

# *Generating a Distilled N-Gram Set*

## *Effective Lexical Multiword Building in the SPECIALIST Lexicon*

Presented By: Dr. Chris J. Lu  
2017.02.22  
(HEALTHINF, Session-6)

- Lexical Systems Group: <http://umlslex.nlm.nih.gov>
- The SPECIALIST NLP Tools: <http://specialist.nlm.nih.gov>
- Chris Lu (E): [chlu@mail.nih.gov](mailto:chlu@mail.nih.gov)

# Table of Contents

1. Introduction
  - What are Lexical Multiwords (LMWs)
  - Why LMWs?
2. Approach
  - MEDLINE N-gram Set
  - Filters and Matchers
3. Tests and Results
4. Applications
5. Conclusion and Future Work

# 1. Introduction

# The SPECIALIST Lexicon

- Lexicon: A fancy synonym for “dictionary”\*
- The SPECIALIST Lexicon:
  - A large syntactic lexicon of biomedical and general English
  - Designed and developed to provide lexical information needed for the SPECIALIST Natural Language Processing (NLP) system
  - Distributed in the Unified Medical Language System (UMLS) Knowledge Sources by the National Library of Medicine (NLM)

\* The Insomniac’s Dictionary, Ivy Books, 1986, by Paul Hellweg, p. 140

# Lexicon - Lexical Records

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

# Lexical Record Example-1

```
{base=seasickness
spelling_variant=sea sickness
spelling_variant=sea-sickness
entry=E0054880
  cat=noun
  variants=uncount
  compl=pphr(in,np)
  nominalization_of=seasick|a
dj|E0054879
}
```

Lexical Information	Example-1
<b>Base</b>	• seasickness
<b>Part of speech</b>	• noun
<b>Inflectional morphology</b>	• seasickness   singular
<b>Orthography (spelling variants)</b>	• sea sickness • sea-sickness
<b>Abbreviation/Acronym</b>	• N/A
<b>Syntax (complementation)</b>	• compl=pphr(of,np)
...	...
<b>Derivational morphology</b>	• <b>seasick   adj</b>
<b>LexSynonyms</b>	• <b>seasick</b> • <b>mal de mer</b>

# Lexical Record Example-2

```
{base=seasick
spelling_variant=sea sick
spelling_variant=sea-sick
entry=E0054879
  cat=adj
  variants=inv;periph
  position=attrib(1)
  position=pred
  stative
  nominalization=seasickness |
noun | E0054880
}
```

Lexical Information	Example-2
Base	• seasick
Part of speech	• adj
Inflectional morphology	• seasick   positive
Orthography	• sea sick • sea-sick
Abbreviation/Acronym	• N/A
Syntax (complementation)	• N/A
...	...
<b>Derivational morphology</b>	• <b>seasickness   noun</b>
<b>LexSynonyms</b>	• <b>mal de mer   noun</b> • <b>seasickness   noun</b> • ...

# Lexical Record Example-3

```
{base=mal de mer  
entry=E0432198  
  cat=noun  
  variants=uncount  
}
```

Lexical Information	Example-3
Base	• mal de mer
Part of speech	• noun
Inflectional morphology	• mal de mer   singular
Orthography	• N/A
Abbreviation/Acronym	• N/A
Syntax (complementation)	• N/A
...	...
Derivational morphology	• N/A
LexSynonyms	• sea sick   adj • sea sickness   noun • ...

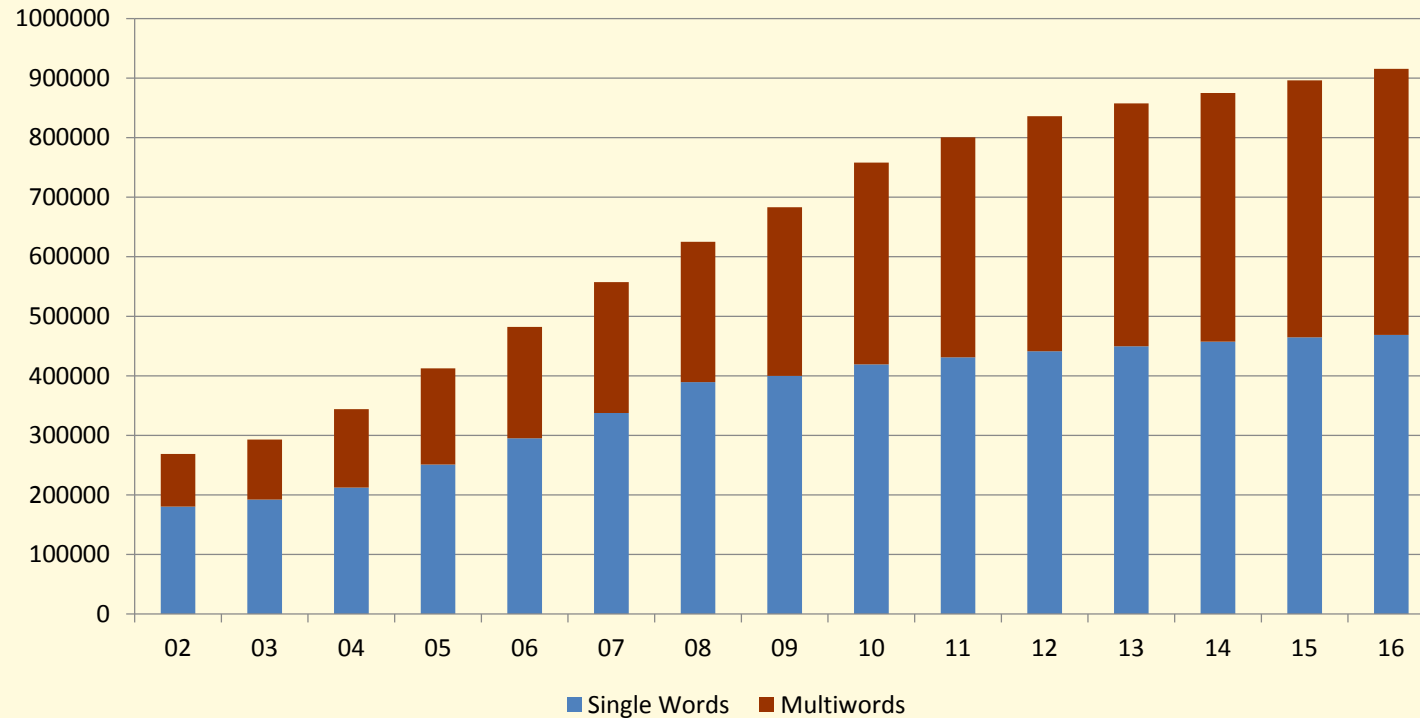


# Lexical Records Example 1-2-3

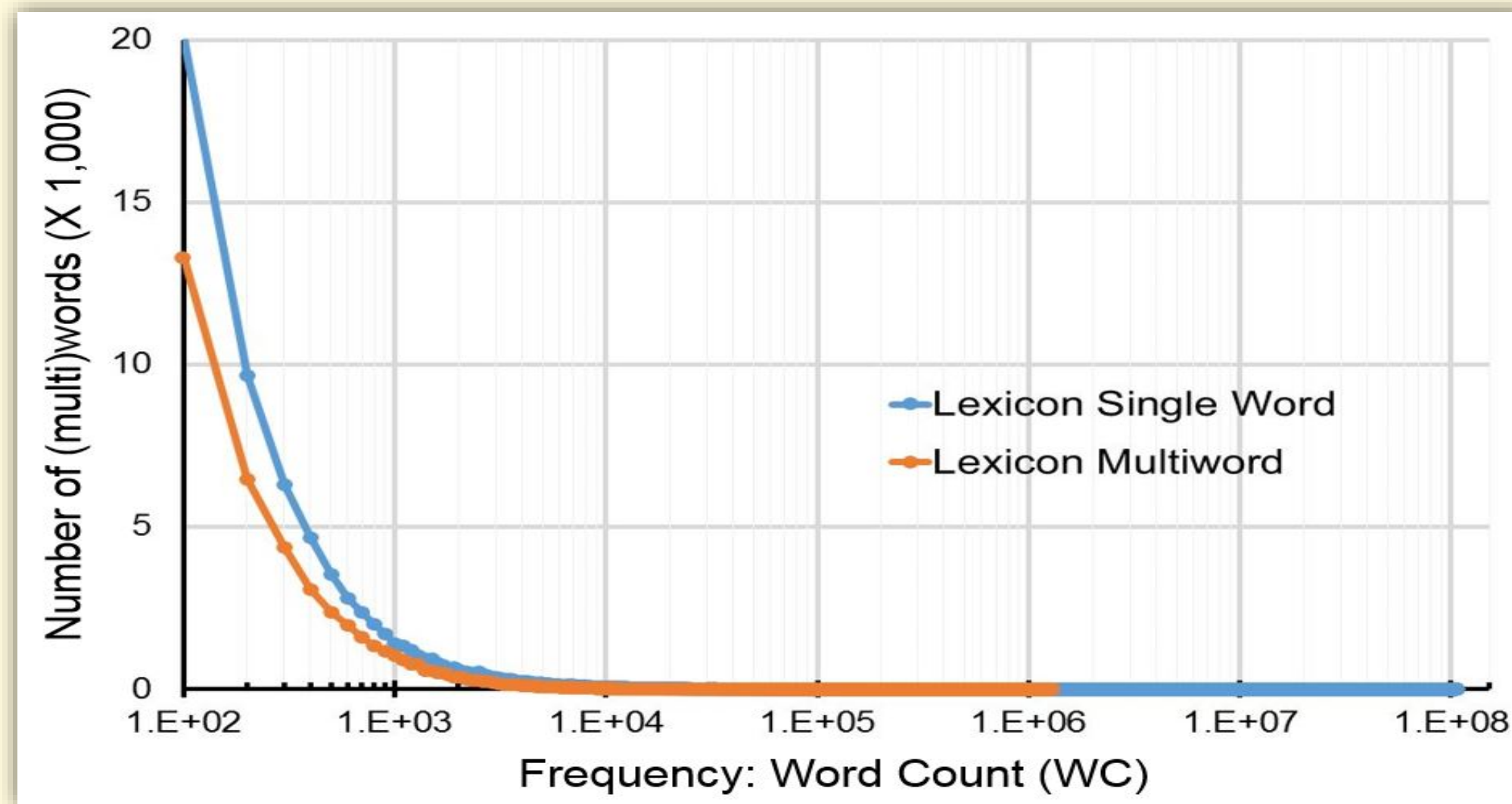
Lexical Information	Example-1	Example-2	Example-3
EUI	• E0054880	• E0054879	• E0432198
Base form of the (multi)word	• seasickness	• seasick	• mal de mer
Part of speech	• noun	• adj	• noun
Inflectional morphology	• seasickness   singular	• seasick   positive	• mal de mer   singular
Orthography	• sea sickness • sea-sickness	• sea sick • sea-sick	• N/A
Abbreviation/Acronym	• N/A	• N/A	• N/A
Syntax (complementation)	• compl=pphr(of,np)	• N/A	• N/A
...	...	...	...
Derivational morphology	• <b>seasick   adj</b>	• <b>seasickness   noun</b>	• <b>N/A</b>
LexSynonyms	• <b>seasick   adj</b> • <b>mal de mer   noun</b>	• <b>mal de mer   noun</b> • <b>seasickness   noun</b> • <b>...</b>	• <b>sea sick   adj</b> • <b>sea sickness   noun</b> • <b>...</b>

# Lexicon Growth – 2002 to 2016

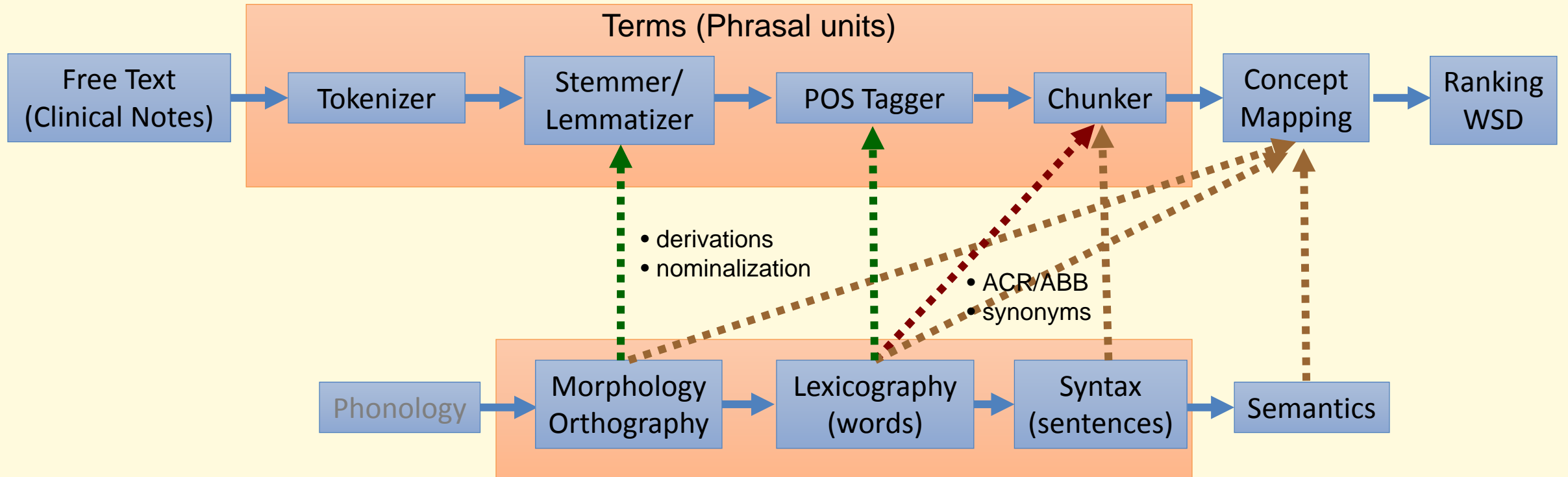
- 491,639 lexical records
- 1,090,050 words (categories and inflections)
- 915,583 forms (spelling only)
  - Single words: 468,655 (51.19%); Multiwords: 446,928 (48.81%)



# The Frequency Spectrum of Lexicon on MEDLINE



# NLP Pipeline



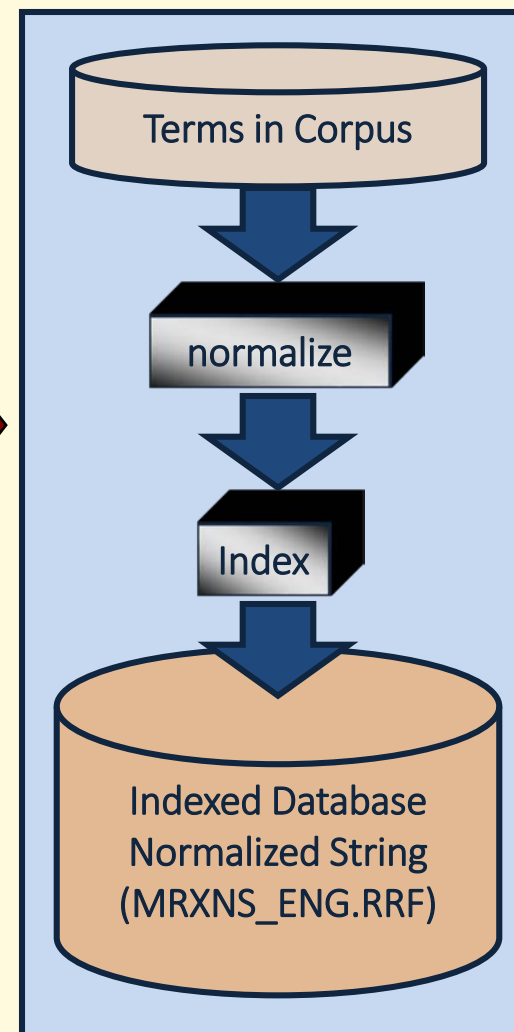
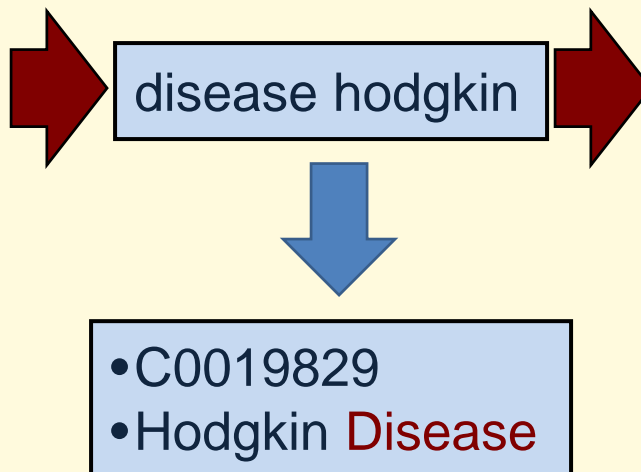
# Words and Multiwords

- 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, pain and fever)
    - exercise training vs. training exercise (military)

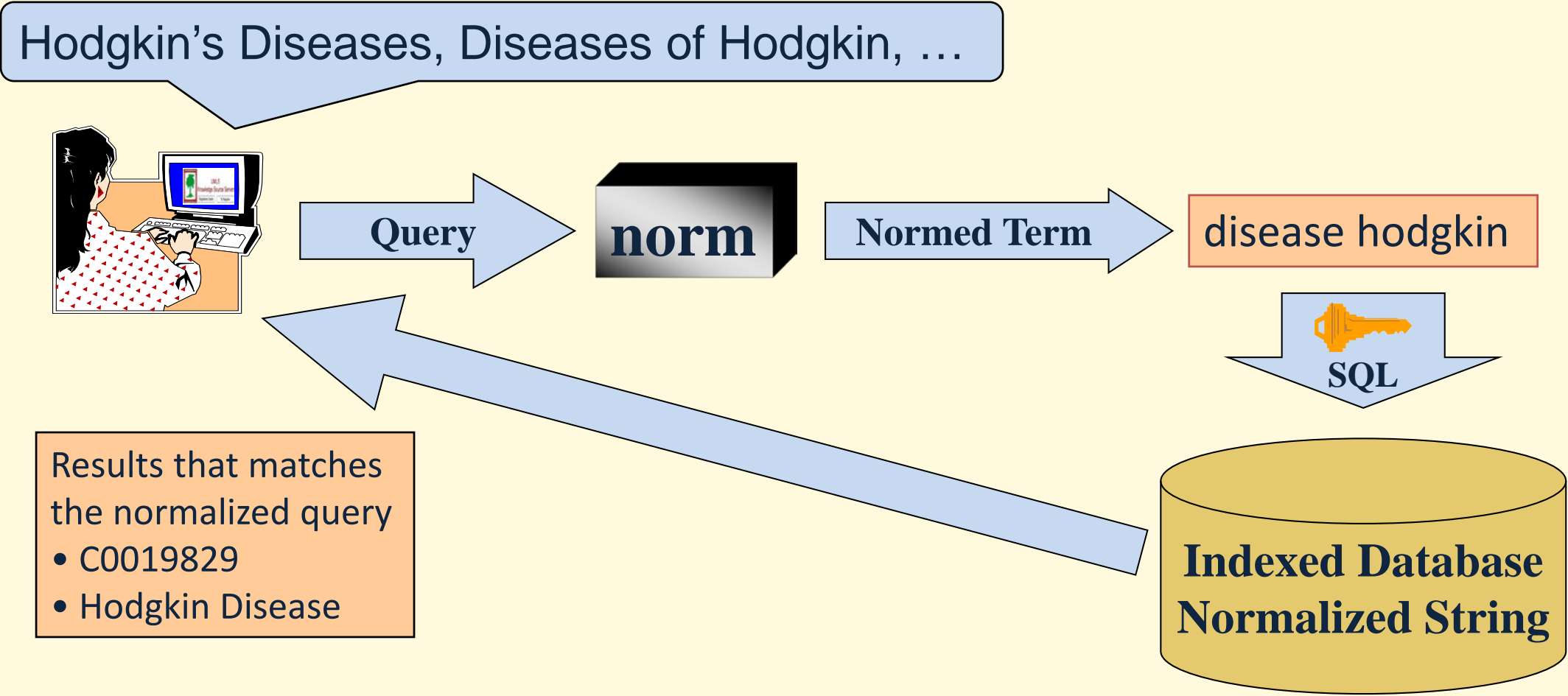


# Criterion 2 – Norm (Uninflection)

- 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
- ...



# Criterion 2 – Norm (Uninflection)





# Criterion 2 – Inflection/Uninflection

➤ Example (PMID 14479709, TI):  
Complete anomalous drainage of left pulmonary veins.

- **Single Word approach** (norm - uninflection):

- left pulmonary vein
- **leave pulmonary vein**

- **Multiword Approach:**

- left pulmonary vein

Inflectional Variants	Adjective (Positive)	Adverb (Positive)	Noun (Singular)	Verb (Infinitive)
left	left	left	left	leave
pulmonary	pulmonary			
veins			vein	vein
left pulmonary veins			left pulmonary vein	

# Criterion 3 – Meaning: Name Entity Recognition (NER)

- Example (PMID 23477346, TI) – Lead-term:
  - Follicular variant of papillary thyroid carcinoma is a unique clinical entity.
- Example (PMID 581461, TI) – End-term:
  - Nucleolar abnormalities in human papillary thyroid carcinoma.
- Example (PMID 6143549 , AB) – Mid-term:
  - Coexisting papillary thyroid carcinoma occurred in three patients with HCA.

E0637059

C0238463  
Papillary thyroid carcinoma

# Criterion 3 – Meaning: Overlap Window

## ➤ 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|...

# Criterion 4 – Word Order (Norm)

- Example (PMID 5820369, TI):

Cardiac arrest during **exercise training**.

E0015232|noun

E0566972|noun

C4279936|Exercise Training

- Example (PMID 14719633, AB):

Military **training exercises** are conducted routinely in the Mojave Desert.

E0359719|noun

E0764715|Noun

C4279936|Exercise Training

?

# LMWs and MWEs

- LMWs are a subset of MWEs (Multiword Expressions)
  - MWEs that have specific inflection morphology, POS, meaning, and word order are LMWs (except for complementation and idioms)

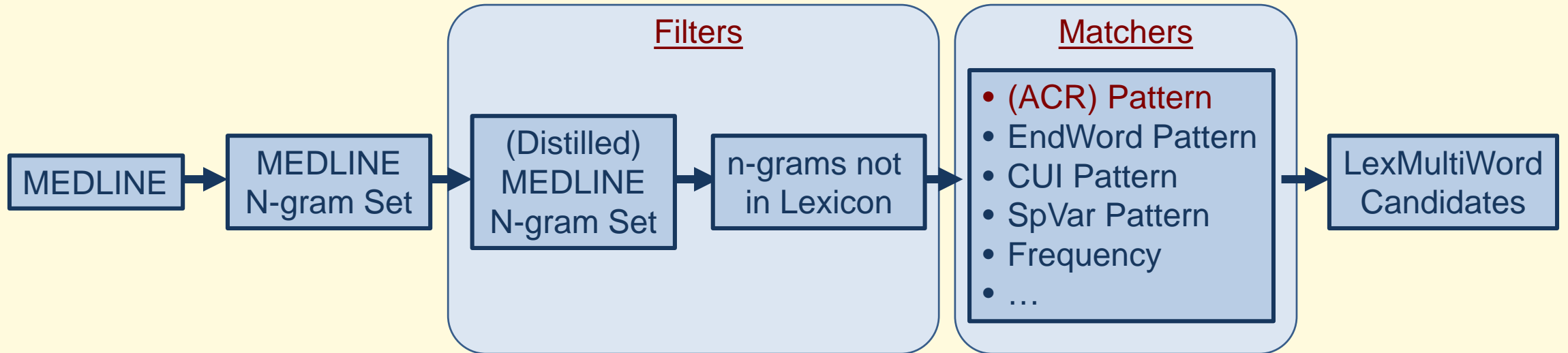
	Descriptions	Examples
LMWs and MWE (Overlap)	<ul style="list-style-type: none"><li>• Phrasal position</li><li>• Adverb, adjective</li><li>• Fixed phrases (non-decomposable)</li></ul>	<ul style="list-style-type: none"><li>• because of, due to</li><li>• face down, in house</li><li>• kingdom come, by and large</li></ul>
LMWs and MWE (Difference)	<ul style="list-style-type: none"><li>• Collocation</li><li>• Specific meaning</li><li>• Complementation</li><li>• Idioms</li><li>• ...</li></ul>	<ul style="list-style-type: none"><li>• undergoing <u>cardiac surgery</u></li><li>• <u>in the house</u></li><li>• <u>beat</u> someone <u>up</u>, <u>give birth</u></li><li>• kick the bucket, shoot the breeze</li></ul>

# Project Objective

- Effective lexical multiword building
  - Develop a systematic approach
  - Add multiwords from the latest MEDLINE to the SPECIALIST Lexicon
  - Generate high precision multiword candidate list

## 2. Approach

# LMWs - Processes





# The MEDLINE N-gram Set – Specifications\*

N-grams	2014	2015	2016
MEDLINE files	1-746	1-779	1-812
Max. length	50	50	50
Min. WC	30	30	30
Min. DC	1	1	1
Total documents	22,356,869	23,343,329	24,358,442
Total sentences	126,612,705	134,834,507	143,471,776
Total tokens	2,610,209,406	2,786,085,158	2,971,013,236

\* Generating the MEDLINE N-Gram Set,

Lu, Chris J.; Tormey, Destinee; McCreedy, Lynn; and Browne, Allen C.,  
AMIA 2015 Annual Symposium, San Francisco, CA, November 14-18, 2015, P1569

- Five-grams: covers 99.38% multiwords in Lexicon.2016
- Max. Word Length: 50 to cover 99.46% words in Lexicon.2016

# The MEDLINE N-gram Set

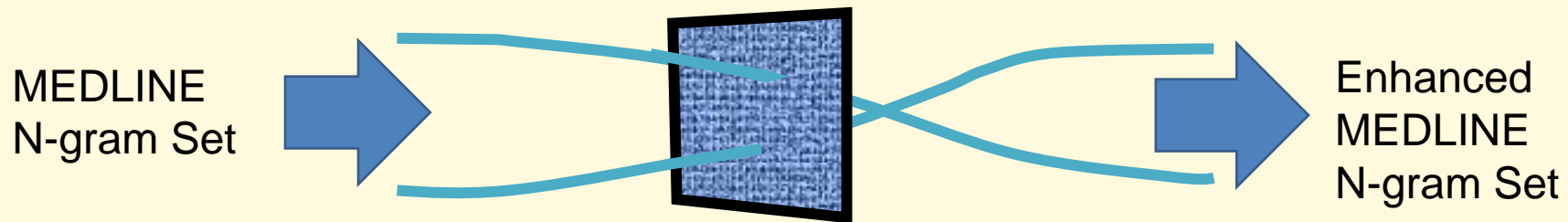
➤ Annual Public Releases: <http://umlslex.nlm.nih.gov/nGram>

N-grams	2014	2015	2016
unigrams	804,382	843,206	883,287
bigrams	4,587,349	4,845,965	5,114,547
trigrams	6,287,536	6,702,194	7,134,807
four-grams	3,799,377	4,082,612	4,380,474
five-grams	1,545,175	1,674,715	1,812,223
n-gram set	17,023,819	18,148,692	19,325,338

# Enhancing N-gram Set

- 17 ~ 19 million is a big number
  - Includes many n-grams that are invalid lexical terms
    - Ex: “increased significantly|78545”
  - Too big for further processes by computer programs with complicated algorithm
- Reduce the size by filtering out invalid multiwords:
  - increase precision
  - without sacrificing recall
  - the distilled MEDLINE n-gram set

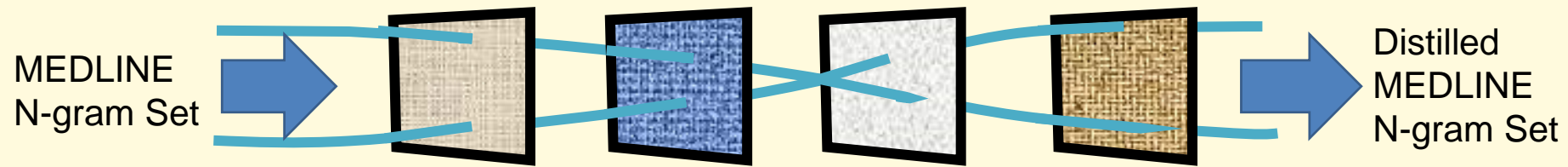
# Exclusive Filter



	Trap (not retrieved)	Pass (retrieved)
Valid (relevant)	FN	TP
Invalid (not relevant)	TN	FP

- Filter passing rate = pass-through terms / total terms
- Recall =  $TP / TP + FN$
- **Recall Test:** apply filters on Lexicon.2016 (valid 915,583 word set where TN & FP are 0)
  - Recall =  $TP / TP + FN$   
= pass-through terms / total terms  
= passing rate

# Serial Filters (High Recall)



	N-gram	Filter-1	Filter-2	...	Filter-N	Distilled
Valid (TP)	TP <sub>0</sub>	TP <sub>1</sub>	TP <sub>2</sub>	...	TP <sub>x</sub>	TP <sub>x</sub>
Invalid (FP)	FP <sub>0</sub>	FP <sub>1</sub>	FP <sub>2</sub>	...	FP <sub>x</sub>	FP <sub>x</sub>

- Applied filters with high recall rate ( $TP_0 \cong TP_1 \cong \dots \cong TP_x$ ;  $FP_0 > FP_1 > \dots > FP_n$ )
- N-gram Precision  $_x = TP_x / (TP_x + FP_x)$ 
  - $\cong TP_0 / (TP_0 + FP_x)$  .....  $TP_x$  is similar to  $TP_0$  (high recall)
  - $> TP_0 / (TP_0 + FP_0)$  .....  $FP_0$  is bigger than  $FP_n$  (high recall)
  - $> Precision_0$
- N-gram Recall  $_x = TP_x / (TP_x + FN_x)$ 
  - $= TP_x / (TP_x + FN_0)$  .....  $FN_n$  is same as  $FN_0$ , a constant
  - $\cong TP_0 / (TP_0 + FN_0)$  .....  $TP_n$  is similar to  $TP_0$  (high recall)
  - $\cong Recall_0$

# Independent High Recall Filters

Filter Type	High Recall Filter
General Exclusive Filters (5)	<ul style="list-style-type: none"><li>• <a href="#">Pipe</a></li><li>• <a href="#">Punctuation or space</a></li><li>• <a href="#">Digit</a></li><li>• <a href="#">Number</a></li><li>• <a href="#">Digit and Stopword</a></li></ul>
Pattern Exclusive Filters (6)	<ul style="list-style-type: none"><li>• <a href="#">Parenthetic acronym - (PAP)</a></li><li>• <a href="#">Indefinite Article</a></li><li>• <a href="#">UPPERCASE Colon</a></li><li>• <a href="#">Disallowed Punctuation</a></li><li>• <a href="#">Measurement</a></li><li>• <a href="#">Incomplete</a></li></ul>
Lead-End-Term Exclusive Filters (5)	<ul style="list-style-type: none"><li>• <a href="#">Absolute Invalid Lead-Term</a></li><li>• <a href="#">Absolute Invalid End-Term</a></li><li>• <a href="#">Lead-End-Term</a></li><li>• <a href="#">Lead-Term No SpVar</a></li><li>• <a href="#">End-Term No SpVar</a></li></ul>

# 1. General Exclusive Filters

Filter	Recall Test* On Lexicon (Trapped)	Passing Rate On MNS (Trapped)	Cumulative Passing Rate On MNS	Trapped Examples
<a href="#">Pipe</a>	100.0000% (0)	100.0000% (7)	100.0000%	<ul style="list-style-type: none"> <li>• 38 44 ( r </li> <li>• 33 37 Ag AgCl</li> </ul>
<a href="#">Punctuation or space</a>	100.0000% (0)	99.9977% (425)	99.9978%	<ul style="list-style-type: none"> <li>• 1259147 3690494 =</li> <li>• 604567 2377864 +/-</li> </ul>
<a href="#">Digit</a>	99.9999% (1)	99.3136% (132,650)	99.3114%	<ul style="list-style-type: none"> <li>• 1404799 2062240 2</li> <li>• 239725 499064 95%</li> </ul>
<a href="#">Number</a>	99.9953% (41)	99.9775% (4,326)	99.2890%	<ul style="list-style-type: none"> <li>• 2463066 3359594 two</li> <li>• 18246 20674 first and second</li> </ul>
<a href="#">Digit and Stopword</a>	99.9991% (8)	99.1777% (157,786)	98.4725%	<ul style="list-style-type: none"> <li>• 3155416 4125616 on the</li> <li>• 11180 12722 1, 2, and</li> </ul>

\* Recall test on Lexicon.2016: 915,583 words

## 2. Pattern Exclusive Filters

Filter	Recall Test* On Lexicon (Trapped)	Passing Rate On MNS (Trapped)	Cumulative Passing Rate On MNS	Trapped Examples
<a href="#">Parenthetic acronym - (PAP)</a>	100.0000% (0)	98.9647% (197,022)	97.4530%	<ul style="list-style-type: none"> <li>• 33117   33381   chain reaction (PCR)</li> <li>• 30095   30315   polymerase chain reaction (PCR)</li> </ul>
<a href="#">Indefinite Article</a>	99.9986% (13)	98.1713% (344,403)	95.6709%	<ul style="list-style-type: none"> <li>• 270384   292590   a case</li> <li>• 40271   40512   A series</li> </ul>
<a href="#">UPPERCASE Colon</a>	99.9999% (1)	99.3838% (113,936)	95.0813%	<ul style="list-style-type: none"> <li>• 2069343   2070116   RESULTS:</li> <li>• 18015   18016   AIM: The</li> </ul>
<a href="#">Disallowed Punctuation</a>	99.9986% (13)	99.2625% (135,508)	94.3801%	<ul style="list-style-type: none"> <li>• 324405   719011   (n =</li> <li>• 86525   133350   (P &lt; 0.05)</li> </ul>
<a href="#">Measurement</a>	99.9920% (73)	98.1572% (336,112)	92.6409%	<ul style="list-style-type: none"> <li>• 154905   181001   two groups</li> <li>• 12160   15197   10 mg/kg</li> </ul>
<a href="#">Incomplete</a>	100.0000% (0)	99.0708% (166,356)	91.7801%	<ul style="list-style-type: none"> <li>• 482021   1107869   (P</li> <li>• 25347   25992   years) with</li> </ul>

\* Recall test on Lexicon.2016: 915,583 words



# 3. Lead-End-Terms\* Exclusive Filters

Filter	Recall Test* On Lexicon (Trapped)	Passing Rate On MNS (Trapped)	Cumulative Passing Rate On MNS	Trapped Examples
<a href="#">Absolute Invalid Lead-Term</a>	99.9943% (52)	73.4329% (4,712,162)	67.3967%	<ul style="list-style-type: none"> <li>• 2780043   3451203   of a</li> <li>• 432921   434591   this study was</li> </ul>
<a href="#">Absolute Invalid End-Term</a>	99.9997% (3)	79.1897% (2,710,470)	53.3713%	<ul style="list-style-type: none"> <li>• 1878109   3534031   patients with</li> <li>• 1062545   1261445   between the</li> </ul>
<a href="#">Lead-End-Term</a>	99.9992% (7)	99.9739% (2,687)	53.3573%	<ul style="list-style-type: none"> <li>• 2578756   3106139   in a</li> <li>• 1733   1744   For one</li> </ul>
<a href="#">Lead-Term No SpVar</a>	<b>99.9913%</b> <b>(80)</b>	85.9342% (1,450,394)	45.8522%	<ul style="list-style-type: none"> <li>• 658430   708246   to determine</li> <li>• 533913   554628   In addition,</li> </ul>
<a href="#">End-Term No SpVar</a>	99.9968% (29)	83.5433% (1,458,246)	<b>38.3064%</b>	<ul style="list-style-type: none"> <li>• 1009451   1295670   number of</li> <li>• 726   734   (HPV) in</li> </ul>

Cum. Recall: 99.96%

\* Recall test on Lexicon.2016: 915,583 words

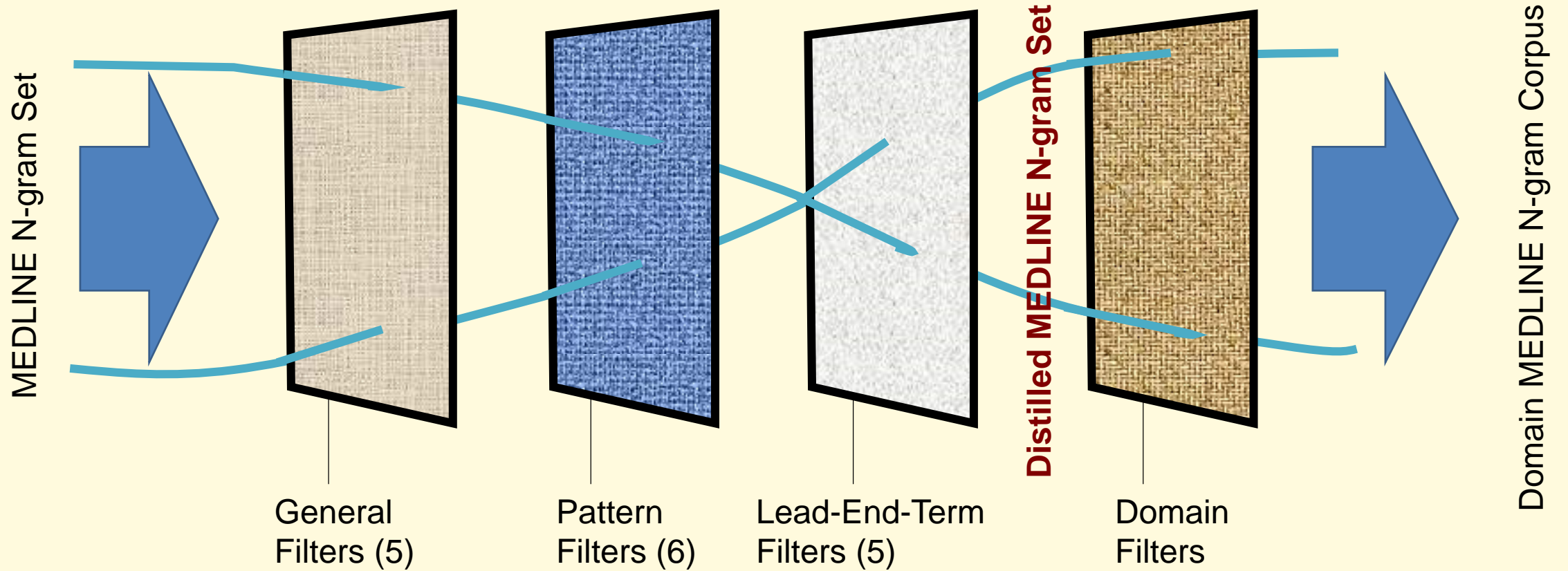
\*\* Lead-end-terms: derived from auxiliaries (be, do), complementizer (that), conjunctions (and, or, but) determiners (a, the, some), modals (may, must, can), pronoun (it, he, they), preposition (to, on, by)

# The Distilled MEDLINE N-gram Set (DMNS)

➤ Available to public: <http://umlslex.nlm.nih.gov/nGram>

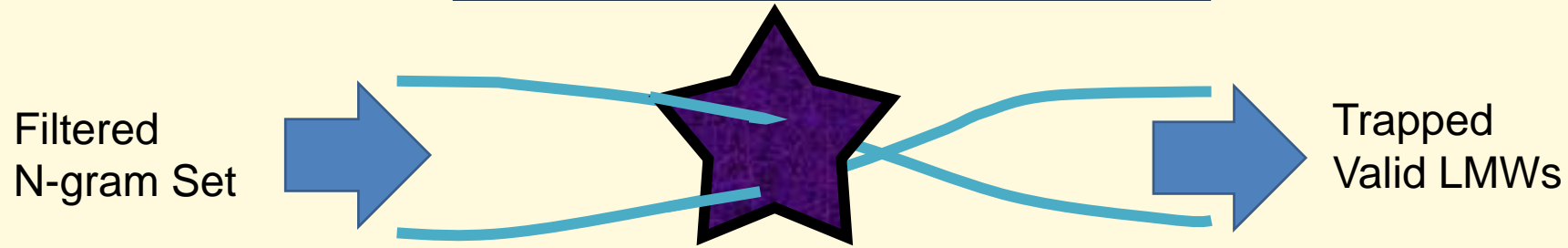
N-grams	2014	2015	2016
unigrams	804,382	843,206	883,287
bigrams	4,587,349	4,845,965	5,114,547
trigrams	6,287,536	6,702,194	7,134,807
four-grams	3,799,377	4,082,612	4,380,474
five-grams	1,545,175	1,674,715	1,812,223
N-gram Set	17,023,819	18,148,692	19,325,338
<b>Distilled N-gram Set</b>	<b>6,351,392</b>	<b>6,793,561</b>	<b>7,402,848</b>
Passing Rate	37.31%	37.43%	38.30%

# Distilled N-gram Set



# 3. Tests and Results

# Test: PAP Matcher



	Trap (retrieved)	Pass (not retrieved)
Valid (relevant)	TP	FN
Invalid (not relevant)	FP	TN

## ➤ Parenthetical Acronym Pattern (PAP) Matcher

Process	Example-1	Example-2	Example-3
1. Apply PAP Filter	clear cell sarcoma (CCS)	cell sarcoma (CCS)	Unified Health System (SUS)
2. Retrieve expansion	clear <u>cell sarcoma</u>	cell sarcoma	<del>Unified Health System</del>
3. Remove invalid expansion	clear cell sarcoma	<del>cell sarcoma</del>	

# Evaluation of Distilled MEDLINE N-gram Set

## ➤ Test on MNS and DMNS

- generate 17,707 LMW candidates from PAP matchers on MNS.2016,
- manually tagged and reviewed by 3 linguists: TP (15,850), FP (1,857)

Case	Test Case - Model	TP	FP	FN	TN	Precision	Recall	F1
1	PAP matcher on MNS (baseline)	15,850	1,857	0	0	0.8951	(1.0000)	0.9447
2	PAP matcher on DMNS (16 filters)	15,840	1,299	10	558	0.9242	0.9994	0.9603

## ➤ Similar results on the recall test and PAP matcher test for 2014-2016

	2014	2015	2016
Recall test: cumulative passing rate	37.31%	37.43%	38.30%
Recall test: cumulative recall (product)	99.97%	99.97%	99.96%
PAP Matcher Test: recall	99.96%	99.94%	99.94%
PAP Matcher Test: precision	94.31%	91.89%	92.42%

# 4. Applications

# Spelling Variant Pattern (SVP) Matcher

➤ Utilize SpVar Norm, Metaphone, edit distance, sorted distance, etc.

Step	Algorithm	ED	Examples	SpVarNorm	Metaphone	ED	SD	Recall On Lexicon	Run Time
1	SpVar Norm	N/A	<ul style="list-style-type: none"> <li>• <i>St. Anthony's fire</i></li> <li>• <i>Saint Anthony's fire</i></li> <li>• <i>lamin-A</i></li> <li>• <i>lamin A</i></li> </ul>	<ul style="list-style-type: none"> <li>• <i>Saint Anthony's fire</i></li> <li>• <i>Lamin A</i></li> </ul>				0.8050	1 min.
2	MES	2	<ul style="list-style-type: none"> <li>• <i>yuppie flu</i></li> <li>• <i>yuppy flu</i></li> </ul>	<ul style="list-style-type: none"> <li>• <i>yuppieflu</i></li> <li>• <i>yuppyflu</i></li> </ul>	<ul style="list-style-type: none"> <li>• [YPFL]</li> </ul>	2	✓	0.9792	7 hr
3	ES	1	<ul style="list-style-type: none"> <li>• <i>zincemia</i></li> <li>• <i>zincaemia</i></li> </ul>	<ul style="list-style-type: none"> <li>• <i>zincemia</i></li> <li>• <i>zincaemia</i></li> </ul>	<ul style="list-style-type: none"> <li>• [SNSM]</li> <li>• [SNKM]</li> </ul>	1	✓	0.9931	23 hr
4	MES	3	...	...	...	...	...	0.9940	8 min
5	ES	2	...	...	...	...	...	0.9970	26 hr
6	MES	4	...	...	...	...	...	0.9972	2 min

56 hr

\* Lexicon: 0.9 M (56 Hr); MNS 20 M (too long); DMNS 7.4 M (20 days)

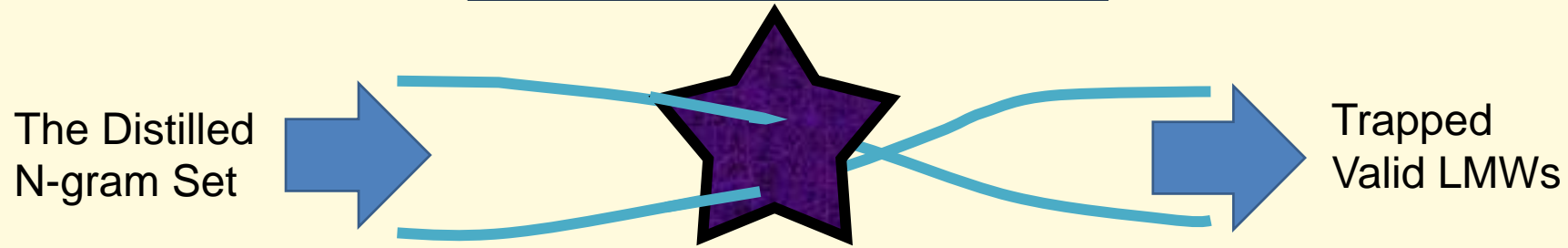


# Practice Results

➤ Baseline: 17,707 LMW Candidates from (ACR) matcher, tagged by linguists

Case	Test Case - Model	TP	FP	FN	TN	Precision	Recall	F1	Accuracy
1	PAP matcher on MNS (baseline)	15,850	1,857	0	0	0.8951	(1.0000)	0.9447	0.8951
2	PAP matcher on DMNS (16 filters)	15,840	1,299	10	558	0.9242	0.9994	0.9603	0.9261
3	SpVar pattern matcher on case 2	8,094	499	7,756	1,358	0.9419	0.5107	0.6623	0.5338

# Other Matchers



- Metathesaurus CUI Pattern Matcher
  - LMW candidate if a term has CUI(s)
  - Apply STMT to retrieve CUIs (2 subterm substitutions by their synonyms)
- EndWord pattern Matcher
  - syndrome: “migraine syndrome”, “contiguous gene syndrome”, etc.
  - disease: “Fabry disease”, “Devic disease”, etc.
  - 33 high frequency end-terms
- Frequency Strategy
- Combination of above models

# Practice Results

➤ Baseline: 17,707 LMW Candidates from (ACR) matcher, tagged by linguists

Case	Test Case - Model	TP	FP	FN	TN	Precision	Recall	F1	Accuracy
1	PAP matcher on MNS (baseline)	15,850	1,857	0	0	0.8951	(1.0000)	0.9447	0.8951
2	PAP matcher on DMNS (16 filters)	15,840	1,299	10	558	0.9242	0.9994	0.9603	0.9261
3	SpVar pattern matcher + Distilled	8,094	499	7,756	1,358	0.9419	0.5107	0.6623	0.5338
4	Metathesaurus CUI pattern matcher	10,056	755	5,794	1,102	0.9302	0.6344	0.7544	0.6301
5	EndWord pattern matcher (Top 33)	2,346	251	13,504	1,606	0.9034	0.1408	0.2544	0.2232
6	<b>Distilled + CUI + SpVar</b>	<b>5,892</b>	<b>212</b>	<b>9,958</b>	<b>1,645</b>	<b>0.9653</b>	<b>0.3717</b>	<b>0.5368</b>	<b>0.4257</b>
7	<b>Distilled + CUI + SpVar + EndWord (33)</b>	<b>992</b>	<b>15</b>	<b>14,858</b>	<b>1,842</b>	<b>0.9851</b>	<b>0.0626</b>	<b>0.1177</b>	<b>0.1600</b>
8	<b>Distilled + CUI + EndWord (33)</b>	<b>1766</b>	<b>113</b>	<b>14,084</b>	<b>1,744</b>	<b>0.9399</b>	<b>0.1114</b>	<b>0.1992</b>	<b>0.1982</b>

# Summary

- Developed high recall filters (> 99.99% on Lexicon recall test)
- Obtained the distilled MEDLINE n-gram set at passing rate of 37-38%
  - smaller data set
  - better precision
  - similar recall
- Applied filters and matchers to generate LMW candidate list
- Improve lexBuilding on multiwords (> 40% efficiency improvement)
  
- Distribute the MEDLINE n-gram set (2014+) to public
- Distribute the Distilled MEDLINE n-gram set (2014+) to public
- Generic filters for other corpora

# Conclusion

- Multiwords are pervasive, challenging and vital to NLP
- Build a biomedical Lexicon to support NLP and MLP
  - POS
  - Morphology (inflections and derivations)
  - Specific meaning
  - Specific order
- Multiword Approach (embedded information)

# Future Work

- Continuously develop filters and matchers for LexBuilding on multiwords
- Develop different matchers for better LMW candidate list
- Develop new SPECIALIST NLP Tools based on multiword approach

# Acknowledgements

- Supported by the Intramural Research Program of the NIH/NLM
- Co-authors (NLM):
  - Destinee Tormey
  - Dr. Lynn McCreedy
  - Allen C. Browne
- Mr. Guy Divita (Utah University)

# Questions



- Lexical Systems Group: <http://umlslex.nlm.nih.gov>
- SPECIALIST NLP Tools: <http://specialist.nlm.nih.gov>
- Chris Lu (E): [chlu@mail.nih.gov](mailto:chlu@mail.nih.gov)