The SPECIALIST Lexicon and NLP Tools
(Consumer Spelling Tool - CSpell)

By: Dr. Chris J. Lu

NLM – LHNCBC - CGSB

June, 2018

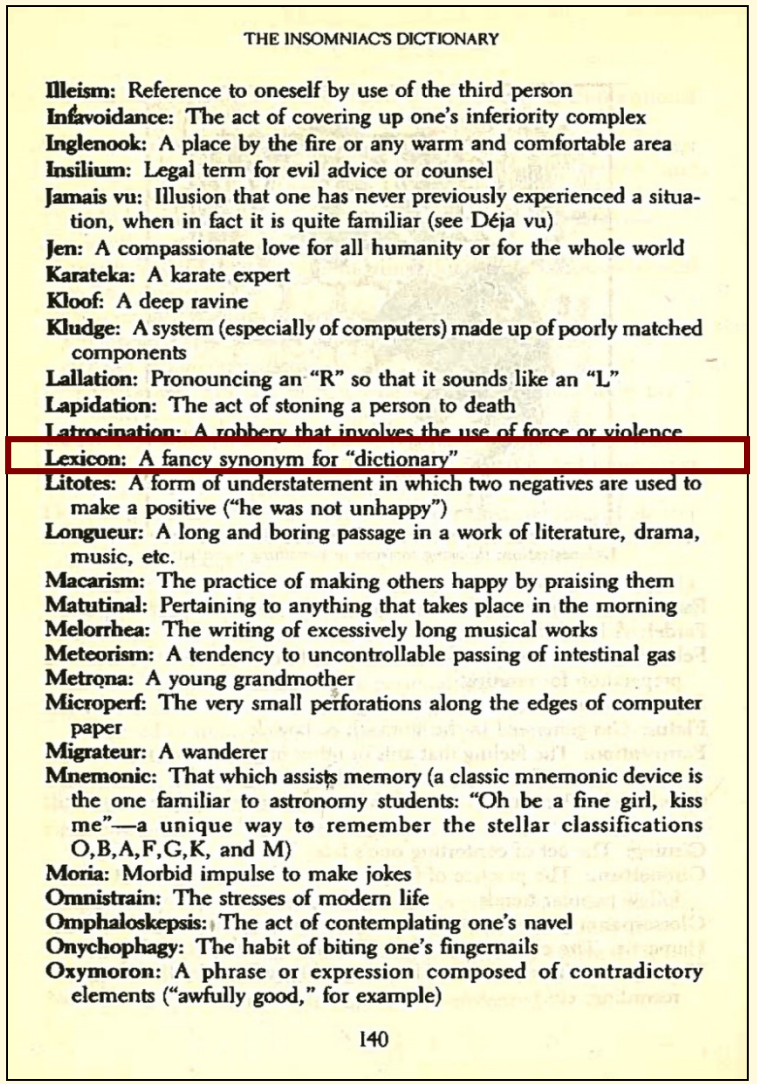
- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov

Outline

- Introduction
  - The SPECIALIST Lexicon
  - The SPECIALIST NLP Tools (Lexical Tools)
- Applications - CSpell
  - Natural Language Processing (NLP)
  - CSpell (Consumer Spelling Tool)
- Questions (anytime)

1. The SPECIALIST Lexicon

- A fancy synonym for “dictionary”
- A syntactic lexicon
- Biomedical and general English
- Over 0.5M 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)



THE INSOMNIAC'S DICTIONARY

Illeism: Reference to oneself by use of the third person
Infavoidance: The act of covering up one's inferiority complex
Inglenook: A place by the fire or any warm and comfortable area
Insilium: Legal term for evil advice or counsel
Jamais vu: Illusion that one has never previously experienced a situation, when in fact it is quite familiar (see Déja vu)
Jen: A compassionate love for all humanity or for the whole world
Karateka: A karate expert
Kloof: A deep ravine
Kludge: A system (especially of computers) made up of poorly matched components
Lallation: Pronouncing an "R" so that it sounds like an "L"
Lapidation: The act of stoning a person to death
Latrocination: A robbery that involves the use of force or violence
Lexicon: A fancy synonym for "dictionary"
Litotes: A form of understatement in which two negatives are used to make a positive ("he was not unhappy")
Longueur: A long and boring passage in a work of literature, drama, music, etc.
Macarism: The practice of making others happy by praising them
Matutinal: Pertaining to anything that takes place in the morning
Melorrhea: The writing of excessively long musical works
Meteorism: A tendency to uncontrollable passing of intestinal gas
Metrona: A young grandmother
Microperf: The very small perforations along the edges of computer paper
Migrateur: A wanderer
Mnemonic: That which assists memory (a classic mnemonic device is the one familiar to astronomy students: "Oh be a fine girl, kiss me"—a unique way to remember the stellar classifications O,B,A,F,G,K, and M)
Moria: Morbid impulse to make jokes
Omnistrain: The stresses of modern life
Omphaloskepsis: The act of contemplating one's navel
Onychophagy: The habit of biting one's fingernails
Oxymoron: A phrase or expression composed of contradictory elements ("awfully good," for example)

140

LexBuild Process (Computer-Aided)

Sources:
- Word candidates from MEDLINE
- Words from consumer data
- 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**

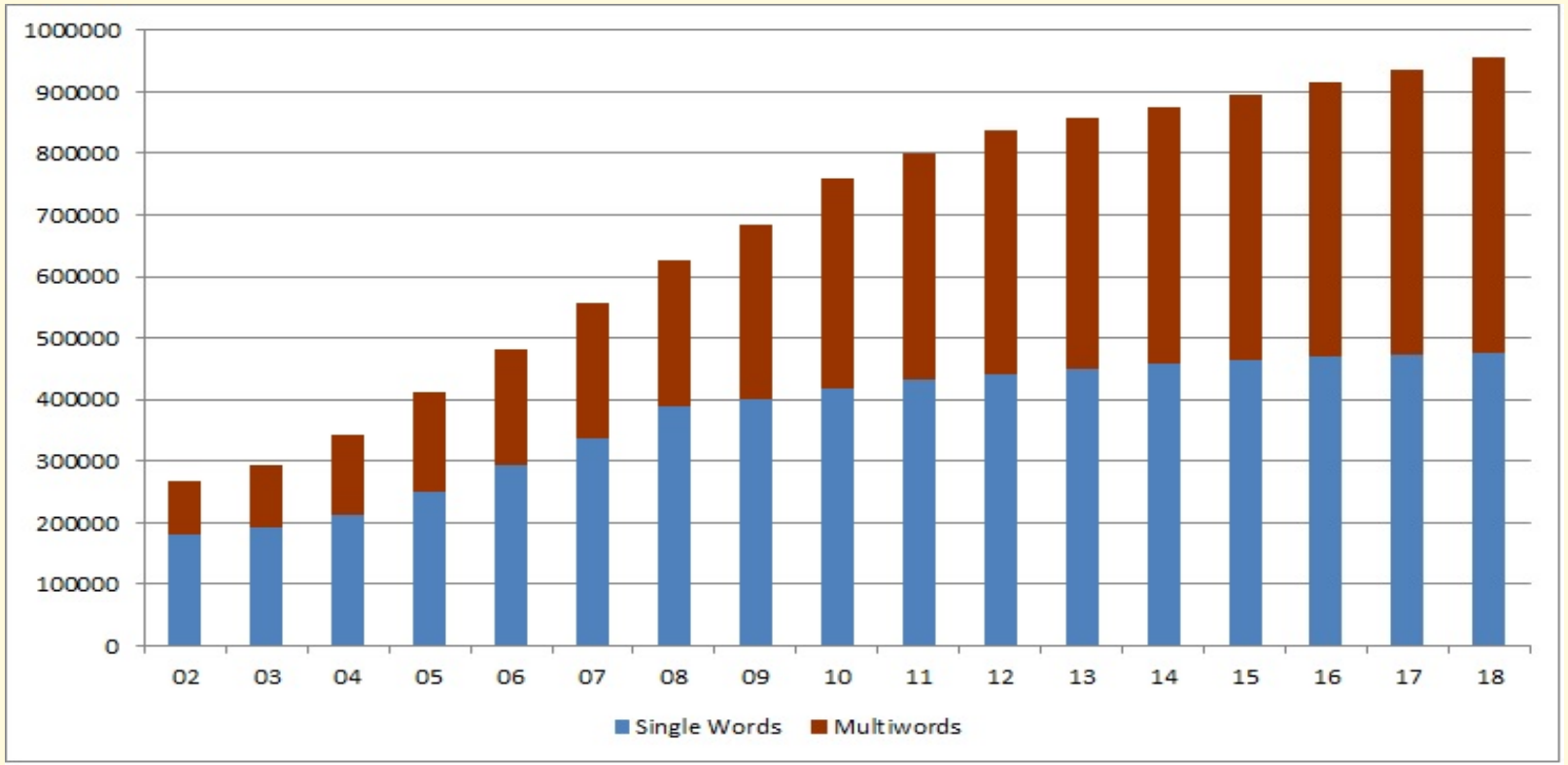
Team of Lexicon Builders

- Dr. Alexa McCray, founded in 1994 (previous LHC Director, 2005-)
- Allen Browne, father of the SPECAILIST Lexicon (retired 2017)

- Dr. Dina Demner Fushman (PI)
- Dr. Chris J. Lu
- Dr. Amanda Payne
- Destinee Tormey
- Francois Lang

Lexicon Growth – 2002 to 2018

- 505,145 lexical records
- 1,31,201 words (categories and inflections)
- 955,564 forms (spelling only)
  - Single words: 476,235 (49.84%); Multiwords: 479,329 (**50.16%**)



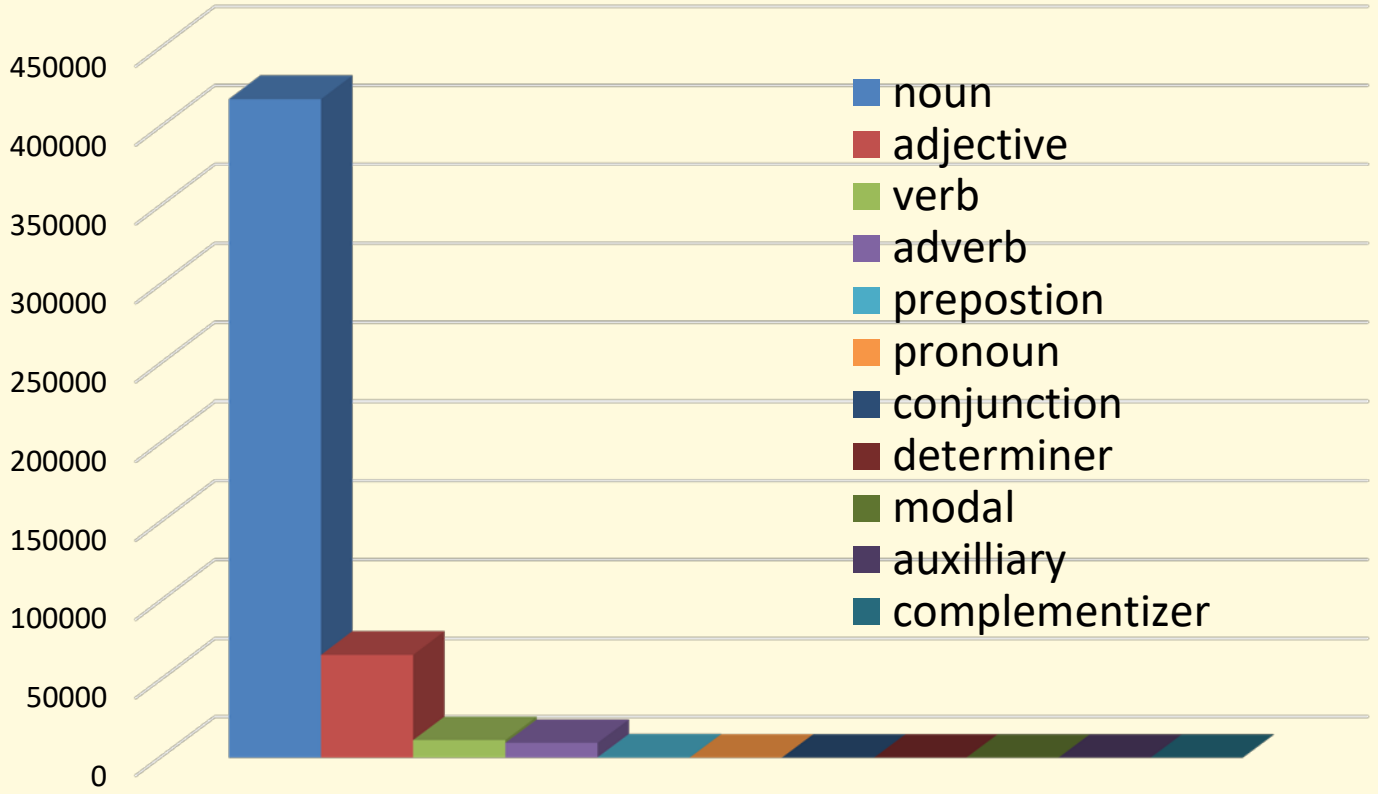
(Multi)Words for Lexical Records

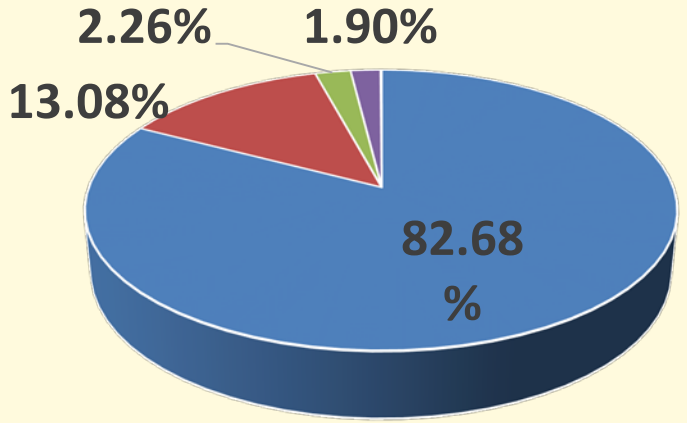
- Lexicon terms: single words and multiwords
  - Space(s): ice-cream vs. ice cream
- Four criteria for Lexicon terms:
  - Part of Speech (POS):
    - tear break up time, frog erythrocytic virus, cardiac surgery
  - Inflection morphology (uninflection):
    - left pulmonary veins (“left pulmonary vein” 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.2018

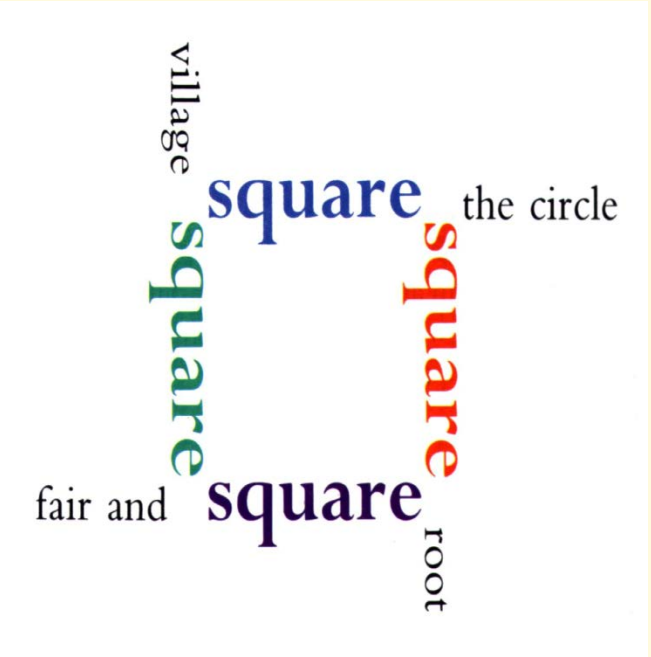
Lexical Records & POS

{base=square
entry=E0057517
cat=verb
variants=reg
intran
intran:part(up)
}

{base=square
entry=E0057516
cat=adj
variants=reg
variants=inv
}

{base=square
entry=E0057518
cat=adv
}

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



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)

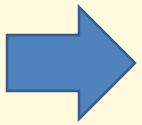
- color|colour
- grey|gray
- align|aline
- Grave's disease|Graves's disease|Graves' disease
- civilize|civilize
- harbor|harbor
- fetus|foetus|fœtus
- centre|center
- spelt|spelled
- ice cream|ice-cream
- xray|x-ray|x ray

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 Information to Coded Lexical Records

Lexical Information | Base	color
Part of speech	• noun
Inflectional morphology (inflections)	• color • colors
Orthography	• colour
Abbreviation/Acronym	• N/A
Syntax (complementation)	• N/A
…	• …
Derivational morphology (derivations)	• colorable • colorful • colorize • colorist • …
LexSynonyms	• chromatic



```
{base=color
spelling_variant=colour
entry=E0017902
        cat=noun
        variants=uncount
        variants=reg
}
```

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=divorcé
entry=E0543077
        cat=noun
        variants=reg
}
```

```
{base=role
spelling_variant=rôle
entry=E0053757
        cat=noun
        variants=reg
}
```

```
{base=cafe
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

Types	Word Count	Percentage %	Accu. %
LEXICON (S)	296,747	8.1978%	8.1978%
NUMBER	62	0.0017%	8.1995%
DIGIT	87,437	2.4155%	**10.6150%**
NW-EW*	43,811	1.2103%	**11.8253%**
NEW	3,191,797	88.1747%	100.0000%
Total	3,619,854		

* NW_EW: an element word only exist in multiword, such as “non”, “vitro”, “vivo”, “intra”, etc.

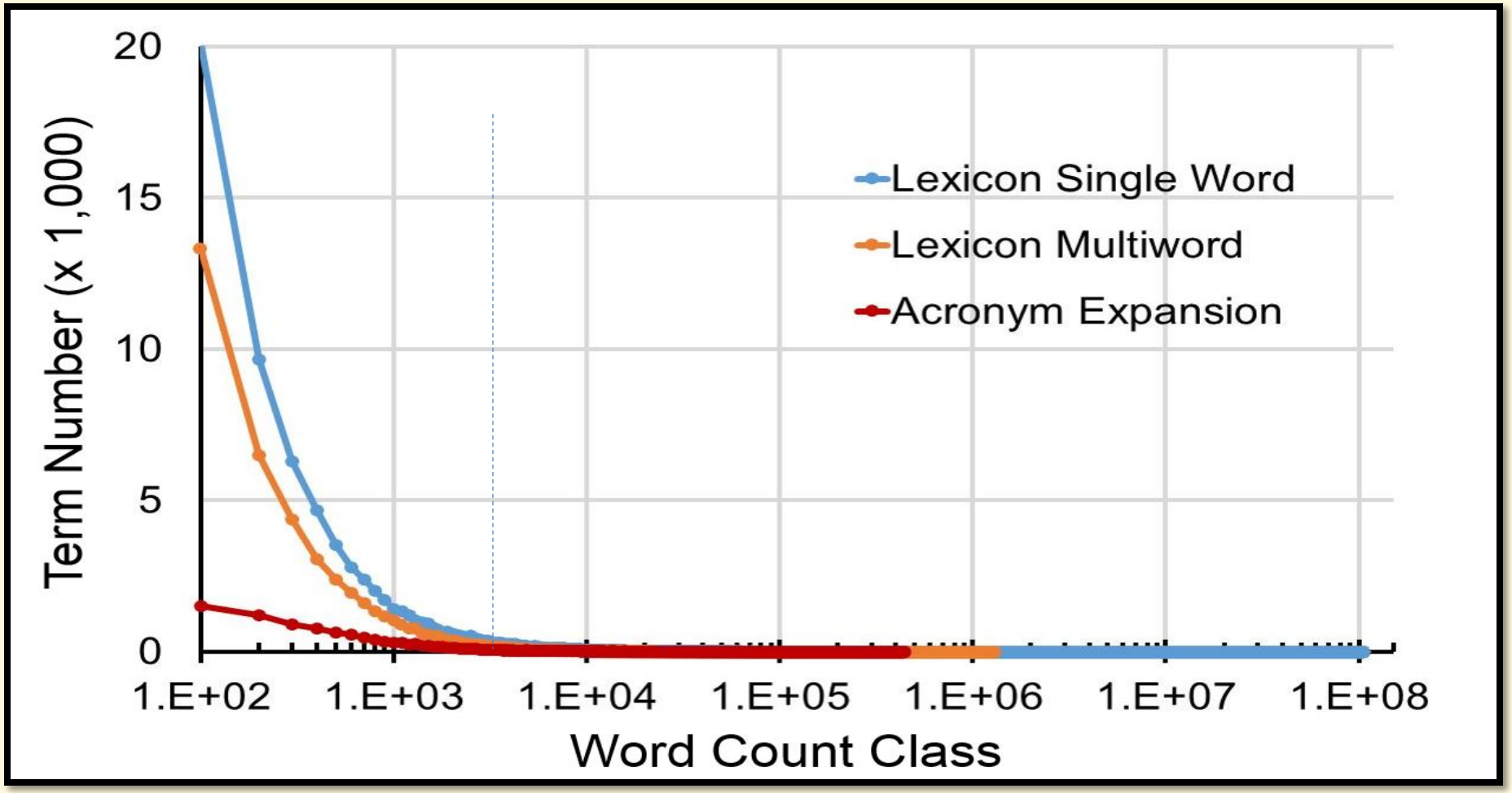
Lexicon Unigram Coverage – With Frequency (WC)

- Total word count for MEDLINE (2016): 3,114,617,940
- Lexicon covers > 98% unigrams from MEDLINE

Types	Word Count	Percentage %	Accu. %
LEXICON	2,911,156,308	93.4675%	93.4675%
NUMBER	8,753,120	0.2810%	93.7485%
DIGIT	145,548,882	4.6731%	**98.4216%**
NW-EW*	19,148,557	0.6148%	**99.0364%**
NEW	30,011,073	0.9636%	100.0000%
Total	3,114,617,940		

* NW-EW: an element word only exist in multiword, such as “non”, “vitro”, “vivo”, “intra”, etc.

The Frequency Spectrum of Lexicon (Multi)words on MEDLINE



The Frequency Spectrum of Alice in Wonderland

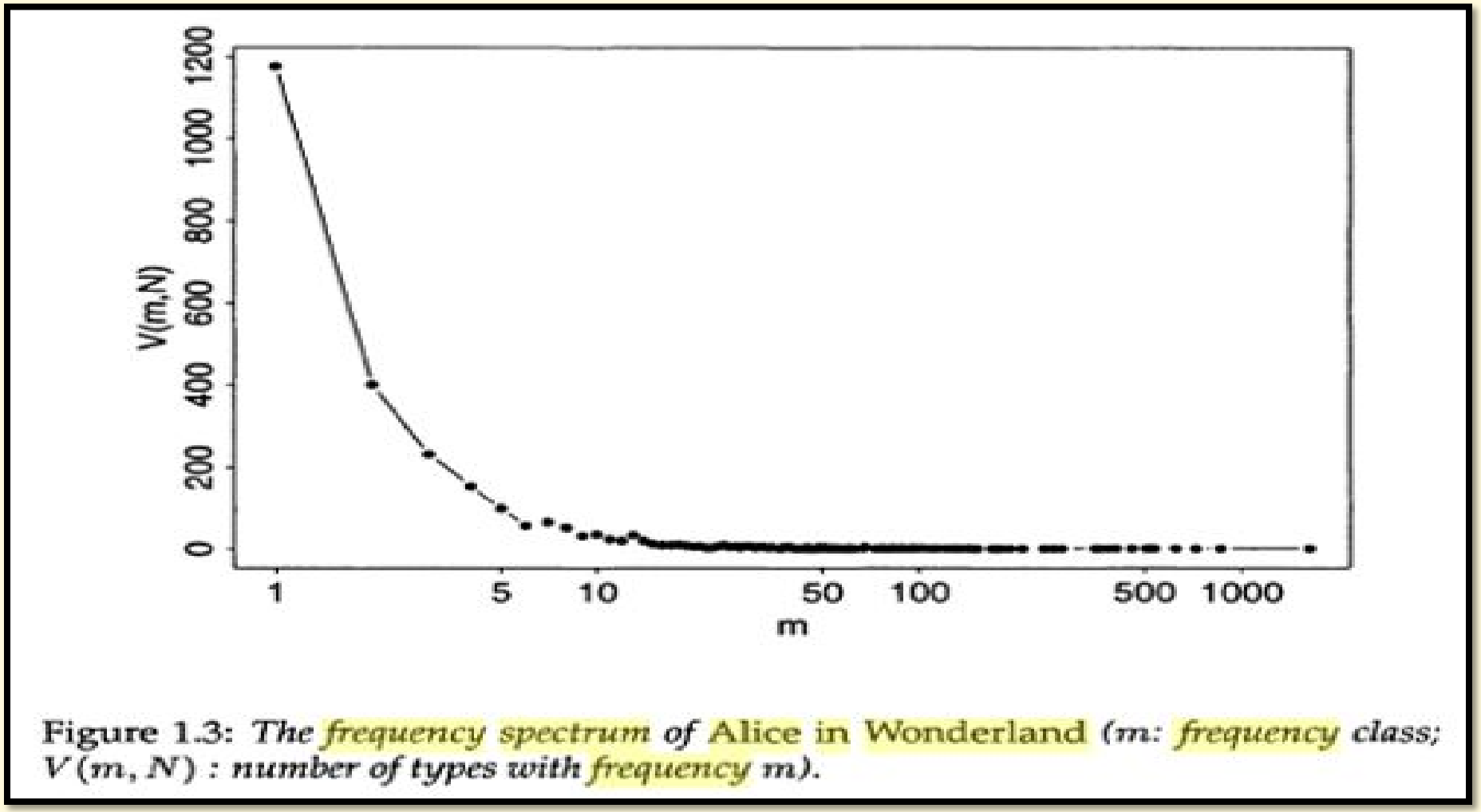


Figure 1.3: *The frequency spectrum of* Alice in Wonderland *(m: frequency class; V (m, N) : number of types with frequency m).*

Lexicon (Data) and Lexical Tools (Software)

LR Tables


```
{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. NLP - 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
- nominalization: generalisation, generalization
- Abbreviation/acronym: n/a

← A LexRecord

Inflectional variants:
- generalises, generalised, generalising

Derivational variants:
- suffixD: generalisation, generalization, generalisable
- prefixD: overgeneralise, over-generalise

← A LexRecord + Rules

Synonyms: generalize

← Multiple LexRecords + Rules

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

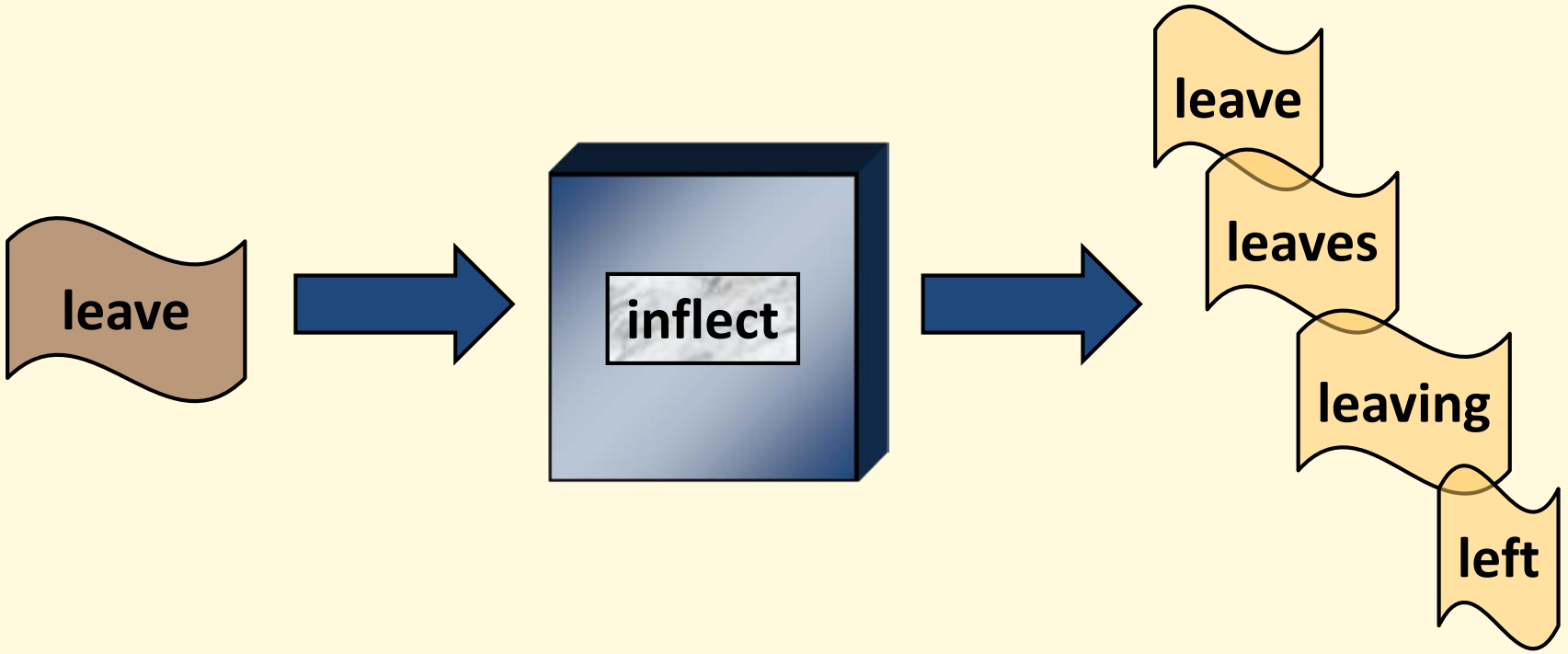
LVG - Lexical Variants Generation

- 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)

Lexical Tools – Flow Components (62)

Lexicon Related – Data (32)	Non-Lexicon related – Algorithm (30)
Inflection (10): b, B, Bn, I, ici, is, L, Ln, Lp, si,	Unicode operation (10): q, q0, q1, q2, q3, q4, q5, q6, q7, q8
Derivation (3): d, dc, R	Tokenizer (3): c, ca, ch
Acronym or abbreviation (3): a, A, fa	Punctuation operation (3): o, p, P
Spelling variant (2): e, s	Lowercase (1): l
Lexicon mapping (3): An, E, f, fp	Metaphone (1): m
Synonym (2): y, r	Remove parenthetic plural forms (1): rs
Nominalization (1): nom	Strip stop word (1): t
Citation (1): Ct	Remove genitive (1): g
Fruitful variant (4): G, Ge, Gn, V	No operation (1): n
Normalization (2): N, N3,	…

LVG Flow Component – Example

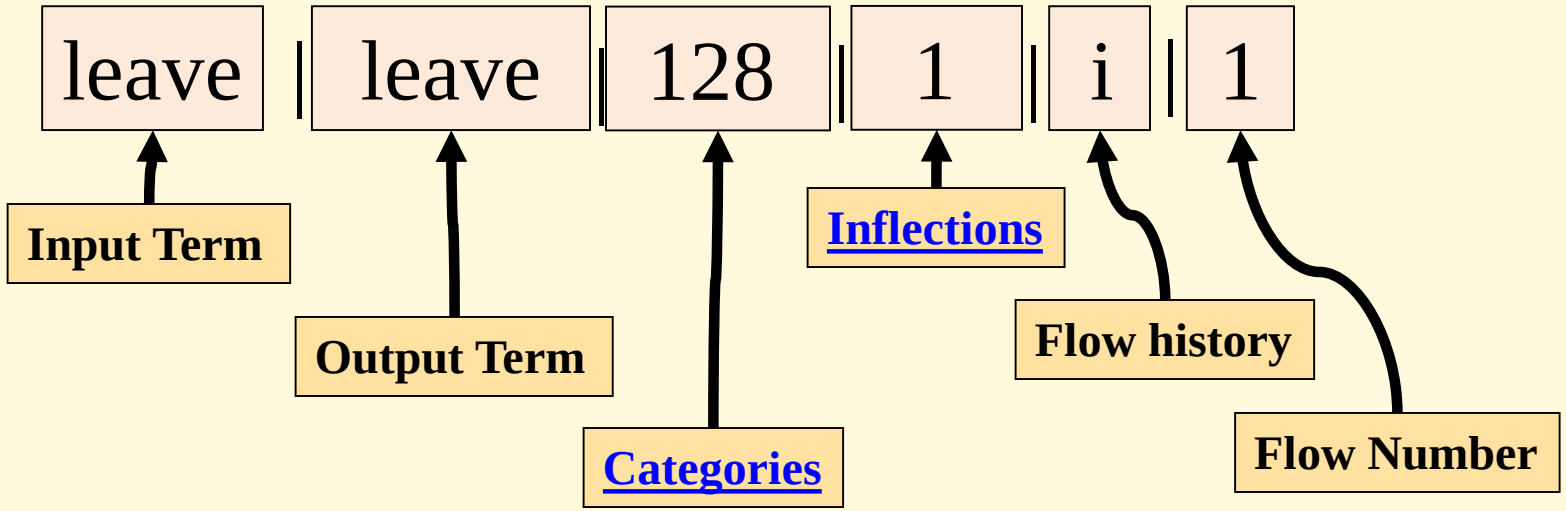


LVG Flow Component – CmdLine

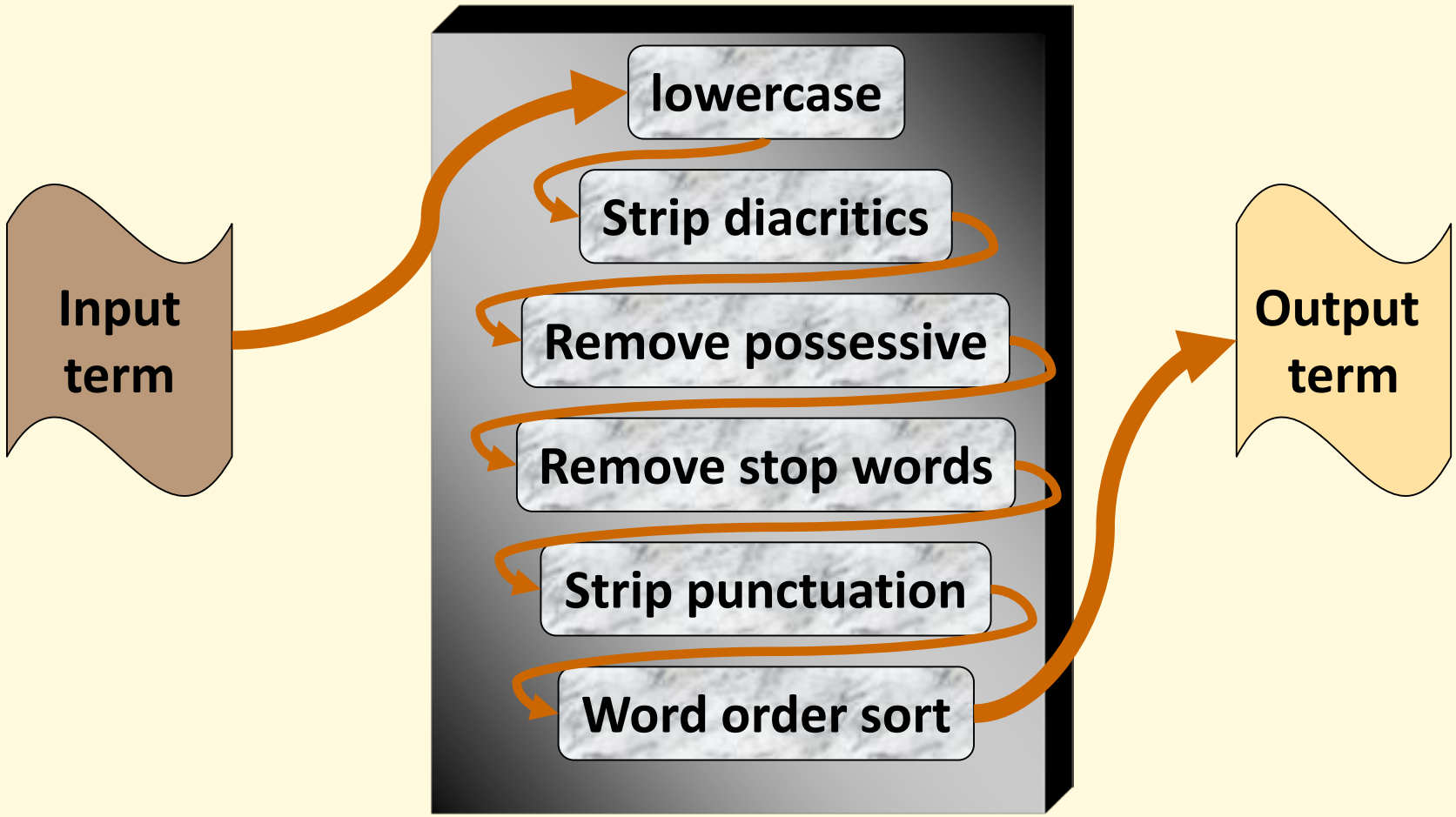
```
> lvg -f:i
leave
leave|leave|128|1|i|1|
leave|leave|128|512|i|1|
leave|leaves|128|8|i|1|
leave|left|1024|64|i|1|
leave|left|1024|32|i|1|
leave|leave|1024|1|i|1|
leave|leave|1024|262144|i|1|
leave|leave|1024|1024|i|1|
leave|leaves|1024|128|i|1|
leave|leaving|1024|16|i|1|
```

LVG Flow Component – Fielded Output

> lvg –f:i
leave



LVG – A Serial Flow



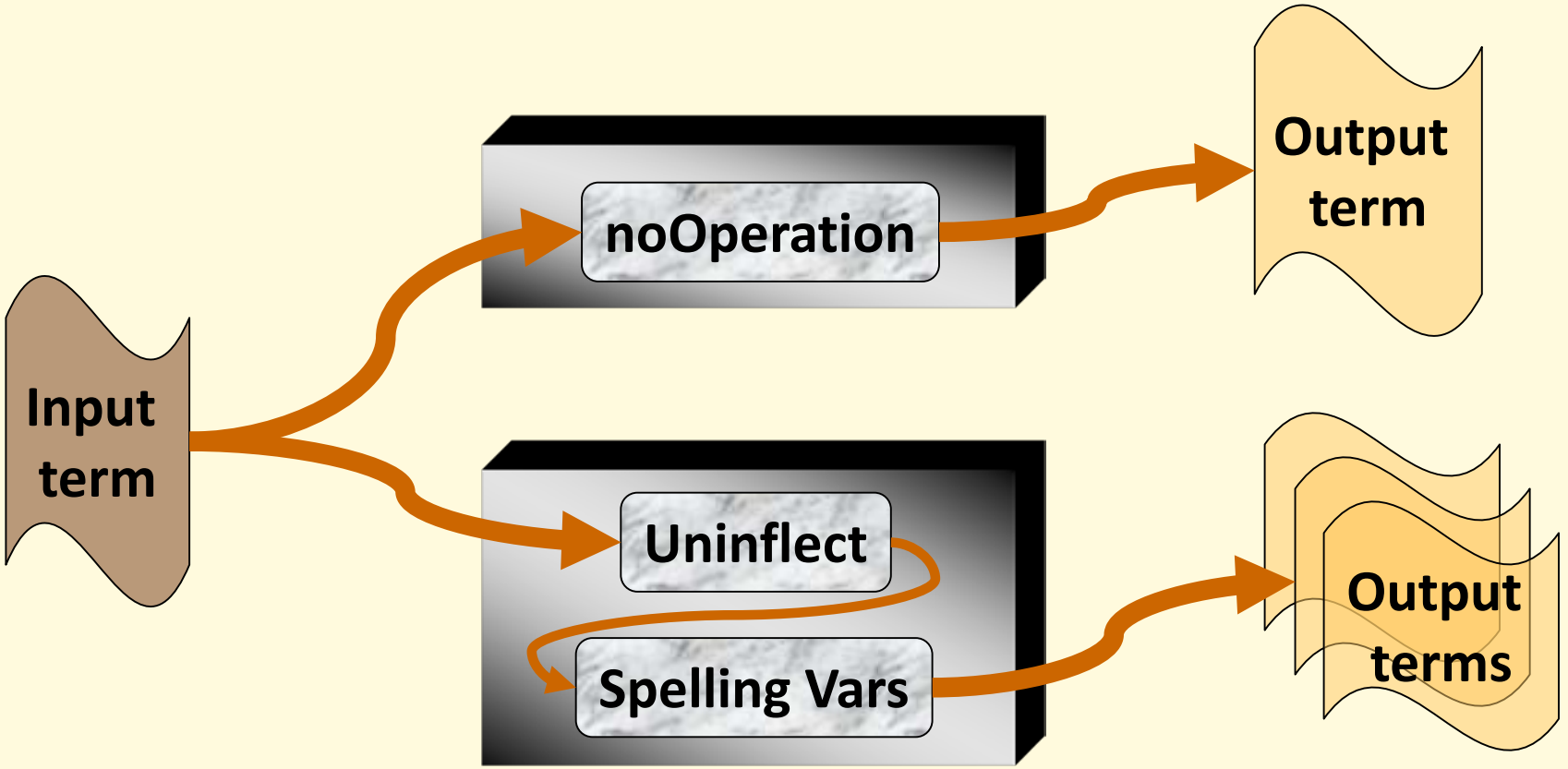
- Flow components can be arranged so that the output of one is the input to another.

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

Parallel Flows - Example

```
> 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 (commonly used flow)

- Composed of 11 Lvg flow components to abstract away from (only keep meaningful words):
  - case
  - punctuation
  - possessive forms
  - inflections
  - spelling variants
  - stop words
  - diacritics & ligatures (non-ASCII Unicode)
  - word order

Ex - Norm

“Fœtoproteins α’s, NOS“

q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order

Norm

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

"Fœtoproteins α's, NOS"

Norm

q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order

“Fœtoproteins α’s, NOS“
"Fœtoproteins α**’s**, NOS"
"Fœtoproteins α, NOS"

Norm

q0: map symbols to ASCII
g: remove genitives
rs: remove parenthetic plural forms
o: replace punctuation with spaces
t: strip stop words
l: lowercase
B: uninflect each words in a term
Ct: retrieve citations
q7: Unicode core Norm
q8: strip or map Unicode to ASCII
w: sort words by order

“Fœtoproteins α’s, NOS“
"Fœtoproteins α’s, NOS"
"Fœtoproteins α, NOS"
"Fœtoproteins α, NOS"

Norm

q0: map symbols to ASCII	“Fœtoproteins α’s, NOS“
g: remove genitives	"Fœtoproteins α’s, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	"Fœtoproteins α, NOS"
t: strip stop words	Fœtoproteins α NOS
l: lowercase	
B: uninflect each words in a term	
Ct: retrieve citations	
q7: Unicode core Norm	
q8: strip or map Unicode to ASCII	
w: sort words by order	

Norm

q0: map symbols to ASCII	“Fœtoproteins α’s, NOS“
g: remove genitives	"Fœtoproteins α’s, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	"Fœtoproteins α, NOS"
t: strip stop words	Fœtoproteins α NOS
l: lowercase	Fœtoproteins α
B: uninflect each words in a term	
Ct: retrieve citations	
q7: Unicode core Norm	
q8: strip or map Unicode to ASCII	
w: sort words by order	

Norm

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

Norm

q0: map symbols to ASCII	“Fœtoproteins α’s, NOS“
g: remove genitives	"Fœtoproteins α’s, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	"Fœtoproteins α, NOS"
t: strip stop words	Fœtoproteins α NOS
l: lowercase	Fœtoproteins α
B: uninflect each words in a term	fœtoproteins α
Ct: retrieve citations	fœtoprotein α
q7: Unicode core Norm	
q8: strip or map Unicode to ASCII	
w: sort words by order	

Norm

q0: map symbols to ASCII	“Fœtoproteins α’s, NOS“
g: remove genitives	"Fœtoproteins α’s, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	"Fœtoproteins α, NOS"
t: strip stop words	Fœtoproteins α NOS
l: lowercase	Fœtoproteins α
B: uninflect each words in a term	fœtoproteins α
Ct: retrieve citations	fœtoprotein α
q7: Unicode core Norm	fetoprotein α
q8: strip or map Unicode to ASCII	
w: sort words by order	

Norm

	“Fœtoproteins α’s, NOS“
q0: map symbols to ASCII	"Fœtoproteins α’s, NOS"
g: remove genitives	"Fœtoproteins α, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	Fœtoproteins α NOS
t: strip stop words	Fœtoproteins α
l: lowercase	fœtoproteins α
B: uninflect each words in a term	fœtoprotein α
Ct: retrieve citations	fetoprotein α
q7: Unicode core Norm	fetoprotein α
q8: strip or map Unicode to ASCII	
w: sort words by order	

Norm

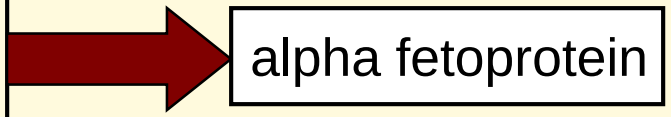
	“Fœtoproteins α’s, NOS“
q0: map symbols to ASCII	"Fœtoproteins α’s, NOS"
g: remove genitives	"Fœtoproteins α, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	Fœtoproteins α NOS
t: strip stop words	Fœtoproteins α
l: lowercase	fœtoproteins α
B: uninflect each words in a term	fœtoprotein α
Ct: retrieve citations	fetoprotein α
q7: Unicode core Norm	fetoprotein α
q8: strip or map Unicode to ASCII	fetoprotein alpha
w: sort words by order	

Norm

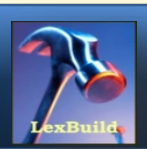
	“Fœtoproteins α’s, NOS“
q0: map symbols to ASCII	"Fœtoproteins α’s, NOS"
g: remove genitives	"Fœtoproteins α, NOS"
rs: remove parenthetic plural forms	"Fœtoproteins α, NOS"
o: replace punctuation with spaces	Fœtoproteins α NOS
t: strip stop words	Fœtoproteins α
l: lowercase	fœtoproteins α
B: uninflect each words in a term	fœtoprotein α
Ct: retrieve citations	fetoprotein α
q7: Unicode core Norm	fetoprotein α
q8: strip or map Unicode to ASCII	fetoprotein alpha
w: sort words by order	alpha fetoprotein

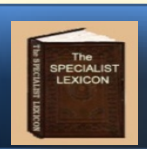
Norm

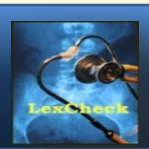
alpha Fetoprotein
alpha Fetoproteins
alpha-Fetoprotein
alpha-Fetoproteins
Alpha fetoproteins
alpha fetoprotein
alpha Foetoprotein
alpha foetoprotein
alpha fetoproteins
Alpha-fetoprotein
alpha-fetoprotein
Alpha Fetoproteins
Alpha-Fetoprotein
Alpha-fetoprotein NOS
Alpha Fetoprotein
alpha-fetoprotein
ALPHA-FETOPROTEIN
Alpha Fœtoprotein
…

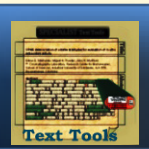


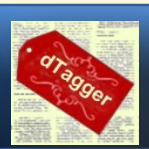
alpha fetoprotein

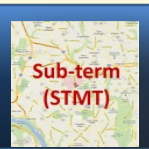










3. Natural Language Processing (NLP)

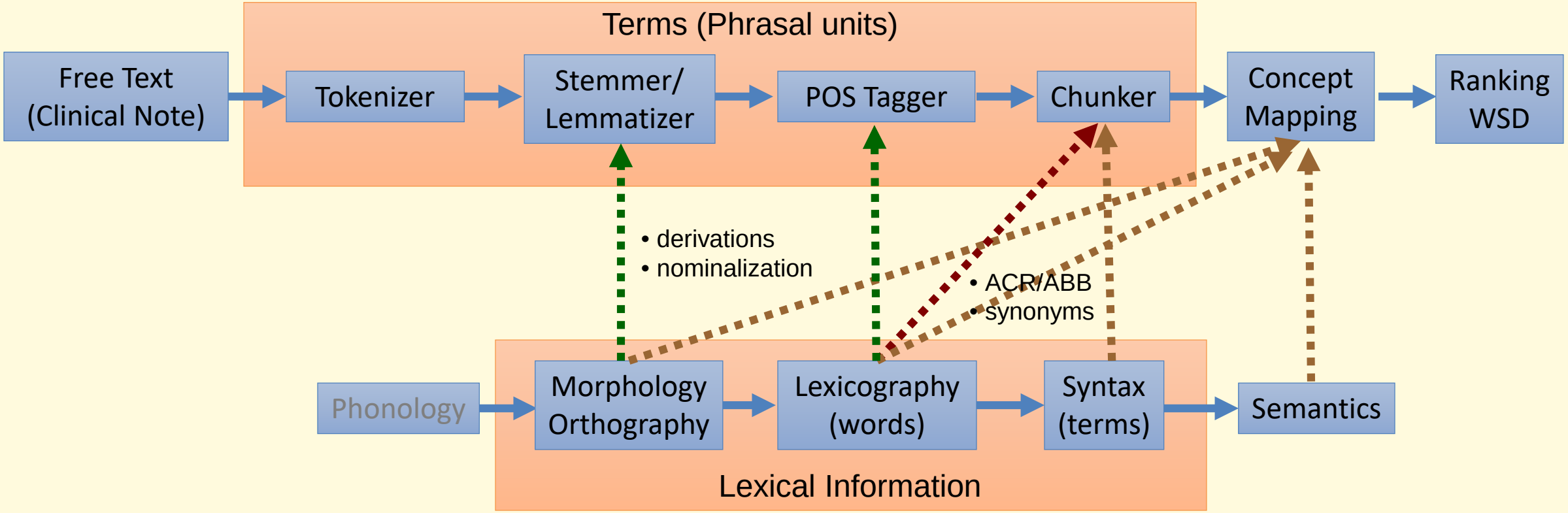
- Natural Language
  - is ordinary language that humans use naturally
  - may be spoken, signed, or written
- Natural Language Processing
  - NLP is to process human language to make their information accessible to computer applications
  - The goal is to design and build software that will analyze, understand, and generate human language
  - NLP includes a board 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.

Concept Mapping Challenges

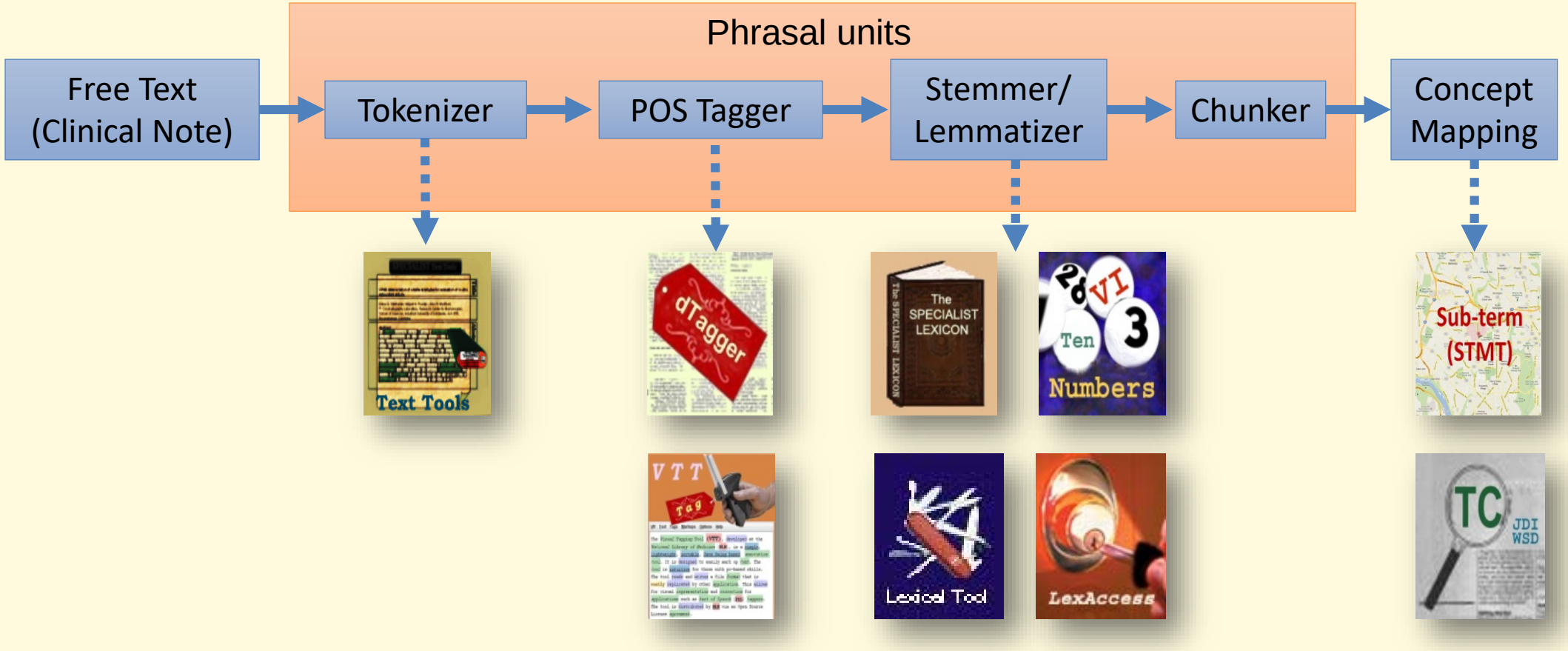
- Challenge 1: Map terms to concepts (meaning)
- Challenge 2: many to many mapping

Terms	Concepts	NLP
• cold • Cold Temperature • Cold Temperatures • Cold (Temperature) • Temperatures, Cold • Low temperature • low temperatures • …	• Cold Temperature|C0009264	• Concept mapping
• cold	• Cold Temperature|C0009264 • Common Cold|C0009443 • Cold Therapy|C0010412 • Cold Sensation|C0234192 • …	• WSD (Word Sense Disambiguation)

NLP Pipe Line – Lexical Information



The SPECIALIST NLP Tools



- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov

NLP Applications

- Syntax:
  - parsers, taggers, POS tagging, etc.
- Semantics:
  - name entity recognition, concept mapping, etc.

- Knowledge extraction:
  - learn relations between entities, recognize events, etc.
- Summarization:
  - sentiment analysis and figure out the topics of a page
- Question answering
  - find answers for queries

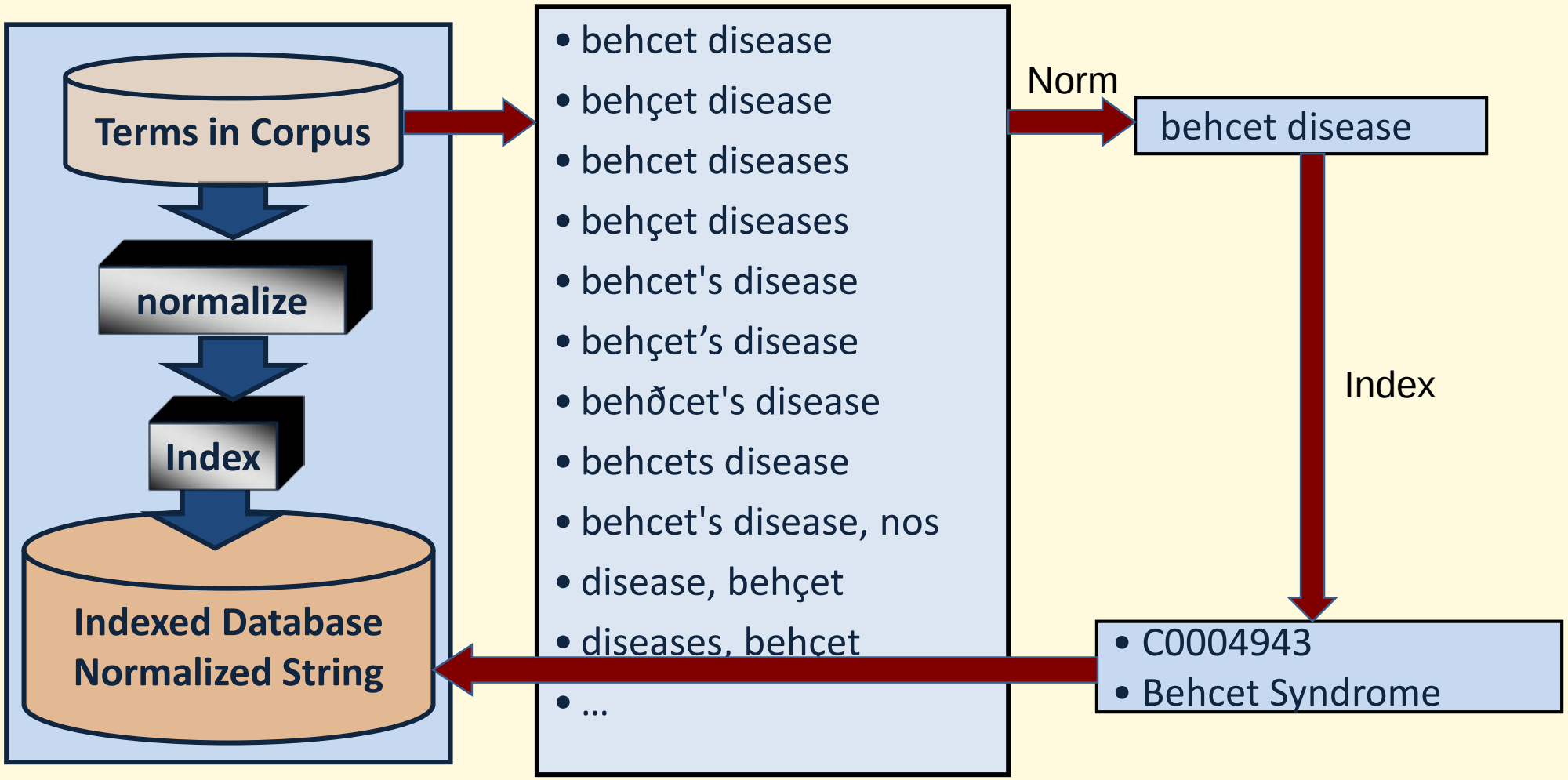
NLP – Concept Mapping

- 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…

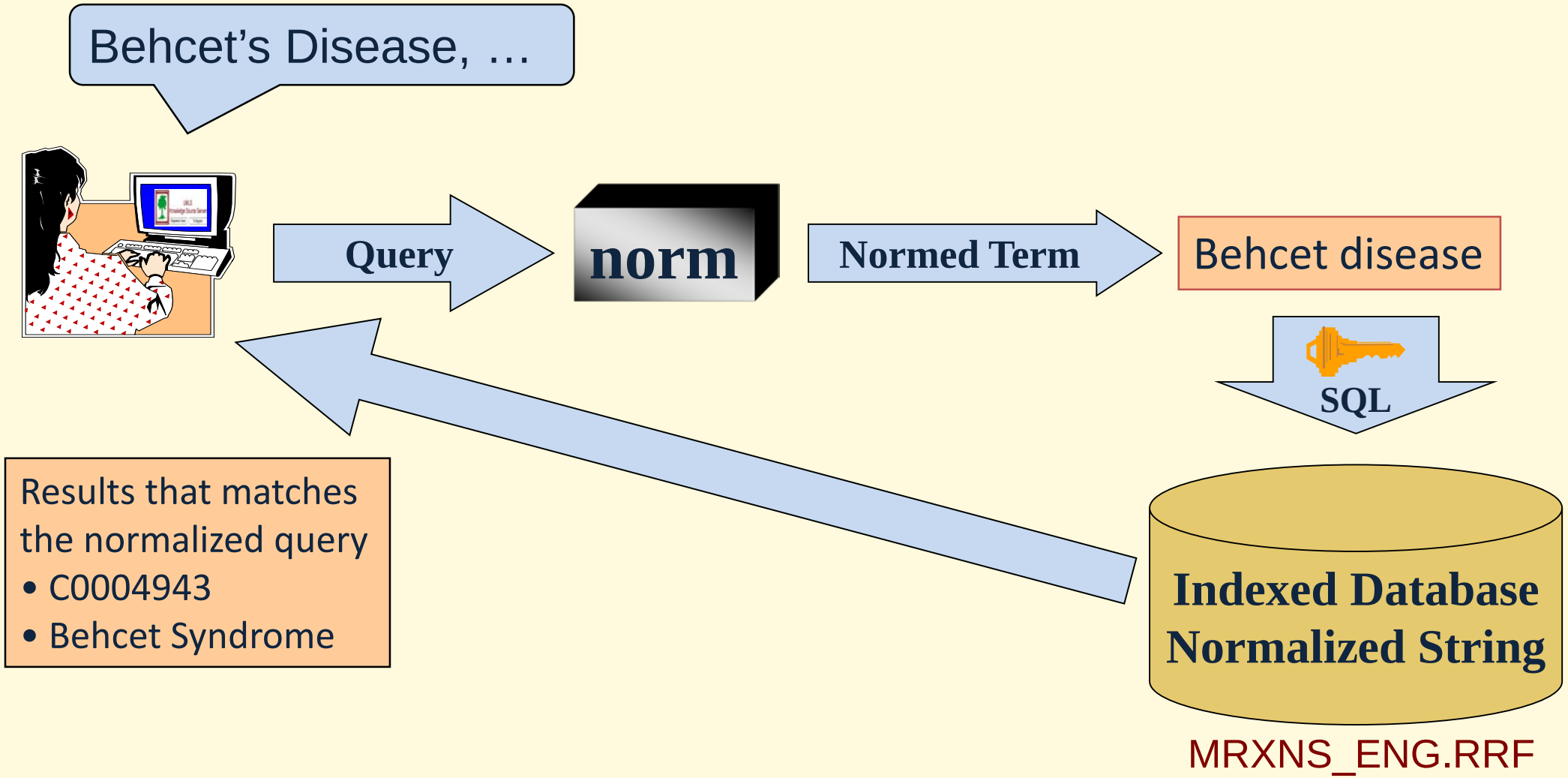
Lexical Tools – Norm

q0: map Unicode symbols to ASCII	Behçet's Diseases, NOS
g: remove genitives	Behçet's Diseases, NOS
rs: remove parenthetic plural forms	Behçet Diseases, NOS
o: replace punctuation with spaces	Behçet Diseases, NOS
t: strip stop words	Behçet Diseases NOS
l: lowercase	Behçet Diseases
B: uninflect each words in a term	behçet diseases
Ct: retrieve citations	behçet disease
q7: Unicode core Norm	behcet disease
q8: strip or map non-ASCII char	behcet disease
w: sort words by order	behcet disease

NLP – Norm (Pre-Process Lexical Variations)

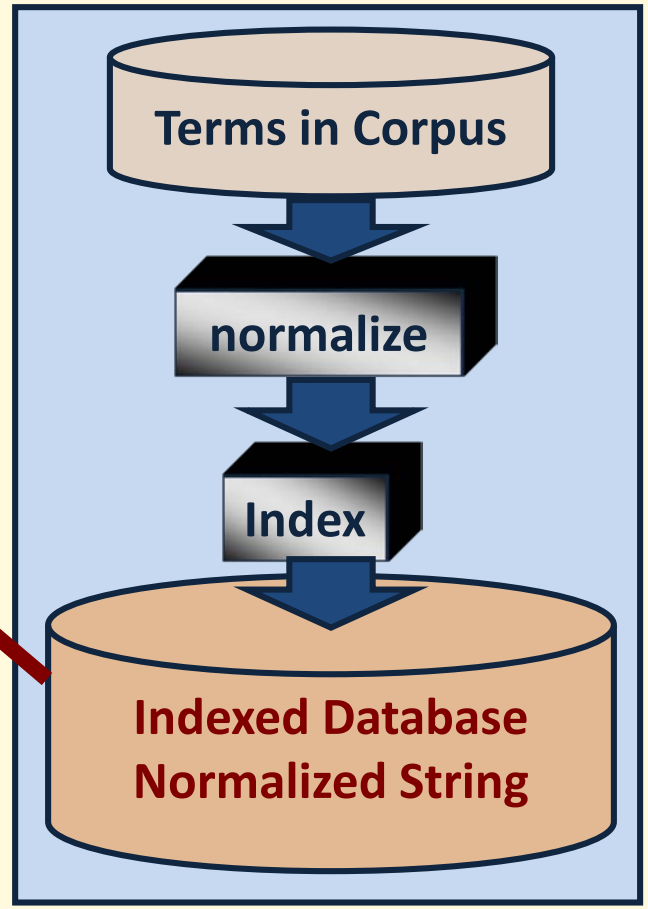


NLP – Norm (Application)

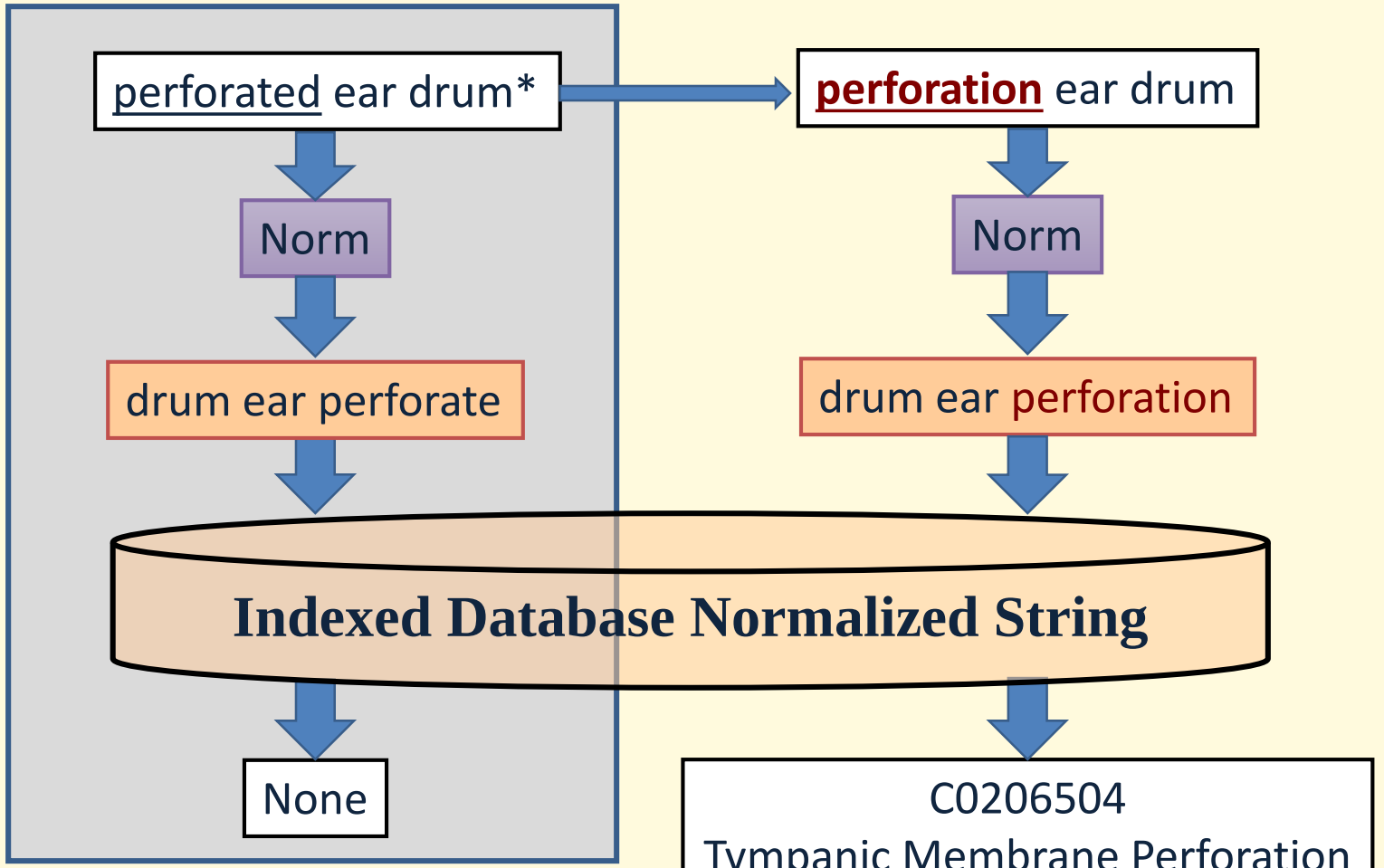


UMLS Metathesaurus

- UMLS Normalized Files
  - Normalized words: MRXNW_ENG.RRF
  - Normalized strings: MRXNS_ENG.RRF

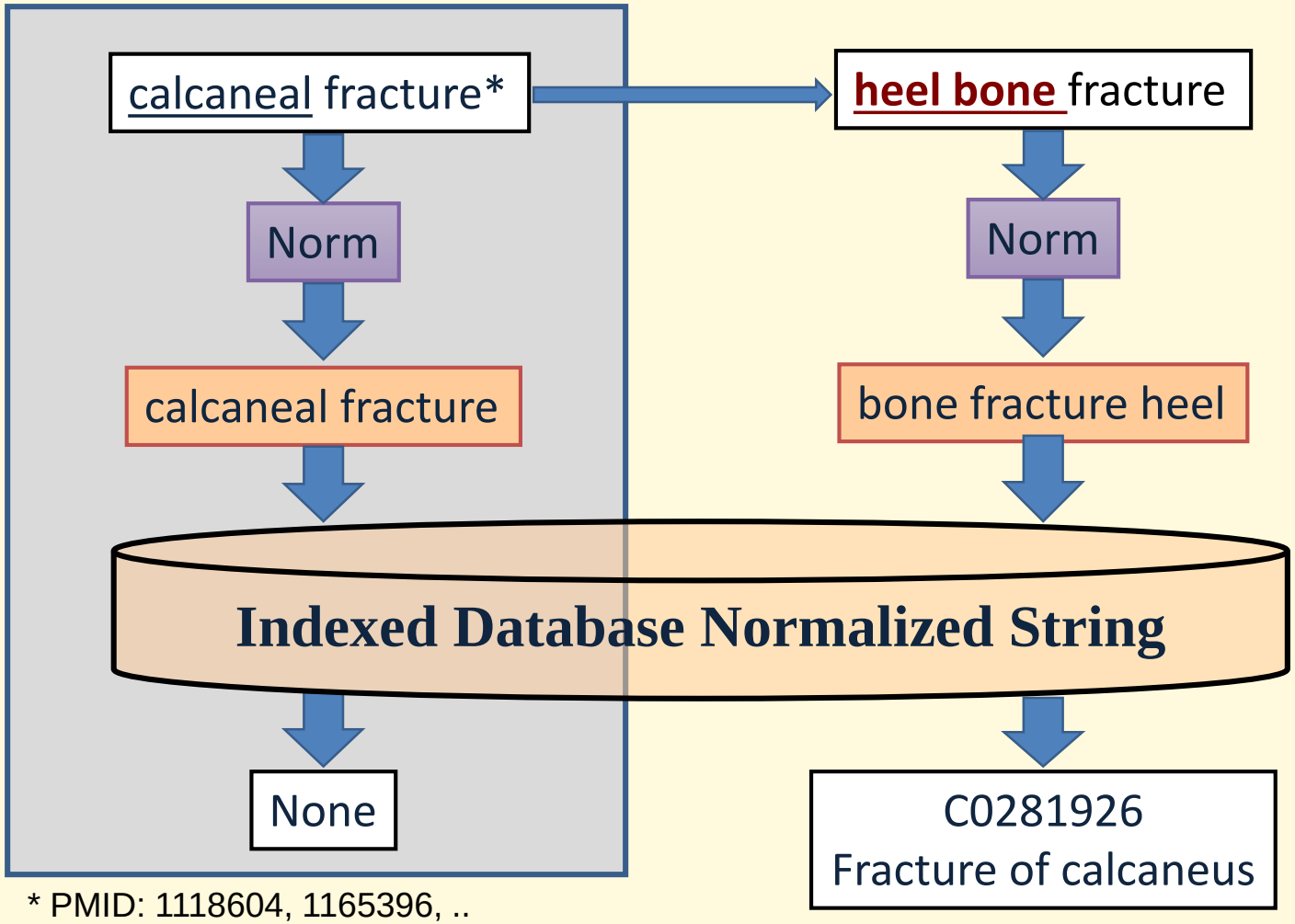


NLP – Query Expansion (derivation)



* PMID: 13114832, 5992689, ..

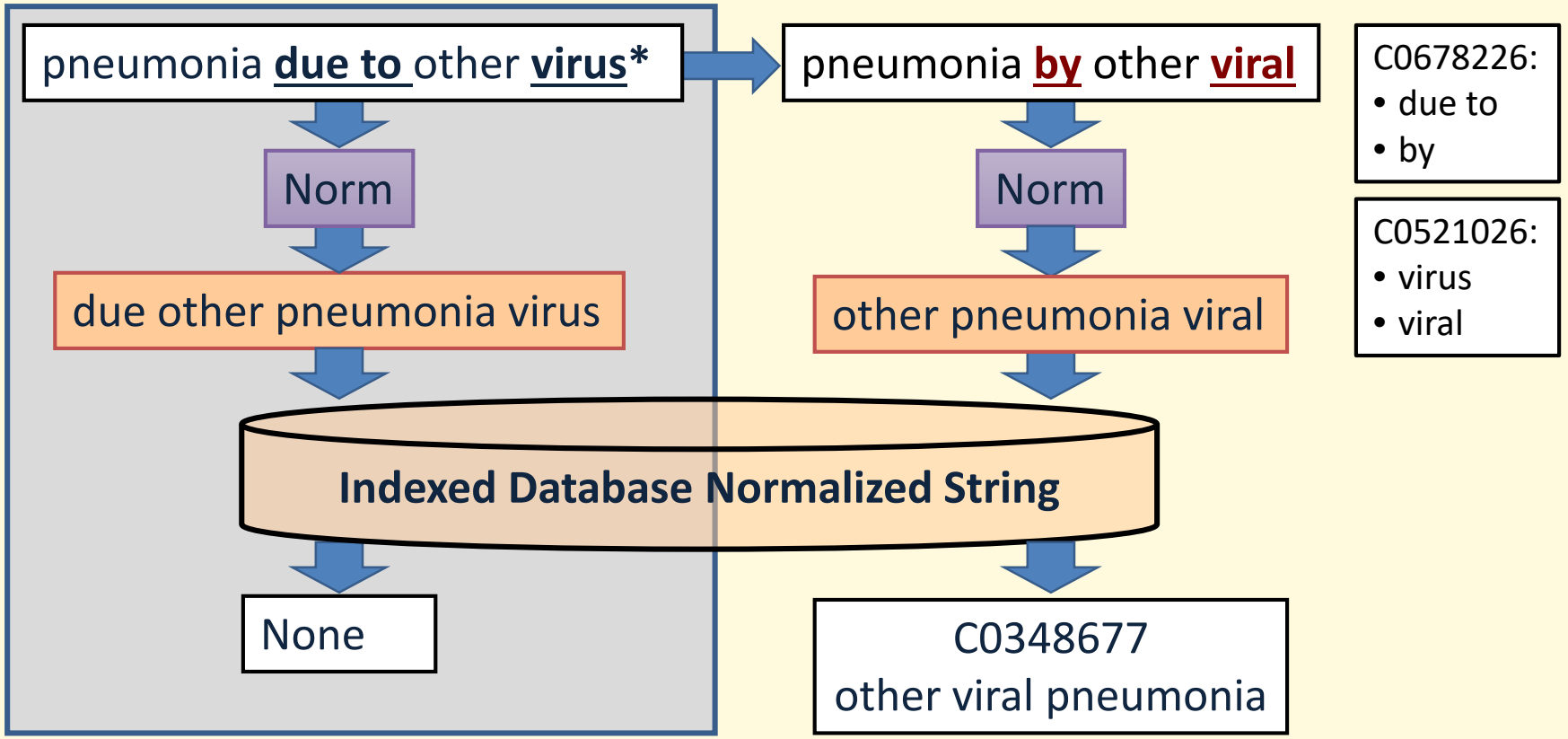
NLP – Query Expansion (Synonym)



C0006655:
- calcaneal
- heel bone

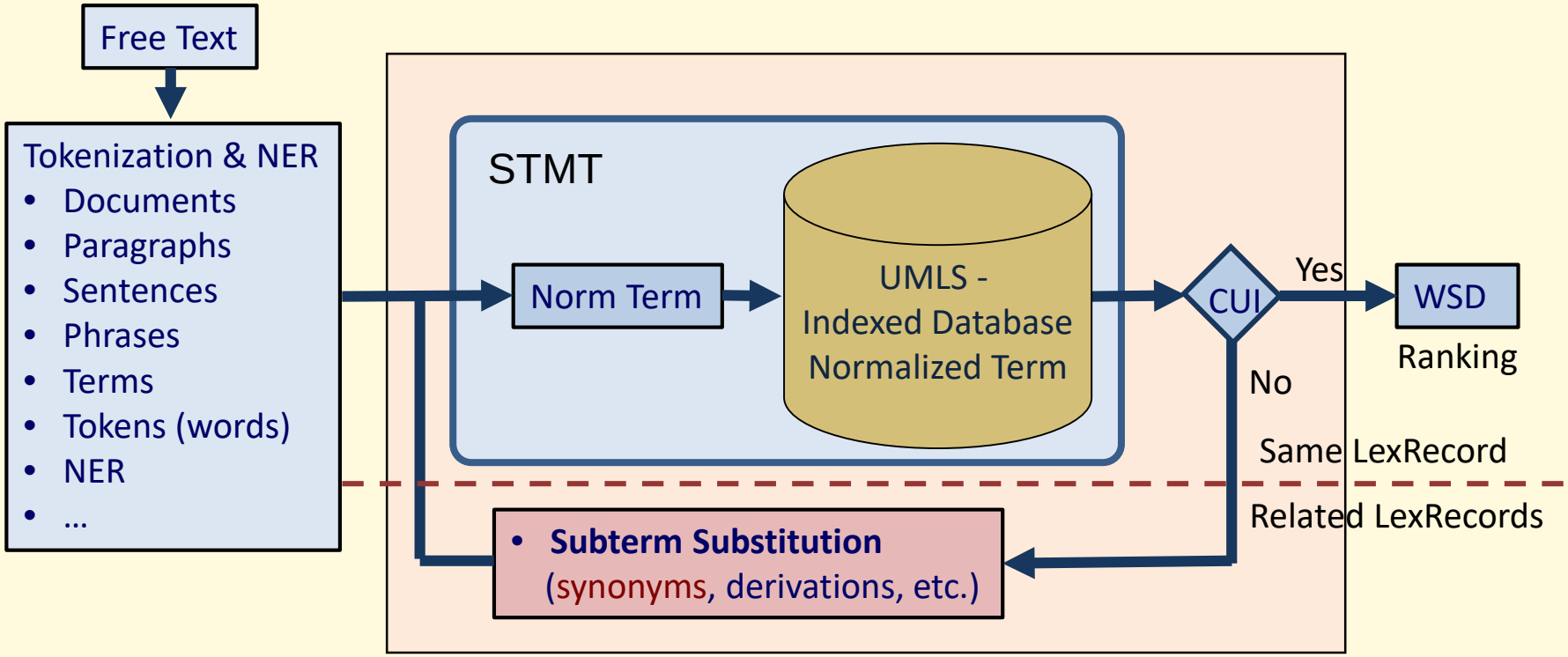
* PMID: 1118604, 1165396, ..

Multiple Substitutions



* VA14760, HA480.80, ..

Real-time Model



4. Applications - CSpell

- CSpell - Consumer Spelling Tool
  - Health information consumers
    - patients, families, caregivers, and the general public
    - seek health information and ask questions online every day
  - Consumer Health Information Question Answering (CHIQA)
    - launched in 2012 by NLM
    - provides reliable health information
  - Consumer Questions:
    - contain many spelling errors, informal expression, medical terminology
    - very few publicly available tools (insufficient features to handle errors)

      => Best: Kilicoglu's Ensemble method outperformed 30+%

Spelling Errors

- Spelling Errors in Consumer Questions:
  - Dictionary bases
    - No-word and Real-word errors
    - 1To1, Split and Merge corrections
  - Non-dictionary based
    - Informal Expression
    - Missing space on adjacent punctuation or digits
    - Others: Xml/Html, Unicode, etc.
  - Combinations of above errors

Error Examples

Ex-1. My mom was dianosed early on set deminita 3 years ago.

diagnosed *onset* *dementia*

NW-1To1 | RW-Merge | NW-1To1

Ex-2. brokenribscantsleepatnight

broken ribs cant sleep at night

NW-Split

Ex-3. A lot of the pain is joint stiff n ess.

stiffness

NW-Merge

Ex-4. Irregular bowl movement

bowel

RW-1To1

Ex-5. Sounding in my ear every time for along time.

a long

RW-Split

Ex-6. Who need to do test?pls guide me thank u.

test? please *you*

ND-Split | ND-Informal Expression

Ex-7. I have a shuntfrom2007.

shunt from 2007

ND-Split, NW-Split

Ex-8. I am permanently depressed and was on 2 or 3 different anti depresants.

antidepressants

NW-1To1, RW-Merge

Objectives

- To develop a spelling tool to detect and correct all types of spelling errors.

Requirements

- An open-source distributable tool package
- A spelling tool for general purpose
  - configurable dictionaries, frequency file, context file, etc.
  - annual release with the latest data (Lexicon, consumer corpus, etc.)
- Provide Java APIs
- Provide other configurable options:
  - functions: non-word and real-word correction
  - merge size, split size, candidate size
  - context window radius
  - …

Non-Dictionary Based Correction

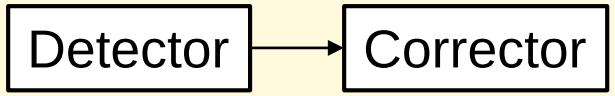
- Xml/Html Handler
- Splitter
  - Ending Punctuation Splitter
  - Leading Punctuation Splitter
  - Leading Digit Splitter
  - Ending Digit Splitter
- Informal Exception Handler

ND - Handlers

- Algorithm: table lookup
- Examples:

Handler Type	Input Text	Output Correction
Xml/Html	"germs"	”germs”
Informal Exception	pls	please

ND - Splitters

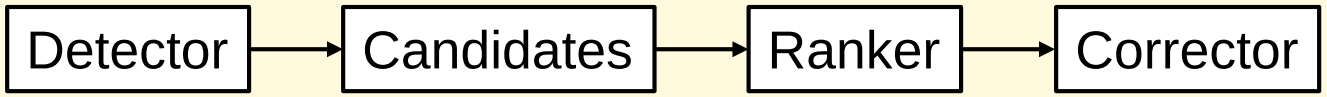
- Algorithm: 

  - Detector - applies matchers and filters (patterns):
    - regular expression + algorithm
    - uses LexRecords and development set to find patterns
  - Corrector:
    - split the token + flat map
- Examples:

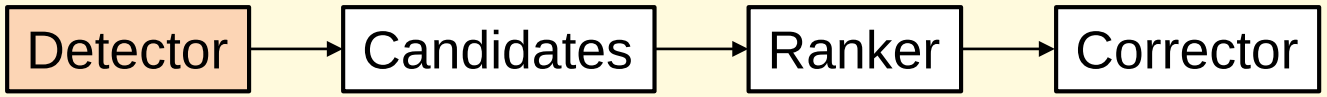
Splitter	Original Text	Corrected Text	Exceptions
Leading Digit	1.5years	1.5 years	42nd, 3Y1, etc.
Ending Digit	from2007.	From 2007.	Alpha1, Co-Q10, etc.
Leading Punctuation	volunteers(healthy)	volunteers (healthy)	R&D, finger(s), etc.
Ending Punctuation	cancer?if so	cancer? If so	Dr.s, 1,200, Beat(2), etc.

Dictionary Based Correction

- Detector: to detect errors (focus token)
- Candidate Generator: to generate correcting candidates
- Ranker: to rank candidates and find the best candidate
- Corrector: to replace errors with the best candidate



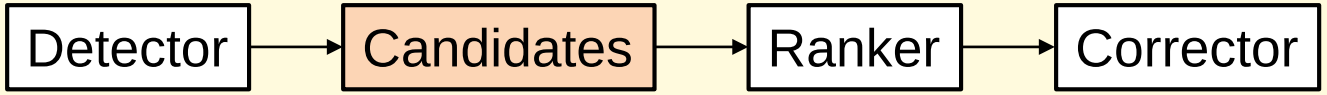
NW, 1To1 - Detector



- Not in the dictionary (non-word)
  - Dictionary includes element words from the Lexicon, numbers, selected unigrams of medical terms from UMLS.
- Not exceptions (non-word, but no need to be corrected)

Exception Type	Example
Digit	12.5
Punctuation	*
Digit/Punctuation	12-25-2016
Email	help@yahoo.com
URL	www.yahoo.com
Measurement (Unit)	30mg/50kg
Single Char	B

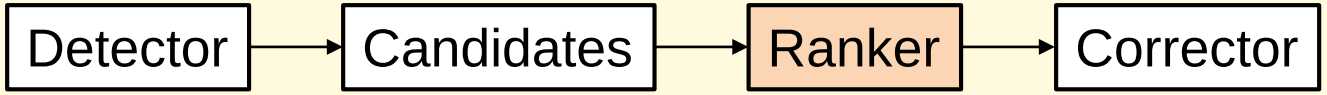
NW, 1To1 - Candidates



- Candidate set, Edit Distance <= 2
- In the suggestion dictionary
- Example: candidates for “havy”:

Edits	ED	Possible Permutations	Candidates (In Dictionary)
Delete	1	{avy, hvy, hay, hav}	{hay}
Insert	1	{ahavy, haavy, havay, havya, …, heavy, …}	{heavy, …}
Substitute	1	{aavy, haay, hava, …, have, …, wavy, …}	{have, wavy, …}
Transpose	1	{ahvy, hvay, hayv}	
…	2	…	{haven, hairy, happy, harry, hacky, …}

NW, 1To1 - Ranker

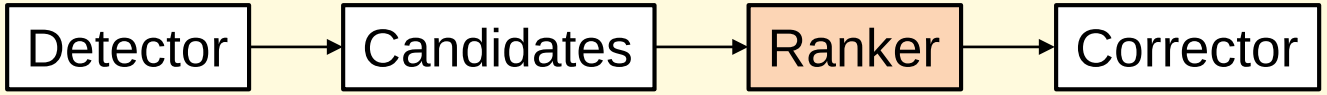


- Orthographic Similarity
  - edit distance similarity, phonetic similarity and overlap similarity
- Frequency Score: (unigrams and WC from consumer corpus)
- Noisy Channel
  - Use orthographic score as the error model score
  - Use frequency score as the language model score
- Context Score (Word Embedding): word2vec (consumer corpus)
- CSpell (2 stages ranking)

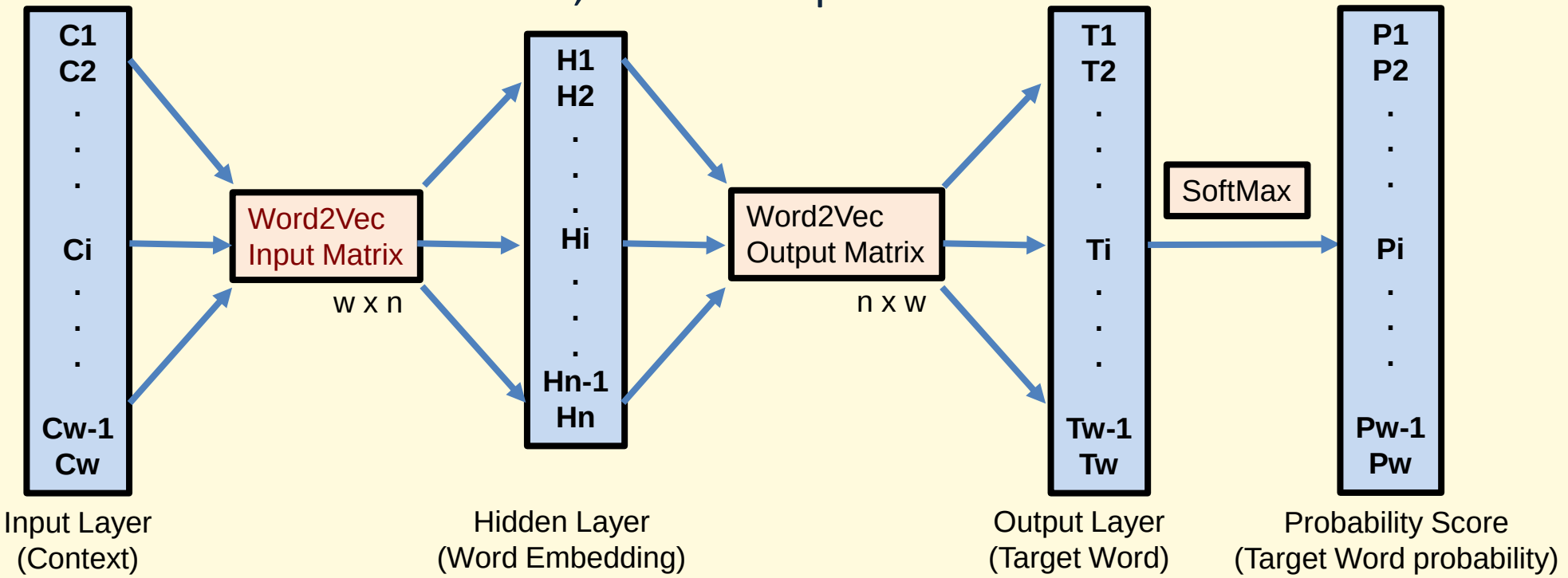
Word2vec: Word Embedding

- Word Vectors (word2vec, Tomas Mikolov):
  - Each word has an associated vector
  - Represent the ‘meaning’ of a word in some abstract way
  - Capture meaningful syntactic and semantic regularities
- Examples
  - Man -> Woman, Uncle -> Aunt
  - France -> Paris, Italy -> Rome
  - Kings – Man + Woman = Queens
- 2 Models:
  - Continuous bag-of-words (CBOW): predict a word from context
  - Continuous Skip-gram: predict context from a word

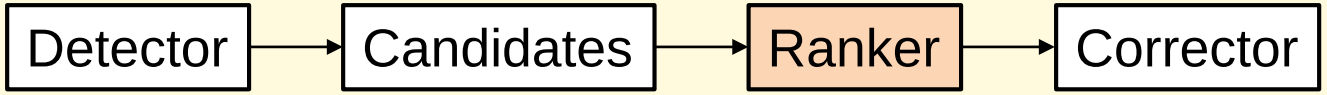
NW, 1To1 – Context Score



- Context Score, Word2vec, CBOW model:
  - Duel embedding: use both input matrix (context) and output matrix (target) to calculate context scores, 10.77% improvement in F1

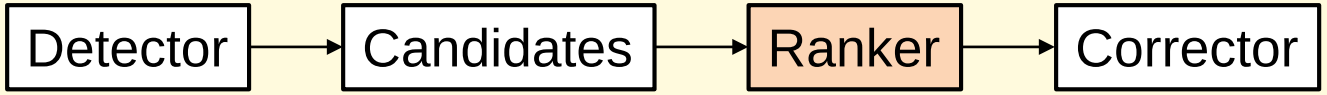


NW, 1To1 – Combined Rules



- CSpell (2 stage ranking)
  - Use orthographic scores to refine candidates in stage 1
    - No phonetic and overlap knowledge were used in candidate generator
  - Stage 2 ranking:
    - 1st: rank candidates by context score (best precision)
    - 2nd: rank candidates by noisy channel score (increase recall)
    - 3rd: if only 1 candidate are qualified in playoff, orthographic score must > threshold
    - Ignore ranking in stage-1
  - Similar to determining a championship in sports through a regular season selection (stage 1) with the best team in the playoffs (stage 2) winning the championship.

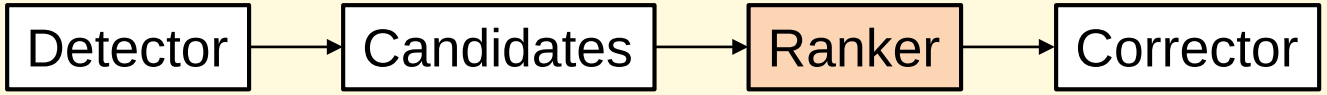
NW, 1To1 – Results



- 2 stage is better than 1 stag

Stage-1	Stage -2	Precision	Recall	F1	
Orthographic	N/A	0.7606	0.7636	0.7621	
Frequency	N/A	0.6970	0.6925	0.6948	
Noisy Channel	N/A	0.7134	0.7171	0.7152	
Context	N/A	0.8035	0.5917	0.6815	
Ensemble	N/A	0.7516	0.7545	0.7531	
Orthographic	Frequency	0.8241	0.7687	0.7955	
Orthographic	Noisy Channel	0.8255	0.7700	0.7968	
Orthographic	Context	**0.8996**	0.5672	0.6957	← Best precision
Orthographic	CS, NC, OR-T	0.8047	0.7842	**0.8115**	← Best F1

NW, 1To1 – Ranker (Context)

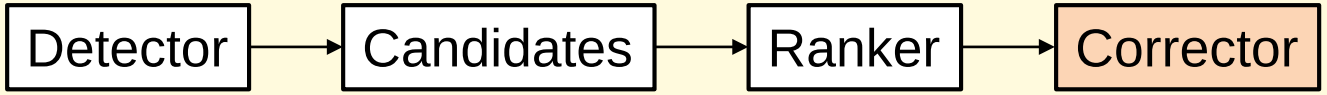


- CSpell examples on “havy”:
- Use context score for qualified candidates to improve precision & F1.

Input Text	Output Correction	Orthographic Scores	Context Scores: heavy|have|hay|wavy
havy	**have***	2.650	0.0000|0.0000|0.0000|0.0000
havy duty	**heavy** duty	2.705	**0.0597**|-0.0302|-0.0053|0.0074
havy diabetes	**have** diabetes	2.650	-0.0667|**0.0586**|-0.0518|-0.0813
havy fever	**hay** fever	2.560	-0.1331|0.2280|**0.2292**|-0.0391
havy lines	**wavy** lines	2.550	-0.0170|-0.0410|-0.0702|**0.1495**
…	…	…	…

* Have the highest noisy channel score

NW, 1To1 - Corrector



- Replace the target token with the selected best candidate (if exists).

NW, Split

- Non-Word, Split:

Correction Step	Descriptions
Detector	Same as NW, 1To1
Candidates	• Split: use space, “ “, to generate candidates • < maxSplitNo • In the Lexicon multiword • split dictionary (no AA, seing -> se i ng)
Ranker	Same as NW 1To1
Corrector	Use Java **flatMap** to flat a multiword token to multiple single word tokens

NW, Merge

- Non-Word, Merge:

Correction Step	Descriptions
Detector	• Not in the split dictionary • Not exceptions (same as NW, 1To1)
Candidates	• Merge with space (“ “) and hyphen (“-”) • < maxMergeNo • Original term not in multiword dictionary • Context not exceptions: digit, punctuation, url, eMail, etc. • In suggestion dictionary (AA: dur ing -> during)
Ranker	• Best context score (!= 0), then best frequency
Corrector	• Reconstruct the whole text (**multiple & overlap merges**)

Non-Word Correction Examples:

- Examples:

Correction Type	Input Text	Output Correction
1To1	• dianosed	• diagnosed
Split	• knowabout	• know about
Merge	• stiff n ess	• stiffness

Real-Word vs. Non-Word Corrections:

- Real Word Correction:
  - Relies on context scores (relevance), more restrictions on detection and correction
  - Strategy is to increase recall on real-word correction and preserve precision
- Comparison:

Correction Step	Non-Word	Real-Word
Detector	• Invalid word (not in dictionary)	• Valid word (in dictionary) • Has Word2Vec score • WC > Threshold
Candidates	• Simple operation	• More restrictions on merge context, split words, and 1To1 candidates • Candidate must have word2Vec, WC > threshold, etc.
Ranker	• Find the best candidate	• Find and validate the best candidate
Corrector	• Replacement/Flat map/Reconstruct	• Same as NW

RW - Merge

- Real-Word, Merge:

Correction Step	Descriptions
Detector	• Not previously corrected • Valid word in the Split dictionary • Not exceptions (same as NW, 1To1)
Candidates	• Merged with " ", not "-" • < maxMergeNo • Original term not in multiword dictionary • Context not exceptions: digit, punctuation, url, eMail, etc. • In suggestion dictionary (not AA) • Has word2Vec and WC > threshold • **No short word merges** (me at -> meat)
Ranker	• Find the best candidate by context score • Validate: best candidate context score * confidence factor > original context score
Corrector	• Reconstruct the whole text

RW - Split

- Real-Word, Split:

Correction Step	Descriptions
Detector	• Not previously corrected • Valid word in the dictionary • Not exceptions (same as NW, 1To1) • Has word2Vec and WC > threshold
Candidates	• Split candidates (same as NW-Split) • **No short word splits** (another -> a not her, an other) • Split word check: has word2Vec, WC > threshold, not units or proper nouns
Ranker	• Same as RW, merge (confident factor = 0.01)
Corrector	• Flat map

RW - 1To1

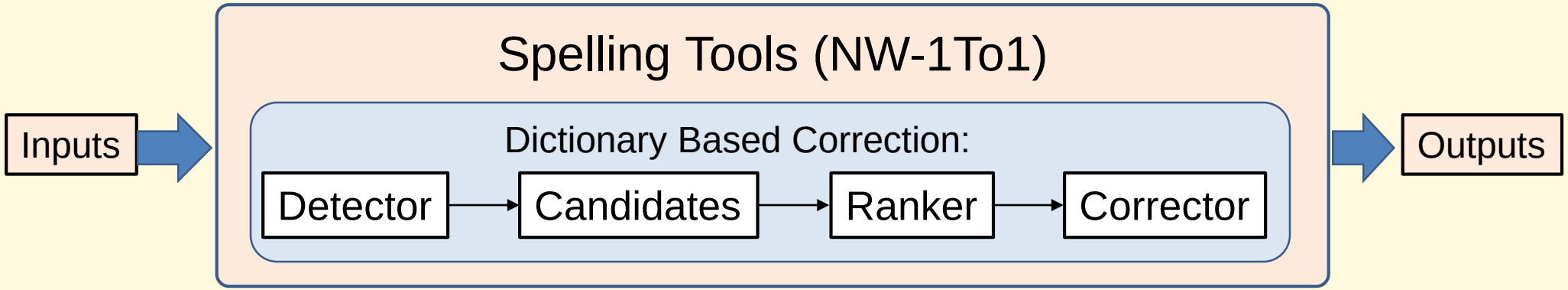
- Real-Word, Split:

Correction Step	Descriptions
Detector	• Not previously corrected • Valid word in the dictionary • Not exceptions (same as NW, 1To1) • Length >= 2 • Has word2Vec and WC > threshold
Candidates	• Generate candidates (same as NW, 1To1) • Candidates have Word2Vec, WC > threshold, length >= 2 • Not inflectional variants of the original token • Heurists rules on phonetic and orthographic similarity
Ranker	• Find: The best candidate has the highest score of orthographic, frequency, edit, phonetic, overlap • Validate: Best candidate context score > 0 and original context score < 0
Corrector	• Replacement

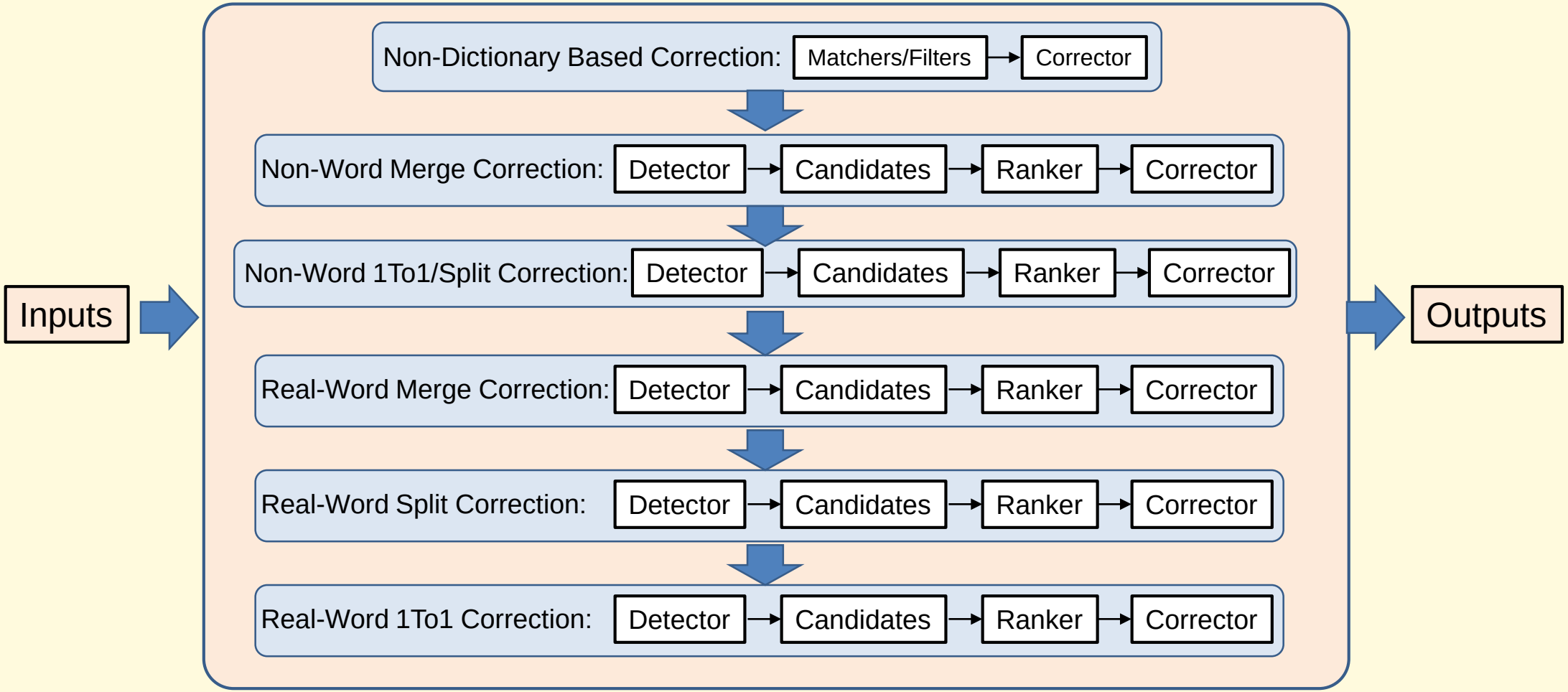
Real-Word Correction Examples:

Correction Type	Input Text	Output Correction
1To1	• bowl movement	• bowel movement
Split	• for along time	• for a long time
Merge	• early on set	• early onset

Traditional Spelling Tools Model



CSpell Architecture



Multiple Corrections on Combined Errors

Input Text	Output Correction
He was dianosed early on set deminita 3years ago.	He was **diagnosed** early **onset** **dementia** **3 years** ago.

NW 1To1 — RW Merge — NW 1To1 — Non-Dic Split

Input Text	Output Correction
I have a shuntfrom2007.	I have a **shunt from 2007**.

NW Split — Non-Dic Split

Input Text	Output Correction
I am permanently depressed and was on 2 or 3 different anti depresants.	I am permanently depressed and was on 2 or 3 different **antidepressants**.

RW Merge — NW 1To1

Development Set & Test Set

	Development Set *	Test Set**
Questions	471	224
Token	24,837	16,707
Tags	1,008	1,946
Error rate	0.04	0.12

- An Ensemble Method for Spelling Correction in Consumer Health Questions, H Kilicoglu, M. Fiszman, K Roberts, D Demner-Fushman, 2015 AMIA Symposium, Chicago, P727-736
- Consumer questions with the highest OOV (out of vocabulary).

Development Set Detection

- Non-word Detection:

Method	Precision	Recall	F1
Baseline (Ensemble)	79.39%	84.63%	0.8193
CSpell	92.38%	86.18%	0.8917

- Real-word Included Detection:

Method	Precision	Recall	F1
Baseline (Ensemble)	80.78%	60.17%	0.6897
CSpell	92.89%	71.78%	0.8098

Test Set Detection

- Non-word Detection:

Method	Precision	Recall	F1
Baseline (Ensemble)	76.19%	75.56%	0.7588
CSpell	87.84%	87.47%	0.8765

- Real-word Included Detection:

Method	Precision	Recall	F1
Baseline (Ensemble)	82.10%	56.45%	0.6690
CSpell	89.00%	71.49%	0.8093

Development Set Correction

- Non-word Correction:

Method	Precision	Recall	F1	Time
Baseline (Ensemble)	66.91%	71.32%	0.6904	~ 1 hr.
CSpell	80.47%	78.42%	0.8115	~ 1 min.

- Real-word Included Correction:

Method	Precision	Recall	F1	Time
Baseline (Ensemble)	72.01%	53.63%	0.6147	~ 1 hr.
CSpell	84.16%	65.04%	0.7338	~ 5 min.

Test Set Results

- Non-word Correction:

Method	Precision	Recall	F1	Time
Baseline (Ensemble)	61.90%	61.40%	0.5684	<1 hr.
CSpell	76.60%	76.28%	0.7644	< 1 min.

- Real-word Included Correction:

Method	Precision	Recall	F1	Time
Baseline (Ensemble)	69.75%	47.96%	0.5684	~ 1 hr.
CSpell	76.07%	63.41%	0.6917	~ 3 min.

Future Work

- Public release
  - Installation package
  - Web site development
  - Documentation (desginDoc, userDoc and JavaDoc)
- Integrate with Consumer Q & A project
- Establish a (more) comprehensive consumer health corpus (for better word2vec and frequency)
- Update dictionaries with the latest release of the SPECIALIST Lexicon
- Other feature enhancements

Questions

- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov