

An Expert Study Evaluating the UMLS Lexical Metaschema

Li Zhang, George Hripcsak², Yehoshua Perl, Michael Halper³, James Geller

Computer Science Dept. NJIT Newark, NJ 07102 {lxz1853, yehoshua.perl, james.geller}@njit.edu	² Dept. of Biomedical Informatics Columbia University New York, NY 10032 hripcsak@columbia.edu	³ Mathematics & Computer Science Dept. Kean University Union, NJ 07083 mhalper@kean.edu
--	--	--

Reprint requests & all communications to:

Yehoshua Perl, Ph.D.
yehoshua.perl@njit.edu
CS Dept., NJIT
NJIT, University Heights, Newark, 07102
phone: (973) 596-3392
fax: (973) 642-7029

Abstract

Objective: A metaschema is an abstraction network of the UMLS's Semantic Network (SN) obtained from a connected partition of its collection of semantic types. A *lexical metaschema* was previously derived based on a lexical partition which partitioned the SN into semantic-type groups using identical word-usage among the names of semantic types and the definitions of their respective children. In this paper, a statistical analysis methodology is presented to evaluate the lexical metaschema based on a study involving a group of established UMLS experts.

Design: In the study, each expert was asked to identify subject areas of the SN based on his or her understanding of the various semantic types. For this purpose, the expert scans the SN hierarchy top-down, identifying semantic types, which are important and different enough from their parent semantic types, as roots of their groups. From the response of each expert, an "expert metaschema" is constructed. The different experts' metaschemas can vary widely. So, additional metaschemas are obtained from aggregations of the experts' responses. Of special interest is the *consensus metaschema* which represents an aggregation of a simple majority of the experts' responses. Statistical analysis comparing the lexical metaschema with the experts' metaschemas and the consensus metaschema is presented.

Results: The analysis results shows that 17 out of the 21 meta-semantic types in the lexical metaschema also appear in the consensus metaschema (about 81%). There are 107 (about 79%) semantic types covered by identical meta-semantic types and refinements. The results show the high similarity between the two metaschemas. Furthermore, the statistical analysis shows that the lexical metaschema did not grossly underperform compared to the experts.

Conclusion: Our study shows that the lexical metaschema provides a good approximation for a partition of meaningful subject areas in the SN, when compared to the consensus metaschema capturing the aggregation of a simple majority of the human experts' opinions.

Keywords: Lexical Partition, Lexical Metaschema, UMLS, Semantic Network, Expert Study, Evaluation

1 Introduction

The Semantic Network (SN) [1, 2, 3] is one of the three knowledge resources of the Unified Medical Language System (UMLS) [4, 5]. SN consists of 135 broad categories called semantic types. Pairs of semantic types are connected by hierarchical (IS-A) and non-IS-A semantic relationships (in short, semantic relationships). The SN provides an overarching abstraction of the Metathesaurus (META) [6, 7] which is the concept repository of the UMLS with about 900,000 concepts in [8]. The SN can help in user orientation into the large META knowledge base, as each concept in the META is categorized as belonging to one or more semantic types.

There are about 7,000 semantic relationships connecting pairs of semantic types in the SN. Hence, although the size of the SN is magnitudes smaller than the size of the META, it is still hard for a user to comprehend the SN.

In order to support orientation into the SN, we introduced the notion of a *metaschema* [9]. A metaschema is a higher-level network that serves as a compact abstraction of the SN. As shown in [9], the notion of a metaschema offers various compact partial views which can help users in their orientation to the SN. In [10] an auditing technique for concept categorizations based on a metaschema was presented. These applications of metaschemas are presented in Section 2.3.

A metaschema is based on an underlying partition of the SN into connected groups of semantic types. For example, the cohesive metaschema in [9] is based on our partition [11] of the SN. In our previous work, we derived the *lexical metaschema* [12] based on a lexical partition using identical word-usage among the names of semantic types and the definitions of their respective children. A more detailed description of the lexical partition and lexical metaschema is presented in Section 2.2.

In this paper we will present techniques to evaluate the lexical metaschema's quality. For this purpose, we conducted a study involving a group of experts who published on UMLS research or related topics. In this study, each expert was asked to manually mark semantic types which are deemed as important and sufficiently different compared to their parents. These semantic types serve as roots of semantic-type groups of his/her partition of the SN. By doing this, each expert derived his/her own partition. A metaschema, called *expert metaschema*, can then be built from each such partition. We found that these expert metaschemas vary so widely that they cannot serve as suitable evaluation yardsticks for our lexical metaschema. Therefore, we built a collection of "cumulative metaschemas," each of which represents a level of aggregation of experts' opinions. Of particular interest is the *consensus metaschema* which was selected from these cumulative metaschemas to represent a simple majority of experts' opinions. The lexical metaschema was then compared in detail to experts' metaschemas and the consensus metaschema using a statistical analysis method. The comparison results are presented and analyzed.

2 Background

2.1 Metaschema of the SN

The notion of metaschema was introduced in [9] as an abstraction of the SN. A metaschema is based on a connected partition of the SN where the SN's IS-A hierarchy is partitioned into disjoint *semantic-type groups*. A partition is said to be *connected* if each of its semantic-type groups satisfies the condition that its semantic types together with their respective IS-A links constitute a connected subgraph of the SN with a unique root. Additionally, while a semantic-type group can be a singleton (i.e., can contain only one semantic type), that singleton semantic type cannot be a leaf in the SN's hierarchy. This condition is imposed because the metaschema should manifest some size reduction, which singletons do not contribute to. However, a singleton containing a non-leaf semantic type is allowed, since it may express an important internal branching point in the metaschema.

In a metaschema, each semantic-type group of the partition is represented by a single node, called a *meta-semantic type* (MST). Two kinds of relationships connect meta-semantic types. The hierarchical *meta-child-of* relationships between meta-semantic types are derived as abstractions of the SN's IS-A links. The non-hierarchical relationships, called *meta-relationships*, are derived from the SN's semantic (non-IS-A) relationships. Details of these derivations were presented in [9, 13].

For example, the hierarchy of the **Event** portion could be partitioned into five semantic-type groups as in Figure 1. Note that one group is a singleton containing only **Event**. Each semantic-type group is represented by a meta-semantic type in the corresponding metaschema, e.g., a meta-

semantic type PHENOMENON OR PROCESS¹ is defined to represent the semantic-type group rooted at **Phenomenon or Process** in Figure 1. The metaschema hierarchy derived from the partition of the **Event** portion is shown in Figure 2.

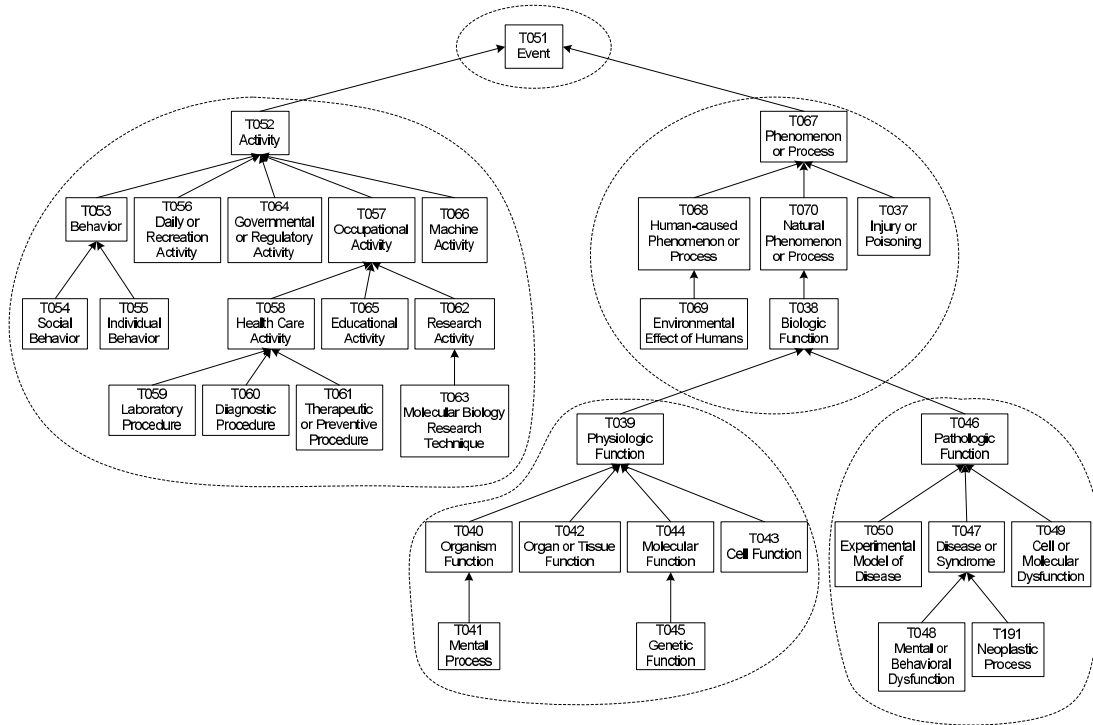


Figure 1: Partition example for the **Event** portion of the SN Hierarchy

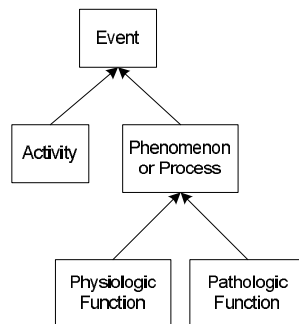


Figure 2: Metaschema hierarchy of the partition of the **Event** portion

In [9] we derived the cohesive metaschema based, with slight modifications, on the structural partition of the SN [11] which groups together semantic types with the same set of semantic relationships. In [13] the notion of a metaschema was extended to a Directed Acyclic Graph (DAG)

¹Meta-semantic types will be written in “small caps” style, except in diagrams.

structure semantic network. This extension was utilized to derive two metaschemas for the DAG-structure UMLS Enriched Semantic Network [14].

2.2 The Lexical Metaschema

The lexical metaschema was derived algorithmically from a lexical partition of the SN which grouped lexically related parent and child semantic types into the same semantic-type group [12].

In [12], a “child/parent pair” (“CP-pair” for short) is defined as a pair of semantic types $(\mathbf{T}_1, \mathbf{T}_2)$ where \mathbf{T}_1 is a child of \mathbf{T}_2 in the SN. A CP-pair $(\mathbf{T}_1, \mathbf{T}_2)$ is a lexically related CP-pair if there exists a string match between \mathbf{T}_1 and \mathbf{T}_2 . Such a string match is defined as a triple $(\mathbf{T}_1; \mathbf{T}_2; S)$ where S is a string appearing both in the definition of \mathbf{T}_1 and the name of \mathbf{T}_2 . That is, if the definition of the child semantic type refers to the name or part of the name of its parent semantic type, then there is a string match between the child and its parent. It is assumed that a string match reflects a lexical relationship between the child and its parent. Therefore, if there is a string match between \mathbf{T}_1 and \mathbf{T}_2 , then the CP-pair $(\mathbf{T}_1, \mathbf{T}_2)$ is lexically related and the child \mathbf{T}_1 is called lexically related to its parent \mathbf{T}_2 ; otherwise, the child was called lexically independent. Based on these definitions, we proposed in [12] an algorithm to identify all lexically related CP-pairs in the SN. Among the 133 CP-pairs, 88 are lexically related. There were 47 lexically independent semantic types (including the two roots **Event** and **Entity**), among which 21 are internal semantic types and 26 are leaves.

Then we defined the lexical partitioning rules as follows. Each non-leaf lexically independent semantic type heads (as a root) a new semantic-type group in the lexical partition. Each lexically independent leaf is assigned to the same semantic-type group as its parent. For each lexically related CP-pair, the child is assigned to the same semantic-type group as its parent. By applying

these partitioning rules, we obtained the lexical partition of the SN containing 21 semantic-type groups.

The lexical metaschema was derived from this lexical partition according to the derivation method used in [9]. Note that the lexical partition for the **Event** portion of SN is the one given in Figure 2. The resulting lexical metaschema contains 21 meta-semantic types, 19 *meta-child-of* relationships, and 86 meta-relationships. The whole lexical metaschema is shown in Figure 3. Each rectangle represents a meta-semantic type. The number in the parenthesis in each rectangle denotes the number of semantic types in the corresponding semantic-type group. Thick arrows denote *meta-child-of* hierarchical relationships between meta-semantic types, while thin arrows denote meta-relationships. Some meta-relationships are represented by numbers in the figure because of space limitations (see legend).

In [12] the lexical metaschema was compared to the cohesive metaschema of [9]. It was shown moderately similar to the cohesive metaschema. On the other hand, when we compared the lexical partition to a partition of the SN derived in [15], we found that those two were quite different. For more detail see [12].

2.3 Applications of Metaschemas

In this section, we briefly describe two applications of a metaschema. These applications were described in detail in [9, 10].

2.3.1 Partial Views Supporting SN Orientation

Overall, a diagram of a metaschema serves as a good visualization mechanism of the SN and, in turn, the META, and helps in the navigation of the UMLS knowledge. In [9] we introduced various

partial graphical views of groups of semantic types supported by the metaschema paradigm. These views can help in orientation of a user to the full scope of the SN's semantic relationships.

In particular we introduce the views of an MST collection subnetwork, an MST focus submetaschema, and bi-collection subnetwork. A collection subnetwork is a subgraph of the SN induced by a semantic-type collection. An MST focus submetaschema contains an MST in which the user is interested (a focus MST) and all its neighboring MSTs. A bi-collection subnetwork is the subgraph of the SN induced by two neighboring collections (i.e., the corresponding MSTs are neighbors) in the metaschema. To illustrate these partial views of the SN in the context of the lexical metaschema, we show the ORGANIC CHEMICAL collection subnetwork, the ORGANIC CHEMICAL focus submetaschema, and the bi-collection subnetwork of PHENOMENON OR PROCESS/ORGANIC CHEMICAL in Figure 4, Figure 5, and Figure 6, respectively.

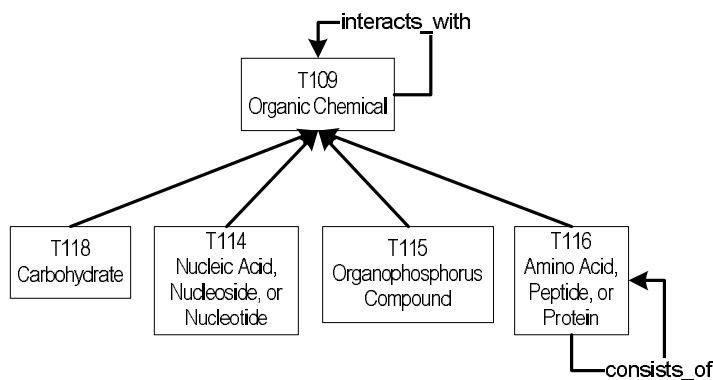


Figure 4: ORGANIC CHEMICAL collection subnetwork

Let us describe a scenario of a user employing these kinds of graphical views to gain an orientation. The user starts by viewing the lexical metaschema (Figure 3) to identify which MST is closest to her interest. Suppose it is ORGANIC CHEMICAL. Then the viewer looks at the ORGANIC CHEMICAL collection subnetwork (Figure 4), and she can see all the semantic types in the collection and all relationships connecting them.

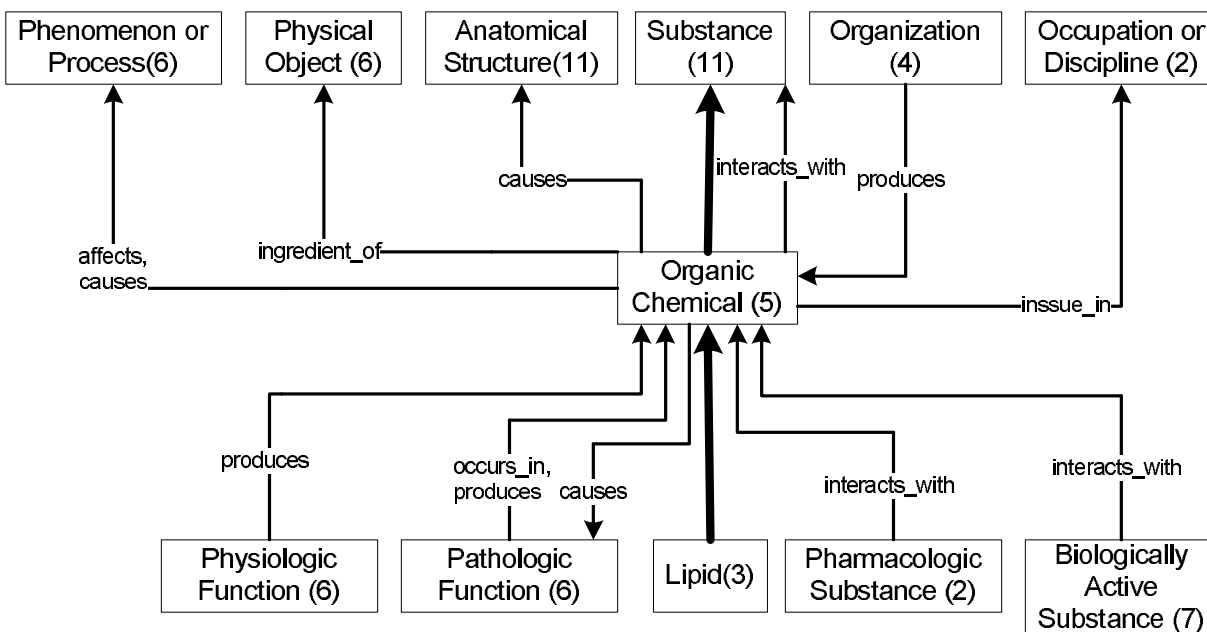


Figure 5: ORGANIC CHEMICAL focus submetaschema

Once the user gains this knowledge, she might want to see the interaction between semantic types of this collection and other external semantic types. But the number of relationships between semantic types of this collection and other semantic types may be overwhelming. Thus, the user can first view an abstraction of this interaction by viewing the ORGANIC CHEMICAL focus submetaschema where the relationships to and from the various neighboring MSTs of ORGANIC CHEMICAL are shown (Figure 5). If, for example, the user identifies an interest in the interaction between PHENOMENON OR PROCESS and ORGANIC CHEMICAL, she can choose to view the PHENOMENON OR PROCESS/ORGANIC CHEMICAL bi-collection subnetwork (Figure 6). The subnetwork contains all the interactions in the SN between the semantic types of these two collections. If the user wants to learn about all the external relationships of the ORGANIC CHEMICAL collection, then she can view a sequence of bi-collection subnetworks, one for each pair of neighboring MSTs in the focus ORGANIC CHEMICAL submetaschema (Figure 5). In this way, the overwhelming task of reviewing all the relationship interactions of one collection is divided into a

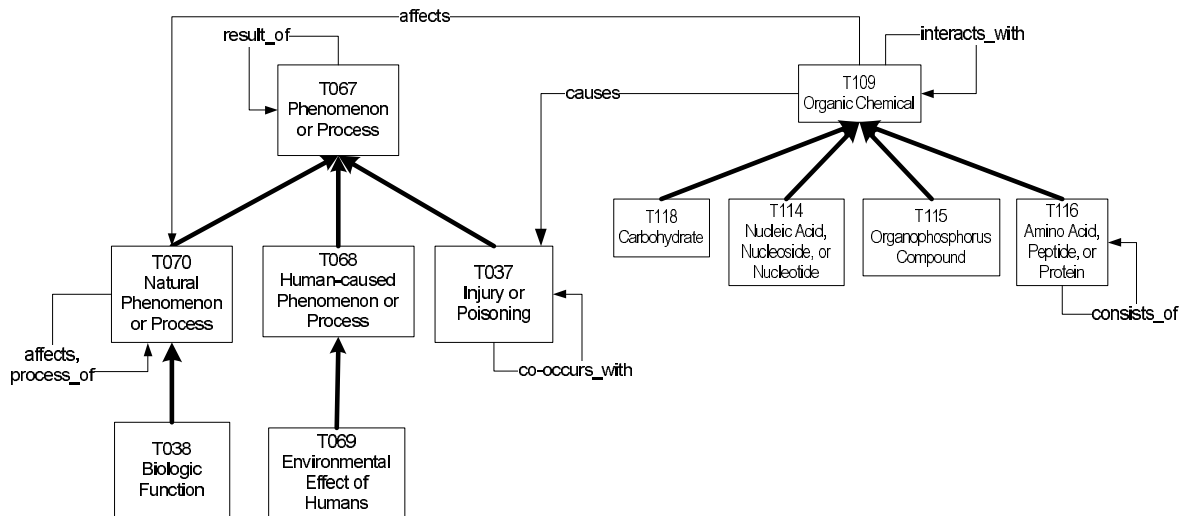


Figure 6: Bi-collection subnetwork of PHENOMENON OR PROCESS/ORGANIC CHEMICAL

sequence of manageable tasks, supporting user comprehension efforts.

Note that we intentionally have picked an MST which was not available in the cohesive metaschema [9]. As is shown in [12], the cohesive metaschema includes instead a larger group named CHEMICAL which contains 16 semantic types. Hence, the lexical metaschema offers the user the new option of concentrating on the ORGANIC CHEMICAL group which is a natural group not earlier available in a metaschema.

2.3.2 Auditing Concept Categorization

The second application uses the metaschema notion for auditing the categorization of concepts in the UMLS, where concepts of the META are assigned to one or more semantic types of the SN. Auditing the META concept categorization is a persistent and overwhelming task for UMLS professionals. There is a need to design auditing techniques for the UMLS which will minimize the effort and maximize the probability of finding errors. Since a concept may be assigned to several semantic types, it may also be associated with several meta-semantic types in a metaschema. As

shown in [10], it is more likely that a concept will be erroneously assigned to several semantic types residing in different meta-semantic types than to several semantic types of the same meta-semantic type. The reason is that, in general, two semantic types of the same meta-semantic type belong to the same domain. While, if two semantic types are in two different meta-semantic types, they belong to two different domains. This observation leads to the idea of an audit that concentrates on concepts which are associated with several meta-semantic types and the effort to review them is limited since their number is not very large.

In [10] we introduced the notions of *intersection of semantic types*, *intersection of meta-semantic types*, and *pure intersection*. An intersection of two or more semantic types is a non-empty set of concepts that are assigned to each of these semantic types and only to them. An intersection of two or more meta-semantic types is a non-empty set of concepts that are associated with each of these meta-semantic types and only with them. A pure intersection of meta-semantic types is a subset of the intersection of the corresponding meta-semantic types, containing all concepts in one intersection of semantic types. The notation for a pure intersection is the list of names of each meta-semantic type followed by its corresponding semantic type in curly brackets, where the intersection symbol \cap appears between any two meta-semantic types in the intersection list (see two examples below). For more elaborate discussion on pure intersection see [10]. Thus, we first identify all concepts of intersections of two or more meta-semantic types. A domain expert reviews each pure intersection containing a small number of concepts of similar semantics to find any categorization errors. This auditing technique is designed to minimize the effort and maximize the likelihood of finding errors.

For example, suppose we use the lexical metaschema to help auditing concept categorization,

the pure intersection $\text{PHARMACOLOGIC SUBSTANCE}\{\mathbf{Pharmacologic Substance}\} \cap \text{ORGANISM}\{\mathbf{Alga}\}$ contains two concepts *Laminaria Tents*² and *Marine Algae*. However, *Laminaria Tents* is a pharmacologic substance produced from a type of marine algae. It should not be assigned to the semantic type **Alga**, while the concept *Marine Algae* should be only assigned to the semantic type **Alga**. Hence, after correcting the errors, there will be no such a pure intersection. As another example, the pure intersection $\text{ACTIVITY}\{\mathbf{Therapeutic or Preventive Procedure}\} \cap \text{PHENOMENON OR PROCESS}\{\mathbf{Natural Phenomenon or Process}\}$ contains only one concept *Diphtheria-Tetanus Vaccine*. However, a vaccine is a pharmacologic substance and immunologic factor. It is neither a procedure nor a process. So it should not be assigned to any one of those two semantic types. This intersection becomes empty. For more details see [10].

3 Methods

In this section, we present the instructions given to the various UMLS experts and describe the derivation of the cumulative metaschemas from their responses. Then, we describe our evaluation techniques used to judge the quality of the lexical metaschema compared to the experts' responses.

3.1 Cumulative Metaschemas Based on Experts' Responses

An important assumption underlying the construction of the lexical metaschema is that even though the lexical partition is the result of an algorithmic process using string matching, it still effectively yields subject areas of the SN similar to those an expert might choose. We conducted a study to evaluate the validity of this assumption. We selected a group of experts with publications in UMLS research or related topics and sent them two pages with diagrams of the SN's IS-A hierarchy, i.e.,

²Concept in META is written in Sans font in the paper.

the two trees rooted at **Event** and **Entity**. Each participant received a page of instructions as follows:

- 1 Start marking by a star the root node of the tree and continue to scan the semantic types downwards.
- 2 While scanning, mark by a star semantic types, which you judge as IMPORTANT AND QUITE DIFFERENT from their parent semantic types.
- 3 There is one exception: Don't mark semantic types which have no children. Thus, you only need to consider the 45 semantic types with children.
- 4 The star markings of each participant can be used to define a metaschema where each semantic type marked by a participant names a meta-semantic-type. The metaschema will be compared with the results of other respondents and with our algorithmically derived metaschema.

The design of the SN is expected to follow the Aristotelian [16] paradigm where categories are specified according to genus and differentiae. The motivation for Step 2 is to partition the SN based on the extent of the difference between the child semantic type and its the parent semantic type.

The instructions heavily utilize the one-to-one correspondence between the semantic-type groups underlying the meta-semantic types, and their root semantic types. By selecting a set of semantic types that are “important and quite different” from their parents, a participating expert induces a partition of the SN and a corresponding metaschema. Each such metaschema is referred to as an “expert metaschema.”

Note that although the instructions seem quite elaborate, they only define structural limitations, such as “don’t mark semantic types which have no children.” These limitations are necessary to make the computation of a valid comparison score between the metaschemas of the participants and the algorithmically obtained lexical metaschema possible. On the other hand, our instructions do not limit the semantic decisions of the participants, who still have the complete freedom to mark semantic types of their choice.

We are interested in quantifying the variability of the experts’ responses. Towards this, we compute the X -by- X agreement matrix (assuming X participating experts) between participants to examine the agreement between any two experts. In the agreement matrix, the number in row i and column j indicates how many meta-semantic types participant i and participant j agree on.

It is to be expected that some choices will be repeated by many participating experts. For our study, we are more interested in metaschemas that represent a kind of aggregation of the experts’ responses rather than the expert metaschemas of the individuals. In particular, we construct a sequence of cumulative metaschemas, each of which reflects a specific level of aggregation of the experts. Suppose we have X experts’ responses. We define a threshold value N in the range $(1, X)$ to represent the level of aggregation. We construct the cumulative metaschema for a given N as follows. For each semantic type marked by at least N participating experts, a meta-semantic type is defined and given the name of the semantic type. Then *meta-child-of*’s and meta-relationships are derived as described in [12]. We denote the cumulative metaschema with the threshold value representing a simple majority [17] of the experts (i.e., $N = \lceil X/2 \rceil$) as the *consensus metaschema*.

3.2 Evaluation Techniques

As noted, our assumption is that our lexical technique can help to capture subject areas of the SN similar to those derived by domain experts. Therefore, we want to evaluate to what degree the lexical metaschema is similar to each expert’s choice, and to what degree the lexical metaschema is similar to each cumulative metaschema. In particular, we want to know how similar the lexical metaschema is to the consensus metaschema representing the simple majority of experts.

We create a gold standard based on the majority vote of the X participating experts (in our study, $X = 11$) on the 45 candidate non-leaf semantic types. To assess the reliability of the gold standard generated by the X experts, we calculate Cronbach’s α [18], which should ideally be greater than or equal to 0.7.

Performance of an expert is calculated using a gold standard composed of the other $X - 1$ (10 in our study) experts. We use majority vote with a random decision for ties. We calculate the agreement between the algorithmic lexical approach and each expert’s choice to show the similarity between the lexical metaschema and each expert metaschema.

Performance of the algorithmic lexical approach is measured in terms of accuracy, sensitivity (recall R), specificity, precision (P), receiver operating characteristic curve trapezoidal area [19], and Rijsbergen’s F measure with equal weighting of recall and precision [20]:

$$F = 2PR/(P + R)$$

Accuracy is the simple proportion of semantic types for which the subject agreed with the gold standard that it either should or should not be in the metaschema. Sensitivity (recall) is the proportion of types in the gold standard metaschema that were also in the subject’s metaschema. It

indicates how good the subject is at detecting types that should be in the metaschema. Specificity is the proportion of types not included in the gold standard metaschema that were also not in the subject's metaschema. It indicates how good the subject is at avoiding types that should not be in the metaschema. Precision is the proportion of types in the subject's metaschema that were also in the gold standard metaschema. It indicates what proportion of the subject's metaschema types are correct. The receiver operating characteristic curve area summarizes the subject's ability to distinguish types that should or should not be in the metaschema. It can be interpreted as the probability that, given a type that belongs in the metaschema and a type that does not belong in the metaschema, the subject will correctly guess which is which. Rijsbergen's F measure also summarizes subject's performance, but as a heuristic combination of recall and precision. It is not directly interpretable as a probability, but it will generally be higher for subjects that perform better.

We compare the performance of our lexical algorithm to the average performance of the experts [21], and we calculate confidence intervals and p-values using bootstrap [22] estimates of variance.

To verify that majority vote rather than another threshold (e.g., 8 out of 11 experts) was a good choice to define the consensus metaschema, we also assess the performance of the algorithm for different values of N . We compute P , R , and Rijsbergen's F measure of the lexical metaschema relative to the corresponding cumulative metaschema, using N as an independent variable. We use the F measure, dependent symmetrically on P and R , as a typical benchmark to evaluate the similarity between the lexical metaschema and the cumulative metaschemas.

Then we compare the lexical metaschema with the consensus metaschema. We not only consider the meta-semantic types' names but also take into account the underlying semantic-type

groups represented by the meta-semantic types. Although the chosen semantic types determined the whole metaschema, we wanted to compare the two metaschemas in more detail. To support this comparison, we need some definitions.

Let M_1 and M_2 be two metaschemas of the SN.

Definition (Identical): A meta-semantic type A in M_1 is *identical* to a meta-semantic type B in M_2 if both meta-semantic types have the same underlying semantic-type group. \square

Definition (Similar): A meta-semantic type A in M_1 is *similar* to a meta-semantic type B in M_2 if the roots of their underlying semantic-type groups are the same. \square

This definition implies that the names of two similar meta-semantic types are equal. To better understand the differences between pairs of similar meta-semantic types, we note that in some cases the difference reflects various levels of granularity in the partition, rather than major disagreements between the metaschemas. A meta-semantic type in one metaschema may be split into several separate meta-semantic types in the other metaschema.

To be formal, we define “refinement” as follows. Let $G_M(A)$ denote the semantic-type group represented by the meta-semantic type A in the metaschema M .

Definition (Refinement): Let A be a meta-semantic type in metaschema M_1 . If there exists a set of meta-semantic types $\{B_1, B_2, \dots, B_k\}$ ($k \geq 2$) in metaschema M_2 such that $G_{M_1}(A) = \cup_{i=1}^k G_{M_2}(B_i)$, then the set $\{B_1, B_2, \dots, B_k\}$ is called a *refinement* of A . \square

We compare the lexical metaschema and the consensus metaschema using the above three terms to measure the closeness between their semantic type coverages.

4 Results

4.1 Experts and Cumulative metaschemas

While studying responses from our eleven UMLS experts, we found that individual participants' responses varied greatly both in the choice of semantic types marked and their numbers. For example, expert 1 chose 21 semantic types to name meta-semantic types in his expert metaschema, while expert 2 chose 34 semantic types as meta-semantic types in his expert metaschema. Table 1 shows the number of meta-semantic types for each expert metaschema, corresponding to the number of semantic types marked by that expert. The average number of meta-semantic types marked by a participant is about 26, with minimum and maximum numbers of 12 and 36, respectively. The standard deviation is 8.10.

Table 1: Number of meta-semantic types in expert metaschemas

Participant	1	2	3	4	5	6	7	8	9	10	11	Average
# Meta-semantic types (Expert)	21	34	21	35	34	35	25	26	12	15	36	26.73

To assess these variations in the responses, we constructed the agreement matrix of all eleven experts and the lexical metaschema (appearing as the 12th column) (Table 2) that demonstrates the agreement as well as the high variability of participant responses. For instance, participants 2 and 5 both marked 34 semantic types and agreed on 27 of them. Similarly, participant 7 and 8 agreed on 14 out of their 25 and 26 marked semantic types, respectively. The average inter-participant agreement is 16.76 (only about 63% of the average number of marked semantic types 26.73), with a high of 30 and a low of six. The large range shows the high variability of participant responses.

We then calculated the agreement between the lexical metaschema and each expert meta-

Table 2: Inter-participant (and Lexical Metaschema) Agreement Matrix

	1	2	3	4	5	6	7	8	9	10	11	12
1		19	15	16	15	19	12	11	11	12	20	13
2			18	28	27	27	20	19	12	14	28	17
3				16	16	17	14	9	10	10	18	13
4					28	26	23	21	8	10	30	15
5						27	20	20	8	10	27	15
6							19	22	10	14	27	16
7								14	8	7	24	12
8									6	9	18	9
9										9	11	7
10											13	12
11												17

schema. The 12th column of Table 2 shows how many semantic types among those marked by this expert (shown in the Table 1) were also chosen by our lexical algorithm. For example, expert 1 marked 21 semantic types, among which 13 also appear as meta-semantic types in the lexical metaschema, since they are roots of semantic-type groups in the lexical partition. The average similarity of the participants with the lexical metaschema shown at the 12th column of Table 2 is 13.27 (about 50% of the average number of marked semantic types 26.73), with a high of 17 and low of seven. The large variation in the choices and numbers of the expert metaschemas’ meta-semantic types raises doubts about the appropriateness of using them to evaluate the lexical metaschema and led us to the consideration of aggregating their responses to obtain the cumulative metaschemas.

In our study, we received responses from eleven experts ($X = 11$) and thus obtained eleven cumulative metaschemas by varying N over the range (1, 11). For $N = 8$, for example, there were 16 semantic types marked by at least eight out of the eleven experts, and so the corresponding cumulative metaschema has 16 meta-semantic types. Table 3 shows the number of semantic types marked for each N . Obviously, the larger the value of N , the smaller the common number of

meta-semantic types.

Table 3: Threshold value N and number of semantic types marked by at least N participants (= # meta-semantic types chosen)

Threshold (N)	1	2	3	4	5	6	7	8	9	10	11
# meta-semantic types	45	45	45	42	36	26	20	16	10	7	2

As we can see from the table, the number of meta-semantic types varies from two (for $N = 11$) to 45 (for $N = 1, 2$, and 3). The corresponding metaschema for the first case contains two meta-semantic types ENTITY and EVENT, each spanning the whole corresponding tree of the SN. For the latter cases, each non-leaf semantic type names a meta-semantic type. The metaschema that emerges in those cases is effectively just the SN itself, without its leaves. No real grouping of related semantic types occurs. Obviously such extreme metaschemas are not interesting. The consensus metaschema ($N = 6$) contains 26 meta-semantic types. Its hierarchy is shown in Figure 7.

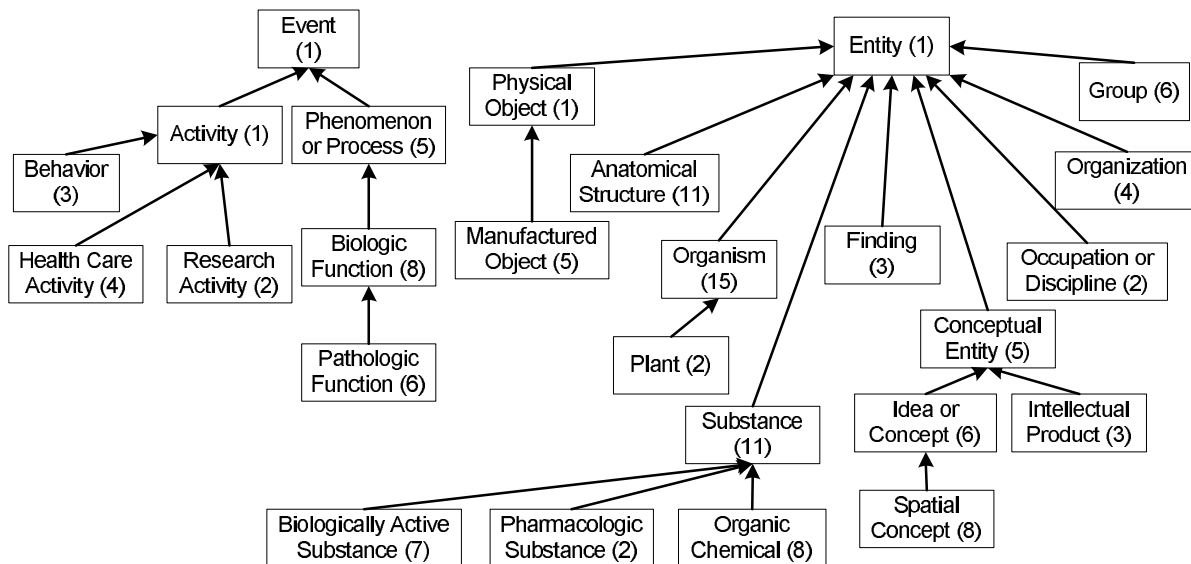


Figure 7: Consensus metaschema hierarchy ($N = 6$)

4.2 Statistical Evaluation Results

Cronbach’s α for the gold standard was 0.62. The performance comparison of the lexical metaschema and of the experts is shown in Table 4, with 95% confidence intervals in parentheses. The values (i.e., accuracy, R , etc.) measure the performance of the lexical metaschema compared to the average performance of the experts. There were no statistically significant differences between the lexical algorithm and the experts. The experiment had sufficient power to detect a difference of 0.15 in ROC area and in the F measure.

Table 4: Performance comparison of lexical algorithm and experts

	Lexical Algorithm	Experts
Accuracy	0.71 (0.53 to 0.84)	0.59 (0.50 to 0.66)
R	0.65 (0.44 to 0.84)	0.66 (0.57 to 0.73)
Specificity	0.79 (0.38 to 0.95)	0.51 (0.44 to 0.56)
P	0.81 (0.53 to 0.95)	0.70 (0.56 to 0.79)
ROC area	0.72 (0.58 to 0.85)	0.59 (0.52 to 0.64)
F measure	0.72 (0.53 to 0.87)	0.65 (0.53 to 0.74)

To verify that the simple majority vote, rather than another threshold, was correctly used in the consensus metaschema evaluation, we assessed the performance of the algorithm for different levels of the threshold. Table 5 shows the results. The second column shows the number of semantic types marked (i.e., number of meta-semantic types chosen) by at least N participants. The third column has the number of semantic types marked by at least N participants that were also identified as roots of groups by the lexical metaschema. For example, the cumulative metaschema with $N = 8$ contains 16 meta-semantic types, among which eleven also appear in the lexical metaschema. Therefore, precision $P = 11/21 = 0.524$, recall $R = 11/16 = 0.688$, and $F = 0.595$.

Table 5: Performance comparison of lexical metaschema for different values of N

Threshold (N)	Marked (B)	Lexical (C)	P= C/21	R = C/B	F=2PR/(P+R)
11	2	2	0.095	1.000	0.174
10	7	5	0.238	0.714	0.357
9	10	8	0.381	0.800	0.516
8	16	11	0.524	0.688	0.595
7	20	13	0.619	0.650	0.634
6	26	17	0.810	0.654	0.723
5	36	20	0.952	0.556	0.702
4	42	21	1.000	0.500	0.667
3	45	21	1.000	0.467	0.636
2	45	21	1.000	0.467	0.636
1	45	21	1.000	0.467	0.636

From the plots in Figure 8, we can see that the larger the value of N , the smaller the number of semantic types marked by at least N experts, and thus the lower the precision value. Also, typically the smaller the value of N , the lower the recall, but not always. An example of an exception appears for $N = 7$. The F measure, reflecting symmetrically both precision and recall, peaks at $N = 6$, with a high precision and a medium recall. This result indicates that the lexical metaschema is most similar to this cumulative metaschema, which, in fact, is actually the consensus metaschema representing a simple majority of the experts. Out of the 26 meta-semantic types in the consensus metaschema, 17 are also in the lexical metaschema with the recall value of 81%, indicating high similarity between the two metaschemas.

4.3 Comparison of the Lexical Metaschema and the Consensus Metaschema

To facilitate the comparison between the lexical and consensus metaschemas, we draw both their hierarchies in Figure 9. Identical meta-semantic types are indicated by black shadows. Similar meta-semantic types are denoted by gray shadows.

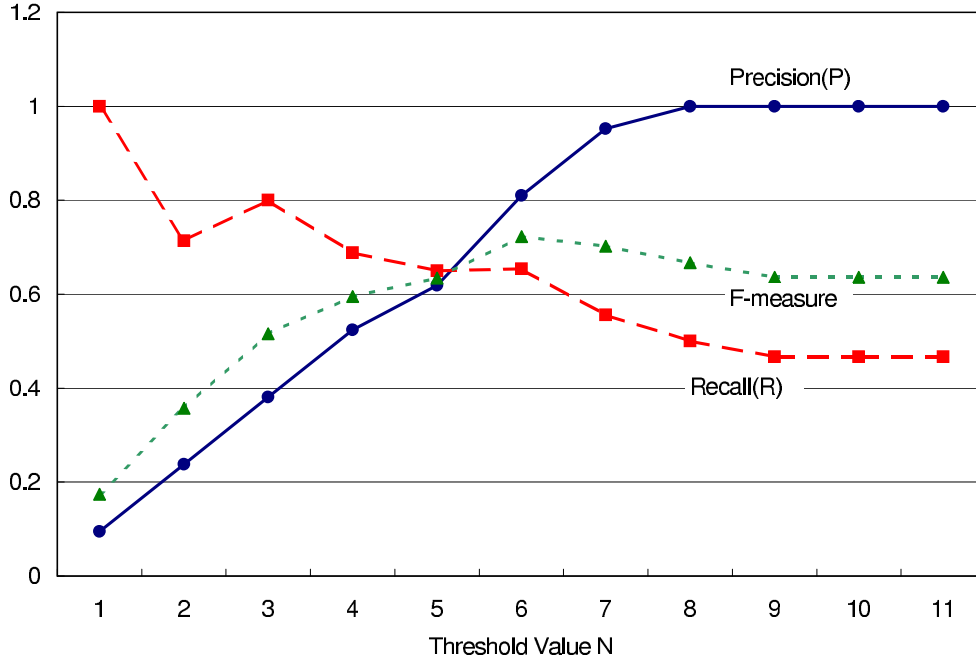


Figure 8: P , R , and F values for different thresholds N

The lexical metaschema contains 21 meta-semantic types, while the consensus metaschema contains 26 meta-semantic types. There are ten identical meta-semantic types between the two metaschemas. For example, FINDING is a meta-semantic type appearing in both metaschemas and representing the same underlying semantic-type group containing three semantic types. Therefore, FINDING in the lexical metaschema is identical to FINDING in the consensus metaschema. Table 6 lists all the ten identical meta-semantic types and their sizes. Altogether, they cover 53 semantic types. That is, both metaschemas agree that these ten meta-semantic types, covering 39.3% of the SN, represent important subject areas in the SN.

There are seven similar meta-semantic types. For example, SPATIAL CONCEPT in the lexical metaschema represents an underlying semantic-type group with four semantic types, while SPATIAL CONCEPT in the consensus metaschema represents a semantic-type group with eight semantic types. Hence, SPATIAL CONCEPT in the lexical metaschema is similar, but not identical, to SPATIAL CONCEPT in the consensus metaschema. Table 7 shows these similar meta-semantic

Table 6: Identical meta-semantic types in lexical and consensus metaschemas

Meta-semantic type	Size
ANATOMICAL STRUCTURE	11
BIOLOGICALLY ACTIVE SUBSTANCE	7
EVENT	1
FINDING	3
IDEA OR CONCEPT	6
OCCUPATION OR DISCIPLINE	2
ORGANIZATION	4
PATHOLOGIC FUNCTION	6
PHARMACOLOGIC SUBSTANCE	2
SUBSTANCE	11

types along with their sizes in each of the two metaschemas. In the lexical metaschema, these seven cover 66 semantic types, which is about 48.9% of the SN. In the consensus metaschema, these seven cover 44 semantic types, which is about 32.6%.

Table 7: Similar meta-semantic types in lexical and consensus metaschemas

Meta-semantic type	Size in lexical metaschema	Size in consensus metaschema
ACTIVITY	15	6
ENTITY	13	1
ORGANIC CHEMICAL	5	8
ORGANISM	17	15
PHENOMENON OR PROCESS	6	5
PHYSICAL OBJECT	6	1
SPATIAL CONCEPT	4	8

As an example for refinement, the meta-semantic type ACTIVITY in the lexical metaschema represents a semantic-type group containing 15 semantic types. These 15 semantic types are split into four semantic-type groups represented by ACTIVITY, BEHAVIOR, HEALTH CARE ACTIVITY, and RESEARCH ACTIVITY in the consensus metaschema. Therefore, {ACTIVITY, BEHAVIOR, HEALTH CARE ACTIVITY, RESEARCH ACTIVITY} in the consensus metaschema is a refinement

of ACTIVITY in the lexical metaschema.

Table 8 shows the cases of refinement from the lexical metaschema to the consensus metaschema. The size of a meta-semantic type is displayed in parentheses following the name. This kind of refinement covers 38 semantic types.

Table 8: Refinements in consensus metaschema

Meta-semantic type in lexical metaschema	Refinement in the consensus metaschema
ACTIVITY (15)	{ACTIVITY (6), BEHAVIOR (3), HEALTH CARE ACTIVITY (4), RESEARCH ACTIVITY (2)}
PHYSICAL OBJECT (6)	{PHYSICAL OBJECT (1), MANUFACTURED OBJECT (5)}
ORGANISM (17)	{ORGANISM (15), PLANT (2)}

There are also some refinements in the other direction from the consensus metaschema to the lexical metaschema. For example, {ORGANIC CHEMICAL, LIPID} in the lexical metaschema is a refinement of ORGANIC CHEMICAL in the consensus metaschema. Table 9 shows all such refinement cases. This kind of refinement covers 16 semantic types. The total number of semantic types covered by refinements in either direction is 54 (about 40%).

Table 9: Refinements in lexical metaschema

Meta-semantic type in consensus metaschema	Refinement in the lexical metaschema
ORGANIC CHEMICAL (8)	{ORGANIC CHEMICAL (5), LIPID (3)}
SPATIAL CONCEPT (8)	{SPATIAL CONCEPT (4), MOLECULAR SEQUENCE (4)}

Besides the identical meta-semantic types, the similar meta-semantic types, and the meta-semantic types appearing in refinements, there are two meta-semantic types that appear exclusively in the lexical metaschema; these are PHYSIOLOGIC FUNCTION and ORGANISM ATTRIBUTE. There are also four meta-semantic types that appear exclusively in the consensus metaschema; these are BIOLOGIC FUNCTION, CONCEPTUAL ENTITY, INTELLECTUAL PRODUCT, and GROUP.

5 Discussion

While the value of 0.62 obtained for Cronbach's α is lower than the target of 0.7 [18], it is not unreasonable. Future studies might benefit from using, say, 15 rather than eleven experts.

Table 4 compared the performance of the lexical metaschema to the average of experts' performance. It shows that while there appears to be a trend of the lexical approach outperforming the experts, none of the differences were statistically significant. One can at least conclude that the algorithmic technique did not grossly underperform compared to the experts.

In the comparison between the lexical metaschema and the consensus metaschema, we note that if there is a refinement case, then there is a meta-semantic type in one metaschema that is similar to one of the meta-semantic types in the refinement. For example, {ORGANIC CHEMICAL, LIPID} in the lexical metaschema is a refinement of ORGANIC CHEMICAL in the consensus metaschema, where the ORGANIC CHEMICAL meta-semantic types in both metaschemas are similar. However, not every case of similar meta-semantic types has a corresponding refinement. For example, ENTITY and PHENOMENON OR PROCESS are both cases of similarity, but they do not have refinements.

If, as in Section 4.2, we consider only the meta-semantic type names and not the underlying semantic-type groups, then 17 out of the 21 meta-semantic types in the lexical metaschema also appear in the consensus metaschema (about 81%). The results in Table 5 show that among all cumulative metaschemas, the lexical metaschema is closest to the consensus metaschema. We note that this is a coincidental result. It is not always expected. Thus, we can conclude that the lexical metaschema can capture subject areas in the SN similar to the ones picked by a simple

majority of the experts. If we consider the semantic coverage as in Section 4.3, the semantic types covered by identical meta-semantic types and refinements together are 107 (about 79%). Both measures show the high similarity between the two metaschemas.

While most of our results bear similarity out, some do not. Consider, for example, the lexically related CP-pair (**Plant, Organism**), in which the definition of **Plant** contains the word “organism.” As such, **Plant** is part of the meta-semantic type ORGANISM in the lexical metaschema. But in the consensus metaschema, PLANT is a separate meta-semantic type, probably due to the difference from the other semantic types in the ORGANISM group. Actually this was a close decision, as six participants marked plant. The same phenomenon appears for **Biologic Function** which was marked by ten of the eleven participants, but is lexically related to its parent **Natural Phenomenon or Process**.

On the other hand, the opposite happens with **Physiologic Function** which is not lexically related to **Biologic Function** and is the root of a lexical meta-semantic type, but was marked only by five experts and thus is not in the consensus metaschema. **Molecular Sequence** is another example of a lexically independent semantic type which was marked by only five experts.

One may be surprised by the extent of the variations between experts. However, one has to bear in mind that this is not similar to experts checking if a patient has a specific disease. The task of partitioning the SN is basically a modeling task where experts are asked to judge the importance and closeness between subject areas. In such issues, experts tend to differ according to their expertise, education, and personal preferences.

To illustrate the differences between experts' opinions, consider the marking of the semantic types **Chemical**, **Chemical Viewed Structurally**, and **Chemical Viewed Functionally**. Four

experts marked only **Chemical**, three others marked both **Chemical Viewed Structurally** and **Chemical Viewed Functionally**, while four other experts did not mark any of the three semantic types. That is, the eleven experts were divided almost equally between three incompatible opinions. As a result, no opinion obtained a simple majority and none of the three semantic types appears in the consensus metaschema. One may say that the majority (seven experts) were of the opinion that the metaschema should reflect some representation of these three semantic types, but this opinion is not reflected in the result of the study, due to the differences of experts' opinions about the details.

Furthermore, in modeling, experts tend to be divided between splitters and lumpers. This phenomenon manifests itself in the variety of the numbers of semantic types marked. Considering a cumulative metaschema, the threshold picked influences the numbers of meta-semantic types in the metaschema. Hence, there is no guidance on which threshold number to choose.

Considering all these issues, one can see the advantage of using an algorithmic approach for partitioning the SN. Here we used the lexical partitioning technique and in [9, 11] the cohesive partitioning technique. These techniques apply rules according to data given for the SN, either lexical data or structural data. Although the partitions and metaschemas obtained suffer from problems as well, they tend to be objective, following the data rather than differing backgrounds of experts. The expert consensus metaschema could be used to evaluate the algorithmically obtained metaschema, as we do here. In our future research, we will consider techniques for utilizing experts' consensus metaschema to guide the improvement of algorithmically obtained metaschemas, modifying them to be more in line with the experts' consensus metaschema.

6 Conclusions

In this paper, we evaluated the lexical metaschema derived via an algorithmic lexical partitioning approach. We also constructed cumulative metaschemas as aggregations of the opinions of eleven UMLS experts participating in an evaluation study we conducted. Of particular interest is the consensus metaschema, representing a simple majority aggregation of the experts' opinions. We used the cumulative metaschemas to evaluate the lexical metaschema. From our evaluation, we can conclude that the results of the lexical algorithmic approach are sufficiently similar to the consensus metaschema, within the limits of the experiment, to warrant further investigation.

Acknowledgement

We thank the following researchers who published on UMLS or related topics for participating in the reported study. The list is in Alphabetical order. We apologize for possible inaccuracies caused by some respondents sending their responses in unnamed self-addressed envelopes:

Dr. G. Octo Barnett, Dr. Charles Barr, Dr. James Brinkley, Dr. Christopher G. Chute, Dr. Gregory F. Cooper, Dr. Gai Elhanan, Dr. Mark Erlbaum, Dr. Robert A. Greenes, Dr. Kenric Hammond, Dr. William R. Hersh, Dr. Perry L. Miller and Dr. Cornelius Rosse.

References

- [1] McCray AT. UMLS Semantic Network. In *Proc. Thirteenth Annual SCAMC*, pages 503–507, Washington, DC, 1989.
- [2] McCray AT and Hole WT. The scope and structure of the first version of the UMLS Semantic Network. In *Proc. Fourteenth Annual SCAMC*, pages 126–130, Los Alamitos, CA, November 1990.

- [3] McCray AT. Representing biomedical knowledge in the UMLS Semantic Network. In Broering NC, editor, *High-Performance Medical Libraries: Advances in Information Management for the Virtual Era*, pages 45–55, Mekler, Westport, CT, 1993.
- [4] Campbell KE, Oliver DE, and Shortliffe EH. The Unified Medical Language System: Toward a collaborative approach for solving terminologic problems. *JAMIA*, 5(1):12–16, 1998.
- [5] Humphreys BL, Lindberg DAB, Schoolman HM, and Barnett GO. The Unified Medical Language System: An informatics research collaboration. *JAMIA*, 5(1):1–11, 1998.
- [6] Tuttle MS, Sherertz DD, Olson NE, Erlbaum MS, Sperzel WD, and Fuller LF et al. Using META-1, the first version of the UMLS Metathesaurus. In *Proc. Fourteenth Annual SCAMC*, pages 131–135, 1990.
- [7] Schuyler PL, Hole WT, Tuttle MS, and Sherertz DD. The UMLS Metathesaurus: Representing different views of biomedical concepts. *Bull Med Libr Assoc*, 81(2):217–222, 1993.
- [8] U. S. Dept. of Health and Human Services, National Institutes of Health, National Library of Medicine. Unified Medical Language System (UMLS), 2003.
- [9] Perl Y, Chen Z, Halper M, Geller J, Zhang L, and Peng Y. The cohesive metaschema: a higher-level abstraction of the UMLS Semantic Network. *Journal of Biomedical Informatics*, 35(3):194 – 212, 2003.
- [10] Gu H, Perl Y, Ethan G, Geller J, Zhang L, and Peng Y. Auditing concept categorizations in the UMLS. *Artificial Intelligence in Medicine*, 31(1):29–44, 2004.
- [11] Chen Z, Perl Y, Halper M, Geller J, and Gu H. Partitioning the UMLS Semantic Network. *IEEE Trans. Information Technology in Biomedicine*, 6(2):102–108, June 2002.
- [12] Zhang L, Perl Y, Halper M, Geller J, and Hripcsak G. A lexical metaschema for the UMLS Semantic Network. *Artificial Intelligence in Medicine*, 33(1):41–59, 2005.
- [13] Zhang L, Perl Y, Halper M, and Geller J. Designing metaschemas for the UMLS Enriched Semantic Network. *Journal of Biomedical Informatics*, 36(6):433–449, Dec 2003.
- [14] Zhang L, Perl Y, Geller J, Halper M, and Cimino JJ. An enriched UMLS Semantic Network with a multiple inheritance hierarchy. *JAMIA*, 11(3):195–206, 2004.
- [15] McCray AT, Burgun A, and Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. In *Proc. Medinfo 2001*, pages 171–175, London, UK, September 2001.
- [16] Aristotle. *The categories*. Cambridge, MA: Harvard University Press, 1973.
- [17] Hripcsak G, Friedman C, Anderson P, DuMouchel W, Johnson S, and Clayton P. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med*, 122:681–688, 1995.

- [18] Dunn G. *Design and Analysis of Reliability Studies*. New York: Oxford University Press, 1989.
- [19] Hanley JA and McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.
- [20] Van Rijsbergen CJ. *Information Retrieval*. London: Butterworth, 1979.
- [21] Hripcsak G and Wilcox A. Reference standards, judges, comparison subjects: roles for experts in evaluating system performance. *JAMIA*, 9:1–15, 2002.
- [22] Efron B and Tibshirani RJ. *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1993.