Filtering the UMLS® Metathesaurus® for MetaMap

2008 Edition

Francois-Michel Lang and Alan R. Aronson

May 6, 2008

1. Overview

The MetaMap program's 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 the 2008AA version of the Metathesaurus, for example, the Metathesaurus includes 3,812,808 English strings, 3,770,763 (98.9%) of them distinct, comprising 1,548,333 concepts. There are 23% more English strings and 23% more concepts than in the 2007AA edition. Many of the strings in the Metathesaurus are of little value to MetaMap for one of four reasons:

1. Some strings either represent general, nonmedical concepts or are unnecessarily ambiguous.
2. Some strings are virtually indistinguishable from each other; for efficiency, only one representative of a set of indistinguishable strings is needed.
3. Some strings have internal structure or meta information and are therefore highly unlikely to appear in regular text.
4. 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:

1. manual filtering,
2. lexical filtering,
3. filtering by type, and

4. 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. Finally, section 7 is an Appendix providing details of filtering decisions for Metathesaurus strings based on term type (TTY).

2. Manual Filtering

A number of Metathesaurus strings are problematic and have been suppressed before performing other forms of filtering. There are 13,203 such strings, 1.3% less than in 2007:

- **Unnecessarily ambiguous terms** [4,991 occurrences]
  'Other' for 'Other location of complaint'
  'Protocols' for 'Protocols: Urinary Elimination'
  …
- **Contextual terms**, i.e., terms whose meaning can only be understood within the context of their vocabulary [7,281 occurrences]
  All terms containing "*NEC"* or an expanded form
- **Brand names**, i.e., short forms of terms containing "*brand"* [349 occurrences]
- **Enzyme Commission (EC) numbers** beginning "*EC <integer>*." [168 occurrences]
- **Single alphabetic strings** (e.g., 'a', 'A', 'b', 'B') [173 occurrences];
- **Numbers** (e.g., '2', '+1', '-4', '98.734', '50000') [237 occurrences];
- **Special cases** [4 occurrences]
  'Periods' and 'Period' for 'Menstruation' (C0025344)
  'Clap' and 'CLAP' for 'Gonorrhea' (C0018081)
  'BRA' for 'Brain' (C0006104)

The Metathesaurus staff marks some terms as suppressible synonyms because they are thought to be inappropriate for any use. These 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 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 in biomedical text. 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;
- *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 anaemias” 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 (and also SNOMEDCT): *[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 *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.

3.3 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.

3.4 Case variation

Two strings which differ from each only because of case variation normally refer to the same thing. For example, the concept 'Abdomen' has strings "abdomen" and "ABDOMEN" in addition to "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*. Similarly, the case variation in the first three letters of the concepts 'CDE genotype', 'CDe genotype', ..., and 'cde genotype' is significant. Despite these observations, case almost never matters within the limited context of the set of all strings for a given concept.

3.5 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.

3.6 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" is such a synonym of "Malignant neoplasm of abdomen." Over 11% (411,146) of the 3,812,808 English Metathesaurus strings are marked as suppressible synonyms. There is also a smaller number (65,298) 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 7.1 below). There are 917,243 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: 251,377 (27%, down from 31% in 2007) of the 917,243 TTY strings are already marked as suppressible.

The filtered out term types together with examples for each are given in the Appendix (section 7). For completeness, the Appendix also includes a section of questionable types and a section of good types.

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, strings containing *of* 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 1,937,745 (58%) of the 3,812,808 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 2,199,211 (66%) 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 3,051,888 (92%) English Metathesaurus strings.

Note that in order to have all Metathesaurus concepts represented in each model, the preferred name of a concept is retained when all of its strings would otherwise be filtered out. For the strict model 458,281 concepts are so spared, for the moderate model 23,159 are spared, and for the relaxed model no concepts need to be spared.

7. Appendix

7.1 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) [108,971 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)

- ACR (Acronym) [50,210 occurrences]
  "FA"
  "BLEO/DOX/DTIC/VCR"

- BPCK (Branded Drug Delivery Device) [150 occurrences]
  "{2 (Risedronate 75 MG Oral Tablet [Actonel]) } Pack [Actonel]"
  "{28 (Norethindrone 0.35 MG Oral Tablet) } Pack [Camila 28 Day]"

- CA2 (ISO 3166-1 standard country code in alpha-2 (two-letter) format) [244 occurrences]
  "AD"
  "ZW"

- CA3 (ISO 3166-1 standard country code in alpha-3 (three-letter) format) [244 occurrences]
  "ABW"
  "ZWE"

- CCN (Chemical code name) [1,964 occurrences]
  "EC 3.5.1.1"
  "T-1220"

- CCS (FIPS 10-4 country code) [236 occurrences]
  "AA"
  "ZI"

- CDD (Clinical drug name in delimited format) [22,197 occurrences]
  "BEESWAX@@MISCELL.@WAX (GM)"
  "CASTOR OIL@@MISCELL.@OIL"

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

- CV (Content view) [2 occurrences]
  "UMLS enhanced VA/KP Problem List Subset of SNOMED (Level 0+SNOMED)"
  "UMLS enhanced VA/KP Problem List Subset of SNOMED (Level 0+SNOMED+MDR)"

- DEV (Descriptor entry version) [16432 occurrences]
  "ACCIDENTS INDUST"
  "techniques BACTERIOL"

- DSV (Descriptor sort version) [3,965 occurrences]
  "ACYLGLYCEROPHOSPHOCHOLINE ACYLTRANSFERASE 001 O"
  "COMPLEMENT C 03 A"

- 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) [1 occurrence]
  "Weak e (lower case e) phenotype (finding)"

- FDAAB (Food and Drug Administration AB) [483 occurrences]
  "BEAD IMP ER"
  "INJ PWD LYO F/SUSPER"

- FN (Full form of descriptor) [311,143 occurrences]
  "Methylphenyltetrahydropyridine (substance)"
  "Acute abdomen (disorder)"

- GPCK (Generic Drug Delivery Device) [93 occurrences]
  "{2 (Risedronate 75 MG Oral Tablet) } Pack"
  "{56 (varenicline 1 MG Oral Tablet) } Pack"

- HS (Short hierarchical term (needed expansion) in ICD 10) [219 occurrences]
  "Agents primarily acting on smooth and skeletal muscles and the respiratory system"
  "Bacterial vaccines"

- 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")

- JAXPT (NCI Mouse Terminology PT) [138 occurrences]
  “AKR”
  “BALB/c”

- JAXSY (NCI Mouse Terminology SY) [18 occurrences]
  “severe combined immune deficiency”
  “C3Smn.C-Prkdc-scid/J”

- LN (LOINC official fully specified name) [49,450 occurrences]
  “ALMECILLIN:SUSC:PT:ISLT:ORDQN:MIC”
  “ALMECILLIN:SUSC:PT:ISLT:ORDQN:AGAR DIFFUSION”

- LO (Obsolete official fully specified name) [1,359 occurrences]
  “AUREOBASIDIUM PULLULANS AB.IGE:ACNC:PT:SER:QN”
  “PARROT AUSTRALIAN DROPPINGS AB.IGE:ACNC:PT:SER:QN”

- LPDN (LOINC parts display name) [202 occurrences]
  “Food Allergen Mix 74 (Herring + Cod + Plaice + Mackerel)”
  “Dust Allergen Mix A (Blatella germanica + Dermatophagoides farinae + Dermatophagoides pteronyssinus + house dust greer) | Bld-Ser-Plas”

- LX (Official fully specified name with expanded abbreviations) [49,447 occurrences]
  “ALMECILLIN:SUSCEPTIBILITY:POINT IN TIME:ISOLATE:QUANTITATIVE:MINIMUM INHIBITORY CONCENTRATION”
  “ALMECILLIN:SUSCEPTIBILITY:POINT IN TIME:ISOLATE:SEMI-QUANTITATIVE:BACTERIAL SENSITIVITY (KIRBY-BAUER)”

- MTH_AB (MTH abbreviation) [1 occurrence]
  “:anti-![GREEK SMALL LETTER ALPHA]!5![GREEK SMALL LETTER BETA]!1 integrin MOAB”

- MTH_ACR (MTH acronym) [1 occurrence]
  “anti-![GREEK SMALL LETTER ALPHA]!5![GREEK SMALL LETTER BETA]!1 integrin MOAB/DTIC”

- MTH_BD (MTH fully-specified drug brand name that can be prescribed) [9 occurrences]
  “TOCOSOL![TRADE MARK SIGN]! Paclitaxel”
  “Xenavex![TRADE MARK SIGN]!”

- MTH_CHN (MTH chemical structure name) [1 occurrence]
  “(![PLUS-MINUS SIGN]!)-1-(3-dimethylaminopropyl)-1- (4-fluorophenyl)-1,3 dihydroisobenzofuran-5-carbonitrile, HBr”

- MTH_FN (MTH Full form of descriptor) [3168 occurrences]
  “beta⁺ Thalassemia, normal Hb A₂, type 1, silent (disorder)”
  “m7 G(5’)pppN pyrophosphatase (substance)”

- MTH_IS (Metathesaurus-supplied form of obsolete synthesized term in the Read Thesaurus) [403occurrences]
  “24,25-Dihydroxyvitamin D₃”
  “beta⁺ Thalassemia, NOS”

- MTH_LV (MTH lexical variant) [1 occurrence]
  “Gelfoam![REGISTERED SIGN]!”

- MTH_OP (Metathesaurus obsolete preferred term) [272 occurrences]
  “Vitamin B4”
  “Vitamin B₄”

- MTH_PTGB (Metathesaurus-supplied form of British preferred term) [136 occurrences]
  “Haemoglobin A₂ Adria”
  “A_{gamma} beta⁺ HPFH AND beta⁰ thalassaemia in cis”

- MTH_RXN_BD (RxNorm Created BD) [146 occurrences]
  “Adenocard 3mg/ml Solution for Injection_#1”
  “Symbicort, 160 mcg-4.5 mcg/inh inhalation aerosol with adapter_#1”

- MTH_RXSN_CD (RxNorm Created CD) [66 occurrences]
  “CROMOLYN SODIUM 800 MCG INHALATION INHALANT [INTAL]_#2”
  “LAMICTAL KIT DOSE ESCALATION 25 MG/100 MG_#1”

- MTH_RXN_DP (RxNorm Created DP) [48 occurrences]
  “Azithromycin 500 MILLIGRAM In 1 TABLET ORAL TABLET, FILM COATED_#1”
  “iopamidol 612 MILLIGRAM In 1 MILLILITER INTRAVASCULAR INJECTION, SOLUTION [Isovue]_#2”

- MTH_SI (MTH MTH Sign or symptom of) [42 occurrences]
  “Communication with community resources - other”
  “Sleep and rest patterns - other”

- MTH_SMQ (Metathesaurus version of Standardised MedDRA Query) [26 occurrences]
  “Gastrointestinal perforation, ulcer, hemorrhage, obstruction non-specific findings/procedures (SMQ)”
  “Adverse pregnancy outcome/reproductive toxicity (including neonatal disorders) (SMQ)

- MTH_SY (MTH Designated synonym) [1,256 occurrences]
  “3-Oxo-5alpha-steroid delta⁴-dehydrogenase”
  “R AW”

- MTH_SYGB (Metathesaurus-supplied form of British synonym) [19 occurrences]
  “Haemoglobin A₂”
  “Hereditary persistence of fetal haemoglobin (HPFH) delta beta⁰ thalassaemia”

- OA (Obsolete abbreviation) [49,962 occurrences]
  "Gastrin secretion abnorm.NOS"
  "Spontaneous abort.incomp.NOS"

- Note that as of 2008, OCD and OF will be classified as questionable types because the great majority of their strings which are not appropriate for text processing are already marked as suppressible.

- OLX (Expanded LOINC obsolete fully specified name) [1,358 occurrences]
  "CRYPTOSPORIDIUM::::"
  "5-FLUOROCYTOSINE:SUSCEPTIBILITY:POINT IN TIME:ISOLATE:QUANTITATIVE OR ORDINAL:MINIMUM INHIBITORY CONCENTRATION"

- ONP (Obsolete non-preferred for language term) [37 occurrences]
  "Retired Code" [the only string--repeated 37 times]

- OOSN (Obsolete official short name) [1,291 occurrences]
  "Deprecated ^"
  "Deprecated 5FC MIC"

- OSN (Official short name) [41534 occurrences]
  "ACTH 1.5H p Ins IV Plas-mCnc"
  "Promyelocytes # Bld Auto"

- OWL (rdf:ID of the owl class) [2137 occurrences]
  "ActualInterruptedBreastfeedingPattern"
  "UnilateralNeglect"

- PTAV (Preferred Allelic Variant) [15735 occurrences]
  "ATRIAL SEPTAL DEFECT WITH ATRIOVENTRICULAR CONDUCTION DEFECTS, 1-BP DEL, 262G"
  "TETRALOGY OF FALLOT, GLU21GLN"

- QAB (Qualifier abbreviation) [83 occurrences]
  "AG"
  "PX"

- QEV (Qualifier entry version) [83 occurrences]
  "HIST"
  "RAD EFF"

- QSV (Qualifier sort version) [7 occurrences]
  "ADMINISTRATION A"
  "LEGISLATION AB"

- RAB (Root abbreviation) [139 occurrences]
  "DSM3R"
  "MSH"

- RENIDN (Registry Nomenclature Information Display Name) [310 occurrences]
  "Carcinoma, islet cell (RENI)"
  "Stomach, NOS (RENI)"

- RHT (Root hierarchical term) [62 occurrences]
  "DSM-III-R"
  "MeSH"

- RPT (Root preferred term) [139 occurrences]
  "DSM-III-R"
  "Medical Subject Headings"

- RSY (Root synonym) [35 occurrences]
  "Diagnostic and Statistical Manual of Mental Disorders: DSM-III-R"
  "LCSH"

- SB (Named subset of a source) [2 occurrences]
  "US English Dialect Subset"
  "GB English Dialect Subset"

- SBD (Semantic branded drug) [18,707 occurrences]
  "salmeterol 0.05 MG/ACTUAT Inhalant Powder [Serevent Diskus]"
  "Acetaminophen 250 MG / Caffeine 30 MG / Chlorpheniramine 2 MG / Hydrocodone 5 MG / Phenylephrine 10 MG Oral Tablet [Hycomine Compound]"

- SBDC (Semantic Branded Drug Component) [18,229 occurrences]
  "Benzoyl Peroxide 0.05 MG/MG [Face Up #2]"
  "Acetaminophen 32 MG/ML / guaiacolsulfonate 60 MG/ML / Hydrocodone 1 MG/ML / Pseudoephedrine 6 MG/ML [Protuss D]"

- SBDF (Semantic branded drug and form) [14,827 occurrences]
  "Acetaminophen / Diphenhydramine / Pseudoephedrine Oral Tablet [Tylenol Allergy Sinus NightTime]"
  "sulfabenzamide / Sulfacetamide Sodium / sulfathiazole Vaginal Cream [Sultrin Triple Sulfa]"

- SMQ (Standardised MedDRA Query) [139 occurrences]
  "Central nervous system haemorrhages and cerebrovascular accidents (SMQ)"
  "Gastrointestinal perforation, ulcer, haemorrhage, obstruction non-specific findings/procedures (SMQ)"

- SN (Official component synonym in LOINC) [319 occurrences]
  “Fy^a AB”
  “SALMONELLA TYPHI d AB”

- SSN (Source short name, used in the UMLS Knowledge Source Server) [139 occurrences]
  “DSM-III-R”
  “MeSH”

- SX (Mixed-case component synonym with expanded abbreviations) [470 occurrences]
  “Blood group antibody P>1< plus P plus p^k^”
  “human histocompatibility complex derived antigens-DQ w7(w3)”

- UCN (Unique common name) [232 occurrences]
  “canaries <genus>”
  “monkeys <#2>”

- UE (Unique equivalent name) [4 occurrences]
  “Tanella <Bacteria>”
  “Wernerella <Bacteria>”

- USN (Unique scientific name) [1,252 occurrences]
  “Absidia <zygomycete fungus>”
  “Aedes <sugbenus>”
  “Bacillus <bacterium>”

- USY (Unique synonym) [108 occurrences]
  “Chlamydia psittaci <Chlamydophila psittaci>”
  “AsGV<agrotis>”

- VAB (Versioned abbreviation) [136 occurrences]
  “SPN2003”
  “GO2004_12_20”
  “MSH2006_2005_11_15”

- VPT (Versioned preferred term) [136 occurrences]
  “Standard Product Nomenclature, 2003”
  “Medical Subject Headings, 2006_2005_11_15”

- VSY (Versioned synonym) [15 occurrences]
  “LCSH, 1990”
  “Glossary of Methodologic Terms for Clinical Epidemiologic Studies of Human Disorders, 1992”

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

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

7.2 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.

- AM (Short form of modifier) [286 occurrences]
  "2nd concurrent infusion ther"
  "Upper right eyelid"

- CC (Trimmed ICPC component process) [1 occurrence]
  "Referral primary care provider"

- CD (Clinical Drug) [132,324 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 99%"

- CDA (Clinical drug name in abbreviated format) [22,917 occurrences]
  "ACYCLOVIR 5 % TOPICAL CREAM (GRAMS)"
  "TURPENTINE DENTAL SPIRIT"

- CDISCPT (Clinical Data Interchange Standards Consortium PT) [913 occurrences]
  "TZA"
  "TRANSDERMAL"
  "SURGICAL AND MEDICAL PROCEDURES"

- CDISCSY (Clinical Data Interchange Standards Consortium SY) [555 occurrences]
  "BELIZE"
  "MYANMAR"
  "Medical Dictionary for Regulatory Activities"

- CP (ICPC component process (in original form)) [25 occurrences]
  "Administrative procedure"
  "Other therapeutic procedure"

- ES (Short form of entry "term") [36 occurrences]
  "periodic light AOD use"
  "AOD tax

- ETAV (Entry Term Allelic Variant) [925 occurrences]
  "ALPERS SYNDROME"
  "VON HIPPEL-LINDAU SYNDROME"

- FDASY (Food and Drug Administration SY) [931 occurrences]
  "Abdominal pain"
  "Viral infection"

- 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."

- LS (Expanded system/sample type (The expanded version was created for the Metathesaurus and includes the full name of some abbreviations.)) [1,684 occurrences]
  About one-third of these terms contain embedded periods.
  "AORTIC VALVE"
  "AORTA.THORACIC.ASCENDING"

- LV (Lexical variant) [434 occurrences]
  "2-Deoxy-5-fluorouridine"
  "monoclonal antibody humanized anti-Tac"

- MP (Preferred names of modifiers) [483 occurrences]
  "ABNORMALITY"
  "PROB"

- MTH_BSY (Metathesaurus broad synonym expanded) [17 occurrences]
  "debranching enzyme complex location"
  "nuclear pore subcomplex location"

- MTH_OF (Metathesaurus-supplied form of obsolete fully specified name) [278 occurrences]
  "Vitamin B₄ (substance)"
  "Vitamin B4 (substance)"

- MTH_PT (Metathesaurus preferred term) [4,726 occurrences]
  "³³Phosphorus"
  "Vitamin D2"

- OAM (Obsolete Modifier Abbreviation) [6 occurrences]
  "FDA investigational device"
  "Item/serv in medicare study"

- OCD (Obsolete clinical drug) [236,241 occurrences]
  "AMOBARBITAL SODIUM 200MG TABLET"
  "FENNEL OIL"

- OF (Obsolete fully specified name) [121,021 occurrences]
  "17-Ketogenic steroid (substance)"
  "alpha-Glucosidase [dup] (substance)"

- PCE (Preferred entry "term" to a Supplementary Concept "term") [63,980 occurrences]
  "GP 130"
  "(+)-5-exo-bromocamphor"

7.3 Good Types

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

- AC (Activities) [10,764 occurrences]
  "Monitor blood pressure"
  "Control bleeding"

- AD (Adjective) [1,383 occurrences]
  "Anorexic"
  "Scarred"

- AS (Attribute type synonym) [25 occurrences]
  "Precipitating factor"
  "Px - Prescription"

- AT (Attribute type) [819 occurrences]
  "Allergen"
  "Association"

- BD (Fully-specified drug brand name that can be prescribed) [30,852 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"

- BioCPT (BioCarta PT) [300 occurrences]
  "Telomeres, Telomerase, Cellular Aging, and Immortality"
  "Chromatin Remodeling by hSWI/SNF ATP-dependent Complexes"

- BN (Fully-specified drug brand name that can not be prescribed) [30,134 occurrences]
  "Parlodel"
  "Aminocaproic Acid"

- BSY (Broad synonym) [747 occurrences]
  "p53-mediated DNA damage response"
  "ER stress response"

- CDCPT (Centers for Disease Control and Prevention PT) [921 pccurrences]
  “Polynesian“
  “Egyptian“

- CE (Entry “term” to a Supplementary Chemical “term”) [208,455 occurrences]
  “2 bromolysergic acid diethylamide”
  “7S RNA”

- CHN (Chemical structure name) [904 occurrences]
  “3-Hydroxy-L-tyrosine”
  “2-Difluoromethylornithine”

- CL (Class) [14 occurrences]
  “Managing the Practice”
  “Ensuring Appropriate Pharmacotherapy”

- CMN (Common name) [8,524 occurrences]
  “acanthocephalans”
  “alfalfa”

- CN (LOINC official component name) [17,883 occurrences]
  “DIPALMITOYLPHOSPHATIDYLCHOLINE”
  “1,4-ALPHA GLUCAN BRANCHING ENZYME”
  “???LEAD” *(sic)*

- CO (Component name (these are hierarchical terms, as opposed to the LOINC component names which are analytes)) [57 occurrences]
  “CARDIAC COMPONENT”
  “HEALTH BEHAVIOR COMPONENT”

- CSN (Chemical Structure Name) [2642 occurrences]
  “Thiocarbamide”
  “1,3-Diamino-4-methylbenzene”

- CTCAEPT (Common Terminology Criteria for Adverse Events Preferred Term) [5598 occurrences]
  “Weight gain”
  “Grade 2 Colon obstruction”

- CU (Common usage) [11 occurrences]
  “Entelev”
  “Jim’s Juice”

- CX (Component process in ICPC with abbreviations expanded) [5,861 occurrences]
  “NOS ANTIBODY”
  “CANCER ANTIGEN 125”

- DCPPT (Division of Cancer Prevention Program PT) [909 occurrences]
  "2-Mercaptoethanesulfonate, Sodium Salt"
  "(17alpha)-Pregna-2,4-dien-20-yno[2,3-d]isoxazol-17-ol"

- DCPSY (Division of Cancer Prevention Program Synonym) [2 occurrences]
  "Ortho-Novum 1/35"
  "Suberanilohydroxamic Acid"

- DE (Descriptor) [10,817 occurrences]
  "synthetic 11-hydroxycorticosteroids"
  "abdomen"

- DF (Dose Form) [873 occurrences]
  "24 Hour Transdermal Patch"
  "Aerosol"

- DI (Disease name) [6,172 occurrences]
  "ABETALIPOPROTEINEMIA"
  "Abortion, Spontaneous"

- DN (Display Name) [8 occurrences]
  "B16 Malignant Melanoma"
  "Chloroacetate Esterase Staining Method"

- DO (Domain) [4 occurrences]
  "Health Systems Management"
  "Ensuring Appropriate Therapy and Outcomes"

- DP (Drug Product) [8004 occurrences]
  "Diazepam 2 MILLIGRAM In 1 TABLET ORAL TABLET"
  "nicardipine hydrochloride 20 MILLIGRAM In 1 CAPSULE ORAL CAPSULE, GELATIN COATED [Cardene]"

- DS (Short form of descriptor) [509 occurrences]
  "AOD withdrawal syndrome"
  "biological AOD dependence"

- DTPPT (NCI Developmental Therapeutics Program Preferred Term) [2 occurrences]
  "(+-)-Hexestrol"
  "PSA Vaccinia (CV1-produced), THERION"

- DTPSY (Developmental Therapeutics Program SY) [2,030 occurrences]
  "Dimethylbenzanthracene"
  "3-Hydroxy-L-tyrosine"

- DX (Diagnosis) [256 occurrences]
  “Anxiety”
  “Blood Pressure Alteration”

- EN (MeSH nonprint entry “term”) [32,997 occurrences]
  “(131)I-MAA”
  “Injuries, Abdominal”

- EP (Entry “term”) [17,935 occurrences]
  “Dipalmitoyllecithin”
  “Branching Enzyme”

- EQ (Equivalent name) [3,313 occurrences]
  “Borrelia burgdorferi sensu stricto”
  “[Prevotella] zoogleoformans”ET (Entry “term”) [48,381 occurrences]
  “MPTP”
  “Abelson’s virus”

- ET (Entry "term") [48381 occurrences]
  “4 nitroquinoline N oxide”
  “vitamin C”

- EX (Expanded form of entry term) [36 occurrences]
  “periodic light Alcohol or Other Drugs use”
  “Alcohol or Other Drugs tax”

- FBD (Foreign brand name) [2,050 occurrences]
  “Helixor”
  “Iscador”

- FDAPT (Food and Drug Administration PT) [5,802 occurrences]
  “MENSA“
  “EPINEPHRINE“

- FI (Finding name) [5,016 occurrences]
  “ABDOMINAL PAIN, CRAMPY”
  “ABDOMINAL DISTENTION”

- GN (Generic drug name) [2,658 occurrences]
  “mesna”
  “paraaminobenzoic acid”

- GT (Glossary “term”) [4,797 occurrences]
  “SYNDROME ABDOMINAL ACUTE”
  “ABDOMINAL CRAMP”

- HC (Hierarchical class) [473 occurrences]
  "Behavior Therapy"
  "Cognition"

- HD (Hierarchical descriptor) [171 occurrences]
  "Transplanted Organ Complication"
  "Neoplasm by Site"

- HG (High Level Group Term) [332 occurrences]
  "Adrenal gland disorders"
  "Benign neoplasms gastrointestinal"

- HT (Hierarchical term) [22,932 occurrences
  "Abdomen"
  "Abdominal pain"

- HTN (HL7 Table Name) [478 occurrences]
  "Bed Status"
  "Body Parts"

- ID (Nursing indicator) [5,292 occurrences]
  "Ankylosed joints"
  "Appetite loss"

- IN (Name for an ingredient) [30,240 occurrences]
  "mesna"
  "beta-Alanine"

- INP (Ingredient preparation) [4,231 occurrences]
  "MESNA PREPARATION"
  "PARA-AMINOBENZOIC ACID PREPARATION"

- IS (Obsolete synthesized "term" in the Read Thesaurus) [84,846 occurrences]
  "Fall - accidental" (for concept "Accidental Falls")
  "MVTA - Motor vehicle traffic accident" (for concept "Accidents, Traffic")

- IT (Index "term", i.e., derived from the index to any non-MeSH source vocabulary) [2,077 occurrences]
  "ACUTE ABDOMEN"
  "CRAMP ABDOMINAL"

- IV (Intervention) [845 occurrences]
  "Activities of Daily Living (ADLs)"
  "Pain Management"

- IVC (Intervention categories) [4 occurrences]
  "Teaching, Guidance, and Counseling"
  "Treatments and Procedures"

- KEGGPT (Kyoto Encyclopedia of Genes and Genomes PT) [126 occurrences]
  "Alzheimer's Disease"
  "tRNA Biosynthesis (Eukaryotes)"

- LPN (LOINC parts name) [43,919 occurrences]
  "DIPALMITOYLPHOSPHATIDYLCHOLINE"
  "1,4-ALPHA GLUCAN BRANCHING ENZYME"

- LT (Lower Level Term) [56,580 occurrences]
  "Acute abdomen"
  "X-ray NOS abnormal"

- MD (CCS multi-level diagnosis categories) [693 occurrences]
  "Abdominal pain"
  "Congenital anomalies"

- MH (Main heading) [24,767 occurrences]
  "1,2-Dipalmitoylphosphatidylcholine"
  "Abdomen"

- MOA (Mechanism of action) [256 occurrences]
  "Adrenergic alpha-Agonists"
  "Bile Acids"

- MS (Multum names of branded and generic supplies or supplements) [8,475 occurrences]
  "Acetone"
  "0.3cc Syringe 29g 1/2""

- MTH_HG (MTH High Level Group Term) [63 occurrences]
  "Hemoglobinopathies"
  "Leukemias"

- MTH_HT (MTH Hierarchical term) [203 occurrences]
  "Amebic infections"
  "Carcinoid tumors"

- MTH_ID (Metathesaurus expanded form of nursing indicator) [47 occurrences]
  "Passive joint movement of left ankle"
  "partial pressure of carbon dioxide in arterial blood"

- MTH_LT (MTH Lower Level Term) [577 occurrences]
  "Myelosis-non-leukemic"
  "Alpha thalassemia"

- MTH_NPT (MTH non-preferred for language term) [2 occurrences]
  "Zuni"
  "Diegueno"

- MTH_NSY (Metathesaurus narrow synonym expanded) [44 occurrences]
  "anaphase promoting complex location inhibition"
  "succinate-CoA ligase complex location (GDP-forming) (sensu Bacteria)"

- MTH_OL (MTH Non-current Lower Level Term) [140 occurrences]
  "Spinal anesthesia (all forms)"
  "Respiratory arrest (excluding neonatal)"

- MTH_OPN (Metathesaurus obsolete preferred term, natural language form) [8 occurrences]
  "megaloblastic anemia"
  "peripheral edema"

- MTH_OS (MTH System-organ class) [1 occurrence]
  "Neoplasms benign, malignant and unspecified (including cysts and polyps)"

- MTH_PTN (Metathesaurus preferred term, natural language form) [512 occurrences]
  "hepatitis A & B immunization"
  "injection of influenza immunization"

- MTH_RLS (Metathesaurus related synonym expanded) [19 occurrences]
  "distal pole complex location"
  "single-stranded DNA-binding protein complex location"

- MV (Multi-level procedure category) [403 occurrences]
  "Adenoidectomy without tonsillectomy"
  "Excision of semilunar cartilage of knee"

- N1 (Chemical Abstracts Service Type 1 name of a chemical) [22,535 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-"

- NCI-GLOSSPT (NCI-GLOSS PT) [2,662 occurrences]
  "mercaptopurine"
  "pernicious anemia"

- NCI-GLOSSSY (NCI-GLOSS SY) [57 occurrences]
  "cell differentiation"
  "IL-1-alfa"

- NM (Supplementary chemical "term", a name of a substance) [175,134 occurrences]
  "2-bromolysergic acid diethylamide"
  "3-hydroxyproline"

- NP (Non-preferred term) [5,437 occurrences]
  "3,4-methylenedioxyamphetamine"
  "congenital defects"

- NPT (HL7 non-preferred for language term) [171 occurrences]
  "Cardiology clinic"
  "health care facility"

- NS (Short form of non-preferred "term") [200 occurrences]
  "neonatal AOD abstinence syndrome"
  "dysfunctional AOD use"

- NSY (Narrow synonym) [2,021 occurrences]
  "surface antigen variation"
  "bone calcification"

- NX (Expanded form of non-preferred "term") [200 occurrences]
  "neonatal Alcohol or Other Drugs abstinence syndrome"
  "dysfunctional Alcohol or Other Drugs use"

- OB (Obsolete term) [57 occurrences]
  "Bupropion"
  "Oral salbutamol preparation"

- OBS (Obsolete broad synonym) [12 occurrences]
  "general amino acid transporter"
  "oligomerization activity"

- OC (Nursing outcomes) [330 occurrences]
  "Thermoregulation"
  "Decision Making"

- OL (Non-current Lower Level Term) [9,025 occurrences]
  "Anomaly anomaly congen"
  "Congenital anomaly, unspecified"

- OM (Obsolete modifiers in HCPCS) [6 occurrences]
  "YING CLINICAL TRIAL"
  "ITEM OR SERVICE PROVIDED IN A MEDICARE SPECIFIED STUDY"

- ONS (Obsolete narrow synonym) [7 occurrences]
  "germination (sensu Saccharomyces)"
  "polarity-dependent cell elongation"

- OP (Obsolete preferred term) [145,878 occurrences]
  "Carbenoxolone sodium [gastro-intestinal use]"
  "Acute abdomen"

- OPN (Obsolete preferred term, natural language form) [85 occurrences]
  "carcinoma of the large intestine"
  "acquired pylorospasm"

- OR (Orders) [1,357 occurrences]
  All terms are from PCDS97 and are full utterances.
  "Discharge patient."
  "Use assistive devices to maintain required position."

- ORS (Obsolete related synonym) [48 occurrences]
  "COPI vesicle"
  "FK506 binding protein"

- OS (System-organ class in the WHO Adverse Reaction Terminology) [58 occurrences]
  "PSYCHIATRIC DISORDERS"
  "AUTONOMIC NERVOUS SYSTEM DISORDERS"

- PC (Preferred "trimmed term" in ICPC) [233 occurrences]
  "arthrogryposis multiplex congenita"
  "Bartter syndrome"

- PE (Physiologic effect) [1,699 occurrences]
  "Vasoconstriction"
  "Cardiac Rate Alteration"

- PEN (Preferred MeSH nonprint entry "term") [15,074 occurrences]
  "(131)I-Macroaggregated Albumin"
  "Abortion Centers"

- PEP (Preferred entry "term") [7,037 occurrences]
  "17 beta-Hydroxysteroid Dehydrogenases"
  "Abdominal Cramps"

- PK (Pharmacokinetics) [59 occurrences]
  "Absorption"
  "Glutathione s-transferases"

- PM (Machine permutation) [83,814 occurrences]
  "1,2 Dipalmitoylphosphatidylcholine"
  "Enzyme, Branching"

- PN (Metathesaurus preferred name) [105,054 occurrences]
  "17-Hydroxysteroid Dehydrogenases"
  "Droxidopa"

- PQ (Qualifier for a problem) [6 occurrences]
  "Family"
  "Health Promotion"

- PR (Name of a problem) [405 occurrences]
  "Placenta abruptio"
  "Dependency on alcohol"

- PS (Short forms that needed full specification) [1740 occurrences]
  "Acoustic nerve"
  "Tonsillar fossa"

- PSC (Protocol selection criteria) [555 occurrences]
  "bladder cancer"
  "basal cell carcinoma of the skin"

- PT (Designated preferred name) [1,043,812 occurrences]
  "Dipalmitoylphosphatidylcholine"
  "Abdomen"

- PTCS (Preferred Clinical Synopsis) [20,364 occurrences]
  "Acute lymphatic leukemia"
  "Ventricular septal defect"

- PTGB (British preferred term) [21,899 occurrences]
  "Abetalipoproteinaemia"
  "Acting out - mental defence mechanism"

- PTN (Preferred term, natural language form) [7,410 occurrences]
  "keratoconus"
  "solar keratosis"

- PX (Expanded preferred terms (pair with PS)) [3,157 occurrences]
  These terms often contain characters indicating a superscript or subscript.
  "Ca^2+^-transporting ATPase"
  "alpha>1< Antichymotrypsin"

- PXQ (Preferred term in preferred qualifier concept) [168 occurrences]
  "agenesis"
  "wounds"

- RLS (Related synonym) [1,427 occurrences]
  "blood coagulation factor activity"
  "glycine betaine/proline porter activity"

- RS (Extracted related names in SNOMED2) [68 occurrences]
  "Aleutian disease virus"
  "Aluminum silicate"

- RT (Designated related "term") [7,129 occurrences]
  "Lumpy jaw"
  "20-Hydroxyprogesterone"

- SC (Special Category term) [41 occurrences]
  "Congenital Malformations"
  "bandages"

- SCALE (Scale) [3 occurrences]
  "BEHAVIOR"
  "KNOWLEDGE"
  "STATUS"

- SCD (Semantic Clinical Drug) [31,582 occurrences]
  "calcium chloride, dihydration 0.2 MG/ML / Potassium Chloride 0.3 MG/ML / Sodium Chloride 6 MG/ML / Sodium Lactate 3.1 MG/ML Irrigation Solution"
  "Hydrogen Peroxide 300 MG/ML Topical Solution"

- SCDC (Semantic Drug Component) [23,720 occurrences]
  "Iron-Dextran Complex 100 MG/ML"
  "ALLERGENIC EXTRACT, DUST, AUTOGENOUS 1 UNT/ML"

- SCDF (Semantic clinical drug and form) [14,548 occurrences]
  "Triamcinolone Oral Paste"
  "Ephedrine / Phenobarbital / Potassium Iodide / Theophylline Oral Tablet"

- SCN (Scientific name) [208,363occurrences]
  "Abelson murine leukemia virus"
  "Absidia"

- SD (CCS single-level diagnosis categories) [284 occurrences]
  "Abdominal pain"
  "Spontaneous abortion"

- SI (Name of a sign or symptom of a problem) [377 occurrences]
  "allergens"
  "anemia"

- SP (CCS single-level procedure categories) [234 occurrences]
  "Diagnostic amniocentesis"
  "Abortion (termination of pregnancy)"

- SS (Synonymous "short" forms) [196 occurrences]
  "Alzheimer disease"
  "adrenoleukodystrophy"

- ST (Step) [132 occurrences]
  "Manage contracts"
  "Interview the patient or patient's representative"

- SU (Active Substance) [1,834 occurrences]
  "Acetylcysteine"
  "vecuronium bromide"

- SY (Designated synonym) [458,131 occurrences]
  "Branching enzyme"
  "Amylo-(1,4,6)-transglycosylase"

- SYGB (British synonym) [8,063 occurrences]
  "Tumour of abdomen"
  "ABL - Abetalipoproteinaemia"

- SYN (Designated alias) [67,497 occurrences]
  "ADULT NEURONAL CEROID LIPOFUSCINOSIS"
  "WEBER-COCKAYNE TYPE EPIDERMOLYSIS BULLOSA SIMPLEX"

- 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]
  "GI_NOS"
  "ABDOMEN"

- TG (Name of the target of an intervention) [76 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 gen-

eral 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]
  "GI_NOS PROBLEM"
  "ABDOMEN PROBLEM"

- 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) [67 occurrences]
  These terms are similar to TQ terms, and the same comments apply.
  "anomalies"
  "teratology"

-