

IMPORTANCE OF MEASURING
SENTENTIAL SEMANTIC KNOWLEDGE BASE
OF A 'FREE TEXT' MEDICAL CORPUS

Lopamudra Chatterjee

Submitted to the faculty of the School of Informatics

in partial fulfillment of the requirements

for the degree of

Master of Science in Health Informatics,

Indiana University

May 2008

Accepted by the Faculty of Indiana University,
in partial fulfillment of the requirements for the degree of Master of Science
in Health Informatics

**Master's Thesis
Committee**

Dr. Josette Jones, Ph.D. Chair

Dr. Mallika Mahoui,, Ph.D.

Dr. Patrick Jamieson, MD, CEO

© 2007
Lopamudra Chatterjee
ALL RIGHTS RESERVED

Dedicated to
My respected parents Mr. Ranit Kr. Banerjee and Mrs. Anima Banerjee
My beloved husband Arindam Chatterjee
&
My life, my son Arko Provo Chatterjee

ACKNOWLEDGEMENTS

A sincere wish of gratitude to:

My Thesis advisor, Dr. Jossette Jones, Professor of Health Informatics, IUPUI for her constant support and encouragement.

Thesis committee member, Dr. Mahoui, Professor, BioInformatics, IUPUI for her valuable advice.

IU Solution Center to support my internship at Logical semantics, Inc.

The Guidant Corporation for granting my fellowship

And

Dr. P.W. Jamieson, MD, CEO of Logical Semantics, IN who have supported me through these years of study. Without his advice and intense supervision my thesis is incomplete.

Table of Contents

TERMS AND DEFINITIONS USED IN THIS THESIS:	1
ABSTRACT:	2
INTRODUCTION:	3
BACKGROUND:	4
HISTORY OF NLP:	4
HISTORY OF MEDICAL LEXICONS AND NLP:	7
MEDLEE--A MEDICAL PARSER USING A SEMANTIC GRAMMAR:	11
OTHER MEDICAL NLP SYSTEMS:	12
SIGNIFICANCE:	18
IMPORTANCE OF NATURAL LANGUAGE PROCESSING (NLP):	18
WHY KNOWLEDGE REPRESENTATION IS IMPORTANT IN MEDICAL NLP:	20
PROBLEM STATEMENT: SENTENCE BASED SEMANTIC ANALYSIS IN MEDICAL	
STATISTICAL SENTENTIAL SEMANTICS:	27
RESEARCH METHODOLOGY AND RESULTS:	28
RDX EDITOR:	29
THE SAMPLE SIZE:	29
PROCEDURES AND STATISTICAL ANALYSIS:	30
DISCUSSION:	36
SEMANTICS AND ZIPF'S LAW:	36
WHY ZIPF'S LAW:	37
SHANNON'S LAW AND SEMANTIC ENTROPY:	41
LIMITATION OF THE STUDY:	43
CONCLUSION:	43
APPENDICES:	45
REFERENCES:	56
VITAE	

Terms and definitions used in this thesis:

- **Proposition:** Atomic unit of semantic meaning capturing in whole or part the knowledge within a declarative sentence.
- **Knowledge Domain:** The set of all propositions that represent the knowledge within a specialized field of study such as radiology.
- **Corpus:** A large collection of related documents or reports from which a semantic knowledge base can be derived.
- **Mapping:** Linking sentences from the corpus to semantic proposition(s).
- **Semantic Hierarchy:** a taxonomic arrangement of semantic propositions, using knowledge categories to facilitate browsing.
- **Discourse level:** The organization of information in paragraphs or a document. . One problem addressed by discourse analysis is resolving anaphora by referring to previous sentences.
- **Computational Linguists (CL):** Researchers who study an interdisciplinary field dealing with the statistical and/or rule-based modeling of natural language from a computational perspective.
- **Ontology:** In computer science, ontology is the product of an attempt to formulate an exhaustive and rigorous conceptual schema about a domain. Ontology is typically a hierarchical data structure containing all the relevant entities and their relationships and rules within that domain (e.g. a domain ontology). The computer science usage of the term ontology is derived from the much older usage of the term ontology in philosophy.

Abstract:

At present, the healthcare industry uses codified data mainly for billing purpose. Codified data could be used to improve patient care through decision support and analytical systems. However to reduce medical errors, these systems need access to a wide range of medical data. Unfortunately, a great deal of data is only available in a narrative or free text form, requiring natural language processing (NLP) techniques for their codification. Structuring narrative data and analyzing their underlying meaning from a medical domain requires extensive knowledge acquired through studying the domain empirically. Existing NLP system like MedLEE has a limited ability to analyze free text medical observations and codify data against Unified Medical Language System (UMLS) codes. MedLEE was successful in extracting meaning from relatively simple sentences from radiological reports, but could not analyze more complicated sentences which appear frequently in medical reports. An important problem in medical NLP is, understanding how many codes or symbols are necessary to codify a medical domain completely. Another problem is determining whether existing medical lexicons like SNOMED-CT and ICD-9, etc. are suitable for representing the knowledge in medical reports unambiguously. This thesis investigates the problems behind current NLP systems and lexicons, and attempts to estimate the number of required symbols or codes to represent a large corpus of radiology reports. The knowledge will provide a greater understanding of how many symbols may be needed for the complete representation of concepts in other medical domains.

Introduction:

Natural Language Processing (NLP) is an interdisciplinary subject of artificial intelligence (AI) of machine learning and linguistics. There are several instances where the NLP techniques have been used to extract the meaning of a particular word of a sentence or simply the occurrence/absence of a word in a language corpus. Most of the earlier NLP systems were heavily depended upon grammatical rule based technology which had many limitations towards representing the corpus completely and thus forming a valid ontology.

In later days the need for application of NLP techniques in medical domain has been realized mainly to process medical claims and billings. At that point importance of NLP to codify data in medical decision support for improved patient care was thought to be hard enough since the medical reports were written in 'Natural Language' or in a free text form. SNOMED is the first well known lexicon based medical coding system that attempted to analyze and codify the medical data, mainly the terminologies or concepts available in a free text form. In more recent time, Medical Language Extraction and Encoding (MedLEE) system was developed to capture the underlying meaning or semantics of relatively simple medical sentences of radiology domain in free text form to some extent. Though there is a considerable amount of time has been invested by the scientists to extract the semantics of the medical 'free text' data to improve patient care with fewer errors in decision making and to build an accurate coding system, no attempts have been made until now to fully and empirically measure the knowledge base to know the number of codes or propositions or symbols to represent the medical information source or corpus. Since the probability of semantic frequency of propositions has been calculated, Zipf-Mandelbrot formula has been used for the symbol estimation purposes,

though incomplete mapping of the corpus doesn't answer the ultimate usefulness of the Zipf-Mandelbrot law in measuring the current corpus and other medical domain too.

Background:

History of NLP:

The notion that natural language could be treated in a computational manner grew out of a research program, back in the mid 1900s, based on Claude E. Shannon's mathematical logic (1948), advanced by Frege, Russell, Wittgenstein, Tarski, Lambek and Carnap. The first Russian- to-English translation projects of the 1950s, led to new research and development in order to automate the analysis and understanding of unstructured or "free" text [Inc., T.A.I. 2001]. Three key developments laid the foundation for natural language processing. In the early 1950s *formal language theory* treated a language as a set of strings allowed by context-free languages and provided the underpinnings for computational syntax [Bird, S. et al. 2005]. In late 1950s the development of *symbolic logic* provided a formal method for capturing selected aspects of natural language relevant in expressing logical proofs. In the mid- 1960s – the ELIZA program was developed at MIT [Weizenbaum, J. 1966]. This was one of the most popular artificial intelligence programs of its time, and versions of it exist for most machines, including personal computers. ELIZA is a question answering system with fixed pattern-matching templates for keywords such as: How many (F) does (N) have?, where F is the feature and N is a noun. Each template had a predefined semantic function, like $count(F,N)$. If user's query matched with the template, it was mapped to the corresponding semantic function, and eventually obtained the answer, $K = count(F,N)$. This answer was substituted into a new template: N has K F . Finally subscripts are

eliminated and the answer in natural language form returned to the user. This approach to NLP is known as *semantic grammar*. Though still widely used in spoken language system, it suffers from brittleness, duplication of grammatical structure in different semantic categories, and lack of domain knowledge and portability [Allen, J. 1995].

In late 1960s the concept of “*conceptual dependency*” was introduced by Roger Schank and Larry Tesler in the field of natural language processing. The theory was implemented in a semantic parser for natural language. The parser was not concerned with the syntactic structure of the input sentence, but rather, it was concerned with underlying meaning of the input [Schank, R.; Tesler, L. 1969]. Schank believed that computers must have an understanding of domain knowledge before they could make any decisions. Schank and Tesler in their paper stated:

The parser utilizes a conceptually-oriented dependency grammar that has at its highest level the network which represents the underlying conceptual structure of a linguistic input. The parser also incorporates a language-free semantics that checks all possible conceptual dependencies with its own knowledge of the world.

This approach tried to correct a major weakness of ELIZA, namely the superficiality of its understanding and lack of focus on the relevant topic at hand.

Another linguistic tool for studying actual human languages was developed by Charles J. Fillmore in 1968. The model is known as *Case Grammar theory*. Appendix 1 represents the basic ideas that define a case structure grammar [Schmidt, C. F]. According to this concept, each verb has a set of named slots that can be filled by nouns. Each slot explains the semantic role of its filler with respect to the verb. The relationship

between the verb and noun phrase is known as ‘case’. The ‘cases’ represent deep structure or semantic relevance even when the surface structure is different. Table 2 describes the difference between the surface vs. deep structure:

	Surface		Deep	
		medicine	John	medicine
John took medicine	subject	direct object	Casual agent	object
Medicine was taken by John	prep.object	subject	Casual agent	object

Table 2: Surface structure vs. deep structure

Casual agents are characteristics of action verbs, in which an agent brings some process about. This type of action verb (took, in example) always consists of a casual agent (e.g. John) and an object (e.g. medicine) [Parunak, V. 1995].

The first commercial research on NLP started in 1980s. Interest grew not only in understanding natural language, but also in the generation of written language.

Researchers concentrated on the goal of discovering a partial understanding of the input rather than extracting the complete meaning of every sentence. In the early 1990s emphasis was placed on using a large corpus for creating natural language processing applications [Bates, M. 1995].

History of Medical Lexicons and NLP:

Research in natural language processing in biomedicine began at the University Of Geneva, Switzerland in 1987 [Baud, R. et al. 1995]. The main directions of development were: a medical language analyzer, a language generator, a query processor, and dictionary building tools to support the Medical Linguistic Knowledge Base (MLKB) depending on conceptual graph knowledge representation [Baud, R. H., A. M. Rassinoux, et al. 1995]. Several methodologies, based on unification grammar appeared promising. Researchers focused on semantic representation in a domain, typically combined with syntax or symbolic driven methods. Researchers also began to formulate theories on discourse processing. “A discourse is an extended sequence of sentences produced by one or more people with the aim of conveying or exchanging information” [Ramsay, A. 2003]. The majority of the domain- specific natural language processing research used either the Unified Medical Language System (UMLS) or the General Architecture for Language and Nomenclatures (GALEN) ontology [Rindflesch, T. 2003].

GALEN was concerned with the computerization of clinical terminologies. The major goals of the research were to:

- 1) Allow clinical information to be captured, represented, manipulated, and displayed in a radically more powerful way [Rector. 2003] and
- 2) Support re-use of information to integrate medical records, decision support and other clinical systems. Their concern with the computerization of clinical terminologies led to the replacement of the static hierarchy of traditional clinical terminologies with a descriptive logic to help make them reusable and therefore better support computerized medical applications using clinical terminology [Rector, A. P.239-252. 1999].

The GALEN project established the ontology, and GALEN Representation and Integration Language (GRAIL) formalism demonstrated the feasibility of combining concepts. GALEN-IN-USE developed the Common Reference Model (CRM) for Medical Procedures, a key element for systems which needed to support knowledge exchange between medical records, decision support, information retrieval and natural language processing systems in healthcare [Rector, A. P.75-78.1994]. (See appendix 2 for detail representation schema)

The UMLS is a major synthesis of biomedical ontologies developed by National Library of Medicine (NLM), and serves as a resource to represent knowledge in across the biomedical domain. The UMLS is basically aggregation of domain specific knowledge bases, such as SNOMED CT, ICD-10, and CPT and can be applied in the development of computer systems which performs a variety of functions involving one or more types of information, i.e., patient records, guidelines, public health data, etc. The UMLS is focused on overcoming two important barriers to the development of information systems which can help health professionals make better decisions. These barriers are the disparity in the terminologies used in different information sources and by different users, and the sheer number and distribution of machine-readable information sources that might be relevant to any user inquiry [Humphreys, B.L.; Lindberg, D.; Schoolman, H.M., and Barnett, G.O.1998].

The UMLS Knowledge Source Server (UMLSKS) is the set of machines, programs and Application Programmer Interfaces (APIs), written in Java, that allow access to the UMLSKS services. There are three types of UMLSKS: the Metathesaurus,

the Semantic Network, and the SPECIALIST lexicon. The Metathesaurus is a large, multi purpose, multi-lingual vocabulary database, which contains information about biomedical and health related concepts, synonyms, and relationships between them. The purpose of the Semantic Network is to provide a consistent categorization of all the concepts in the Metathesaurus and thus provide a set of useful relationships between concepts. The lexical entry for each word or phrase from the Metathesaurus stores the syntactic and semantic information needed by the SPECIALIST lexicon system. (Appendix 3 shows the results of queries from these three UMLS components)

The first attempt at classifying diseases systematically was made by Sauvages with his comprehensive classification published under the title '*Nosologia Methodica*' [Knibbs, G.H. 1929]. Today the major classification for diseases in the International Classification of Diseases, Tenth Revision, Clinical Modification, ICD-10-CM, published by the Center of Medicare and Medicaid Services. All codes in ICD-10-CM are alphanumeric, i.e., one letter followed by two numbers. Of the 26 available letters, all but the letter U is used, which is reserved for additions and changes that may need to be incorporated in the future, or for classification difficulties that may arise between revisions. Some three-character categories have been left vacant for future expansion and revision (<http://www.ingenixonline.com/content/icd10/structure.asp>). (Appendix 4 shows the diagnosis codes derived from ICD-10 CM.)

As more health care professionals agree upon the adoption of Electronic Medical Record (EMR) for sharing patient care information, the Federal Government has taken the initiative in endorsing standards for EMR interoperability. Since health information

coders use the narrative information from patient's reports, the government's effort to promote robust but complex coding standards will require new technology to assist coders. The NLM strongly believes the SNOMED CT lexicon will serve as lead clinical language standard for the national health information infrastructure [Jamieson, P. 2006]; however, after studying SNOMED-CT, several weaknesses have been found which may create certain constraints for proper coding of medical conditions:

1. The example in figure 2 shows there is multiple concept IDs for 'place'. Another instance of multiple mapping is: 'Displaced fracture' (134341006) and 'Fracture with displacement' (123735002). For information retrieval, only one code should represent the sentence semantically, otherwise some reports will not be indexed properly for data mining, and some rules would not trigger in decision support applications.
2. SNOMED CT is not corpus-driven. There are many terms that have no relevancy with any medical domain. For example, concept id 257653003 represents 'open sea'. The utility of 368,000 terms for medical data mining is unclear. The sheer number of codes makes it difficult to correctly assign the correct codes to a medical document. SNOMED does not offer training to coders to efficiently and reliably code over its entire code set.
3. SNOMED CT is biased towards pathological analysis of the medical report. Domains such as radiology are not adequately represented through this terminology.

4. SNOMED CT is fairly good when representing one or two word phrases. For example, to represent a sentence such as “*there is left lower lobe pneumonia*”, a well defined SNOMED concept is “left lower zone pneumonia”. This concept is totally unambiguous. On the other hand, concepts combination like “new”, “lung structure”, “radiographic opacity”, “no”, “abnormal”, “radiographic infiltrate of lung” do not precisely represent the meaning of the sentence like “there are no focal areas of abnormal opacity overlying the lungs to suggest infiltrate” for extraction of related data. SNOMED CT is difficult to use when many concepts must be used to represent a single sentence in a medical report [Jamieson, 2003]

These unanswered questions restrict the utility of SNOMED and create hindrances for building a strong NLP extraction methodology for data mining and decision support [cross reference: Jamieson unpublished work]

MedLEE--A Medical Parser using a Semantic Grammar:

A medical NLP system for information extraction was developed by Carol Friedman at Columbia University, known as Medical Language Extraction and Encoding (MedLEE). The goal was to help physicians to extract information to communicate with the decision support system, in order to reduce health care cost, and eliminate coding errors. Her method involves taking the structured output generated by MedLEE and matching both findings and modifiers to obtain the most specific UMLS code. This system is guided by a semantic grammar consisting of patterns of semantic classes, such as degree + change + finding, which would match ‘*mild increase in pleural effusion*’. These classes are built based on UMLSKS [cross ref. Rindfleisch, T.C. et. al. 2003].

Currently, MedLEE parses a medical text document into a series of observations, with associated modifiers and modifier values. These observations are organized into sections corresponding to sections of the medical document. The result of the parser is an XML document of observations, with these observations linked to the corresponding narrative text [Nielson, J. & Wilcox, A. 2004]. The main components of this system are pre-processor, parser, error recovery module, phrase regularizer and encoder. Fig. 1 shows the knowledge components and the workflow for creating structured data:

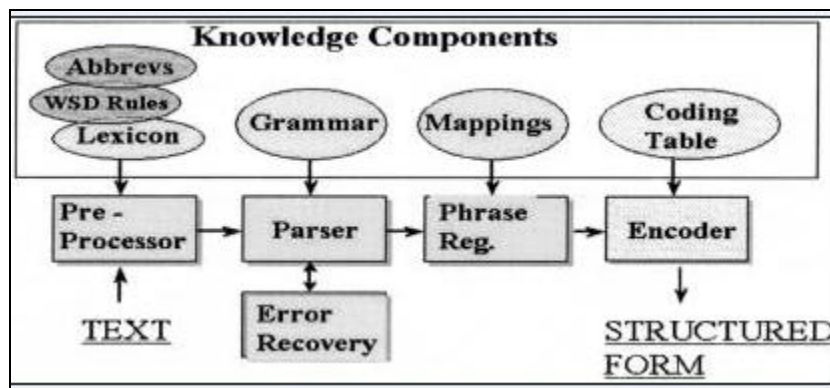


Figure 1: MedLEE Knowledge components:

From processing free text to creating structured data

Other Medical NLP Systems:

There are some other NLP systems developed in medicine such as A Query Analyzer (AQUA), RECIT (an acronym for Representation du Contenu Informationnel des Textes médicaux), etc. [cross ref. Rindfleisch, 2003]. Most of these systems are rule based, where the narrative text follows some pre-defined grammar rules to extract the desired information; but there are certain problems associated with the rule based frameworks. Pre-defined rules can be successful for interpreting a simple sentence, however, a

challenge all medical NLP systems face is that valid semantic knowledge is commonly found in syntactically incorrect sentences, and that less commonly invalid semantics (nonsense) can be found in syntactically correct sentences.

As sentence complexity increases, some computational linguists add new formulas in order to make the logical reasoning complete. This leads to the well known “Frame Problem” in artificial intelligence, namely, additive functions make it impossible to express a rule with an open ended set of exceptions. For example, it is possible to create rule to frame the semantics of the sentence like ‘*The ventricles are prominent*’ with first order predicate logic (FOPL), where the relations or predicates (i.e., *prominent*) are placed on the left hand side and terms or body parts (*cerebral sulci and ventricles*) are placed to the right, such as (Prominent: Cerebral Sulci). FOPL can modify only a single subject at a time. If the sentence becomes larger with more complex concepts like ‘*The cerebral sulci and ventricles are prominent, compatible with mild diffuse cerebral atrophy*’, FOPL must use more complex structures or it may fail to capture the entire meaning of the sentence.

The extraction of key words or noun phrases from the medical text is not complete enough to represent the knowledge required in decision support and data mining. For example, the natural language parser can pullout terms from clinical reports using shallow parsing, and then alert clinicians about a possible adverse drug reaction.

However, to go beyond simple concept spotting requires a deeper knowledge of the domain. MedLEE lacks a knowledge representation framework for representing many medical concepts and their relations. A domain-independent knowledge representation must be granular enough to represent all the information expressed in the free text. If not,

precision will be lost for both researchers performing text analysis and informaticists designing decision support systems. Appendix 5 shows the input and output of MedLEE, whose primary goal was to analyze free text radiological reports. It is capable of capturing the radiological phrases in a structured format, but at the same time, its lack of domain knowledge does not make allow it to reject semantic nonsense, or recognize when it simply can't fully analyze a given sentence. Table 3 shows a few meaningless but real radiological sentences and their representation in MedLEE. Structuring any syntactically correct word irrespective of its semantic value raises questions about the competence of the NLP system, and the range of phenomena it can truly extract.

Actual 'no-sense' radiological sentence	MedLEE representation
<p>1. Heart size is mediastinal contour.</p>	<pre> finding: size of heart bodyloc>> mediastinum idref>> 14 code>> UMLS:C0025066_mediastinum idref>> [14] code>> UMLS:C1278909_entire mediastinum idref>> [14] certainty>> high certainty idref>> 12 idref>> 8 parsemode>> mode4 sectname>> report summary item sid>> 2 code>> UMLS:C0744689_heart size idref>> [8] </pre>

<p>2. The patient has had prior surgical removal of the ballbladder .</p>	<pre> procedure:surgery idref>> 18 parsemode>> mode4 sectname>> report summary item sid>> 2 procedure:procedure idref>> 18 parsemode>> mode4 sectname>> report summary item sid>> 2 bodyloc>> biliary tract idref>> 18 code>> UMLS:C0005423_biliary tract structure idref>> [18] code>> UMLS:C1278928_entire biliary tract idref>> [18] code>> UMLS:C0005427_biliary tract surgical procedures idref>> [18] </pre>
<p>3. The lungs show lung volumes .</p>	<pre> bodymeas:lung volume idref>> 14 parsemode>> mode4 sectname>> report summary item sid>> 2 code>> UMLS:C0231953_lung volume idref>> [14] </pre>
<p>4. The lung volumes are somewhat lobe.</p>	<pre> problem: low lung volume certainty>> high certainty idref>> 14 degree>> low degree idref>> 16 idref>> [10,16] parsemode>> mode4 sectname>> report summary item sid>> 2 bodymeas:lung volume certainty>> high certainty idref>> 14 degree>> low degree idref>> 16 idref>> [10,16] parsemode>> mode4 sectname>> report summary item sid>> 2 degree>> low degree idref>> [10,16] code>> UMLS:C0231953_lung volume idref>> [10,16] </pre>

Table 3: MedLEE's information extraction and coding of some real 'no-sense' sentences

Thus, to develop a high precision semantic retrieval system, it is very important to have a deep understanding of the domain knowledge. There are many ways in which knowledge can be represented in a computer system. One way is to use a set of propositions that represent the knowledge within a specialized field of study, such as radiology. Currently computational linguists are only able to analyze 30% of English sentences and transform them into structured forms [Rebholz-Schuhmann D, et al. 2005]. Unfortunately, there are a limited tools and methods available for systematically categorizing a large domain knowledge into ‘knowledge elements’. BioTeKS, a tool developed by IBM researchers, is capable in pointing out some semantic categories and their relations using automated annotators [Mack R. et al. 2004], but fail to extract the broader semantic relationships found in medical reports without creating and refining a large rule-based grammar especially within a complicated domain [Jamieson, P. 2004].

Table 4 shows a comprehensive list of some existing NLP systems and lexicons including their characteristics and limitations:

NLP systems and lexicons	Characteristics	Limitations
ELIZA	Question answering system with fixed pattern-matching templates for keywords.	Brittleness, duplication of grammatical structure in different semantic categories, and lack of domain knowledge and portability.
GALEN Ontology	Allows clinical information to be captured, represented, manipulated, and displayed in a radically more powerful way. Demonstrates	Only pre defined, restricted combinations of concepts are allowed.

	feasibility of combining concepts.	
UMLS lexicons	Can be applied in the development of computer systems which performs a variety of functions involving one or more types of information, i.e., patient records, guidelines, public health data, etc.	Only one-to-one relationships are mapped – Only terms from source vocabularies present; no new terms added – No unifying hierarchy is present, only those that exist in source vocabularies – Not extensible (i.e., in the SNOMED sense)
ICD-10 CM lexicons	All codes are alphanumeric, i.e., one letter followed by two numbers. Used in disease classification.	Limited coding systems. Granularity often inadequate.
SNOMED-CT lexicons	A standard with more than 368,000 codes to analyze health records. Compositional in nature.	Coding ambiguity. Ignores context. Granularity often inadequate, e.g., no coding for ‘gross’ soft tissues. Highly conceptual based hierarchy. (Chute, 2005)
MedLEE : IR And Coding system	Most sophisticated natural language information extraction system.	Knowledge base is not clearly specified. Mainly focused on absence or presence of key phrases in the sentence.

Table 4: Different NLP systems including characteristics and limitations

Significance:

Importance of Natural Language Processing (NLP):

Informatics can help solve one of the most vexing problems in US health care -- rising cost. The public demands high quality care at an affordable price which is increasingly difficult to deliver. Quality improvement is particularly important in chronic disease management since chronic diseases account for 75% of total health care costs in the US. Prevention of long term complications depends on the implementation of essential evidence-based services [CDC. 2004] which could dramatically lower costs.

Researchers have developed decision support tools such as antibiotic advisors, which have lowered costs and reduced hospital stays [Kuperman, G.J. et al. 2003]. However, the lack of semantic and conceptual understanding of structured healthcare terminology is a barrier to deploying decision support applications and improving clinical management.

Medical documents –electronic or paper, are rich in biomedical information but most of the information is recorded in a narrative text (natural language) form. W. Giere stated, “... although it is possible to structure medical records and to use codes or abbreviations, for much of the data which is frequent and typical, there can be no medical record or no useful electronic patient information without narrative text, i.e. “free text” [Giere, 2004].

Drawing medical conclusions from this enormous repository of unstructured free text data and creating a knowledge base for a high quality treatment is one of the grand challenges in health informatics.

For computer supported decision making, all data relating to health care events, including free text data should be analyzed. Free text information must be semantically understood, extracted and converted into a structured coded form [Hripcsak, G; Friedman, C et al. 1995]; Natural language processing (NLP) is a computational field that

attempts to convert free text into structured form to perform tasks like translation, speech recognition, summarization, information extraction, and document categorization. A NLP system must be able to represent knowledge in a form appropriate for computer manipulation. Statistical and logical modeling of natural language is studied by computational linguists (CL). Their goal is the development of a description of natural language, where a theory guides the descriptive format, and a methodology establishes the procedures for obtaining the description. Both the descriptive format and methodology significantly impact the system's design [Nirenburg S, Raskin, V. 2004]. According to computational linguists a text string with one given semantic meaning should be represented by unique symbol(s), and two text strings with the same semantic meaning should be represented by the same unique symbol(s) [Cimino, JJ. et al.1994]. For example, words like 'normal' and 'unremarkable' should share the same unique symbol since both of them represent the same meaning semantically.

To create a knowledge base or ontology, a specification of conceptualization (Gruber, T. 1993) is required. CLs require a keen understanding of the way knowledge is organized in the free text, and NLP techniques which can extract that knowledge in order to represent the document's semantics. Knowledge representation (KR) is not typically concerned with the physical details of how knowledge is encoded, but the overall conceptual scheme [Jackson, P. 1999]. Jamieson (in press) states that unless a computational system knows how and what to represent in free text documents, users cannot mine the free text successfully. Most existing coding systems are focused on building lexicons, overlooking the importance of the KR at the semantic level. Natural language processing must develop new schemes for knowledge representation to

accurately reflect document semantics. However, current NLP systems rarely describe methods for evaluation their KR schemes [Robert H. Baud et al.1997].

Why Knowledge Representation is Important in Medical NLP:

Despite the rigorous research in medical NLP, Lee and Bryant (2002) have stated that there had been limited successful attempts made to *automate* the extraction of knowledge from documents written in a natural language. First, NLP involves the integration of many forms of knowledge, including syntactic, semantic, lexical, pragmatic and domain knowledge [Chen, H. ET. Al., 2005]. Another reason is the ambiguity of some terms, and the complicated hierarchies of existing classifications and concepts such as the Systematized Nomenclature of Medicine- Clinical Terms (SNOMED CT). For example, the phrase '*displaced fracture*' and '*fracture with displacement*' have two separate identifiers in SNOMED-CT, even though the semantics of those phrases are equivalent. The design problems of existing classifications and coding systems must be addressed before one can extract knowledge through natural language processing. Redundancy and inconsistent vocabularies as well as the lack of granularity are obstacles which make interoperability difficult among different information systems. For example, 'yes' and 'no' can be represented in different ways in different databases like y = yes and n= no, 1= yes, 2= no or 0=y, 1=n. Without creating a unique identifier for 'yes' and 'no', the operation of a query system could be impaired. For example, there is no such concept identifier available to describe "*focal infiltrate*" in ICD-10-CM. because the term "*infiltrate*" does not exist in the ICD-9 lexicon since it doesn't imply disease.

Beside the hurdles of ambiguity, granularity and hierarchical complexity, existing coding systems are not corpus based. The word ‘corpus’ is defined as a large collection of related written documents. A corpus should aim for balance and comprehensiveness within a specific sampling frame, in order to allow a variety of language to be studied, because without limiting the domain boundary, it is impossible to collect all of the utterances of a natural language within one system [McEnery, Tony. 2003]. Studying a medical corpus empirically helps to better understand and analyze the pattern of the related documents in free text and build a strong knowledge base to process the free text with few limitations. Some clinical applications (table1) that have used corpus based NLP technology are as follows: [Chen, H. 2005]

<i>Clinical Domain</i>	<i>Application</i>
Progress notes	Quality Assessment
Pathology	Key diagnoses for indexing
Radiology, Emergency Medicine	Coding for billing
Discharge summary	ICD-9 CM for indexing

Table1: Some clinical applications that use corpus based NLP technology

Problem Statement: Sentence Based Semantic Analysis in Medical Domain:

As previously mentioned, medical reports are being actively researched by computational linguists for various reasons:

- The need to improve medical decision making
- The availability of manpower to manually structure and codify free text.

- The limitations of free text for data mining and decision support
- There is a large corpus of published reports

A review of the literature [Greenes, RA. 2003, Calzolari, N. 2003, Zarri, GP. 1996] confirms that medical knowledge representation is a key component in medical natural language processing. The knowledge construction activity has consumed significant effort among investigators trying to develop a comprehensive standardized health lexicon. The Unified Medical Language System (UMLS) sponsored by the National Library of Medicine has been in development for over 20 years. Despite this concentrated effort, there is not a well structured code set by which most medical report sentences can be represented semantically. The methodology for constructing an adequate knowledge representation system is a challenging research area that has received little formal analysis. Most knowledge representation schemes have not been empirically driven and few have been scaled to encompass a substantial medical domain such as radiology. A good NLP system grounded with a comprehensive sentential semantic knowledge base would allow the physicians to use the free text in a report for text mining and decision support and must meet the following criteria:

- Semantically equivalent text string(s) should have same unique symbol(s).
- One symbol = one meaning only.
- Identify sentences with only valid semantics.
- Document the system's abilities and limitations to the end users to avoid confusion.

Most knowledge representation research has focused only on terminology or lexicons. Alan Rector in his article “Clinical Terminology: Why Is It so hard?” points out several obstacles to terminology development. Scaling terminology by an order of magnitude or more from current code sets like ICD-9 or CPT is notoriously difficult, and requires a change in methodology and technique [Rector AL. P.239-252. 1999, Rector AL. P.1-4.1999]. Beside scalability, there are fundamental conflicts between the needs of users and the requirements for developing software. Few medical computational linguists have articulated a coherent semantic theory or described in detail their formal concept representation system for modeling the semantics of medical sentences. Without articulating a semantic theory it is impossible to extract the free text sentences from the medical reports that are equivalent in meaning to each other [Jamieson P.W. 2006].

The sentential semantic theory is based on sentential logic or propositional logic. It is that branch of logic that studies ways of combining or altering statements or propositions to form more complicated statements or propositions. Joining two simpler propositions with the word "and", “or” is two common ways of combining statements. When two statements are joined together with "and", “or” the complex statement formed by them is true if and only if *both* the component statements are true. For example, ‘*No abnormal filling defects or extravasation of contrast was noted*’ can be more completely represented by the propositions: ‘*There are no abnormal filling defects*’ and ‘*There was no extravasation of contrast was noted*’.

There are several benefits of the knowledge representation of descriptive text at the sentential semantic level. First of all, it eliminates ambiguity underlying the

knowledge contained in the sentence, e.g. the phrase ‘*There is mild atrophy*’ may indicates both ‘mild cerebral atrophy’ and ‘mild muscle atrophy’; however, understanding the correct context makes it possible to identify and code the sentence in an unambiguous way. Next, a corpus-driven knowledge base is an efficient way to express concepts with the fewest codes. For example, to represent a sentence ‘*NG tube is in place with its tip not seen in the film*’ in SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms) six different types of code are required (see figure 2); while a sentential knowledge representation technique has an advantage to express the meaning only with one unique proposition (see figure 3). These ideas, if properly applied will facilitate text mining and semantic queries with high precision, and make decision support easier for health informaticists to implement.

Text line in a medical report: “*NG tube is in place with its tip not seen in the film*”.

NG – Nasogastric tube, NGT- Nasogastric tube -> 17102003

Place -> 246297005, 257557008

Catheter Tip -> 116204000 (closest match for ‘tip’)

Not seen -> 47492008

Plain film -> 168537006 (closest match for ‘film’)

Figure 2: Example use of SNOMED-CT codes to represent typical medical sentence.

UnMapped Sentences	Suggested Propositions	Similar Sentences	Review Maps	Review Categories	Review Propositions
UnMapped Text Lines					Freq
■	LEFT ELBOW				19
■	VERTEBRAE				19
■	The left kidney measures at least 11.6 cm in length .				19
■	Baseline with no comparison .				19
■	The lungs are well ventilated .				19
■	These findings are consistent with rheumatoid arthritis .				19
■	The ribs appear intact .				19
■	Worsening pulmonary edema .				19
■	No images were obtained .				19
■	No active pulmonary disease .				19
■	history : knee pain .				19
■	Soft tissue planes are appropriate .				19
■	The remainder of the bones are normal in alignment and configuration .				19
■	There has been interval decrease in the right pleural effusion .				19
■	There is no lymph adenopathy .				19
■	The peel away sheath was removed .				19
■	These are unchanged from the prior examination .				19
<input type="button" value="Create Maps"/>					
Proposition	System Context	Modality Context			
There is no lymphadenopathy.	All	All			

Figure 3: Usage of proposition in Rdx Editor to represent one unmapped medical text line.

Numbers on the right hand side showing frequency of occurrence of that line in the corpus

One of the reasons it is difficult for a NLP system's to extract and codify the meaning of free text is that language understanding is very knowledge intensive. One way to improve a NLP system is to create a rich semantic knowledge base that mirrors the content of the domain it is trying to analyze. Existing NLP systems frequently do not describe their semantic knowledge in enough detail, so one can determine exactly what they are capable of extracting. It is critically important to have a thorough understanding of the system's knowledge structures including predicates, terms, and operators. The data structures can reveal whether a NLP system can represent the semantics of a phrase,

sentence, or a larger textual unit like a paragraph or document. Even when a NLP system's data structures are disclosed, it is still important to understand precisely which concepts can be represented. For example, consider the following concepts: 'The patient is status post craniotomy', 'The patient is status post left craniotomy', 'The patient is status post left frontal parietal craniotomy', 'The patient is status post left temporo-parietal craniotomy', and 'The patient is status post left temporo-occipital craniotomy.' Clearly, there are differences in "granularity" between these concepts. If a researcher was only interested in finding patients that have had a craniotomy, or a procedure where the neurosurgeon entered the skull, then the top level concept, 'The patient is status post craniotomy' would be the only concept the NLP system would be required to represent. But what if the researcher were a neurosurgeon that wanted to find all cases in which a left temporo-parietal craniotomy was performed.'? In this case, if the NLP did not have an entry in its knowledge base which describes this concept, it is highly unlikely that the NLP would be able to locate reports that described this procedure.

Another problem most NLP systems fail to confront is the inevitable mismatch between the documents they wish to analyze and the knowledge base they wish to use. Knowledge is very context specific. If the semantic knowledge base is not derived from the source documents being analyzed it is highly likely some concepts will not be represented. These observations frame the power of a natural language processing system in a new light. The heart of the system is much more than the NLP text extraction algorithms. The quality and depth of knowledge representation are critically important. Given how important the semantic knowledge base is for the overall success of the NLP system it is surprising that very little information exist to describe and quantify how large

a knowledge base is needed to represent a knowledge domain. While we know the frequency of words in various collections of documents, the informatics community has no understanding of how many concepts are needed to represent the knowledge in a domain like radiology. This has critical ramifications for the development of code sets like SNOMED CT, which are attempting to model the clinical domain.

Statistical Sentential Semantics:

This thesis attempts to answer how many semantic symbols are needed to represent the radiology domain. The research will examine, which symbols are most important from the standpoint of representing most of the content. It will attempt to empirically determine the frequency distribution of sentences mapped to sentential propositions (a convenient way of representing statements in natural language). I will examine Zipf's Law and a related equation, the Zipf-Mandelbrot equation to see if one can predict the distribution of the semantics of a domain. If this is successful, a prediction will be made of how many propositions will be required to represent a certain percentage of the domain. It also shows how well Zipf-Mandelbrot law represents the actual data of this particular corpus.

The keys to this research are:

- (1) Empirically deriving from a large corpus of radiology documents the propositions needed to cover a portion of the sentences in the domain. For this research, I will use a knowledge base with over 2 million sentences that have had their semantics fully characterized.

- (2) Understanding the mathematics of information communication and coding. Probably the greatest mathematician to formulate a coherent theory in this area is Claude Shannon. Shannon considered a source of information as one that generates words composed from a finite number of symbols. These are transmitted through a channel, with each symbol spending a finite time in the channel. He used statistics with the assumption that if x_n is the n th symbol produced by the source the x_n process is a stationary stochastic process. Our problem is not so very different. However, instead of words, we are working with larger units of semantic information called propositions. Instead of transmitting symbols through a channel, we need to map sentences to propositions. Each symbol (proposition) used by the Rdx editor has a certain probability of occurrence based on the frequency distribution of concepts in the corpus.

Research Methodology and Results:

The frequency of propositions and other linguistic units play a central role in corpus linguistics. Indeed, the use of frequency information distinguishes the corpus-based methodology from other non-empirical approaches in computational linguistics. In order to study the proposition frequency distribution, we counted all the instances of semantically equivalent sentences that were mapped to a given proposition that occur in the corpus of interest [Baroni, M. 2006], for our case the radiology domain.

Rdx Editor:

Semantically annotating sentences is a labor intensive process. Rdx, is a semantic annotation tool (see figure 3), that makes it easier to semantically tag each sentence in a domain. The first step is to segment each report into sentences. For our domain there are 4.4 million sentences of which slightly over 2 million are unique. Propositions are created and then arranged in a knowledge base to represent the underlying meaning of the segmented sentences from the radiological reports. Sentences equivalent in meaning (semantics) are mapped to the same proposition(s). Mapping sentences and creating propositions is done through a semi-automated process using different statistical methods such as K nearest neighbor (K-NN) method. This method computes the nearest neighbor or the nearest matching to the unknown target sentence in the corpus. The K highest ranked sentences that were previously mapped to the same proposition(s) were retrieved, sorted by the 7 most relevant pre-mapped sentences first in the list. Propositions in the knowledge base have been arranged according to the semantic hierarchy, not the conceptual hierarchy. The most general propositions are presented at the higher level in the knowledge base followed by the more specific one. Sentences with high ambiguity and personal information have been marked as 'skipped'. The mapping is checked, revised and approved by the senior medical editor, who has the sufficient domain knowledge.

The sample size:

The radiology corpus contains over 4 million sentences. Currently, 50% of total corpus sentences have been mapped to approximately 5700 propositions of the semantic knowledge base. About 63,800 sentences (1.45% of entire corpus sentences) has been marked as 'skip' because a sentence contains only a single word (none, otherwise, Dr.,

etc.) which can be easily ignored for the semantic analysis purposes and about 12,053 (0.27% of 4 million sentences) sentences containing personal information (Doctor's name, patient's name, etc.) which have been deleted to maintain the HIPAA privacy rules and regulations. Other sentence like "*Otherwise normal exam*" has been skipped due to its highly ambiguous nature. The sentence is mainly dependent on the prior sentence of the report and without discourse understanding, this sentence can't be analyzed.

Procedures and Statistical analysis:

A SQL stored procedure has been developed (see appendix 6) to count the current semantic frequencies of sentences mapped to propositions. Figure 4 shows the outcome of the stored procedure. The numbers in 'totfreq' column demonstrate the weighted frequencies of all text lines in the corpus that have been mapped. The more frequently a proposition is used to map sentences, the more weight it has. 'Statement' is the ranked propositions in a descending order.

To compare the actual and predicted trend of frequency vs. ranks of first 10,000 data points, a graph (figure 5a) has been drawn. Note that each axis is plotted on a log 10 scale.

After reviewing existing power-law formulas, to model the behavior of the semantic frequency accurately and estimate an approximate number of total propositions required to cover whole corpus, The Zipf's law has been applied. The Zipf's-Mandelbrot formula (a special case of Zipf's original formula) was used for first 100 propositions to adjust the higher frequency more closely to fit the Zipfian straight line. The formula is as follows:

$$\text{Term (semantic) frequency} = \frac{C}{K + r^\phi}, \text{ where } C (2113489), K (6.3) \text{ and } \phi (1.33) \text{ are}$$

the three parameters that have been used to provide a fit for the actual data (partially

shown in figure 4. The rank of the propositions is represented by the symbol 'r' (1, 2, 3...n).

In deriving the formula, ϕ has been calculated to get the slope of the line. Three data points from the excel spreadsheet have been selected to calculate the slope accurately. Parameter C symbolizes the point where the slope cuts the Y axis. It has been estimated using Y intercept formula: $Y=mx +b$, where m is the slope, x is the value of X axis (rank) and b indicates Y-interception point. The final step is to calculate the K value to minimize the distance error between actual data point and predicted data point at the higher frequency level, i.e., first 100 highest ranked propositions. Various values of K (0, 4, 5.5, 6.3, and 7) have been examined to fit the actual data best (see appendix 7).The total number of mapped lines has been summed up to 2259944(adjusted predicted value).

The linear portion of the curve derived from the actual data, follows the formula originally developed by Zipf, since Mandelbrot's formula doesn't correspond to a straight line in double logarithmic space. The Zipf's equation is as follows:

$$\text{Frequency (semantic sentences)} = \frac{C}{r^{\phi}}$$

(2113489) is the Y-intercept and $r (1...n)$ is the rank of the proposition. The total number of actual adjusted mapped lines is 2194133.

Finally, comparison between the actual data curve and estimated data curves based on K value (figure 6) has been performed to prove the efficiency of the formula in representing a model for estimating the extent of knowledge base in the radiology corpus. In addition to the frequency counting, statistics regarding the percentage of total corpus area covered by the current propositions were measured. The area covered by each proposition was estimated based on the cumulative semantic frequency of ranked

propositions, i.e., if the first proposition in rank covers 141,851 semantically equivalent sentences and second highest proposition is linked to 107,812 sentences, together they cover about 4.36% of the corpus. To analyze the data, first 5,500 ranked propositions and their corresponding predicted frequency counting has taken into account. The data predicts about 52.26% of the corpus can be mapped to 10,000 propositions (figure 7). The nature of the graph suggests that approximately first 3,000 propositions have the most frequent text lines that rapidly cover almost half of the corpus area. From data mining perspective, these propositions contain the most important concepts or knowledge. The linear portion of the graph represents rarer sentences with lower frequencies that may need a new proposition to represent the underlying meaning. Increase in the rate of creating propositions does not imply greater coverage of the corpus, since semantic frequencies are in lower ranges (in most of the cases less than 5 occurrences).

In addition to the Zipf's law, a brief attempt has been made to calculate the average entropy of the propositions using Shannon's theory of communication (see appendix 8) and correlate it with Zipf's law. The formula of information entropy derived by Shannon is:

$$H = - \sum P_i * \log_2 P_i \text{ (bits per symbol)}$$

Where H = informativeness per symbol or uncertainty, P_i = probability of i^{th} symbol. The entropy has been measured from the sentential semantic probability distribution in the corpus. The calculated average semantic entropy is 8 bits per symbol in this radiology corpus, which suggests 2 to the power 8 or 256 symbols, represent most of the semantics

of the 2 million sentences in the corpus. Discussing the theory of communication in detail beyond the entropy calculation and definition are out of scope of this research paper.

	totfreq	statement
1	141851	A comparison was made to prior films.
2	107812	No comparison was made to other studies.
3	77666	A portable film was obtained.
4	68402	An image was obtained of the right side of the b...
5	62026	An image was obtained of the left side of the bo...
6	61126	There are no comparison films available.
7	57812	The heart size is normal.
8	55522	A lateral chest x-ray was performed.
9	53660	A posterior-anterior chest x-ray was performed.
10	53426	The pulmonary vasculature is normal.
11	49552	The lungs are clear.
12	41736	There are no fractures.
13	40717	Three views of the area were obtained.
14	35802	A frontal chest x-ray was performed.
15	35710	There is no pneumothorax.
16	31383	There are no focal pulmonary infiltrates.
17	30615	The examination is normal.
18	28964	There are no pleural effusions.

Figure 4: Table showing top ranked propositions by frequency of mapped sentences in the corpus

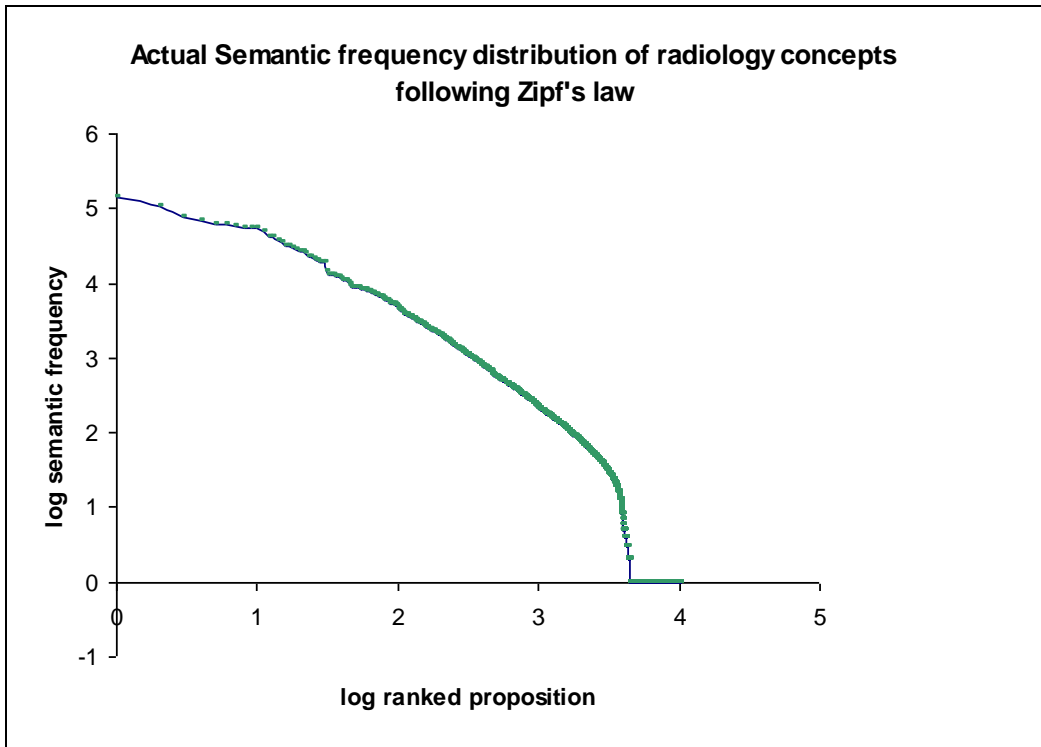


Figure 5a: Distribution of actual and predicted semantic frequency vs. rank of first 10,000 propositions in log-log scale

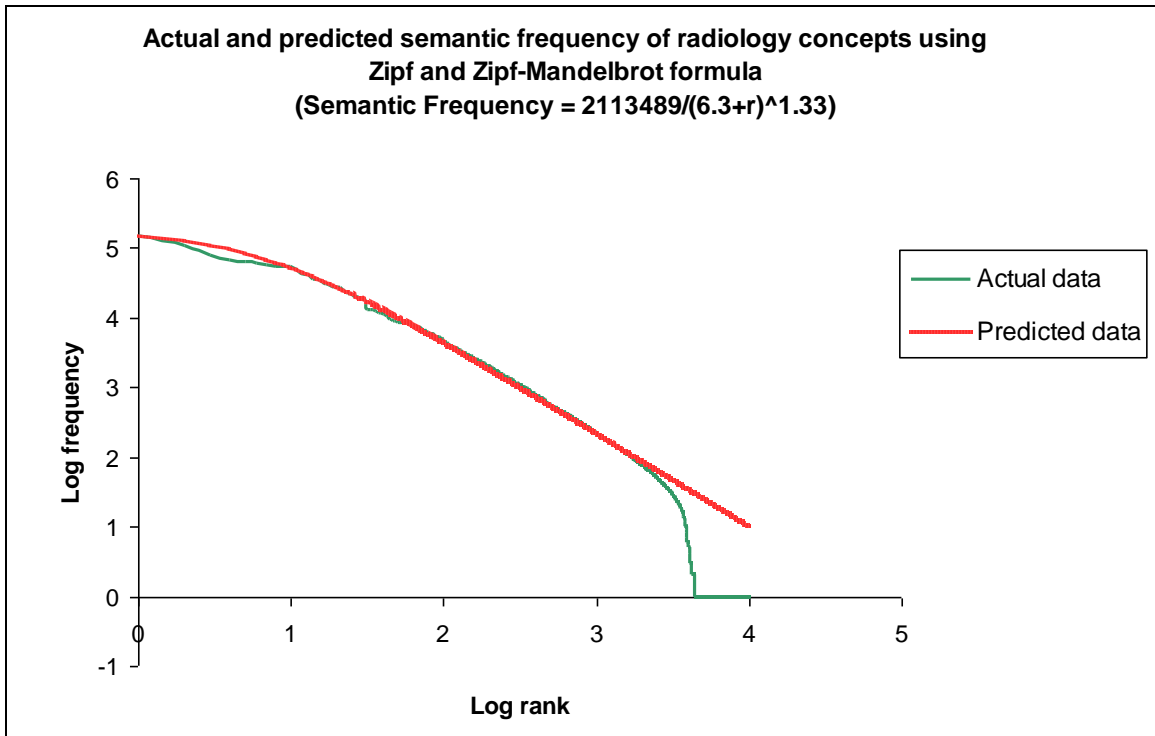


Figure 6: Actual and predicted semantic frequency distribution using Zipf and Zipf-Mandelbrot law

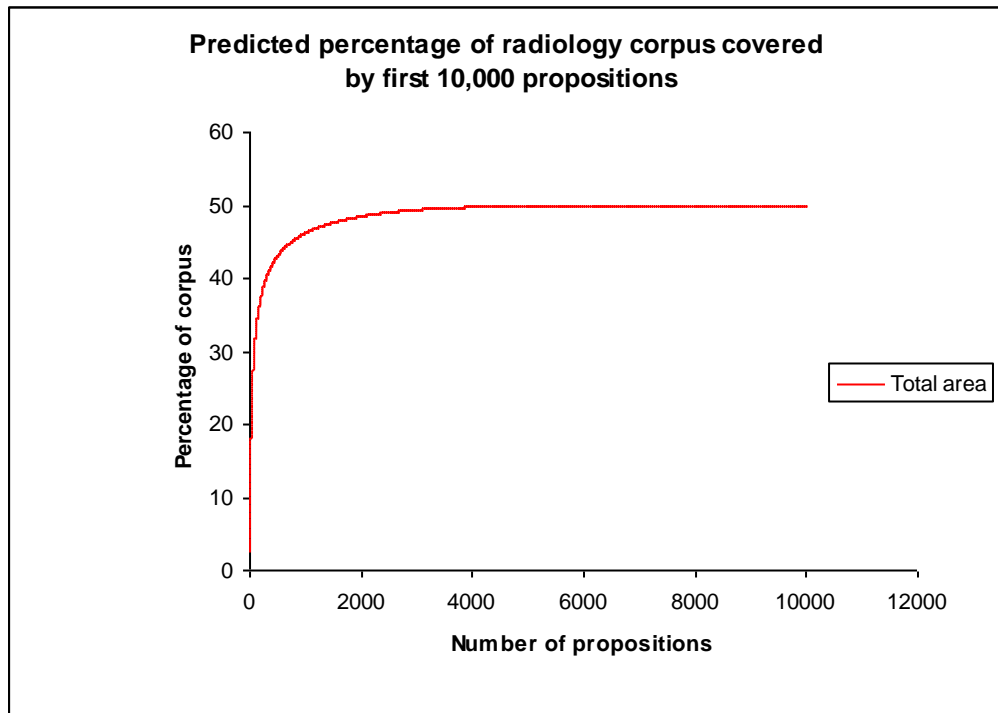


Figure 7: Prediction of corpus coverage by 10,000 propositions.

Discussion:

Semantics and Zipf's Law:

Currently, medical natural language processors do not adequately extract the semantics of medical reports at the sentence level. Many NLP systems like MedLEE identify key phrases using simple “chunkers” that are based on part of speech tagging. Syntactical and/or semantic analysis is mostly performed using rule base technology. However, we propose that semantic analysis of free text can be replaced by stochastic method or statistical modeling to facilitate the scaling of semantic extraction in a complex domain. Rule based technology is efficient for knowledge extraction of a limited number of concepts or analyzing a small and simple sentence. As the sentences grow in complexity, it is impractical to add rules because the entire system becomes unmanageably complex. Returning to the example given by Jamieson, the NLP system that wants to semantically interpret the phrase ‘*There are no new breast masses to suggest malignancy*’ must make it clear which terms are modifiers and what terms they modify, or placing the predicates into a more general structure. Only the NLP system with deep understanding of a medical domain can perform the semantic analysis completely. It is very easy to delete some portion of the sentence where the granular analysis is not intended in the NLP system, but eliminating a significant part from the sentence doesn’t reflect system’s semantic analytical ability at the sentence-level. For example, if the system is only interested in clinical finding like ‘*breast masses*’ from the sentence ‘*There are no new breast masses to suggest malignancy*’ it might ignore the

malignancy portion. Thus, leaving a significant chunk of the sentence from processing degrade the overall NLP power of the system. It is also important to know that a sentence is not only a ‘bag of words’; rather, it is a combination of ‘inter-related words or concepts’ that express the full semantic of the sentence. For example, considering the above radiological phrase, sentential semantics is not the aggregation of each individual term in the sentence; instead, it is the reproduction of the same complex concept represented in different ways in medical reports. Finding this similarity in meaning among various sentences in a medical domain avoids ambiguity and misclassification of codes [White, M. D., Kolar, L. M., & Steindel, S. J. 1999] for data mining, decision support and interoperability.

Why Zipf’s Law:

We have calculated the frequency of symbols (propositions) according to their rank in this paper and the result shows a decrease in frequency as the ranks increase and interested in predicting the number of propositions required to map the corpus, we tested our data using Zipf’s law to see if the trend of the line also follows the Zipf’s law. Originally, this law was utilized in bibliometrics (library and information science) for quantitative analysis and statistics to describe the pattern of words occurring in a decreasing frequency manner within the text. The rank of a word on that text multiplied by its frequency will equal a constant. Zipf’s law is often used to predict the frequency of words within a text. This law shows the probability of occurrence of words or other items starts high and tapers off. Thus, a few items occur very often while many others occur rarely. However, Zipf’s law can also model the World Wide Web surfing process [Cunha,

C. R., Bestavros, A., Crovella, M. E. 1995] to show the relationship between page “hits” and page rank. Figure 8 shows a Zipf distribution for incoming page requests to www.sun.com during a one-month period [Nielsen, J. 1997]

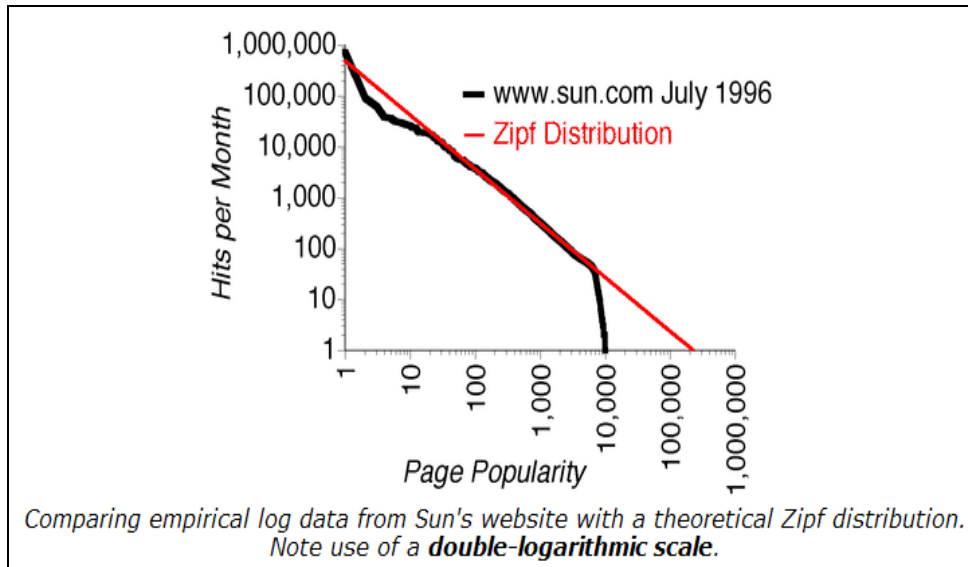


Figure 8: Zipf distribution for incoming page requests to www.sun.com during a one-month period

Baroni (2006) showed that Zipf-Mandelbrot law correctly modeled the frequency distribution of words in various corpora. The Brown Corpus of Present-Day American English consists of 1,014,312 words of running text of edited English prose printed in the United States during the calendar year 1961. The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written. The latest edition is the *BNC XML Edition*, released in 2007. ‘The world of the Wars’ was written by Herbert George wells in 1897 consists of 60,308 words. This novel was written in response to several historical events.

The la Repubblica corpus has 325,290,035 tokens of Italian newspaper text. The corpus of Japanese webpage contains 2,175,736 tokens.

The usefulness of the Zipf's law in figuring out the required symbols in representing sentential knowledge of the partial corpus is a unique and important outcome of this research. Zipf's law stated that, in a corpus of natural language utterances, the frequency of any word is roughly inversely proportional to its rank in the frequency table. So, the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc. Zipf's law is a consequence of independently categorizing items, and rank ordering the categories.

Therefore, it can be applied to many natural distributions [Wheeler, E. S., 2002] and falls under a more general rubric of scaling phenomena. Property of the subject matter is trivial at this point.

We propose that instead of using 'words' as symbols representing the semantic meaning in a corpus, 'propositions' better model the semantic meaning of sentences in the corpus. We use propositions as our atomic unit and calculate the rank-frequency distribution. The graph of rank-frequency distribution in figure 13 actually shows that the semantic distribution of the corpus follows Zipf's law except in the higher frequency area. The Zipf's-Mandelbrot formula has been applied to make the curve more look like a straight line as Zipf mentioned in his theory. The green solid line and red dotted line using Zipf-Mandelbrot law shows the current and future trend of the semantic distribution. At present, there are little over 5,600 propositions in the knowledge base and the green solid line follows the Zipfian distribution. If this distribution pattern continues in the future (red dotted line), it will be easier for us to predict the approximate number of propositions

required to cover the radiology corpus (at least 70- 80% of the corpus is the target). The deviation of the actual empirical data at the low end is due to a variety of factors, including the fact the corpus is not fully mapped yet to have enough accumulated propositions of low-frequency interest.

Prediction of estimating the probability of semantic symbols in a certain corpus is unique from another aspect. Most of the existing coding systems like SNOMED CT, CPT, ICD - 9 CM, do not try to derive the symbols empirically from a particular corpus at the sentential semantic level. Sometimes unclear meanings of some of the concepts fail to represent the semantics of the sentence found in free text form within a corpus and thus make the usage of coding system difficult. For example, the phrase '*right heart murmur*' might suggest EITHER Cardiac murmur with PMI on right chest wall OR murmur diagnosed as originating from right side of heart [Wilcke, 2000]. Also, we do not know exactly how many codes are necessary to cover all the information in the clinical domain, though SNOMED CT has 368,000 codes primarily designed to cover all aspects of pathological concepts. Their coding system does not define the code 'open sea' or how it is related to a medical concept. The lack of a gloss or 'use case' makes it difficult to apply the codes correctly. The designers of SNOMED CT frequently create circumstances where two different symbols can describe the same semantic meaning, e.g., '*fractured dislocation*' and '*dislocation with fracture*' which makes it extremely difficult to find out the proper number of symbols needed to cover the semantics of any domain.

Zipf's law can determine the rank-frequency distribution of propositions through calculating the probability of the semantic occurrence of a particular proposition by simply dividing the frequency of similar semantic sentences mapped to given proposition by total number of mapped lines. This law is not deterministic in nature; it completely depends upon empirical study of the domain contains lot of training data to examine the probability distribution. Other power laws like Lotka's law and Bradford's law are very domain specific, which are not applicable to serve the purpose of this research paper.

Shannon's Law and Semantic Entropy:

Entropy is a measure of the randomness of a random variable. It is also a measure of the amount of information of a random variable. The field of information theory was developed by Claude Shannon in the 1940s [Manning, C. D. a. S. H., 2000.] He was interested in the problem of maximizing the amount of information that one can transmit over an imperfect communication channel such as a noisy phone line. For any 'information' source and 'communication channel', Shannon wanted to determine theoretical limit of data compression, i.e., the least information required for maximum communication- which turns out to be given by entropy 'H'¹. If all the symbols are equiprobable, then the entropy or uncertainty reaches its maximum. With a variable probability, the uncertainty of occurrence of symbols diminishes.

However, According to Weaver [Shannon, C. E. a. W., W.,1949], the 'information' in Shannon's theory of communication can be semantically valid or invalid; in fact, a non sense message (at word level) contains same weight as the semantically valid message. Shannon used Zipf's law to calculate the entropy of English

text that outputs words independently with Zipf's probabilities [Schroeder, M., 2002]. Later, others have used this formula to measure semantic entropy, i.e., the measurement of semantic ambiguity and unformativeness of words in text corpora [Melamed, I. D., 1998]. Yarowsky (1993) compared the entropy of homophones (aid/aide, censor/sensor, cellar/seller, etc.) based on different conditional contexts. Resnik (1995) suggested that measuring conceptual semantic similarity using information content provides quite reasonable results, significantly better than the traditional method of simply counting the number of intervening is-a links (Gold *is a* metal, dime *is a* coin, etc.). The main limitation of all these papers is their emphasis on semantic entropy analysis from the lexical point of view; none of these researches have attempted to measure the semantic informativeness (entropy or uncertainty) at the sentence level.

Existing medical natural language processing systems depend heavily on lexicons and outside commonsense knowledge rather than an empirical analysis of the domain to formulate their semantic knowledge bases. This research paper attempts to calculate the least number of propositions (symbols) required to represent the semantic information of a medical corpus by calculating sentential semantic probability and entropy of the corpus using Zipf's law and Shannon's theory of communication. Refinement of this entropy calculation depends upon more complete annotation of our corpus.

Limitation of the study:

This research paper is focused on estimation of required propositions to cover an entire corpus, however, certain limitations with respect to the analysis and data that may affect the accuracy of the results.

- A fully built knowledge base is not available yet.
- Current trend of the semantic frequency distribution line in fig. 5a doesn't guarantee to follow Zipf's law; thus, current estimation of the propositions might become incorrect in future.
- The system is still challenged by the lumping and splitting of the information in the knowledge base which may affect the number of required propositions to map the entire corpus.
- Semantic mapping needs medical editorial skill to insure accuracy. It might affect the frequency counting.

Conclusion:

The need for an automated biomedical NLP system has been realized by the medical professionals for past several years to provide inexpensive but quality patient care with an improved decision support system. Using binary classification and FOPL, most of the existing Medical NLP systems are interested in extracting presence or absence of certain clinical findings from medical reports and coding against SNOMED, UMLS. On the other hand, limited efforts have been made to codify medical concepts through semantic understanding of the free text. As a NLP system, MedLEE has somewhat succeeded in capturing 'deep surface' meaning from comparatively simple sentences of the medical reports, but there are several instances where it fails to pull out the correct understanding of the complex sentences with multiple medical concepts. By extracting underlying meaning from each sentence of each medical report, Rdx Editor is the first tool to attempt structuring the free text and codify it at the sentence level.

Semantic analysis has been made possible only by the extensive knowledge of the domain and computational linguistics.

Besides finding the gaps in the current NLP systems, this paper specifically addresses the need for measuring the symbols or propositions to semantically cover a medical corpus like radiology using Zipf's-Mandelbrot law. This law has been variously used to count the word frequency, thus retrieving most frequent words from a given corpus. Applying it to a medical corpus to find the most frequent proposition that has the maximum number of semantically equivalent sentence is a unique attempt in the world of NLP research. In addition, attempting to use Shannon's theory of communication in calculation of semantic entropy of a small medical corpus like radiology is another important aspect of this research paper. Both Zipf's law and Shannon's theory are helpful in deriving the most frequent or informative proposition from the corpus. The most distinct part of this entire research paper is finding a way to estimate the number of propositions to cover a small corpus like radiology. From the current trend of the frequency line (fig. 5a), it has been predicted that to semantically map this particular corpus, we need around 78,000 -84,000 propositions in our knowledge base. If this radiology corpus follows the Zipf's law, which is currently going in that direction, calculating the total number of symbols will be easier for the researcher to predict the required number of codes for other medical domain with no hassle.

Appendices:

Appendix 1

Case Structure Grammar
alla Fillmore

Basic Idea:

- There are a finite number of Cases
- Each "Verb Sense"
 - accepts a subset of these Cases
 - this subset is partitioned into
 - the obligatory set
 - and
 - the optional set

Examples of elements of the set of Cases:

- AGENTIVE agent or instigator
- INSTRUMENTAL causally involved
- DATIVE the affected entity
- FACTITIVE the result
- LOCATIVE the location
- etc.

Example:

GIVE
{AGENTIVE,DATIVE,OBJECTIVE,
{TIME,LOCATIVE,FREQ) }

John gave the book to Jim.

↓

give	agent	John	WHAT was the event?
agent	dative	Jim	WHO did the event?
dative	objective	book	to WHOM was it done?
objective	time	<i>past</i>	WHAT was involved in the event?
time			WHEN was the event done?

↑

Jim was given the book by John.


Appendix 2

The screenshot displays the OpenGALEN Browser interface. At the top, there are navigation links for 'Home', 'Forum', and 'Browser'. The main header reads 'OpenGALEN Browser'. Below this, there are search and filter controls: a 'BodySystem' dropdown set to 'Heart', a 'hierarchical downwards browser' dropdown, and a search box containing 'Heart' with 'Search' and 'Top' buttons. On the left side, a tree view shows the hierarchy: 'Heart' (expanded) with sub-items 'FoetalHeart', 'UniventricularHeart', 'Pneumocardia', and 'HeartAllograft'. On the right side, the 'FoetalHeart' concept is detailed. It includes a 'Definition (necessary and sufficient)' section with a link to 'Heart' and a relationship 'isStructuralComponentOf Foetus'. Below this is a 'Conventional (necessary) criteria' section listing numerous relationships such as 'hasSpecificSurfaceDivision WholeSurfaceOfHeart', 'hasFunctionalComponent Myocardium', and 'isServedBy CoronaryVein'.

GALEN representation of concepts:

Left hand side is the terminology representation and right hand side implies various relationships associated with terminologies

Appendix 3



UMLSKS Version 5.0 UMLS Releases: 2002 2002AB 2002AC 2002AD 2003AA 2003AB 2003AC 2004AA 2004AB 2004AC 2005AA 2005AB 2005AC 2006AA 2006AB 2006AC

UMLS Knowledge Source Server (UMLSKS)

[Home](#) [Advanced Search](#) [Logout](#)

[Metathesaurus](#) [Semantic Network](#) [SPECIALIST Lexicon](#)

Metathesaurus Search for: **Murmur** in UMLS Release **2006AC**

Concept

- Definition
- Synonyms
- Other Languages
- Suppressible Synonyms
- Sources

Context

- Ancestors
- Parents
- Siblings
- Children

Relations

- Narrower
- Broader
- Similar
- Other
- Related and possibly synonymous
- Source asserted synonymy
- Allowable Subheadings
- Associated Expressions

Co-occurring Concepts

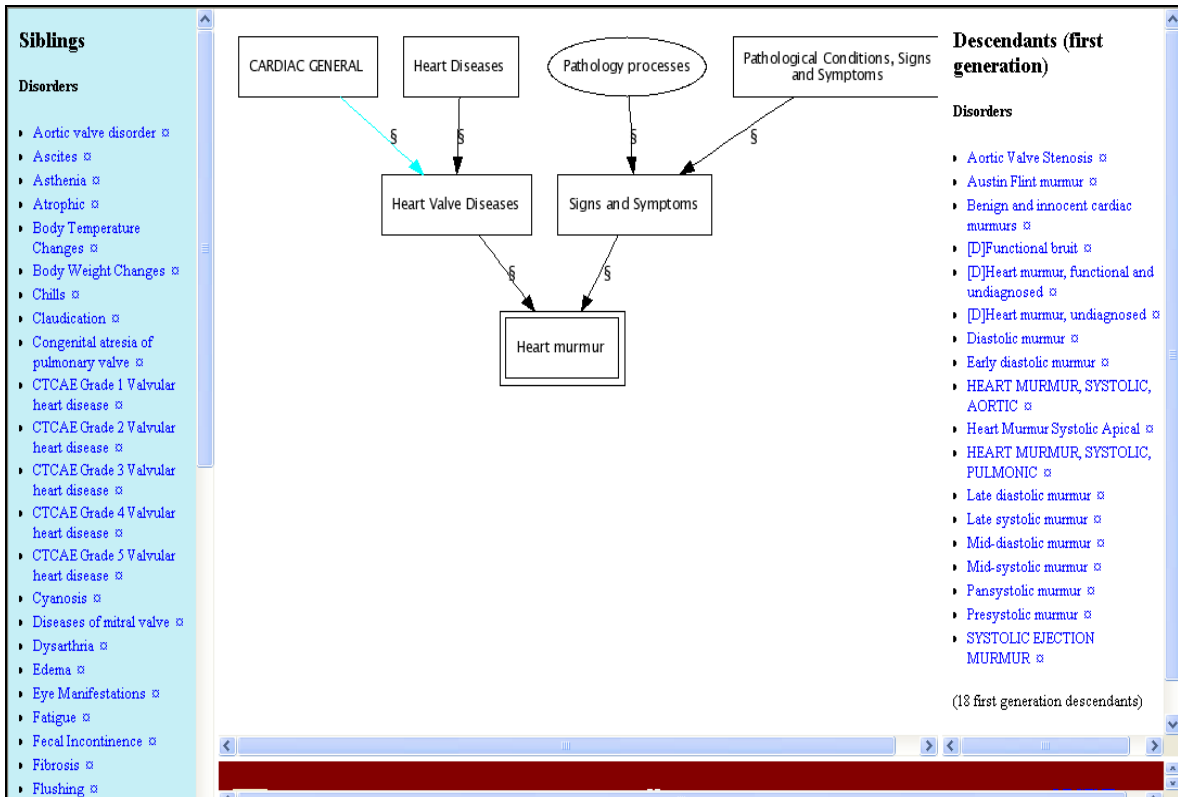
- Co-occurring MeSH
- Co-occurring AI/RHEUM

Concept: Heart murmur
CUI: [C0018808](#)
Semantic Type: [Sign or Symptom](#)

Definition:
Heart sounds caused by vibrations resulting from the flow of BLOOD through the HEART. They are classified by time of occurrence and by the intensity of sound on a scale of I to VI. They may be normal or abnormal. ([MeSH](#))

Synonyms:
[Heart murmur](#)
[\[D\]Heart murmur](#)
[\[D\]Heart murmur NOS \(context-dependent category\)](#)
[Cardiac murmur, unspecified](#)
[Cardiac Murmurs](#)
[Finding of heart murmur \(finding\)](#)
[Heart/arterial murmur nos](#)
[Heart murmur \(finding\)](#)
[Heart murmur \[D\] \(finding\)](#)
[heart valve; murmur](#)
[Murmur](#)
[Murmur \(finding\)](#)
[Observation of heart murmur](#)

UMLSKS Methathesaurus



UMLS Semantic Network



UMLS Knowledge Source Server (UMLSKS)

UMLSKS Version: 5.0 UMLS Releases: 2002 2002AB 2002AC 2002AD 2003AA 2003AB 2003AC 2004AA 2004AB 2004AC 2005AA 2005AB 2005AC 2005AD

[Metathesaurus](#) [Semantic Network](#) [SPECIALIST Lexicon](#)

[Logout](#)

About the UMLSKS

- ▶ [Home](#)
- ▶ [Overview](#)
- ▶ [Frequently Asked Questions](#)
- ▶ [Edit Views/Profile](#)

Downloads

- ▶ [UMLS Knowledge Sources](#)
- ▶ [RxNorm Files](#)
- ▶ [Mappings](#)
- ▶ [Developer's API](#)

Documentation

- ▶ [User's Guide](#)
- ▶ [Developer's Guide](#)
- ▶ [Developer's API Javadocs](#)
- ▶ [UMLS Documentation Set](#)
- ▶ [RxNorm Documentation](#)

Resources

- ▶ [NLP & Lexical Resources](#)
- ▶ [Semantic Network Resources](#)
- ▶ [Metathesaurus Resources](#)

Specialist Lexical Record

```
{base=heart murmur
entry=E0201273
cat=noun
variants=reg
variants=uncount
}
```

View [heart murmur](#) in relational format.

UMLSKS SPECIALIST Lexicon

Appendix 4

Case: Cerclage for Cervical Incompetence

A 30-year old female presents to the outpatient surgery center for placement of a cervical cerclage to treat an incompetent cervix. She is currently at 14 weeks gestation.

PMH: The patient's previous pregnancy was terminated due to her incompetent cervix, resulting in a second-trimester spontaneous abortion. The patient has had no further complications. Her current pregnancy has been otherwise uncomplicated.

PMSH: Patient is a Caucasian female, nonsmoker, ETOH negative. She is currently unemployed.

HPI: The patient is G2P0Ab1 (Gravida: 2 Parity: 0 Abortus:1) Prenatal records and tests have been received for review. Cervical exam and obstetric ultrasound confirm the diagnosis of incompetent cervix; Cervix length < 25 mm.

Procedure: Under epidural anesthetic, a band of strong 0.2 in (5 mm) suture thread was stitched around the cervix, and the thread was tightened to hold the cervix firmly closed.

The patient tolerated the procedure well and was discharged with instructions.

Pre-procedure Diagnosis: Pregnancy complicated by incompetent cervix

Post-procedure Diagnosis: same

Maternal care for cervical incompetence, second trimester	O34.32
--	--------

Diagnosis codes derived by ICD-10 CM from a medical report

Appendix 5

MedLEE input:

The screenshot displays the 'WWW MedLEE' web interface. On the left side, there are three vertical panels for category selection: 'D/C' with radio buttons for 'new', 'ex.1', 'ex.2', 'ex.3', and 'ex.4'; 'CXR' with radio buttons for 'new', 'ex.1', 'ex.2', and 'ex.3'; and 'Mammo' with radio buttons for 'new', 'ex.1', 'ex.2', and 'ex.3'. The main content area is titled 'RADIOLOGY REPORT' and contains the following elements:

- Control fields: 'FORMAT' set to 'indented', 'PARSE MODE' set to 'best', and a checkbox for 'Show positive only'.
- A button labeled 'Click here to process the report by MedLEE'.
- A text area with three sections:
 - CLINICAL INFORMATION:** followed by the placeholder text '<replace with your sentences or delete the whole section >'. This section is currently empty.
 - .DESCRIPTION:** followed by the text 'There is increased atelectasis with blue pleural effusion.'.
 - .IMPRESSION:** followed by the placeholder text '<replace with your sentences or delete the whole section >'. This section is currently empty.
- A 'Clear' button at the bottom of the text area.

MedLEE output:

Output Generated by MedLEE

```
procedure:replacement
  idref>> 7
  parsemode>> mode5
  sectname>> report clinical information item
  sid>> 1
  code>> UMLS:C0035139_surgical replantation
    idref>> [7]

procedure:division - action
  idref>> 23
  parsemode>> mode4
  region>> whole
    idref>> 21
  sectname>> report clinical information item
  sid>> 1
  code>> UMLS:C1293097_division - action
    idref>> [23]

problem:atelectasis
  change>> increase
    idref>> 47
  idref>> 49
  parsemode>> mode2
  sectname>> report clinical information item
  sid>> 1
  code>> UMLS:C0004144_atelectasis
    idref>> [49]

problem:pleural effusion
  certainty>> high certainty
    idref>> 51
  descriptor>> blue
    idref>> 54
  idref>> 56
  parsemode>> mode2
  sectname>> report clinical information item
  sid>> 1
  code>> UMLS:C0032227_pleural effusion disorder
    idref>> [56]
  code>> UMLS:C1253943_pleural effusion fluid
    idref>> [56]
```

Appendix 6

```
set ANSI_NULLS ON
set QUOTED_IDENTIFIER ON
GO
-- =====
-- Author:          Lopa
-- Create date:    02/20/2007
-- Description:    <freq count>
-- =====
ALTER PROCEDURE [dbo].[freqcount]
■ Add the parameters for the stored procedure here
@current_date datetime
AS
BEGIN
-- SET NOCOUNT ON added to prevent extra result sets from
-- interfering with SELECT statements.
SET NOCOUNT ON;

■ Insert statements for procedure here
select sum (med.freq) as totfreq, map.propid into #t1
from maplines map, medline med, proposition p
where
map.medlineid = med.id and
map.propid = p.id
group by map.propid

select totfreq, p.statement from #t1, proposition p
where
#t1.propid = p.id and
@current_date >= 02202007
order by totfreq desc
END
```

SQL source code for semantic frequency count

Appendix 7

F	G	H	I	J	K	L	M	N	O
K=0(1through 99 original value)	log10 K=0	K=7	log10 K=7	K=6.3	log10 K=6.3	K=5.5	log10 K=5.5	k=4	log10 K=4
141851	5.151832402	133011.8424	5.123890309	160238.2748	5.176780588	175317.5	5.24382523	248524.9	5.39537
107811	5.032663074	113725.3951	5.055857454	126656.1052	5.102626129	144933.4	5.16116851	195010.9	5.290059
77666	4.890230938	98855.3076	4.994999992	108872.3683	5.03691767	122708	5.08887282	158861.8	5.20102
68402	4.8350688	87085.87515	4.939947721	95044.38896	4.977926483	105834.6	5.0246276	133011.8	5.12389
62026	4.792573775	77569.13251	4.889688935	84024.44683	4.924405662	92644.2	4.96681822	113725.4	5.055857
61126	4.786225977	69735.72707	4.843455333	75063.0477	4.875426194	82086.52	4.91427186	98855.31	4.995
57812	4.762017994	63190.19628	4.800649704	67651.48114	4.830277309	73469.94	4.86610967	87085.88	4.939948
55519	4.744441635	57649.90783	4.760798617	61433.18926	4.788403062	66321.77	4.82165608	77569.13	4.889689
53658	4.729634481	52907.8867	4.723520415	56151.37414	4.749360389	60308.78	4.78038055	69735.73	4.843455
53392	4.727476189	48809.33901	4.688502926	51616.72982	4.712790478	55189.8	4.74185883	63190.2	4.80065
49547	4.695017364	45236.35046	4.65548756	47686.85428	4.678398675	50786.27	4.70574635	57649.91	4.760799
41723	4.620375528	42097.63541	4.624257703	44252.7274	4.645940042	46963.38	4.67175939	52907.89	4.72352
40717	4.609775772	39321.50191	4.594630098	41229.56672	4.615208771	43617.58	4.63966159	48809.34	4.688503
35802	4.553907288	36850.91959	4.56644833	38550.52185	4.586030261	40668.11	4.60925396	45236.35	4.655488
35710	4.552789885	34639.99545	4.539577826	36162.22294	4.558255119	38051.11	4.58036736	42097.64	4.624258
31383	4.496694457	32651.41222	4.51390197	34021.58353	4.531754524	35715.51	4.55285684	39321.5	4.59463
30594	4.485636262	30854.53746	4.48931904	32093.46561	4.506416617	33619.96	4.52659724	36850.92	4.566448
28964	4.461858539	29224.00815	4.46573978	30348.94845	4.482143648	31730.71	4.50147973	34640	4.539578
27066	4.432424077	27738.65755	4.443085439	28764.02772	4.458849699	30019.89	4.4774091	32651.41	4.513902
27000	4.431363764	26380.69212	4.421286185	27318.62561	4.436458846	28464.37	4.45430155	30854.54	4.489319
26995	4.431283332	25135.05328	4.40027981	25995.82898	4.414903671	27044.75	4.43208298	29224.01	4.46574
24296	4.385534779	23988.91771	4.380010655	24781.29699	4.394124033	25744.68	4.41068751	27738.66	4.443085
22724	4.356484781	22931.30248	4.360428723	23662.79584	4.374066057	24550.27	4.39005633	26380.69	4.421286
22069	4.343782655	21952.75037	4.341488939	22629.83023	4.354681296	23449.67	4.37013671	25135.05	4.40028
21503	4.332499055	21045.07708	4.323150521	21673.34893	4.335926023	22432.68	4.35088121	23988.92	4.380011
19863	4.298044843	20201.16667	4.305376452	20785.50771	4.317760637	21490.52	4.33224696	22931.3	4.360429
19456	4.289053558	19414.80497	4.288133032	19959.47718	4.300149161	20615.56	4.31419512	21952.75	4.341489
18810	4.274388796	18680.54286	4.271389493	19189.28584	4.283058812	19801.15	4.2966904	21045.08	4.323151
18127	4.258325935	17993.58359	4.255117666	18469.69117	4.266459634	19041.48	4.2797006	20201.17	4.305376
17639	4.24647396	17349.68918	4.239291699	17796.07308	4.250324181	18331.43	4.26319628	19414.8	4.288133
14791	4.169997537	16745.10242	4.223887808	17164.34541	4.234627246	17666.5	4.24715048	18680.54	4.271389
12943	4.112034951	16176.48147	4.208884064	16570.88191	4.219345622	17042.7	4.23153841	17993.58	4.255118
12917	4.11116166	15640.84468	4.194260203	16012.45417	4.2044579	16456.49	4.21633722	17349.69	4.239292
12859	4.109207196	15135.52405	4.179997462	15486.1792	4.18994428	15904.71	4.20152585	16745.1	4.223888
12374	4.092510112	14658.1255	4.166078436	14989.47503	4.175786423	15384.55	4.18708481	16176.48	4.208884
12231	4.087461966	14206.4951	4.152486946	14520.02296	4.161967303	14893.47	4.17299602	15640.84	4.19426
11857	4.07397482	13778.69009	4.139207932	14075.73525	4.14847109	14429.22	4.15924271	15135.52	4.179997
11813	4.072360204	13372.95393	4.126227348	13654.72745	4.135283036	13989.73	4.14580928	14658.13	4.166078
11532	4.061904634	12987.69483	4.113532076	13255.29456	4.122389383	13573.17	4.13268118	14206.5	4.152487
10686	4.02881517	12621.46713	4.101109841	12875.89047	4.109777274	13177.86	4.11984483	13778.69	4.139208
10637	4.026819159	12272.95512	4.088949146	12515.1101	4.097434675	12802.29	4.10728756	13372.95	4.126227
10622	4.026206297	11940.95895	4.077039205	12171.67389	4.085350308	12445.07	4.09499749	12987.69	4.113532
10569	4.024033898	11624.38234	4.065369886	11844.41428	4.073513589	12104.96	4.08296348	12621.47	4.10111
10012	4.000520841	11322.22178	4.053931658	11532.26389	4.061914572	11780.81	4.07117508	12272.96	4.088949
9454	3.975615598	11033.55705	4.042715545	11234.24514	4.050543896	11471.56	4.05962246	11940.96	4.077039
8844	3.951696308	10753.51888	4.032471605	10945.12418	4.039273214	11176.22	4.04829122	11624.38	4.06537

Table showing different K values to be tested in Zipf's-Mandelbrot Law

Appendix 8

<u>Frequency</u>	<u>Ranks</u>	<u>Probability</u>	<u>log₂ (C)</u>	<u>Entropy</u>	<u>Max. probability</u>	<u>Log₂ (col.F)</u>	<u>Max.Entropy</u>
144654	1	0.050764286	-4.300042315	-0.218288577	0.0001	-13.28771238	-0.001328771
108225	2	0.037980041	-4.718614732	-0.17921318	0.0001	-13.28771238	-0.001328771
77940	3	0.027351946	-5.192212696	-0.142017123	0.0001	-13.28771238	-0.001328771
68416	4	0.024009632	-5.380242869	-0.129177654	0.0001	-13.28771238	-0.001328771
62048	5	0.021774872	-5.521191919	-0.12022325	0.0001	-13.28771238	-0.001328771
61266	6	0.021500441	-5.539489965	-0.119101475	0.0001	-13.28771238	-0.001328771
58061	7	0.020375691	-5.617007207	-0.114450403	0.0001	-13.28771238	-0.001328771
55539	8	0.019490631	-5.681075426	-0.110727742	0.0001	-13.28771238	-0.001328771
53821	9	0.018887723	-5.726407431	-0.108158795	0.0001	-13.28771238	-0.001328771
53675	10	0.018836486	-5.730326341	-0.107939212	0.0001	-13.28771238	-0.001328771
49816	11	0.017482224	-5.837967443	-0.102060656	0.0001	-13.28771238	-0.001328771
42563	12	0.014936886	-6.064976786	-0.090591866	0.0001	-13.28771238	-0.001328771
40769	13	0.014307307	-6.127104057	-0.087662358	0.0001	-13.28771238	-0.001328771
36345	14	0.012754766	-6.292819722	-0.080263445	0.0001	-13.28771238	-0.001328771
35844	15	0.012578947	-6.312844984	-0.079408945	0.0001	-13.28771238	-0.001328771
31800	16	0.011159763	-6.485549862	-0.072377196	0.0001	-13.28771238	-0.001328771
30709	17	0.010776891	-6.535915094	-0.070436847	0.0001	-13.28771238	-0.001328771
29393	18	0.01031506	-6.599104012	-0.068070152	0.0001	-13.28771238	-0.001328771
27621	19	0.009693201	-6.688811075	-0.064835992	0.0001	-13.28771238	-0.001328771
27335	20	0.009592834	-6.703827252	-0.064308699	0.0001	-13.28771238	-0.001328771
26995	21	0.009473515	-6.72188441	-0.063679875	0.0001	-13.28771238	-0.001328771
24594	22	0.008630918	-6.856270231	-0.059175908	0.0001	-13.28771238	-0.001328771
23057	23	0.00809153	-6.949371814	-0.056231048	0.0001	-13.28771238	-0.001328771
22228	24	0.007800604	-7.002198482	-0.054621376	0.0001	-13.28771238	-0.001328771
21503	25	0.007546175	-7.050038675	-0.053200828	0.0001	-13.28771238	-0.001328771
20124	26	0.007062235	-7.145659533	-0.050464324	0.0001	-13.28771238	-0.001328771
19900	27	0.006983625	-7.161808196	-0.050015383	0.0001	-13.28771238	-0.001328771
19207	28	0.006740426	-7.21294443	-0.048618321	0.0001	-13.28771238	-0.001328771
19073	29	0.006693401	-7.223044844	-0.048346735	0.0001	-13.28771238	-0.001328771
19055	30	0.006687084	-7.224407019	-0.048310217	0.0001	-13.28771238	-0.001328771
14791	31	0.005190693	-7.589857033	-0.03939662	0.0001	-13.28771238	-0.001328771
13118	32	0.004603578	-7.763028847	-0.035737705	0.0001	-13.28771238	-0.001328771
13062	33	0.004583925	-7.769200814	-0.035613435	0.0001	-13.28771238	-0.001328771
13060	34	0.004583223	-7.76942173	-0.035608994	0.0001	-13.28771238	-0.001328771
12956	35	0.004546726	-7.780956254	-0.035377875	0.0001	-13.28771238	-0.001328771
12468	36	0.004375469	-7.836346567	-0.034287693	0.0001	-13.28771238	-0.001328771
11960	37	0.004197194	-7.896359238	-0.033142549	0.0001	-13.28771238	-0.001328771
11906	38	0.004178243	-7.902887828	-0.033020187	0.0001	-13.28771238	-0.001328771
11730	39	0.004116478	-7.924373614	-0.032620513	0.0001	-13.28771238	-0.001328771
11004	40	0.003861699	-8.016548582	-0.030957497	0.0001	-13.28771238	-0.001328771
10773	41	0.003780633	-8.047156568	-0.030423344	0.0001	-13.28771238	-0.001328771
10732	42	0.003766244	-8.052657668	-0.030328277	0.0001	-13.28771238	-0.001328771
10651	43	0.003737819	-8.063587739	-0.030140228	0.0001	-13.28771238	-0.001328771
10215	44	0.003584811	-8.123887423	-0.029122597	0.0001	-13.28771238	-0.001328771
9527	45	0.003343367	-8.224482733	-0.027497461	0.0001	-13.28771238	-0.001328771
9225	46	0.003237384	-8.270955811	-0.02677626	0.0001	-13.28771238	-0.001328771
8899	47	0.003122979	-8.322861496	-0.02599212	0.0001	-13.28771238	-0.001328771
8832	48	0.003099466	-8.33376455	-0.025830221	0.0001	-13.28771238	-0.001328771
8753	49	0.003071742	-8.346727152	-0.025638994	0.0001	-13.28771238	-0.001328771

Portion of the table showing Entropy calculation according to Shannon's formula

References:

- ICD-10 Corner. from <http://www.ingenixonline.com/content/icd10/structure.asp>
- The Burden of Chronic diseases and Their Risk Factors. (2004). *CDC-National and State Perspectives*(February).
- Allen, J. (1995). *Natural Language Understanding* (Second ed.): The Benjamin/cummings Publishing Company, Inc.
- Baroni, M. (2006). 39 Distributions in text. from http://sslmit.unibo.it/~baroni/publications/hsk_39_dist_rev2.pdf
- Bates, D. W., Ebell, M., Gotlieb, E., Zapp, J., & Mullins, H. C. (2003). A proposal for electronic medical records in U.S. primary care. *J Am Med Inform Assoc*, *10*(1), 1-10.
- Bates, D. W., Kuperman, G. J., Wang, S., Gandhi, T., Kittler, A., Volk, L., et al. (2003). Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc*, *10*(6), 523-530.
- Bates, M. (1995). Models of natural language understanding. *Proc Natl Acad Sci U S A*, *92*(22), 9977-9982.
- Baud, R. H., Rassinoux, A. M., & Scherrer, J. R. (1992). Natural language processing and semantical representation of medical texts. *Methods Inf Med*, *31*(2), 117-125.
- Baud, R. H., Rassinoux, A. M., Wagner, J. C., Lovis, C., Juge, C., Alpay, L. L., et al. (1995). Representing clinical narratives using conceptual graphs. *Methods Inf Med*, *34*(1-2), 176-186.
- Baud, R. H. R., Jean-Marie; Wagner, Judith C; Rassinoux, Anne-Marie; Lovis, Christian; Rush, Philippe; Trombert-Paviot, Beatrice; Scherrer, Jean-Raoul.(1997). Validation of concept Representation Using Natural Language Generation. AMIA Annual Fall Symposium (formerly SCAMC),841.
- Bird, S. K., E.; Loper, E. (2005). NLTK: Introduction to Natural Language Processing. from <http://nltk.sourceforge.net/tutorial/introduction/index.html>
- Bodenheimer, T., & Grumbach, K. (2003). Electronic technology: a spark to revitalize primary care? *Jama*, *290*(2), 259-264.

-
- Calzolari, N. (2003). Natural Language Processing and Knowledge Engineering. Proceedings. 2003 International Conference on Volume , Issue , 26-29 Oct. 2003 Page(s): 16 – 18.
- CDC. (2004). The Burden of Chronic Diseases and Their Risk Factors. *National and State Perspectives* (February).
- Chen, H., Fuller, S.S., Friedman, C. & Hersh, W. (2005). *Medical Informatics: Knowledge Management and Data Mining in Biomedicine* (Vol. 2): Springer.
- Chute, C. (2005). *Advances in Knowledge Management and Data Mining in Biomedicine*. New York: Springer-Verlag.
- Cimino, J. J., Clayton, P. D., Hripcsak, G., & Johnson, S. B. (1994). Knowledge-based approaches to the maintenance of a large controlled medical terminology. *J Am Med Inform Assoc, 1*(1), 35-50.
- Cunha, C. R., Bestavros, A. , Crovella, M. E. (1995). Characteristics of WWW Client-based Traces.
- Doyle, P. (1997). AI Qual summary - Natural Language [Electronic Version].
- Eiselt, K. H., Jennifer. (1998). Augmented Transition Networks.
- Evans, R. S., Classen, D. C., Pestotnik, S. L., Lundsgaarde, H. P., & Burke, J. P. (1994). Improving empiric antibiotic selection using computer decision support. *Arch Intern Med, 154*(8), 878-884.
- Friedman, C., & Hripcsak, G. (1999). Natural language processing and its future in medicine. *Acad Med, 74*(8), 890-895.
- Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc, 11*(5), 392-402.
- Giere, W. (2004). Electronic Patient Information -- Pioneers and MuchMore. A vision, lessons learned, and challenges. *Methods Inf Med, 43*(5), 543-552.
- Greenes, R.A. (2003). Decision support at the point of care: Challenges in knowledge representation, management and patient-specific access. *Adv Dent Res 17*:69-73.
- Gruber, T. (1993). A translation approach to portable ontologies. *Knowledge Acquisition, 5*(2):199-220,

-
- Hripcsak, G., Friedman, C., Alderson, P. O., DuMouchel, W., Johnson, S. B., & Clayton, P. D. (1995). Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med*, 122(9), 681-688.
- Humphreys, B. L., Lindberg, D., Schoolman, H.M., and Barnett, G.O. (1998). The Unified Medical Language System: An Informatics Research Collaboration. *J Am Med Inform Assoc.*, 5 (1)(Jan–Feb).
- Inc., T. A. I. (2001). Integrated Development Environments for Natural Language Processing [Electronic Version].
- Jackson, P. (1999). *Introduction to Expert Systems* (2nd ed.): Addison-Wesley Longman Limited.
- Jamieson, P.W. (2003). *Process for constructing a semantic knowledge base using a document corpus*. Unpublished manuscript, Indianapolis.
- Jamieson, P. W. (2004). USA Patent No.
- Jamieson, P. W. (2006). USA Patent No.
- Jamieson, P. W. (2006). *Representing and Extracting Knowledge from Free Text Medical Records*. Unpublished manuscript, Indianapolis.
- Joachims, T. (2001). *Learning to classify text using support vector machines: methods, theory and algorithms.*: KAP.
- Johnson, S. B. (1999). A semantic lexicon for medical language processing. *J Am Med Inform Assoc*, 6(3), 205-218.
- Knibbs, G. H. (1929). The International classification of Disease and Causes of Death and its revision. *Medical Journal of Australia*.
- Knobby, G. H. (1929). The International classification of Disease and Causes of Death and its revision: *Medical Journal of Australia*.
- Kuperman, G. J., & Gibson, R. F. (2003). Computer physician order entry: benefits, costs, and issues. *Ann Intern Med*, 139(1), 31-39.
- Lee, B.-S. B., Barrett R. (2002). Contextual Knowledge Representation for Requirements Documents in Natural Language. *American Association for Artificial Intelligence*.
- Mack, R. e. a. (2004). Text analytics for life science using the Unstructured Information Management Architecture. *IBM Systems Journal*(September).

-
- Manning, C. D. a. S. H. (2000). *Foundation of Statistical Natural Language Processing* (2nd ed.): The MIT Press Cambridge, MA.
- McEnery, T. (2003). *Corpus Linguistics*: Oxford University Press.
- Melamed, I. D. (1998). Measuring Semantic Entropy [Electronic Version] from <http://citeseer.ist.psu.edu/cache/papers/cs/1921/ftp:zSzzSzftp.cis.upenn.eduzSzpubzSzmelamedzSzpaperszSzSemEnt.pdf/measuring-semantic-entropy.pdf>.
- Nielsen, J. (1997). Zipf Curves and Website Popularity. from www.useit.com
- Nielson, J., Wilcox, A. (2004). Linking Structured Text to Medical Knowledge. MEDINFO. IMIA.
- Nirenburg, S. R., V. (2004). *Ontological Semantics*: The MIT Press, Cambridge, MA.
- Parunak, V. (1995). Case Grammar: A linguistic tool for engineering Agent-Based Systems [Electronic Version] from www.iti.org/~van.
- Ramsay, A. (2003). *Discourse*: Oxford University Press.
- Rebholz-Schuhmann, D., Kirsch, H., & Couto, F. (2005). Facts from text--is text mining ready to deliver? *PLoS Biol*, 3(2), e65.
- Rector, A. L. (1999). Clinical terminology: why is it so hard? *Methods Inf Med*, 38(4-5), 239-252.
- Rector, A. L. (1999). Terminology and concept representation languages: where are we? *Artif Intell Med*, 15(1), 1-4.
- Rector, A. L., & Nowlan, W. A. (1994). The GALEN project. *Comput Methods Programs Biomed*, 45(1-2), 75-78.
- Rector, A. L., Rogers, J. E., Zanstra, P. E., & Van Der Haring, E. (2003). OpenGALEN: open source medical terminology and tools. *AMIA Annu Symp Proc*, 982.
- Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language [Electronic Version], 11, 95-130 from <http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume11/resnik99a.pdf>.
- Rindflesch, T. C., & Fisman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform*, 36(6), 462-477.

-
- Schank, R. C. T., L. (1969). *A Conceptual Dependency Parser for Natural Language*.
Paper presented at the International Conference On Computational Linguistics;
Proceedings of the 1969 conference on Computational linguistics
- Schmidt, C. F. Case Grammar. from
http://www.rci.rutgers.edu/~cfs/305_html/Understanding/CaseGram1.html
- Schroeder, M. (2002). Power laws: from Alvarez to Zipf. *Glottometrics*, 4, 39-44.
- Shannon, C. E. a. W., W. (1949). *The Mathematical Theory of Communication* (Illini
book edition, 1963 ed.): University of Illinois Press, Urbana and Chicago.
- Shapiro, S. C. (1982). Generalized Augmented Transition Network Grammars for
Generation from Semantic Networks. *American Journal of Computational
Linguistics*, 8(1).
- Stuhlinger, W. H., Oliver; Stoyan, Herbert; Muller, Michael. Intelligent Data Mining for
Medical Quality Management.
- Weizenbaum, J. (1966). ELIZA - A computer program for the study of natural language
communication between man and machine.
- Wheeler, E. S. (2002). Zipf's Law and why it works everywhere. *Glottometrics*, 4, 45-48.
- White, M. D., Kolar, L. M., & Steindel, S. J. (1999). Evaluation of vocabularies for
electronic laboratory reporting to public health agencies. *J Am Med Inform Assoc*,
6(3), 185-194.
- Wilcke, J. R. a. H., Allen W. (2000). Evaluating Performance for Summarizing
Veterinary Cardiovascular Findings. [Electronic Version].
- Wilson, D. a. S., D. (2002). Relevance Theory. In G. a. H. Ward, L. (Ed.), *Handbook of
Pragmatics* (pp. 607-632): Oxford:Blackwell.
- Yarowsky, D. (1993). *One sense per collocation*. Paper presented at the DARPA
Workshop on Human Language Technology.
- Zarri, G.P. (1996). NKRL, a Knowledge Representation Language for Narrative
Natural Language Processing.

CURRICULUM VITAE

NAME: Lopamudra Chatterjee

EDUCATION: MS, Health Informatics, IUPUI, 2008

BA, Education, University of Calcutta, India, 1995

HONORS, AWARDS, FELLOWSHIP: Recipient of the Guidant Fellowship 2006 for advance research in data mining and natural language processing.

RESEARCH and TRAINING EXPERIENCE: Advance research in Natural Language Processing and statistical analysis of radiological knowledge base.

WORK EXPERIENCE: Data Analyst, March 2008 – present, MDWise Hoosier Alliance

Data Analyst, 2007, American Health Data Institute

Graduate intern, May 2006 – September 2006, Logical

Semantics, Inc.

PUBLICATION (in progress): Usefulness of Nursing Management Minimum Data Set (NMMDS) in finding rare nursing management articles that meets the nurse executive's need to produce accurate, reliable, and useful data for decision making.