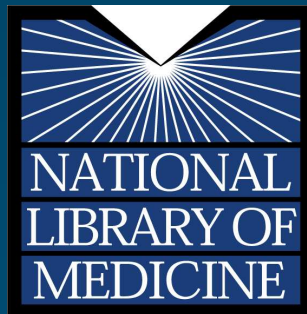


Laboratoire d'Informatique de Paris-Nord
December 7, 2004

The Unified Medical Language System

Identifying relations among biomedical terms



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



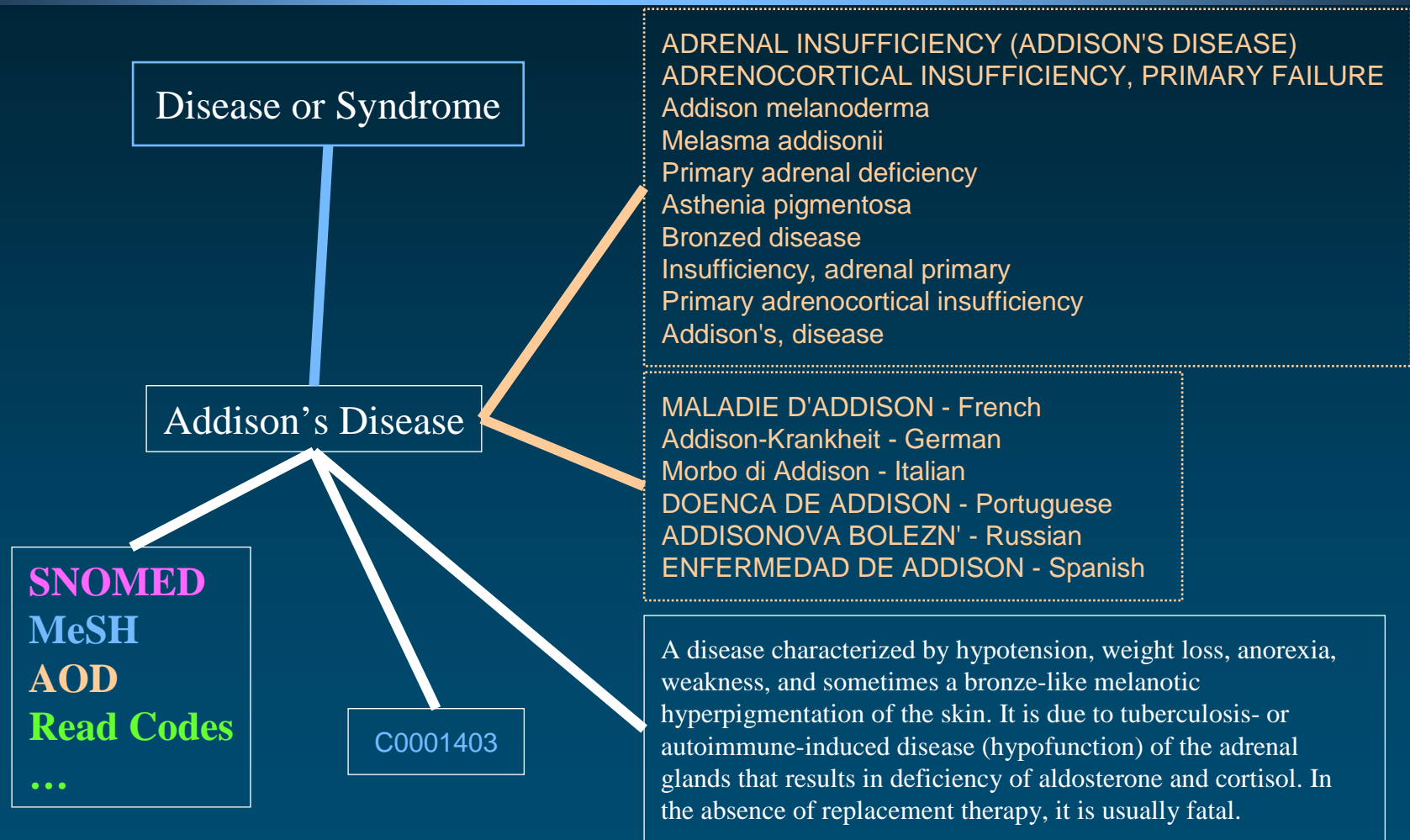
Source Vocabularies

(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: IIR-IV)
- ◆ Broad coverage of biomedicine
- ◆ Common presentation



Addison's Disease: Concept



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	<p>S0011232 <i>Adrenal Gland Diseases</i></p> <p>S0011231 Adrenal Gland Disease</p> <p>S0000441 Disease of adrenal gland [...]</p> <p>S0481705 Disease of adrenal gland, NOS</p> <p>S0220090 Disease, adrenal gland</p> <p>S0044801 Gland Disease, Adrenal</p>
Term L0041793	<p>S0860744 <i>Disorder of adrenal gland, unspecified</i></p> <p>S0217833 Unspecified disorder of adrenal glands</p>
Term L0161347	<p>S0225481 <i>ADRENAL DISORDER</i> [...]</p> <p>S0627685 DISORDER ADRENAL (NOS)</p>
Term L0181041	<p>S0632950 <i>Disorder of adrenal gland</i> [...]</p> <p>S0354509 Adrenal Gland Disorders</p>
Term L0368399	<p>S0586222 <i>Adrenal disease</i> [...]</p> <p>S0466921 ADRENAL DISEASE, NOS</p>
Term L1279026	<p>S1520972 <i>Nebennierenkrankheiten</i> GER</p>
Term L0162317	<p>S0226798 <i>SURRENALE, MALADIES</i> FRE [...]</p>



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



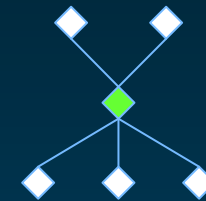
Symbolic relationships Type

- ◆ Hierarchical

- Parent / Child
- Broader / Narrower than

PAR / CHD

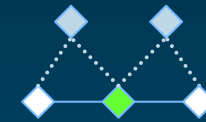
RB / RN



- ◆ Derived from hierarchies

- Siblings (children of parents)

SIB



- ◆ Associative

- Other

RO



- ◆ Various flavors of near-synonymy

- Similar
- Source asserted synonymy
- Possible synonymy

RL

SY

RQ

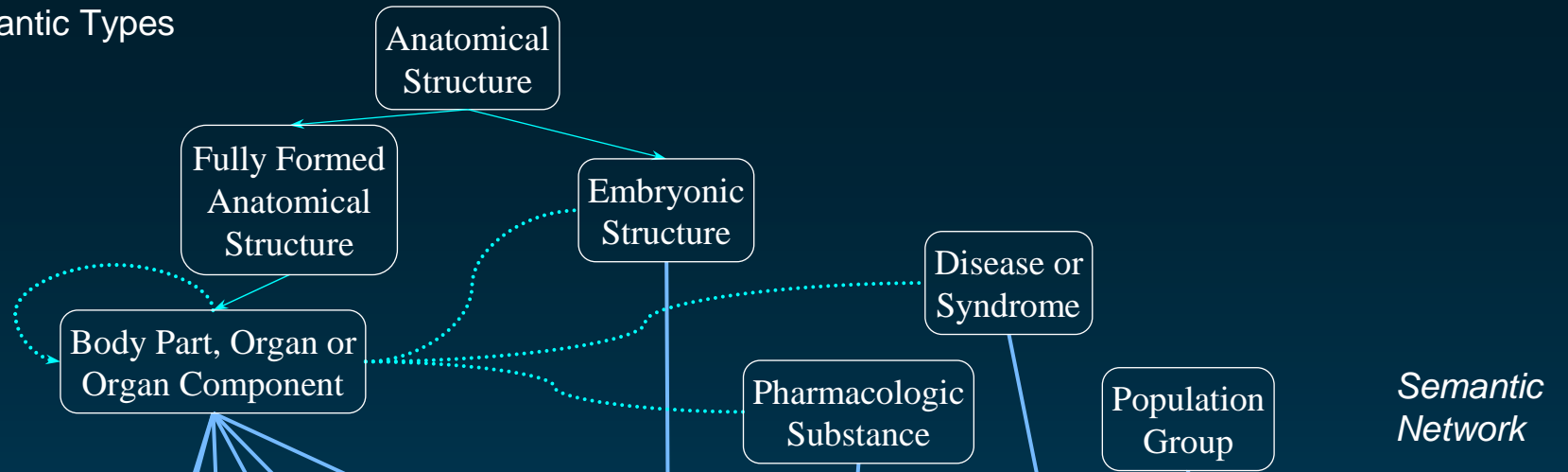


Symbolic relationships Attribute

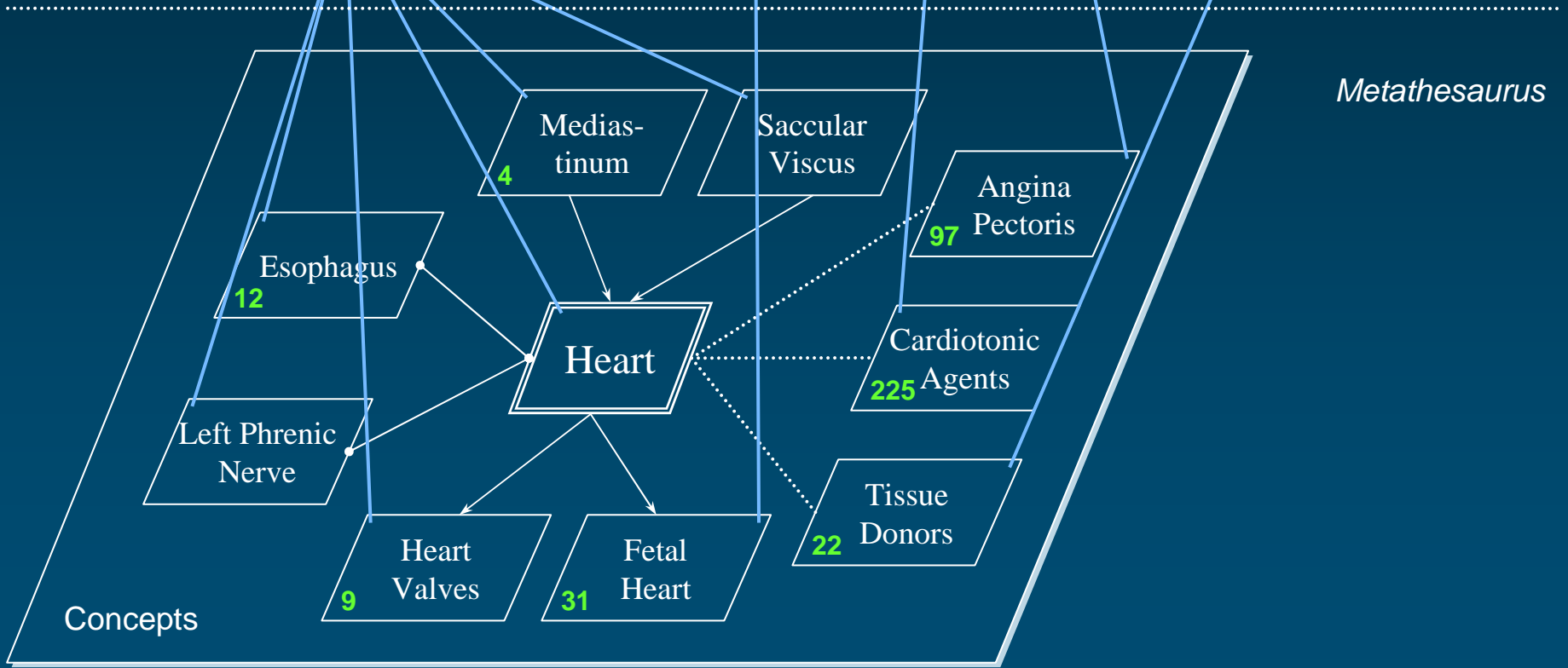
- ◆ Hierarchical
 - isa (is-a-kind-of)
 - part-of
- ◆ Associative
 - location-of
 - caused-by
 - treats
 - ...
- ◆ Cross-references (mapping)



Semantic Types



Semantic Network



Metathesaurus

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



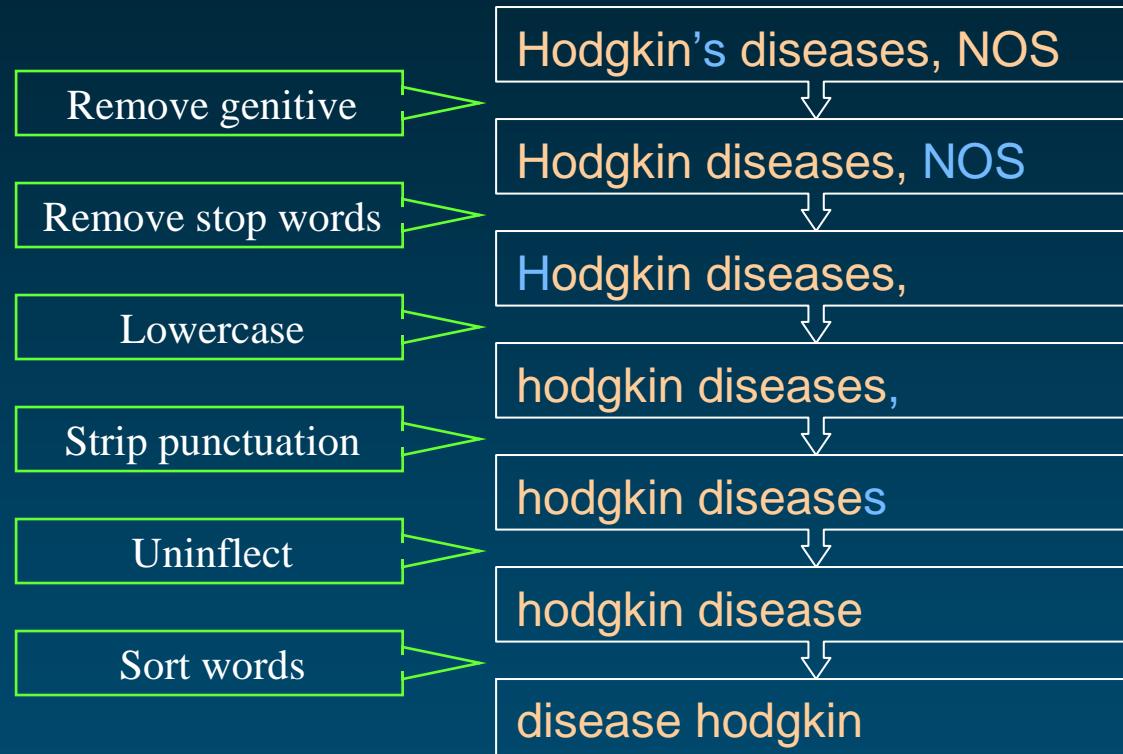
Lexical tools

- ◆ 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



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

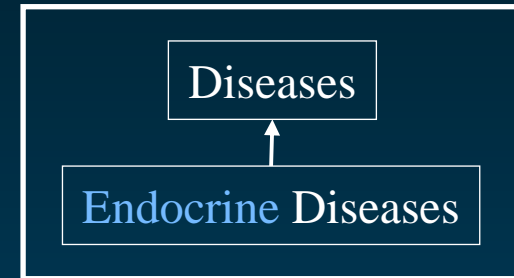


Compare hierarchical relations represented in the UMLS to hyponymic relations acquired independently



Acquiring hyponymic relations

- ◆ Adjectival modification generally induces hyponymy



- ◆ 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

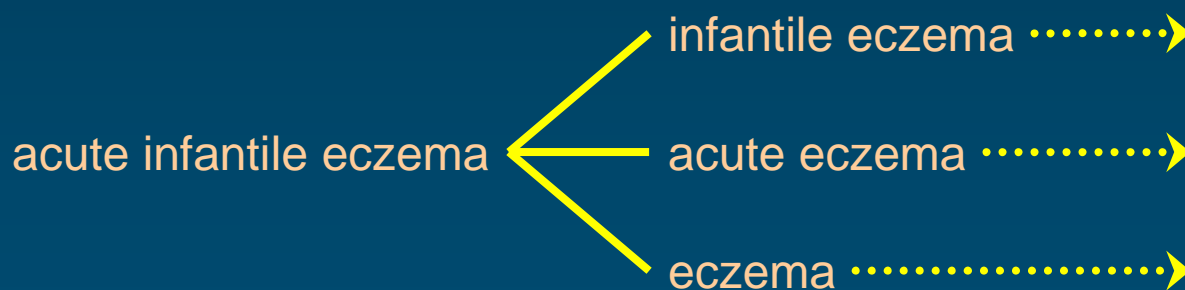
acute eczema

eczema



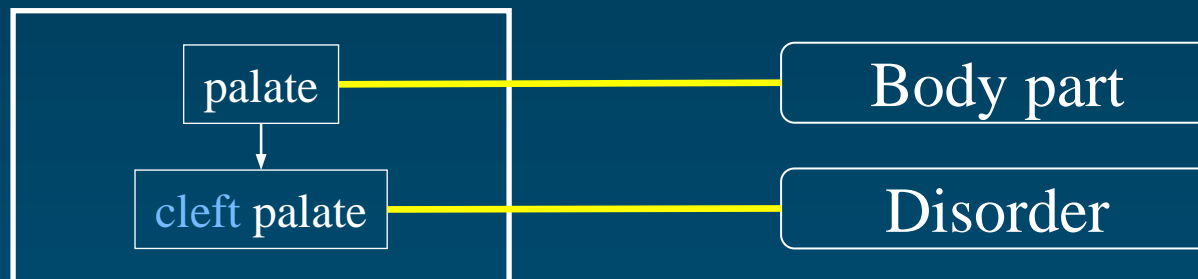
Mapping transformed terms to UMLS

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



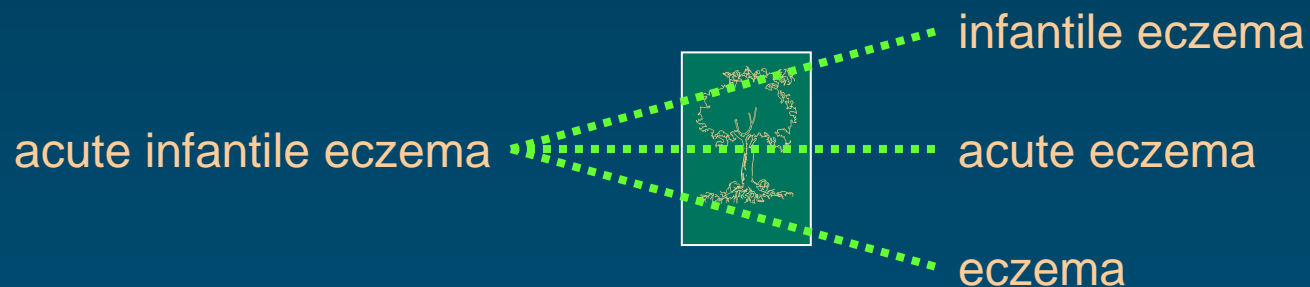
Excluding non-hyponymic relations

- ◆ 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 (inter-concept relationship)
 - Siblings (inter-concept relationship)
 - Otherwise related (inter-concept relationship)

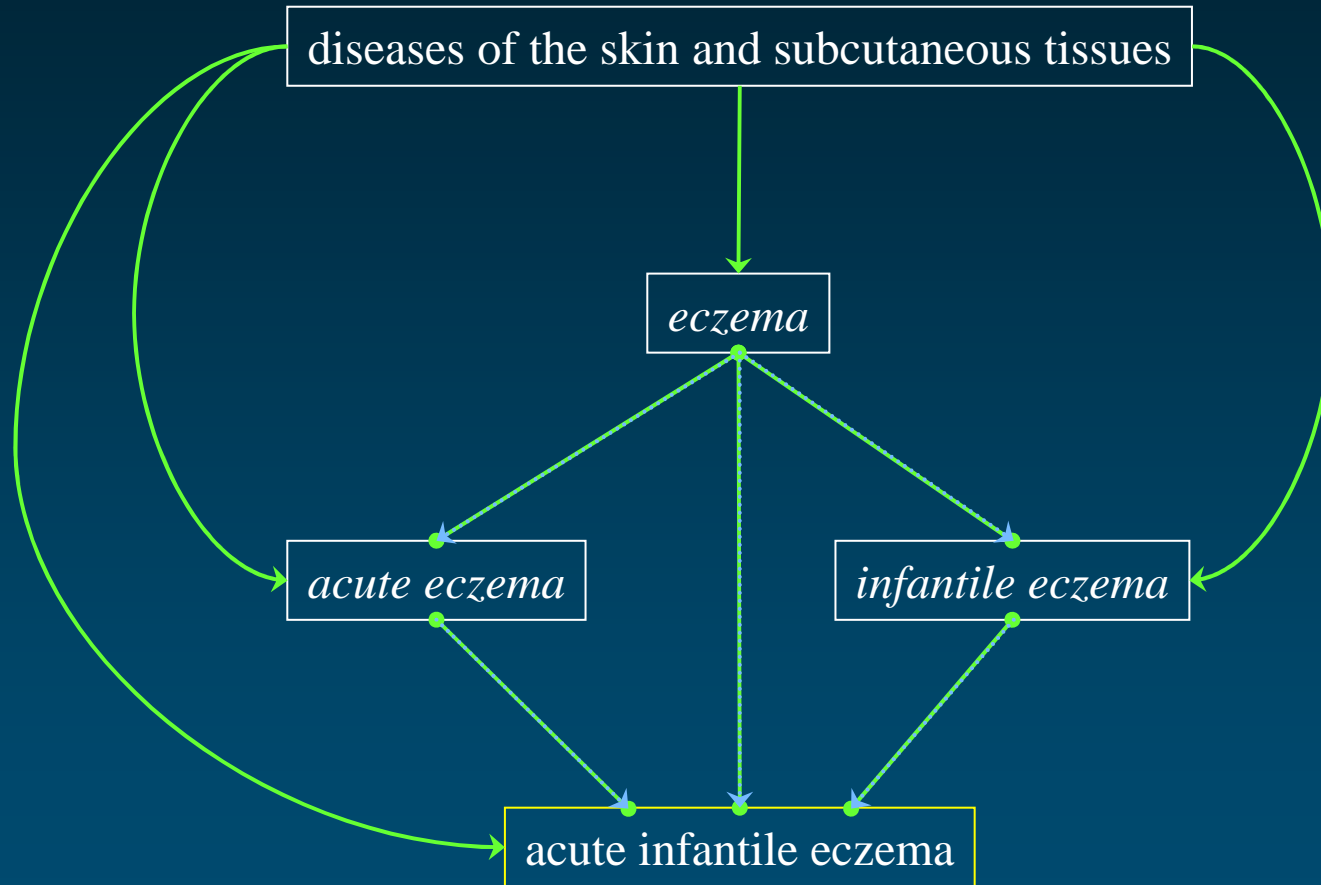


Lexically-suggested relationships / UMLS

- ◆ 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

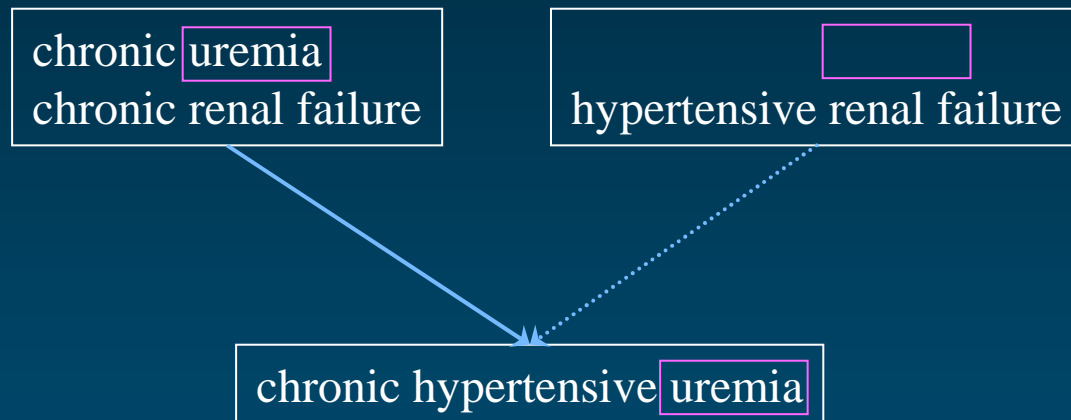


Plesionymy

posttransfusion hepatitis
posttransfusion viral hepatitis



Missing synonymy



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



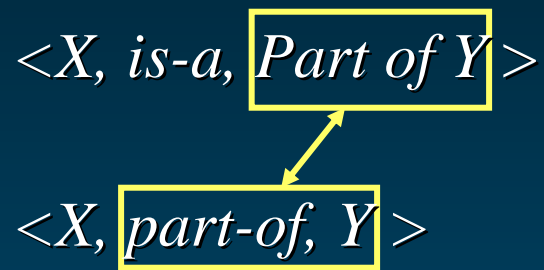
Step 3: Acquiring semantic relations

- ◆ Semantic relations
 - $\langle \text{concept}_1, \text{relationship}, \text{concept}_2 \rangle$
 - Hierarchical relationships: *is-a* and *part-of*
 - $\langle \text{Arm}, \text{part-of}, \text{Proximal segment of upper limb} \rangle$
- ◆ 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



◆ Augmenting reified *part-of* relations

- Reified: $\langle \text{Cardiac chamber}, \text{is-a}, \text{Subdivision of heart} \rangle$
- No explicit (direct or indirect) *part-of* relationships between *Cardiac chamber* and *Heart* in FMA
- Augmented: $\langle \text{Cardiac chamber}, \text{part-of}, \text{Heart} \rangle$



Implicit knowledge Others

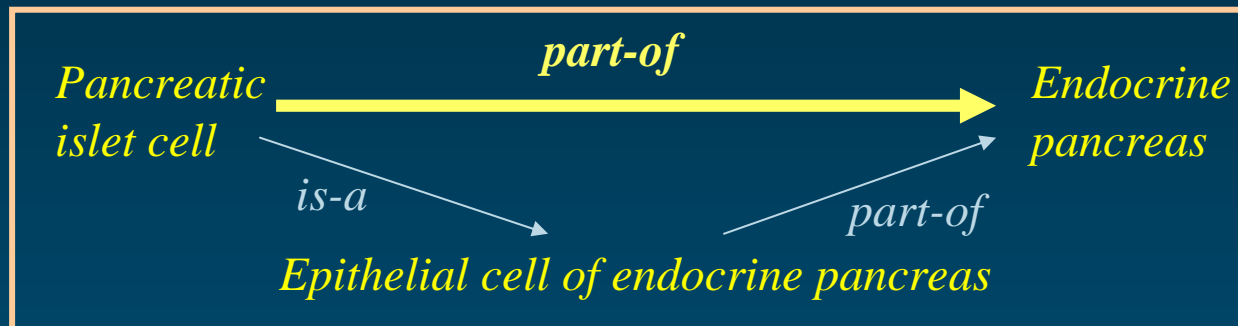
- ◆ Noun-noun compounds (X Y)
 - X Y and Y exist as concepts
 - $\langle X Y, isa, Y \rangle$ generated
 - $\langle Sweat\ gland, isa, Gland \rangle$
- ◆ Prepositional attachment with “of” (X of Y)
 - X and Y exist as concepts
 - $\langle X\ of\ Y, part-of, Y \rangle$ generated
 - $\langle Neck\ of\ femur, part-of, Femur \rangle$

- No syntactic analysis
- Constraint by domain



Implicit knowledge Inferring

- ◆ Generating new inter-concept relationships by applying inference rules



Semantic relations acquired

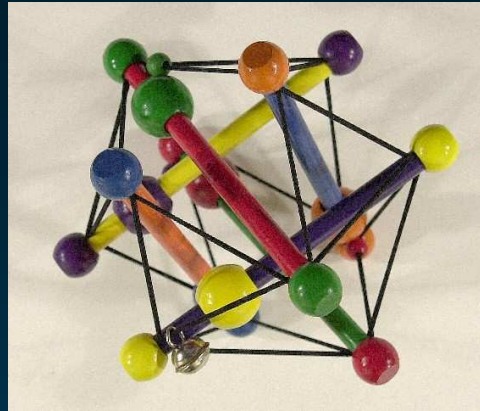
Types of relations	FMA	GALEN
Explicitly represented	238,135	214,403
Complemented	104,754	107,689
Augmented	315,860	27,274
Inferred	5,172,668	1,661,824
Total	5,831,417	2,011,190



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

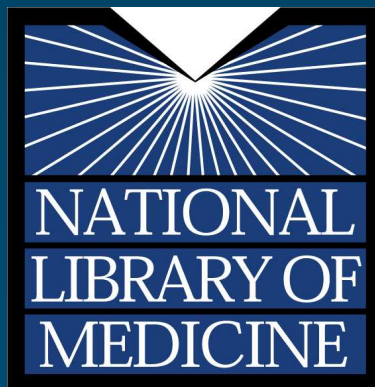




Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: mor.nlm.nih.gov



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*)

