Lexical Tools
Sub-Term Mapping Tools (STMT)

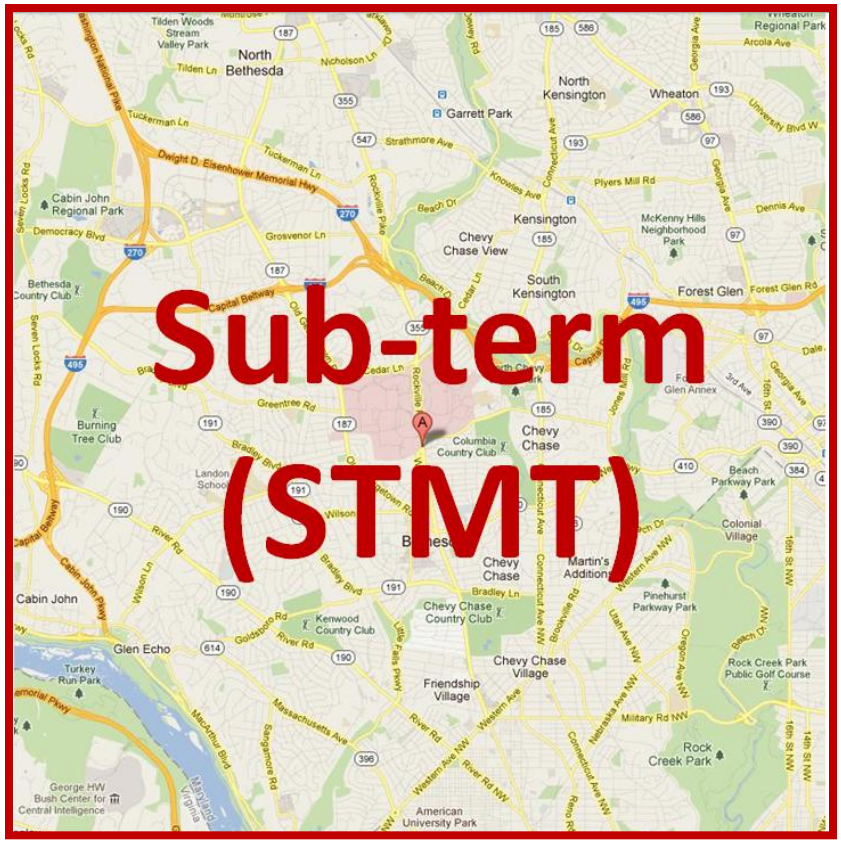
Dr. Chris J. Lu

The Lexical Systems Group
NLM. LHNCBC. CGSB

June, 2012

Table of Contents

STMT - Introduction



Sub-term

- A sub-term is a term that is a subset of another term
- A term is composed of words
- A word is delimited by space(s) or tab(s)
- A sub-term could be a word or a term

Example: Sub-term

- Term: Otitis externa, chronic infectious
- Sub-terms:
  - otitis
  - externa
  - chronic
  - infectious

  - otitis externa
  - chronic infectious
  - …
  - otitis externa chronic infectious

Example: Prefix

- Term: Otitis externa, chronic infectious
- Prefix: a sub-term starts with the same starting words of the term
- Sub-terms:
  - otitis => prefix
  - externa
  - chronic
  - infectious
  - otitis externa => prefix
  - chronic infectious
  - otitis externa chronic infectious => prefix

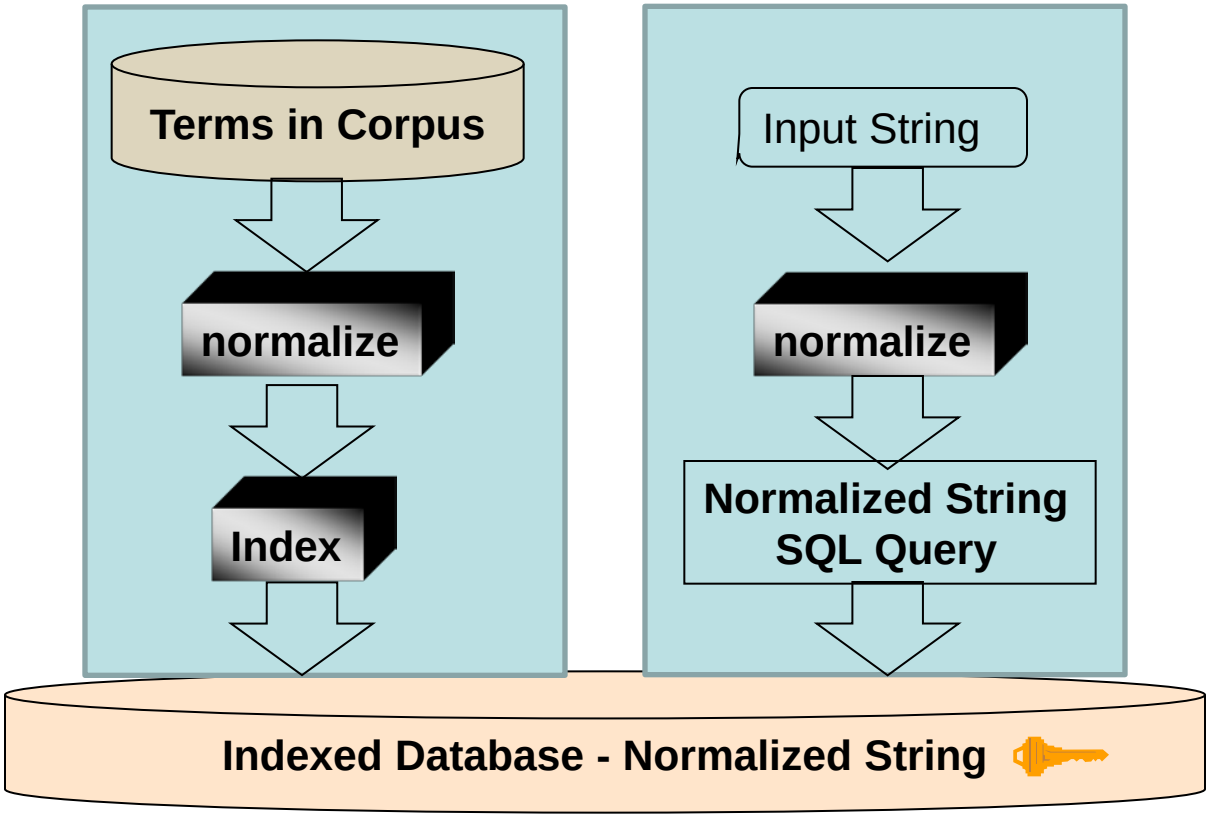
Example: Corpus

- Term: Otitis externa, chronic infectious
- Corpus: The SPECIALIST Lexicon
- Sub-terms:
  - otitis (E0044452)
  - otitis externa (0044453)
  - chronic (E0219527|E0016869)
  - infectious (E0044452)

  - ~~externa~~
  - ~~chronic infectious~~
  - ~~otitis externa chronic infectious~~

Example: Normalization

- Term: Otitis externa, chronic infectious
- Normalization: ignore case and punctuation
- Normalized Corpus: The SPECIALIST Lexicon



Example: Normalization

- Term: Otitis externa, chronic infectious
- Normalization for LexItem:
  ignore punctuation and case (lvg –f:o:l)
- PreProcess – normalized Lexicon terms in database
- Normalized term: otitis externa chronic infectious
- Performed normalized database query

STMT - Applications

Application – MetaMap*

- The Longest Prefix:
  - Find all prefixes and the longest prefix sub-term on an input term that is a lexical item (LexItem) in the SPECIALIST Lexicon
  - The longest prefix usually has the best match sense than other sub-terms
    - Example:
      University of California, Los Angeles
  - Ignore case and punctuation

* A.R. Aronson and F.M. Lang, “An Overview of MetaMap: historical perspective and recent advances”, JAMIA, 2010, Vol. 17, p.229-236

Lexical Mapping

- Algorithm:
  1. Normalization (LexItem Norm, -f:g:rs:o:l)
     - One to one
  2. Lexicon Mapping (LexAccess)
     - Table lookup (key - values)

- Example:
  University of California, Los Angeles

⇒ university of california los angeles (E0063261)

⇒

Original Term from Lexicon	Normalized Term (key)	EUI
…	…	…
University of California, Los Angeles	university of california los angeles	E0063261
University of California Los Angeles	university of california los angeles	E0063261
University of California-Los Angeles	university of california los angeles	E0063261
…	…	…

Lexical Mapping (cont.)

- University of California, Los Angeles
- University of California Los Angeles
- University of California-Los Angles
- university of california los angeles
- UNIVERSITY OF CALIFORNIA LOS ANGELES
- …

Normalization (LexItem Norm: -f:g:rs:o:l)

university of california los angeles

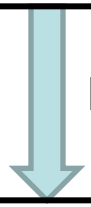
LexAccess (lvg –f:E)

Original Term from Lexicon	Normalized Term (key)	EUI
…	…	…
University of California, Los Angeles	university of california los angeles	E0063261
University of California Los Angeles	university of california los angeles	E0063261
University of California-Los Angeles	university of california los angeles	E0063261
…	…	…

E0063261

Longest Prefix

- University of California, Los Angeles is in USA …
- University of California Los Angeles is in USA …
- University of California-Los Angles is in USA …
- university of california los angeles is in USA …
- UNIVERSITY OF CALIFORNIA LOS ANGELES is in USA …
- …



Normalization (LexItem Norm: -f:g:rs:o:l)

university of california los angeles is in usa …

Lexical Mapping on each prefix sub-term

Longest Prefix (cont.)

- University of California, Los Angeles is in USA …
  - Normalize (LexItem Norm)

  => university of california los angeles is in usa …
  - Word by word
    1. university (E0063257)
    2. university of (none)
    3. university of california (E0702384)
    4. university of california los (none)
    5. university of california los angles (E0063261)
    6. university of california los angles is (none)
    7. …

Longest Prefix (cont.)

- university of california los angeles is in usa …
  I. Normalized
  II. Word by word
    1. university (E0063257)
    2. university of (none)
    3. university of california (E0702384)
    4. university of california los (none)
    5. university of california los angles (E0063261)
    6. university of california los angles is (none)
    7. …
- Issues:
  - Slow performance
  - Limited to 10 prefixes

Application – UMLS-CORE*

- Query expansion: to find concepts in 3 steps:
  1) Normalized lexical matches
     - MRXNS_ENG.RRF: normalized string to CUI
     - Lexical tools – Norm (-f:q0:g:rs:o:t:l:B:Ct:q7:q8:w)
  2) One synonymous word or phrase substitution
  3) Two synonymous word or phrase substitutions
- Collected synonyms:
  - acronyms: AB|ANTIBODY
  - British: aeroplane|airplane
  - Greco-latin: cat|feline
  - ECRI (devices): Airway|Tube
  - Lvg: zoic|animal life
  - …

* K.W. Fung, C. McDonald, S. Srinivasan, “The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions", JAMIA, 2010, Vol. 17, p.675-680

UMLS-Core: Example

- OTITIS EXTERNA, CHRONIC INFECTIOUS
  1) Normalized lexical matches
     - chronic externa infectious otitis
     - chronic externon infectious otitis
     - chronic externum infectious otitis

     $\Rightarrow$ No CUI found

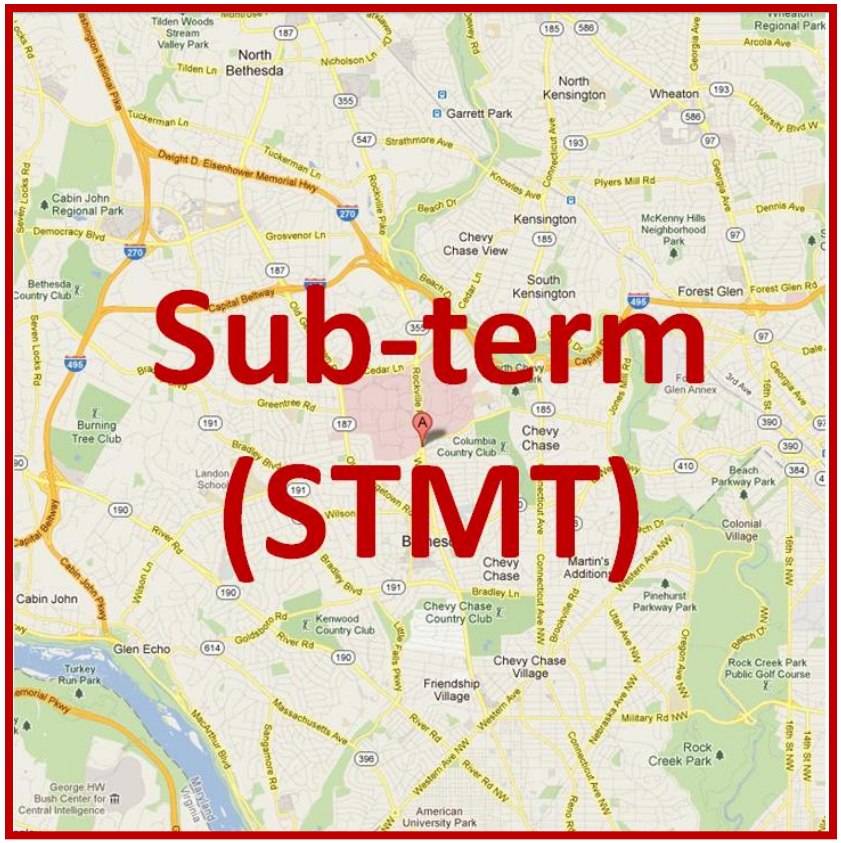
  2) One synonymous word or phrase substitution

     $\Rightarrow$ No CUI found

  3) Two synonymous word or phrase substitutions

     $\Rightarrow$ **...**

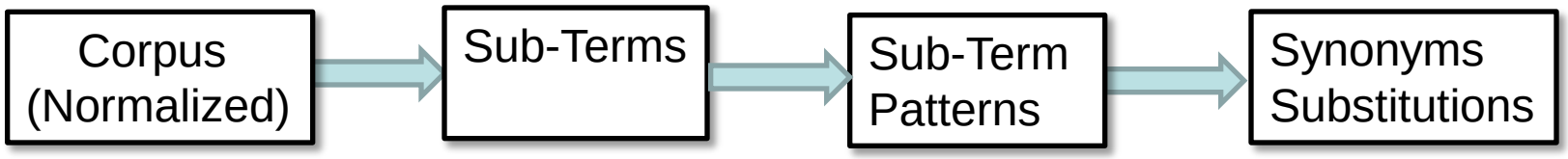
STMT - Details



STMT Release

- First release: STMT, 2013
- Java APIs and 5 command line tools
  - Generic tool
    - STMT (Sub-Term Mapping Tool)
  - Preloaded corpus & predefined norm
    - LSF (LexItem Sub-Term Finder)
    - SMT (Synonym Mapping Tool)
  - Others:
    - NT (Normalization Tool)
    - MT (Mapping Tool)
- Web site with full documents and supports
  http://umlslex.nlm.nih.gov/stmt

STMT


Corpus
(Normalized)
Sub-Terms
Sub-Term
Patterns
Synonyms
Substitutions

LSF - MetaMap

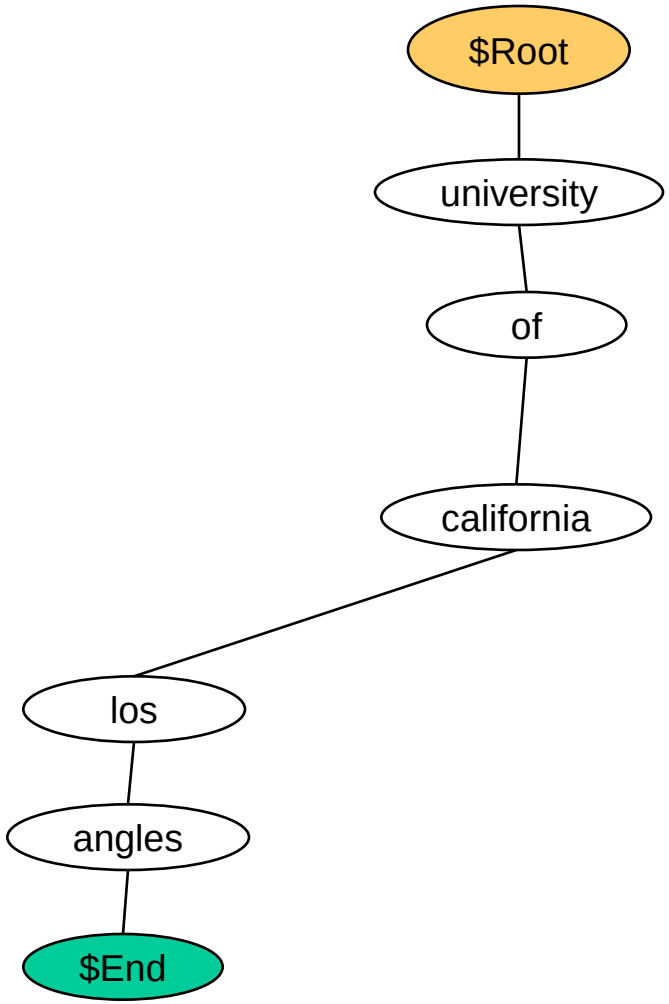
- Corpus (Lexicon) – LexItem Norm

Original Term from Lexicon	Normalized Term (key)	EUI
…	…	…
University of California, Los Angeles	university of california los angeles	E0063261
University of California Los Angeles	university of california los angeles	E0063261
University of California-Los Angeles	university of california los angeles	E0063261
University of California, San Francisco	university of california san francisco	E0063262
University of California SanFrancisco	university of california sanfrancisco	E0063262
University of California-SanFrancisco	university of california sanfrancisco	E0063262
University of California	university of california	E0702384
University of California, Davis	university of california davis	E0702385
University of California Davis	university of california davis	E0702385
University of California-Davis	university of california davis	E0702385
University of California, San Diego	university of california san diego	E0702386
…	…	…

- Load the corpus to a tree structure with each term as a branch in the tree and each word in the term as a node in the branch

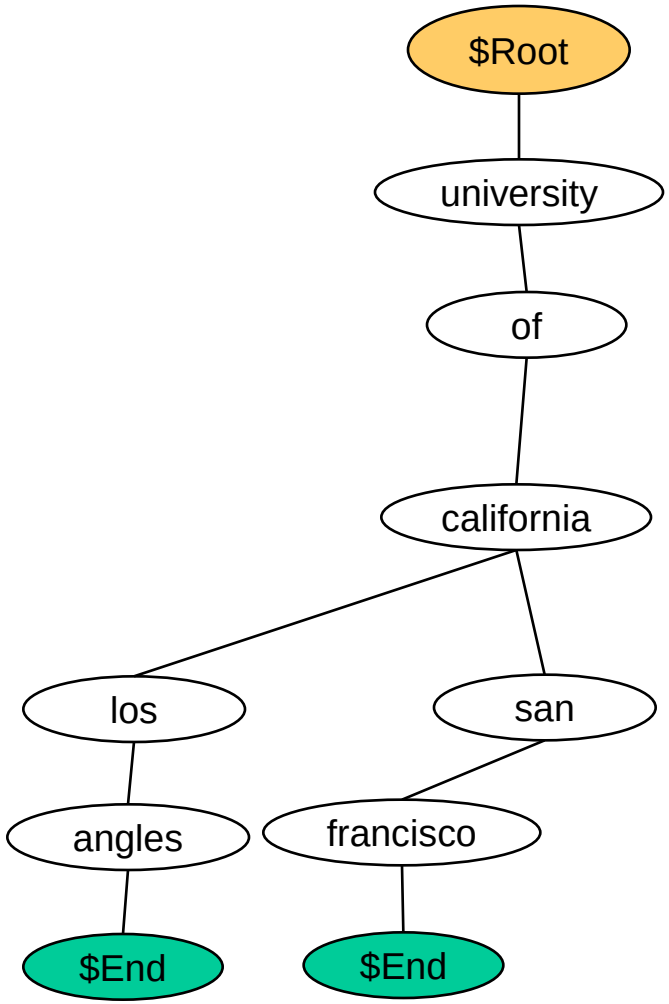
LSF


Corpus (Normalized)
Sub-Terms
Sub-Term Patterns
Synonyms Substitutions


$Root
university
of
california
los
angles
$End

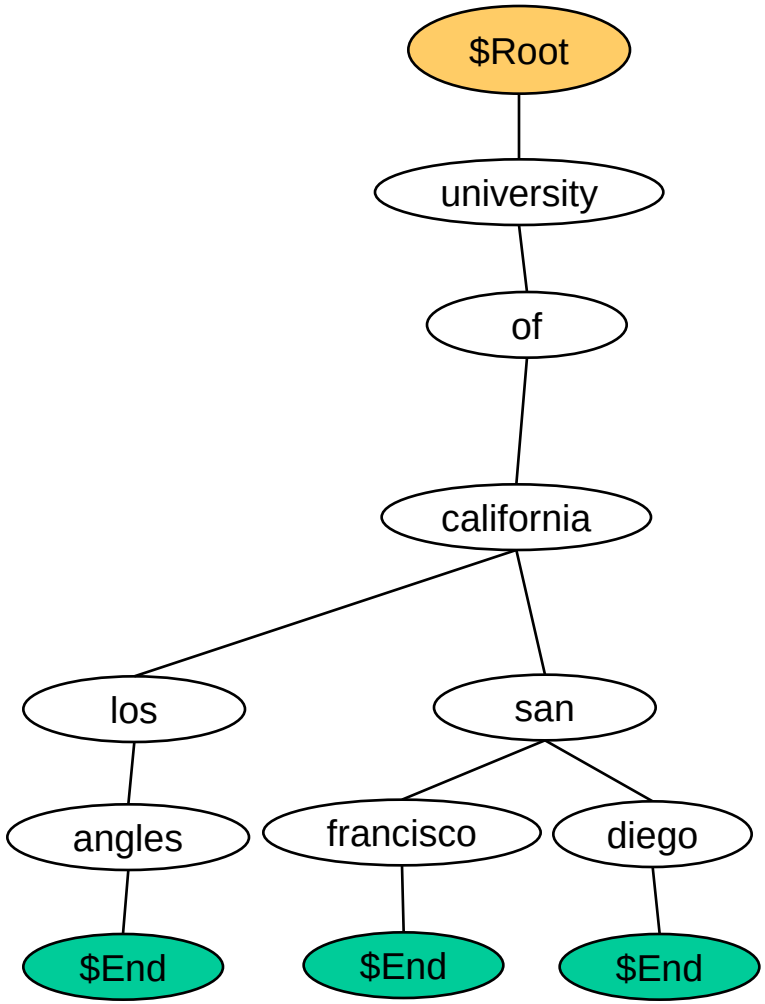
LSF


Corpus (Normalized)
Sub-Terms
Sub-Term Patterns
Synonyms Substitutions

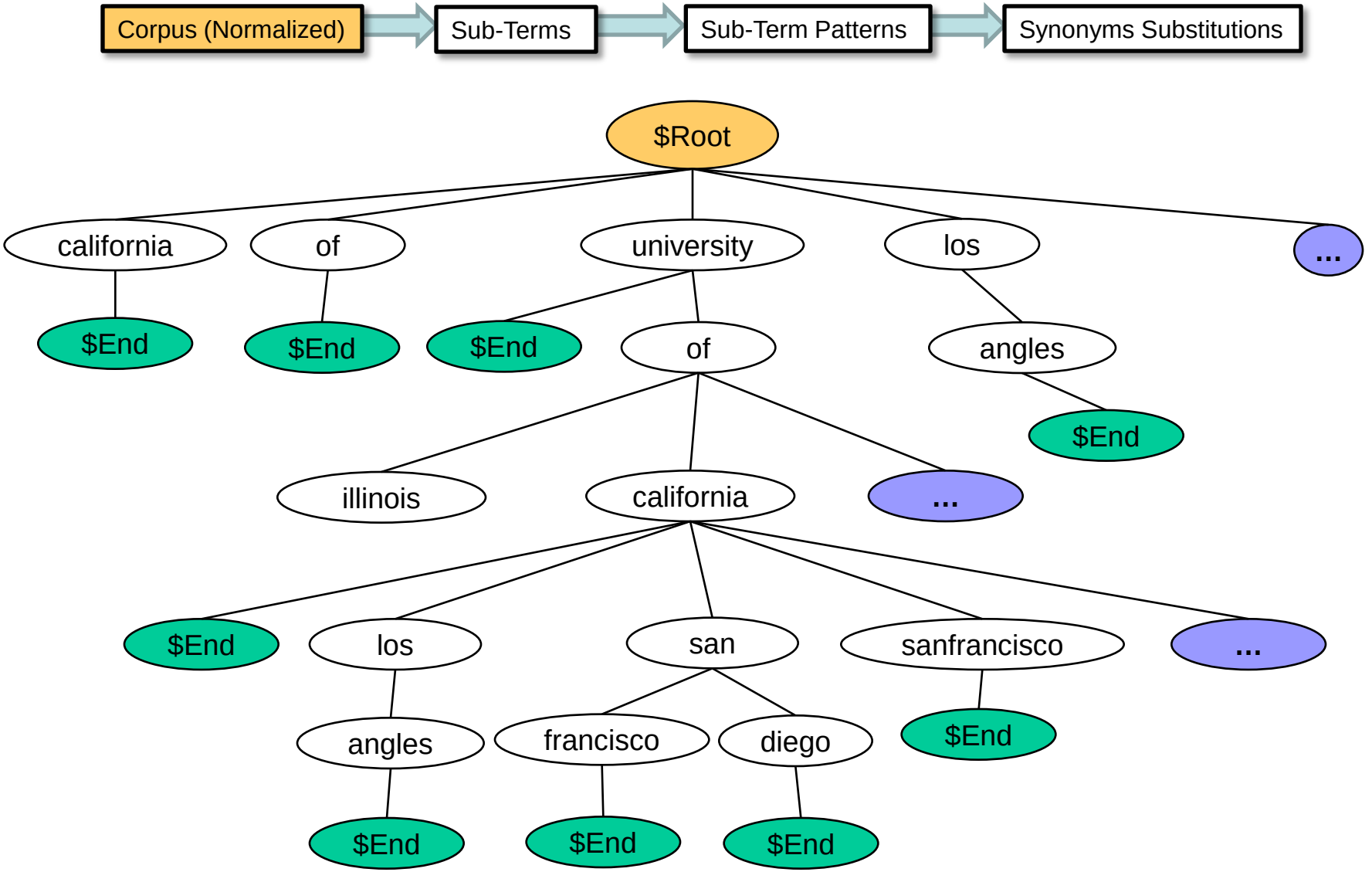

$Root
university
of
california
los
san
angles
francisco
$End
$End

LSF


Corpus (Normalized)
Sub-Terms
Sub-Term Patterns
Synonyms Substitutions


$Root
university
of
california
los
san
angles
francisco
diego
$End
$End
$End

LSF

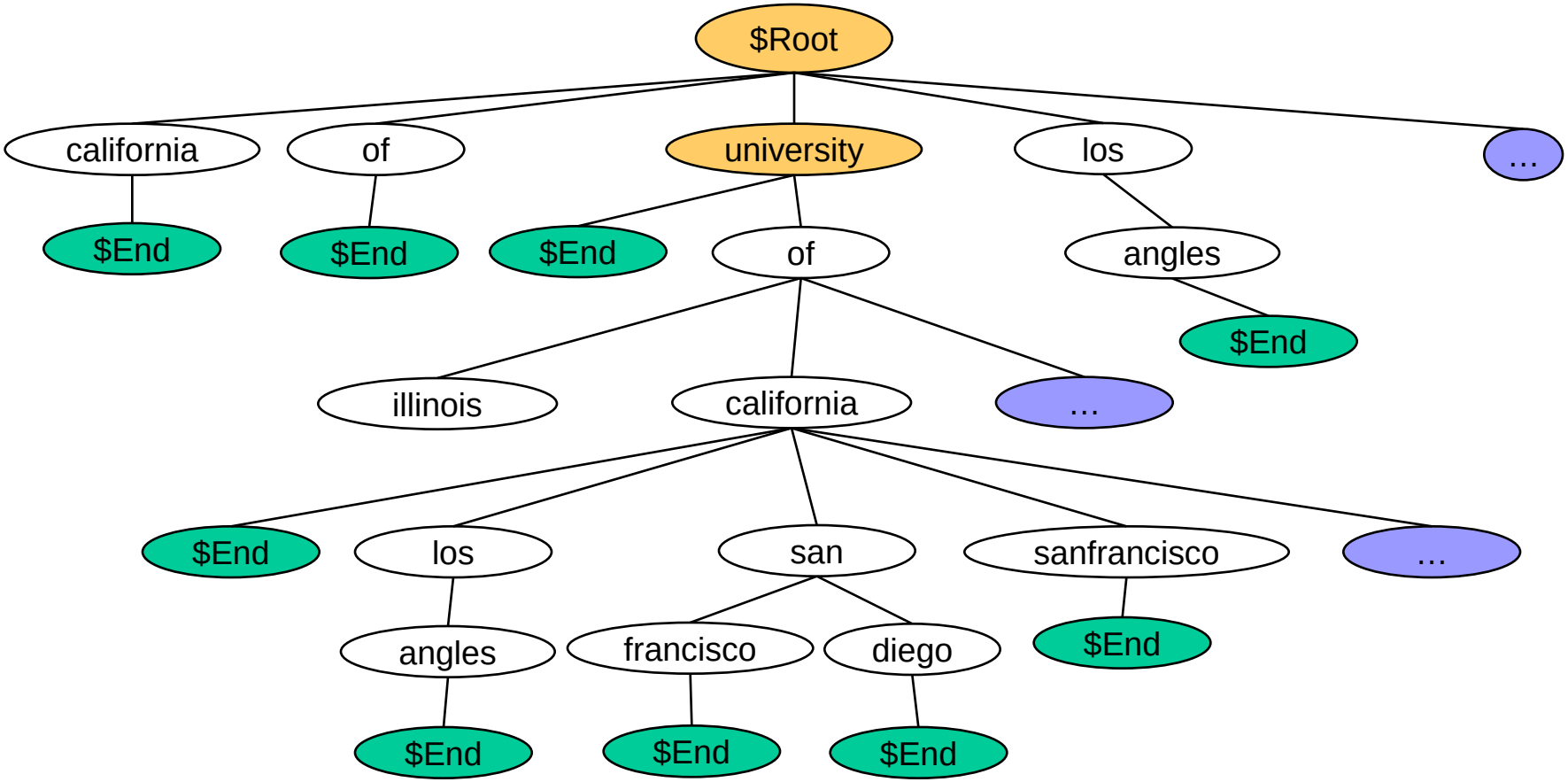

Corpus (Normalized)
Sub-Terms
Sub-Term Patterns
Synonyms Substitutions
$Root
california
of
university
los
...
$End
$End
$End
of
angles
$End
illinois
california
...
$End
los
san
sanfrancisco
...
angles
francisco
diego
$End
$End
$End
$End

LSF – Longest Prefix

university of california los angeles is in usa …

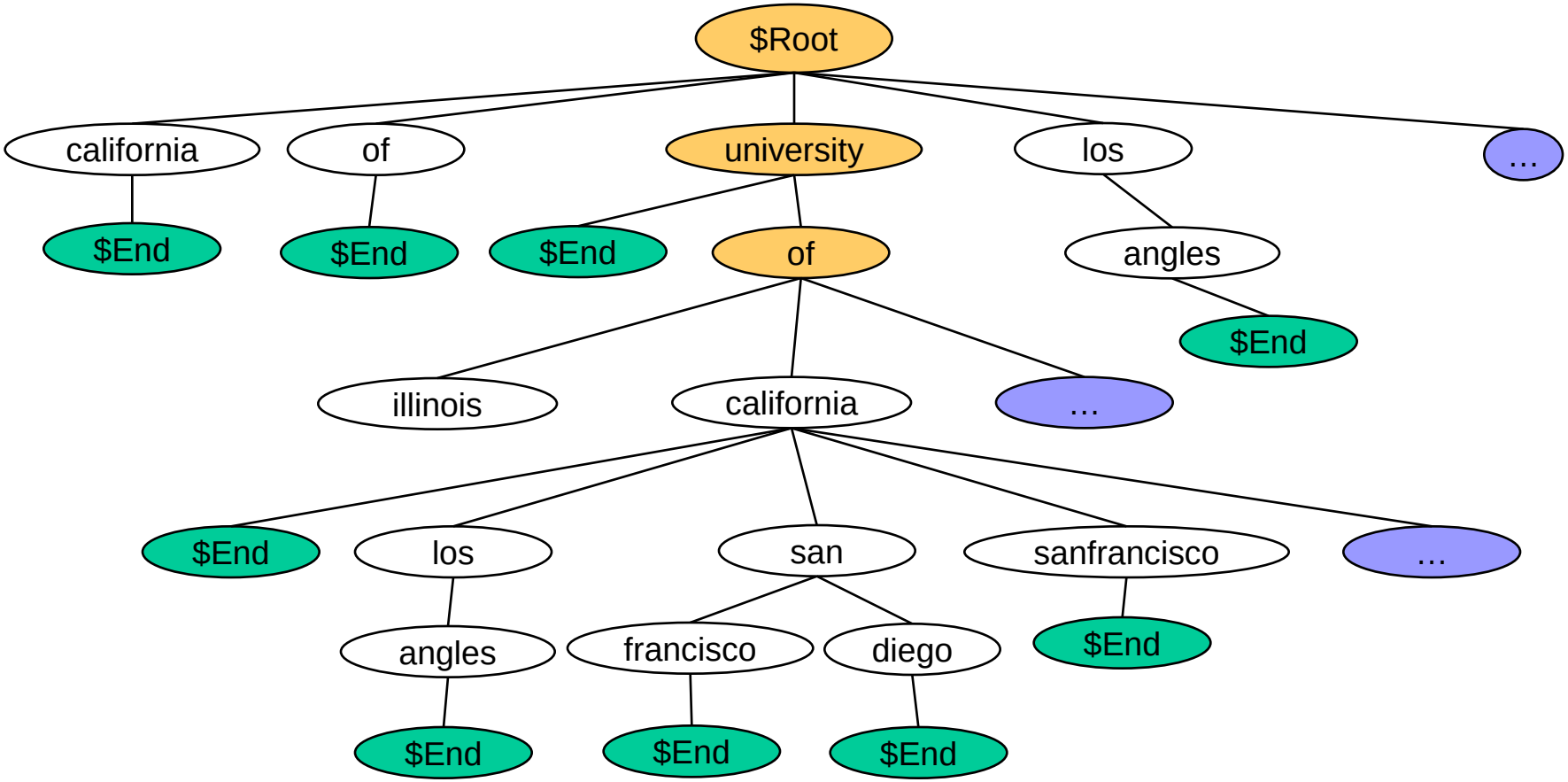
LSF – Longest Prefix

university of california los angeles is in usa …



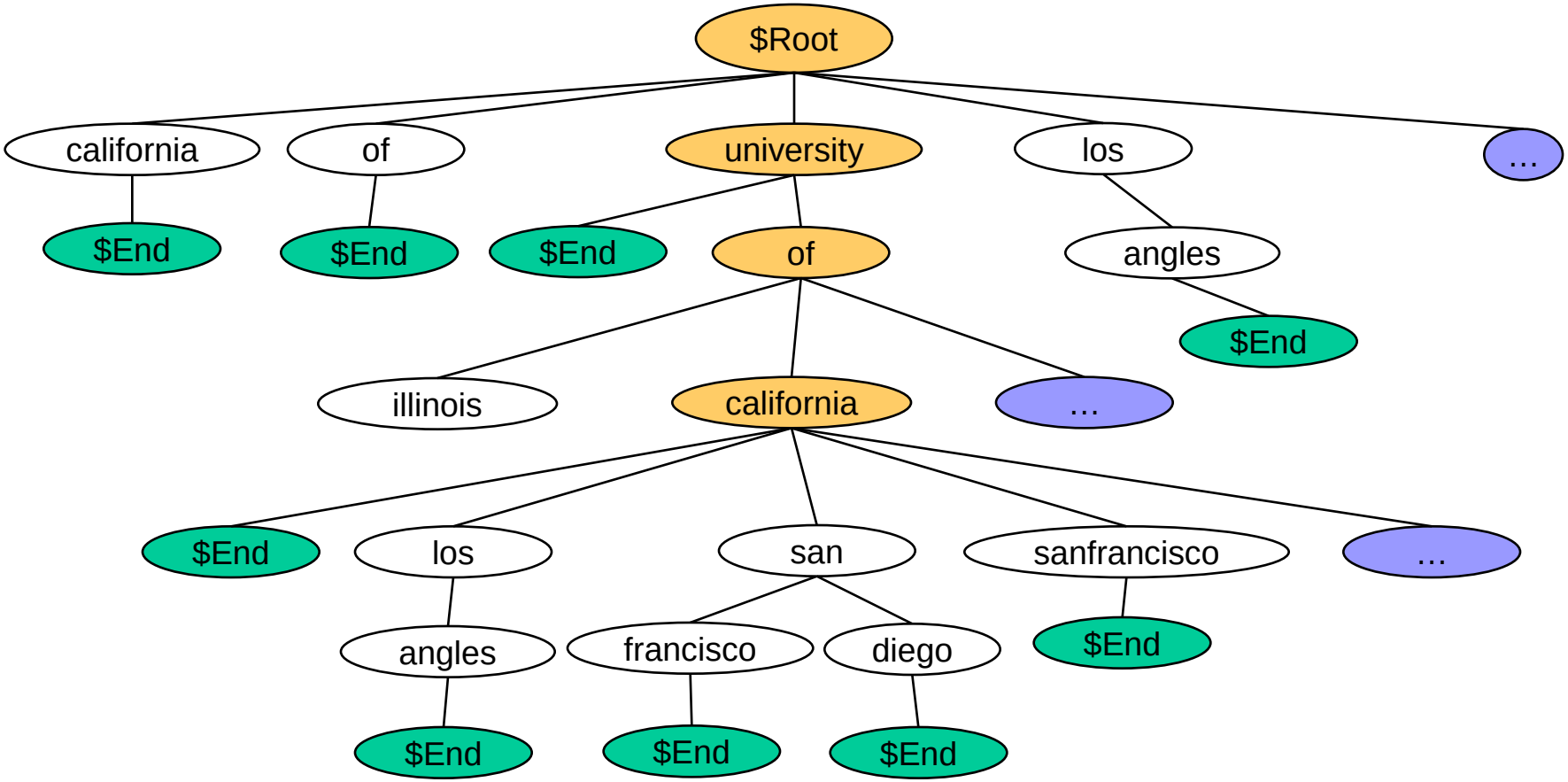
LSF – Longest Prefix

university of | california los angeles is in usa …



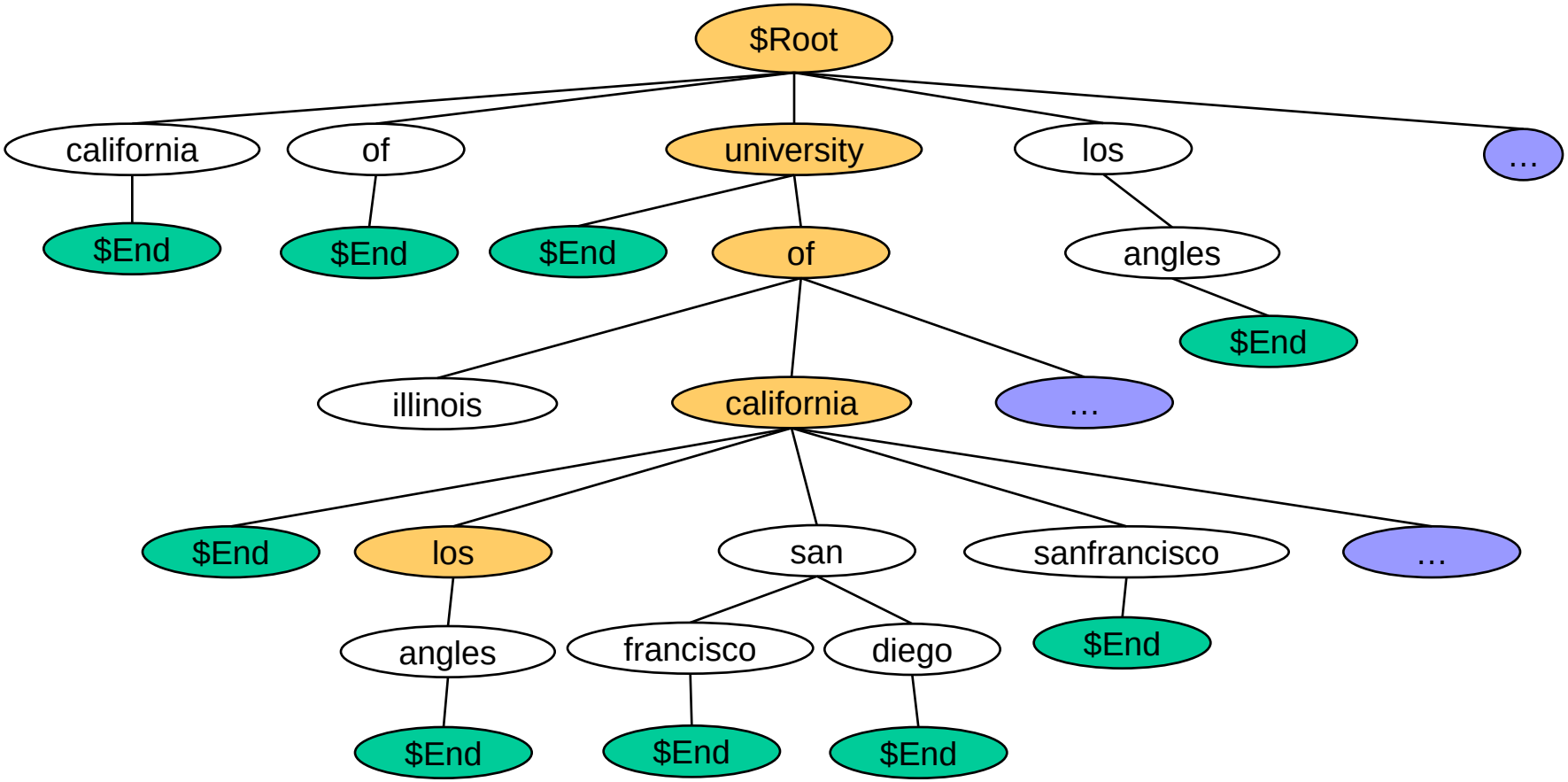
LSF – Longest Prefix

university of california los angeles is in usa …



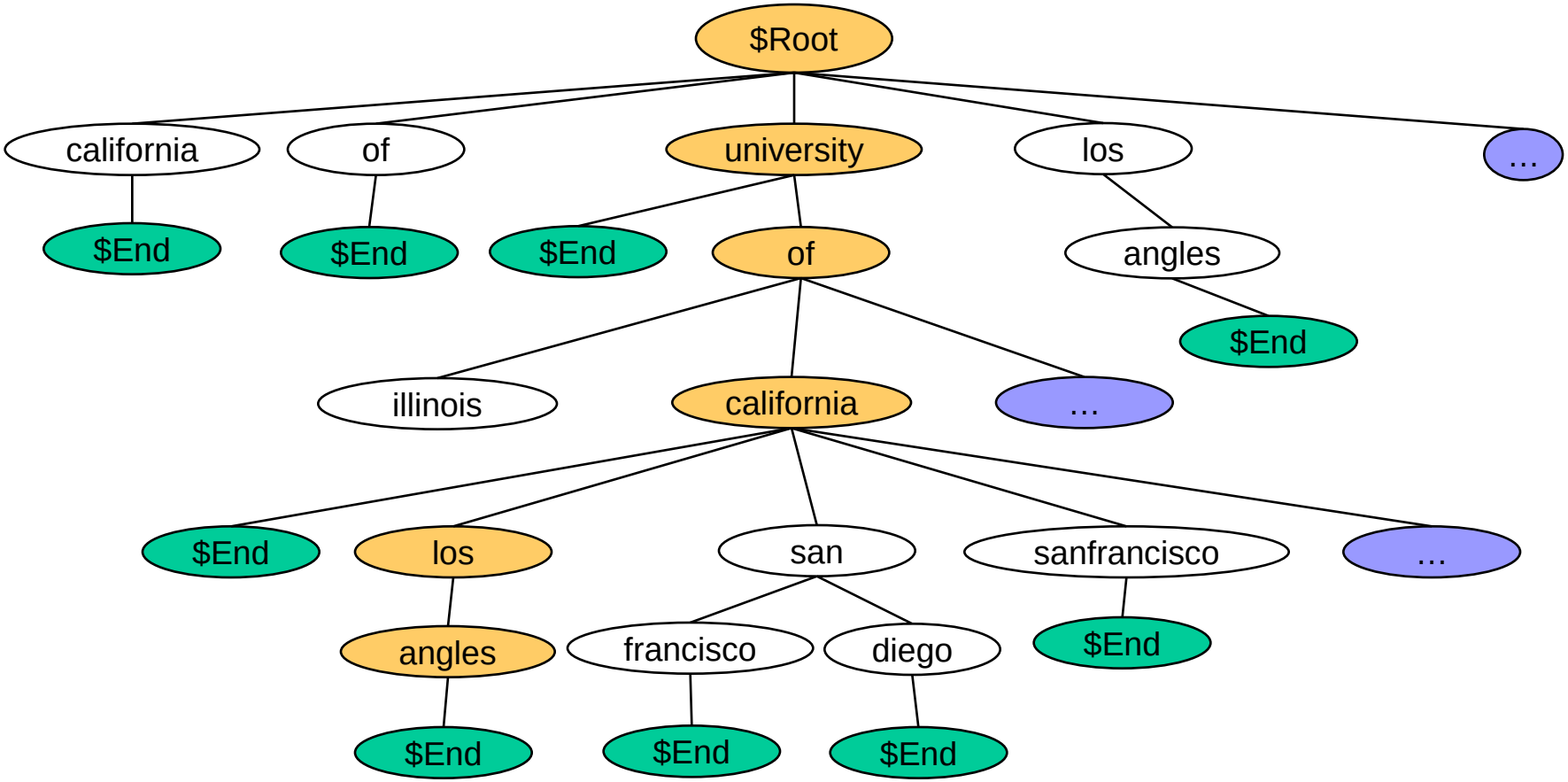
LSF – Longest Prefix

university of california los angeles is in usa …



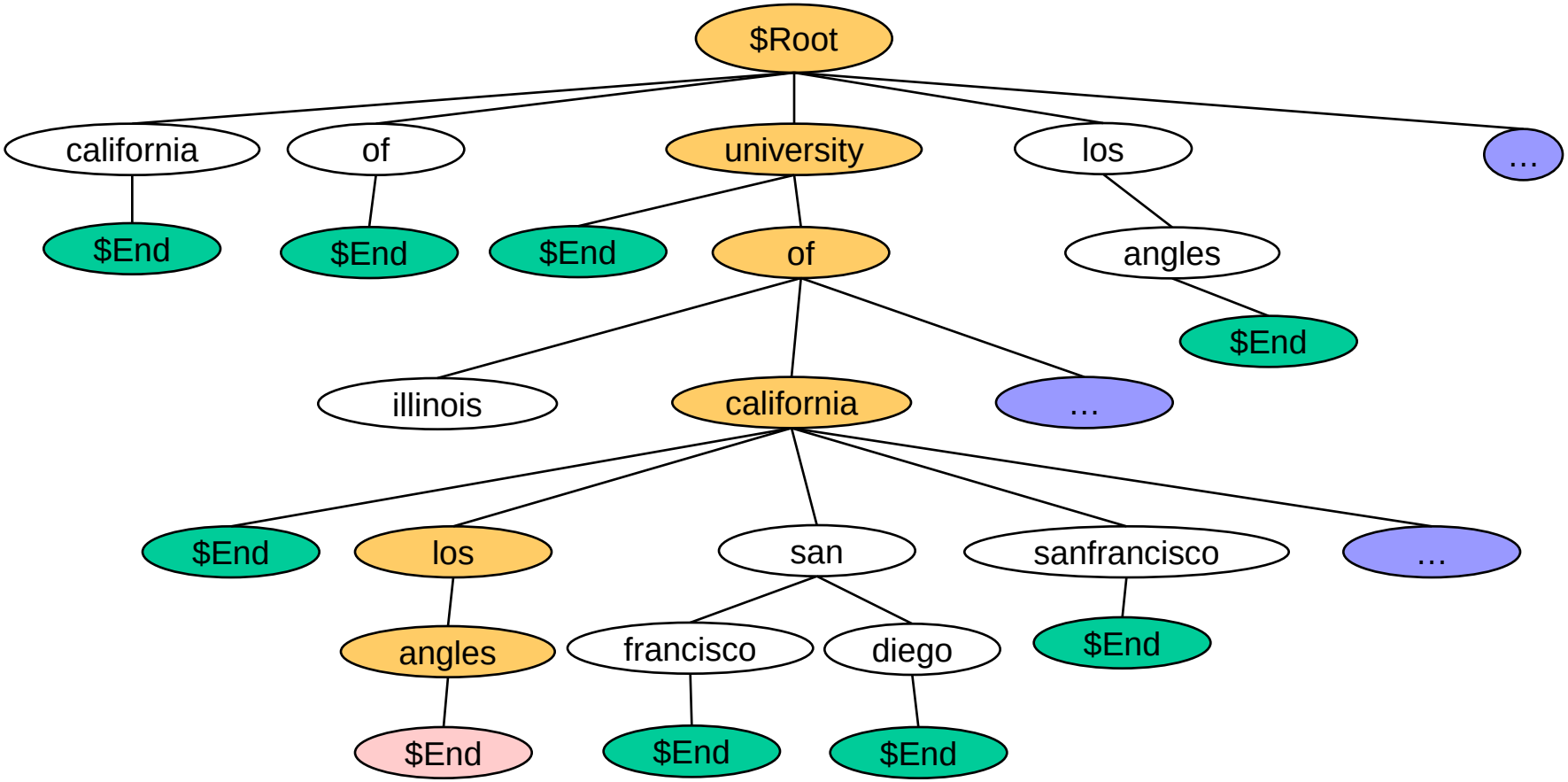
LSF – Longest Prefix

university of california los angeles is in usa …



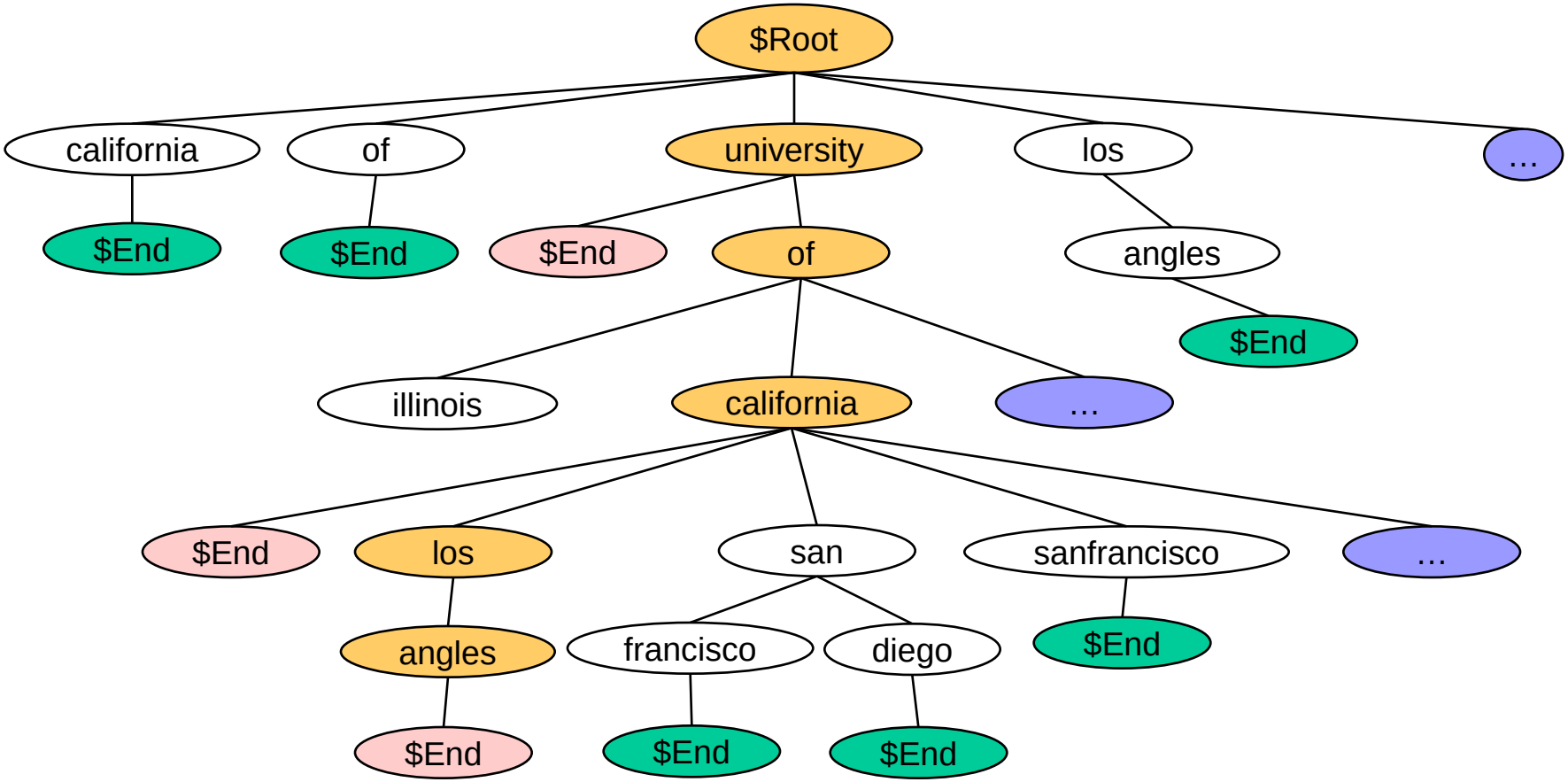
LSF – Longest Prefix

university of california los angeles is in usa …



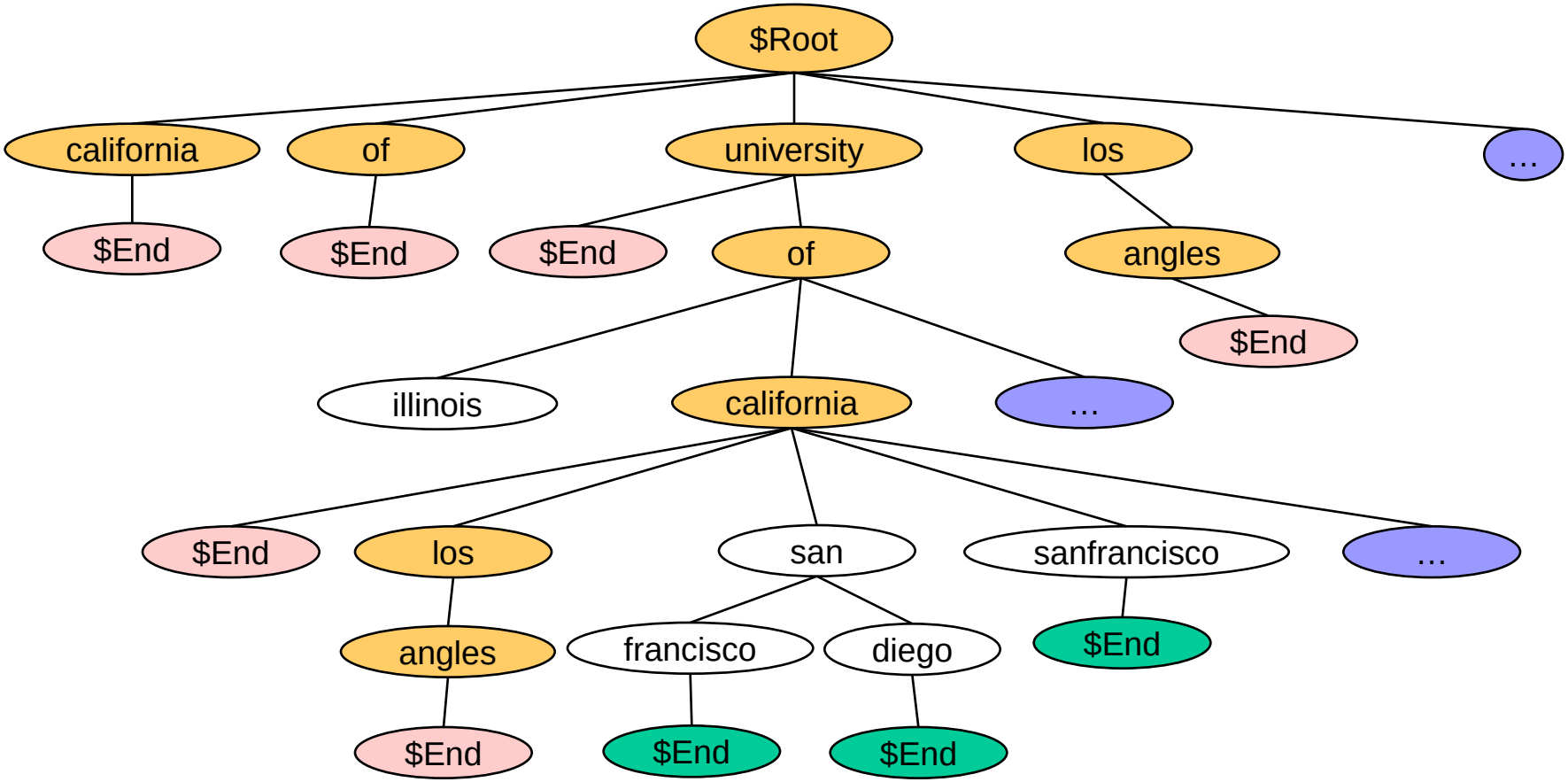
LSF – Prefixes

university of california los angeles is in usa …



LSF – Sub-terms

university of california los angeles is in usa …



SMT – UMLS-Core

- OTITIS EXTERNA, CHRONIC INFECTIOUS
  1) Normalized lexical matches
     - chronic externa infectious otitis => no CUI found
     - chronic externon infectious otitis => no CUI found
     - chronic externum infectious otitis => no CUI found
  2) One sub-term (synonymous word or phrase) substitution
  3) Two sub-term substitutions

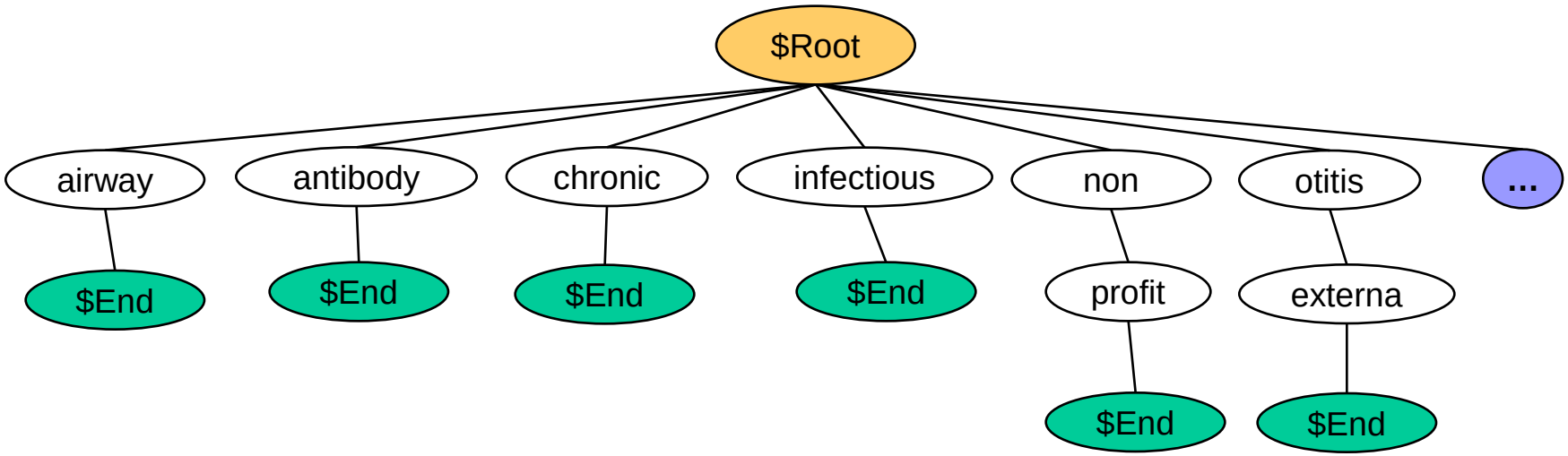
SMT – UMLS-Core

- Corpus (8 Synonyms files):
  - Acronyms, British, Greco-latin, ECRI, Lvg, etc.
- Synonym Norm (-f: g:rs:Ct:o:l)
  - Abstract away from punctuation, cases, spelling variants, inflectional variants
  - One to many
  - Remove from synonym list if the synonym pair are the same after norm (side-to-side | side to side)
- Load the normalized synonyms to corpus tree

SMT – UMLS-Core

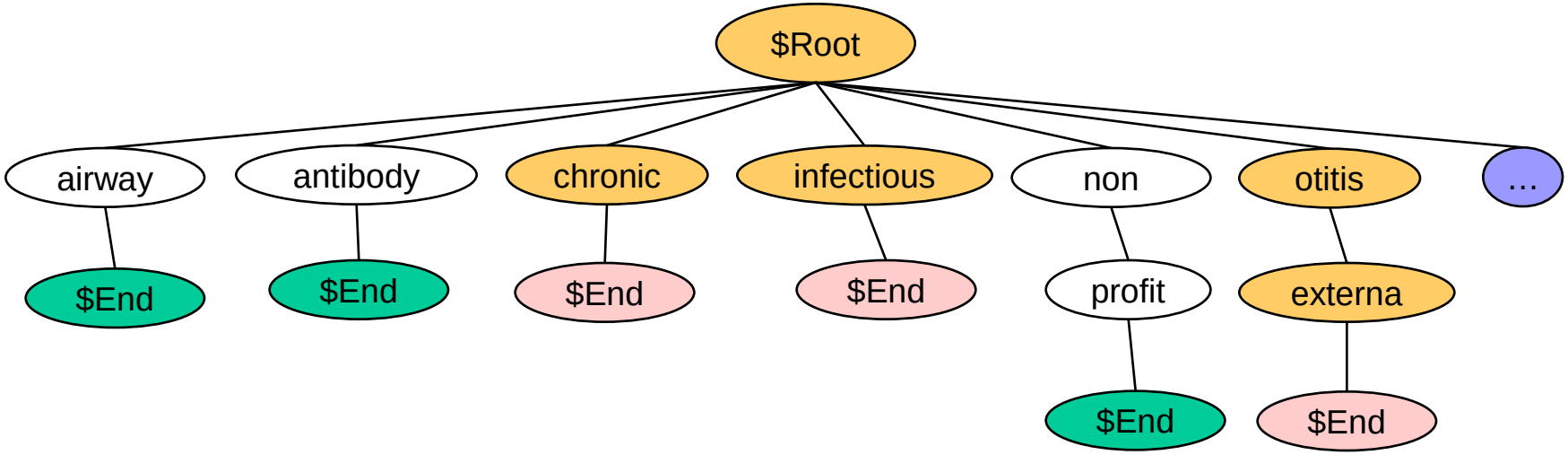
Original Term from Synonyms	Normalized Term (key)	synonyms
…	…	…
chronic	chronic	chron
chronic	chronic	long-term
chronic	chronic	persistent
chronic	chronic	recurrent periodic
chronic	chronic	relapsing
infectious	infectious	communicable
infectious	infectious	contagious
infectious	infectious	infection
otitis externa	otitis externa	auditory
otitis externa	otitis externa	auditory canal
otitis externa	otitis externa	aural
otitis externa	otitis externa	ear
ANTIBODY	antibody	AB
Airway	airway	Tube
non-profit	non profit	not for profit
…	…	…

SMT – UMLS-Core



SMT – UMLS-Core

- OTITIS EXTERNA, CHRONIC INFECTIOUS
  Normalized to => otitis externa chronic infectious
- Three sub-terms found
  - otitis externa|0|2
  - chronic|2|3
  - infectious|3|4



SMT – UMLS-Core

- OTITIS EXTERNA, CHRONIC INFECTIOUS
- => otitis externa chronic infectious

- Sub-term patterns with one substitution
  - **otitis externa** chronic infectious
  - otitis externa **chronic** infectious
  - otitis externa chronic **infectious**

- Sub-term patterns with two substitutions
  - **otitis externa** **chronic** infectious
  - **otitis externa** chronic **infectious**
  - otitis externa **chronic** **infectious**

- It could be really complicated (multiple layers of overlap or inclusion)

SMT – UMLS-Core

- otitis externa chronic infectious
- Synonyms:
  - otitis externa (4): auditory | auditory canal | aural | ear
  - chronic (5): chron | long-term | persistent | relapsing | recurrent periodic
  - infectious (3): communicable | contagious | infection

- Sub-term patterns with one substitution (12)
  - **otitis externa** chronic infectious (4)
  - otitis externa **chronic** infectious (5)
  - otitis externa chronic **infectious** (3)

- Sub-term patterns with two substitutions (47)
  - **otitis externa** **chronic** infectious (20)
  - **otitis externa** chronic **infectious** (12)
  - otitis externa **chronic** **infectious** (15)

SMT – UMLS-Core

- otitis externa chronic infectious

- Sub-term patterns with one substitution (12)
  - **otitis externa** chronic infectious (4)
  - otitis externa **chronic** infectious (5)
  - otitis externa chronic **infectious** (3)

  => No CUI Found

- Sub-term patterns with two substitutions (47)
  - **otitis externa** **chronic** infectious (20)
  - **otitis externa** chronic **infectious** (12)
  - otitis externa **chronic** **infectious** (15)

  => ear chronic infection | C0743359 | EAR INFECTION CHRONIC

SMT – Ambiguity Study

- Use for word sense disambiguity (WSD) project
- To find/compare the degree of concepts between UMLS-Metathesaurus releases for manually review suppress in MetaMap.
- Example:
  - shell> smt -p -pt –x:smt.properties.2011AA
    - Please input a term (type "Ctl-d" to quit) >
    herbal medicine
    herbal medicine|herbal medicine|C0025125|Medicinal Herbs|0
    herbal medicine|herbal medicine|C1533719|Discipline of Herbal Medicine|0
    herbal medicine|herbal medicine|C2240391|Herbal medicine (product)|0

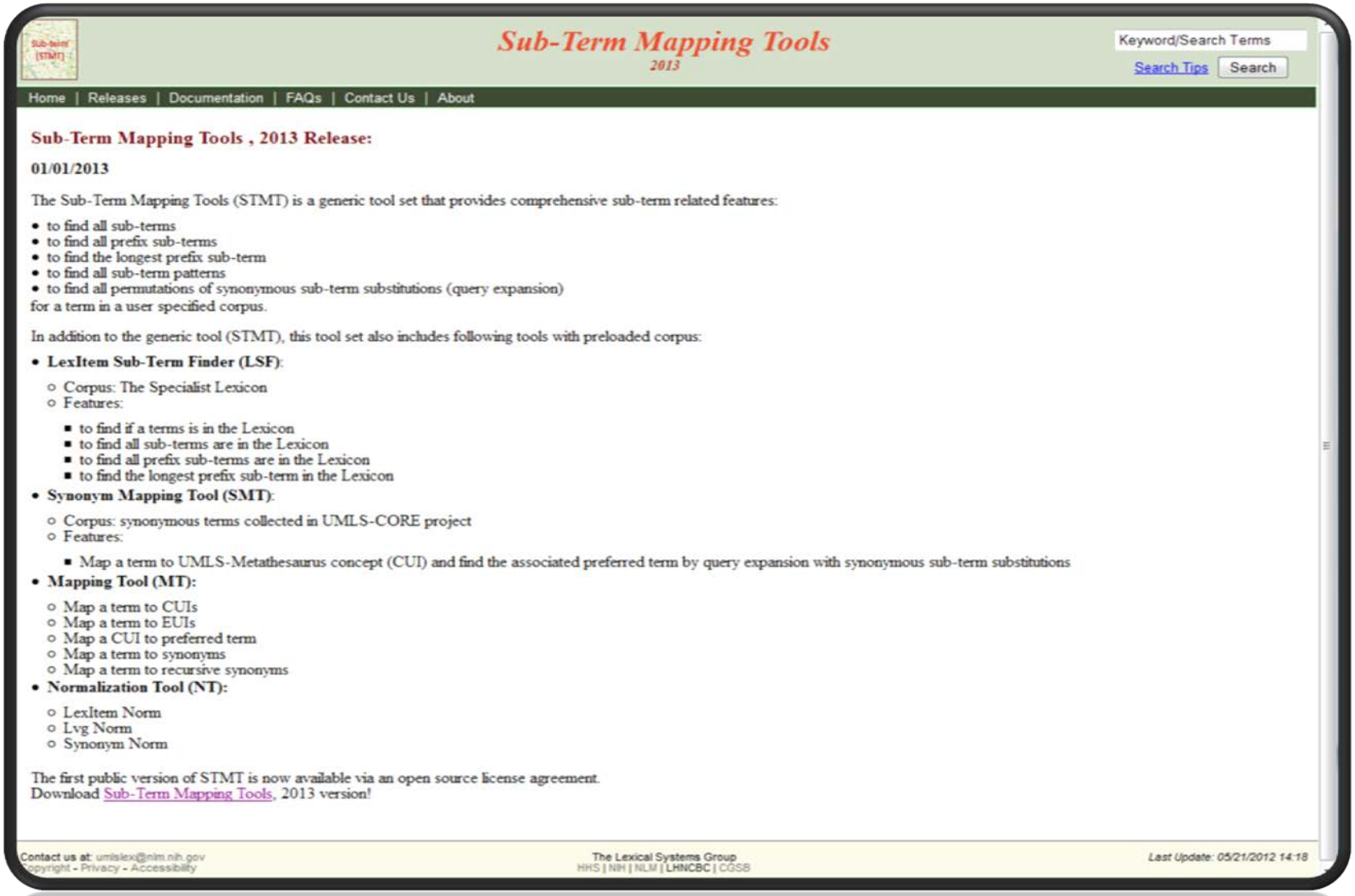
  - shell> smt -p -pt –x:smt.properties.2012AA
    - Please input a term (type "Ctl-d" to quit) >
    herbal medicine
    herbal medicine|herbal medicine|C0025125|Medicinal Herbs|0
    herbal medicine|herbal medicine|C0242388|Phytotherapy|0
    herbal medicine|herbal medicine|C1533719|Discipline of Herbal Medicine|0
    herbal medicine|herbal medicine|C2240391|Herbal medicine (product)|0

Future Work

- Users' Feedback
- Persistent corpus

STMT Web-Site

http://umlslex.nlm.nih.gov/stmt



Sub-Term Mapping Tools
2013

Keyword/Search Terms
Search Tips | Search

Home | Releases | Documentation | FAQs | Contact Us | About

Sub-Term Mapping Tools , 2013 Release:

01/01/2013

The Sub-Term Mapping Tools (STMT) is a generic tool set that provides comprehensive sub-term related features:

- to find all sub-terms
- to find all prefix sub-terms
- to find the longest prefix sub-term
- to find all sub-term patterns
- to find all permutations of synonymous sub-term substitutions (query expansion)

for a term in a user specified corpus.

In addition to the generic tool (STMT), this tool set also includes following tools with preloaded corpus:

- **LexItem Sub-Term Finder (LSF)**:
  - Corpus: The Specialist Lexicon
  - Features:
    - to find if a terms is in the Lexicon
    - to find all sub-terms are in the Lexicon
    - to find all prefix sub-terms are in the Lexicon
    - to find the longest prefix sub-term in the Lexicon
- **Synonym Mapping Tool (SMT)**:
  - Corpus: synonymous terms collected in UMLS-CORE project
  - Features:
    - Map a term to UMLS-Metathesaurus concept (CUI) and find the associated preferred term by query expansion with synonymous sub-term substitutions
- **Mapping Tool (MT):**
  - Map a term to CUIs
  - Map a term to EUIs
  - Map a CUI to preferred term
  - Map a term to synonyms
  - Map a term to recursive synonyms
- **Normalization Tool (NT):**
  - LexItem Norm
  - Lvg Norm
  - Synonym Norm

The first public version of STMT is now available via an open source license agreement.
Download Sub-Term Mapping Tools, 2013 version!

Contact us at: umlslex@nlm.nih.gov
Copyright - Privacy - Accessibility

The Lexical Systems Group
HHS | NIH | NLM | LHNCBC | CGSB

Last Update: 05/21/2012 14:18

References

- Lu, Chris J. and Browne, Allen C., “Development of Sub-Term Mapping Tools", Submitted for publication in Proceeding of AMIA 2012 Annual Symposium, Nov. 3-7, 2012, Chicago, IL

- http://umlslex.nlm.nih.gov/stmt

Questions

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