# Filtering the UMLS® Metathesaurus® for MetaMap

1999 Edition

**Alan R. Aronson**

March 12, 1999

## 1. Overview

MetaMap's primary purpose is to provide a basis for further processing of biomedical text by finding the Metathesaurus concepts referred to in the 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 1999, for example, the Metathesaurus has 1,163,141 distinct English strings comprising 626,893 concepts. 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 540 such strings:

- Numbers (e.g., 2, +1, -4, 98.734, 50000) [142 occurrences];
- Single alphabetic characters (e.g., a, A, b, B <3>; note the ambiguity designator) [125 occurrences];
- Unnecessarily ambiguous terms [272 occurrences]
  "Other <2>", … (however, "Other" itself is retained)

"Protocols <1>', … (however, "Protocols" itself is retained)
…

- Special cases [1 occurrence]
  "0" for "TPBF protein" (from RCD98) [included in numbers]
  "Periods" for "Menstruation" (from RCD98)

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;
- NEC/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]", "[Q]" and "[SO]". 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 1999, 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.)

### 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 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; 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 10% (116,112) of the 1,163,141 distinct English Metathesaurus strings are suppressible synonyms.

Similarly, some Term Type (TTY) values indicate strings which are normally inappropriate for text processing. There are 159,526 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 44,341 (28%) of the 159,526 TTY strings are *not* suppressible synonyms.) The term types filtered out together with examples for each are given in the next section.

### 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) [73,500 occurrences]
  “ABN INVOLUN MOVEMENT NEC” (a concept with no other strings)
  “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) [13 occurrences]
  “Med exam/health evalua/complete”
  “Microbio/other immunol test”

- HX (Expanded version of short hierarchical term) [5,577 occurrences; added in 1999]
  “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”)

- IS (Obsolete synthesized term in the Read Thesaurus) [3,551 occurrences]
  “Fall - accidental” (for concept “Accidental Falls”)
  “MVTA - Motor vehicle traffic accident” (for concept “Accidents, Traffic”)

- LN (LOINC official fully specified name) [13,300 occurrences]
  “ALMECILLIN:SUSC:PT:ISLT:QN:MIC”
  “ALMECILLIN:SUSC:PT:ISLT:SQ:AGAR DIFFUSION”

- LO (Obsolete official fully specified name) [165 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) [13,465 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) [48,441 occurrences]
  “Gastrin secretion abnorm.NOS”
  “Spontaneous abort.incomp.NOS”

- OM (Obsolete modifiers in HCPCS) [16 occurrences]
  “POWDERED ENTERAL FORMULAE (THIS SHOULD BE USED WHEN ENTERAL POWDERED PRODUCTS ARE SUPPLIED)”
  “DISTINCT PROCEDURAL SERVICE: THE PHYSICIAN MAY NEED TO INDICATE THAT A PROCEDURE OR SERVICE WAS DISTINCT OR SEPARATE FROM OTHER SERVICES PERFORMED ON THE SAME DAY. THIS MAY REPRESENT A DIFFERENT SESSION OR PATIENT ENCOUNTER, DIFFERENT PROCEDURE OR SURGERY, DIFFERENT SITE, SEPARATE LESION, OR SEPARATE INJURY (OR AREA OF INJURY IN EXTENSIVE INJURIES).”

- PS (Short forms that needed full specification) [1,464 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.

- CD (Clinical Drug) [6,874 occurrences]
  These terms describe a quantity of some drug; they may or may not frequently occur in text.

"HYDROGEN PEROXIDE 3% solution"
"ICHTHAMMOL 20% ointment"

- LS (Expanded system/sample type (The expanded version was created for the Metathesaurus and includes the full name of some abbreviations.)) [449 occurrences]
  About one-third of these terms contain embedded periods.
  "HEART.AORTIC VALVE"
  "AORTA.THORACIC.ASCENDING"
- 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)) [2,881 occurrences]
  These terms often contain characters indicating a superscript.
  "Ca^2+^-transporting ATPase"
  "alpha>1< Antichymotrypsin"

## 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. As of 1999, however, *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 been included in the baseline syntactic filtering because of their tractability. Less strict filtering might involve considering both the number of phrases and the number of syntactic units in the phrases and would be useful for term processing or browsing. An analysis of the syntactic properties of all Metathesaurus strings is in progress.

## 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 608,887 (52%) of the 1,163,141 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 846,841 (73%) 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 991,809 (85%) English Metathesaurus strings.