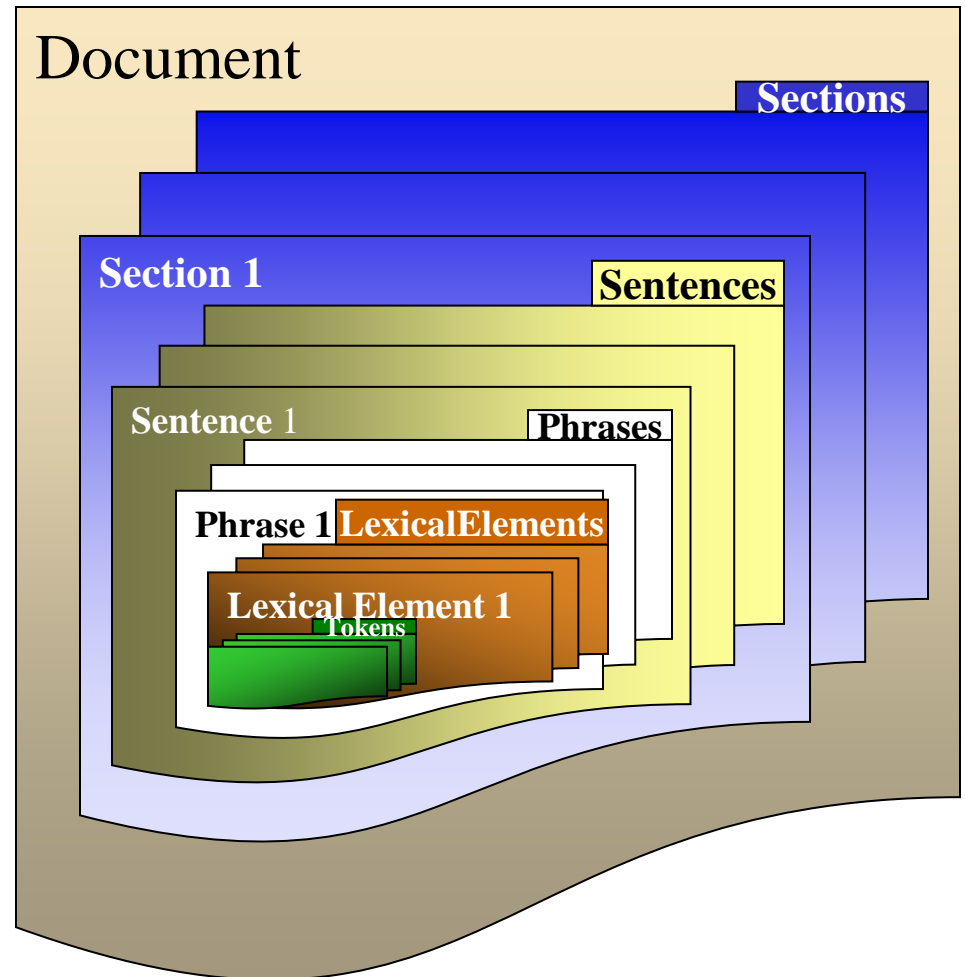


Additional NLS Tools

- NLS's Java NLP tools
- MMTx
- GSpell

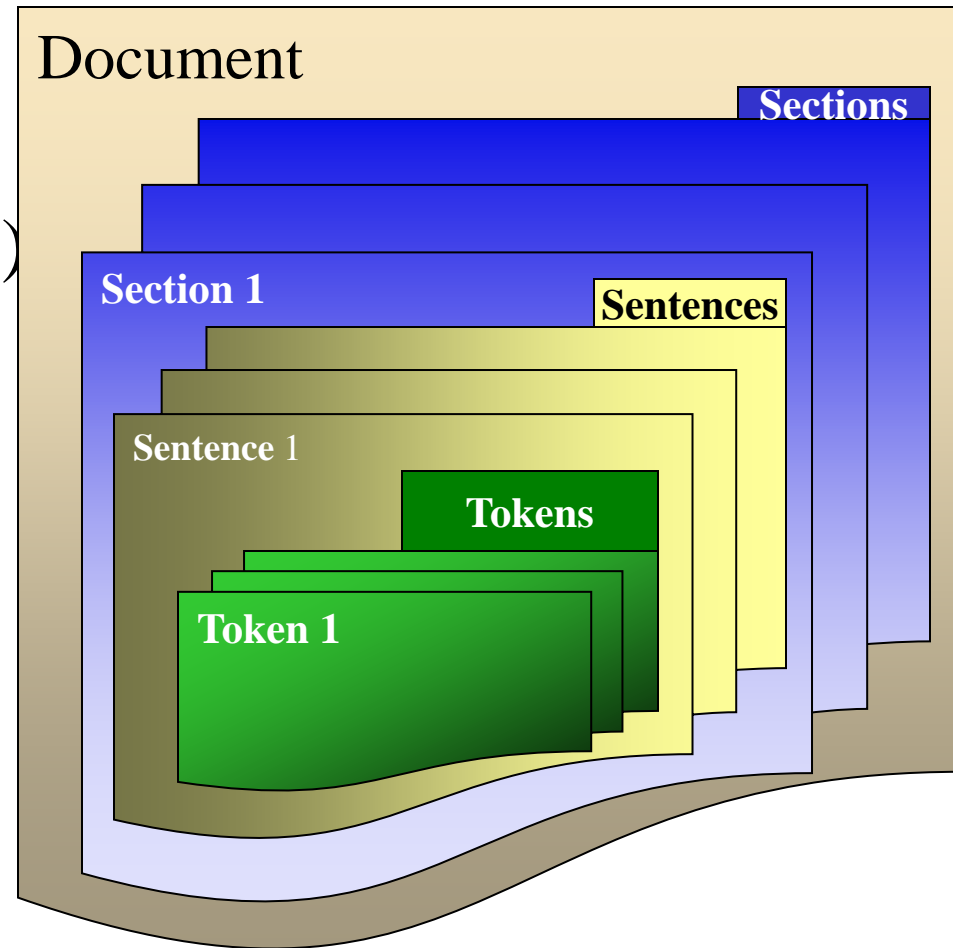
NLS Java NLP Tools

- Tokenizer
- Lexical Lookup
- NP Parser
 - Document Centric
 - Java Programs and API's



Java NLP Tools: Tokenizer

- Tokenizes text into
 - Sections (paragraphs)
 - Sentences
 - Tokens
- Can handle
 - FreeText
 - HTML
 - MedLINE Abstracts



Java NLP Tools: Tokenizer

Usage

tokenize.*[bat|sh]* [*Options*]

--fileName=*fileName*

--outputFileName=*fileName*

--inputType=[*freeText|HTML|medlineCitations*]

--sections

--sentences

--tokens

--pipedOutput

--indicate_citation_end

Java NLP Tools: Tokenizer

```
tokenize.bat --inputFile=5.txt --inputType=freeText --sentences --tokens  
--pipedOutput
```

Sentence|1|97|182|But those follow-up tests have been inconclusive, state
and federal officials said.

Token|16|97|99|0|0|**But**|||

Token|17|101|105|1|0|**those**|||

Token|18|108|113|2|0|**follow**|||

Token|19|114|114|2|0|-|||

Token|20|115|116|3|0|**up**|||

Token|21|118|122|4|0|**tests**|||

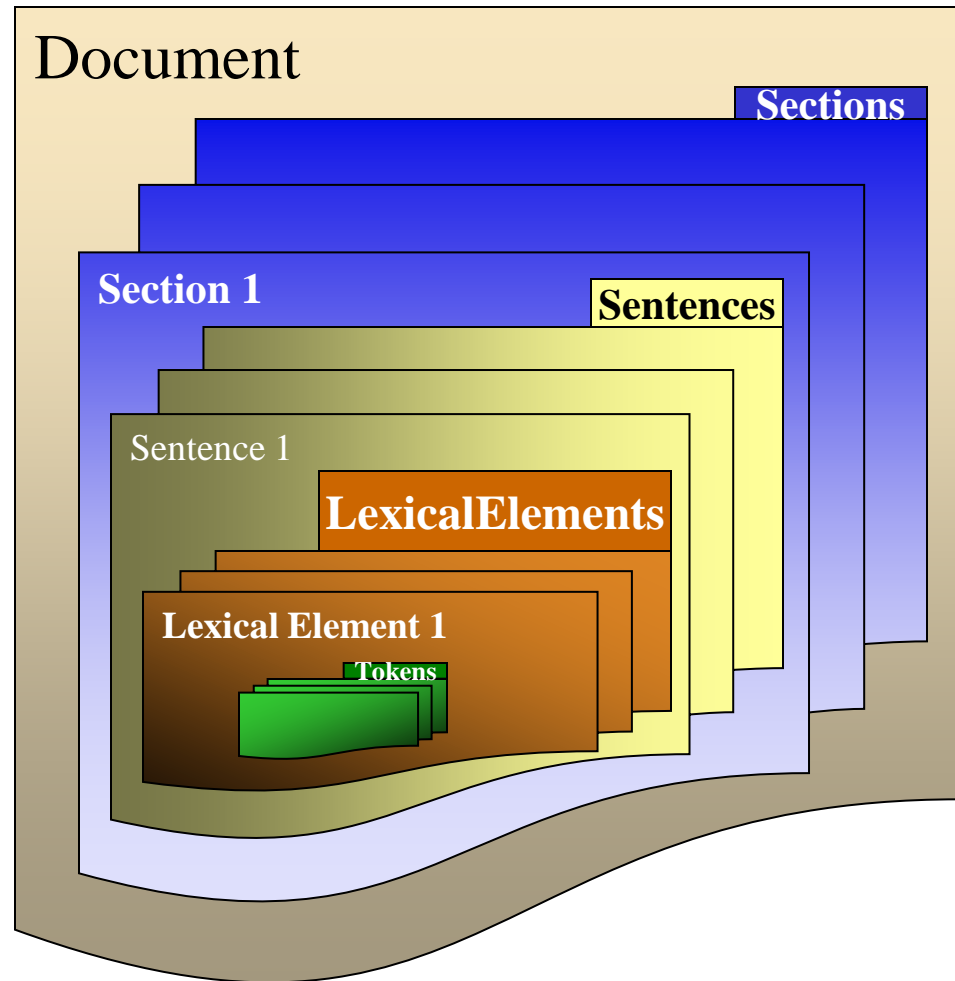
Token|22|124|127|5|0|**have**|||

Token|23|129|132|6|0|**been**|||

Token|24|134|145|7|0|**inconclusive**|||

NLP Tools: Lexical Lookup

- Chunks tokens into terms
 - From SPECIALIST Lexicon
 - From regular expressions



Java NLP Tools: Lexical Lookup

Usage

LexicalLookup.*[bat|sh]* [*Options*]

--fileName=*fileName*

--outputFileName=*fileName*

--inputType=[*freeText|HTML|medlineCitations*]

--sections

--sentences

--lexicalElements

--lexicalEntries

--tokens

--pipedOutput

Java NLP Tools: Lexical Lookup

```
LexicalLookup.bat --inputFile=5.txt --inputType=freeText  
--lexicalElements --lexicalEntries --pipedOutput
```

Lexical Element|17|LEXICON|prep|**But**|97|99

LexicalEntry|but|conj|base|E0014465

LexicalEntry|but|prep|base|E0014464

Lexical Element|18|LEXICON|det|**those**|101|105

LexicalEntry|those|det|plural|E0060728

LexicalEntry|those|pron|base|E0060729

Lexical Element|20|LEXICON|adj|**follow-up**|108|116

LexicalEntry|follow-up|adj|base|E0028422

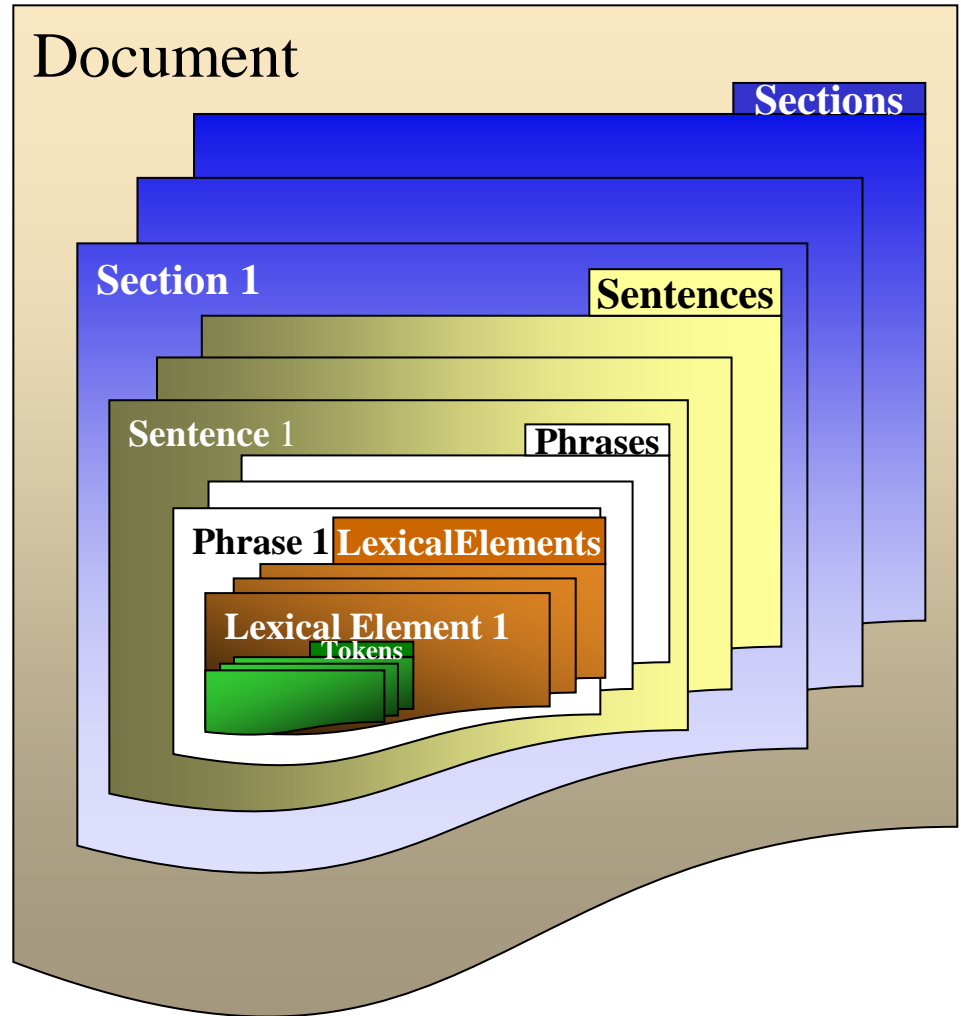
Lexical Element|23|LEXICON|noun|**tests**|118|122

LexicalEntry|tests|verb|pres3s|E0060349

LexicalEntry|tests|noun|plural|E0060348

NLP Tools: NpParser

- Chunks sentences into simple phrases



Java NLP Tools: NpParser

Usage

npParser.*[bat|sh]* *[Options]*

--**fileName**=*fileName*

--**outputFileName**=*fileName*

--**inputType**=[*freeText|HTML|medlineCitations*]

--**sections**

--**sentences**

--**phrases**|--**nps**|--**mincoMan**

--**lexicalElements**

--**lexicalEntries**

--**tokens**

--**pipedReader**

Java NLP Tools: NpParser

```
npParser.bat --inputFile=5.txt --inputType=freeText --phrases  
--pipedOutput
```

Phrase|0|0|10|**The company**|*company*

Phrase|1|12|14|**has**|

Phrase|2|16|24|**forwarded**|

Phrase|3|26|39|**some materials**|*materials*

Phrase|4|41|62|**to a state laboratory**|*state laboratory*

Phrase|5|64|74|**in Richmond**|*Richmond*

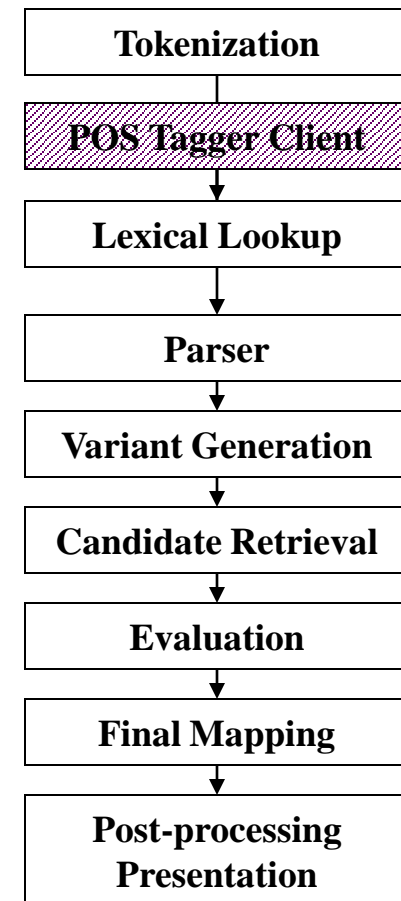
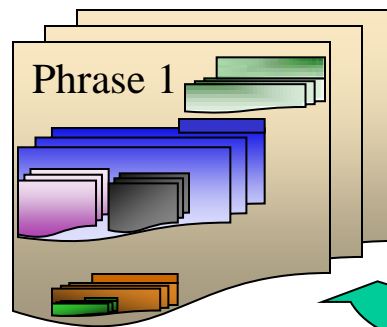
Phrase|6|76|86|**for further**|*further*

Phrase|7|88|94|**testing**|

MMT_x

MetaMapTechnology Transfer

- Maps text phrases to Metathesaurus concepts
- Java Implementation of MetaMap



MMTx

Usage

MMTx [*<options>*] [--**fileName**=*infile*]
[**outputFileName**=*outfile*]
--**strict_model**|--**moderate_model**|--**relaxed_model**
--**KYear**=*year*|--**mm_data_version**=*customName*
--**threshold**=*lowestScore*
--**truncate_candidates_mappings**
--**term_processing**|--**allow_overmatches**|--**allow_concept_gaps**
--**composite_phrases**
--**prefer_multiple_concepts**
--**fielded_output**

MMTx

```
MMTx --inputFile=5.txt --inputType=freeText
```

Processing 00000000.tx.3: ***One problem*** is caused by the *VecTest* itself, which uses a *dipstick* to measure the *presence* of a *protein* associated with the *parasite* that causes *malaria*.

Phrase: "**One problem**"

Meta Candidates (2)

861 Problem, NOS [Finding,Pathologic Function]

694 One [Quantitative Concept]

Meta Mapping (888)

694 **One** [Quantitative Concept]

861 **Problem, NOS** [Finding,Pathologic Function]

GSpell



“Gottlieb, I think I know why we’ve been receiving so few commissions.”

GSpell

- Spelling suggestion tool
- Pure Java application with Java API's
- Support for multi word dictionary entries

GSpell: Usage

Usage

GSpellFind.*[sh|bat]*

--dictionary=*NameOfDictionary*

[--inputFile=*Source*] **[--outputFile**=*target*]

[--truncate=*N*] **[--considerNCandidates**=*N*]

[--maxEditDistance=*N*]

[--fieldedText] **[--termField**=*X*] **[--correctField**=*Y*]

[--reportTime] **[--version]****[--help]**

GSpell: Example

Input Term	Suggestion	Edit Distance	Rank	Method	Message
------------	------------	---------------	------	--------	---------

anonomous|**anonymous**|1.0|0.8734230160180236|NGrams|
anonomous|**allonamous**|2.0|0.5819672267388108|NGrams|
anonomous|**autonomous**|2.0|0.5819672267388108|NGrams|
anonomous|**anadromous**|3.0|0.2958160192082048|NGrams|
anonomous|**analogous**|3.0|0.2958160192082048|NGrams|
anonomous|**anomalous**|3.0|0.2958160192082048|NGrams|
anonomous|**anonymously**|3.0|0.295816019208248|NGrams|
anonomous|**anonymes**|3.0|0.2958160192082048|Metaphone|
anonomous|**anonyms**|3.0|0.2958160192082048|Metaphone|
anonomous|**acoprous**|4.0|0.11470810702102521|NGrams|

GSpell: Indexing

Usage

GSpellIndex.*[sh|bat]*

--dictionary=*NameOfDictionary*

--inputFile=*SourceFile*

[--reportTime] **[--version]****[--help]**

- Format for the input file
 - One word per line

Downloadable Resources

- umlslex.nlm.nih.gov
 - Lvg
 - Java NLP Tools
 - GSpell
- mmtx.nlm.nih.gov



Lexical Tools for UMLS Developers

Allen C. Browne, Guy Divita, Chris Lu
Lister Hill National Center for Biomedical Communications
National Library of Medicine

Lexical Systems:

umlsLex.nlm.nih.gov

Email:

umlslex@nlm.nih.gov

Knowledge Source Server: **<http://umlsks.nlm.nih.gov>**

UMLS Information:

<http://umlsInfo.nlm.nih.gov>