# The Descent of Hierarchy, and Selection in Relational Semantics*

**Barbara Rosario**
**SIMS**
**UC Berkeley**
**Berkeley, CA 94720**
*rosario@sims.berkeley.edu*

**Marti A. Hearst**
**SIMS**
**UC Berkeley**
**Berkeley, CA 94720**
*hearst@sims.berkeley.edu*

**Charles Fillmore**
**ICSI**
**UC Berkeley**
**Berkeley, CA 94720**
*fillmore@icsi.berkeley.edu*

## Abstract

In many types of technical texts, meaning is embedded in noun compounds. A language understanding program needs to be able to interpret these in order to ascertain sentence meaning. We explore the possibility of using an existing lexical hierarchy for the purpose of placing words from a noun compound into categories, and then using this category membership to determine the relation that holds between the nouns. In this paper we present the results of an analysis of this method on two-word noun compounds from the biomedical domain, obtaining classification accuracy of approximately 90%. Since lexical hierarchies are not necessarily ideally suited for this task, we also pose the question: how far down the hierarchy must the algorithm descend before all the terms within the subhierarchy behave uniformly with respect to the semantic relation in question? We find that the topmost levels of the hierarchy yield an accurate classification, thus providing an economic way of assigning relations to noun compounds.

## 1 Introduction

A major difficulty for the interpretation of sentences from technical texts is the complex structure of noun phrases and noun compounds. Consider, for example, this title, taken from a biomedical journal abstract:

> *Open-labeled long-term study of the subcutaneous sumatriptan efficacy and tolerability in acute migraine treatment.*

An important step towards being able to interpret such technical sentences is to analyze the meaning of noun compounds, and noun phrases more generally.

*With apologies to Charles Darwin.

Interpretation of noun compounds (NCs) is highly dependent on lexical information. Thus we explore the use of a large corpus (Medline) and a large lexical hierarchy (MeSH, Medical Subject Headings) to determine the relations that hold between the words in noun compounds.

Surprisingly, we find that we can simply use the juxtaposition of category membership within the lexical hierarchy to determine the relation that holds between pairs of nouns. For example, for the NCs *leg paresis, skin numbness,* and *hip pain*, the first word of the NC falls into the MeSH A01 (Body Regions) category, and the second word falls into the C10 (Nervous System Diseases) category. From these we can declare that the relation that holds between the words is "located in". Similarly, for *influenza patients* and *aids survivors*, the first word falls under C02 (Virus Diseases) and the second is found in M01.643 (Patients), yielding the "afflicted by" relation. Using this technique on a subpart of the category space, we obtain 90% accuracy overall.

In some sense, this is a very old idea, dating back to the early days of semantic nets and semantic grammars. The critical difference now is that large lexical resources and corpora have become available, thus allowing some of those old techniques to become feasible in terms of coverage. However, the success of such an approach depends on the structure and coverage of the underlying lexical ontology.

In the following sections we discuss the linguistic motivations behind this approach, the characteristics of the lexical ontology MeSH, the use of a corpus to examine the problem space, the method of determining the relations, the accuracy of the results, and the problem of ambiguity. The paper concludes with related work and a discussion of future work.

## 2 Linguistic Motivation

One way to understand the relations between the words in a two-word noun compound is to cast the words into a head-modifier relationship, and assume that the head noun has an argument structure, much the way verbs do, as well as a qualia structure in the sense of Pustejovsky (1995). Then the meaning of the head noun determines what kinds of things can be done to it, what it is made of, what it is a part of, and so on.

For example, consider the noun *knife*. Knives are created for particular activities or settings, can be made of various materials, and can be used for cutting or manipulating various kinds of things. A set of relations for knives, and example NCs exhibiting these relations is shown below:

*(Used-in): kitchen knife, hunting knife*
*(Made-of): steel knife, plastic knife*
*(Instrument-for): carving knife*
*(Used-on): meat knife, putty knife*
*(Used-by): chef's knife, butcher's knife*

Some relationships apply to only certain classes of nouns; the semantic structure of the head noun determines the range of possibilities. Thus if we can capture regularities about the behaviors of the constituent nouns, we should also be able to predict which relations will hold between them.

We propose using the categorization provided by a lexical hierarchy for this purpose. Using a large collection of noun compounds, we assign semantic descriptors from the lexical hierarchy to the constituent nouns and determine the relations between them. This approach avoids the need to enumerate in advance all of the relations that may hold. Rather, the corpus determines which relations occur.

## 3 The Lexical Hierarchy: MeSH

MeSH (Medical Subject Headings)[1] is the National Library of Medicine's controlled vocabulary thesaurus; it consists of set of terms arranged in a hierarchical structure. There are 15 main sub-hierarchies (trees) in MeSH, each corresponding to a major branch of medical terminology. For example, tree A corresponds to Anatomy, tree B to Organisms, tree C to Diseases and so on. Every branch has several sub-branches; Anatomy, for example, consists of Body Regions (A01), Musculoskeletal System (A02), Digestive System (A03) etc. We refer to these as "level 0" categories.

These nodes have children, for example, Abdomen (A01.047) and Back (A01.176) are level 1 children of Body Regions. The longer the ID of the MeSH term, the longer the path from the root and the more precise the description. For example migraine is C10.228.140.546.800.525, that is, C (a disease), C10 (Nervous System Diseases), C10.228 (Central Nervous System Diseases) and so on. There are over 35,000 unique IDs in MeSH 2001. Many words are assigned more than one MeSH ID and so occur in more than one location within the hierarchy; thus the structure of MeSH can be interpreted as a network.

Some of the categories are more homogeneous than others. The tree A (Anatomy) for example, seems to be quite homogeneous; at level 0, the nodes are all *part of* (meronymic to) Anatomy: the Digestive (A03), Respiratory (A04) and the Urogenital (A05) Systems are all part of anatomy; at level 1, the Biliary Tract (A03.159) and the Esophagus (A03.365) are part of the Digestive System (level 0) and so on. Thus we assume that every node is a (body) part of the parent node (and all the nodes above it).

Tree C for Diseases is also homogeneous; the child nodes are a *kind of* (hyponym of) the disease at the parent node: Neoplasms (C04) is a *kind of* Disease C and Hamartoma (C04.445) is a *kind of* Neoplasms.

Other trees are more heterogeneous, in the sense that the meanings among the nodes are more diverse. Information Science (L01), for example, contains, among others, Communications Media (L01.178), Computer Security (L01.209) and Pattern Recognition (L01.725). Another heterogeneous sub-hierarchy is Natural Science (H01). Among the children of H01 we find Chemistry (parent of Biochemistry), Electronics (parent of Amplifiers and Robotics), Mathematics (Fractals, Game Theory and Fourier Analysis). In other words, we find a wide range of concepts that are not described by a simple relationship.

These observations suggest that once an algorithm descends to a homogeneous level, words falling into the subhierarchy at that level (and below it) behave similarly with respect to relation assignment.

## 4 Counting Noun Compounds

In this and the next section, we describe how we investigated the hypothesis:

> For all two-word noun compounds (NCs) that can be characterized by a category pair (CP), a particular semantic relationship holds between the nouns comprising those NCs.

The kinds of relations we found are similar to those described in Section 2. Note that, in this analysis we focused on determining which sets of NCs fall into the same relation, without explicitly assigning names to the relations themselves. Furthermore, the same relation may be described by many different category pairs (see Section 5.5).

First, we extracted two-word noun compounds from approximately 1M titles and abstracts from the Medline collection of biomedical journal articles, resulting

[1] http://www.nlm.nih.gov/mesh/meshhome.html; the work reported in this paper uses MeSH 2001.

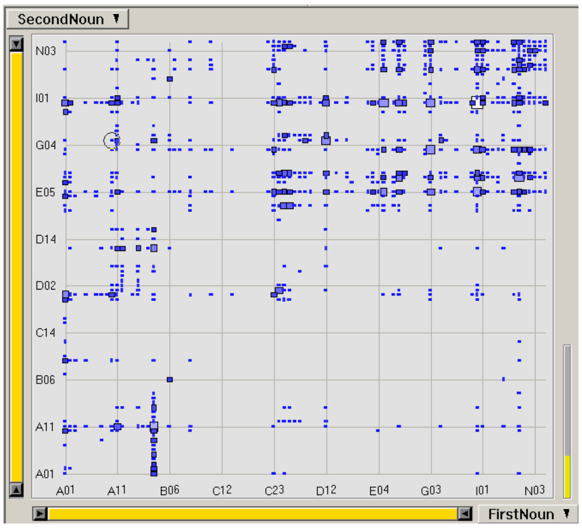


Figure 1: Distribution of Level 0 Category Pairs. Mark size indicates the number of unique NCs that fall under the CP. Only those for which $>$ 50 NCs occur are shown.

in about 1M NCs. The NCs were extracted by finding adjacent word pairs in which both words are tagged as nouns by a tagger and appear in the MeSH hierarchy, and the words preceding and following the pair do not appear in MeSH[2] Of these two-word noun compounds, 79,677 were unique.

Next we used MeSH to characterize the NCs according to semantic category(ies). For example, the NC *fibroblast growth* was categorized into A11.329.228 (Fibroblasts) and G07.553.481 (Growth).

Note that the same words can be represented at different levels of description. For example, *fibroblast growth* can be described by the MeSH descriptors A11.329.228 G07.553.481 (original level), but also by A11 G07 (Cell and Physiological Processes) or A11.329 G07.553 (Connective Tissue Cells and Growth and Embryonic Development). If a noun fell under more than one MeSH ID, we made multiple versions of this categorization. We refer to the result of this renaming as a category pair (CP).

We placed these CPs into a two-dimensional table, with the MeSH category for the first noun on the X axis, and the MeSH category for the second noun on the Y axis. Each intersection indicates the number of NCs that are classified under the corresponding two MeSH categories.

A visualization tool (Ahlberg and Shneiderman, 1994) allowed us to explore the dataset to see which areas of the category space are most heavily populated, and to get a feeling for whether the distribution is uniform or not (see Figure 1). If our hypothesis holds (that NCs that fall within the same category pairs are assigned the same relation), then if most of the NCs fall within only a few category pairs then we only need to determine which relations hold between a subset of the possible pairs. Thus, the more clumped the distribution, the easier (potentially) our task is. Figure 1 shows that some areas in the CP space have a higher concentration of unique NCs (the Anatomy, and the E through N sub-hierarchies, for example), especially when we focus on those for which at least 50 unique NCs are found.

[2]Clearly, this simple approach results in some erroneous extractions.

## 5 Labeling NC Relations

Given the promising nature of the NC distributions, the question remains as to whether or not the hypothesis holds. To answer this, we examined a subset of the CPs to see if we could find positions within the sub-hierarchies for which the relation assignments for the member NCs are always the same.

### 5.1 Method

We first selected a subset of the CPs to examine in detail. For each of these we examined, by hand, 20% of the NCs they cover, paraphrasing the relation between the nouns, and seeing if that paraphrase was the same for all the NCs in the group. If it was the same, then the current levels of the CP were considered to be the correct levels of description. If, on the other hand, several different paraphrases were found, then the analysis descended one level of the hierarchy. This repeated until the resulting partition of the NCs resulted in uniform relation assignments.

For example, all the following NCs were mapped to the same CP, A01 (Body Regions) and A07 (Cardiovascular System): *scalp arteries, heel capillary, shoulder artery, ankle artery, leg veins, limb vein, forearm arteries, finger capillary, eyelid capillary, forearm microcirculation, hand vein, forearm veins, limb arteries, thigh vein, foot vein*. All these NCs are "similar" in the sense that the relationships between the two words are the same; therefore, we do not need to descend either hierarchy. We call the pair (A01, A07) a "rule", where a rule is a CP for which all the NCs under it have the same relationship. In the future, when we see an NC mapped to this rule, we will assign this semantic relationship to it.

On the other hand, the following NCs, having the CP A01 (Body Regions) and M01 (Persons), do not have the same relationship between the component words: *abdomen patients, arm amputees, chest physicians, eye patients, skin donor*. The relationships are different depending on whether the person is a patient, a physician or a donor. We therefore descend the M01 sub-hierarchy, obtaining the following clusters of NCs:

A01 M01.643 (Patients): *abdomen patients, ankle inpatient, eye outpatient*

A01 H01 (Natural Sciences):
A01 H01 *abdomen x-ray, ankle motion*
  A01 H01.770 (Science): *skin observation*
  A01 H01.548 (Mathematics): *breast risk*
  A01 H01.939 (Weights and Measures): *head calibration*
  A01 H01.181 (Chemistry): *skin iontophoresis*
  A01 H01.671 (Physics)
    A01 H01.671.538 (Motion): *shoulder rotations*
    A01 H01.671.100 (Biophysics): *shoulder biomechanics*
    A01 H01.671.691 (Pressure): *eye pressures*
    A01 H01.671.868 (Temp.): *forehead temperature*
    A01 H01.671.768 (Radiation): *thorax x-ray*
    A01 H01.671.252 (Electricity): *chest electrode*
    A01 H01.671.606 (Optics): *skin color*

Figure 2: Levels of descent needed for NCs classified under A01 H01.

A01 M01.526 (Occupational Groups): *chest physician, eye nurse, eye physician*
A01, M01.898 (Donors): *eye donor, skin donor*
A01, M01.150 (Disabled Persons): *arm amputees, knee amputees.*

In other words, to correctly assign a relationship to these NCs, we needed to descend one level for the second word. The resulting rules in this case are (A01 M01.643), (A01, M01.150) etc. Figure 2 shows one CP for which we needed to descend 3 levels.

In our collection, a total of 2627 CPs at level 0 have at least 10 unique NCs. Of these, 798 (30%) are classified with A (Anatomy) for either the first or the second noun. We randomly selected 250 of such CPs for analysis.

We also analyzed 21 of the 90 CPs for which the second noun was H01 (Natural Sciences); we decided to analyze this portion of the MeSH hierarchy because the NCs with H01 as second noun are frequent in our collection, and because we wanted to test the hypothesis that we do indeed need to descend farther for heterogeneous parts of MeSH.

Finally, we analyzed three CPs in category C (Diseases); the most frequent CP in terms of the total number of non-unique NCs is C04 (Neoplasms) A11 (Cells), with 30606 NCs; the second CP was A10 C04 (27520 total NCs) and the fifth most frequent, A01 C04, with 20617 total NCs; we analyzed these CPs.

We started with the CPs at level 0 for both words, descending when the corresponding clusters of NCs were not homogeneous and stopping when they were. We did this for 20% of the NCs in each CP. The results were as follows.

For 187 of 250 (74%) CPs with a noun in the Anatomy category, the classification remained at level 0 for both words (for example, A01 A07). For 55 (22%) of the CPs we had to descend 1 level (e.g., A01 M01: A01 M01.898, A01 M01.643) and for 7 CPs (2%) we descended two levels. We descended one level most of the time for the sub-hierarchies E (Analytical, Diagnostic and Therapeutic Techniques), G (Biological Sciences) and N (Health Care) (around 50% of the time for these categories combined). We never descended for B (Organisms) and did so only for A13 (Animal Structures) in A. This was to be able to distinguish a few non-homogeneous subcategories (e.g., milk appearing among body parts, thus forcing a distinction between *buffalo milk* and *cat forelimb*).

For CPs with H01 as the second noun, of the 21 CPs analyzed, we observed the following (level number, count) pairs: (0, 1) (1, 8) (2, 12).

In all but three cases, the descending was done for the second noun only. This may be because the second noun usually plays the role of the head noun in two-word noun compounds in English, thus requiring more specificity. Alternatively, it may reflect the fact that for the examples we have examined so far, the more heterogeneous terms dominate the second noun. Further examination is needed to answer this decisively.

### 5.2 Accuracy

We tested the resulting classifications by developing a randomly chosen test set (20% of the NCs for each CP), entirely distinct from the labeled set, and used the classifications (rules) found above to automatically predict which relations should be assigned to the member NCs. An independent evaluator with biomedical training checked these results manually, and found high accuracies: For the CPs which contained a noun in the Anatomy domain, the assignments of new NCs were 94.2% accurate computed via intra-category averaging, and 91.3% accurate with extra-category averaging. For the CPs in the Natural Sciences (H01) we found 81.6% accuracy via intra-category averaging, and 78.6% accuracy with extra-category averaging. For the three CPs in the C04 category we obtained 100% accuracy.

The total accuracy across the portions of the A, H01 and C04 hierarchies that we analyzed were 89.6% via intra-category averaging, and 90.8% via extra-category averaging.

The lower accuracy for the Natural Sciences category illustrates the dependence of the results on the properties of the lexical hierarchy. We can generalize well if the sub-hierarchies are in a well-defined semantic relation with their ancestors. If they are a list of "unrelated" topics, we cannot use the generalization of the higher levels; most of the mistakes for the Natural Sciences CPs occurred in fact when we failed to descend for broad terms such as Physics. Performing this evaluation allowed us to find such problems and update the rules; the resulting categorization should now be more accurate.

### 5.3 Generalization

An important issue is whether this method is an economic way of classifying the NCs. The advantage of the high level description is, of course, that we need to assign by hand many fewer relationships than if we used all CPs at their most specific levels. Our approach provides generalization over the "training" examples in two ways. First, we find that we can use the juxtaposition of categories in a lexical hierarchy to identify semantic relationships. Second, we find we can use the higher levels of these categories for the assignments of these relationships.

To assess the degree of this generalization we calculated how many CPs are accounted for by the classification rules created above for the Anatomy categories. In other words, if we know that A01 A07 unequivocally determines a relationship, how many possible (i.e., present in our collection) CPs are there that are "covered by" A01 A07 and that we do not need to consider explicitly? It turns out that our 415 classification rules cover 46001 possible CP pairs[3].

This, and the fact that we achieve high accuracies with these classification rules, show that we successfully use MeSH to generalize over unique NCs.

### 5.4 Ambiguity

A common problem for NLP tasks is ambiguity. In this work we observe two kinds: lexical and "relationship" ambiguity. As an example of the former, *mortality* can refer to the state of being mortal or to death rate. As an example of the latter, *bacteria mortality* can either mean "death of bacteria" or "death caused by bacteria".

In some cases, the relationship assignment method described here can help disambiguate the meaning of an ambiguous lexical item. *Milk* for example, can be both Animal Structures (A13) and Food and Beverages (J02). Consider the NCs *chocolate milk, coconut milk* that fall under the CPs (B06 -Plants-, J02) and (B06, A13). The CP (B06, J02) contains 180 NCs (other examples are *berry wines, cocoa beverages*) while (B06, A13) has only 6 NCs (4 of which with *milk*). Assuming then that (B06, A13) is "wrong", we will assign only (B06, J02) to *chocolate milk, coconut milk*, therefore disambiguating the sense for milk in this context (Beverage). Analogously, for *buffalo milk, caprine milk* we also have two CPs (B02, J02) (B02, A13). In this case, however, it is easy to show that only (B02 -Vertebrates-, A13) is the correct one (i.e. yielding the correct relationship) and we then assign the MeSH sense A13 to *milk*.

Nevertheless, ambiguity may be a problem for this method. We see five different cases:

[3] Although we began with 250 CPs in the A category, when a descend operation is performed, the CP is split into two or more CPs at the level below. Thus the total number of CPs after all assignments are made was 415.

1) Single MeSH senses for the nouns in the NC (no lexical ambiguity) and only one possible relationship which can predicted by the CP; that is, no ambiguity. For instance, in *abdomen radiography*, *abdomen* is classified exclusively under Body Regions and *radiography* exclusively under Diagnosis, and the relationship between them is unambiguous. Other examples include *aciclovir treatment* (Heterocyclic Compounds, Therapeutics) and *adenocarcinoma treatment* (Neoplasms, Therapeutics).

2) Single MeSH senses (no lexical ambiguity) but multiple readings for the relationships that therefore cannot be predicted by the CP. It was quite difficult to find examples of this case; disambiguating this kind of NC requires looking at the context of use. The examples we did find include *hospital databases* which can be *databases* **regarding** (topic) *hospitals*, *databases* **found in** (location) or **owned by** hospitals. *Education efforts* can be *efforts* **done through** (*education*) or **done to achieve** *education*. *Kidney metabolism* can be *metabolism* **happening in** (location) or **done by** the *kidney*. *Immunoglobulin staining*, (D12 -Amino Acids, Peptides-, and Proteins, E05 -Investigative Techniques-) can mean either *staining* **with** *immunoglobulin* or *staining* **of** *immunoglobulin*.

3) Multiple MeSH mappings but only one possible relation. One example of this case is *alcoholism treatment* where *treatment* is Therapeutics (E02) and *alcoholism* is both Disorders of Environmental Origin (C21) and Mental Disorders (F03). For this NC we have therefore 2 CPs: (C21, E02) as in *wound treatments, injury rehabilitation* and (F03, E02) as in *delirium treatment, schizophrenia therapeutics*. The multiple mappings reflect the conflicting views on how to classify the condition of alcoholism, but the relationship does not change.

4) Multiple MeSH mappings and multiple relations that *can* be predicted by the different CPs. For example, *Bread diet* can mean either that a person usually eats *bread* or that a physician prescribed *bread* to treat a condition. This difference is reflected by the different mappings: *diet* is both Investigative Techniques (E05) and Metabolism and Nutrition (G06), *bread* is Food and Beverages (J02). In these cases, the category can help disambiguate the relation (as opposed to in case 5 below); word sense disambiguation algorithms that use context may be helpful.

5) Multiple MeSH mappings and multiple relations that *cannot* be predicted by the different CPs. As an example of this case, *bacteria mortality* can be both "death of bacteria" or "death caused by bacteria". The multiple mapping for *mortality* (Public Health, Information Science, Population Characteristics and Investigative Techniques) does not account for this ambiguity. Similarly, for *inhibin immunization*, the first noun falls under Hormones and Amino Acids, while *immunization* falls under

Environment and Public Health and Investigative Techniques. The meanings are *immunization* **against** *inhibin* or *immunization* **using** *inhibin*, and they cannot be disambiguated using only the MeSH descriptors.

We currently do not have a way to determine how many instances of each case occur. Cases 2 and 5 are the most problematic; however, as it was quite difficult to find examples for these cases, we suspect they are relatively rare.

A question arises as to if representing nouns using the topmost levels of the hierarchy causes a loss in information about lexical ambiguity. In effect, when we represent the terms at higher levels, we assume that words that have multiple descriptors under the same level are very similar, and that retaining the distinction would not be useful for most computational tasks. For example, *osteosarcoma* occurs twice in MeSH, as C04.557.450.565.575.650 and C04.557.450.795.620. When described at level 0, both descriptors reduce to C04, at level 1 to C04.557, removing the ambiguity. By contrast, *microscopy* also occurs twice, but under E05.595 and H01.671.606.624. Reducing these descriptors to level 0 retains the two distinct senses.

To determine how often different senses are grouped together, we calculated the number of MeSH senses for words at different levels of the hierarchy. Table 1 shows a histogram of the number of senses for the first noun of all the unique NCs in our collection, the average degree of ambiguity and the average description lengths.[4] The average number of MeSH senses is always less than two, and increases with length of description, as is to be expected.

We observe that 3.6% of the lexical ambiguity is at levels higher that 2, 16% at L2, 21.4% at L1 and 59% at L0. Level 1 and 2 combined account for more than 80% of the lexical ambiguity. This means that when a noun has multiple senses, those senses are more likely to come from different main subtrees of MeSH (A and B, for example), than from different deeper nodes in the same subtree (H01.671.538 vs. H01.671.252). This fits nicely with our method of describing the NCs with the higher levels of the hierarchy: if most of the ambiguity is at the highest levels (as these results show), information about lexical ambiguity is not lost when we describe the NCs using the higher levels of MeSH. Ideally, however, we would like to *reduce* the lexical ambiguity for similar senses and to *retain* it when the senses are semantically distinct (like, for example, for *diet* in case 4). In other words, ideally, the ambiguity left at the levels of our rules accounts for only (and for all) the semantically different senses. Further analysis is needed, but the high accuracy we obtained in the classification seems to indicate that this indeed is what is happening.

[4]We obtained very similar results for the second noun.

| # Senses | Original | L2 | L1 | L0 |
|---|---|---|---|---|
| 1 (Unambiguous) | 51539 | 51766 | 54087 | 58763 |
| 2 | 18637 | 18611 | 18677 | 17373 |
| 3 | 5719 | 5816 | 4572 | 2177 |
| 4 | 2222 | 2048 | 1724 | 1075 |
| 5 | 831 | 827 | 418 | 289 |
| 6 | 223 | 262 | 167 | 0 |
| 7 | 384 | 254 | 32 | 0 |
| 8 | 2 | 2 | 0 | 0 |
| 9 | 61 | 91 | 0 | 0 |
| 10 | 59 | 0 | 0 | 0 |
| Total(Ambiguous) | 28138 | 27911 | 25590 | 20914 |
| Avg # Senses | 1.56 | 1.54 | 1.45 | 1.33 |
| Avg Desc Len | 3.71 | 2.79 | 1.97 | 1 |

Table 1: The number of MeSH senses for N1 when truncated to different levels of MeSH. Original refers to the actual (non-truncated) MeSH descriptor. Avg # Senses is the average number of senses computed for all first nouns in the collection. Avg Desc Len is the average description length; the value for level 1 is less than 2 and for level 2 is less that 3, because some nouns are always mapped to higher levels (for example, *cell* is always mapped to A11).

### 5.5 Multiple Occurrences of Semantic Relations

Because we determine the possible relations in a data-driven manner, the question arises of how often does the same semantic relation occur for different category pairs. To determine the answer, we could (i) look at all the CPs, give a name to the relations and "merge" the CPs that have the same relationships; or (ii) draw a sample of NC examples for a given relation, look at the CPs for those examples and verify that all the NCs for those CPs are indeed in the same relationship.

We may not be able to determine the total number of relations, or how often they repeat across different CPs, until we examine the full spectrum of CPs. However, we did a preliminary analysis to attempt to find relation repetition across category pairs. As one example, we hypothesized a relation **afflicted by** and verified that it applies to all the CPs of the form (Disease C, Patients M01.643), e.g.: *anorexia (C23) patients, cancer (C04) survivor, influenza (C02) patients*. This relation also applies to some of the F category (Psychiatry), as in *delirium (F03) patients, anxiety (F01) patient*.

It becomes a judgement call whether to also include NCs such as *eye (A01) patient, gallbladder (A03) patients*, and more generally, all the (Anatomy, Patients) pairs. The question is, is "afflicted-by (unspecified) Disease in Anatomy Part" equivalent to "afflicted by Disease?" The answer depends on one's theory of relational semantics. Another quandary is illustrated by the

NCs *adolescent cancer, child tumors, adult dementia* (in which *adolescent, child* and *adult* are Age Groups) and the heads are Diseases. Should these fall under the afflicted by relation, given the references to entire groups?

# 6 Related Work

## 6.1 Noun Compound Relation Assignment

Several approaches have been proposed for empirical noun compound interpretation. Lauer & Dras (1994) point out that there are three components to the problem: identification of the compound from within the text, syntactic analysis of the compound (left versus right association), and the interpretation of the underlying semantics. Several researchers have tackled the syntactic analysis (Lauer, 1995), (Pustejovsky et al., 1993), (Liberman and Church, 1992), usually using a variation of the idea of finding the subconstituents elsewhere in the corpus and using those to predict how the larger compounds are structured.

We are interested in the third task, interpretation of the underlying semantics. Most related work relies on hand-written rules of one kind or another. Finin (1980) examines the problem of noun compound interpretation in detail, and constructs a complex set of rules. Vanderwende (1994) uses a sophisticated system to extract semantic information automatically from an on-line dictionary, and then manipulates a set of hand-written rules with hand-assigned weights to create an interpretation. Rindflesch et al. (2000) use hand-coded rule-based systems to extract the factual assertions from biomedical text. Lapata (2000) classifies nominalizations according to whether the modifier is the subject or the object of the underlying verb expressed by the head noun.

Barker & Szpakowicz (1998) describe noun compounds as triplets of information: the first constituent, the second constituent, and a marker that can indicate a number of syntactic clues. Relations are initially assigned by hand, and then new ones are classified based on their similarity to previously classified NCs. However, similarity at the lexical level means only that the same word occurs; no generalization over lexical items is made. The algorithm is assessed in terms of how much it speeds up the hand-labeling of relations. Barrett et al. (2001) have a somewhat similar approach, using WordNet and creating heuristics about how to classify a new NC given its similarity to one that has already been seen.

In previous work (Rosario and Hearst, 2001), we demonstrated the utility of using a lexical hierarchy for assigning relations to two-word noun compounds. We use machine learning algorithms and MeSH to successfully generalize from training instances, achieving about 60% accuracy on an 18-way classification problem using a very small training set. That approach is bottom up and requires good coverage in the training set; the approach described in this paper is top-down, characterizing the lexical hierarchies explicitly rather than implicitly through machine learning algorithms.

## 6.2 Using Lexical Hierarchies

Many approaches attempt to automatically assign semantic roles (such as case roles) by computing semantic similarity measures across a large lexical hierarchy; primarily using WordNet (Fellbaum, 1998). Budanitsky & Hirst (2001) provide a comparative analysis of such algorithms.

However, it is uncommon to simply use the hierarchy directly for generalization purposes. Many researchers have noted that WordNet's words are classified into senses that are too fine-grained for standard NLP tasks. For example, Buitelaar (1997) notes that the noun *book* is assigned to seven different senses, including *fact* and *section, subdivision*. Thus most users of WordNet must contend with the sense disambiguation issue in order to use the lexicon.

The most closely related use of a lexical hierarchy that we know of is that of Li & Abe (1998), which uses an information-theoretic measure to make a cut through the top levels of the noun portion of WordNet. This is then used to determine acceptable classes for verb argument structure, and for the prepositional phrase attachment problem and is found to perform as well as or better than existing algorithms.

Additionally, Boggess et al. (1991) "tag" veterinary text using a small set of semantic labels, assigned in much the same way a parser works, and describe this in the context of prepositional phrase attachment.

# 7 Conclusions and Future Work

We have provided evidence that the upper levels of a lexical hierarchy can be used to accurately classify the relations that hold between two-word technical noun compounds. In this paper we focus on biomedical terms using the biomedical lexical ontology MeSH. It may be that such technical, domain-specific terminology is better behaved than NCs drawn from more general text; we will have to assess the technique in other domains to fully assess its applicability.

Several issues need to be explored further. First, we need to ensure that this technique works across the full spectrum of the lexical hierarchy. We have demonstrated the likely usefulness of such an exercise, but all of our analysis was done by hand. It may be useful enough to simply complete the job manually; however, it would be preferable to automate some or all of the analysis. There are several ways to go about this. One approach would be to use existing statistical similarity measures (Budanitsky and Hirst, 2001) to attempt to identify which subhierarchies are homogeneous. Another approach would be to see if, after analyzing more CPs, those categories found to be heterogeneous should be assumed to be heterogeneous across classifications, and similarly for those that seem to be homogeneous.

The second major issue to address is how to extend the technique to multi-word noun compounds. We will need to distinguish between NCs such as *acute migraine treatment* and *oral migraine treatment*, and handle the case when the relation must first be found between the leftmost words. Thus additional steps will be needed; one approach is to compute statistics to indicate likelihood of the various CPs.

Finding noun compound relations is part of our larger effort to investigate what we call statistical semantic parsing (as in (Burton and Brown, 1979); see Grishman (1986) for a nice overview). For example, we would like to be able to interpret titles in terms of semantic relations, for example, transforming *Congenital anomalies of tracheobronchial branching patterns* into a form that allows questions to be answered such as "What kinds of irregularities can occur in lung structure?" We hope that by compositional application of relations to entities, such inferences will be possible.

**Acknowledgements** We thank Kaichi Sung for her work on the relation labeling, Steve Maiorano for his support of this research, and the anonymous reviewers for their comments on the paper. This research was supported by a grant from ARDA.

## References

Christopher Ahlberg and Ben Shneiderman. 1994. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *Proceedings of ACM CHI'94*, pages 313–317.

Ken Barker and Stan Szpakowicz. 1998. Semi-automatic recognition of noun modifier relationships. In *Proceedings of COLING-ACL '98*, Montreal, Canada.

Leslie Barrett, Anthony R. Davis, and Bonnie J. Dorr. 2001. Interpreting noun-noun compounds using wordnet. In *Proceedings of 2001 CICLing Conference*, Mexico City.

Lois Boggess, Rajeev Agarwal, and Ron Davis. 1991. Disambiguation of prepositional phrases in automatically labelled technical text. In *AAAI 91*, pages 155–159.

Alexander Budanitsky and Graeme Hirst. 2001. Semantic distance in wordnet: an experimental, application-oriented evaluation of five measures. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA, June.

P. Buitelaar. 1997. A lexicon for underspecified semantic tagging. In *Proceedings of ANLP 97, SIGLEX Workshop*, Washington DC.

R. R. Burton and J. S. Brown. 1979. Toward a natural-language capability for computer-assisted instruction. In H. O'Neil, editor, *Procedures for Instructional Systems Development*, pages 273–313. Academic Press, New York.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Timothy W. Finin. 1980. *The Semantic Interpretation of Compound Nominals*. Ph.d. dissertation, University of Illinois, Urbana, Illinois.

Ralph Grishman. 1986. *Computational Linguistics*. Cambridge University Press, Cambridge.

Maria Lapata. 2000. The automatic interpretation of nominalizations. In *Proceedings of AAAI*.

Mark Lauer and Mark Dras. 1994. A probabilistic model of compound nouns. In *Proceedings of the 7th Australian Joint Conference on AI*.

Mark Lauer. 1995. Corpus statistics meet the compound noun. In *Proceedings of the 33rd Meeting of the Association for Computational Linguistics*, June.

Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDI principle. *Computational Linguistics*, 24(2):217–244.

Mark Y. Liberman and Kenneth W. Church. 1992. Text analysis and word pronunciation in text-to-speech synthesis. In Sadaoki Furui and Man Mohan Sondhi, editors, *Advances in Speech Signal Processing*, pages 791–831. Marcel Dekker, Inc.

James Pustejovsky, Sabine Bergler, and Peter Anick. 1993. Lexical semantic techniques for corpus analysis. *Computational Linguistics*, 19(2).

James Pustejovsky, editor. 1995. *The Generative Lexicon*. MIT Press.

Thomas Rindflesch, Lorraine Tanabe, John N. Weinstein, and Lawrence Hunter. 2000. Extraction of drugs, genes and relations from the biomedical literature. *Pacific Symposium on Biocomputing*, 5(5).

Barbara Rosario and Marti A. Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*. ACL.

Lucy Vanderwende. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of COLING-94*, pages 782–788.