# Finding UMLS Metathesaurus Concepts in MEDLINE

Suresh Srinivasan, M.S., Thomas C. Rindflesch, Ph.D., William T. Hole, M.D., Alan R. Aronson, PhD.,
James G. Mork, M.S.,
National Library of Medicine, Bethesda, MD

**Abstract**

*The entire collection of about 11.5 million MEDLINE abstracts was processed to extract 549 million noun phrases using a shallow syntactic parser. English language strings in the 2002 and 2001 releases of the UMLS Metathesaurus were then matched against these phrases using flexible matching techniques. 34% of the Metathesaurus names occurring in 30% of the concepts were found in the titles and abstracts of articles in the literature. The matching concepts are fairly evenly chemical and non-chemical in nature and span a wide spectrum of semantic types. This paper details the approach taken and the results of the analysis.*

## Introduction

This study is concerned with the extent to which the concept names of the UMLS® Metathesaurus® are represented in the free text of the full MEDLINE® collection. The Metathesaurus is one of the knowledge sources of the UMLS and organizes biomedical information by meaning. In its January 2002 edition, it contains 776,940 concepts and well over 2 million names for these concepts. The strings representing these names come from 108 sources in 15 languages [1]. MEDLINE is the U.S. National Library of Medicine's premier bibliographic database that contains over 11 million references to scientific journal articles. It contains citations from over 4,600 journals from 1966 to the present [2].

There is considerable interest in the use of phrases for information management technology applications, such as phrase browsing [3], indexing or provision of hyperlinks in text [4]. Research in query expansion ([5, 6], for example) seeks to provide the best match between input phrases and indexing terms. Earlier work has extracted noun phrases from the complete MEDLINE database in order to support research on content analysis for information retrieval [7]. In this study we investigate the potential for identifying phrases in the biomedical research literature in order to enhance the currency and quality of the UMLS Metathesaurus when editing costs can be justified. Such an enhancement could have a positive impact on retrieval systems that rely on the Metathesaurus, such as MeSH-centric query expansion in PubMed [8]. The methodology used can be generalized to other applications that look for biomedical content in free text corpora such as patient records [9].

Our phrase extraction program (PhraseX) relies on an underspecified syntactic parse as well as a parallel processing architecture (SKR scheduler) for efficient extraction of noun phrases from the complete MEDLINE database. Subsequently, these phrases are compared to strings of interest from the Metathesaurus. Matching is conducted at varying degrees of strictness, from exact matching to a flexible matching scheme that ignores case, punctuation, word order and inflectional variation.

## PhraseX – Extracting Noun Phrases

PhraseX is a program that extracts noun phrases from text such as MEDLINE abstracts. It does so by referring to the syntactic structure provided by the SPECIALIST minimal commitment parser [10], which relies on the SPECIALIST lexicon [11] as well as the XEROX stochastic tagger to resolve part-of-speech ambiguity [12]. The notion of barrier words within sentences is then used to delimit phrase boundaries. PhraseX uses the underspecified structure from the parser to output three kinds of noun phrases. For example, based on the structure assigned to the sentence *Characteristics of the affection were studied and different therapies applied*, "**simp**" phrases are those with a head noun, e.g., *characteristics*, "**macro**" phrases are those that have prepositional modification to the right, such as *characteristics of affection*. The first preposition is unconstrained, but the rest must be *of*. The third kind of phrase, denoted "**mega**" includes all the content words in the sentence to the left and right of a finite verb, e.g., *characteristics of affection studied and different therapies applied.* In many cases, the **mega** phrases are not syntactically well-formed noun phrases, but neither are many Metathesaurus names, so this has proved to be a useful method of analysis. PhraseX tokenizes its input using non-alphanumeric characters as token separators and preserves case. It drops determiners and pronouns in the resulting phrases.

### SKR Scheduler

The SKR scheduler (henceforth the "scheduler") is a program to support parallel execution of other long-running programs. It was developed as part of the Semantic Knowledge Representation (SKR) project at NLM. The scheduler was crucial for being able to process large quantities of MEDLINE abstracts.

The scheduler starts daemon processes on one or more client machines using the UNIX remote shell (rsh) program. Once started, the clients communicate their results back to the scheduler via UNIX-domain sockets. The input is distributed to the available client pool in round-robin fashion. The results are reported back to the scheduler to be collated and assembled in correct order. The system is remarkably fault-tolerant and can deal with client workstations being rebooted or encountering processing errors.

There is an easy-to-use Web interface to configure the scheduler for the available client machines and their times of availability. The clients run the daemons with the lowest priority so as not to interfere with normal processes.

### MEDLINE Processing

The full complement of MEDLINE citations was obtained from PubMed in the Fall of 2001 for this analysis. These were in 24 files using 14GB of disk space, each containing either a million, or half a million records for a total of 11,541,221 citations. The scheduler used 24 client workstations for 28 days (some all day, and some off-hours only) for a cumulative average processing rate of 25,000 citations per hour.

A grand total of 549,057,568 phrases (175,147,356 unique) were extracted from text in the titles and abstracts of these citations (taking up about 20GB of disk space). Figure 1 shows the percentages of different phrase types generated.

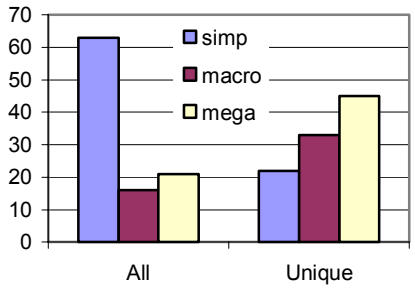


**Figure 1 Percentage of phrase types**

For all the phrases, the majority (63%) are **simp** phrases, with **mega** phrases slightly more frequent (21%) than **macro** phrases (16%). If duplicates are eliminated, the proportion of *unique* **simp** phrases drops off to 22% while **macro** and **mega** phrases rise to 33% and 45% respectively. This mirrors the fact that the simpler phrases occur in many citations. For example, for the sentence: *The average for delay of diagnosis was 7 days and for hospitalization was 24 days*, PhraseX produces the following phrases in each category (separated by a semicolon):

| | |
|---|---|
| **simp** | average; delay; diagnosis; days; hospitalization; days |
| **macro** | average for delay of diagnosis |
| **mega** | average for delay of diagnosis 7 days and for hospitalization 24 days |

The MEDLINE phrases and Metathesaurus strings were matched as exact, lowercase (lc) or normalized string matches. The notion of normalization is in the sense defined by the LVG suite's **luiNorm** program as distributed with the 2001 release of the UMLS. Essentially, a normalized form of a string abstracts away from case variation, punctuation, possessive markers, inflectional variation and word order. For example, both "Bone Losses, Age-Related" and "Age-Related Bone Loss" would normalize to "age bone loss relate". So a normalized match would consider these strings as equivalent. Lowercase and normalized versions of the phrases for both MEDLINE phrases and Metathesaurus strings were computed.

### Extracting Metathesaurus Names

We looked at two versions of the UMLS Metathesaurus: 2002 and 2001 and matched the concept names (or strings) to the 549 million phrases extracted from MEDLINE.

There are 1,912,340 names for 776,940 concepts in 2002. From these, we eliminated non-English strings and strings with a suppressible term type (TTY="s"). This latter group is predominantly made up of strings that are idiosyncratic abbreviations, obsolete synonyms, or names with little or no face validity. Table 1 shows the counts of concepts and strings for the two versions of the UMLS and the number that resulted after applying the above criteria.

| Version | Concepts | Distinct Strings | Concepts Used | Strings Used |
|---|---|---|---|---|
| 2002 | 776,940 | 1,912,340 | 776,940 | 1,451,824 |
| 2001 | 797,359 | 1,728,075 | 797,357 | 1,322,616 |

**Table 1 Metathesaurus Concept and String Counts**

### Match Results

Overall for 2002, we found that using normalized matching, **30%** of the concepts have at least one matching name that was in the MEDLINE phrase list. For 2001, the match proportion is nearly as high - **29%**. In general, the 2001 results mirror the 2002 results closely. Here are some examples of strings that matched for illustrative purposes.

| Type | Metathesaurus string | Matched string |
|---|---|---|
| **exact** | ribonuclease | ribonuclease |
| | ethnic differences | ethnic differences |
| **lc** only | Stage IC | stage ic |
| | FSH Receptors | fsh receptors |
| **norm** only | Lyell's syndrome | lyell syndrome |
| | Conjugates, Cytotoxin-Antibody | antibody conjugate cytotoxin |

In Table 2 below are shown the counts and percentages of concept and strings that were matched from the 2002 Metathesaurus. The counts are broken down by the type of match and can be interpreted as follows.

The "Total" column is the number of distinct concepts or strings that matched for a given match type. The "%" column is the appropriate percentage of the concepts or strings while the column labeled "Added" is the additional contribution to the number of matches for the given match type.

| | Match Type | Total | % | Added |
|---|---|---|---|---|
| Distinct concepts matched | exact | 51,743 | 6.65 % | – |
| | lc | 175,145 | 22.54 % | 123,402 |
| | norm | 233,312 | 30.03 % | 58,167 |
| Distinct strings matched | exact | 61,444 | 4.23 % | – |
| | lc | 296,114 | 20.39 % | 234,670 |
| | norm | 498,165 | 34.31 % | 202,051 |

**Table 2 Match results for 2002 Metathesaurus**

For example, there were 175,145 concepts whose names matched MEDLINE phrases ignoring case. This is 22.54% of the 776,940 concepts that we started with. This number also includes the 51,743 concepts whose names matched some MEDLINE phrase exactly. But the contribution of just the case-insensitive matches is an additional 123,402 concepts.

The data for the 2001 UMLS Metathesaurus is shown in Table 3 and is largely similar to that for 2002.

| | Match Type | Total | % | Added |
|---|---|---|---|---|
| Distinct concepts matched | exact | 52,846 | 6.62 % | – |
| | lc | 171,792 | 21.54 % | 118,946 |
| | norm | 233,910 | 29.33 % | 62,118 |
| Distinct strings matched | exact | 59,999 | 4.53 % | – |
| | lc | 279,358 | 21.12 % | 219,359 |
| | norm | 466,949 | 35.30 % | 187,591 |

**Table 3 Match results for 2001 Metathesaurus**

The proportion of successful matches for each phrase type is shown in Figure 2 for the 2002 Metathesaurus. It is interesting to note that while the **simp** algorithm was the most successful for all three types of matches, the **mega** phrases did far better than the **macro** phrases.

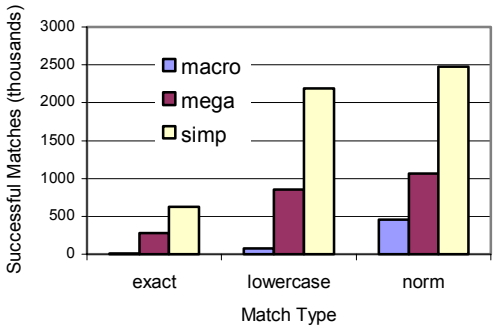


**Figure 2 Counts by match and phrase types for 2002 Metathesaurus**

### Matches by Semantic Type

The Semantic Network, which is another knowledge source of the UMLS, contains 134 broad categories or Semantic Types (STY's) for classifying concepts in the Metathesaurus [1]. A concept that is multifaceted or has inherent ambiguity is assigned more than one STY. For example "Febrile Convulsion" is both a "Finding" and a "Disease or Syndrome". STY's can be broadly categorized into chemical or non-chemical.

We investigated whether the techniques we used were more successful for non-chemicals. Table 4 shows that nearly equal proportions of chemicals as non-chemicals matched when considering the number of unique strings tagged as chemicals and the proportion of them that matched.

| | Strings | Matches | % |
|---|---|---|---|
| Chemical | 579,421 | 177,636 | 30 % |
| Non-chemical | 873,280 | 321,386 | 36 % |

**Table 4 String matches for 2002 by chem/non-chem**

This is somewhat surprising given the large number of formulaic chemical names in the Metathesaurus. An informal analysis of the chemical names indicates that about half are systematic names (roughly estimated by the presence of numeric characters) and only 12% of these matched. So the bulk of the chemical matches were to ordinary names e.g., "Arsenic". But overall the results indicate that the phrase extraction and matching algorithms work well for chemical terms.

Another question that can be asked is how do the matches break down by individual STY's. If we normalize the number of matches for each STY by the frequency of that STY in MRSTY [1], and plot the frequency histogram, we see, as shown in Figure 3, that the bulk of the STY's have a match percentage between 30% and 80% with a mean of 58%.

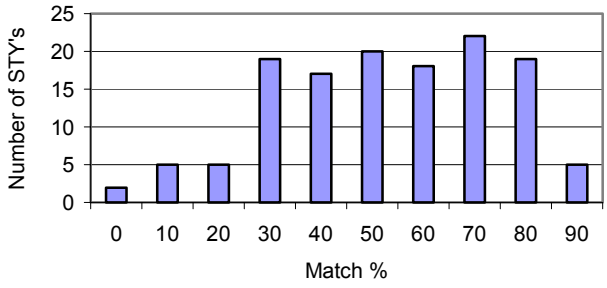


**Figure 3 Frequency histogram of match % for all STY's**

For example, 20 STY's had a match % between 50% and 60%. The data show that the coverage is relatively uniform across a range of STY's and that it is not skewed toward a few STY's.

The following table, Table 5, shows the data for some selected STY's for 2002. The column marked "Concepts" is the frequency of this STY in the MRSTY file, the "Matches" column is the number of matches that had this STY, and the "%" column is the match percentage. For example, 42,526 strings in concepts assigned the STY "Organic Chemical" were matched in MEDLINE.

Other types of broad categorization of the matches are possible, say by the semantic type aggregations proposed by McCray, et al [13].

| Semantic Type | Concepts | **Match** | % |
|---|---|---|---|
| Organic Chemical | 107,157 | 42,526 | 40 |
| Amino Acid, Peptide, or Protein | 83,556 | 39,168 | 47 |
| Pharmacologic Substance | 88,431 | 36,057 | 41 |
| Disease or Syndrome | 48,286 | 20,062 | 42 |
| Finding | 43,658 | 13,870 | 32 |
| Therapeutic or Preventive Procedure | 63,962 | 12,219 | 19 |
| Body Part, Organ, or Organ Component | 49,201 | 9,847 | 20 |
| Medical Device | 33,443 | 4,869 | 15 |

**Table 5 Matches for selected STY's (2002).**

### Analysis of Non-Matches

It is instructive to study and categorize the Metathesaurus strings that *did not* match the phrases extracted from MEDLINE. Broadly, of the 953,659 strings (from 622,805 concepts) that weren't matched, 401,785 (42%) were chemicals and 551,894 were not (there were 20 problematic cases with both chemical and non-chemical STY's, as of this writing). Here are some examples and reasons why they didn't match:

1. The Metathesaurus represents constituent sources' content without modifications. In some cases, the strings may not be suitable for natural language use [14] and are thus unlikely to appear in running text. For example, terms from LOINC[1] are highly structured identifiers for laboratory and other clinical observations, e.g., "BACAMPICILLIN:SUSC:PT:ISLT:QN:MLC". In other instances, terms are descriptive and may run to great lengths; this one has 31 words: "ELECTROCARDIOGRAPHIC MONITORING FOR 24 HOURS BY CONTINUOUS ORIGINAL ECG WAVEFORM RECORDING AND STORAGE …". Yet others are formal IUPAC[2] chemical names, e.g., "5,7-dihydro-1,2,3,9,10,11-hexahydroxydibenz(c,e)oxepin". We recognize that we may not be able to match some of these Metathesaurus terms in a natural language corpus.
2. A purely syntactic analysis, such as that provided by PhraseX is not able to completely address the ambiguity of natural language. For example PhraseX failed to identify the noun phrase "oxonin" in the following MEDLINE title: *Staining behavior and applicability of spectrally pure Capriblue GN, Stella Blue, Oxonin and Punky Blue*. Consequently,

[1] Logical Observations Identifiers, Names, and Codes
[2] http://www.iupac.org

the Metathesaurus string "Oxonin" was not matched. The reason for this error is that PhraseX analyzed *stella blue oxonin* as a noun phrase, rather than the two phrases *stella blue* and *oxonin*. The presence of the comma after the word *Blue* caused the parser to interpret it incorrectly as an adjective instead of a noun..

3. Some Metathesaurus terms, e.g., some trade names, are simply not found in MEDLINE titles and abstracts. Examples: "Gammolin", "Bensulfide".
4. The constraint on the **macro** algorithm which requires that all but the first prepositional phrase in the structure be introduced by "of" is may be too conservative.
5. The PhraseX algorithm attempts to structure the input stream into syntactically well-formed phrases. We then constrained the matching algorithm to strictly match only complete phrases identified by PhraseX; there was no attempt to find partial or subsumptive matches. For example, the Metathesaurus string "duodenal string" was not allowed to match the MEDLINE phrase "duodenal string capsule", which was in fact found by PhraseX. One reason for our caution in this regard is that the internal semantic structure of noun phrases like "duodenal string capsule" is ambiguous between "duodenal string" modifying "capsule" and "duodenal" modifying "string capsule". Determining partial matches in cases such as these remains a topic for future research.

## Conclusions

The scope of this study encompasses all of MEDLINE. The millions of phrases extracted from the MEDLINE corpus will be a continuing area of investigation. Our methodology, which was able to successfully identify 34% of Metathesaurus strings in the free text of titles and abstracts, indicates that many concept names in the Metathesaurus are being used in the literature. The results of this research create opportunities for enhancing the Metathesaurus. Meaningful high frequency phrases can potentially be added to the Metathesaurus (with expert human review) for use in NLP and other applications. Statistical comparisons of MEDLINE and Metathesaurus content may point to areas in the Metathesaurus that are problematic or have weak coverage. It can also lead to co-occurrence analysis analogous to MRCOC, but broader in scope. The techniques described here can be used in applications that seek to provide hooks to biomedical concepts, for instance, in free text elements of patient record systems.

But there is also gold in what did not match. Further analysis of these strings is warranted, to improve both parser performance and our understanding of Metathesaurus content. Clearly, string equality does not necessarily mean identity of meaning, especially when the strings are normalized. We do know, however, that only 5612 strings (0.3%) in the 2002 Metathesaurus are ambiguously in more than one concept (see the AMBIG.SUI file). So we feel confident that the scale of such false positives is small.

## References


1. National Library of Medicine. Documentation, UMLS Knowledge Sources, 13th Edition, January 2002.
2. National Library of Medicine: MEDLINE factsheet: http://www.nlm.nih.gov/pubs/factsheets
3. Wacholder N, Nevill-Manning C. The technology of phrase browsing applications. SIGIR Forum 2001;35(1):16-20.
4. Kim W, Wilbur WJ. Corpus-based statistical screening for phrase identification. JAMIA 2000;7(5):499-511.
5. Aronson AR, Rindflesch TC. Query expansion using the UMLS Metathesaurus. Proc. AMIA Symp., 1997, 485-9.
6. Srinivasan P. Query expansion and MEDLINE. Inf. Proc & Man. 1996;32(4):431-443.
7. Bennet NA, He Q, Powell K, Schatz BR. Extracting noun phrases for all of MEDLINE. Proc. AMIA Symp., 1999, 671-5.
8. National Library of Medicine. PubMed: from http://www.nlm.nih.gov
9. van Mulligen, E.M., UMLS-based access to CPR data, MEDINFO 1998; 9 Pt 1;166-70
10. Rindflesch TC, Rajan JV, Hunter L. Extracting molecular binding relationships from biomedical text. Appl. Nat. Lang. Process., 2000:188-95.
11. McCray A. T., Srinivasan S., and Browne A. C. Lexical methods for managing variation in biomedical terminologies. Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 235-9.
12. Cutting DR, Kupiec J, Pedersen JO, Sibun P. A practical part-of-speech tagger. Proceedings of the Third Conference on Applied Natural Language Processing, 1992.
13. Alexa T. McCray, Anita Burgun, Olivier Bodenreider. Aggregating UMLS Semantic Types for Reducing Conceptual Complexity, Medinfo 2001;10(Pt 1):216-20
14. Alexa T. McCray, Olivier Bodenreider, James D. Malley, Allen C. Browne. Evaluating UMLS Strings for Natural Language Processing, Proceedings of AMIA Annual Symposium 2001:448-452