Laboratoire d’Informatique de Paris-Nord
December 7, 2004

The Unified Medical Language System

Identifying relations among biomedical terms

NATIONAL
LIBRARY OF
MEDICINE

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

Outline

- The Unified Medical Language System *Olivier Bodenreider*
  - Overview
  - Identifying relations among biomedical terms

- Extension of the UMLS to processing French language *Pierre Zweigenbaum*

The Unified Medical Language System

Overview

Bodenreider O.
The Unified Medical Language System (UMLS): Integrating biomedical terminology.
Nucleic Acids Research; 2004. p. D267-D270.

Motivation

- Started in 1986
- National Library of Medicine
- “Long-term R&D project”
- Complementary to IAIMS

(Integrated Academic
Information Management Systems)

«[…] the UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable information.

- The first is the variety of ways the same concepts are expressed in different machine-readable sources and by different people.
- The second is the distribution of useful information among many disparate databases and systems.»

The UMLS in practice

- Database
  - Series of relational files
- Interfaces
  - Web interface: Knowledge Source Server (UMLSKS)
  - Application programming interfaces (Java and XML-based)
- Applications
  - lvg (lexical programs)
  - MetamorphoSys (installation and customization)

The UMLS is *not* an end-user application

UMLS 3 components

- Metathesaurus
  - Concepts
  - Inter-concept relationships
- Semantic Network
  - Semantic types
  - Semantic network relationships
- Lexical resources
  - SPECIALIST Lexicon
  - Lexical tools

UMLS Metathesaurus

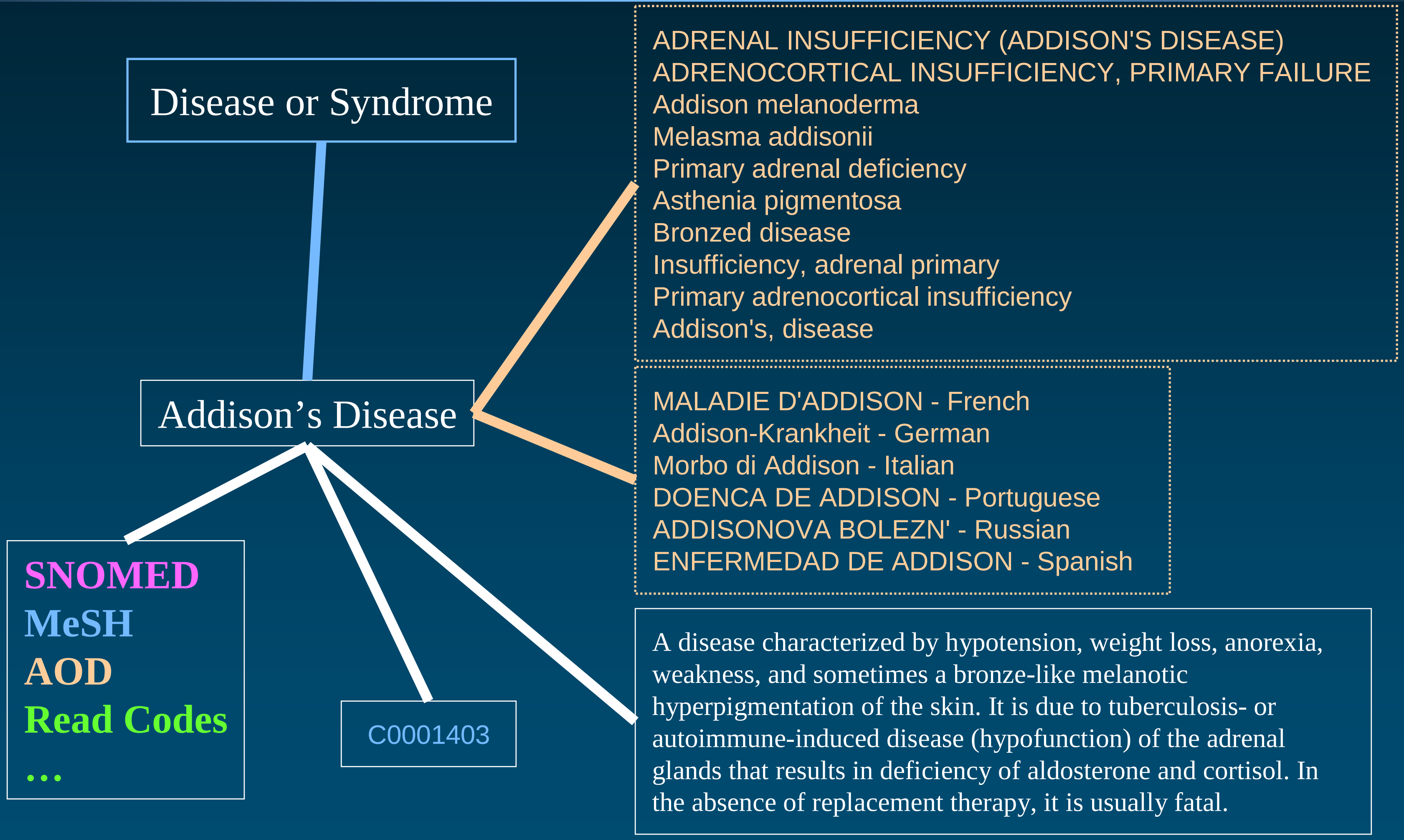
Metathesaurus Basic organization

- Concepts
  - Synonymous terms are clustered into a concept
  - Properties are attached to concepts, e.g.,
    - Unique identifier
    - Definition
- Relations
  - Concepts are related to other concepts
  - Properties are attached to relations, e.g.,
    - Type of relationship
    - Source

Source Vocabularies (2004AB)

- 134 source vocabularies
  - 126 contributing concept names
- 73 families of vocabularies
  - multiple translations (e.g., MeSH, ICPC, ICD-10)
  - variants (American-English equivalents, Australian extension/adaptation)
  - subsequent editions usually considered distinct families (ICD: 9-10; DSM: IIIR-IV)
- Broad coverage of biomedicine
- Common presentation

Addison’s Disease: Concept



Disease or Syndrome

Addison’s Disease

ADRENAL INSUFFICIENCY (ADDISON'S DISEASE)
ADRENOCORTICAL INSUFFICIENCY, PRIMARY FAILURE
Addison melanoderma
Melasma addisonii
Primary adrenal deficiency
Asthenia pigmentosa
Bronzed disease
Insufficiency, adrenal primary
Primary adrenocortical insufficiency
Addison's, disease

MALADIE D'ADDISON - French
Addison-Krankheit - German
Morbo di Addison - Italian
DOENCA DE ADDISON - Portuguese
ADDISONOVA BOLEZN' - Russian
ENFERMEDAD DE ADDISON - Spanish

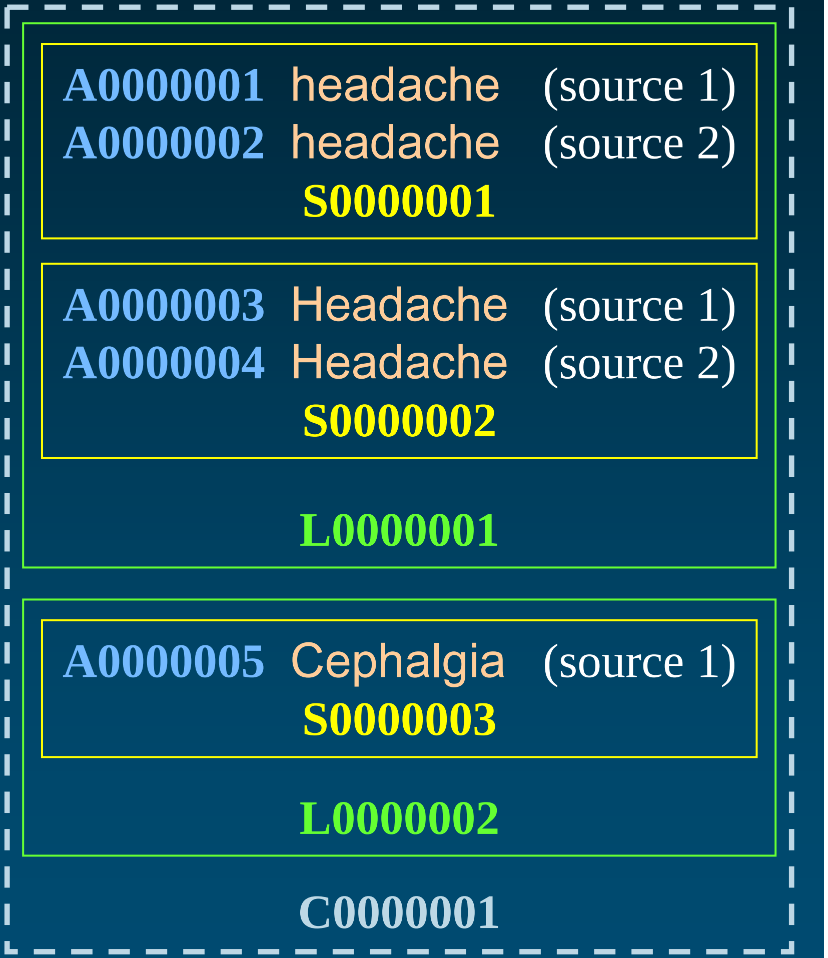
SNOMED
MeSH
AOD
Read Codes
…

C0001403

A disease characterized by hypotension, weight loss, anorexia, weakness, and sometimes a bronze-like melanotic hyperpigmentation of the skin. It is due to tuberculosis- or autoimmune-induced disease (hypofunction) of the adrenal glands that results in deficiency of aldosterone and cortisol. In the absence of replacement therapy, it is usually fatal.

Metathesaurus Concepts (2004AB)

- Concept (> 1M) CUI
  - Set of synonymous concept names
- Term (> 3.8 M) LUI
  - Set of normalized names
- String (> 4.3M) SUI
  - Distinct concept name
- Atom (> 5.1M) AUI
  - Concept name in a given source



A0000001 headache (source 1)
A0000002 headache (source 2)
S0000001

A0000003 Headache (source 1)
A0000004 Headache (source 2)
S0000002

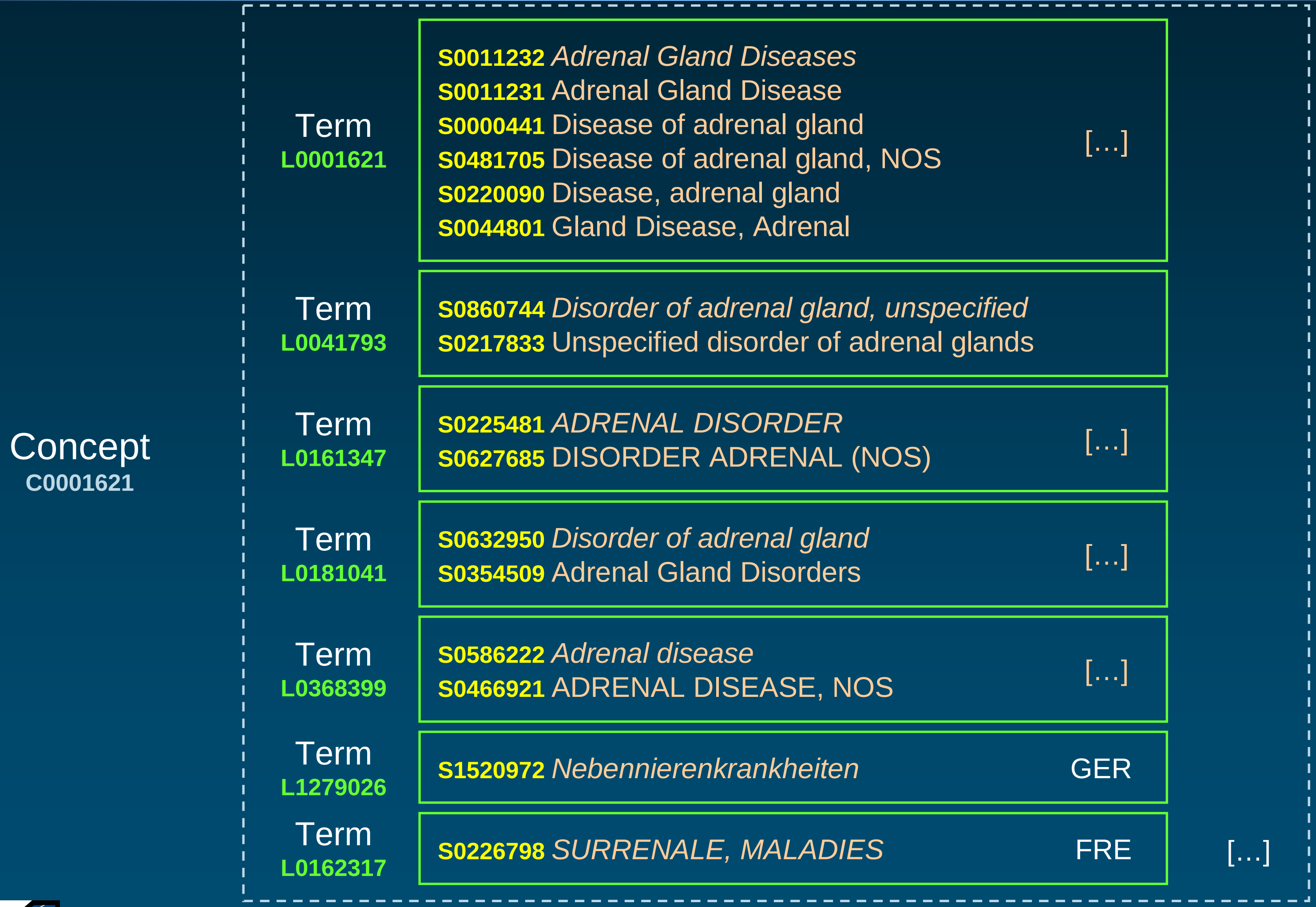
L0000001

A0000005 Cephalgia (source 1)
S0000003

L0000002

C0000001

Cluster of synonymous terms



Concept
C0001621

Term
L0001621

S0011232 *Adrenal Gland Diseases*
S0011231 Adrenal Gland Disease
S0000441 Disease of adrenal gland
S0481705 Disease of adrenal gland, NOS
S0220090 Disease, adrenal gland
S0044801 Gland Disease, Adrenal
[…]

Term
L0041793

S0860744 *Disorder of adrenal gland, unspecified*
S0217833 Unspecified disorder of adrenal glands

Term
L0161347

S0225481 *ADRENAL DISORDER*
S0627685 DISORDER ADRENAL (NOS)
[…]

Term
L0181041

S0632950 *Disorder of adrenal gland*
S0354509 Adrenal Gland Disorders
[…]

Term
L0368399

S0586222 *Adrenal disease*
S0466921 ADRENAL DISEASE, NOS
[…]

Term
L1279026

S1520972 *Nebennierenkrankheiten* GER

Term
L0162317

S0226798 *SURRENALE, MALADIES* FRE

[…]

Metathesaurus Relationships

- Symbolic relations: ~9 M pairs of concepts
- Statistical relations : ~7 M pairs of concepts (co-occurring concepts)
- Mapping relations: 100,000 pairs of concepts

- Categorization: Relationships between concepts and semantic types from the Semantic Network

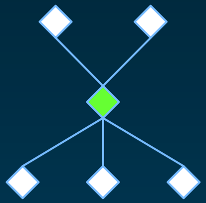
Symbolic relations

- Relation
  - Pair of “atom” identifiers
  - Type
  - Attribute (if any)
  - List of sources (for type and attribute)
- Semantics of the relationship: defined by its type [and attribute]

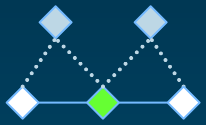
Source transparency: the information is recorded at the “atom” level

Symbolic relationships Type

- Hierarchical
  - Parent / Child PAR/CHD
  - Broader / Narrower than RB/RN



- Derived from hierarchies
  - Siblings (children of parents) SIB

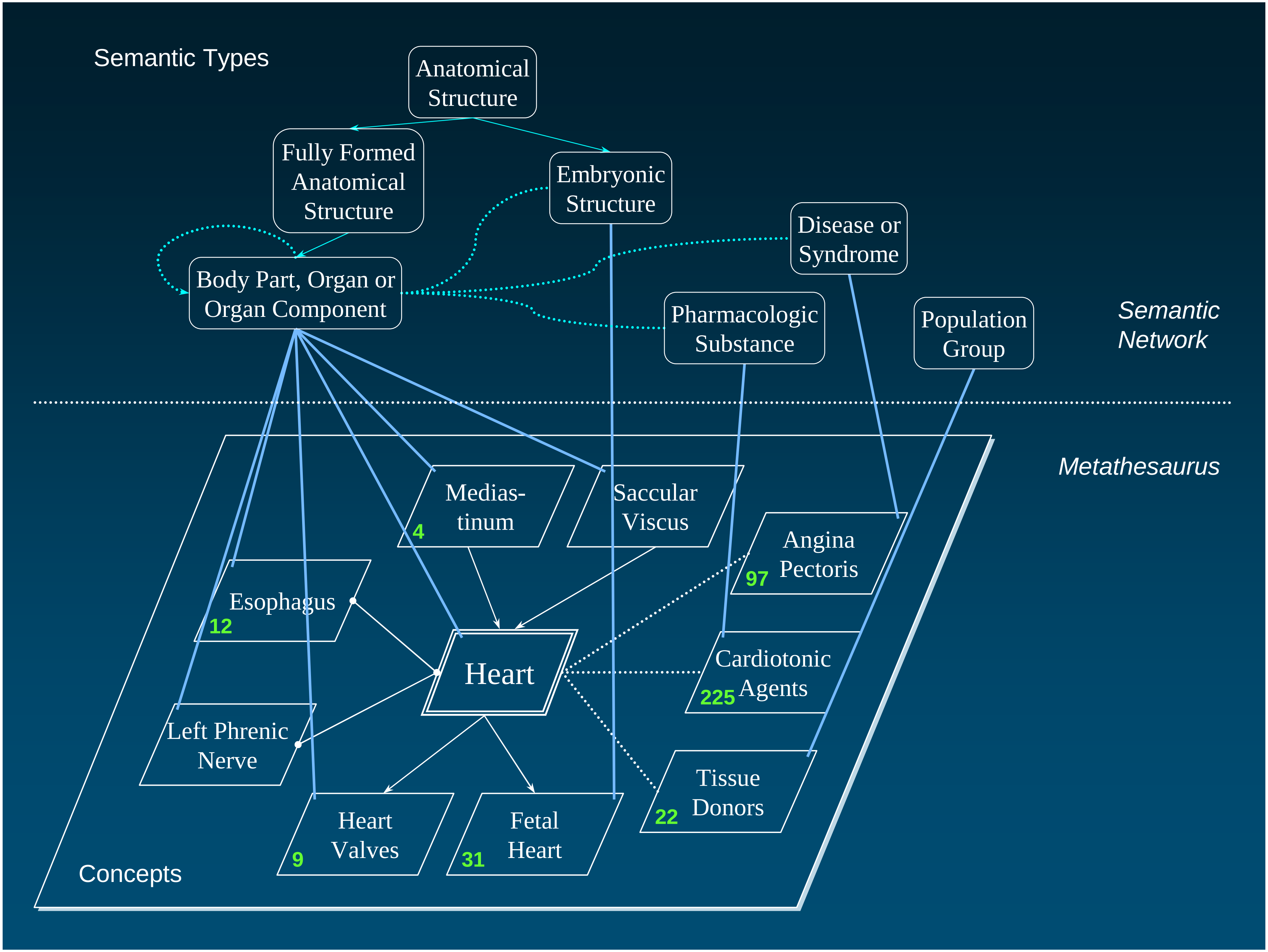


- Associative
  - Other RO

- Various flavors of near-synonymy
  - Similar RL
  - Source asserted synonymy SY
  - Possible synonymy RQ

Symbolic relationships Attribute

- Hierarchical
  - isa (is-a-kind-of)
  - part-of
- Associative
  - location-of
  - caused-by
  - treats
  - …
- Cross-references (mapping)


Semantic Types
Anatomical Structure
Fully Formed Anatomical Structure
Embryonic Structure
Body Part, Organ or Organ Component
Disease or Syndrome
Pharmacologic Substance
Population Group
Semantic Network
Metathesaurus
Medias-tinum
4
Saccular Viscus
Angina Pectoris
97
Esophagus
12
Heart
Cardiotonic Agents
225
Left Phrenic Nerve
Tissue Donors
22
Heart Valves
9
Fetal Heart
31
Concepts

SPECIALIST Lexicon and lexical tools

SPECIALIST Lexicon

- Content
  - English lexicon
  - Many words from the biomedical domain
- 200,000+ lexical items
- Word properties
  - morphology
  - orthography
  - syntax
- Used by the lexical tools

Not available
in other languages

Morphology

- Inflection

noun	nucleus, nuclei
verb	cauterize, cauterizes, cauterized, cauterizing
adjective	red, redder, reddest

- Derivation

verb ⟺ noun	cauterize -- cauterization
adjective ⟺ noun	red -- redness

Orthography

- Spelling variants

oe/e	oesophagus - esophagus
ae/e	anaemia - anemia
ise/ize	cauterise - cauterize
genitive mark	Addison's disease Addison disease Addisons disease

Syntax

- Complementation
  - verbs
    - intransitive — I'll treat.
    - transitive — He treated the patient.
    - ditransitive — He treated the patient with a drug.
  - nouns
    - prepositional phrase

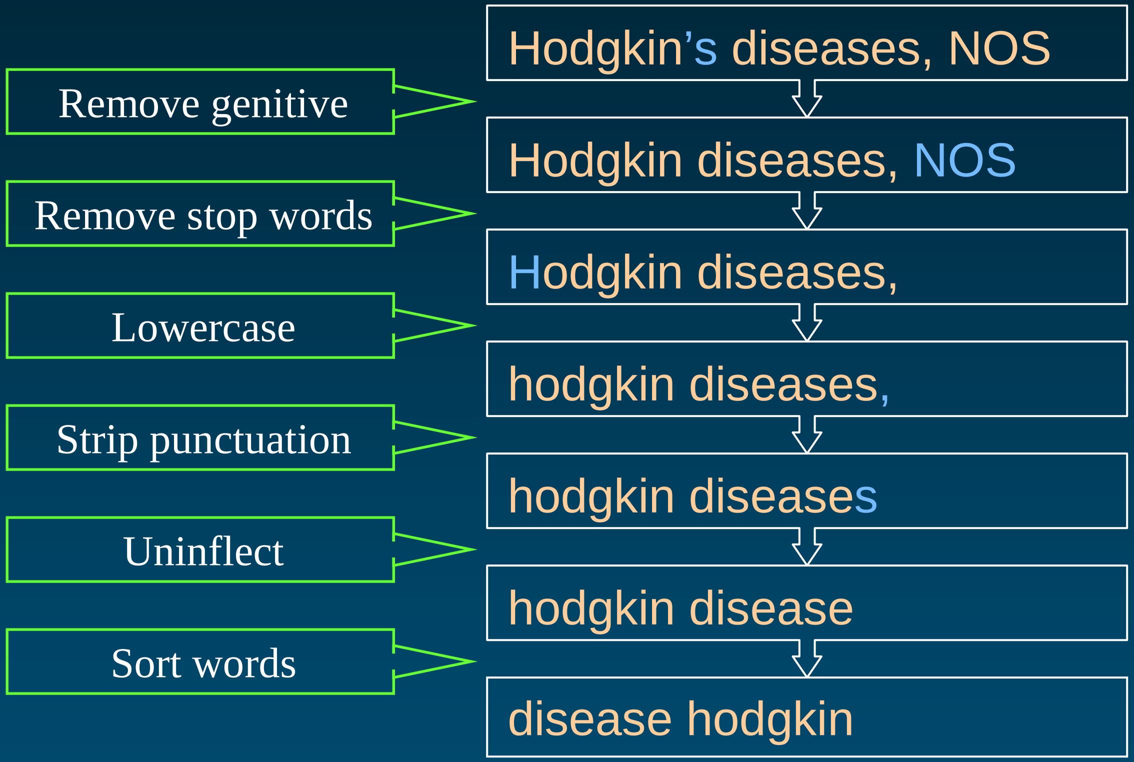
      Valve of coronary sinus
- Position for adjectives

Lexical tools

- To manage lexical variation in biomedical terminologies
- Major tools
  - Normalization
  - Indexes
  - Lexical Variant Generation program (lvg)
- Based on the SPECIALIST Lexicon
- Used by noun phrase extractors, search engines

Not available in other languages

Normalization



Hodgkin’s diseases, NOS

Remove genitive

Hodgkin diseases, NOS

Remove stop words

Hodgkin diseases,

Lowercase

hodgkin diseases,

Strip punctuation

hodgkin diseases

Uninflect

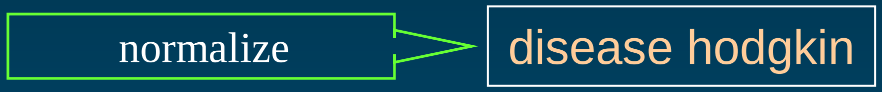
hodgkin disease

Sort words

disease hodgkin

Normalization: Example

Hodgkin Disease
HODGKINS DISEASE
Hodgkin's Disease
Disease, Hodgkin's
Hodgkin's, disease
HODGKIN'S DISEASE
Hodgkin's disease
Hodgkins Disease
Hodgkin's disease NOS
Hodgkin's disease, NOS
Disease, Hodgkins
Diseases, Hodgkins
Hodgkins Diseases
Hodgkins disease
hodgkin's disease
Disease, Hodgkin



normalize

disease hodgkin

Normalization Applications

- Model for lexical resemblance
- Help find lexical variants for a term
  - Terms that normalize the same usually share the same LUI
- Help find candidates to synonymy among terms
- Help map input terms to UMLS concepts

Indexes

- Word index
  - word to Metathesaurus strings
  - one word index per language
- Normalized word index
  - normalized word to Metathesaurus strings
  - English only
- Normalized string index
  - normalized term to Metathesaurus strings
  - English only

Lexical Variant Generation program

- Tool for specialists (linguists)
- Performs atomic lexical transformations
  - generating inflectional variants
  - lowercase
  - …
- Performs sequences of atomic transformations
  - a specialized sequence of transformations provides the normalized form of a term (the *norm* program)

Identifying relations among biomedical terms

- Adjectival modification
- Reification of *part-of* relations

Adjectival modification

Bodenreider O, Burgun A.
Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS.
Terminologie & Intelligence Artificielle; 2001. p. 11-21.

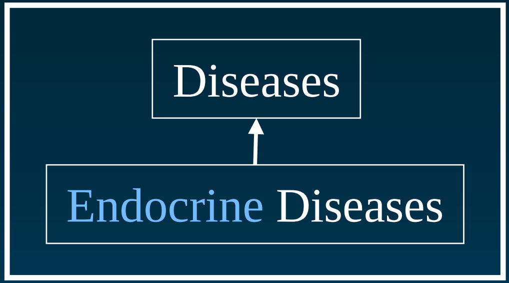
Objective

- Compare
  - Lexically-suggested hyponymic relations among medical terms
  - Inter-concept relationships represented in the UMLS
- Motivation
  - Not systematically represented
  - Some relationships are inaccurately hierarchical

➡ Compare hierarchical relations represented in the UMLS to hyponymic relations acquired independently

Acquiring hyponymic relations

- Adjectival modification generally induces hyponymy



Diseases

Endocrine Diseases

- Removing modifiers from a term should produce a term in hypernymic relation (*isa*)
- This relation should be recorded in the Metathesaurus

Material

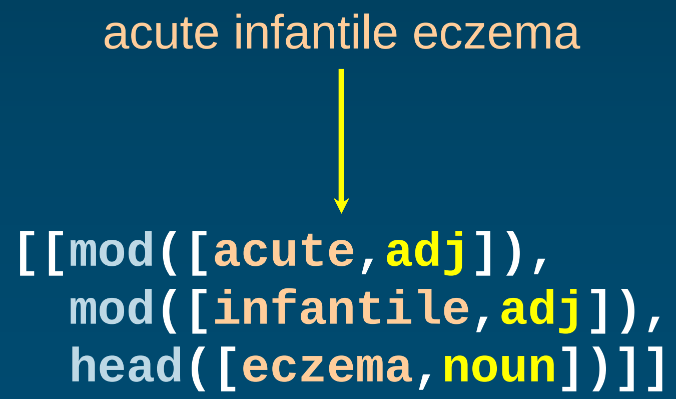
- SNOMED International
- Significant subset of the clinical domain
  - Diseases
  - Procedures
- Filtered out terms containing a comma
  - Permuted terms
  - Complex terms
- 63,000 SNOMED terms
- 42,000 UMLS concepts

Methods Overview

- Syntactic analysis to identify adjectival modifiers
- Generate transformed terms by removing adjectival modifiers
- Map transformed terms to the UMLS
- Study the relationship between original term and transformed term in the UMLS, if any

Identify adjectival modifiers

- Underspecified syntactic analysis
  - Xerox part of speech tagger
  - SPECIALIST Lexicon (UMLS)
- Modifiers used: adjectives (+ adverbs)
- Modifiers identified in 64% of the terms
- Usually 1 to 2 modifiers
- Unique modifiers
  - 5400 adjectives
  - 69 adverbs



acute infantile eczema

```
[[mod([acute,adj]),
  mod([infantile,adj]),
  head([eczema,noun])]]
```

Transforming terms

- Remove any combination of modifiers found in the original term
- 2^n-1 transformed terms
  when the original term has n modifiers
- 104,000 transformed terms generated

acute infantile eczema

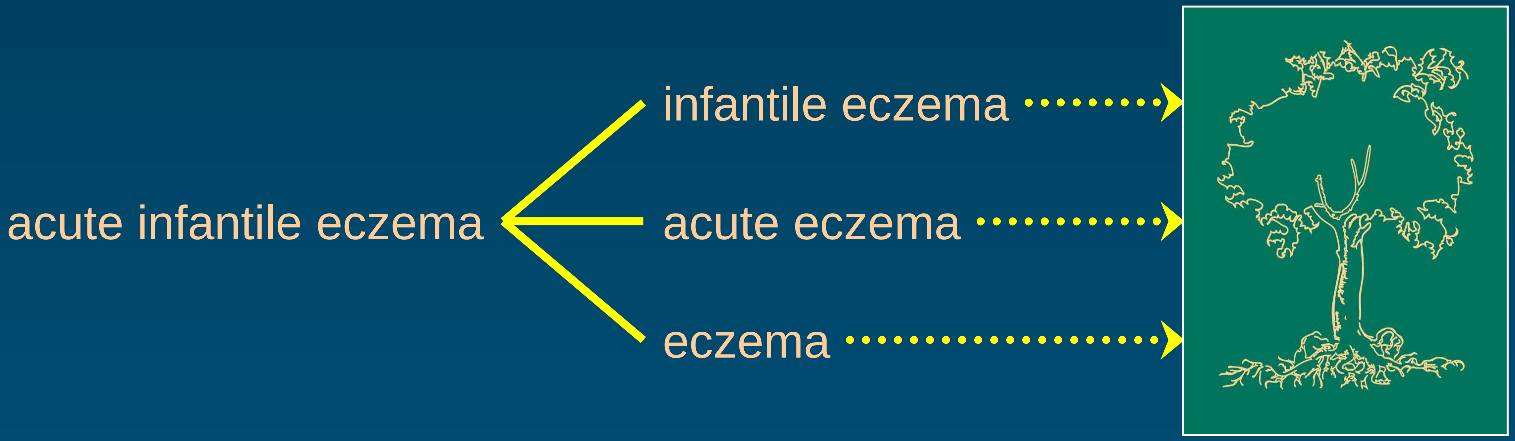
infantile eczema

acute eczema

eczema

Mapping transformed terms to UMLS

- Increasing aggressiveness
  - Exact match
  - After normalization
- 25% of the transformed terms successfully mapped to UMLS



acute infantile eczema

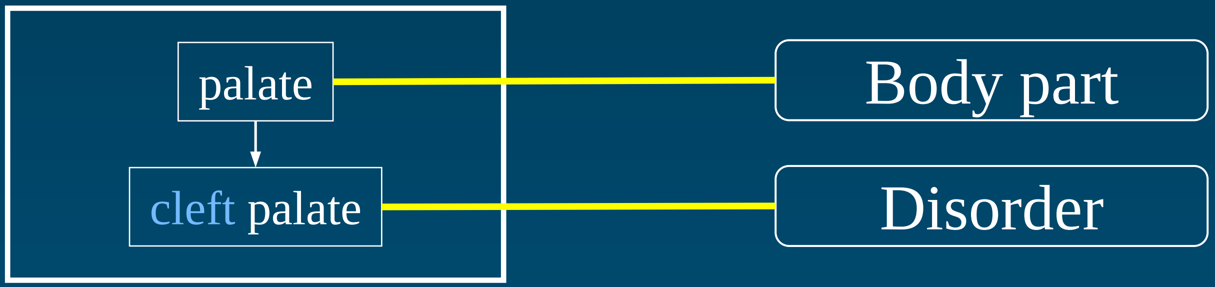
infantile eczema

acute eczema

eczema

Excluding non-hyponymic relations

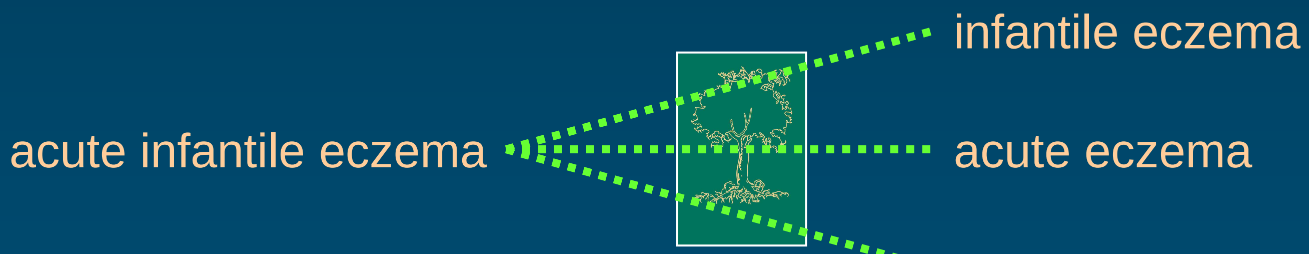
- If in hyponymic relation, original term and the transformed term should have the same semantic type (both Disease or both Procedure)
- Different semantic types in 10%



Checking relationships against UMLS

- Original term (OT) Transformed term (TT)
  - Synonyms (same concept)
  - TT ancestor of OT
  - Siblings
  - Otherwise related

  (inter-concept relationship)



acute infantile eczema

infantile eczema

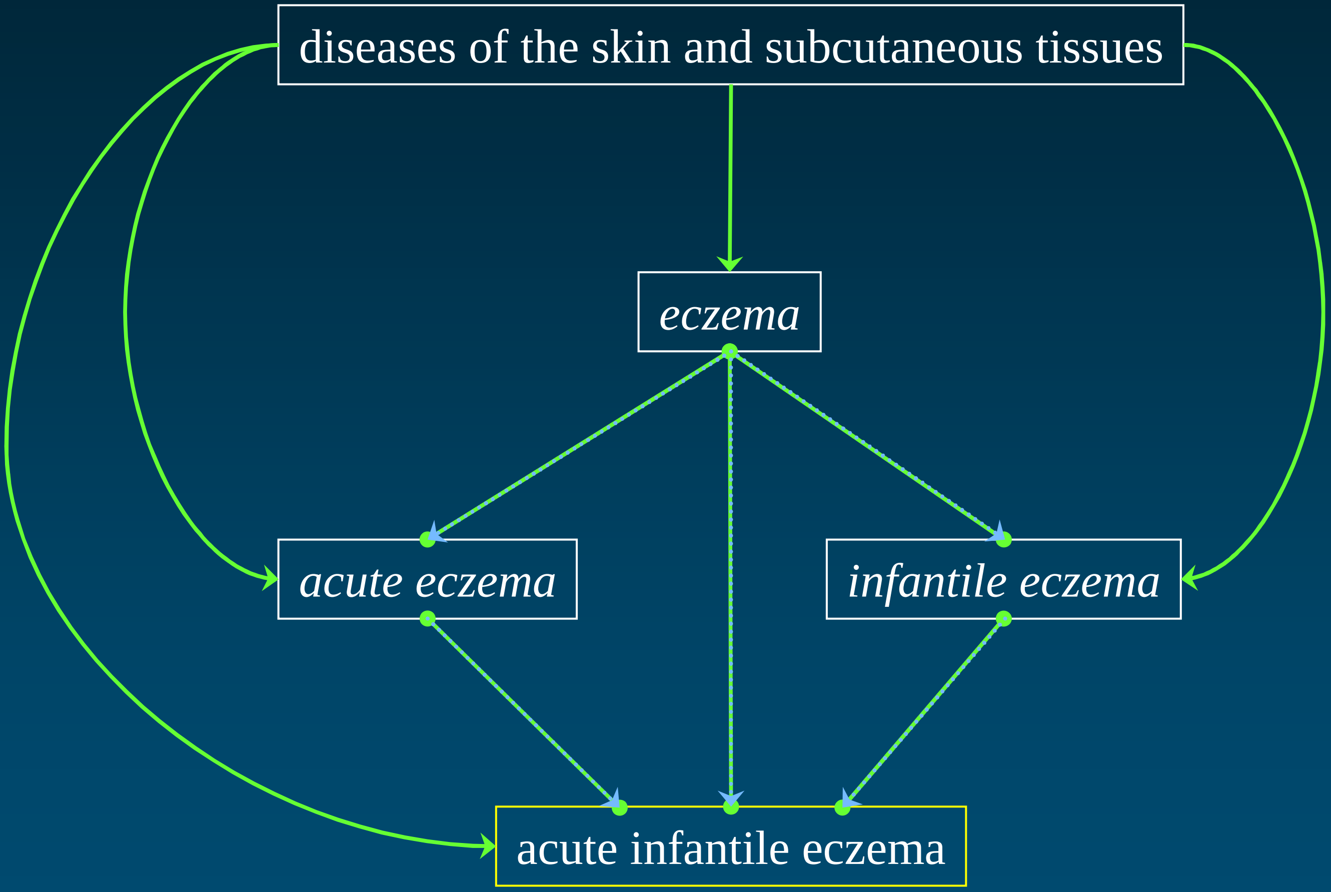
acute eczema

eczema

Lexically-suggested relationships / UMLS

- 28,851 pairs of terms
  - Original SNOMED term
  - Transformed term (found in UMLS)
- Corresponding relationship in the Metathesaurus
  - Hierarchical in 50% of the cases
  - « Sibling » in 25% of the cases
  - Missing in 25% of the cases

Lack of structure within a source

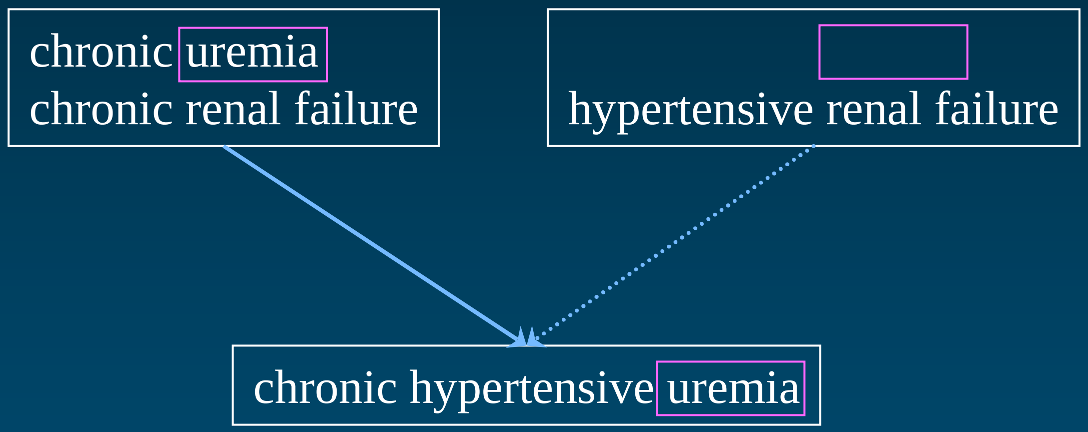


Plesionymy

posttransfusion hepatitis

posttransfusion viral hepatitis

Missing synonymy



chronic uremia
chronic renal failure

hypertensive renal failure

chronic hypertensive uremia

Reification of *part-of* relations

Zhang S, Bodenreider O.
Aligning representations of anatomy using lexical and structural methods.
AMIA; 2003. p. 753-757.

Two representations of anatomy

- FMA
  - Foundational Model of Anatomy
  - University of Washington, 1994
  - Conceptualization of the physical objects and spaces that constitute the human body

- GALEN common reference model
  - Generalized Architecture for Languages, Encyclopaedias and Nomenclatures in medicine
  - University of Manchester, 1991
  - Development of a compositional and generative formal system for modeling all and only sensible medical concepts

Aligning steps

Lexical alignment

- Step 1: Acquiring terms
- Step 2: Identifying anchors (i.e., shared concepts) lexically

Structural alignment

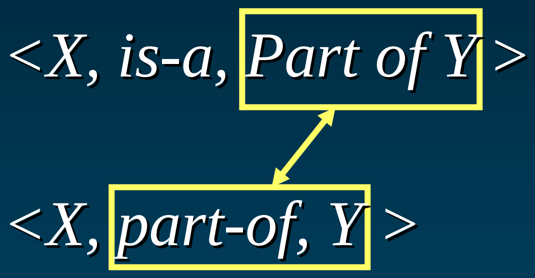
- Step 3: Acquiring (explicit and implicit) semantic relations
- Step 4: Identifying anchors structurally

Step 3: Acquiring semantic relations

- Semantic relations
  - *<concept$_1$, relationship, concept$_2$>*
  - Hierarchical relationships: *is-a* and *part-of*
    - *<Arm, part-of, Proximal segment of upper limb>*
- Extracting the explicit relations
- Acquiring implicit knowledge
  - Complementing missing inverse relations
  - Augmenting relations embedded in concept names
  - Inferring relations from a combination of relations

Implicit knowledge Reification

- Reification of *part-of* relationships



<X, is-a, Part of Y >

<X, part-of, Y >

- Augmenting reified *part-of* relations
  - Reified: <*Cardiac chamber, is-a, Subdivision of heart*>
  - No explicit (direct or indirect) *part-of* relationships between *Cardiac chamber* and *Heart* in FMA
  - Augmented: <*Cardiac chamber, part-of, Heart*>

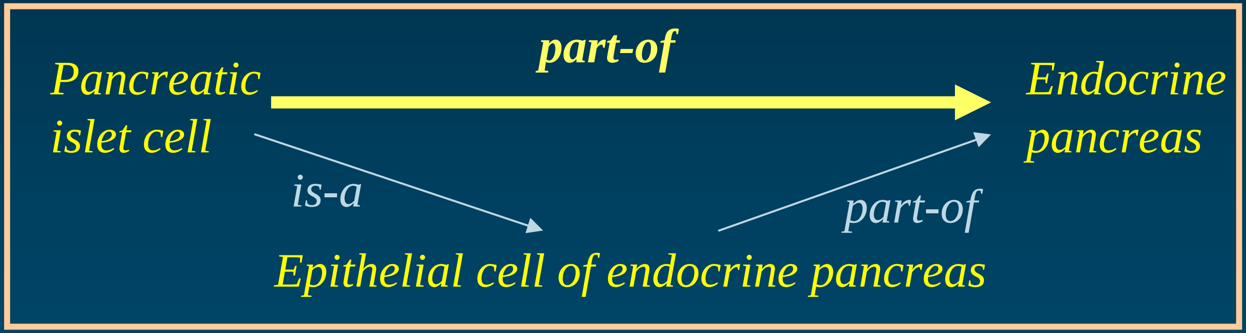
Implicit knowledge Others

- Noun-noun compounds (X Y)
  - X Y and Y exist as concepts
  - <*X Y*, *isa*, *Y*> generated
  - <*Sweat gland*, *isa*, *Gland*>
- Prepositional attachment with “of” (X of Y)
  - X and Y exist as concepts
  - <*X of Y*, *part-of*, *Y*> generated
  - <*Neck of femur*, *part-of*, *Femur*>

- No syntactic analysis
- Constraint by domain

Implicit knowledge Inferring

- Generating new inter-concept relationships by applying inference rules



Pancreatic islet cell

part-of

Endocrine pancreas

is-a

part-of

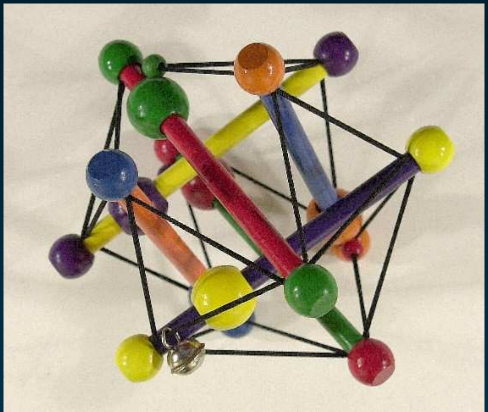
Epithelial cell of endocrine pancreas

Semantic relations acquired

Types of relations	FMA	GALEN
Explicitly represented	238,135	214,403
Complemented	104,754	107,689
Augmented	315,860	27,274
Inferred	5,172,668	1,661,824
Total	5,831,417	2,011,190

Explicit vs. implicit knowledge

- More positive structural evidence found for anchors
- Augmentation accounted for 74% of 523 anchors acquiring positive evidence
- More conflicting relations found for anchors



Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

NATIONAL LIBRARY OF MEDICINE

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

References UMLS

- UMLS
  `umlsinfo.nlm.nih.gov`
- UMLS browser
  - Knowledge Source Server: **`umlsks.nlm.nih.gov`**
  - Semantic Navigator: **`http://mor.nlm.nih.gov/perl/semnav.pl`**
  - (free, but UMLS license required)
- UMLS and information integration
  - O. Bodenreider. The UMLS: Integrating biomedical terminology. *Nucl. Acids Res. 2004;32(1) (in press)*