# Discovering Missed Synonymy in a Large Concept-Oriented Metathesaurus

William T. Hole, M.D.,
Suresh Srinivasan, M.S.,
National Library of Medicine, Bethesda, MD

**Abstract**

The Unified Medical Language System® (UMLS®) [1, 2] Metathesuarus® is concept-oriented; its goal is to unite all names with identical meaning in a single Concept. The names come from its constituent vocabularies or "sources" - a wide variety of biomedical terminologies including many controlled vocabularies and classifications used in patient records, administrative health data, bibliographic, research, full-text, and expert systems. Many offer little definitional information, and many are not themselves concept-oriented, so identifying synonymy is a challenging semantic task [3]. The rapidly increasing size of the Metathesaurus makes the task daunting, demanding effective computational support; there are more than 1.5 million names for 730,000 concepts in the January 2000 release.

Vocabularies are added and updated using sophisticated lexical matching, selective algorithms, and expert review [4, 5, 6]. Yet the result is imperfect; we have discovered and corrected missed synonymy in approximately 1% of previously released concepts each year. This paper reviews general methods for finding missed synonymy and describes several specific novel approaches which we have found effective.

**An Overview of the Metathesaurus Process**

New sources or updates to existing sources are inserted into the Metathesaurus after mapping the source's schema into the Metathesaurus schema – a process called *inversion*. After insertion, a merge process attempts to merge the new terms with terms in existing concepts subject to a variety of constraints and controls. Constraints are generally source specific, an example being "don't merge CPT-4 procedures with chemical names, since CPT-4 will mean the laboratory test for the chemical".

In a source update, where the merging is largely between versions, it helps if the source is itself strongly concept-oriented, e.g., if the source is consistent about assigning and maintaining a unique identifier or code for each meaning of its terms. In such cases new terms with codes identical to a previous version's, already in Metathesaurus, may be correctly merged into existing concepts.

For merging between sources, flexible lexical matching using the *norm* program (see below) is generally employed. If a new term is *norm*- identical with an existing term, the new term is tentatively merged and the concept is marked as needing review. An editor makes the final decision on whether the terms in a Reviewed[1] concept are indeed synonymous, or whether one (or more) terms need to be moved to another concept.

*Norm* is part of the Lexical Variant Generation (LVG) [7] package that is distributed with the UMLS. It is a tool for creating a canonical lexical form of an English string. *Norm* abstracts away from differences in case, punctuation, word order and inflectional variation. For example "protein deficiency" and "Deficiencies, Protein" would have the same normalized form ("deficiency protein"). The Metathesaurus includes two files, MRXNS.ENG and MRXNW.ENG, containing respectively the normalized form of every string and every word in the Metathesaurus. These files and LVG are very useful in searching for missed synonymy. We recognize that LVG is less effective for chemical naming; research on a chemical version is under way at NLM.

Expert editors make final decisions about synonymy for all reviewed concepts in the UMLS Metathesaurus. The principle is to preserve all distinctions important to any significant area of biomedicine as different concepts. In our experience

[1] It should be noted that the Metathesaurus contains 242,013 Unreviewed concepts from the MeSH Supplementary Concepts. Unreviewed concepts are clearly labeled in the Metathesaurus and may be excluded; they are included as a valuable source of information not available elsewhere. This source names less frequently used chemicals, biomedical materials, and protocols; the Unreviewed concepts did not match any other vocabulary. This source is not yet concept-oriented so there are missed synonyms, particularly within the source itself; the closely related clusters are related as "RL" ("like"). As MeSH assumes its new concept orientation the correct concept structure will be assigned [9].

it is not difficult for an editor to apply this principle in most cases.

**Determining Intended Meaning**

The Metathesaurus represents meanings in its sources. In some ways, this is an easier task than determining face meaning, common usage, or scientific truth. All these may change with time, increased knowledge, or may differ for particular users or disciplines. Yet few vocabularies provide much information about the meanings of their terms. Where present, definitions, scope notes, hierarchies, relationships, other attributes, and the nature and purposes of the vocabulary provide clues to the intended meaning.

Many non-concept-oriented vocabularies have entry terms (sometimes called "synonyms") which map related meanings to a core set of controlled terms; these may represent either synonymy or somehow related meanings.

In Metathesaurus construction, identical or lexically similar names are assumed to represent the same meaning, absent information to the contrary. Similarly, synonymy asserted in one vocabulary is applied transitively to all unless there is contradictory evidence. Expert review then confirms or undoes these assumptions. When sources contradict each other, the editor must determine scientifically correct meanings.

**What is a Synonym?**

In our experience, most cases are simple with little disagreement; more difficult are cases with subtle distinctions in context or cases where some may view a very broad aggregation of meanings as a single concept. Different thesauri with different purposes or interpretations of "concepts" may also represent differing views.

The Metathesaurus asserts that there exists a useful level of granularity, one that maintains distinctions important to any area of biomedicine as separate concepts. Alternate views may be derived by differing aggregations of Metathesaurus concepts, using Metathesaurus relationships. For example, mappings to a less granular classification such as ICD-9-CM can form a classification view; or relationships such as those from the Canonical Clinical Problem Statement System may be used to form a clinical problem-oriented view.

**Examples of Synonymy Problems**

In the simplest case, names are identical and there is no disagreement; in other cases of identical names, hierarchies show clear differences in granularity.

In some cases, identical names clearly have differing meanings, most egregiously when the hierarchy is not expressed:

```
ICD10 D07.5
 International Statistical
 Classification of Diseases and
 Related Health Problems, Tenth
 Revision (ICD-10)
  Neoplasms
   In situ neoplasms
    <Prostate>
```

The context may include the nature of the vocabulary, for example that it contains procedures:

```
CPT2000 82728
 Current Procedural Terminology
  Pathology and Laboratory Tests
   Chemistry Pathology and
   Laboratory Tests
     <Ferritin>
```

Other cases are more involved, for example when synonymy reveals differing views; UWDA, a detailed anatomical source, asserts that "Posterior descending artery" is a synonym for the "Posterior interventricular branch of right coronary artery"; yet the Read Codes use "Posterior descending artery" as a parent for the two variant forms arising from the left or right coronary arteries.

```
Read thesaurus RCD99 X74eR
 …
 Cardiac structure
  Coronary artery
   <Posterior descending artery>
     Left dominant posterior descending
        artery
     Right dominant posterior descending
        artery
```

This example hints at the complexities which may occur when the same string has different meanings in different vocabularies, polysemy or "multiple meanings" [8].

**General Methods to Identify Synonymy**

1. *General approximate matching algorithms*

We have tested several approaches to approximate matching which identified massive numbers of potential synonyms but very few actual synonyms. No general algorithm has yet proved effective.

2. *Creation of standardized synonymous names prior to lexical matching*
These names, for example, express implied context; eliminate extraneous parenthetical information; or create Americanized versions of British forms. An example is "Hemolytic anemia" for "Haemolytic anaemia."

3. *Targeted mapping between vocabulary pair*
This systematic approach creates useful relationships (including "not related") which enhance the Metathesaurus. Multiple efforts mapping to a common target will yield transitive synonymy, but productivity is not high. An example is the mapping of ICD-9-CM to MeSH.

4. *Editor-directed searches for selected normalized words*
This approach can be very effective and can lead to useful algorithmic techniques. Unfortunately, any editor who takes pride in his or her work is tempted to spend a great deal of effort in the search, which can only be justified for high priority areas.

5. *Exploiting source semantics*
Editor training about the nature of sources, their hierarchies, and their naming styles allows the most effective use of expertise in searching and can help editors discover effective algorithms. It also may consume large amounts of effort with limited yield.

6. *Users' contributions*
Comments, small or large sets of missed synonyms, reports of patterns suggesting useful algorithms, or algorithms themselves are vital contributions to the quality of the Metathesaurus.

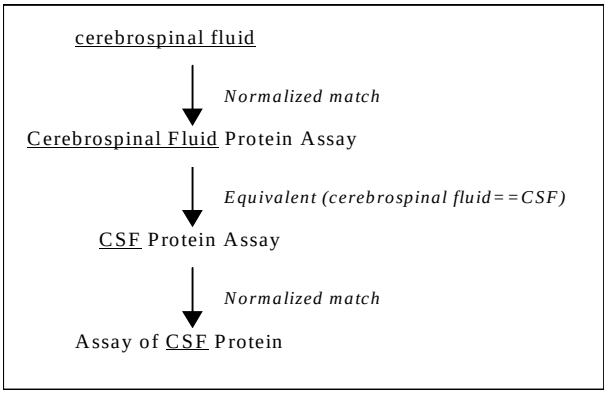
**Unique Identifiers when Concepts are Merged**

Since Concept Unique Identifiers ("CUIs") must track meanings over time, the CUIs of merged concepts must be explicitly mapped. Since only one CUI is allowed in the Metathesaurus schema, a file (MERGED.CUI) mapping all "losing" merged CUIs to the corresponding CUI that is still present is part of each Metathesaurus release.

**Novel Algorithmic Approaches**

Several algorithmic approaches used to identify potential missed synonymy rely on a phrase substitution program called *phrasesub* that internally uses *norm*. This program forms the basic building block for the different approaches described below.

The logic of the program is illustrated by the following diagram for the example of "Cerebrospinal Fluid Protein":




*Lexical Algorithm*

The lexical algorithm for finding missed synonymy starts where *norm* leaves off. As editors notice patterns of missed synonymy that *norm* failed to identify, we fold these into our missed synonymy suite of queries. This is best illustrated with a few examples.

1. We noted different ways of specifying dosage patterns in newly inserted drug thesauri and existing Metathesaurus vocabularies. The difference may be in the units, e.g., "5 mg per 5ml" vs "1 mg/ml" or in the absence of a space between the amount and unit, "5mg" vs. "5 mg". These differences, while trivial at first glance, are beyond the scope of *norm* to detect as equivalent so the post-insertion merging did not occur – a classic case of missed synonymy. One or more abbreviations can also be dealt with using this technique. For example, *tablet* is equivalent to *tab* and *citrate* is equivalent to *cit* in the names "Tamoxifen citrate CP 20mg tablets" and "Tamoxifen cit CP 20mg tab".

2. In some sources, the parenthetical expression "(all forms)" is appended to a drug name to mean all forms of delivery – tablets, liquid, etc. It was judged that the meaning is identical to other terms without this appended expression, allowing algorithmic candidate merges.

3. Many vocabularies represent the same general meaning at two levels of their hierarchies, often adding "NOS" (Not Otherwise Specified) to the parent term. Matching without these or similar differences allows candidate merges.

In each case the equivalent phrases are fed to *phrasesub* to find potential pairs of synonymous concepts. Our queries initially found 1230 pairs of this type of potential missed synonymy, of which editors decided 525 were actual synonyms.

*Word Level Synonymy*

This approach involved using word level synonymy to infer term-level synonymy. Several sources of word level synonymy were used including those extracted from the SPECIALIST lexicon, another of the UMLS Knowledge Sources.

The words are fed directly to the *phrasesub* program to extract terms that only differ in the synonymous words. This approach was not as productive and produced only 340 concept pairs of which 9 were judged to be actual synonyms by editors. In many cases some other type of relationship was assigned. A simple but salient example is the merge of "Renal failure" with "Kidney failure."

*Inferred Phrase Level Synonymy*

This heuristic approach was generously contributed by Randolph A. Miller, MD, of Vanderbilt University. It uses the Metathesaurus itself to derive possible phrase level synonymy. Knowing that all names within a concept are identical in meaning, equivalent word clusters are obtained by removing words in common between all concept names, examined in pairs.

For example, the names "Relatives died" and "Relatives deceased" are present in concept C0557091. Dropping the common word "Relative", allows us to infer potential synonymy between "deceased" and "died".

If the result maps a single word to one or more words, the case is considered for further review; mappings between multi-word phrases are discarded at this time to keep the result to a manageable number. Only English language names were used and some punctuation and stop words were also ignored in the process. There were many dubious and incorrect suggestions from this algorithm, necessitating human review of the resulting phrases. Examples of these incorrect cases are "Automatic" and "computer" or "Birth" and "sibling."

74,159 word or phrase synonyms were identified by this algorithm; 24,005 (32%) were selected by human review as worthy of further investigation. Those selected generated 65,477 concept pairs for review, of which 4,024 (6.2%) were merged by editors.

The following table shows the comparative merits of each of these methods in initial use:

| Method | Potential | Merged | % |
|---|---|---|---|
| Lexical Techniques | 1230 | 525 | 43% |
| Word Synonymy | 340 | 9 | 3% |
| Phrase Synonymy | 65,477 | 4024 | 6% |

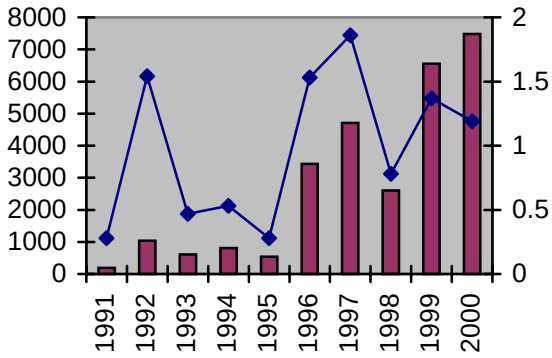
Subsequent incremental runs had smaller yields as is to be expected, since our Editing Management System (EMS) tracks the review of each concept pair and does not schedule repeat reviews for missed synonymy.

These approaches are now used regularly in each editing cycle. They are computationally expensive, requiring days or weeks on backup systems. Yet they are cost-effective since they leverage our most expensive resource: the expert editors.

**Conclusions**

Figure 1 shows the number of merges of previously released concepts for all releases of the UMLS Metathesaurus. The approaches described in this paper account for the majority of merges in 1999 and 2000 while previous years appear to represent primarily additional synonymy in new or updated sources.

**Figure 1: Merges of Previously Released Metathesaurus Concepts, by Year**




Legend: *Bars indicate number of concepts merged; the line indicates the percentage of concepts merged.*

While we have made significant progress in attacking this problem with assistance from the UMLS user

community, much more can be done. This task requires unremitting attention as new sources and updates are inserted. We are currently exploring automated methods to exploit combinations of these methods and to develop ways to mine source semantics and external information sources effectively to predict possible synonymy.

Improvements in science, vocabulary standards, and practices in biomedical vocabularies will lead to concept-oriented thesauri with better naming, which eliminates implied context and other idiosyncrasies which obscure meaning; more explicit definitional information; and concept-oriented links to other vocabularies, supplied by the authors - who clearly understand their own meanings best. These improvements will assist all who grapple with biomedical meaning in the service of science and health.

**References**


1. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993 Aug; 32 (4): 281-91.

2. Humphreys BL, Lindberg DA. Building the Unified Medical Language System. Proc Annu Symp Comput Appl Med Care 1989:475-80.

3. Cimino JJ. Auditing the Unified Medical Language System with Semantic Methods. Department of Medical Informatics. Proc Annu Symp Comput Appl Med Care 1994:145-9.

4. Tuttle MS, Suarez-Munist ON, Olson NE, et al. Merging Terminologies. Medinfo 1995; 8 Pt 1:162-6.

5. Sherertz DD, Olson NE, Tuttle MS, Sperzel WD, Erlbaum MS, Fuller LF. The META-1 Engine: a Database Methodology Used in Building the UMLS METATHESAURUS. Medinfo 1992; 7(Pt 1):144-9.

6. Suarez-Munist ON, Tuttle MS, Olson NE, et al. MEME II Supports the Cooperative Management of Terminology. Proc AMIA Fall Symp 1996:84-8.

7. McCray AT, Srinivasan S, Browne AC. Lexical Methods for Managing Variation in Biomedical Terminologies. Proc Annu Symp Comput Appl Med Care 1994:235-9.

8. Roth LA, Hole WT. Managing Name Ambiguity in the UMLS Metathesaurus. Submitted in Proc AMIA Annu Fall Symp 2000.

9. Johnston D, Nelson SJ, Schulman JA, Savage AG, Powell TP. Redefining a Thesaurus: Term-Centric No More. J Am Med Informatics Assoc (Symposium Suppl). 1998:1025.