Remove Parenthesis Plural Forms of (s), (es), and (ies)

By:

Chris Lu

Guy Divita

Allen Browne

Date: 12.13.2004

Table of Content

- Background
- Problems
- Objective
- Methods
- Results
- Future work

Background

Norm:

- is the most common used program in Lvg
- is used to create the normalized string and word indexes to UMLS Metathesaurus
- is used to access those indexes in UMLS Metathesaurus
- includes 10 lvg flows (2004)

Background – Cont.

Norm:

1. Remove genitives
2. Replace punctuations with space
3. Remove stop words
4. Strip diacritic
5. Split ligatures
6. Lowercase
7. Uninflect each words
8. Retrieve citation
9. Word sort
10. Retrieve Unicode symbol

Background – Cont.

Plural forms with parenthesis

- (s):
  - Accessory finger(s)
  - Addiction, drug(s)
  - Burn of wrist(s) and hand(s)
- (es):
  - Abdomen CT Adrenal Mass(es) Bilateral
  - Provide picture of fetus(es), as appropriate
  - sequelae of; injury, nerve, roots and plexus(es), spinal
- (ies):
  - Donor pneumonectomy(ies) with preparation and maintenance pf allograft (cadaver)
  - Orthotic(s) fitting and training, upper extremity(ies), lower extremity(ies), and/or trunk, each 15 minutes

Problems

- No flow in lvg to handle this issue
- Can we just simply remove (s), (es), (ies) ?
    - to get the uninflected form
    - without change the word
- (es), (ies): no problem
- (s): ?

Challenge

How about:

- 1-N-(s)-4-amino-2-hydroxybutyryl-3'4'-deoxyneamine
- 9(s)-erythromycylamine
- anatoxin-b(s)
- Ap(s)pCHClpp(s)A
- Bacillus phage rho11(s)
- Cbz-AAPhepsi((s)-CH(OH)CH2)GlyVV-OMe
- EAV G(s) glycoprotein
- G(s), alpha Subunit
- Histone H1(s)
- J(s)(b) ANTIBODY
- N(alpha)-benzoylarginineamide monohydrochloride, (s)-isomer
- natoxin-a(s)
- Salmonella II 6,7:(g),m,(s),t:1,5
- (s)-(+)-citreofuran
- su(s) protein, Drosophila
- XLalpha(s) protein
- [X]O spontn disrptn/lig(s)knee
- O spontn disrptn/lig(s)knee

Challenge – Cont.

- Not to remove (s) in chemical, Protein, Gene, mathematics, etc.
- Sometimes, (s) should be replaced by a space instead of removal

Objective

- Remove parenthesis plural forms of (s), (es), (ies)
- Do not remove (s) in chemical, protein, gene, etc..
- Replace (s) with a space appropriately
- Fast performance
- High precision

Scope

- UMLS Metathesaurus: 2.8 M terms
- Lexicon: 0.8 M inflected terms
- Total: 3.6 M terms
- Terms with (s), (es), (ies) patterns: ~ 2800

Methods - Pattern Observation

- 1-N-(s)-4-amino-2-hydroxybutyryl-3'4'-deoxyneamine
- 9(s)-erythromycylamine
- anatoxin-b(s)
- Ap(s)pCHClpp(s)A
- Bacillus phage rho11(s)
- Cbz-AAPhepsi((s)-CH(OH)CH2)GlyVV-OMe
- EAV G(s) glycoprotein
- G(s), alpha Subunit
- Histone H1(s)
- J(s)(b) ANTIBODY
- N(alpha)-benzoylarginineamide monohydrochloride, (s)-isomer
- natoxin-a(s)
- Salmonella II 6,7:(g),m,(s),t:1,5
- (s)-(+)-citreofuran
- su(s) protein, Drosophila
- XLalpha(s) protein

Pattern Observation – (1)

- 1-N-(s)-4-amino-2-hydroxybutyryl-3'4'-deoxyneamine
- 9(s)-erythromycylamine
- anatoxin-b(s)
- Ap(s)pCHClpp(s)A
- Bacillus phage rho11(s)
- Cbz-AAPhepsi((s)-CH(OH)CH2)GlyVV-OMe
- EAV G(s) glycoprotein
- G(s), alpha Subunit
- Histone H1(s)
- J(s)(b) ANTIBODY
- N(alpha)-benzoylargininеamide monohydrochloride, (s)-isomer
- natoxin-a(s)
- Salmonella II 6,7:(g),m,(s),t:1,5
- (s)-(+)-citreofuran
- su(s) protein, Drosophila
- XLalpha(s) protein

Pattern Observation – (1)

Sample Term	Word Size	Distance
9(s)-erythromycylamine	1	1
Ap(s)pCHClpp(s)A	2	1
EAV G(s) glycoprotein	1	1
G(s), alpha Subunit	1	1
Histone H1(s)	2	1
J(s)(b) ANTIBODY	1	1
N(alpha)-benzoylarginineamide monohydrochloride, (s)-isomer	0	1
(s)-(+)-citreofuran	0	1
su(s) protein, Drosophila	2	1

- The size of the word in front of (s) must be less than/equal to 2

Pattern Observation – (2)

- 1-N-(s)-4-amino-2-hydroxybutyryl-3'4'-deoxyneamine
- 9(s)-erythromycylamine
- anatoxin-b(s)
- Ap(s)pCHClpp(s)A
- Bacillus phage rho11(s)
- Cbz-AAPhepsi((s)-CH(OH)CH2)GlyVV-OMe
- EAV G(s) glycoprotein
- G(s), alpha Subunit
- Histone H1(s)
- J(s)(b) ANTIBODY
- N(alpha)-benzoylargininеamide monohydrochloride, (s)-isomer
- natoxin-a(s)
- Salmonella II 6,7:(g),m,(s),t:1,5
- (s)-(+)-citreofuran
- su(s) protein, Drosophila
- XLalpha(s) protein

Pattern Observation – (2)

Sample Term	Character	Distance
9(s)-erythromycylamine	Arabic number 9	1
Bacillus phage rho11(s)	Arabic number 1	1
Histone H1(s)	Arabic number 1	1

- The character in front of (s) is an Arabic number

Pattern Observation – (3)

- 1-N-(s)-4-amino-2-hydroxybutyryl-3'4'-deoxyneamine
- 9(s)-erythromycylamine
- anatoxin-b(s)
- Ap(s)pCHClpp(s)A
- Bacillus phage rho11(s)
- Cbz-AAPhepsi((s)-CH(OH)CH2)GlyVV-OMe
- EAV G(s) glycoprotein
- G(s), alpha Subunit
- Histone H1(s)
- J(s)(b) ANTIBODY
- N(alpha)-benzoylargininеamide monohydrochloride, (s)-isomer
- natoxin-a(s)
- Salmonella II 6,7:(g),m,(s),t:1,5
- (s)-(+)-citreofuran
- su(s) protein, Drosophila
- XLalpha(s) protein

Pattern Observation – (3)

Sample Term	Character	Distance
1-N-(s)-4-amino-2-hydroxybutyryl-3'4'-deoxyneamine	Punctuation -	1
anatoxin-b(s)	Punctuation -	2
Cbz-AAPhepsi((s)-CH(OH)CH2)GlyVV-OMe	Punctuation (	1
natoxin-a(s)	Punctuation -	2
Salmonella II 6,7:(g),m,(s),t:1,5	Punctuation ,	1

- Punctuation is in front of (s) within distance 1 or 2

Pattern Observation – (4)

- 1-N-(s)-4-amino-2-hydroxybutyryl-3'4'-deoxyneamine
- 9(s)-erythromycylamine
- anatoxin-b(s)
- Ap(s)pCHClpp(s)A
- Bacillus phage rho11(s)
- Cbz-AAPhepsi((s)-CH(OH)CH2)GlyVV-OMe
- EAV G(s) glycoprotein
- G(s), alpha Subunit
- Histone H1(s)
- J(s)(b) ANTIBODY
- N(alpha)-benzoylarginineamide monohydrochloride, (s)-isomer
- natoxin-a(s)
- Salmonella II 6,7:(g),m,(s),t:1,5
- (s)-(+)-citreofuran
- su(s) protein, Drosophila
- XLalpha(s) protein

Pattern Observation – (4)

Sample Term	Pattern	Distance
Ap(s)pCHClpp(s)A	pp	1
XLalpha(s) protein	alpha	1

- The word in front of (s) ends with:
  - pp
  - alpha

Pattern Observation – (5)

Sample Term	Pattern	Distance
[X]O spontn disrptn/lig(s)knee	Followed by a word	1
O spontn disrptn/lig(s)knee	Followed by a word	1

- (s) followed with an English word
- An English word begins with a letter

➢ if (s) followed with a letter, replace (s) with a space

- Exceptions:
    - Ap(s)pCHClpp(s)A
    - G(s)alpha

Implementation – Wild Cards

Wild Card Definition:

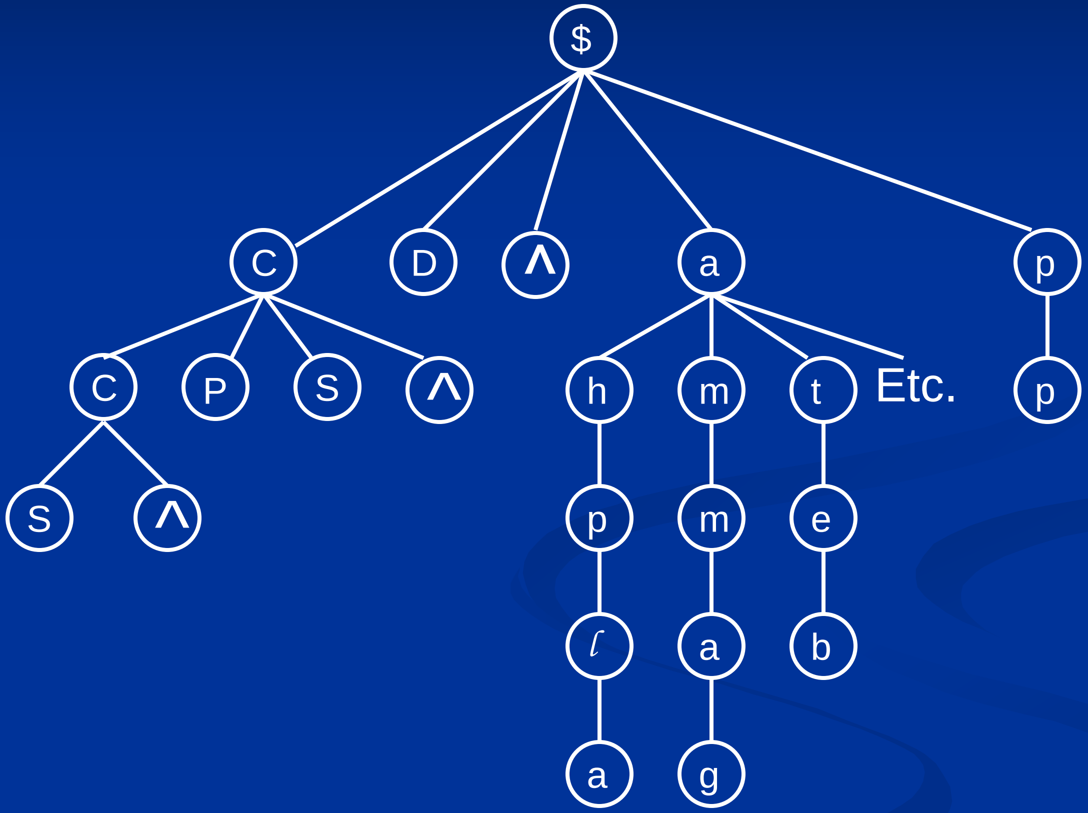
- **^**: start, starting mark of the term
- **$**: end, ending mark of the term right before (s)
- **C**: any character
- **D**: any digit, [0-9]
- **L** any letter, [a-z]
- **P**: punctuation: [- (,]
- **S**: space: [ ]

Implementation – Rule Representations

Pattern	Sample Term	Rule
1	(s)-(+)-citreofuran	^$
1	J(s)(b) ANTIBODY	^C$
1	EAV G(s) glycoprotein	SC$
1	su(s) protein, Drosophila	^CC$
1	Histone H1(s)	SCC$
2	9(s)-erythromycylamine	D$
3	Salmonella II 6,7:(g),m,(s),t:1,5	P$
3	natoxin-a(s)	PC$
4	Ap(s)pCHClpp(s)A	pp$
4	XLalpha(s) protein	alpha$
..	…	…

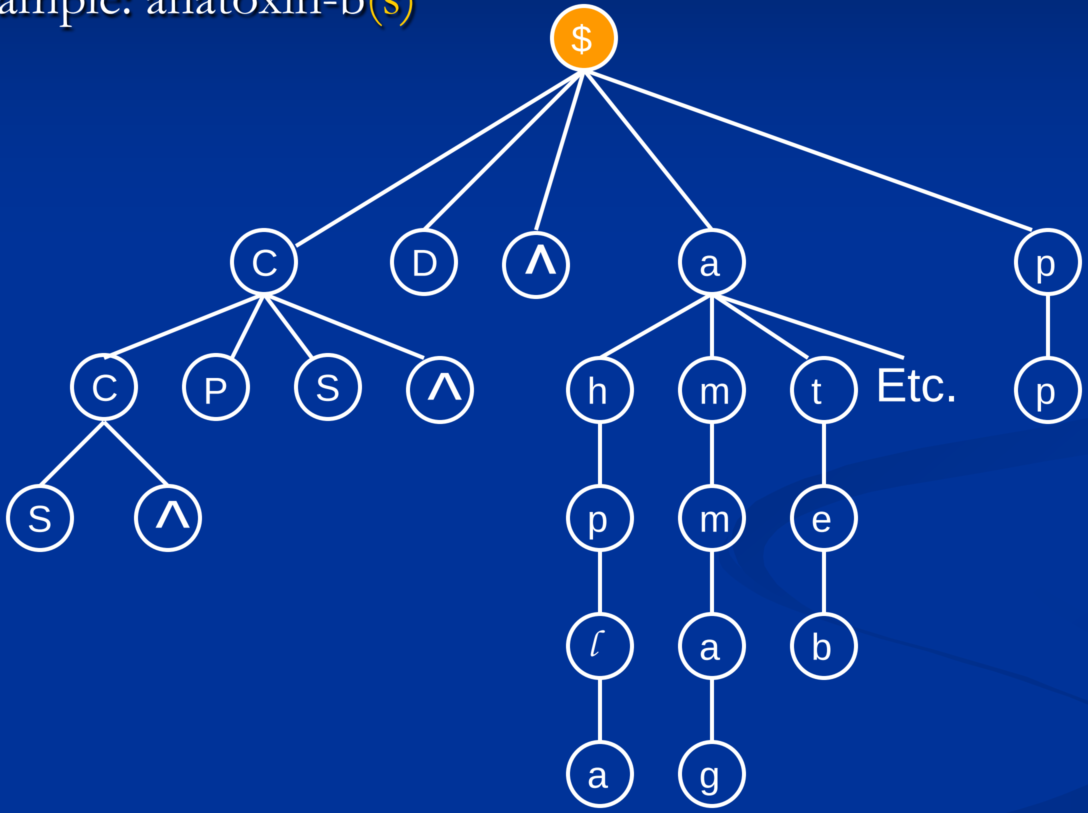
Implementation – Reversed Trie Tree

		Rule
		^$
		^C$
		SC$
		^CC$
		SCC$
		D$
		P$
		PC$
		pp$
		alpha$
		…



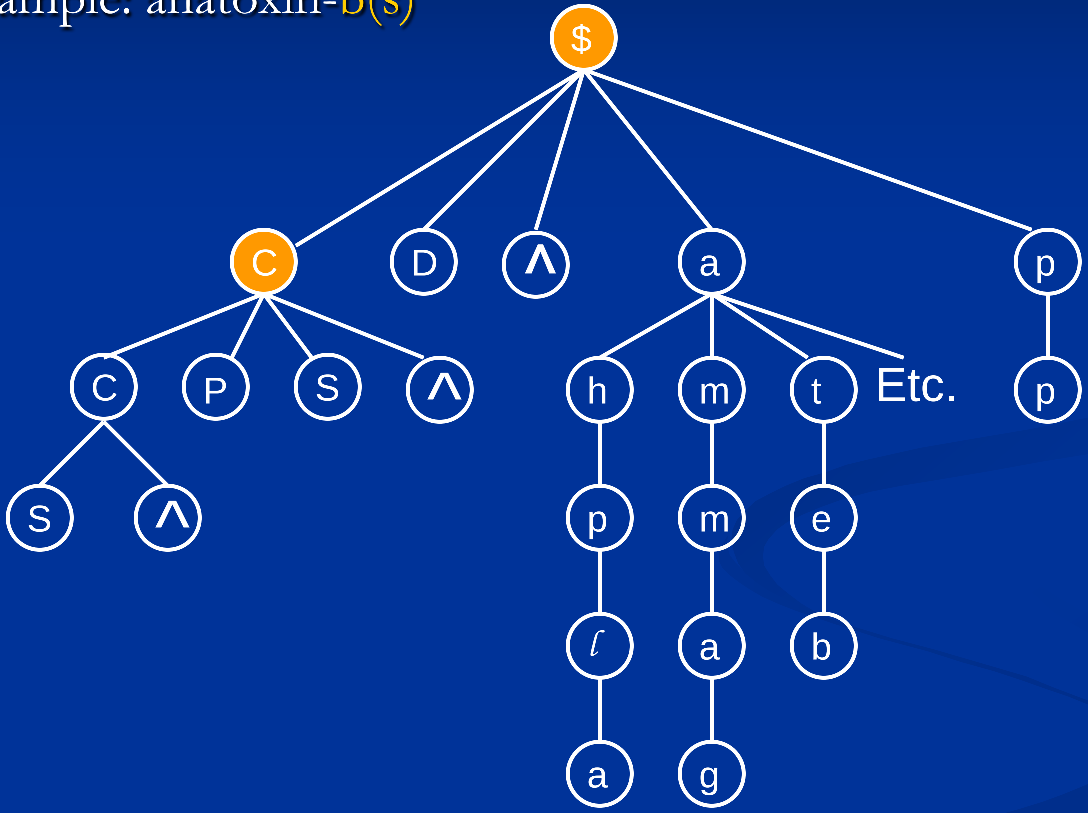
Implementation – Reversed Trie Tree

- Example: anatoxin-b(s)



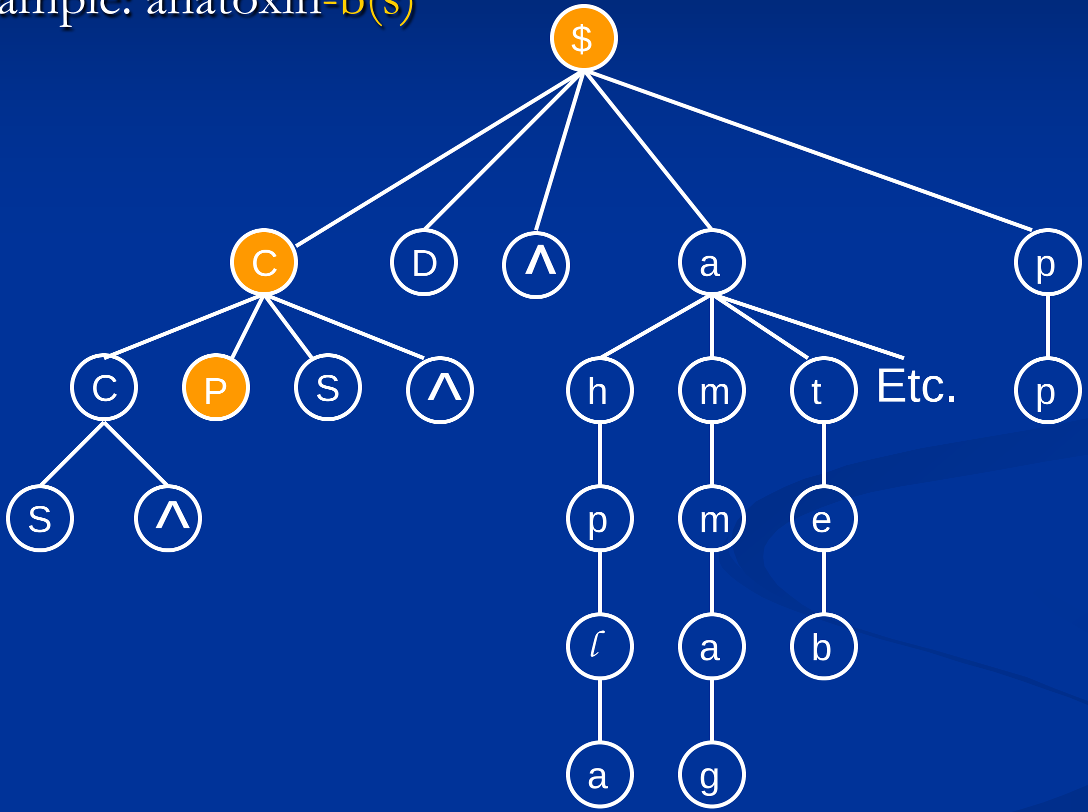
Implementation – Reversed Trie Tree

- Example: anatoxin-b(s)

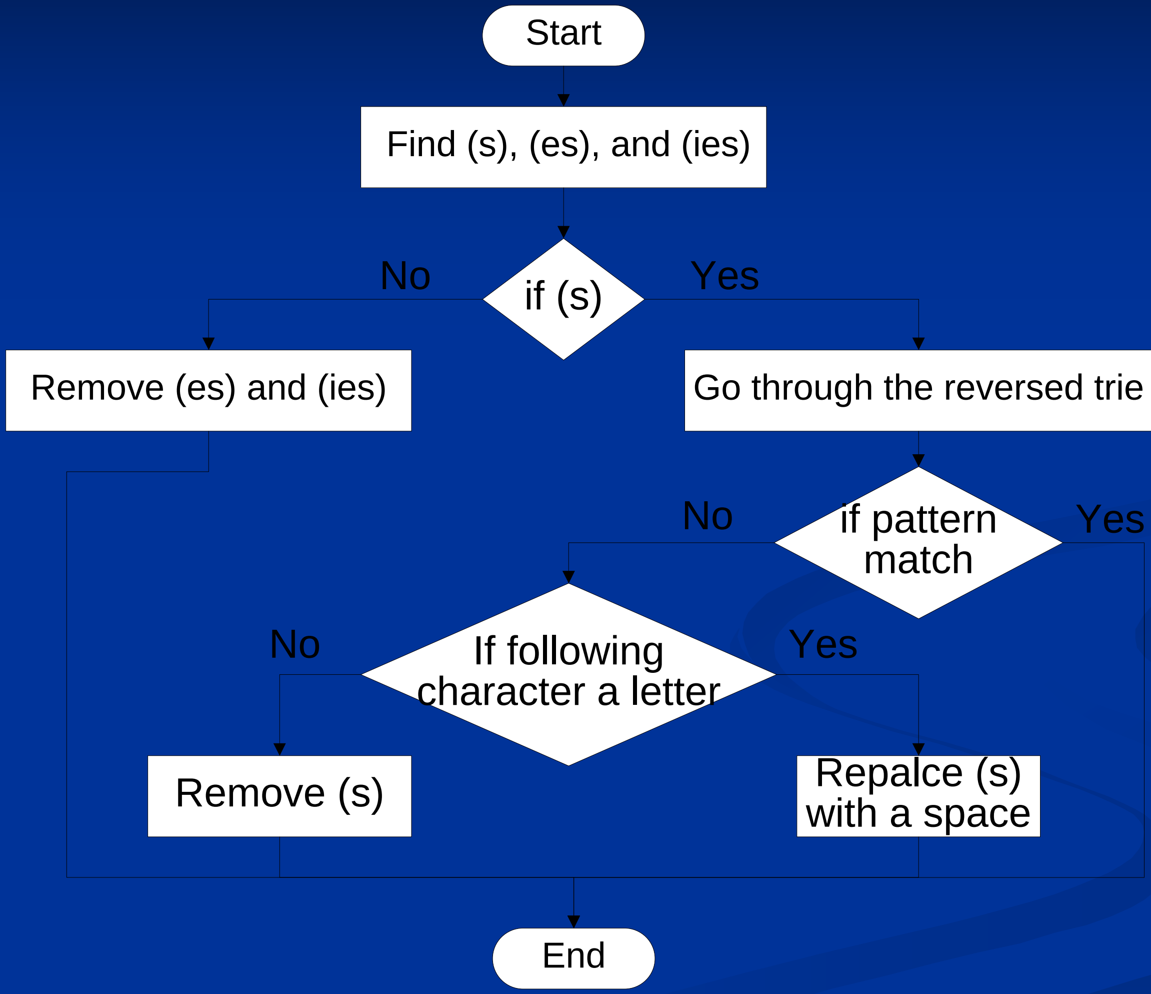


Implementation – Reversed Trie Tree

- Example: anatoxin-b(s)



Implementation – Algorithm Flow



Results

- Remove (s) properly
- Remove (es) properly
- Remove (ies) properly
- Replace (s) with space properly

- A fast, precise, and expandable system

Future Work

- More testing cases, update more rules
- Implement this feature to both Norm and LuiNorm
- Apply to (ing), (ed), (en)

Thank you !

- lu@nlm.nih.gov
- http://umlslex.nlm.nih.gov/lvg/2005