The SPECIALIST Lexicon and NLP Tools
(CSpell – Spell Checker for Consumer Language)

By: Dr. Chris J. Lu

NLM – LHNCBC - CGSB

Sept., 2018

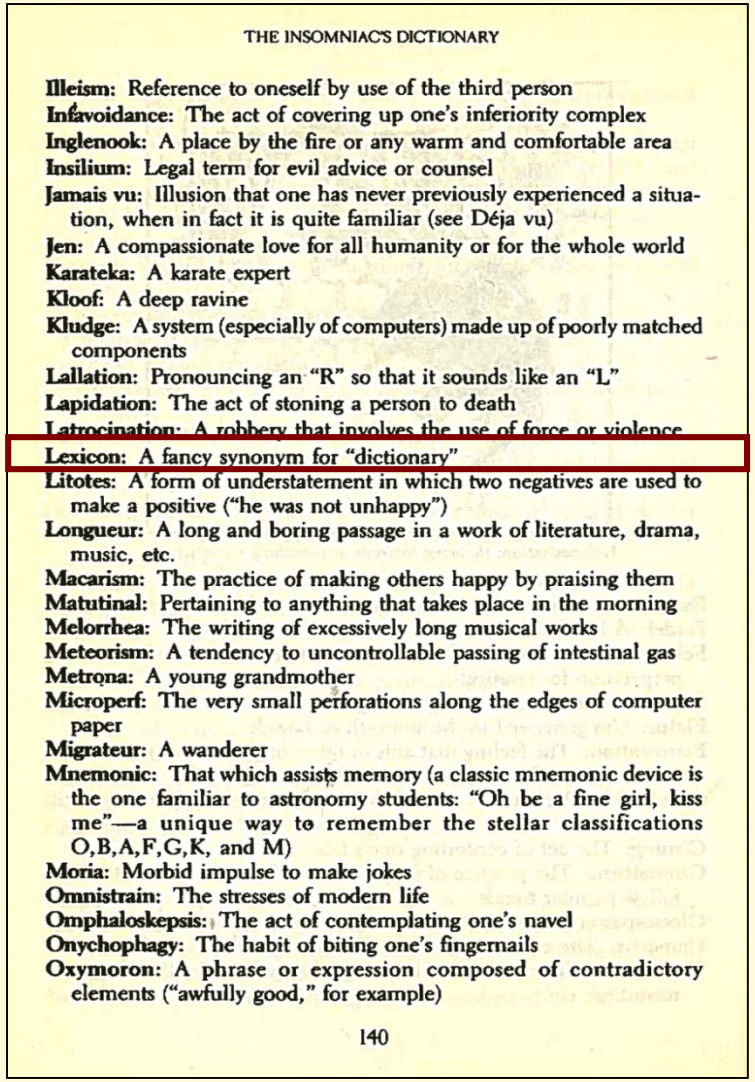
- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov

Outline

- Introduction
  - The SPECIALIST Lexicon
  - The SPECIALIST NLP Tools (Lexical Tools)
- Applications - CSpell
  - Natural Language Processing (NLP)
  - CSpell (Spell Checker for Consumer Language)
- 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**

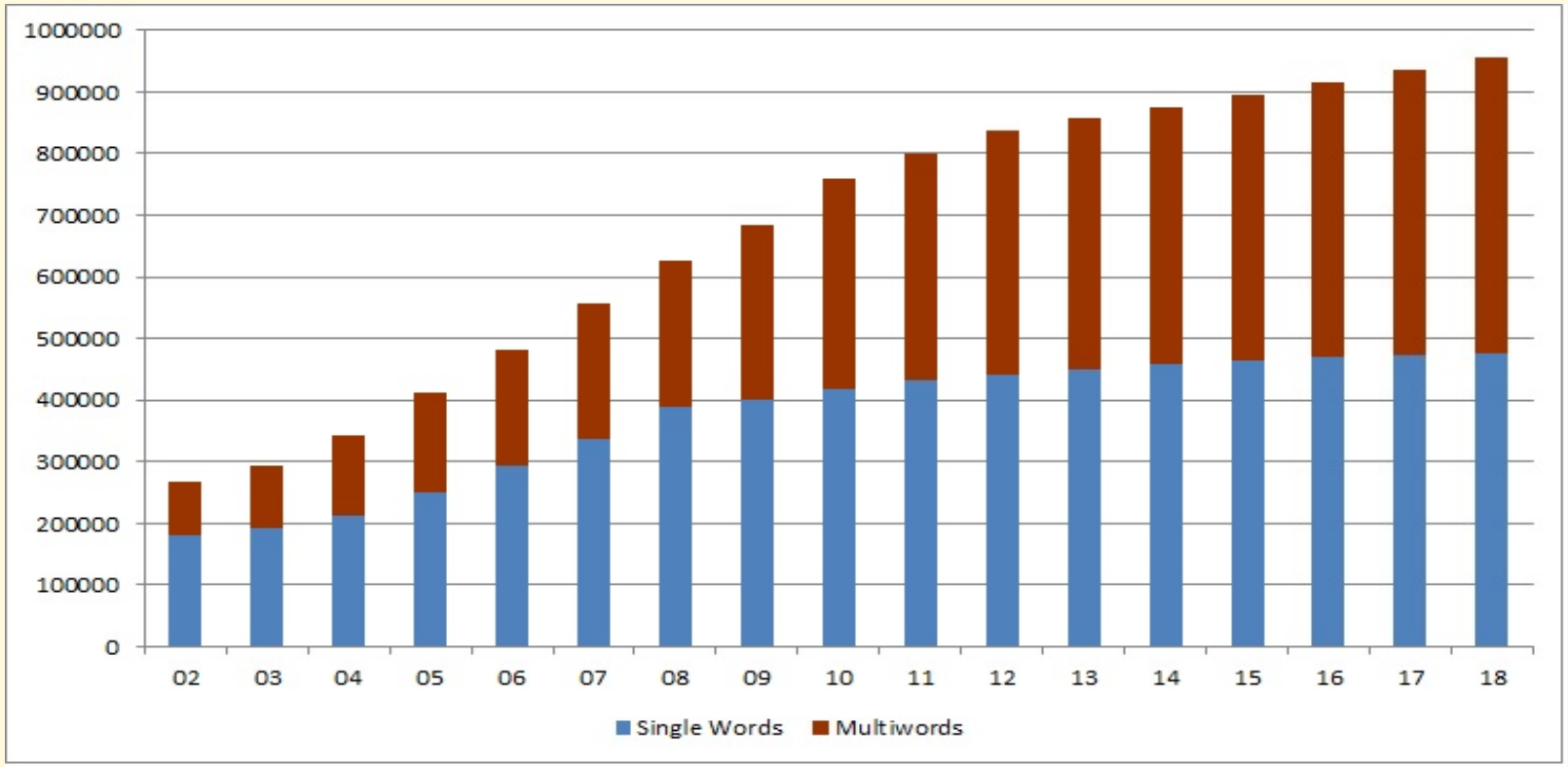
Team of Lexicon Builders

- 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 in Lexical Records

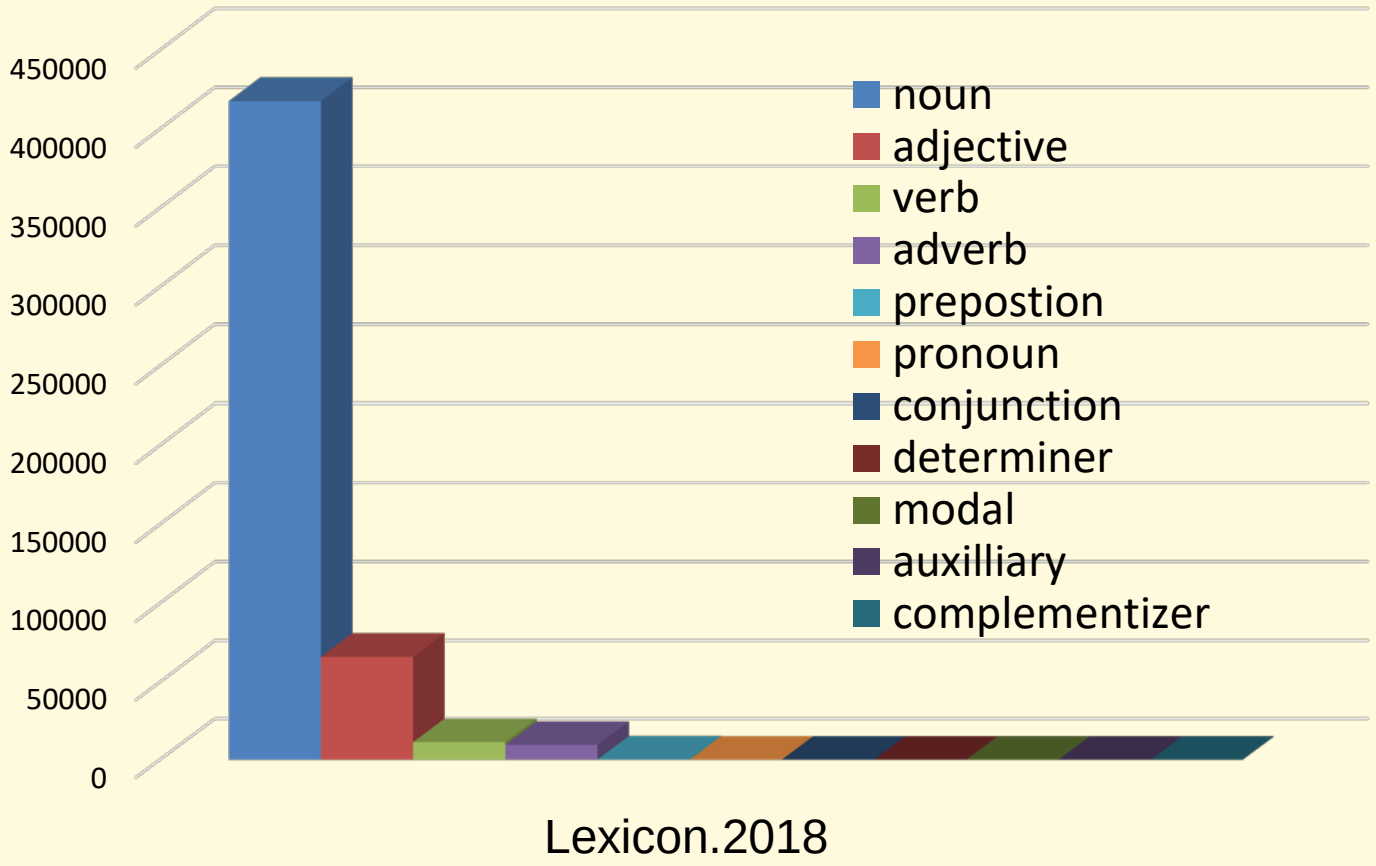
- What is a word?
  - spelling, POS, spelling variants, inflectional variants, meaning, etc.

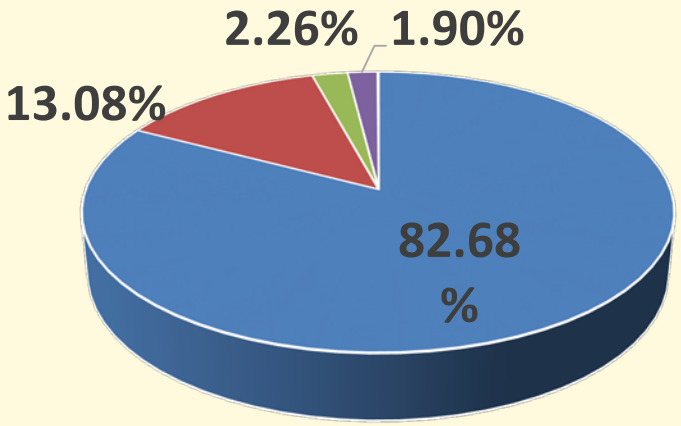
- Lexical term (lexeme): single words and multiwords
  - Space(s): ice-cream vs. ice cream; x ray, x-ray, xray,
- Four criteria for Lexicon terms:
  - Part of Speech (POS):
    - tear break up time, frog erythrocytic virus, cardiac surgery
  - Inflection morphology (uninflection):
    - left pulmonary veins (“left pulmonary vein” and “leave pulmonary vein”)
  - Specific meaning:
    - <u>hot</u> dog (<u>high temperature</u> 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)





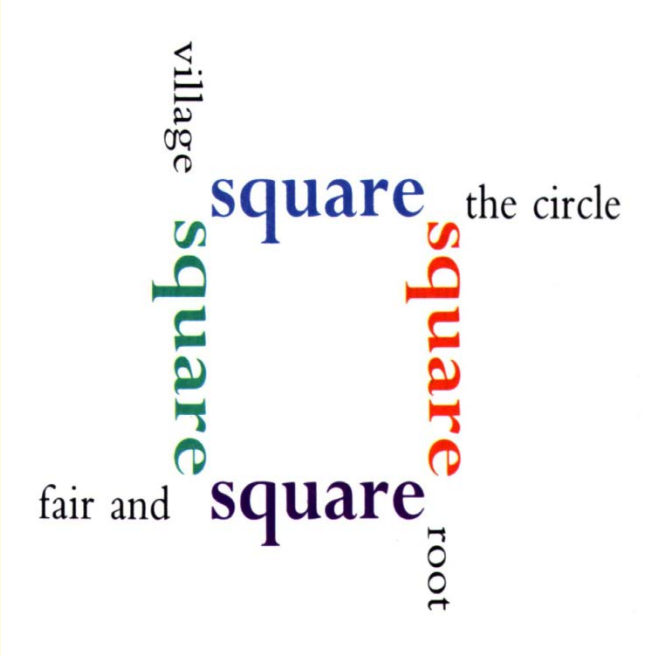
Lexical Records & POS

```
{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 - transport<u>ation</u>, transport<u>able</u>, transport<u>er</u>, …
  - prefix – <u>auto</u>transport, <u>intra</u>transport, <u>pre</u>transport, …
  - conversion (zero) - transport (verb), transport (noun)

Orthography (Spelling Variation)

- Same meaning and pronunciation, different spelling (words in British English, Australian English and Canadian English)
  - color|colour
  - grey|gray
  - align|aline
  - Grave's disease|Graves's disease|Graves' disease
  - civilize|civilise
  - fetus|foetus|fœtus
  - centre|center
  - spelt|spelled
  - ice cream|ice-cream
  - xray|x-ray|x ray

Syntax - Verb Complements

- intran
  - He will treat.
- tran=np
  - He treated the patient.
- ditran=np,pphr(with,np)
  - He 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
}
```

```
{base=colorable
spelling_variant=colourable
entry=E0523388
}
```

```
{base=chromatic
entry=E0016801
        cat=adj
        variants=inv
        position=attrib(3)
        position=pred
        stative
        nominalization=chromaticness|noun|E0226323
        nominalization=chromaticity|noun|E0332902
}
```

UTF-8 (Since 2006)

```
{base=resume
spelling_variant=résumé
spelling_variant=resumé
entry=E0053099
        cat=noun
        variants=reg
}
```

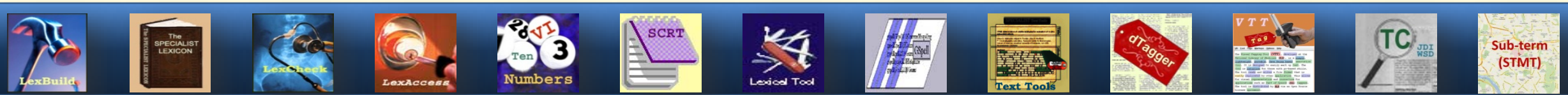
```
{base=deja vu
spelling_variant=deja-vu
spelling_variant=déjà vu
entry=E0021340
        cat=noun
        variants=uncount
}
```

```
{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.

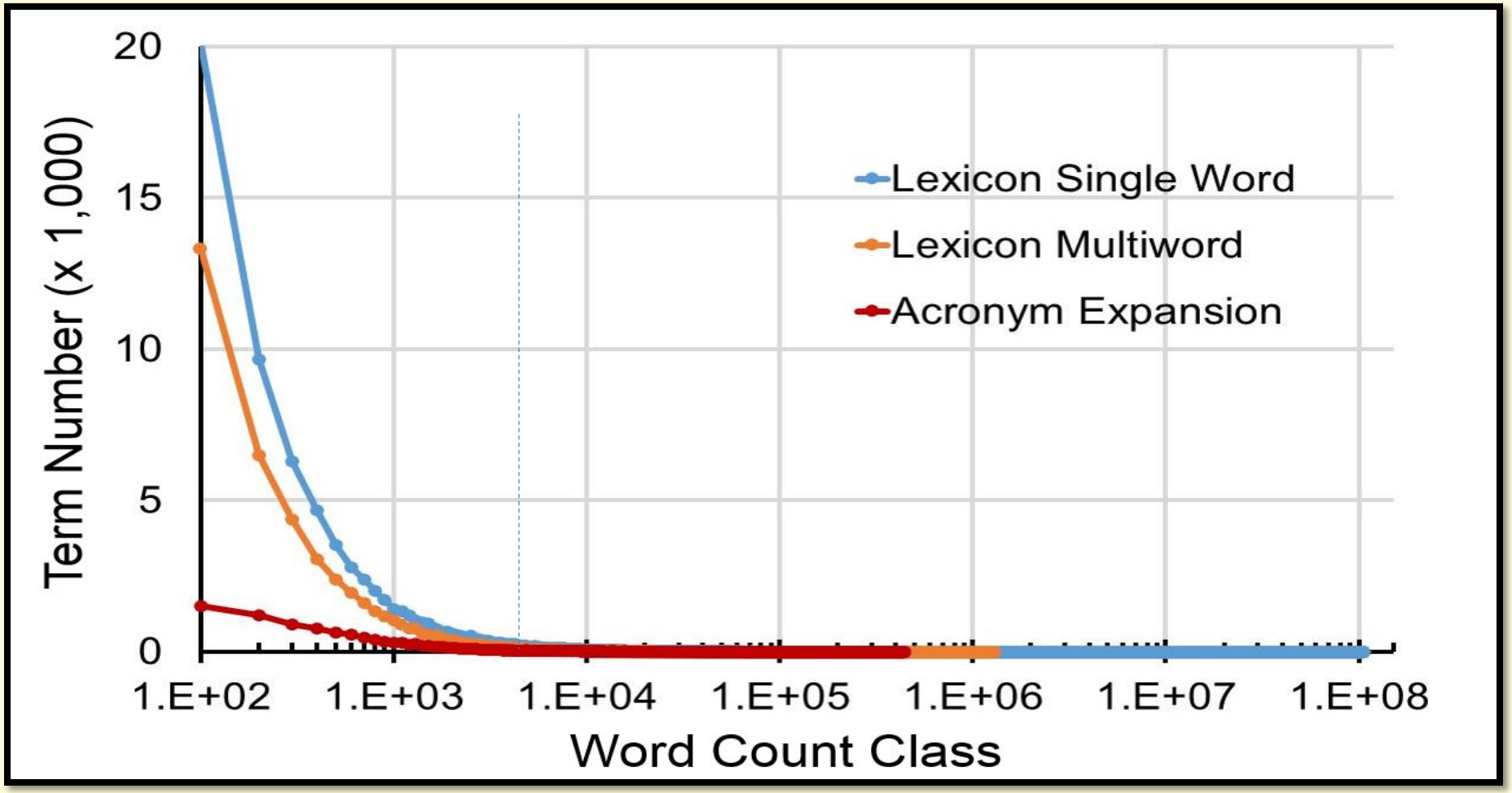
Lexicon Unigram Coverage – With Frequency (WC)

- Total word count for MEDLINE (2016): 3,114,617,940
- Lexicon covers > 98.4% 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.
** Words with highest count: of, the, in, and to, a, with, The, etc…

The Frequency Spectrum of Lexicon (Multi)words on MEDLINE



The Frequency Spectrum of Alice in Wonderland

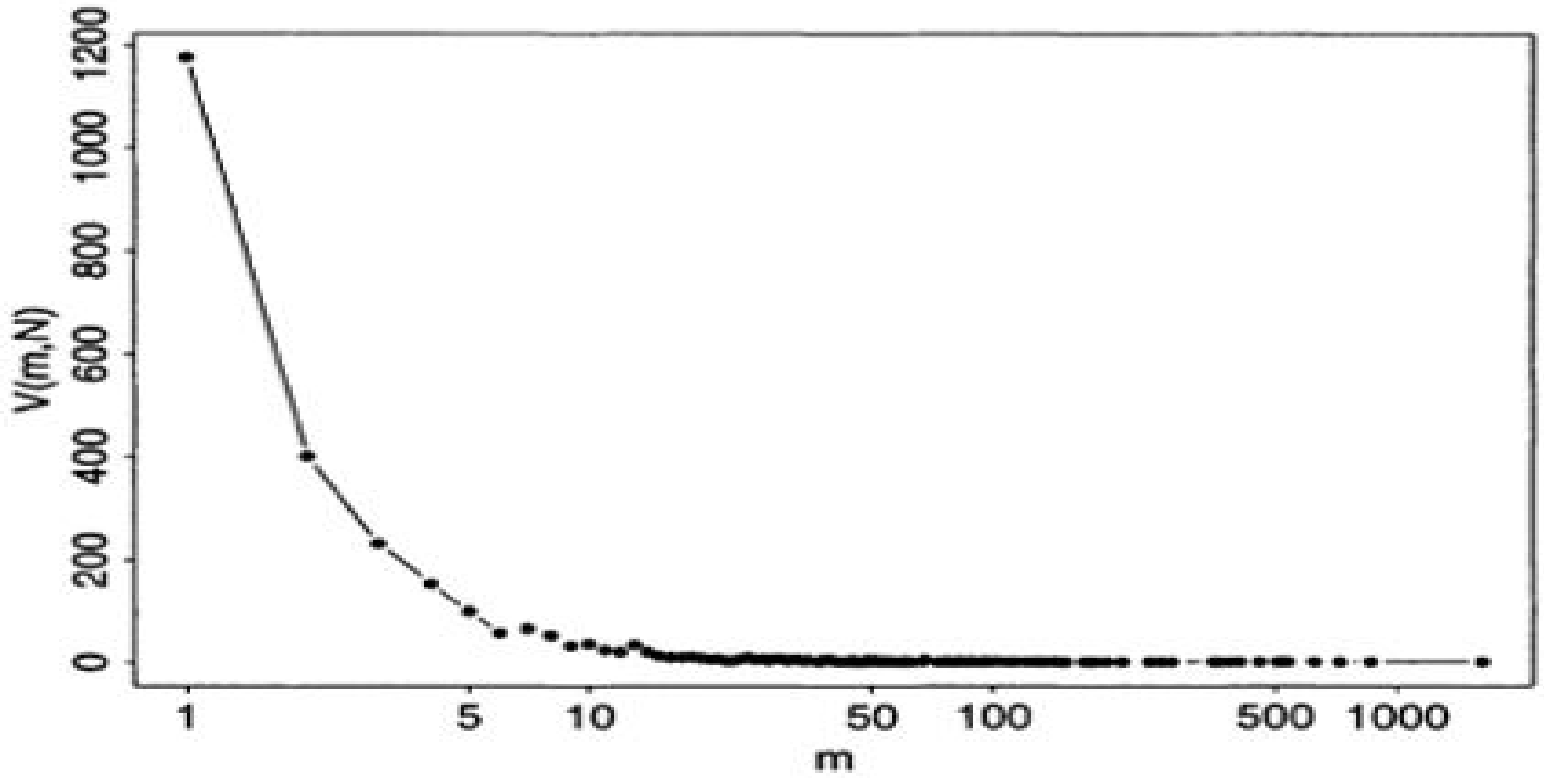


Figure 1.3: *The frequency spectrum of Alice in Wonderland (m: frequency class; V (m, N) : number of types with frequency m).*

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 → spelling variant
- cat=verb → part of speech
- variants=reg → inflectional variant
- tran=pphr(from,np) → chunker
- nominalization=generalisation|noun|E0029525 → derivational variant, synonym

2. 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 - 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

Lexical Variants Generation (LVG)

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.

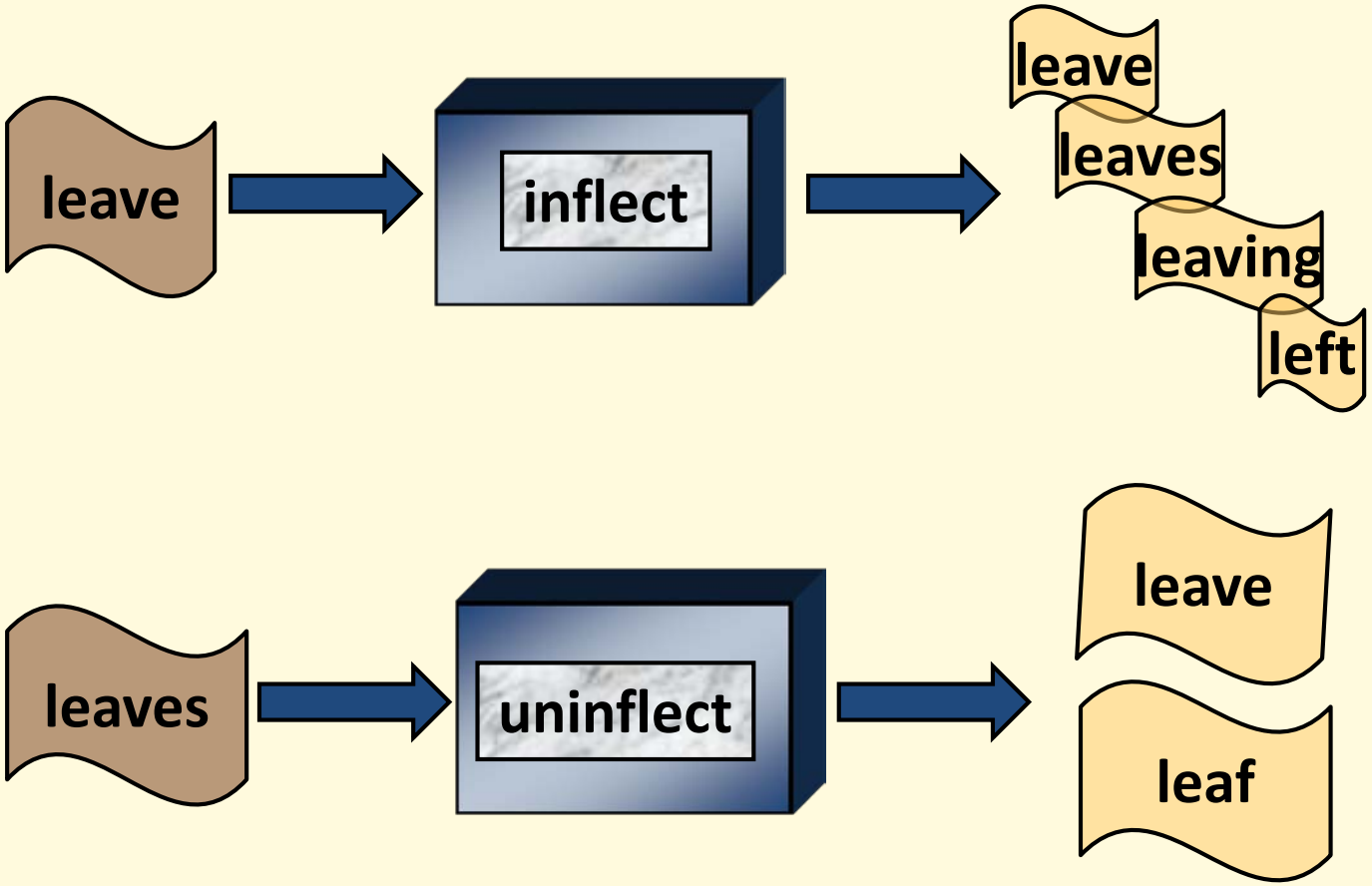
LVG - Lexical Variants Generation

- 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

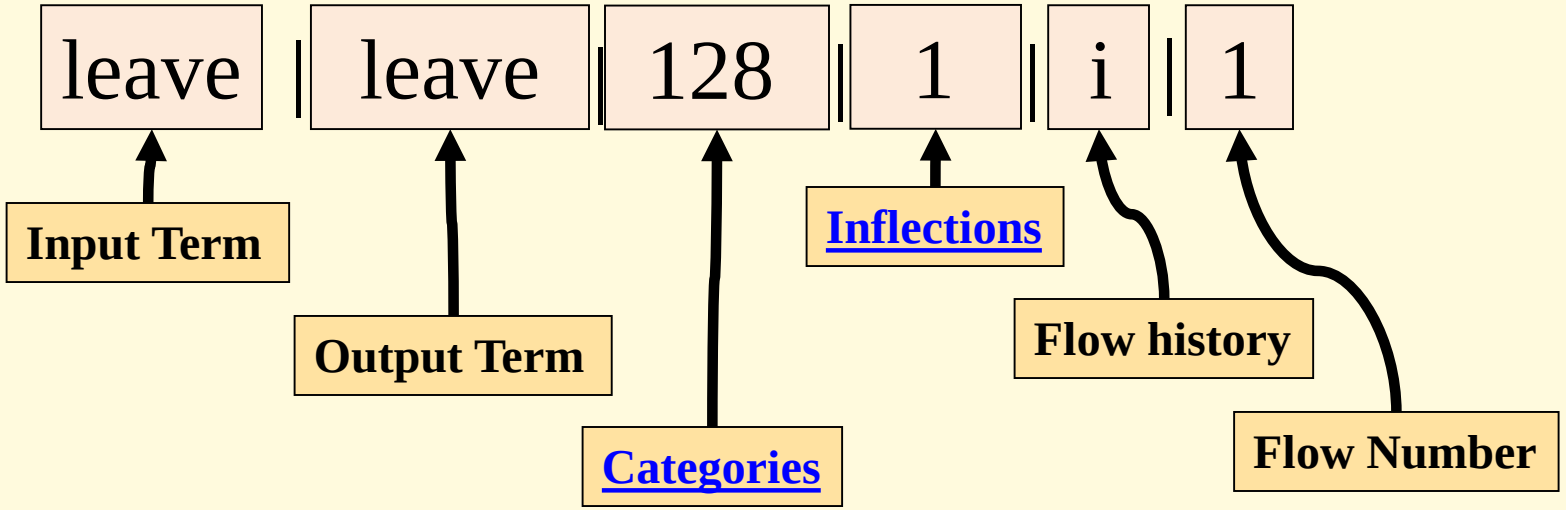


LVG Flow Component – CmdLine

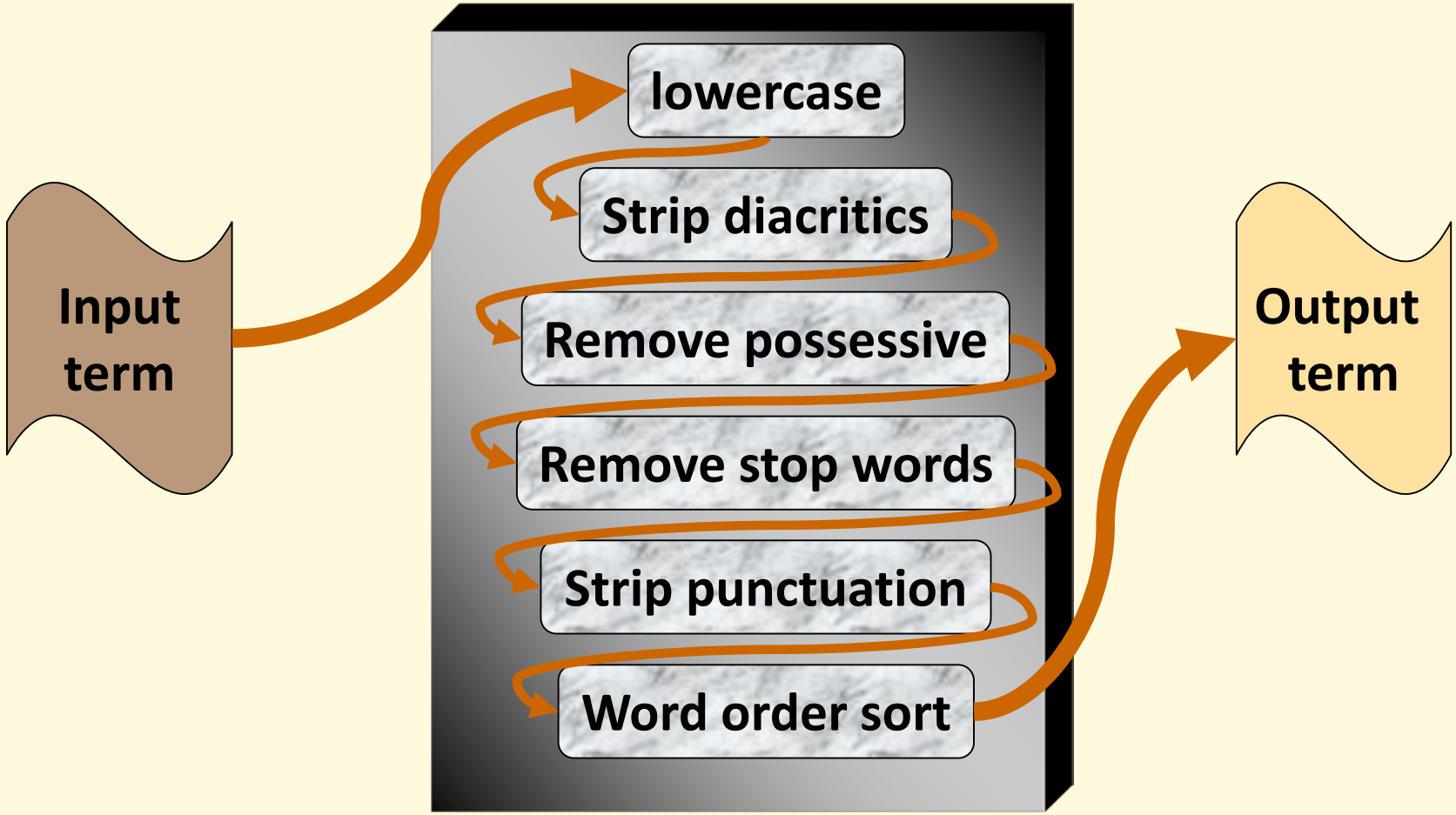
```
> lvg -f:i
leave
leave|leave|128|1|i|1|
leave|leave|128|512|i|1|
leave|leaves|128|8|i|1|
leave|left|1024|64|i|1|
leave|left|1024|32|i|1|
leave|leave|1024|1|i|1|
leave|leave|1024|262144|i|1|
leave|leave|1024|1024|i|1|
leave|leaves|1024|128|i|1|
leave|leaving|1024|16|i|1|
```

LVG Flow Component – Fielded Output

```
> lvg –f:i
  leave
```



LVG – A Serial Flow



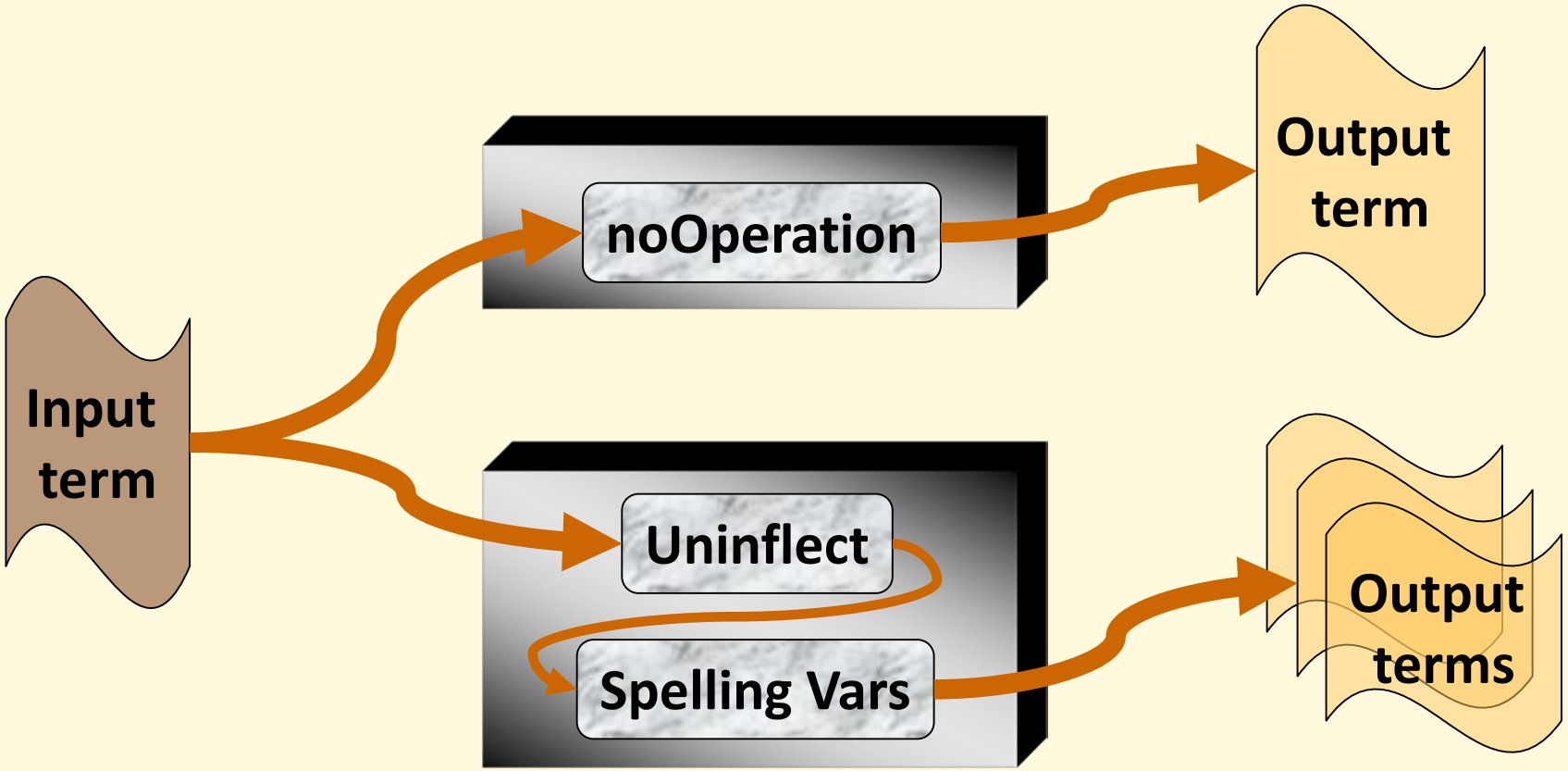
- Flow components can be arranged so that the output of one is the input to another.

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
	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

	“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	
q7: Unicode core Norm	
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	
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, written or signed
- 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.

☺

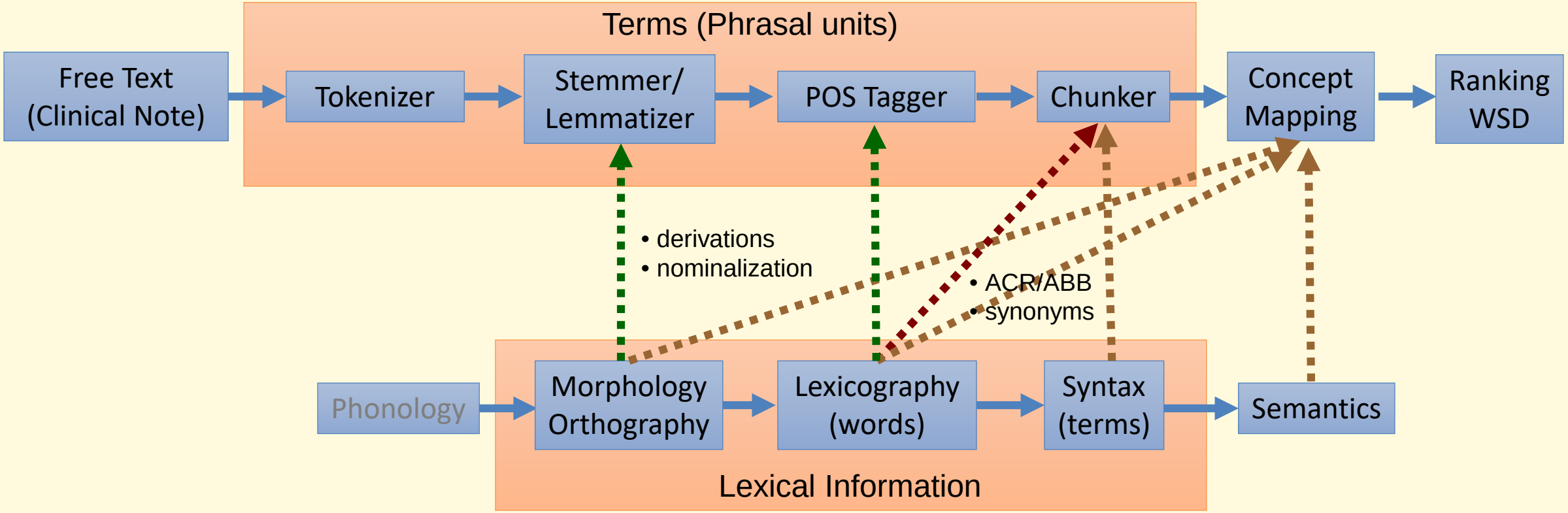
Concept Mapping Challenges

- 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

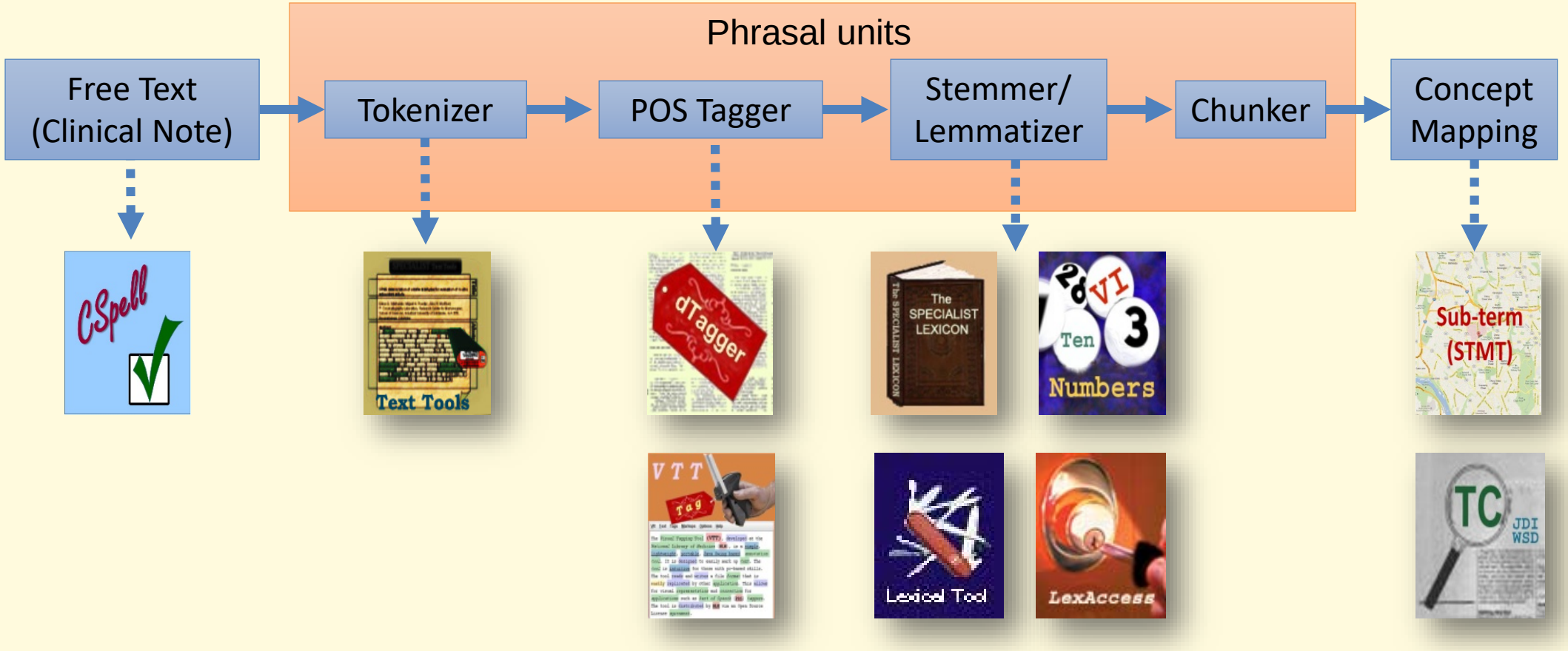


NLP Applications

- 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

The SPECIALIST NLP Tools



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

☺

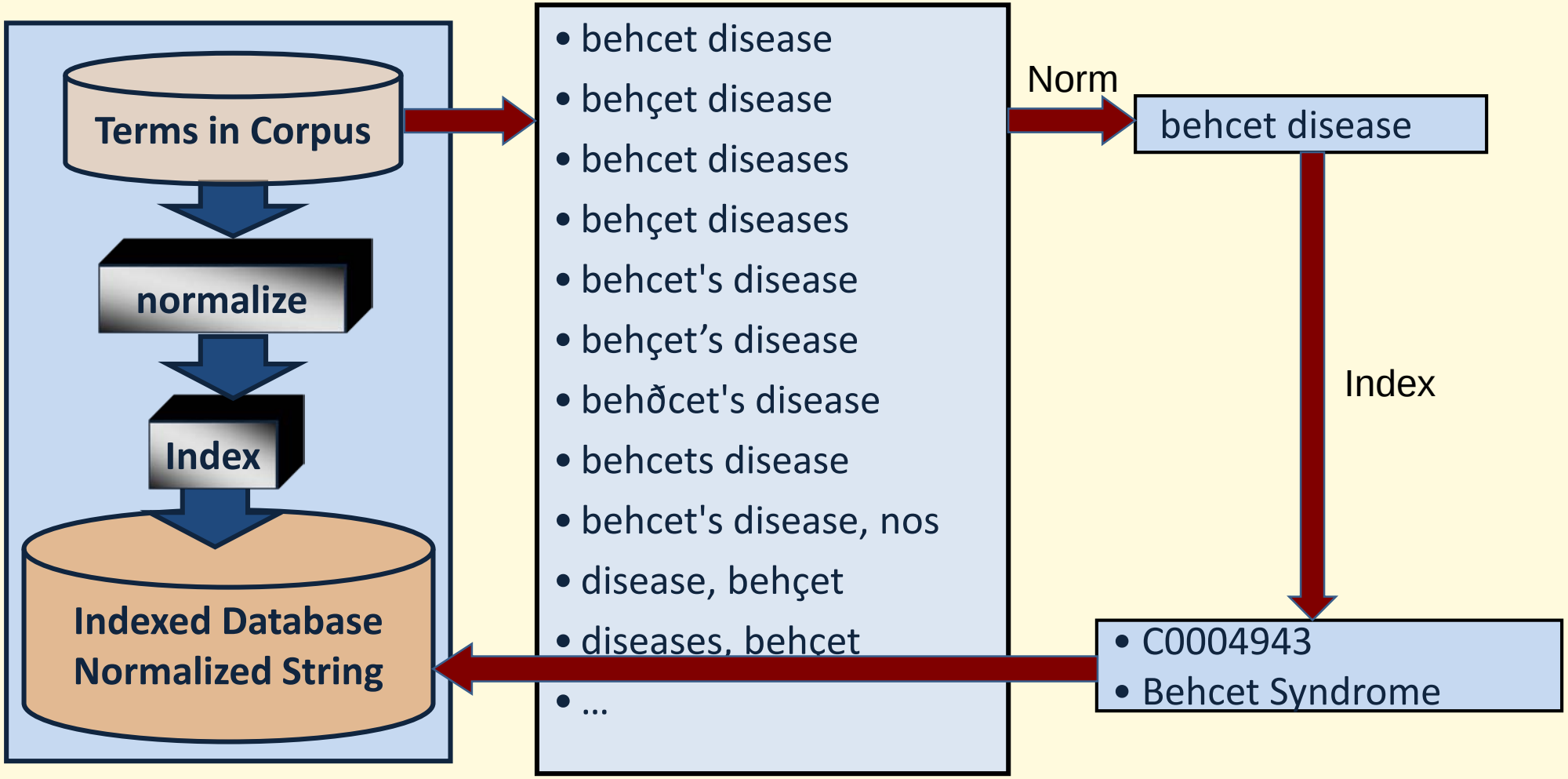
NLP – Concept Mapping

- Normalization (same lexical 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 lexical 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…

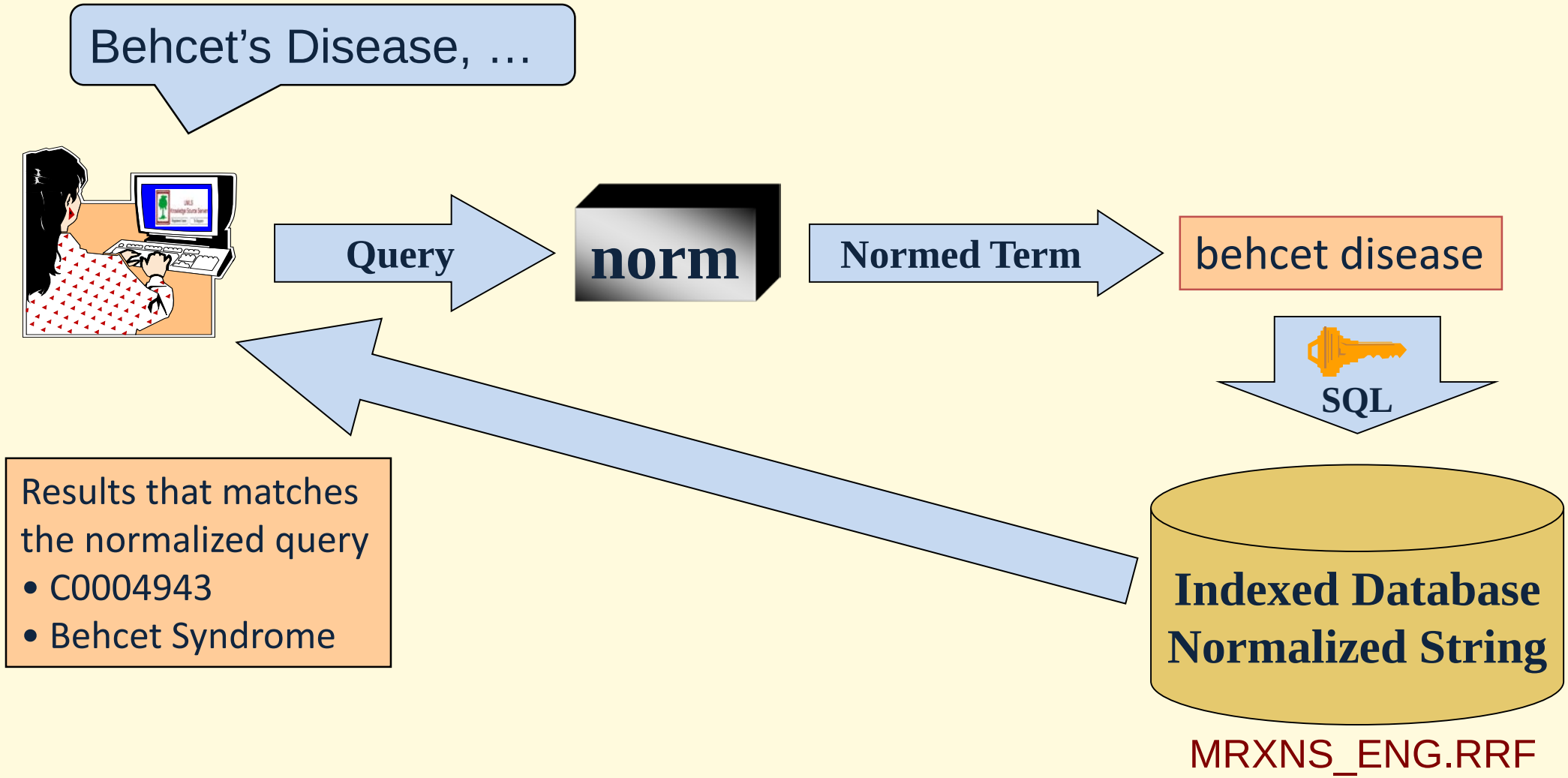
Lexical Tools – Norm

	Behçet’s Diseases, NOS
q0: map Unicode symbols to ASCII	Behçet's Diseases, NOS
g: remove genitives	Behçet Diseases, NOS
rs: remove parenthetic plural forms	Behçet Diseases, NOS
o: replace punctuation with spaces	Behçet Diseases NOS
t: strip stop words	Behçet Diseases
l: lowercase	behçet diseases
B: uninflect each words in a term	behçet disease
Ct: retrieve citations	behcet 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)

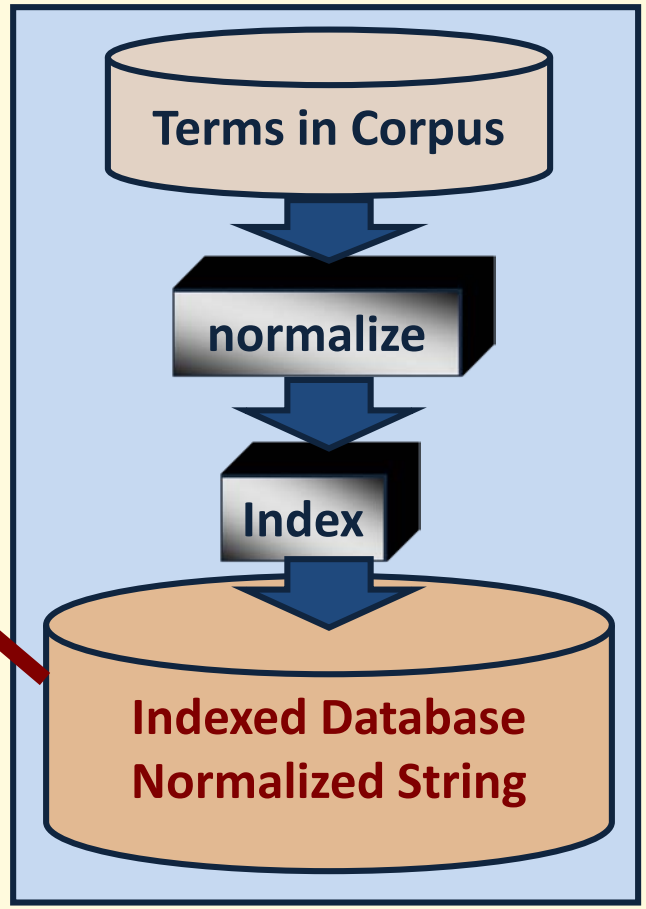


NLP – Norm (Application)

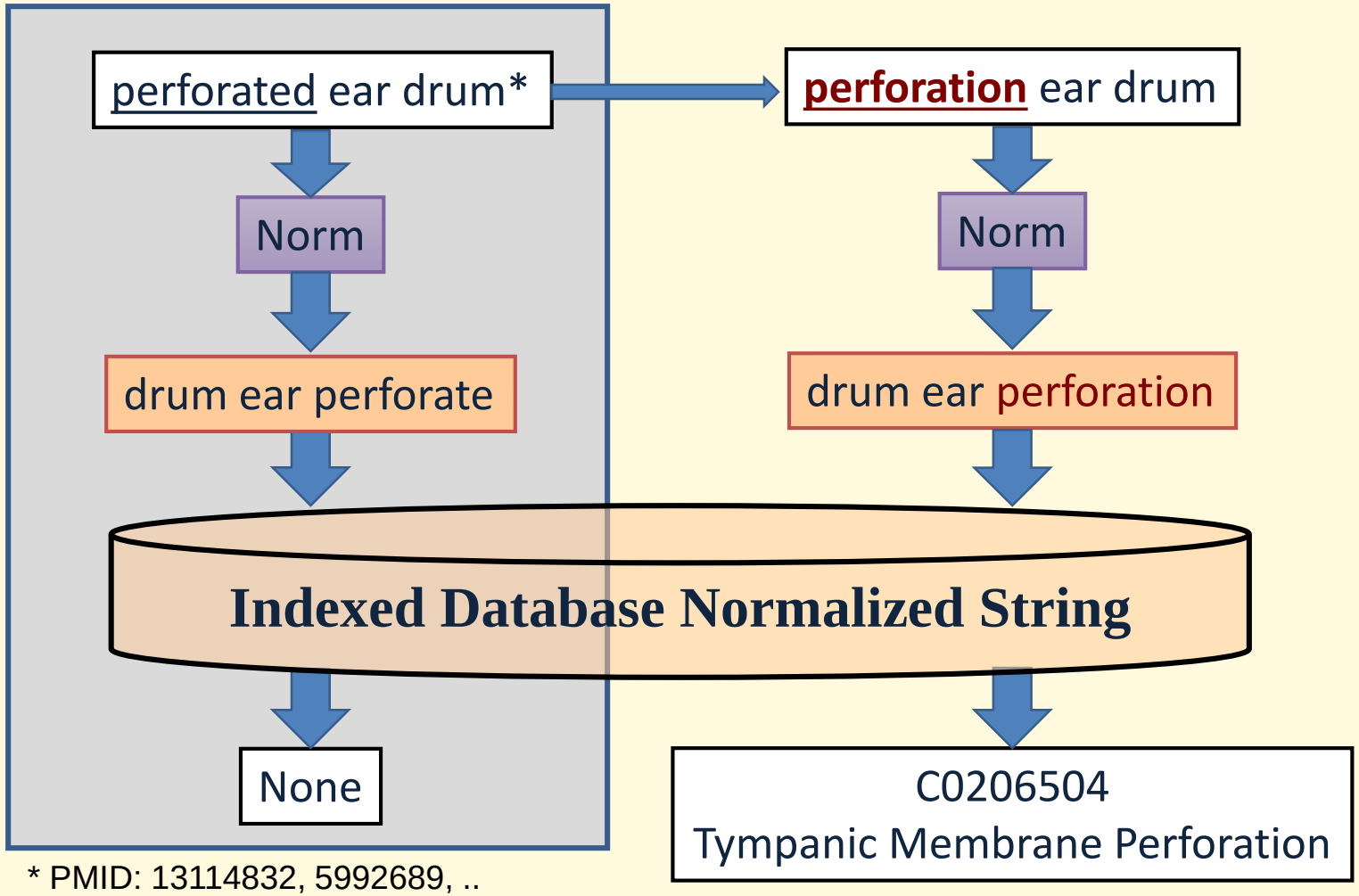


UMLS Metathesaurus

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

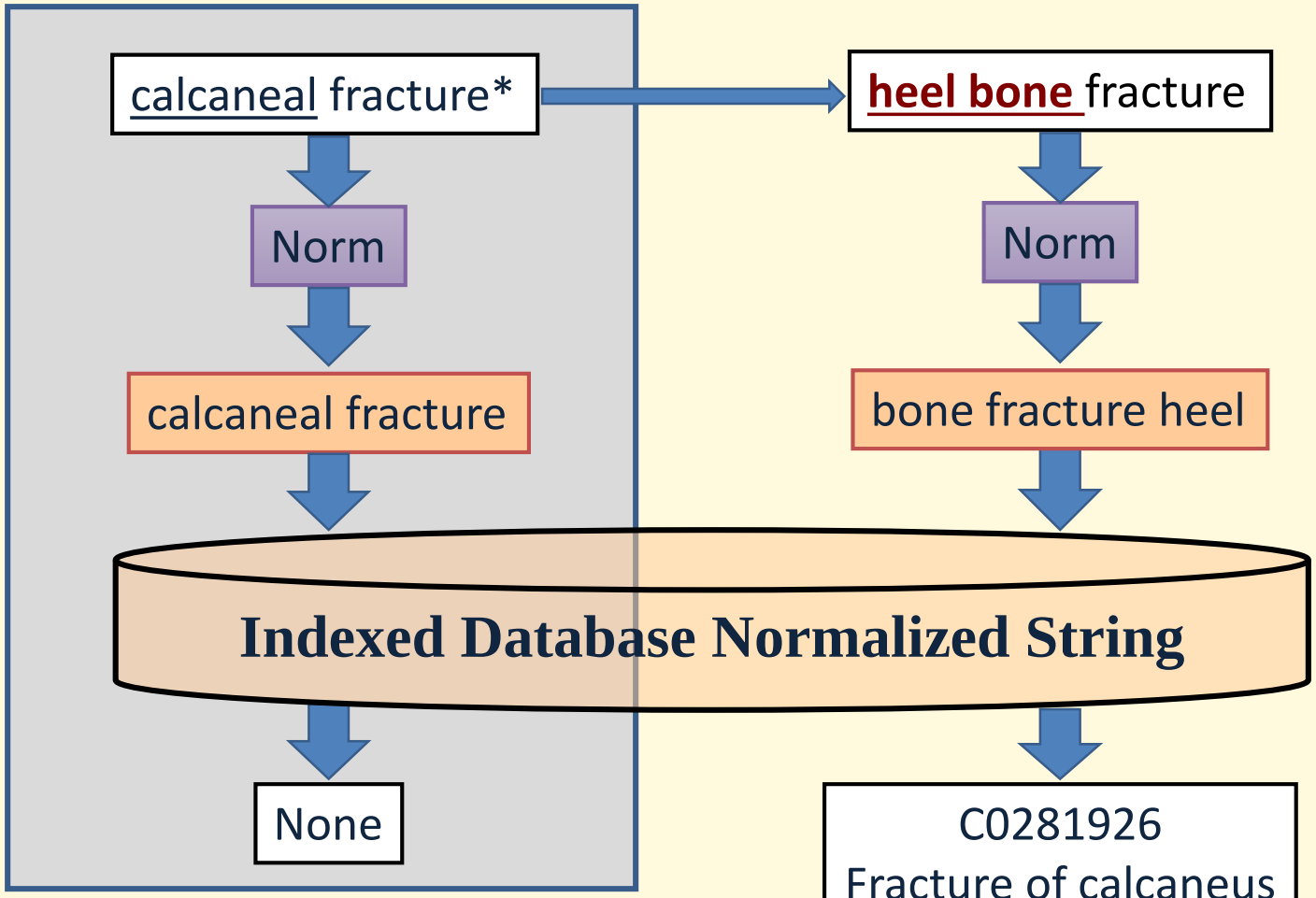


NLP – Query Expansion (Derivation)



* PMID: 13114832, 5992689, ..

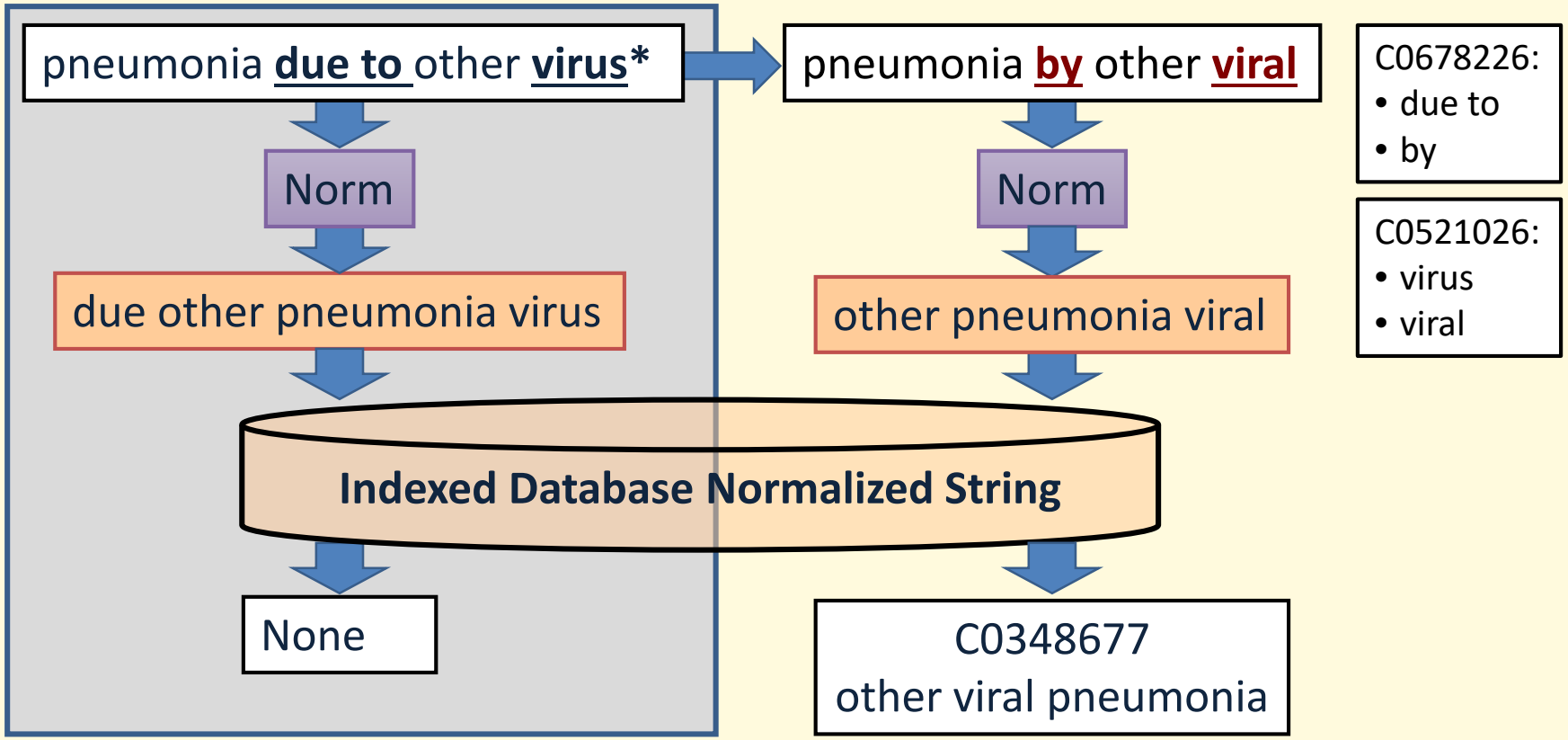
NLP – Query Expansion (Synonym)



C0006655:
- calcaneal
- heel bone

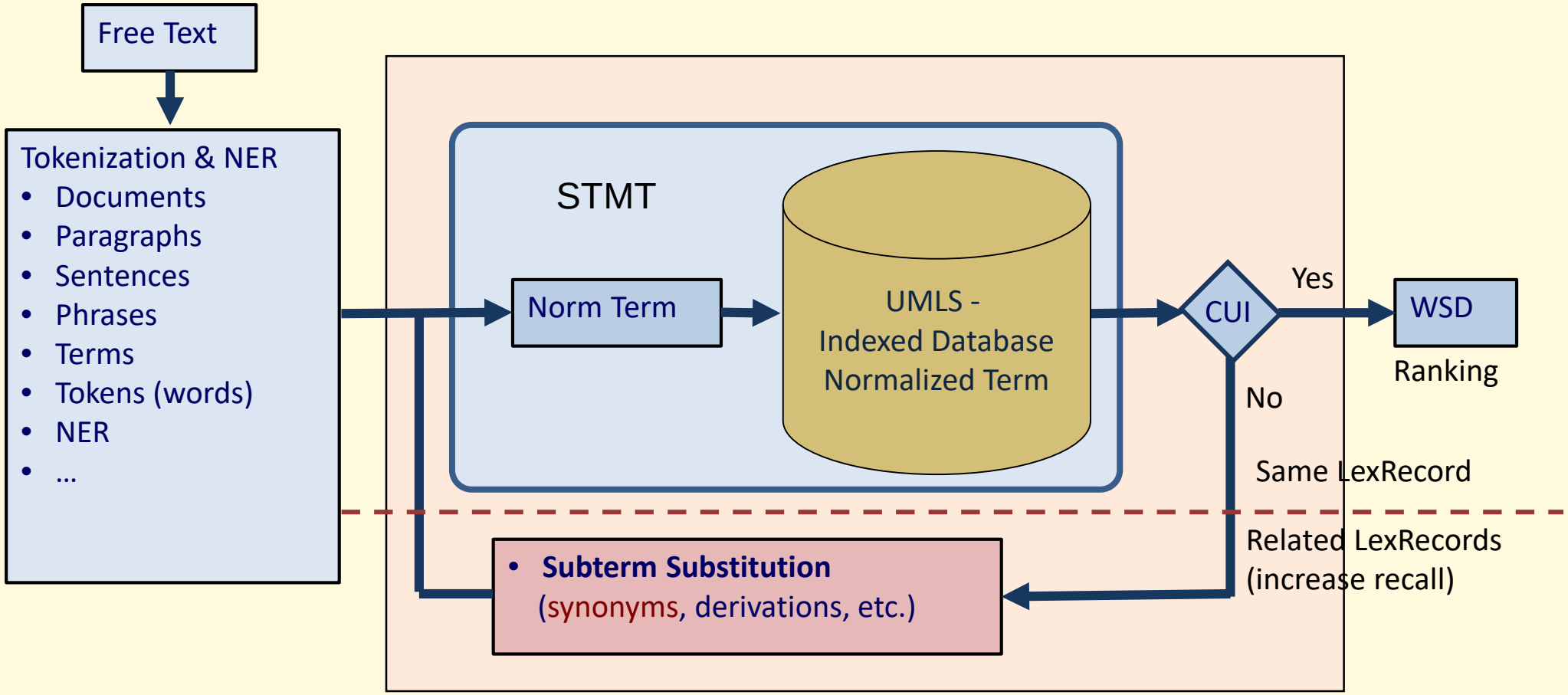
* PMID: 1118604, 1165396, ..

Multiple Substitutions



* VA14760, HA480.80, ..

Real-time Model



4. Applications - CSpell

- CSpell – Spell Checker for Consumer Language
  - 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: Dr. Kilicoglu's Ensemble method outperformed 30+% (2015)

☺

Examples

- Input Text: My mom was dianosed early on set deminita 3years ago.
- Correction:

Focus Token	Current Corrected Text	Notes
My	My	
mom	My mom	
was	My mom was	
dianosed	My mom was *diagnosed*	non-word spelling
early	My mom was *diagnosed* early	
on	My mom was *diagnosed* early on	
set	My mom was *diagnosed* early *onset*	real-word merge
deminita	My mom was *diagnosed* early *onset dementia*	non-word spelling
3years	My mom was *diagnosed* early *onset dementia 3 years*	ND split
ago.	My mom was *diagnosed* early *onset dementia 3 years* ago.	Lazy tokenization

Spelling Errors and Corrections

- Spelling Errors:
  - Non-word error, real-word errors, word boundary infraction, punctuation errors, informal expression, and combinations of the above.
- Spelling Corrections:
  - **Dictionary-based** vs. non-dictionary-based
  - **Spelling** (1-to-1), split and merge
  - **Isolated-word** vs. context-dependent
  - **Non-word** vs. real-word
  - Others (informal expression, Html/Xml tag, etc.)
  - Multiple corrections of the above

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
  - …

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

Detector → Candidates → Ranker → Corrector

Dictionary-based Correction Example

- Detector: not in the dictionary & not exceptions (digit, punctuation, email, URL, measurement, single character, etc.)
- Candidate Generator: looks and sounds alike, Church's reverse minimum edit distance
- Ranker: edit-distance, phonetic, overlap, word frequency, context, combination
- Corrector: replacement for spelling, flatmap for split, reconstruct text for merge

Input	Detector	Candidates	Ranker	corrector
diagnost	non-word	• diagnose • diagnosed • diagnostic • diagnosis • diagnoses • diagnoser • …	1) diagnosis 2) diagnosed 3) …	diagnosis

Correction Techniques

- Isolated-error corrections
  - Orthographic Similarity (looks and sounds alike)
    - Edit distance (Damerau and Levenshtein)
    - Phonetic algorithm
    - Leading/trailing character overlap
  - Word Frequency
- Context-dependent corrections
  - Semantic Distance
  - N-gram based
  - Word Embedding
- Combined corrections
  - Noisy Channel
  - Ensemble Method
  - CSpell 2-stage ranking

Edit Distance (Token) Similarity

- Edit Distance:

  the minimum number of operations required to transform one string into the other

  - **D**eletion (0.95): testt -> test (delete t at 5)
  - **I**nsertion (0.95): tet -> test (insert s at 2)
  - **S**ubstitution (1.00): tast -> test (substitute e for a at 1)
  - **T**ransposition (0.90): tset -> test (transpose se to es at 1)

- Church's reverse minimum Edit distance + Dictionary (candidate generator)
- Example:

diagnost	Operation	Edit Distance	Dictionary	Candidate
diagnosi	substitution	1	No	No
disagnosis	insertion	2	Yes	Yes
diagnose	substitution	1	Yes	Yes

Phonetic Similarity

- Phonetic Algorithm:

  Algorithm converts strings to codes for indexing by pronunciation (most in English)
  - Refined Soundex
  - Metaphone
  - Caverphone 2
  - Double Metaphone
- Phonetic Similarity: phonetic code + edit distance
- Example:

	diagnost	diagnosis	Similarity
Refined Soundex	D60480306	D60480303	2
Metaphone	TNST	TNSS	1
Caverphone 2	TKNST11111	TKNSS11111	1
Double Metaphone	TKNST	TKNSS	1

Overlap Similarity

➢Overlap Similarity:

Calculate the number of matching characters at the beginning and the end of two strings.

- Similarity score is between 0.00 and 1.00
- Use this with other orthographic similarity to improve the precision

➢Example:

- minLength = 8 (diagnost)
- maxLength = 9 (diagnosis)
- leadOverlap = 7 (diagnos)
- trailOverlap = 0
- Overlap similarity = 7/8 = 0.7778

diagnost	diagnosis	0.7778

Word Frequency

- Word Frequency

  Words grouped by frequency (word count) of occurrence within some given text corpus
  - Word: unigram (lowercase)
  - Score range: 0.0 ~ 1.0 (normalized by the max. frequency)
  - Use with other knowledge sources

- Example:

Word	Word Count
diagnosis	9083
diagnosed	1948
diagnose	2115
diagnostic	769
diagnoses	203
diagnoser	0

Context-dependent Corrections

- Isolated-word Corrections

Technique	Input	Correction
Orthographic	diagnost	diagnose
Word frequency	diagnost	diagnosis
Noisy Channel	diagnost	diagnosis

- Context-dependent Corrections

Input	Correction
the diagnost	the diagnosis
was diagnost	was diagnosed

Context Information

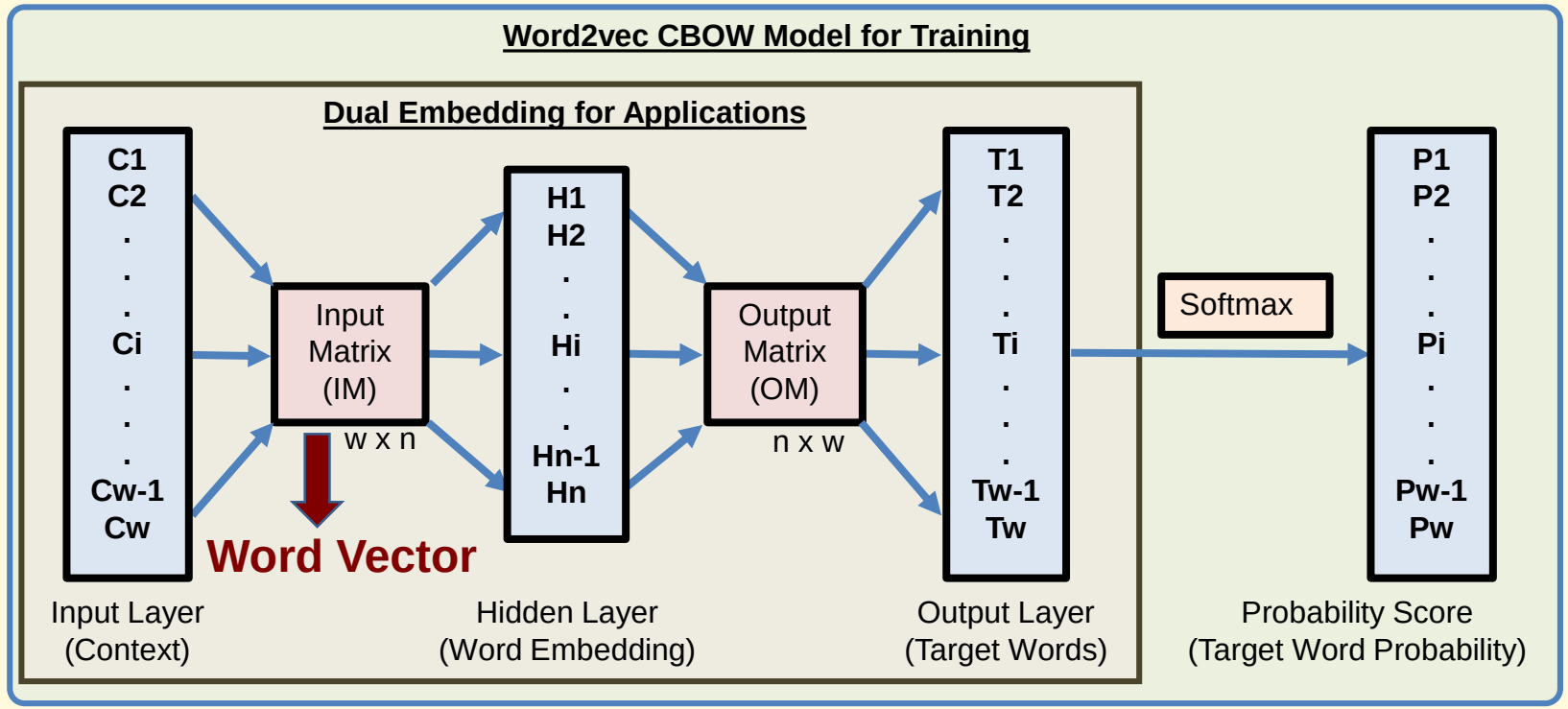
- N-gram based model
  - Bigrams and trigrams with frequency
- Word embedding
  - Word vector
    - Represent the meaning of a word in some abstract way
    - words that have the same meaning have a similar representation
    - individual words are represented as real-valued vectors in a predefined (dense) vector space
    - Unsupervised model captures certain **syntactical and semantic regularities**
  - Word2vec
  - GolVe (Global Vectors)

Word2vec: Word Embedding

- Word Vectors (word2vec, Tomas Mikolov, 2013):
  - Each word has an associated vector
  - Represent the meaning of a word in some abstract way
  - Capture meaningful syntactic and semantic regularities
- Examples (Prediction by similarity)
  - 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

Context Score – CBOW

- Context Score, Word2vec, CBOW model:
  - Duel embedding: use both input matrix (context) and output matrix (target) to calculate context scores, ~10% improvement in F1



Context-dependent Corrections

➢ Non-word corrections

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**

➢ Real-word Corrections

Input Text	Correction
smell size	small size
smell amount	small amount
smell intestine	small intestine

Input Text	Correction
foul small	foul smell
small an order	smell an order
taste and small	taste and smell

Word Boundary Infractions (Split & Merge)

- Non-word Corrections

Input Text	Split Correction
viceversa	vice versa
knowabout	know about
aftercreemail	aftercare email

Input Text	Merge Correction
anyt ime	anytime
dur ing	during
stiff n ess	stiffness

- Real-word Corrections

Input Text	Split Correction
along	a long
because	be cause
another	a not her

Input Text	Merge Correction
on set	onset
non drug	nondrug
some what	somewhat

Combined Techniques

➢ Noisy Channel

- Use orthographic score as the error model score
- Use frequency score as the language model score
- Noisy Channel Score = Language model score * Error model score

➢ Ensemble method

- Ensemble Score = 0.15 * Context score + 0.25 * Frequency Score + 0.2 * (Orthographic score)
- Orthographic score = Edit distance similarity score + Phonetic similarity score + overlap similarity score

➢ CSpell 2 Stage Ranking

- Stage 1 (regular season): use orthographic similarity scores to exclude irrelevant non-word candidates
- Stage 2 (playoff): used chain comparators to rank the selected candidates by the context score, then the noisy channel score in a sequential order. Rank in stage-1 is disregarded.

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:

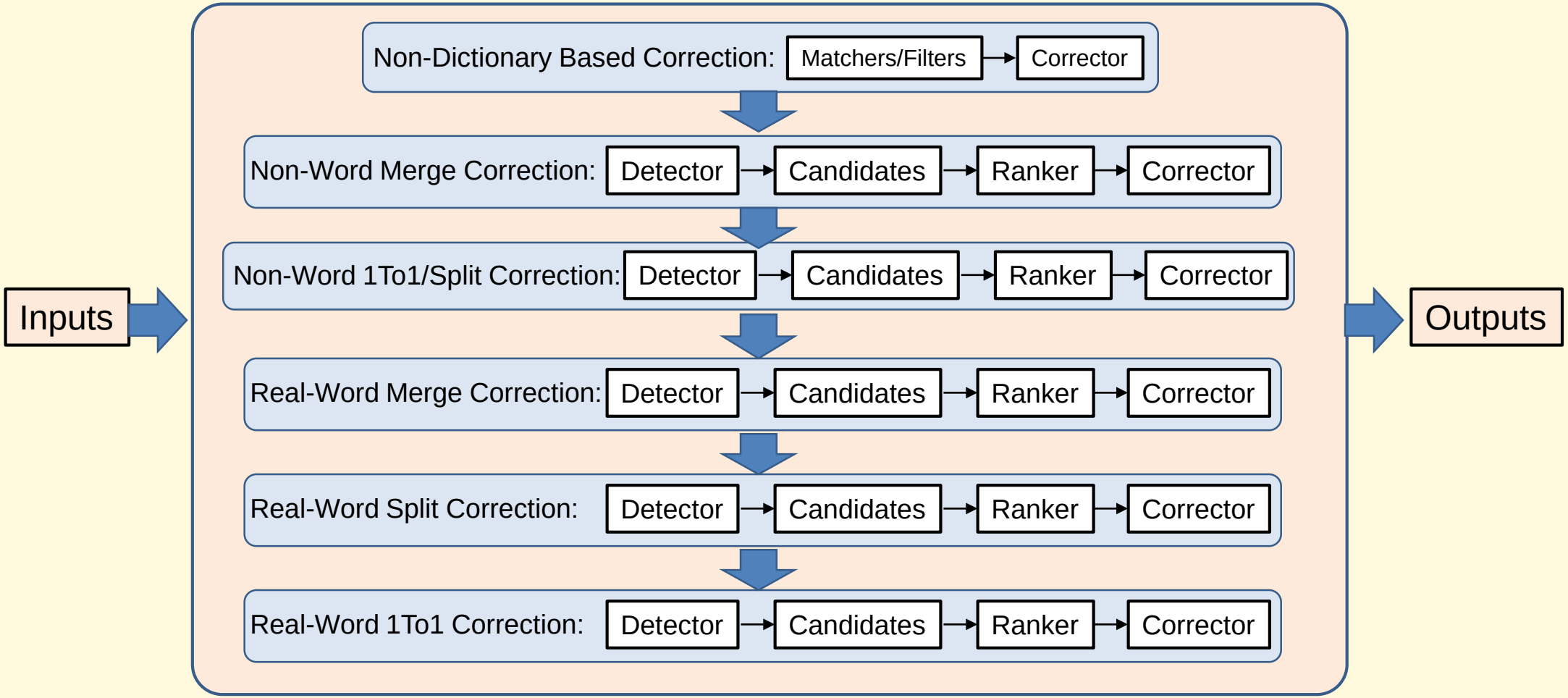
Handler Type	Input Text	Output Correction
Xml/Html	"germs"	”germs”
Informal Exception	pls	please

ND Splitters

- Algorithm: Detector → Corrector
  - 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.

Pipeline for Multiple Corrections



Examples – Errors and Corrections

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.

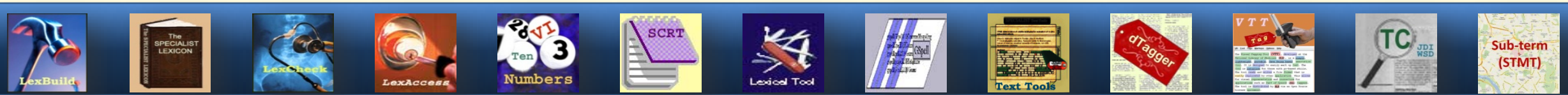
shunt from 2007

ND-Split, NW-Split

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

antidepressants

NW-1To1, RW-Merge



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

Evaluation: Training Set & Test Set

	Training Set *	Test Set**
Questions	471	224
Tokens	24,837	16,707
Tags	1,008	1,946
Error rate	0.04	0.07

* 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).

Dictionary

- The Lexicon-based dictionary has the best F1 score

Dictionary	Size	Precision	Recall	F1
Jazzy	159,345	0.2085	0.6835	0.3195
Ensemble	450,536	0.7139	0.7610	0.7367
MEDLINE	496,387	0.7506	0.7468	0.7487
CSpell (Lexicon)	597,649	0.8407	0.7842	0.8115

← Best F1

Corpus (for Word Frequency)

- Corpus:
  - Consumer Health Corpus: smaller relevant corpus
  - The MEDLINE N-gram Set (unigrams): general large collections

Corpus	Size	Precision	Recall	F1
MEDLINE	496,388	0.8085	0.7907	0.7995
Consumer Health Corpus	109,818	0.8407	0.7842	0.8115

← Better F1

Dual Embedding

- Single embedding:
  - use the [IM] as word vectors
  - Use the similarity between context and candidates by calculating context scores as cosine similarity with word vectors
- Double embedding:
  - Use CBOW model to predict a word from a given context by calculating context scores with both [IM] and [OM]

Embedding	Matrixes	Precision	Recall	F1
Single	IM	0.5887	0.5917	0.5902
Dual	IM & OM	0.8035	0.5917	**0.6815**

← Better precision and F1

CSpell 2-stage Ranking System

- 2 stage is better than 1 stage

The idea of 2 stage ranking is similar to the regular season/playoff in sports championship

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

Context Window Size

- Context window size should be the same as in the training model

Context Radius	Precision	Recall	F1
1	0.8380	0.7817	0.8088
2	**0.8407**	**0.7842**	**0.8115**
3	0.8366	0.7804	0.8075
4	0.8352	0.7791	0.8061
5	0.8352	0.7791	0.8061
6	0.8296	0.7739	0.8008
7	0.8310	0.7752	0.8021
8	0.8310	0.7752	0.8021
9	0.8310	0.7752	0.8021
10	0.8296	0.7739	0.8008
25	0.8283	0.7726	0.7995
50	0.8283	0.7726	0.7995
100	0.8283	0.7726	0.7995

← Best F1

Results: Training Set Detection

- Non-word Detection:

Method	Precision	Recall	F1
ESpell	33.47%	51.03%	0.4043
Jazzy	82.44%	41.86%	0.5553
Ensemble	79.39%	84.63%	0.8193
CSpell	92.38%	86.18%	0.8917

- Real-word Included Detection:

Method	Precision	Recall	F1
ESpell	34.75%	42.53%	0.3825
Jazzy	84.99%	34.65%	0.4923
Ensemble	80.78%	60.17%	0.6897
CSpell	92.89%	71.78%	0.8098

Results: Training Set Correction

- Non-word Correction:

Method	Precision	Recall	F1
ESpell	20.08%	30.62%	0.2426
Jazzy	47.58%	24.16%	0.3205
Ensemble	66.91%	71.32%	0.6904
CSpell	80.47%	78.42%	0.8115

- Real-word Included Correction:

Method	Precision	Recall	F1
ESpell	20.76%	25.41%	0.2285
Jazzy	48.60%	19.81%	0.2815
Ensemble	72.01%	53.63%	0.6147
CSpell	84.16%	65.04%	0.7338

Results: Test Set Detection

- Non-word Detection:

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

- Real-word Included Detection:

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

Results: Test Set Correction

- Non-word Correction:

Method	Precision	Recall	F1	Time
Ensemble	61.90%	61.40%	0.6165	< 1 hr.
CSpell	76.60%	76.28%	0.7644	< 1 min.

- Real-word Included Correction:

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

Questions

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

☺