<th>sub-term</th>
<th>EUI</th>
<th>CUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>infected</td>
<td>E0034360, ..</td>
<td>C0439663</td>
</tr>
<tr>
<td>artificial</td>
<td>E0010589</td>
<td>C2004457</td>
</tr>
<tr>
<td>heart</td>
<td>E0030957</td>
<td>C0018787, ..</td>
</tr>
<tr>
<td>valve</td>
<td>E0063958</td>
<td>C0184252, ..</td>
</tr>
<tr>
<td>artificial heart</td>
<td>E0010602</td>
<td>C0018829</td>
</tr>
<tr>
<td>heart valve</td>
<td>E0030978</td>
<td>C0018826, ...</td>
</tr>
<tr>
<td><strong>artificial heart valve</strong></td>
<td>E0584205</td>
<td>C0018825, ...</td>
</tr>
</tbody>
</table>
CR: The Longest Word in the Sentence Level?

- Example (PMID 23477346, TI) – Beginning:
  - Follicular variant of papillary thyroid carcinoma is a unique clinical entity.

- Example (PMID 581461, TI) – Ending:
  - Nucleolar abnormalities in human papillary thyroid carcinoma.

- Example (PMID 6143549 , AB) – Middle:
  - Coexisting papillary thyroid carcinoma occurred in three patients with HCA.
Strings, MWEs, or Words

- Example of UMLS String (PMID 15528223, AB):
  - Right heart failure due to pulmonary hypertension causes significant morbidity and mortality.
    - UMLS String: right heart failure due to pulmonary hypertension|C1960038
    - Search: “Right heart failure because of pulmonary hypertension”
    - LMW: right heart failure|C0235527, pulmonary hypertension|C0020542

- Example of MWE (PMID 23477346, TI):
  - Thirty patients undergoing cardiac surgery and 7 patients undergoing thoracic surgery not involving the heart were studied.
    - MWE: Undergoing cardiac surgery (MWE), no meaning, no POS, no morphology
    - LMW: cardiac surgery (MWE): C0018821

- Strings and MWEs do not have all 4 criteria of LMWs: POS, morphology, order, specific meaning
Example (PMID: 47945, AB): Two of ten women whose blood contained Mycoplasma hominis gave birth to stillborn infants.

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

- Example (PMID: 47945, AB):
  Two of ten women whose blood contained Mycoplasma hominis gave birth to stillborn infants.

- Multiword: “Mycoplasma hominis”
- Patterns:
  - complementation: “gave birth”
  - Measurement: (0.1-2.3 mg/day)
  - title: Mr. Song, Dr. Coma
  - ...
- StopWords: preposition, auxiliary, modal, determiner, conjunction, complementizer, pronoun, numbers, ...

Diagram:
- Number Pattern: Two
- Pronoun: of ten women
- Mycoplasma hominis
- Light Verb: gave birth
- to
- Stillborn infants
- ppbr(to, NP)
- C0317838: Mycoplasma hominis
- C0005615: Birth
- C0595939: Stillbirth
Summary

Coverage:
• How many multiwords do we need?
• Query Expansion
• Hybrid: use multiword as default, single word as supplement

Other Resources (MWE? UMLS String?)
• Linguistics concerns: POS, morphology, order, specific meaning
• Technical concerns (size – performance):

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

- Single word approach vs. multiword approach:
  - Data driven – embedded knowledge (Facts instead of Rules)
  - The concept of approach (identify words), not the algorithm

- Other Components:
  - Morphology: inflections, uninflections, derivations
  - POS tagger (frog erythrocytic virus)
  - Norm (left pulmonary veins)
  - Query Expansion (zona vaccine)
  - Element Synonyms (zona -> herpes zoster)
  - Standard data set
  - etc.