Filtering the UMLS® Metathesaurus® for MetaMap

2004 Edition

Alan R. Aronson

July 27, 2004

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 2004, for example, the Metathesaurus has 2,370,524 English strings, 2,351,545 (99.2%) of them distinct, comprising 1,008,958 concepts. There were 36% more English strings and 17% more concepts than in the 2003 edition, largely due to the inclusion of SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) for the first time this year. Many of the strings in the Metathesaurus are of little value to MetaMap for one of four reasons. First, some strings either represent general, 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 or meta information 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 number of Metathesaurus strings are problematic and have been suppressed before performing other forms of filtering. There are 15,451 such strings, 19% more than in 2003:

Unnecessarily ambiguous terms [5,205 occurrences]
Other <2> for Other location of complaint
Protocols <3> for Protocols: Urinary Elimination

Contextual terms, i.e., terms whose meaning can only be understood within the context of their vocabulary [8,873 occurrences]
All terms containing *NEC* or an expanded form
Brand names, i.e., all terms containing the word *brand* [1,116 occurrences]
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) [111 occurrences];
Special cases [2 occurrences]
Periods (from RCD) for Menstruation (slang)
BRA (from RCD) for Brain (abbreviation)

Note that the occurrence numbers for the individual categories above are approximate since some strings could be classified into more than one category.

The Metathesaurus staff has simply chosen some terms that metamorphosys always suppresses because they are thought to be inappropriate for any use. The unnecessarily ambiguous terms are terms that, for one reason or another, do not adequately describe the concept that contains them. They are given a Term Status (TS) of lowercase s (or p) and are fully discussed in the annual editions of *Ambiguity in the UMLS Metathesaurus*. Contextual terms are actually a specific kind of ambiguous term. They are identifiable by the presence of *NEC* or one of its expansions (e.g., not elsewhere classified). Brand names are problematic because they often consist of a common word (e.g., *Cold*) which almost never has the brand name meaning. 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 type Element, Ion, or Isotope . A final class of special cases includes the string Periods for Menstruation and BRA for Brain . Both of these are problematic because they are ambiguous with other concepts which occur far more frequently in biomedical text.

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.

3.3 *NOS* variation

Many of the Metathesaurus vocabularies incorporate the acronym *NOS* (*Not Otherwise Specified*) into their terms. Examples include Abdomen, NOS and X-RAY NEC AND NOS . As with case variation, the presence of *NOS* (except when also accompanied by *NEC*) does not generally have a significant effect on the meaning of the term. The argument for ignoring *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 inversions, 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; such 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 or p are suppressible synonyms or suppressible preferred names (see section 2 above). Abdomen and Abdomen <2> are such synonyms of Malignant neoplasm of abdomen. About 6% (144,034) of the 2,370,524 English Metathesaurus strings are suppressible synonyms. There are also a small number (133) of strings with TS of lowercase p.

Similarly, some Term Type (TTY) values indicate strings which are normally inappropriate for text processing, often because they are abbreviatory in nature, have some internal structure or include some meta information (see section 4.1 below). There are 506,775 unique occurrences of strings with such term types in the Metathesaurus. Actually there is a non-trivial overlap between

these inappropriate term types and the suppressible synonyms: 131,144 (26%) of the 506,775 TTY strings are already suppressible synonyms. Note that the percentage of overlap fell from 64% in 2003 due to the inclusion of the TTY FN (Full form of descriptor) which mainly consists of terms such as Acute abdomen (disorder) . There are 301,022 unique FN terms which include meta information but are not marked as suppressible.

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) [75,255 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)

CO (ICPC component names (these are hierarchical terms, as opposed to the LOINC component names which are analytes)) [57 occurrences]
Like topical qualifiers (TQ), these terms seem to have a meaning too specific to be useful for most biomedical applications
ACTIVITY COMPONENT
CARDIAC COMPONENT

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

DFA (Dose Form Abbreviation) [106 occurrences]
AER
CAP

FN (Full form of descriptor) [301,034 occurrences]
Dipalmitoylphosphatidylcholine (substance)
Acute abdomen (disorder)

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) [33,166 occurrences]
  ALMECILLIN:SUSC:PT:ISLT:ORDQN:MIC
  ALMECILLIN:SUSC:PT:ISLT:ORDQN:AGAR DIFFUSION

LO (Obsolete official fully specified name) [978 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) [34,113 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) [8,596 occurrences]
  Cranial nerves (for concept Benign neoplasm of cranial nerves)
  Bladder, unspecified (for concept Malignant neoplasm of bladder, NOS)

SB (Named subset of a source) [1 occurrence]
  US English Dialect Subset

UCN (Unique common name) [149 occurrences]
  algae <Chlorophyta>
  monkeys <#2>

USN (Unique scientific name) [667 occurrences]
  Acremonium <Hypocreaceae>
  Aedes <sugbenus>
  Bacillus <bacterium>

USY (Unique synonym) [37 occurrences]
  Chlamydia psittaci <Chlamydophila psittaci>
  AsGV<agrotis>

VAB (Versioned abbreviation) [113 occurrences]
  SPN02
  GO2002_09_01
  MSH2003_2002_10_24

XM (Cross mapping set) [4 occurrences]
  MSH Associated Expressions
  SNOMEDCT mappings to ICD-9-CM

XX (Expanded string) [69 occurrences]
superior frontal sulcus (human only)
ectocalcarine sulcus (macaque only)

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) [15,503 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) [93,321 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%
CP (ICPC component process (in original form)) [25 occurrences]
Administrative procedure
Other therapeutic procedure
DS (Short form of descriptor) [509 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.
HS (Short hierarchical term (needed expansion) in ICD 10) [32 occurrences]
Agents primarily acting on smooth and skeletal muscles and the respiratory system
Bacterial vaccines
IS (Obsolete synthesized term in the Read Thesaurus) [3,644 occurrences]
Fall - accidental (for concept Accidental Falls)
MVTA - Motor vehicle traffic accident (for concept Accidents, Traffic)
IX (Expanded forms of indicators (embedded abbreviations expanded)) [257 occurrences]
Systolic blood pressure
Exercise stress test within normal limits
LS (Expanded system/sample type (The expanded version was created for the Metathesaurus and includes the full name of some abbreviations.)) [1,237 occurrences]
About one-third of these terms contain embedded periods.

AORTIC VALVE
AORTA.THORACIC.ASCENDING
MP (Preferred names of modifiers) [197 occurrences]
ABNORMALITY
PROB
MT (An alternate form of a concept name from one of the source vocabularies created for the Metathesaurus) [104 occurrences]
coma
incontinence of stool
MV (Multi-level procedure category) [404 occurrences]
Adenoidectomy without tonsillectomy
Excision of semilunar cartilage of knee
NS (Short form of non-preferred term) [200 occurrences]
neonatal AOD abstinence syndrome
dysfunctional AOD use
NX (Expanded form of non-preferred term) [200 occurrences]
neonatal Alcohol or Other Drugs abstinence syndrome
dysfunctional Alcohol or Other Drugs use
OBD (Obsolete branded drug) [448 occurrences]
EC Naprosyn, 500 mg oral tablet
Intralipid, 10% intravenous suspension
OC (Nursing outcomes) [193 occurrences]
Thermoregulation
Decision Making
OCD (Obsolete clinical drug) [12,300 occurrences]
Adhesive Tape
Glucose Oral Soln 50%
OL (Non-current Lower Level Term) [9,070occurrences]
Lab test abnormality
Laboratory test abnormality
OP (Obsolete preferred term) [71,456 occurrences]
Carbenoxolone sodium [gastro-intestinal use]
Acute abdomen
OR (Orders) [1,357 occurrences]
All terms are from PCDS97 and are full utterances.
Discharge patient.
Use assistive devices to maintain required position.
OSN (Official short name) [24,680 occurrences]
Calculus Analysis
Drugs Ur Scn
Amoxicillin MIC
PR (Name of a problem) [407 occurrences]
Placenta abruptio
Dependency on alcohol
PX (Expanded preferred terms (pair with PS)) [10,021 occurrences]
These terms often contain characters indicating a superscript or subscript.

Ca^2+^-transporting ATPase
alpha>1< Antichymotrypsin
RAB (Root abbreviation) [116 occurrences]
DSM3R
MSH
RHT (Root hierarchical term) [38 occurrences]
DSM-III-R
MeSH
RPT (Root preferred term) [116 occurrences]
DSM-III-R
Medical Subject Headings
RSY (Root synonym) [28 occurrences]
Diagnostic and Statistical Manual of Mental Disorders: DSM-III-R
LCSH
SA (Short forms of activities) [608 occurrences]
Adhere to agency protocol for donor screening and acceptance
Administer agents to expand intravascular volume, as appropriate
SBD (Semantic branded drug) [13,707 occurrences]
Sodium Fluorescein 250 MG/ML Injectable Solution [AK-Fluor]
Acetaminophen 250 MG / Caffeine 30 MG / Chlorpheniramine 2 MG / Hydrocodone 5 MG / Phenylephrine 10 MG Oral Tablet [Hycomine Compound]
SBDF (Semantic branded drug and form) [10,445 occurrences]
Furacin Topical Cream
Dianeal Low Calcium with 1.5% Dextrose Intraperitoneal Solution
SC (Special Category term) [53 occurrences]
Congenital Malformations
bandages
SCD (Semantic Clinical Drug) [20,950 occurrences]
ichthammol 0.2 MG/MG Topical Ointment
Potassium Chloride 0.3 MG/ML / Sodium Chloride 6 MG/ML / Sodium Lactate 3.1 MG/ML Intravenous Solution
SCDC (Semantic Drug Component) [17,132 occurrences]
Iron-Dextran Complex 100 MG/ML
ALLERGENIC EXTRACT,HOUSE DUST 1 UNT
SCDF (Semantic clinical drug and form) [9,939 occurrences]
Triamcinolone Oral Paste
Ephedrine / Phenobarbital / Potassium Iodide / Theophylline Oral Tablet
SD (CCS single-level diagnosis categories) [281 occurrences]
Abdominal pain
Spontaneous abortion
SI (Name of a sign or symptom of a problem) [311 occurrences]
allergens
anemia
SN (Official component synonym in LOINC) [319 occurrences]
Most of these terms are abbreviatory.

ETOH
O NOS AG
SP (CCS single-level procedure categories) [231 occurrences]
Abortion (termination of pregnancy)
Diagnostic amniocentesis
SSN (Source short name, used in the UMLS Knowledge Source Server) [116 occurrences]
DSM-III-R
MeSH
ST (Step) [132 occurrences]
Manage contracts
Order laboratory tests
SX (Mixed-case component synonym with expanded abbreviations) [453 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
TC (Term class) [61 occurrences]
ABDOMEN
GI_NOS
TG (Name of the target of an intervention) [63 occurrences]
Behavior modification
Communication
TQ (Topical qualifier) [83 occurrences]
The meaning of these terms is specific to MeSH indexing and may not be appropriate for general use, but they are not currently being excluded because the Medical Text Indexer (MTI) includes them in its recommendations.
abnormalities
administration & dosage
TX (CCPSS synthesized problems for TC termgroup) [61 occurrences]
ABDOMEN PROBLEM
GI_NOS PROBLEM
VPT (Versioned preferred term) [113 occurrences]
Standard Product Nomenclature, 2002
Medical Subject Headings, 2002_10_24
VS (Value Set) [13 occurrences]
Report Priority Value Set
Sex Value Set
VSY (Versioned synonym) [24 occurrences]
Computer Retrieval of Information on Scientific Projects Thesaurus, 2002
LCSH, 1990
XD (Expanded descriptor in AOD) [509 occurrences]
identification and screening for Alcohol or Other Drugs use
Alcohol or Other Drug Disorder

XQ (Alternate name for a qualifier) [235 occurrences]
These terms are similar to TQ terms, and the same comments apply.
anomalies
teratology

4.3 Good Types

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

AC (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) [19,461 occurrences]
Parlodel
Aminocaproic Acid
CE (Entry term to a Supplementary Chemical term) [214,297 occurrences]
2 bromolysergic acid diethylamide
7S RNA
CL (Class) [14 occurrences]
Managing the Practice
Ensuring Appropriate Pharmacotherapy
CMN (Common name) [4,738 occurrences]
acanthocephalans
alfalfa
CN (LOINC official component name) [12,728 occurrences]
DIPALMITOYLPHOSPHATIDYLCHOLINE
BEHAVIOR
???LEAD *(sic)*
CX (Component process in ICPC with abbreviations expanded) [5,148 occurrences]
NOS ANTIBODY
CANCER ANTIGEN 125
DE (Descriptor) [10,817 occurrences]
synthetic 11-hydroxycorticosteroids
abdomen
DF (Dose Form) [141 occurrences]
Aerosol
24 Hour Transdermal Patch

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) [182 occurrences]
  Anxiety
  Blood Pressure Alteration
EN (MeSH nonprint entry term) [33,402 occurrences]
  (131)I-Macroaggregated Albumin
  Injuries, Abdominal
EP (Entry term) [23,887 occurrences]
  Dipalmitoyllecithin
  Branching Enzyme
EQ (Equivalent name) [1,629 occurrences]
  Borrelia burgdorferi sensu stricto
  Flavibacterium
ET (Entry term) [45,757 occurrences]
  MPTP
  Abelson s virus
EX (Expanded form of entry term) [809 occurrences]
  periodic light Alcohol or Other Drugs use
  Alcohol or Other Drugs tax
FI (Finding name) [5,016 occurrences]
  ABDOMINAL PAIN, CRAMPY
  ABDOMINAL DISTENTION
GN (Generic drug name) [2,267 occurrences]
  mesna
  aminocaproic acid
GT (Glossary term) [4,797 occurrences]
  SYNDROME ABDOMINAL ACUTE
  ABDOMINAL CRAMP
HC (Hierarchical class) [196 occurrences]
  Behavior Therapy
  Cognitive Therapy
HG (High Level Group Term) [383 occurrences]
  Adrenal gland disorders
  Benign neoplasms gastrointestinal
HT (Hierarchical term) [22,152 occurrences]
  Abdomen
  Abdominal pain

ID (Nursing indicator) [2,576 occurrences]
  Ankylosed joints
  Appetite loss
IN (Name for an ingredient) [20,455 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
IV (Intervention) [685 occurrences]
  Activities of Daily Living (ADLs)
  Pain Management
LT (Lower Level Term) [52,163 occurrences]
  Acute abdomen
  X-ray NOS abnormal
MD (CCS multi-level diagnosis categories) [693 occurrences]
  Abdominal pain
  Congenital anomalies
MH (Main heading) [22,568 occurrences]
  1,2-Dipalmitoylphosphatidylcholine
  Abdomen
MM (Metathesaurus string created to distinguish different meanings of the same lexical string.) [21,295 occurrences]
  17-hydroxysteroid dehydrogenase <1>
  DOPS <1>
MS (Multum names of branded and generic supplies or supplements) [4,652 occurrences]
  Acetone
  0.3cc Syringe 29g 1/2"
N1 (Chemical Abstracts Service Type 1 name of a chemical) [22,879 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) [138,777 occurrences]
  2-bromolysergic acid diethylamide
  3-hydroxyproline
NP (Non-preferred term) [5,437 occurrences]
  3,4-methylenedioxyamphetamine
  congenital defects
OS (System-organ class in the WHO Adverse Reaction Terminology) [58 occurrences]
  AUTONOMIC NERVOUS SYSTEM DISORDERS
  PSYCHIATRIC DISORDERS
PC (Preferred trimmed term in ICPC) [233 occurrences]
  arthrogryposis multiplex congenita
  Bartter syndrome

PM (Machine permutation) [73,249 occurrences]
1,2 Dipalmitoylphosphatidylcholine
Enzyme, Branching
PN (Metathesaurus preferred name) [33,130 occurrences]
17-Hydroxysteroid Dehydrogenases
Droxidopa
PQ (Qualifier for a problem) [9 occurrences]
Family
Health Promotion
PT (Designated preferred name) [810,026 occurrences]
Dipalmitoylphosphatidylcholine
Brancher enzyme
PTGB (British preferred term) [20,666 occurrences]
Abetalipoproteinaemia
Acting out - mental defence mechanism
RS (Extracted related names in SNOMED2) [68 occurrences]
Aleutian disease virus
Aluminum silicate
RT (Designated related term) [6,404 occurrences]
20-Hydroxyprogesterone
Lumpy jaw
SCN (Scientific name) [154,568 occurrences]
Abelson murine leukemia virus
Absidia
SF (Synonym made by replacing ; with no spaces around it with , in ICPCP) [5,985 occurrences]
Cramps, abdominal
Loss (of) appetite
SS (Synonymous short forms) [969 occurrences]
adrenoleukodystrophy
Alzheimer disease
SY (Designated synonym) [290,933 occurrences]
Branching enzyme
Amylo-(1,4,6)-transglycosylase
SYGB (British synonym) [6,555 occurrences]
Tumour of abdomen
ABL - Abetalipoproteinaemia

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 syn-

tactic 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 models 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 1,449,616 (61%) of the 2,370,524 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 1,939,851 (82%) 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 2,126,673 (90%) English Metathesaurus strings.