Filtering the UMLS® Metathesaurus® for MetaMap

2001 Edition

Alan R. Aronson

May 9, 2001

1. Overview

The MetaMap program's primary purpose is to discover the Metathesaurus concepts referred to in arbitrary text. A given Metathesaurus concept can have many alternative names (Metathesaurus strings) which originate in the many source vocabularies included in the Metathesaurus. As the number of strings has grown over the years, MetaMap's performance has suffered. In 2001, for example, the Metathesaurus has 1,462,202 English strings, 1,457,129 (99.7%) of them distinct, comprising 797,359 concepts representing a 9% increase over the 2000 edition. Many of these strings are of little value to MetaMap for one of four reasons. First, some strings either represent generic, nonmedical concepts or are unnecessarily ambiguous. Second, some strings are virtually indistinguishable from each other. For efficiency, only one representative of a set of indistinguishable strings is needed. Third, some Metathesaurus strings have internal structure which makes it highly unlikely to find them in regular text. Fourth, some strings, including lengthy descriptions of things such as procedures, health activities or medical devices, are so complicated that it is again unlikely to find them in normal text. Corresponding to the four classes of strings are four filtering methods for discovering and removing them: manual filtering, lexical filtering, filtering by type, and syntactic filtering. These methods are discussed in sections 2-5. Then section 6 describes ways to selectively combine the filtering methods to produce a range of alternative views of the Metathesaurus appropriate for various purposes.

2. Manual Filtering

A small number of Metathesaurus strings are problematic and have been suppressed before performing other forms of filtering. There are 6,561 such strings, almost 900% more than in 2000.

Note, however, that the growth from 2000 occurs partly because of a more thorough analysis of ambiguous terms but mainly because of the inclusion of terms containing NEC:

- Numbers (e.g., 2, +1, -4, 98.734, 50000) [144 occurrences];
- Single alphabetic characters (e.g., a, A, b, B <3>; note the ambiguity designator) [128 occurrences];
- Unnecessarily ambiguous terms [976 occurrences]
  "Other <2>", … (however, "Other" itself is retained)
  "Protocols <1>', … (however, "Protocols" itself is retained)
  …
- Special cases [1 occurrence]
  "Periods" for "Menstruation" (from RCD99)
- Contextual terms, i.e., terms whose meaning can only be understood within the context of their vocabulary [5,312 occurrences]
  All terms containing NEC or an expanded form

Although a few numbers correspond to biomedical entities ("98.734" has semantic types 'Steroid' and 'Pharmacologic Substance'), they generally have semantic types 'Quantitative Concept' or 'Intellectual Product'. Similarly, the single alphabetics generally mean the letter itself (the concept for "a" is "Lower case ay") and have semantic type 'Intellectual Product'. Several single alphabetics, however are biomedical: "B" has concept "Boron" with semantic types 'Pharmacologic Substance', 'Biologically Active Substance' and 'Element, Ion, or Isotope'. The third class of problematic strings are the unnecessarily ambiguous ones (see *Ambiguity in the UMLS® Metathesaurus®* for details). A final class of special cases includes the string "0" for "TPBF protein" and "Periods" for "Menstruation". The first example is subsumed by the numbers class, and the second example arises because the word *periods* occurs in biomedical text but has been associated with "Menstruation" due to slang usage.

3. Lexical Filtering

Lexical filtering is the most benign type of filtering and consists of removing strings for a concept which are effectively the same as another string for the concept. Properties which can make strings effectively the same are:

- non-essential parentheticals;
- Metathesaurus multiple meaning designators;
- NOS variation;
- syntactic uninversion;
- case variation;
- hyphen variation; and
- possessives.

Lexical filtering is accomplished by normalizing all strings for a given concept and removing all but one string for each set of strings that normalize to the same thing.

3.1 Non-essential parentheticals

Non-essential parentheticals are parenthetical expressions within a string which provide meta information about the string. As such they are not useful for text processing. Non-essential parentheticals can occur at the left or right end of a string and can be delimited by either parentheses or brackets. For example the concept “Anemia, Hemolytic” has synonyms “[X]Haemolytic anemias” and “[X]Hemolytic anemias” both of which contain the left parenthetical “[X]”. Previous editions of the Metathesaurus only contained right parentheticals which seemed to be relatively well-behaved in the sense that a string without the parenthetical was almost always present in the set of strings for a given concept. Thus, “Drug Toxicity (Non MeSH)” had a string “Drug Toxicity”. Now right parentheticals are much less well-behaved and only a few left parentheticals can be reliably removed without altering the string’s meaning. These left parentheticals come from the Read Codes: “[X]”, “[V]”, “[D]”, “[M]”, “[EDTA]”, “[SO]” and “[Q]”. These are the only parentheticals declared to be non-essential and removed from strings. The problem of detecting non-essential parentheticals has changed as the Metathesaurus has matured. The current practice of removing the few left parentheticals listed above is by no means adequate. The problem requires further analysis.

3.2 Metathesaurus multiple meaning designators

Strings such as “Cold <1>” and “Cold <2>” end with a multiple meaning designator, i.e., a number within angle brackets. These designators are essentially parenthetical expressions and are likewise removed. (Note that for 2000, each string “AAA <n>” ending with a multiple meaning designator has a corresponding string “aaa” without it where “AAA” and “aaa” differ only by case. There are three exceptions to this rule in 2001.[1])

3.3 NEC/NOS variation

Many of the Metathesaurus vocabularies incorporate the acronyms NEC (Not Elsewhere Covered) and NOS (Not Otherwise Specified) into their terms. Examples include “Psychotherapy, NEC”, “Abdomen, NOS”, “X-RAY NEC AND NOS”, and “INJURY NEC/NOS”. As with case variation, the presence of NEC and/or NOS does not generally have a significant effect on the meaning of the term. The argument for ignoring NEC/NOS variation is not as strong as that for case variation, but it still seems reasonable for most text processing.

3.4 Syntactic uninversion

Inversion refers to the practice of inverting words of a term and inserting a comma to signal the inversion. It is normally done to index the original term under each of its important words and thereby make it more accessible. Inverted forms of a term, however, are not useful for processing text since inverted forms rarely appear in text. The concept “1,4-alpha-Glucan Branching Enzyme” has some interesting inversions. It has a synonym “Branching Enzyme” with inversion “Enzyme, Branching”, and it also has a synonym “Starch Branching Enzyme” with two inver-

1. The strings **‘Antitussive <2>**’, ‘**Antigonadotrophins, antiestrogens, antiandrogens, not elsewhere classified <1>’** and **‘Other psychotropic drugs, not elsewhere classified <1>’** have no corresponding string without the ambiguity designator.

sions, “Branching Enzyme, Starch” and “Enzyme, Starch Branching”. The process of uninversion simply undoes inversion, i.e., it searches for a comma followed by a space, inverts the term at that point and removes the comma and space. Syntactic uninversion is just uninversion which is inhibited if the term contains a preposition or conjunction. This prevents terms such as “Biological Phenomena, Cell Phenomena, and Immunity” or “Legal blindness, as defined in U.S.A.” from being incorrectly uninverted. Note that the concept “1,4-alpha-Glucan Branching Enzyme” mentioned earlier is also not uninverted because the comma within it is not followed by a space; embedded commas do not call for uninversion.

3.5 Case variation

Two strings which differ from each only because of case variation normally refer to the same thing. For example, the concept “Abdomen <1>” has strings “Abdomen”, “abdomen” and “ABDOMEN” which differ from each other only by case. Similarly, the concept “beta-Alanine” has strings, “beta Alanine”, “beta alanine” and “BETA ALANINE”, which differ from each other only by case. Note, however, that case *does* matter for some aspects of text processing. Text containing the pronoun *us* is not referring to the acronym *US* for the *United States*; and the verb *aids* does not refer to the disease *AIDS*. Despite this observation, case almost never matters within the limited context of all strings for a given concept.

3.6 Hyphen variation

As with case variation, the presence of a hyphen rather than a space normally means little especially in the context of all strings for a given concept. For example, the concept “1,4-alpha-Glucan Branching Enzyme” used in the last section has a variant “1,4 alpha Glucan Branching Enzyme” in which both hyphens have been replaced by spaces.

3.7 Possessives

Alternatives such as “Down’s Syndrome” and “Down Syndrome” or “American Nurses’ Association” and “American Nurses Association” differ only by a possessive.

4. Filtering by Type

Some Metathesaurus strings can be filtered out based solely on their type. For example, strings with a Term Status (TS) of lowercase s are suppressible synonyms; they are an abbreviated form of another term. “Abdomen” and “Abdomen <2>” are such synonyms of “Malignant neoplasm of abdomen.” About 9% (127,736) of the 1,462,202 English Metathesaurus strings are suppressible synonyms.

Similarly, some Term Type (TTY) values indicate strings which are normally inappropriate for text processing. There are 180,845 unique occurrences of strings with such term types in the Metathesaurus. (Actually there is significant overlap between the inappropriate term types and the suppressible synonyms: only 56,929 (31%) of the 180,845 TTY strings are *not* already suppressible synonyms.) The filtered out term types together with examples for each are given in the next

section. For completeness, this is followed by a section of questionable types and a final section of good types.

4.1 Filtered out Types

- AA (Attribute type abbreviation) [34 occurrences]
  "Route administration of drug" (for concept "Drug Administration Routes")
  "Type-partial denture connector" (for concept "Type of partial denture connector")

- AB (Abbreviation in any source vocabulary) [75,402 occurrences]
  "Cluster of diff antigen 73" (for term "Cluster of differentiation antigen 73" of concept "5'-Nucleotidase"
  "AIDS dementia complex" (for concept "AIDS Dementia Complex"; note that in this case the AB string is not unusual, but it is redundant)

- CS (Short component process in ICPC, i.e. include some abbreviations) [13 occurrences]
  "Med exam/health evalua/complete"
  "Microbio/other immunol test"

- HX (Expanded version of short hierarchical term) [5,577 occurrences]
  "D40-D44 ABDOMEN" (for concept "Abdomen"; the HX form occurs in SNMI98 and has an HT form "ABDOMEN")
  "SECTION 2 CONGENITAL ANOMALIES" (for concept "Congenital Abnormality"; the HX form occurs in SNMI98 and has an HT form "CONGENITAL ANOMALIES")

- LN (LOINC official fully specified name) [24,194 occurrences]
  "ALMECILLIN:SUSC:PT:ISLT:ORDQN:MIC"
  "ALMECILLIN:SUSC:PT:ISLT:ORDQN:AGAR DIFFUSION"

- LO (Obsolete official fully specified name) [536 occurrences]
  "AUREOBASIDIUM PULLULANS AB.IGE:ACNC:PT:SER:QN"
  "PARROT AUSTRALIAN DROPPINGS AB.IGE:ACNC:PT:SER:QN"

- LX (Official fully specified name with expanded abbreviations) [24,730 occurrences]
  "ALMECILLIN:SUSCEPTIBILITY:POINT IN TIME:ISOLATE:QUANTITATIVE:MINIMUM INHIBITORY CONCENTRATION"
  "ALMECILLIN:SUSCEPTIBILITY:POINT IN TIME:ISOLATE:SEMI-QUANTITATIVE:BACTERIAL SENSITIVITY (KIRBY-BAUER)"

- OA (Obsolete abbreviation) [49,131 occurrences]
  "Gastrin secretion abnorm.NOS"
  "Spontaneous abort.incomp.NOS"

- PS (Short forms that needed full specification) [3,348 occurrences]
  "Cranial nerves" (for concept "Benign neoplasm of cranial nerves")
  "Bladder, unspecified" (for concept "Malignant neoplasm of bladder, NOS")

4.2 Questionable Types

The Term Types listed in this section are not actually filtered out during type filtering, but they present problems for processing text adequately.

- BD Fully-specified drug brand name that can be prescribed [13,856 occurrences]
  These terms describe a quantity of some drug; they may or may not occur in text.
  "Adalat, 10 mg oral capsule"
  "Afrin, 0.05% nasal spray"
- CC (Trimmed ICPC component process) [1 occurrence]
  "Referral primary care provider"
- CD (Clinical Drug) [44,518 occurrences]
  Like BD terms, these terms often describe a quantity of some drug and may or may not occur in text.
  "Lavender Oil"
  "Hydrogen Peroxide Soln 3%"
  "Isopropyl Alcohol 70%"
- CO (ICPC component names (these are hierarchical terms, as opposed to the LOINC component names which are analytes) [56 occurrences]
  Like topical qualifiers (TQ) below, these terms seem to have a meaning too specific to be useful for most biomedical applications
  "ACTIVITY COMPONENT"
  "CARDIAC COMPONENT"
- DS (Short form of descriptor) [508 occurrences]
  "AOD abuse"
  "biological AOD dependence"
- ES (Short form of entry term) [36 occurrences]
  "periodic light AOD use"
  "AOD tax"
- GO (Goal) [311 occurrences]
  Like orders (OR) below, the terms are full utterances.
  "Mobility, exercise, and activity will increase to optimal or return to baseline."
  "Patient's activity tolerance will increase or progress."
- IS (Obsolete synthesized term in the Read Thesaurus; used to be filtered out) [3,644 occurrences]
  "Fall - accidental" (for concept "Accidental Falls")
  "MVTA - Motor vehicle traffic accident" (for concept "Accidents, Traffic")
- LS (Expanded system/sample type (The expanded version was created for the Metathesaurus and includes the full name of some abbreviations.)) [806 occurrences]
  About one-third of these terms contain embedded periods.
  "HEART.ATRIA"
  "AORTA.THORACIC.ASCENDING"
- NS (Short form of non-preferred term) [200 occurrences]
  "neonatal AOD abstinence syndrome"
  "dysfunctional AOD use"
- OM (Obsolete modifiers in HCPCS; used to be filtered out) [17 occurrences]
  "MICROSURGERY"

"MEDICAL DIRECTION OF OWN EMPLOYEE(S) BY ANESTHESIOLOGIST (NOT MORE THAN FOUR EMPLOYEES)"

- OR (Orders) [1,357 occurrences]
  All terms are from PCDS97 and are full utterances.
  "Discharge patient."
  "Use assistive devices to maintain required position."
- PX (Expanded preferred terms (pair with PS)) [4,773 occurrences]
  These terms often contain characters indicating a superscript or subscript.
  "Ca^2+^-transporting ATPase"
  "alpha>1< Antichymotrypsin"
- SA (Short forms of activities) [608 occurrences]
  "Adhere to agency protocol for donor screening and acceptance"
  "Administer agents to expand intravascular volume, as appropriate"
- SN (Official component synonym in LOINC) [314 occurrences]
  Most of these terms are abbreviatory.
  "ETOH"
  "O NOS AG"
- ST (Step) [132 occurrences]
  "Manage contracts"
  "Order laboratory tests"
- SX (Mixed-case component synonym with expanded abbreviations) [448 occurrences]
  Like PX above, these terms contain special characters to indicate superscripts or subscripts.
  24,25-Dihydroxyvitamin D>3<
  Vitamin B>4<
- TA (Task) [170 occurrences]
  "Introduce self to patient and explain services"
  "Determine patient's primary spoken language and communications ability/limitations"
- TQ (Topical qualifier) [82 occurrences]
  The meaning of these terms is specific to MeSH indexing and may not be appropriate for general use.
  "abnormalities"
  "administration & dosage"

4.3 Good Types

This section contains the remaining types, i.e., those appropriate for text processing.<

- AC (Nursing activities) [9,119 occurrences]
  "Monitor blood pressure"
  "Control bleeding"
- AD (Adjective) [879 occurrences]
  "Anorexic"
  "Scarred"
- AS (Attribute type synonym) [25 occurrences]
  "Precipitating factor"
  "Px - Prescription"

- AT (Attribute type) [819 occurrences]
  "Allergen"
  "Association"
- BN (Fully-specified drug brand name that can not be prescribed) [9,183 occurrences]
  "Parlodel"
  "Aminocaproic Acid"
- CE (Entry "term" to a Supplementary Chemical "term") [189,136 occurrences]
  "2 bromolysergic acid diethylamide"
  "BOL 148"
- CL (Class) [14 occurrences]
  "Managing the Practice"
  "Ensuring Appropriate Pharmacotherapy"
- CN (LOINC official component name) [10,447 occurrences]
  "DIPALMITOYLPHOSPHATIDYLCHOLINE"
  "BEHAVIOR"
  "???LEAD"
- CP (ICPC component process (in original form)) [25 occurrences]
  "Administrative procedure"
  "Other therapeutic procedure"
- CX (Component process in ICPC with abbreviations expanded) [4,508 occurrences]
  "NOS ANTIBODY"
  "CANCER ANTIGEN 125"
- DE (Descriptor) [10,808 occurrences]
  "synthetic 11-hydroxycorticosteroids"
  "abdomen"
- DI (Disease name) [2,088 occurrences]
  "ABETALIPOPROTEINEMIA"
  "ABORTION, SPONTANEOUS"
- DO (Domain) [4 occurrences]
  "Health Systems Management"
  "Ensuring Appropriate Therapy and Outcomes"
- DT (Definitional term, present in the Metathesaurus because of its connection to a Dorland's definition or to a definition created especially for the Metathesaurus) [176 occurrences]
  "Acetylcholinesterase <1>
  "Animal"
- DX (Diagnosis) [146 occurrences]
  "Anxiety"
  "Blood Pressure Alteration"
- EN (MeSH nonprint entry "term") [28,578 occurrences]
  "(131)I-Macroaggregated Albumin"
  "Injuries, Abdominal"
- EP (Entry "term") [21,082 occurrences]
  "Dipalmitoyllecithin"
  "Branching Enzyme"

- ET (Entry “term”) [40,4307 occurrences]
  “MPTP”
  “Abelson’s virus”
- EX (Expanded form of entry term) [36 occurrences]
  “periodic light Alcohol or Other Drugs use”
  “Alcohol or Other Drugs tax”
- FI (Finding name) [5,016 occurrences]
  “ABDOMINAL PAIN, CRAMPY”
  “ABDOMINAL DISTENTION”
- FN (Full form of descriptor) [20 occurrences]
  “Data sources and data collection methods.”
  “Lorrs’s Inpatient Multidimensional Psychiatric Rating Scale”
- GN (Generic drug name) [2,101 occurrences]
  “mesna”
  “aminocaproic acid”
- GT (Glossary “term”) [4,797 occurrences]
  “SYNDROME ABDOMINAL ACUTE”
  “ABDOMINAL CRAMP”
- HC (Hierarchical class) [54 occurrences]
  “Behavior Therapy”
  “Cognitive Therapy”
- HS (Short hierarchical term (needed expansion) in ICD 10) [28 occurrences]
  “Agents primarily acting on smooth and skeletal muscles and the respiratory system”
  “Bacterial vaccines”
- HT (Hierarchical term) [21,185 occurrences]
  “Abdomen”
  “Abdominal pain”
- ID (Nursing indicator) [2,576 occurrences]
  “Ankylosed joints”
  “Appetite loss”
- IN (Name for an intervention) [11,567 occurrences]
  “mesna”
  “BETA-ALANINE”
- IT (Index “term”, i.e., derived from the index to any non-MeSH source vocabulary) [2,077 occurrences]
  “ACUTE ABDOMEN”
  “CRAMP ABDOMINAL”
- IX (Expanded forms of indicators (embedded abbreviations expanded)) [257 occurrences]
  “Systolic blood pressure”
  “Exercise stress test within normal limits”
- MD (CCS multi-level diagnosis categories) [691 occurrences]
  “Abdominal pain”
  “Congenital anomalies”
- MH (Main heading) [19,942 occurrences]
  “1,2-Dipalmitoylphosphatidylcholine”
  “Abdomen”

- MM (Metathesaurus string created to distinguish different meanings of the same lexical string.) [12,840 occurrences]
  "17-hydroxysteroid dehydrogenase <1>"
  "DOPS <1>"
- MP (Preferred names of modifiers) [790 occurrences]
  "ABNORMALITY"
  "Abortion (termination of pregnancy)"
- MS (Multum names of branded and generic supplies or supplements) [4,338 occurrences]
  "Acetone"
  "0.3cc Syringe 29g 1/2"“
- MT (An alternate form of a concept name from one of the source vocabularies created for the Metathesaurus) [104 occurrences]
  "coma"
  "incontinence of stool"
- N1 (Chemical Abstracts Service Type 1 name of a chemical) [22,904 occurrences]
  "1,4-alpha-D-Glucan:1,4-alpha-D-glucan 6-alpha-D-(1,4-alpha-D-glucano)-transferase"
  "1,1,3-Propanetricarboxylic acid, 3-amino-"
- NM (Supplementary chemical "term", a name of a substance) [114,857 occurrences]
  "2-bromolysergic acid diethylamide"
  "3-hydroxyproline"
- NP (Non-preferred term) [5,428 occurrences]
  "3,4-methylenedioxyamphetamine"
  "congenital defects"
- NX (Expanded form of non-preferred term) [200 occurrences]
  "neonatal Alcohol or Other Drugs abstinence syndrome"
  "dysfunctional Alcohol or Other Drugs use"
- OC (Nursing outcomes) [193 occurrences]
  "Thermoregulation"
  "Decision Making"
- OP (Obsolete preferred term) [72,018 occurrences]
  "Carbenoxolone sodium"
  "Acute abdomen"
- OS (System-organ class in the WHO Adverse Reaction Terminology) [32 occurrences]
  "AUTONOMIC NERVOUS SYSTEM DISORDERS"
  "PSYCHIATRIC DISORDERS"
- PC (Preferred "trimmed" term in ICPC) [233 occurrences]
  "arthrogryposis multiplex congenita"
  "Bartter syndrome"
- PM (Machine permutation) [70,702 occurrences]
  "1,2 Dipalmitoylphosphatidylcholine"
  "Enzyme, Branching"
- PN (Metathesaurus preferred name) [10,004 occurrences]
  "17-Hydroxysteroid Dehydrogenases"
  "Droxidopa"

- PQ (Qualifier for a problem) [9 occurrences]
  "Family"
  "Health Promotion"
- PR (Name of a problem) [407 occurrences]
  "Placenta abruptio"
  "Dependency on alcohol"
- PT (Designated preferred name) [475,116 occurrences]
  "Dipalmitoylphosphatidylcholine"
  "Brancher enzyme"
- RN (Official component related name in LOINC) [6,872 occurrences]
  "ALPHA-1,4-GLUCAN BRANCHING ENZYME"
  "17-KGS"
- RT (Designated related "term") [6,061 occurrences]
  "20-Hydroxyprogesterone"
  "Lumpy jaw"
- RX (Alternate name of preferred name) [290 occurrences]
  "Aleutian disease virus"
  "Aluminum silicate"
- SC (Special Category term) [36 occurrences]
  "Congenital Malformations"
  "bandages"
- SD (CCS single-level diagnosis categories) [280 occurrences]
  "Abdominal pain"
  "Spontaneous abortion"
- SF (Synonym made by replacing ";" with no spaces around it with "," in ICPCP) [5,985 occurrences]
  "Cramps, abdominal"
  "Loss (of) appetite"
- SI (Name of a sign or symptom of a problem) [311 occurrences]
  "allergens"
  "anemia"
- SP (CCS single-level procedure categories) [231 occurrences]
  "Abortion (termination of pregnancy)"
  "Diagnostic amniocentesis"
- SS (Synonymous "short" forms) [196 occurrences]
  "adrenoleukodystrophy"
  "adrenomyeloneuropathy"
- SY (Designated synonym) [137,753 occurrences]
  "Branching enzyme"
  "Amylo-(1,4,6)-transglycosylase"
- TC (Term class) [61 occurrences]
  "ABDOMEN"
  "GI_NOS"
- TG (Name of the target of an intervention) [63 occurrences]
  "Behavior modification"
  "Communication"

- TX (CCPSS synthesized problems for TC termgroup) [61 occurrences]
  "ABDOMEN PROBLEM"
  "GI_NOS PROBLEM"
- VS (Value Set) [13 occurrences]
  "Report Priority Value Set"
  "Sex Value Set"
- XD (Expanded descriptor in AOD) [508 occurrences]
  "identification and screening for Alcohol or Other Drugs use"
  "Alcohol or Other Drug Disorder"
- XQ (Alternate name for a qualifier) [237 occurrences]
  "anomalies"
  "teratology"

5. Syntactic Filtering

The final kind of filtering considered here is based on a high-level syntactic parse of the Metathesaurus strings. Since normal MetaMap processing involves mapping the simple noun phrases found in text, it is highly unlikely that a complex Metathesaurus string will be part of a good mapping. For example, the concept "Accident caused by caustic and corrosive substances" has high-level syntactic analysis [[head],[verb],[prep,head],[conj],[mod,head]] which contains seven syntactic units (head, verb, etc.) broken into five simple phrases ([head], [verb], etc.) Any text which resembles the concept will be broken up into several phrases each of which is processed separately. Thus, the text might map to constituent concepts (such as "Accident"); but the entire text will not map to the full concept. The strictest form of syntactic filtering, then, would be to filter out any string consisting of more than one simple phrase. However some tractable strings with more than one simple phrase are not filtered out. As of 1999, for example, *of strings* such as "Acute necrosis of liver" and "Radical resection of tumor of soft tissue of leg area", which consist of a simple phrase followed by one or more *of* prepositional phrases, have not been excluded in syntactic filtering because of their tractability. In 2001 this condition was relaxed further to include phrases consisting of a simple phrase followed by any prepositional phrase followed by zero or more *of* prepositional phrases. An example of such a phrase is "Other operations on vessels of heart".

6. Filtered Metathesaurus Models

The filtering described in the previous sections can be selectively applied to provide different views of the Metathesaurus. Three such model are

- Strict Model: All forms of filtering, manual, lexical, type-based and syntactic, are applied. This view is most appropriate for semantic processing where the highest level of accuracy is needed. The Strict Model consists of 801,612 (55%) of the 1,462,202 English Metathesaurus strings;
- Moderate Model: Manual, lexical and type-based filtering, but not syntactic filtering, are used. This view is appropriate for term processing where input text should not be divided into simple

phrases but considered as a whole. The Moderate Model consists of 1,092,308 (75%) English Metathesaurus strings; and

- Relaxed Model: Only manual and lexical filtering are performed. This provides access to virtually all Metathesaurus strings and is appropriate for browsing. The Relaxed Model consists of 1,266,569 (87%) English Metathesaurus strings.