# Advanced NLP

Slides from Josh Denny, Adi Bejan, and Robert Carroll

# Study of Natural Language

- Human language (vs. formal and computer language)
- Linguistics - a description of language - used by theoretical linguists.
- Psycholinguistics - a cognitive model of how people understand and generate language.
- Computational linguistics - build computational models to understand and generate language.

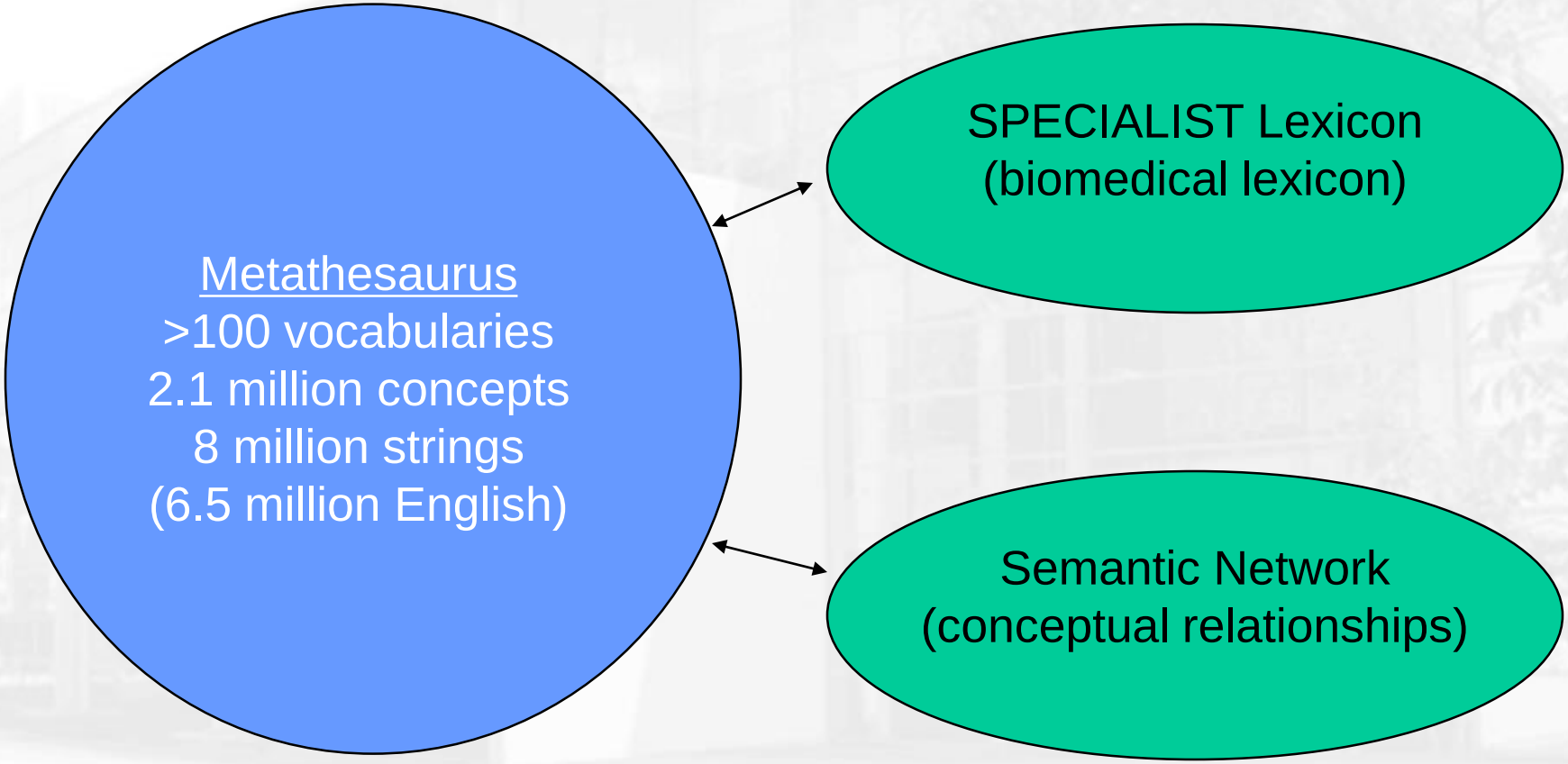
# Overview of Linguistic Levels

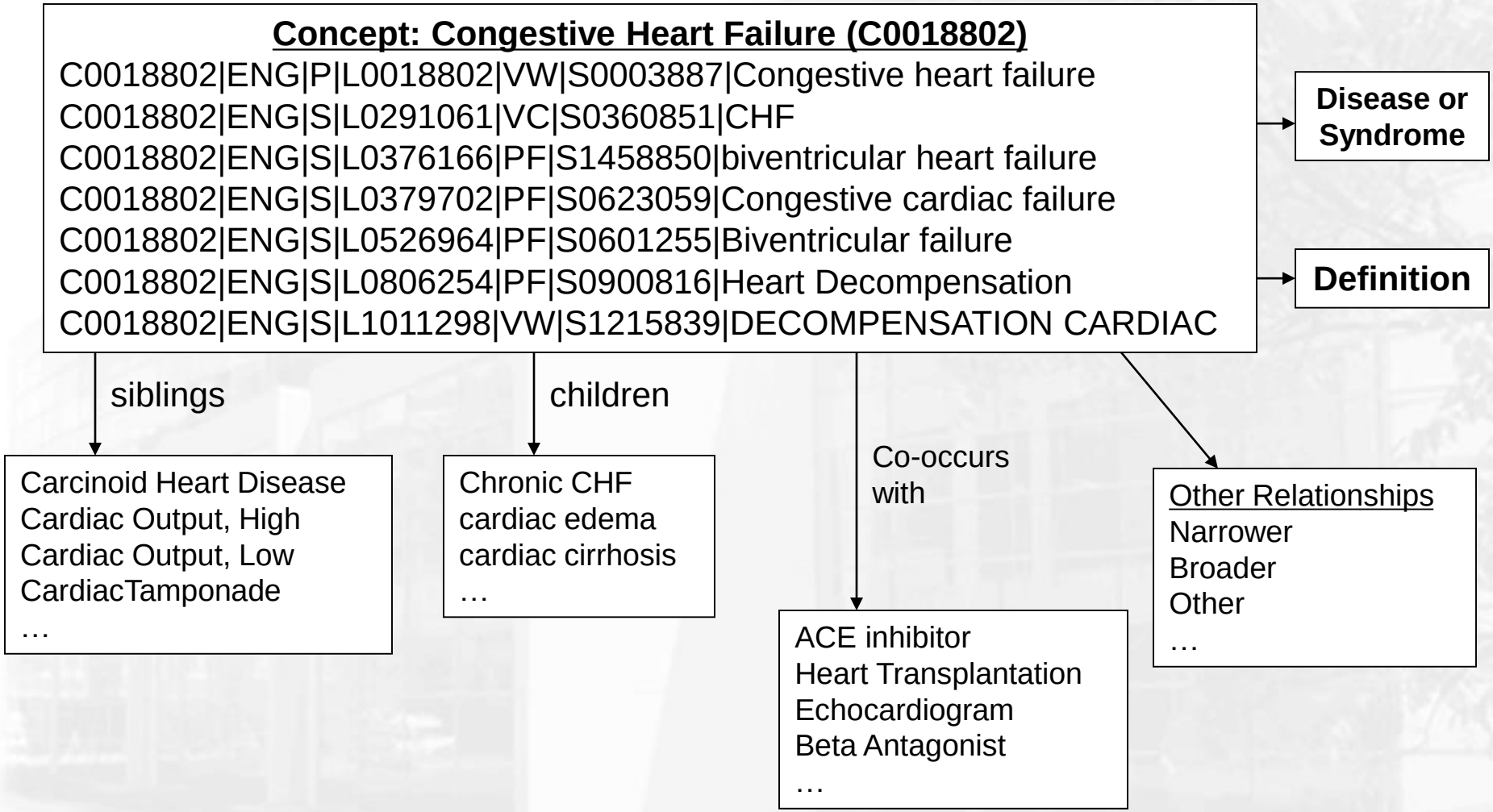
- **Phonology**: units of sound combine to produce words (will not cover)
- **Morphology:** basic units combine to produce words
- **Lexicography**: syntactic (part of speech) and semantic categories of words
- **Syntax:** structures combine to produce sentences
- **Semantics:** meaning/interpretations combine
- **Pragmatic:** context affects the interpretation of meaning
- **Discourse** – previous information affects the interpretation of the current information
- **World knowledge** – knowledge about the world affects interpretation

# Controlled Vocabularies

- Unified Medical Language System (UMLS)
  - Contains >100 individual vocabularies
    - Incl. SNOMED, LOINC, RxNorm… and many others
  - Sometimes "messy" and imprecise, but most "robust" coverage
- SNOMED-CT
  - current "gold standard" for medical problems and signs/symptoms
  - Well-defined, curated relationships
- LOINC – "gold standard" for labs and tests
- RxNorm – currently most robust freely available drug database

# Unified Medical Language System (UMLS)




Concept: Congestive Heart Failure (C0018802)
C0018802|ENG|P|L0018802|VW|S0003887|Congestive heart failure
C0018802|ENG|S|L0291061|VC|S0360851|CHF
C0018802|ENG|S|L0376166|PF|S1458850|biventricular heart failure
C0018802|ENG|S|L0379702|PF|S0623059|Congestive cardiac failure
C0018802|ENG|S|L0526964|PF|S0601255|Biventricular failure
C0018802|ENG|S|L0806254|PF|S0900816|Heart Decompensation
C0018802|ENG|S|L1011298|VW|S1215839|DECOMPENSATION CARDIAC
Disease or Syndrome
Definition
siblings
children
Co-occurs with
Other Relationships
Narrower
Broader
Other
…
Carcinoid Heart Disease
Cardiac Output, High
Cardiac Output, Low
CardiacTamponade
…
Chronic CHF
cardiac edema
cardiac cirrhosis
…
ACE inhibitor
Heart Transplantation
Echocardiogram
Beta Antagonist
…

# "Major tasks" in clinical NLP

- Concept identification
- Negation detection
- Section identification
- Word sense disambiguation
- Temporal resolution

# Concept recognition (KM)

**Documents**

KM Sentence Parser

POS identification

Noun phrase identification, approximate NLP

Normalization and variant generation

Get concept candidates from UMLS

Find best matches using probabilities, linguistic and semantic rules, and other document concepts

# Lexical Tools

- KM Lexicon creation
  - Used SPECIALIST word inflections to map each source word to a unique base form for normalization
    - Generated all period, possessive, and plural variants of words and acronyms
    - Hyphens normalized to non-hyphenated form (if present)
  - > 200,000 single word forms
- Developed a list of >2000 unique prefixes, roots, and terminals

# POS tagging

- Necessary to parse a sentence
- KM originally used a library from a company (rule-based)
- Now uses freely available Perl module (HMM)
- Dictionary for both the company module and Perl module have been customized
- UIMA and GATE have freely available POS tagging modules

- Is a word is not known, what should be the default POS?

# Processing of the UMLS

- Used KM normalization routines to normalize strings
- Removed strings deemed “not useful” for NLP:
  1. Suppressible synonyms
  2. Non-matching strings if containing >1 word
  3. Strings longer than 6 non-stopwords
  4. Strings beginning with “other,” or containing variants of “without mention” or “not elsewhere classified”
  5. Certain semantic types (i.e., all of “Clinical Drug” and any “Medical Device” containing a number followed by a unit (e.g., “Silicone foam 20g dressing”))

# "Form rules"

- Manually created 144 "form rules" to convert non-KM-base-form word variants to related KM base-forms
  - Included part-of-speech conversions, common Latin declensions, and Greek variants

| **Rule:** | | | | **Example:** |
|---|---|---|---|---|
| -ae | NO | -a | NO | fimbriae (NO) => fimbria (NO) |
| -as | NO | -atic | AJ | pancreas (NO) => pancreatic (AJ) |
| -nce | NO | -nt | AJ | absence (NO) => absent (AJ) |

# Processing of the UMLS

- Used KM normalization routines to normalize strings
- Removed strings deemed “not useful” for NLP
  - e.g., concepts containing non-matching words, concepts longer than 6 non-stopwords
- Created inverted indexes for words to concepts

# Processing of the UMLS

- Acronym database
  - Extracted acronyms from UMLS
  - Scanned document corpus for other acronyms
- Derived Semantic Type (DSTY)
  - Used for "semantic regularization"
  - Includes semantic types from:
    - MRCON
    - 12,000 additional single words and 90,000 new multiwords extracted from MRCON
    - Some manually defined entries

# Document Analysis

- Sentence identification
  - Removal of outline headings
  - Elimination of newline characters, tabs, and multiple spaces from within identified sentences
- Noun phrase identification
  - A simple noun phrases must contain at least an adjective or a noun but can also include past participles, gerunds, or adverbs
  - Linkages between noun phrases are identified

# Noun phrase identification

**Sentence:**

The patient's heart and liver are enlarged.

**Noun Phrase Analysis:**

NP1 ("patient", link_type="Prep(of)", link_to=NP1 )

NP2 ("heart", link_type="Conj(and)", link_to=NP2 )

NP3 ("liver", link_type="LinkVerb(are)", link_to=NP3)

NP4 ("enlarged")

# Semantic Regularization

- Conjunctive regularization is the process of determining proper linkages between noun phrases joined by conjunctions
- Performed if DSTY identical
- Also involves distribution of modifying adjectives
  - "aortic and mitral valve" → "aortic valve" and "mitral valve" not "aortic" and "mitral valve"

# Concept Identification

- Simple noun phrases
  - Sequentially intersect possible Metathesaurus entries for each word
  - Acronyms found in the document are immediately expanded
  - If intersection yields a null set:
    - semantic and derivational word variants are generated for each word
    - Certain parts-of-speech are eliminated

# Concept identification

- Composite noun phrases
  - Noun phrases must match completely
  - Linked noun phrases are intersected by priority of linkage
    - “heart failure from stenosis of aortic valve”
  - Word variants and acronyms are expanded

# Disambiguation

- Semantic type information
- Exacted-matched concepts in documents
  - “beta receptor” is more likely to be “beta adrenergic receptor” instead of “beta C receptor” if KM matched former in document
- Words appearing together in a document
  - “donor screening” probably means “blood donor screening” if “blood donor” is a concept in the document
- MEDLINE co-occurences of concepts

I. **Monoamine Oxidase Inhibitors**

A. **History**

1. In 1951 isoniazid and iproniazid were developed for the treatment of tuberculosis. It was observed serendipitously that TB patients treated with these compounds developed an unexpected elevation of mood including mania. Clinical trials followed and by the mid-1950's it was established that iproniazid had significant, sustained antidepressant effects.

2. After introduction, the monoamine oxidase inhibitor drugs gained wide acceptance. Problem side effects, including hypertensive reactions (see below) began to be recognized by the early 1960's and the introduction of the tricyclic antidepressants lead to diminished use of the MAOIs.

# Concept Categorization

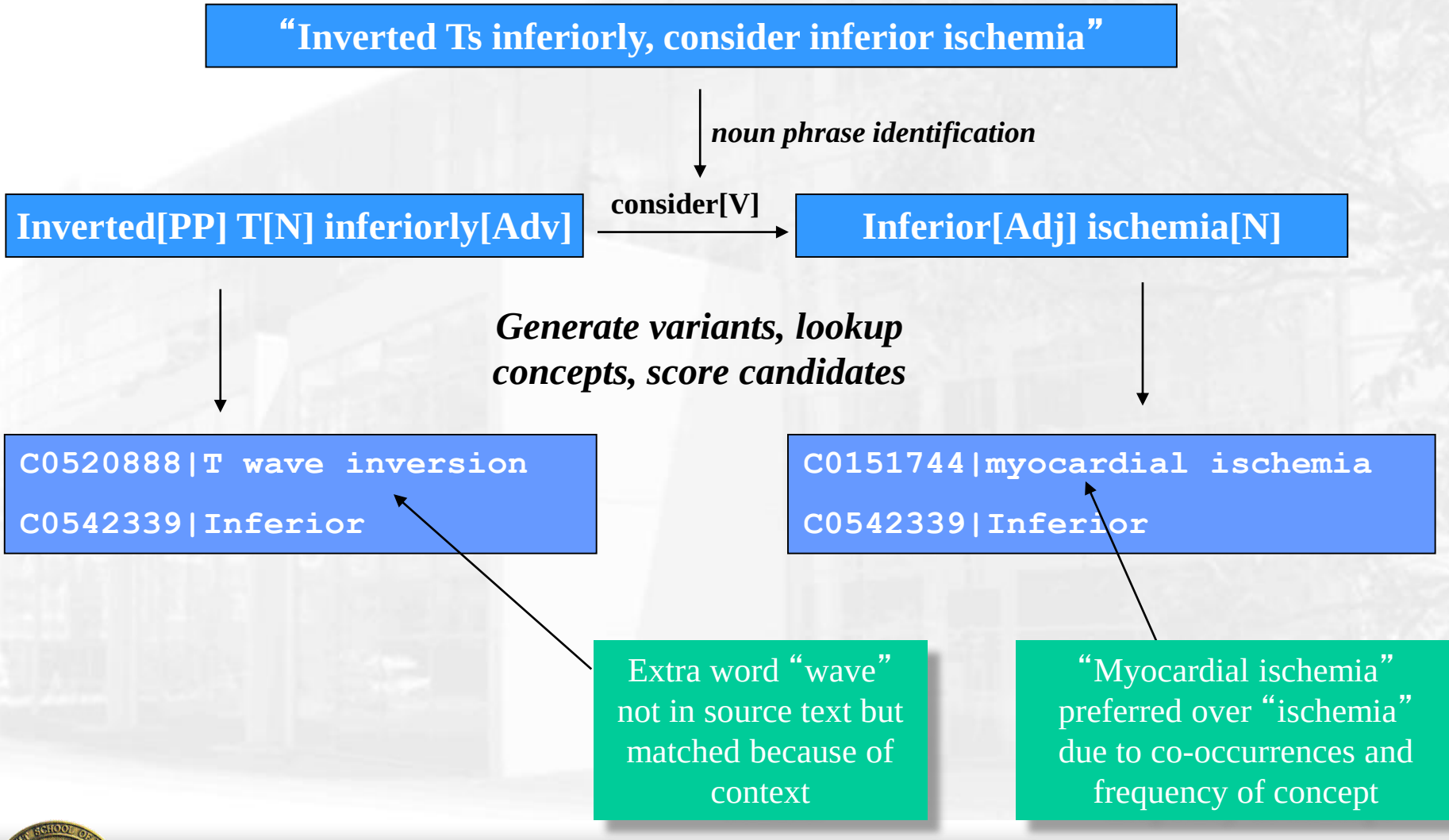
- "meaningful term" vs. "composite phrase"
  - *Meaningful term* – e.g. "heart", "lung", "congestive heart failure", "Stevens Johnson Syndrome"
  - *Composite phrase* – e.g. "elevation of blood pressure", "carcinoma of the head of the pancreas", "amino acid precursor"
- *A priori* categorization

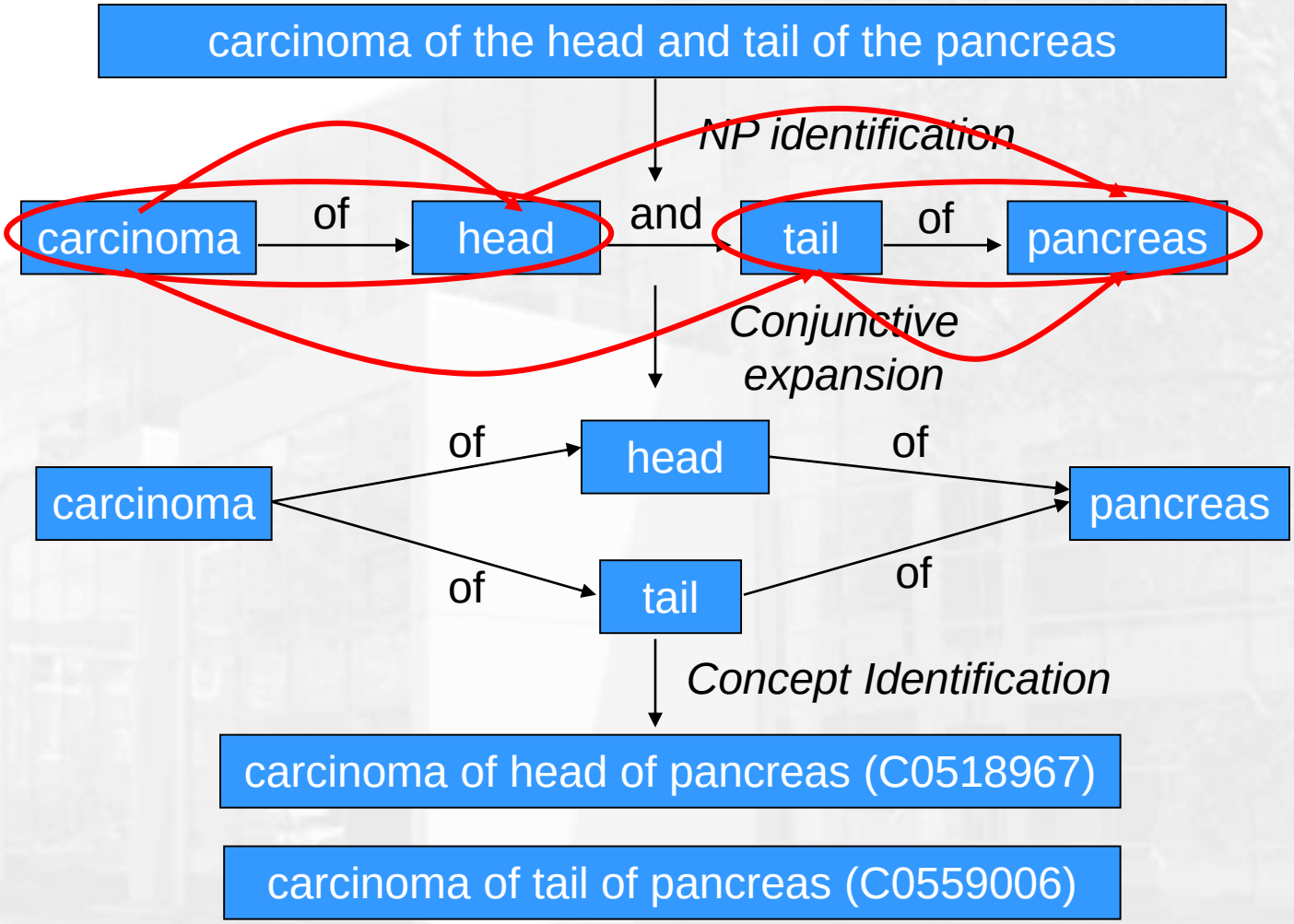
# Electrocardiograms – a limited domain still requiring NLP

- Atrial fibrillation expressed in 170 different ways

```
underlying rhythm is atrial fib
probable atrial fibrillation
afib has replaced sinus rhythm
a fib/ flutter with ventricular rate
probably afibrilation with rbbb
atr. fibrillation, transient
underlying rhythm is a-fib
af is new
```

- Over 200 different “negation phrases”
  - “?”, “no evidence of”, “possible”
- Accuracy of automated image analysis (to produce codified results) is poor


"Inverted Ts inferiorly, consider inferior ischemia"
noun phrase identification
Inverted[PP] T[N] inferiorly[Adv]
consider[V]
Inferior[Adj] ischemia[N]
Generate variants, lookup concepts, score candidates
C0520888|T wave inversion
C0542339|Inferior
C0151744|myocardial ischemia
C0542339|Inferior
Extra word "wave" not in source text but matched because of context
"Myocardial ischemia" preferred over "ischemia" due to co-occurrences and frequency of concept


carcinoma of the head and tail of the pancreas
NP identification
carcinoma
of
head
and
tail
of
pancreas
Conjunctive expansion
carcinoma
of
head
of
pancreas
of
tail
of
Concept Identification
carcinoma of head of pancreas (C0518967)
carcinoma of tail of pancreas (C0559006)

# Scoring Process

**Input Sentence #1:** I . Drugs Used to Treat gram - Negative Infections

| A | | B | |
|---|---|---|---|
| | C0013227:Drugs (Pharmaceutical Preparations) [Pharmacologic Substance]<br><br>C0042153:Using (utilization) [Quantitative Concept]<br><br>C0439208:Gram (Gram) [Quantitative Concept]<br>C0205160:Negative (Negative) [Finding]<br>C0021311:Infections (Infection) [Disease or Syndrome] | | C0449889:Drug used (Drug used) [Functional Concept]<br><br>C0085423:Gram-negative bacterial infection (Gram-Negative Bacterial Infections) [Disease or Syndrome]<br>C0332154:Treat (Received therapy or drug for) [Functional Concept] |

MetaMap

KnowledgeMap

# MetaMap vs. KnowledgeMap

- Both do sentence ID, shallow parsing
- Both do variant generation, semantic (lungs and pulmonary) and derivational variants
- MM looks at more verbs than KM
- MM does stemming
- KM does more WSD than MM
- KM does semantic regularization
- KM allows matching of underspecified terms ("ST" -> "ST segment")

# MedLEE (Medical Language Extraction and Encoding) System

- Developed by Dr. Carol Friedman in 1991
- A mainly semantic rule-based system, based on sublanguage theory, implemented in Quintus Prolog
- Starting with radiological reports, extended to mammography, discharge summaries and pathology reports (Friedman C. et al., 1994; Friedman C. et al., 2000)
- Integrated with CIS at NYPH, with multiple applications (Jain N et al., 1996; Mendonça E. et al., 2005)

# MedLEE Semantic Types

| | |
|---|---|
| Behavior | User, drinks |
| Bodyfunc | Breathing, movement |
| Bodymeas | Pulse, weight |
| Device | Catheter, atrial electronic pacemaker |
| Finding | |
| – Cfinding | Cardiomyopathy, diabetes mellitus |
| – Descriptor | Patchy, egg shaped |
| – Organism | E. coli, Staphylococcus |
| – Pfinding | Enlarged, opacity |
| Labproc | Liver function test, SMAC |
| Labtest | Sodium, alkaline phosphatase |
| Med | Aspirin, ace inhibitor |
| Substance | Cigarettes, illegal substance |
| …. | |

# MedLEE – An Example

- **Input**:

  ***Spleen appears to be moderately enlarged.***

- **Output**:

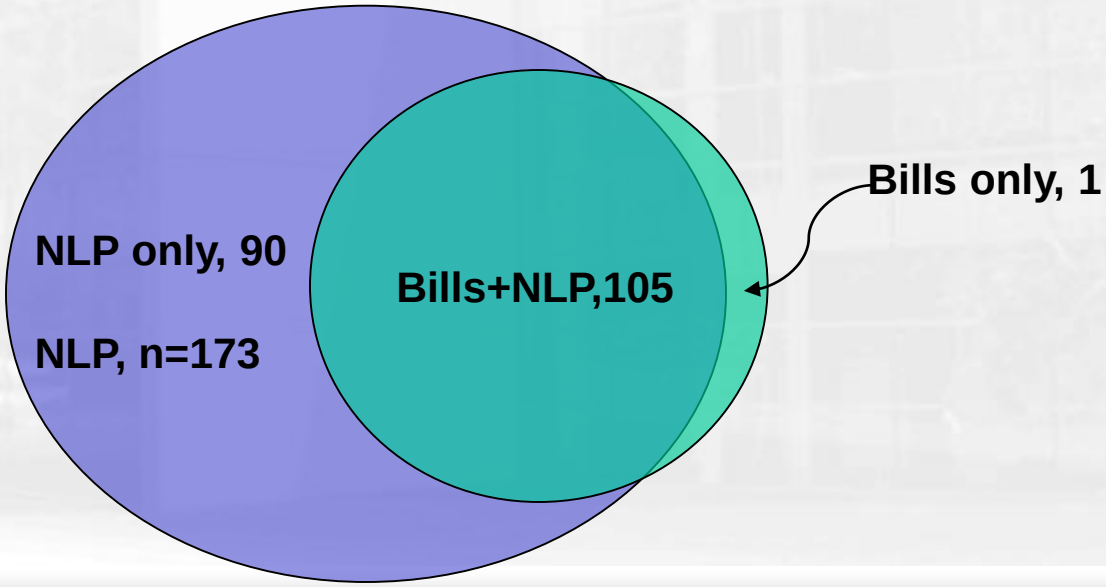
```
<problem v = "enlarged spleen" code =
    "UMLS:C0038002_splenomegaly" idref = "p1 p6">
    <bodyloc v = "spleen" code = "UMLS:C0037993_spleen" idref
    = "p1">
        <code v = "UMLS:C0037993_spleen" idref = "p1"></code>
    </bodyloc>
    <certainty v = "moderate certainty" idref = "p2"></certainty>
    <degree v = "moderate degree" idref = "p5"></degree>
    <parsemode v = "mode1"></parsemode>
    <sectname v = "report unknown section item"></sectname>
    <sid idref = "s1"></sid>
</problem>
```

# Identifying colorectal cancer screening

- Procedure billing codes (CPT) – has false negatives
  - Procedure may have been done at another institution
  - Sometimes procedures may not be billed at own institution

**CPT Codes vs NLP for Colonoscopies Performed at VUMC**

N=200 charts reviewed



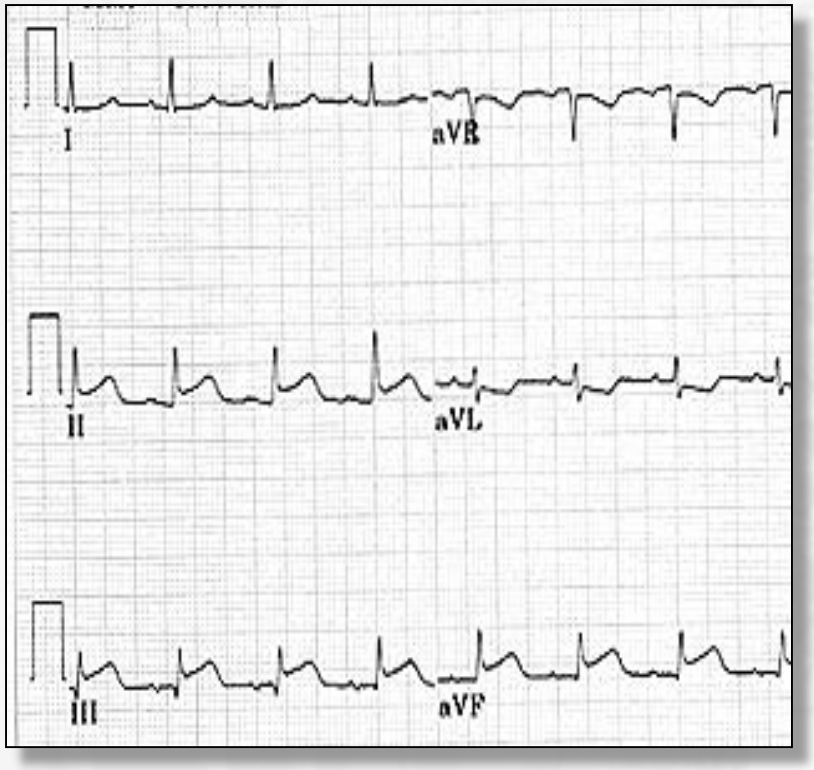
# NLP has better recall than chart review or CPT codes for CRC screening tests

| | All CRC Tests (n=265) | Colonoscopy (n=190) | FSIG (n=10) | DCBE (n=19) | FOBT (n=46) |
|---|---|---|---|---|---|
| **NLP** | | | | | |
| Recall | **93% *** | 92% | 100% | 100% | 96% |
| Precision | **94% £** | 94% | 91% | 100% | 92% |
| **Chart Review** | | | | | |
| Recall | **74%** | 72% | 80% | 53% | 91% |
| Precision | **98%** | 98% | 89% | 100% | 98% |
| **Billing Records** | **44%** | 56% | 20% | 42% | 2% |
| Recall | **83%** | 99% | 67% | 100% | 4% |
| Precision | | | | | |

*p<0.001 when comparing recall of NLP to both chart review and billing records
£ p=0.001 when comparing precision of NLP to billing records

# Electrocardiograms (ECGs): A little background

- ECG impressions contain two types of information:
  - *Morphologic finding*
    - "ST elevation"
  - *Interpretation of these findings*
    - "Inferior myocardial infarction"
- KMCI detects with 90% recall, 94% precision
  - ~40% of matches are "underspecified" in UMLS



# Methods: QT Prolongation

**ECG impression**

"Atrial fib; no evidence of MI QT is prolonged"

→ **Negation Tagging** → **KnowledgeMap Concept Identifier** →

**UMLS matches:**

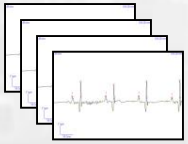
C0004238: Atrial fibrillation (asserted)

**C0151878 Prolonged QT interval (asserted)**

C0027051: Myocardial infarction (negated)

---

**Image Analysis of ECG**



→

**Automated Intervals**

PR interval: 146
QRS duration: 90
QT interval: 440
**QTc: 481**

# Finding QT Prolongation: NLP vs ECG-machine

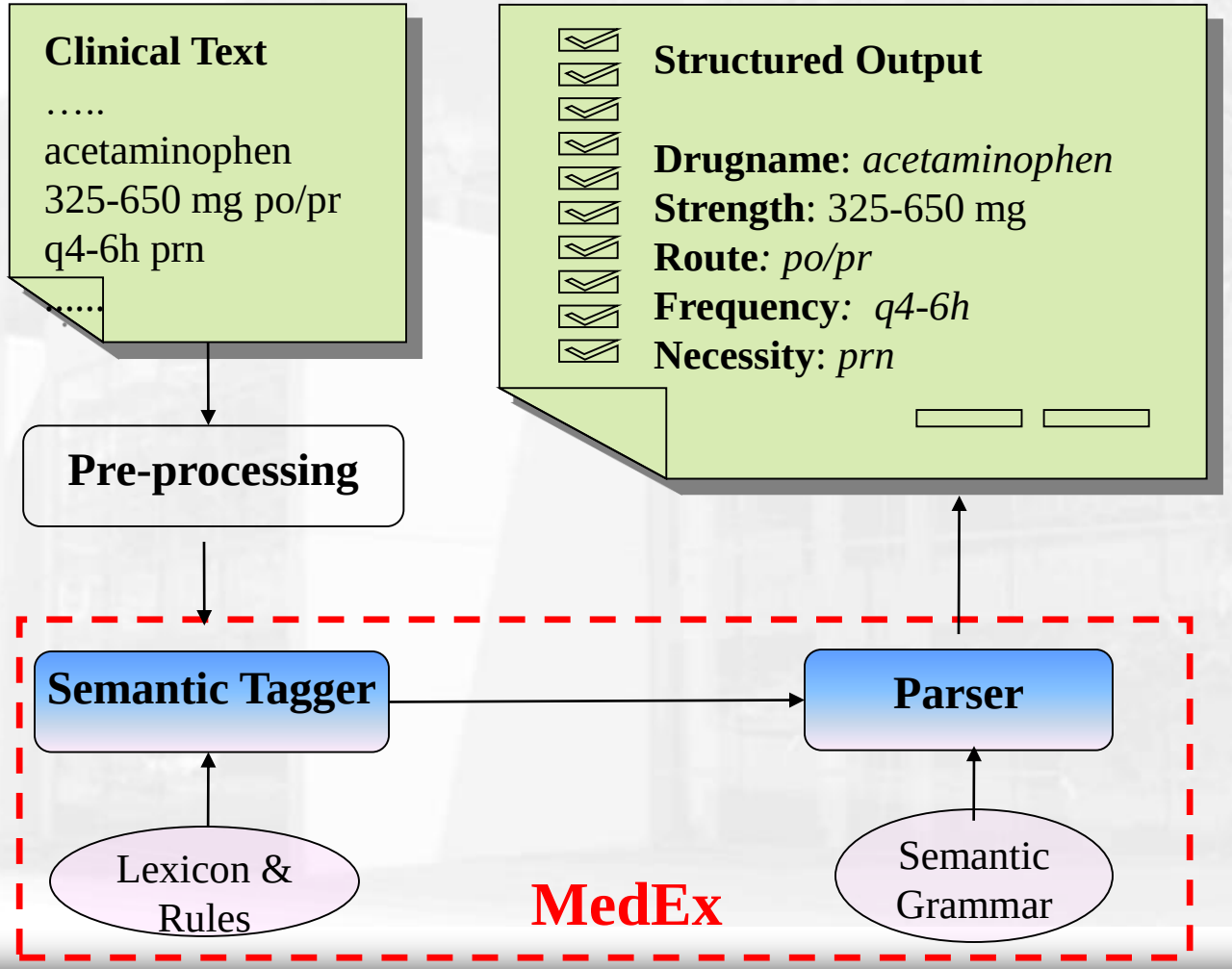
| | Concept query | ECG-machine-calculated intervals | | | |
|---|---|---|---|---|---|
| | | QTc > 400 | **QTc > 450** | QTc > 500 | QTc > 550 |
| ECGs matching criteria | 2,364 | 34,059 | **11,804** | 2,518 | 620 |
| % total ECGs | 5.3% | 77% | **26.8%** | 5.7% | 1.4% |
| Matched "Prolonged QT" | 2,364 | 2,357 | **2,304** | 539 | 117 |
| Sensitivity | **0.996** | 1.00 | **0.98** | 0.23 | 0.05 |
| Specificity | **1.000** | 0.19 | **0.77** | 0.95 | 0.99 |
| Positive predictive value | **1.000** | 0.06 | **0.20** | 0.21 | 0.19 |
| Negative predictive value | **1.000** | 1.00 | **1.00** | 0.96 | 0.95 |
| F-measure | **0.998** | 0.124 | **0.329** | 0.219 | 0.077 |

**Clinical question**: How many people have QT prolongation on an ECG, *then* receive a QT-prolonging medication?
**Answer**: 599 patients out of 1586 patients (38%) with QT prolongation over 4 years of admissions

# MedEx

- A Medication Information Extraction System developed at Vanderbilt
  - Strength “50mg”, “500/50”
  - Route “by mouth”, “iv”
  - Frequency “b.i.d.”, “every 2 days”
  - Form “tablet”, “ointment”
  - Dose Amount “take one tablet”
  - IntakeTime “cc”, “at 10am”
  - Duration “for 10 days”
  - Dispense Amount “dispensed #30”
  - Refill “refills: 2”
  - Necessity “prn”, “as needed
- Second best system in the 2009 i2b2 NLP challenge

# MedEx Overview



# Results – Discharge Summaries

| Findings Types | Total # | Precision (%) | Recall (%) | F-measure (%) |
|---|---|---|---|---|
| **DrugName** | **377** | **95.0** | **91.5** | **93.2** |
| **Strength** | **179** | **98.8** | **90.5** | **94.5** |
| **Route** | **182** | **98.8** | **89.6** | **93.9** |
| **Frequency** | **192** | **98.9** | **93.2** | **96.0** |
| Form | 39 | 97.0 | 82.1 | 88.9 |
| Dose Amt | 36 | 100 | 77.8 | 87.5 |
| IntakeTime | 23 | 83.3 | 43.5 | 57.1 |
| Duration | 22 | 76.2 | 72.7 | 74.4 |
| Disp. Amt | 7 | 100 | 71.4 | 83.3 |
| Refill | 4 | 100 | 75.0 | 85.7 |
| Necessity | 42 | 100 | 83.3 | 90.9 |

**Table 2**. Results of MedEx on 50 discharge summaries.

# Morphology

- Definition: The study of how words are composed from smaller, meaning-bearing units (morphemes)
  - Inflection: Word stem + grammatical morpheme
    - like → likes, liked, liking
  - Derivation: Word stem + syntactic/grammatical morpheme
    - generalize → generalization
  - Compounding: Two base forms join to form a new word
    - bedtime
- Application: spelling check, stemming, POS tagging, speech recognition

# Lexicography - Words

- Recognize word – Tokenization (determine the word boundary)
  - "H.Pylori gastritis", "Lortab 5 take one b.i.d"
- Identify word – Lookup (map to dictionary entry)
- Categorize word – Tagging
  - Syntactic – Assign Part-of-Speech Tags
  - Semantic – Assign semantic categories

# Part of Speech

- Part of speech - POS (syntactic category)
  - Noun, Verb, Adjective, Adverb, Preposition, Pronoun, Determiner, Auxiliary verbs, Particle, Conjunction
- POS tagging:
  - Time/N flies/V like/Prep an/Det arrow/N
- Challenge : Ambiguity
  - saw (noun / verb)
  - like (noun / verb / adv / prep / part / conj)

# POS Tagging

- Corpus-based methods: likelihoods based on sample corpus where tags have been manually annotated
  - Rule-based methods: use most frequent POS for a word or grammatical patterns
  - Statistical-based methods:
    - HMM (Hidden Markov Model)
  - Hybrid methods:
    - TBL (transformation-based learning)
- Issues
  - Corpus must be representative of domain
  - Annotation is costly

# Syntax

- Definition: study of the structure of a sentence.
  - Categories combine with others to produce a well-formed structure with underlying relations
- Representation
  - Bracketed notation
  - XML
  - Tree notation

# An Example

*The patient had pain in lower extremities*

Bracketed notation:

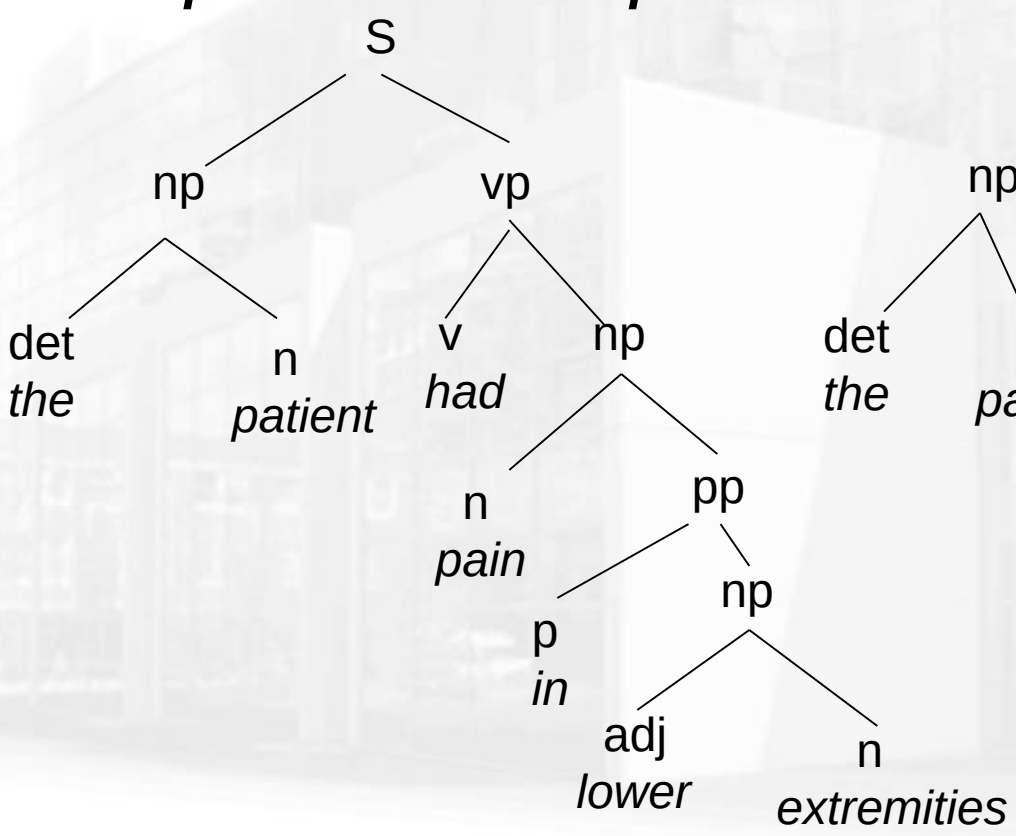
$[_{s}[_{np}[_{det}$the] $[_{n}$patient]] $[_{vp}[_{v}$had] ]$[_{np}[_{n}$pain] $[_{pp}[_{prep}$in$[_{np}[_{adj}$lower]$[_{n}$extremities]]]]]]

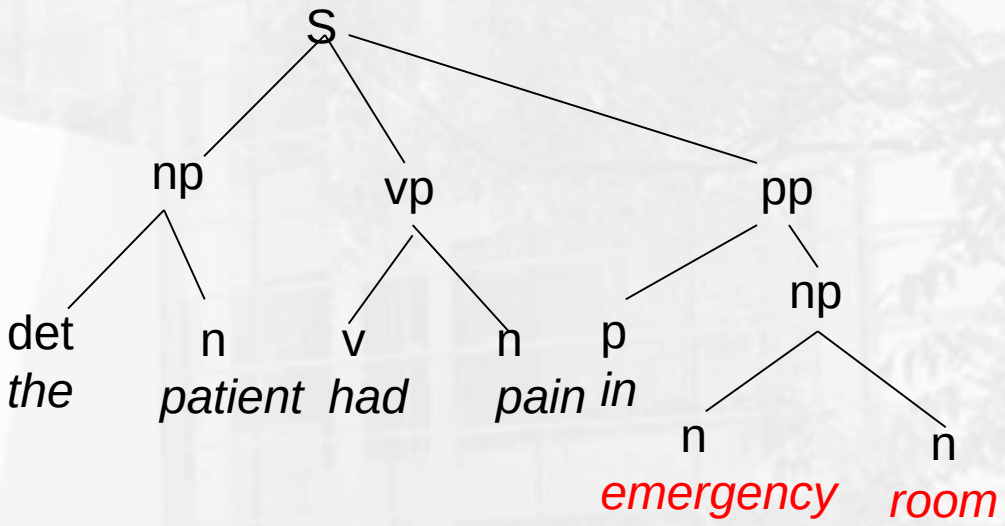
XML:

```
<s><np><det>the</det><n>patient</n></np>
  <vp><v>had</v><np><n>pain</n><pp><pr
  ep>in</prep><np><adj>lower</adj><n>extr
  emities</n></np></pp></np></vp></s>
```

# An Example

*The patient had pain in lower extremities*



# Parsing: Determining Syntax

- Formalisms:
  - Regular expressions
  - Context-free grammar or equivalent formalisms
- Parsing Methods:
  - Partial Parsing/Chunking– find simple phrases but not complex nested relations
  - Full Parsing– determine structure of complete sentence
  - Machine Learning - use manually annotated corpus to automatically detect phrases

# Semantics: Lexical

- Semantic categories of a word
  - *Abdomen* – body location
  - *Fever* – symptom
  - *IL2* – gene  (“IL2” [non-italics] – protein)
- Ambiguity (one word has multiple meanings)
  - “pt”: *patient, physical therapy, prothrombin time assay*
- Word Sense Disambiguation
  - Rule-based
  - Machine Learning based

# Semantics - Grammatical

- Word senses in a structure combine to form a meaning of the whole structure
- Methods:
  - Semantic grammar – combine syntax and semantics into one CFG grammar
  - Build template – frame with slots and fill slots based on semantic word categories
  - Have a separate syntactic parse; integrate it with semantic patterns

# Context Free Grammar

- Very simple context-free grammar of English – underlined terms are syntactic parts of speech that must correspond to a word in sentence when parsing)

s → np vp '.'

np → [det][adj]n [pp]

vp → v np| v np pp

pp → prep np

# Semantics: Lexical

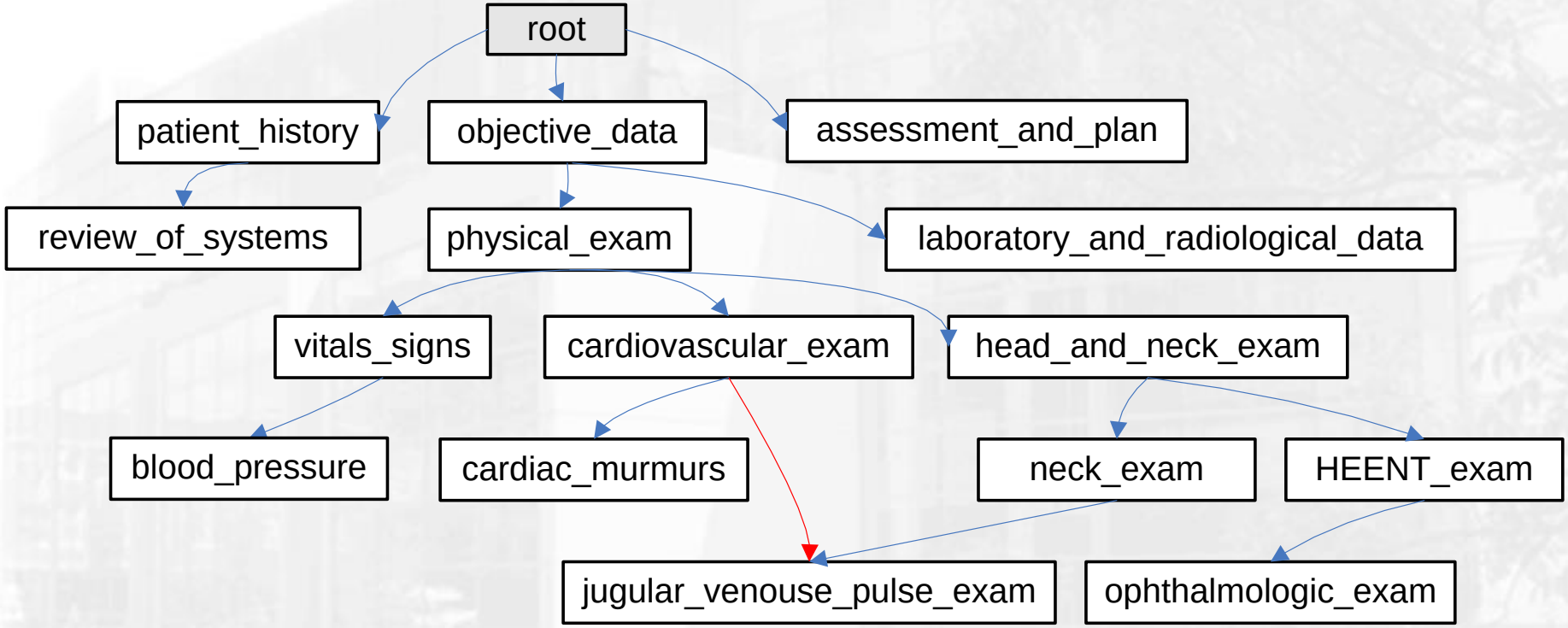
- Lexical level – to determine the meaning of a word
- Semantic categories of a word
  - *Abdomen* – body location
  - *Fever* – symptom
  - *IL2* – gene
- Difficult – ambiguity (one word has multiple meanings

# Word Sense Disambiguation

- WSD Methods
  - Can choose most likely sense
  - Use context to disambiguate (neighboring words, domain, section of report)
    - Develop rules manually
      - Discharge from hospital, discharge from eye
    - Corpus-based machine learning – develop classifier for each ambiguous word (supervised vs. unsupervised/semi-supervised)

# A portion of SecTag terminology

## (to recognize sections in clinical notes)



# Natural Language Extraction

- Text: Improved patchy opacity in the left lower lobe, no effusions seen.
- Output:
  - Finding: opacity
    - Descriptor: patchy
    - Body location: left lower lobe of lung
    - Change: improved
  - Finding: effusion
    - Certainty: no

# Pragmatics

- Context affect meaning
  - Domain: *A mass was observed*
  - Section of Report: past history vs. hospital course
  - Prior information

# Discourse

- Previous information in text affects current text
  - Correct reference (**co-reference** or **anaphora**) for pronouns, definite noun phrases, bridging noun phrases.
    - *Mass noted in left upper lobe. It was well-marginated.*
  - Time of events
  - Determining topic
  - Coherence of text

# General Purpose Clinical NLP Systems

- SAPHIRE (Hersh), PostDoc (Miller) – historical
- HITEx (Zeng) – recent but retired - clinical
- **MedLEE** (Friedman and Hripcsak) – clinical, not commericialized
- **MetaMap** (Aronson) – focused on literature
- KnowledgeMap/SecTag (Denny) – education (literature?), clinical
- **cTAKES** (Savova and Chute) - clinical
- MedEx (Xu) - clinical
- SemRep (Rindflesch) - literature
- Ytex – derivative of cTAKES/UIMA - clinical
- Mgrep (NCBO) - simple but fast
- RapTAT (Gobbel, Matheny) – ML to learn tags from other systems

VU systems **Outside systems to know**

# Example - Clinical Notes

CC: SOB

HPI: 71 yo woman h/o DM, HTN, Dilated CM/CHF, Afib s/p embolic event, chronic diarrhea, admitted with SOB. CXR pulm edema. Rx'd Lasix.

All: none

Meds Lasix 40mg IV bid, ASA, Coumadin 5, Prinivil 10, glucophage 850 bid, glipizide 10 bid, immodium prn

A/P:

- increase lasix to 80 bid
- maintain sao2 > 92% per RT protocol
- no fever and wbc wnl, so cont immodium prn
- dm2: ccm

# Example - Literature

To define the mechanism underlying signal transducer and activator of transcription (STAT) protein recruitment to the interleukin 10 (IL-10) receptor the STAT proteins activated by IL-10 in different cell populations were first defined using electrophoretic mobility shift assay.

# Clinical NLP - Requirements

- Good performance
  - running time, precision, recall, specificity
- Intra-operability: within a EMR
  - Integration with EMR
  - Shouldn't disrupt system
  - Interoperability - Map to standard controlled vocabulary (SNOMED,UMLS, ICD-9)
- Portability to other institutions
    - Generalizable representational model
    - Extensible to other domains
    - Useable for diverse applications