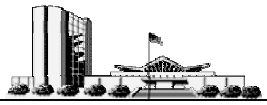
LISTER HILL NATIONAL CENTER
FOR BIOMEDICAL COMMUNICATIONS



**TECHNICAL REPORT**
**LHNCBC-TR-2003-004**

# Semantic Knowledge Representation Project; A report to the Board of Scientific Counselors

September 2003

Thomas C. Rindflesch
Marcelo Fiszman
Halil Kilicoglu
Bisharah Libbus

**U.S. National Library of Medicine, LHNCBC**
**8600 Rockville Pike, Building 38A**
**Bethesda, MD 20894**

## 1. Introduction

Although work in computational lnguistics began soon after the development of the first computers (Booth et al. 1958), significant advancement in automatic understanding of natural language in the general case remains elusive. The Semantic Knowledge Representation project (SKR) is pursuing research aimed at investigating principles that might underpin general progress in natural language processing, including identification of semantic concepts and relationships in biomedical text. The complexity of natural language poses a formidable challenge to realizing this goal.

Formal linguistics, which provides guidance for computational methods, has focused on syntactic structures that encode core semantic phenomena, such as propositions, attitudes, quantification, negation, and tense and modality (Stockwell et al. 1973; McCawley 1988). Traditional methods in computational linguistics (Allen 1995) include a prominent concentration on formal mechanisms that address these structures. Although large systems have been developed in this paradigm (Sager 1981; Palmer et al. 1986; Alshawi 1993, for example), they have not enjoyed lasting practical application. An alternative, semantics-oriented, approach to understanding natural language focused on the use of domain knowledge (Schank 1975; Wilks 1976; Riesbeck 1981, Hahn 1989, for example). The lack of large amounts of readily-accessible structured knowledge, however, prevented this appealing methodology from being widely implemented. More recently, research in natural language processing has relied on statistical methods (Manning and Schütze 1999); however, semantic interpretation has not been the focus of this approach.

Two general strategies seem promising for achieving progress in natural language processing, including semantic interpretation: a) increased attention to large amounts of lexical and domain knowledge (Bates and Weischedel 1993) and b) processing limited by domain or by linguistic phenomena (or both) (Grishman and Kittredge 1986).

Several research groups in the biomedical domain are developing and applying natural language processing methodologies focused in scope, typically by domain of discourse. Many applications are designed to interpret clinical text, including discharge summaries (Zweigenbaum et al. 1995) or imaging reports, such as chest X-rays (Friedman et al. 1994; Fiszman et al. 2000) or mammograms (Jain and Friedman 1997) as well as clinical guidelines (Fiszman and Haug 2000). Others are directed at molecular biology and genetics (Friedman et al. 2001; Pustejovsky et al. 2002, Gaizauskas et al. 2003; Leroy et al. 2003). The majority of this work is knowledge based, and the specific domain guides the type and amount of knowledge used (Baud et al. 1998). Often this is drawn from existing resources, such as the Unified Medical Language System® (UMLS)® (Humphreys et al. 1998) or the GALEN ontology (Amaral et al. 1998), but several systems rely largely on locally developed knowledge bases. Further, system restrictions may be imposed on the basis

of syntactic structure; some process only noun phrases, for example (Rassinoux et al. 1995; Rosario et al. 2002). Finally, various linguistic formalisms are used, including augmented transition networks (Haug et al. 1994) semantic grammars (Friedman et al. 1994), definite clause (Johnson et al. 1993) and dependency grammars (Hahn et al. 1999; Hahn et al. 2000), as well as bottom-up chart parsers (Christensen et al. 2002). (For more comprehensive reviews see Spyns 1996 and Friedman and Hripcsak 1999.)

The SKR approach to natural language processing is being developed in several Prolog applications that are modifications of a core program called SemRep, which is divided into two major conceptual parts: a) basic linguistic processing that is meant to apply to English in general and b) analysis specific to a subdomain of medicine. SemRep differs from other approaches regarding the linguistic formalism used and the way it is limited to only some syntactic structures, as well as the source of the domain knowledge. Syntactic structures are represented by two mechanisms, a shallow categorial parser and an underspecified dependency grammar. Although these are both incomplete, they apply to English syntax in general and are not crafted for the biomedical domain. (See Grishman et al. 2002 for a related approach, although one that does not use the UMLS.)

The domain knowledge for SemRep is taken directly from the UMLS Metathesaurus® and Semantic Network, rather than being compiled manually. The Metathesaurus was not developed as an ontology and will not ultimately support extensive inferencing without enhancement; nonetheless, the breadth of coverage it provides enables the application of SemRep in a variety of medical subdomains with a minimum of effort.

## 2. Project Objectives

The SKR project seeks to exploit UMLS resources in developing effective natural language processing systems that can provide underspecified semantic interpretation to support innovative information management applications in the biomedical domain.

## 3. Project Significance

The development of programs for semantic interpretation underlies increased understanding of viable strategies for effective natural language processing and serves as the basis for ongoing research initiatives in biomedical information management. These include projects for extracting medical and molecular biology information from text, processing clinical data in patient records, and research in knowledge summarization and visualization.

## 4. Methods and Procedures

SKR research fits within the context of the Natural Language Systems (NLS) program and draws on resources being developed in the SPECIALIST system (McCray et al. 1993), which provides a framework for exploiting the UMLS knowledge sources for natural language processing. In addition to the Metathesaurus and Semantic Network, the SPECIALIST lexicon and associated lexical tools (McCray et al. 1994) as well as the UMLS Knowledge Source Server (McCray et al. 1996; Bangalore et al. 2003) support syntactic analysis and semantic interpretation of free text in the biomedical domain.

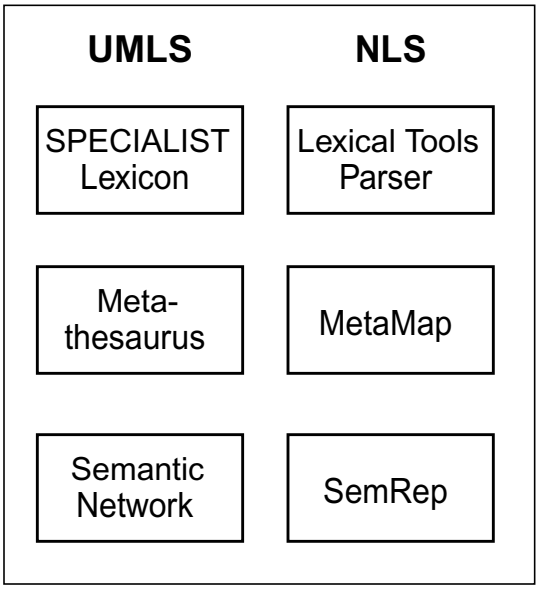



Figure 1. Semantic Knowledge Representation

As shown in Figure 1, the lexicon and associated tools underpin an underspecified syntactic analysis, which serves as input to MetaMap (Aronson 2001), a program that provides access to Metathesaurus concepts. SemRep relies on this processing as well as the relationships in the Semantic Network to identify semantic predications in text.

### 4.1 System overview

During SemRep processing, after input and tokenization, lexical look-up routines give access to syntactic information in the SPECIALIST lexicon. Lexically analyzed text with part-of-speech labels is submitted to the Xerox Tagger (Cutting et al. 1992) for disambiguation. The underspecified parser uses the output of the Tagger as well as additional lexical information, such as inflectional variant forms of verbs and nouns. An example of parser output is given in (2) for the text in (1).

(1) New fluoroquinolones such as ofloxacin are beneficial in the treatment of chronic obstructive airways disease exacerbation requiring mechanical ventilation.

(2) [mod(adj(new)),head(noun(fluoroquinolones),metaconc(Fluoroquinolones:[orch,phsu]))],
[prep(such as),head(noun(ofloxacin),metaconc(Ofloxacin:[orch,phsu]))],
[aux(are)],
[head(adj(beneficial))],
[prep(in),det(the),head(noun(treatment))],
[prep(of),mod(adj(chronic)),mod(adj(obstructive)),mod(noun(airways)),
mod(noun(disease),
head(noun(exacerbation),metaconc(Chronic obstructive airways disease exacerbated:[dsyn]))]
[verb(requiring)],
[head(noun([mechanical ventilation)),punc(.) ]]

Referring expressions such as fluoroquinolones in (2) are augmented with Metathesaurus concepts and semantic types. (The semantic types are abbreviated: Disease or Syndrome (dsyn); Organic Chemical (orch); Pharmacologic Substance (phsu).) This domain knowledge is acquired through MetaMap, a flexible, knowledge-based application that uses the SPECIALIST lexicon along with rules for morphological variants to determine the best mapping between the text of a noun phrase and a concept in the Metathesaurus.

The interpretation of semantic propositions depends on this underspecified analysis enriched with domain knowledge and is driven by syntactic phenomena that indicate semantic predicates, including verbs, prepositions, nominalizations, and the head-modifier relation in simple noun

phrases. Rules are used to map syntactic indicators to predicates in the Semantic Network. For example, there is a rule that links the nominalization *treatment* with the predicate TREATS.

Domain restrictions are enforced by a meta-rule stipulating that all semantic propositions identified by SemRep must be sanctioned by a predication in the Semantic Network. This rule ensures that syntactic arguments associated with *treatment* in the analysis of (2) must have been mapped to Metathesaurus concepts with semantic types that match one of the permissible argument configurations for TREATS, such as 'Pharmacologic Substance' and 'Disease or Syndrome'.

Further syntactic constraints on argument identification are controlled by statements expressed in a dependency grammar. For example, the rules for nominalizations state that one possible argument configuration is for the object to be marked by the preposition *of* occurring to the right of the nominalization and that one possible location for the subject is anywhere to the left of the noun phrase containing the nominalization.

During semantic interpretation of the predication on *treatment* in (2), choosing the noun phrase *ofloxacin* (which maps to a concept with semantic type 'Pharmacologic Substance') as the subject and *chronic obstructive airways disease exacerbation* (mapped to a concept with semantic type 'Disease or Syndrome') as object allows all constraints to be satisfied. The final interpretation is the semantic proposition in (3), where the Metathesaurus concepts are arguments of the predicate from the Semantic Network.

(3) Ofloxacin-TREATS-Chronic obstructive airways disease exacerbated

A final step in SemRep processing submits all the semantic propositions identified in a sentence to local inferencing rules. One of these states that if it is asserted that a drug treats a patient and that the patient suffers from a disorder, then it can be inferred that the drug treats the disorder. For example, SemRep identifies the predications in (5) for (4).

(4) Clinical trials have shown calcipotriol ointment to be an effective and well tolerated topical agent in adult patients with psoriasis

(5) calcipotriene-TREATS-Patients
Psoriasis-OCCURS_IN-Patients

Inferencing then constructs (6).

(6) calcipotriene-TREATS(INFER)-Psoriasis

Inferences are labeled as such in order to signal that they must be treated with caution. For example, the inference (9) drawn from the predications in (8) as the interpretation of (7) does not have the strength of an assertion.

(7) aspirin for patients with AIDS

(8) Aspirin-TREATS-Patients
Acquired Immunodeficiency Syndrome-OCCURS_IN-Patients

(9) Aspirin-TREATS(INFER)-Acquired Immunodeficiency Syndrome

We now examine several aspects of SemRep processing in more detail.

### 4.2 Shallow categorial parser

The output of the underspecified parser is in the tradition of partial parsing (Hindle 1983, McDonald 1992) and concentrates on the simple noun phrase—what Weischedel et al. (1993) call the "core noun phrase," that is, a noun phrase with no modification to the right of the head. Several approaches provide similar output based on statistics (Church 1988, Zhai 1997, for example), a finite-state machine (Ait-Mokhtar and Chanod 1997), or a hybrid approach combining statistics and linguistic rules (Voutilainen and Padro 1997).

Processing is based on the notion of barrier words (Tersmette et al. 1988), which indicate boundaries between phrases. Complementizers, conjunctions, modals, prepositions, and verbs are marked as boundaries. Subsequently, boundaries are considered to open a new phrase (and close the preceding phrase). Any phrase containing a noun is considered to be a (simple) noun phrase, and in such a phrase, the right-most noun is labeled as the head; all other items (other than determiners) are labeled as modifiers.

As seen in (2) above, simple noun phrases are identified and given a partial internal analysis. The head is identified and modifiers occurring to the left of the head other than determiners are marked as modifiers regardless of their part-of-speech label. Prepositional phrases are treated as simple noun phrases whose first element is a preposition. Other syntactic categories, including verbs, auxiliaries, and conjunctions are simply given their part-of-speech label and put into a separate phrase.

This structure alone is not meant to support semantic interpretation. It forms the basis for interpretation of the referential vocabulary. Phrases enriched with Metathesaurus concepts underpin argument identification controlled by the dependency grammar.

### 4.3 Processing the referential vocabulary

Part of the strategy in developing SemRep is to take advantage of the general structure of English as much as possible and to devise specialized processing only when necessary; much of the latter can be isolated to issues in interpreting the referential vocabulary. Terminology varies in definable ways by subdomain of medicine. For example, the concepts in a coronary catheterization report differ dramatically from the concepts in the research literature on molecular biology.

SemRep addresses this phenomenon by incorporating processing that augments MetaMap in interpreting noun phrases as semantic concepts. We illustrate such processing with examples from molecular biology. We have developed a similar strategy for addressing the specialized vocabulary found in clinical text (coronary catheterization reports and gastrointestinal endoscopy reports, for example).

When processing text in the molecular biology domain, before SemRep begins, the input is sent to ABGene (Tanabe and Wilbur 2002) for independent identification of gene and protein names. During SemRep processing of the referential vocabulary, each noun phrase from the underspecified parse is subjected to three steps. First, MetaMap attempts to identify concepts in the Metathesaurus. If the corresponding semantic type belongs to a stipulated set of semantic types (such as 'Gene or Genome') that noun phrase is considered to refer to a molecular biology phenomenon. If the phrase does not map to a Metathesaurus concept, the text tokens in the phrase are compared to

the list of gene names received earlier from ABGene. In the third step, words in a noun phrase not having met one of the first two criteria are matched against a small list of characteristic words for the domain, such as *codon, exon, deletion*, etc. Finally, contiguous simple noun phrases are coalesced into a single macro noun phrase, which is considered to be a potential argument in a semantic relationship.

### 4.4 Indicator rules

An "indicator" is a syntactic phenomenon that expresses a semantic relationship. Examples of indicators in English are nominalizations (and other relational nouns, including gerunds) verbs, and prepositions, as seen in (10), (11), and (12), respectively.

(10) **treatment** of headache with aspirin

(11) aspirin **treats** headache

(12) aspirin **for** headache

One syntactic indicator is structural (rather than lexical): the modifier-head relation in the simple noun phrase. In (13), for example, the fact that *chest* and *pain* are in a head-modifier relation indicates the semantic predicate LOCATION_OF.

(13) [mod(chest),head(pain)]

An indicator is the predicate of a syntactic predication and requires arguments.

The UMLS Semantic Network (McCray 1993) has 54 semantic predicates hierarchically grouped into classes such as SPATIALLY_RELATED_TO and FUNCTIONALLY_RELATED_TO. Included are relations such as TREATS, LOCATION_OF, PREVENTS, and OCCURS_IN.

Indicator rules stipulate a link between an indicator and a relation in the Semantic Network, and are related to Jackendoff's (1997) (and Alshawi's (1992)) correspondence rules. For a particular Semantic Network relation, such rules specify a word along with the relevant part of speech as well as argument cues (for verbs and nominalizations), as in (14). Or the token "mod_head" appears in rules that apply in the simple noun phrase, as in (15). Since part-of-speech and argument cues are not relevant for this indicator, they are left blank in the rule.

(14) (treatment,noun,for-with,treats)

(15) (mod_head,_,_,location_of)

Currently, SemRep has 227 indicator rules linking syntactic phenomena to 58 of the 108 relations (direct and inverse) in the Semantic Network. 180 rules refer to verbs and nominalizations, 40 to prepositions, and 17 to the head-modifier relation in the simple noun phrase. The number of rules referring to a particular relation varies from one rule each for relations such as DEGREE_OF and INDICATES to 29 for TREATS/TREATED_BY. Examples of the number of rules for some relations (and their inverses) are given in (16) and some of the rules for TREATS can be seen in (17).

(16) AFFECTS / AFFECTED_BY - 17
BRANCH_OF / HAS_BRANCH - 11
CAUSES / CAUSED_BY - 14
LOCATION_OF / HAS_LOCATION - 9
OCCURS_IN / HAS_OCCURRENCE - 19

PART_OF / HAS_PART - 13
PROCESS_OF / HAS_PROCESS - 11
TREATS / TREATED_BY - 29
USES / USED_BY - 11

(17) (alleviate,verb,_,treats)
(cure,verb,_,treats)
(for,prep,_,treats)
(in,prep,_,treats)
(management,noun,of,treats)
(relieve,verb,_,treats)

Language genre varies with regard to semantic indicators as well as in the referential vocabulary. We are developing rules for particular genres in addition to the rules above that apply generally in the medical domain. Subdomain-specific rules include those in (18) for coronary catheterization reports and in (19) for molecular biology.

(18) (arise, verb,from,branch_of)
bifurcate,verb,_,has_branch)

(19) (activate,verb,_,stimulate)
(upregulate,verb,_stimulate)
(repress,verb,_,inhibit)
(downregulate,verb,_,inhibit)

An example for the rules in (18) is *the right coronary artery* **arises** *from the aorta* and for those in (19) *RAL-GTP* **activates** *the phospholipase D by hydrolyzing phosphatidilcholines*.

### 4.5 Argument identification

SemRep uses an underspecified dependency grammar supported by semantic propositions from the Semantic Network for argument identification. As is common in dependency grammars, there is a general constraint disallowing crossing dependency lines. A further principle prevents multiple use of arguments, without license (discussed below).

Based on the predicates it contains, the Semantic Network asserts predications linking UMLS semantic types in associative semantic propositions expressing commonly accepted medical knowledge, as seen in (20).

(20) Pharmacologic Substance-TREATS-Disease or Syndrome
Body Part, Organ, or Organ Component-LOCATION_OF-Disease or Syndrome

All semantic predications interpreted by SemRep must be sanctioned by a proposition from the Semantic Network.

#### 4.5.1 Constraints by class of indicator

Subject to the general principles noted above, further constraints for syntactic argument identification are based on classes of indicators. The subject of a preposition must be to its left and its object must be in the same noun phrase as the preposition. The application of these rules is illustrated by the analysis of (21).

(21) Cephalexin in young children with cystic fibrosis

The arguments of *in* in this example are *Cephalexin* and *young children,* and the interpretation for this syntactic predication is the first proposition in (22).

(22) Cephalexin-TREATS-young child
Cystic Fibrosis-OCCURS_IN-young child

The interpretation of the predication on *with* in (21) is the second proposition in (22).

Subjects of verbs must occur to the left of the verb and objects to the right. (We have not yet addressed objects that occur to the left of the subject.) Complement information from the SPECIALIST lexicon further constrains the identification of objects of verbs. If a verb is recognized as being in passive voice, the order of its arguments is reversed. Applying these rules to the passive verb *was upregulated* in (23) identifies the subject as *amphiregulin mRNA* and the coordinated objects as *amphiregulin* and *alpha-thrombin.*

(23) Amphiregulin mRNA was upregulated by amphiregulin itself as well as alpha-thrombin.

The application of the passive rule allows the propositions in (24) to be generated.

(24) amphiregulin-STIMULATE-amphiregulin mrna
alpha thrombin-STIMULATE-amphiregulin mrna

Either one or both of the arguments of a nominalization or gerund may be left unasserted. If both arguments are specified, they can fall into one of the patterns shown below. The prepositional pattern *of-by* is always checked as cuing the object and subject of a nominalization or gerund. Other patterns are stipulated in the indicator rules for specific nominalizations.

(25) a. treatment of aneurysms with surgery

b. [PRED] [Prep1 Arg1] [Prep2 Arg2]

(26) a. surgical treatment of aneurysms

b. [Arg1 PRED] [Prep Arg2]

(27) a. surgery for the treatment of aneurysms

b. [Arg1] [Prep-x PRED] [Prep Arg2]

An example of SemRep processing based on the pattern in (26) applies in the analysis of (28).

(28) A correlation was observed between the level of rhodopsin phosphorylation and the amount of arrestin binding to these mutants.

The gerund *binding* in this example cues its object with the preposition *to,* and its subject occurs immediately to the left of the gerund in the same noun phrase. The application of the indicator rule for *binding* and associated argument identification principles allows the predication in (29) to be generated for (28).

(29) arrestin-BINDS-mutant

### 4.5.2 Argument reuse

As noted above, no argument can be used in the interpretation of more than one predication, with out license. The two licensing phenomena are coordination and relativization. Processing proceeds in two phases for both structure types: a) identification of coordination or relativization and b) construction of predications involving reused arguments.

SemRep addresses noun phrase coordination (Rindflesch 1995) by taking advantage of semantic types. This processing begins before the interpretation of semantic propositions. On the basis of the underspecified syntax enhanced with domain knowledge, an attempt is made to determine whether each coordinator is conjoining noun phrases or something other than noun phrases. For a coordinator determined to be conjoining noun phrases, the semantic type of the noun phrase immediately to the right of that coordinator is examined. The noun phrase immediately to the left of the coordinator and noun phrases occurring to the left of that noun phrase (and separated from it either by another coordinator or by a comma) are examined to see whether they are semantically consonant. In the current formulation of the coordination algorithm, semantic consonance means that the semantic types belong to the same semantic group in the Semantic Network (McCray et al. 2002).

For example in (30), *inflammatory bowel disease* has been mapped to a concept with semantic type 'Disease or Syndrome'; *allergic rhinitis* and *asthma* have also been mapped to concepts with semantic types in the group Disorders, and thus these three noun phrases are considered to be coordinate.

(30) ... a new class of anti-inflammatory drugs that have clinical efficacy in the management of asthma, allergic rhinitis and inflammatory bowel disease

During the process of semantic interpretation, if a coordinate noun phrase is found to be an argument of a semantic predicate, then all noun phrases coordinate with that noun phrase must also be arguments of a predication with that predicate. During the semantic processing of (30), for example, once the first predication in (31) has been constructed, the other two are automatically generated by virtue of the coordinate status of *asthma*. The concept "Anti-Inflammatory Agents" is reused in each predication.

(31) Anti-Inflammatory Agents-TREATS-Asthma
Anti-Inflammatory Agents-TREATS-Allergic rhinitis, NOS
Anti-Inflammatory Agents-TREATS-Inflammatory Bowel Diseases

Heads of relative clauses may also be used in more than one predication. Currently, SemRep recognizes the head of a relative clause when it precedes an overt relativizer (32) or when it precedes a prepositional phrase, of which it is an argument, as in (33).

(32) We report the case of a 75-year-old patient with **onychomycosis** which was treated with ciclopirox.

(33) calcipotriol in **patients** with chronic plaque psoriasis

Reuse of arguments licensed by relativization allows the predications in (34) to be generated from (32), and those in (35) for (33).

(34) Onychomycosis-OCCURS_IN-Patients
ciclopirox-TREATS-Onychomycosis

(35) calcipotrial-TREATS-Patients
Plaque psoriasis-OCCURS_IN-Patients
calcipotriene-TREATS(INFER)-Plaque psoriasis

### 4.6 Hypernymic propositions

We have implemented a mechanism for interpreting hypernymic propositions as a separate module (called SemSpec) in SemRep (Rindflesch and Fiszman 2003). A hypernymic proposition involves two concepts in a taxonomic ('ISA') relationship, one semantically more specific, the hyponym, and the other more general, the hypernym. In English, there are three major syntactic strategies for encoding a hypernymic proposition: with verbs as in (36), appositives (37), or nominal modification (38).

(36) **Modafinil** is a novel **stimulant** that is effective in the treatment of narcolepsy.

(37) **Non-steroidal anti-inflammatory drugs** such as **indomethacin** attenuate inflammatory reactions.

(38) The **anticonvulsant gabapentin** has proven effective for neuropathic pain.

SemSpec takes advantage of the linguistic processing in SemRep by first identifying the syntactic structures that potentially indicate hypernymic propositions. After potential syntactic arguments have been identified, regardless of the structure in which they were found, they are subjected to uniform semantic constraints based on the UMLS. Due to the semantic characteristics of the hypernymic proposition, this knowledge is exploited differently than it is in interpreting associative propositions. Rather than using the overt stipulations in the Semantic Network for semantic constraints on argument identification, SemSpec calls on semantic groups from the Semantic Network and hierarchical relationships from the Metathesaurus to constrain the arguments of the hypernymic proposition.

The interpretation of hypernymic propositions enhances semantic interpretation generally, and in addition increases the usefulness of the associative relationships produced by SemRep. Without this module, SemRep was deficient in interpreting sentences such as (36) above, for which it only identified the proposition "Stimulants-TREATS-Narcolepsy." Although this is correct, it is more useful to identify "Modafinil" (hyponym of "Stimulant") as the semantic subject of TREATS in this sentence. SemSpec determines that "Modafinil" is a hyponym of "Stimulant" in this sentence and thus supplies the information that SemRep needs to provide the more precise semantic interpretation.

We have devised a rule to exploit SemSpec output in SemRep (Fiszman et al. 2003). The rule schema in (39) essentially states that if the argument of an associative semantic proposition is asserted to be in a hypernymic relationship in the sentence being interpreted, the more specific hyponym of the relationship may be substituted for the hypernym in the associative proposition.

(39) I. <Hypernym Y> ARGUMENT_OF $PRED_1$
II. <Hyponym X>-ISA-<Hypernym Y>
III. IF (I and II) THEN IV
IV. <Hyponym X> ARGUMENT_OF $PRED_1$(SPEC)

The application of this rule is illustrated with an analysis of (40).

(40) Market authorization has been granted in France for pilocrapine, an old parasympathomimetic agent, in the treatment of xerostomia.

SemRep retrieves (41) for (40) and SemSpec retrieves (42). From (39), (I) and (II) are true, which allows (43) to be generated. (The label "SPEC" indicates that the predication is due to rules producing the more specific argument.)

(41) Parasympathomimetic Agents-TREATS-Xerostomia

(42) Pilocrapine-ISA-Parasympathomimetic Agents

(43) Pilocrapine-TREATS(SPEC)-Xerostomia

### 4.7 Evaluation

We are developing a test collection based on 2,000 sentences drawn from MEDLINE abstracts discussing (mainly drug) therapies for several diseases, including pneumonia, psoriasis, atopic dermatitis, onychomycosis, hypercholesterolemia, and sinusitis. Initially, our markup concentrates on a few core predicates, primarily TREATS and PREVENTS, but also OCCURS_IN, LOCATION_OF, CO-OCCURS_WITH, and ISA.

This evaluation methodology is limited in the sense that the number of sentences is modest and not necessarily representative of any genre. The judge (Fiszman) is one of the system developers and the set of predications scrutinized is limited. These efforts nonetheless provide valuable guidance in further developing SemRep. Preliminary evaluation limited to TREATS predications indicates Recall of 49% and Precision of 78%, while hypernymic propositions involving chemicals and drugs yielded Precision of 83%.

## 5. Error Analysis

SemRep output from 330 sentences from the test collection were submitted to error analysis. Determining the exact etiology of semantic interpretation errors, especially in sentences with complicated structure, is difficult. We are not yet confident enough of our analyses to provide exact figures; however, certain trends were observed. In the following discussion, we point out major error categories and comment on proposed solutions.

### 5.1 False negatives

About a third of the false negative errors were due to factors other than syntactic processing, and most of these were caused by acronyms. The remainder of the false negatives not due to syntactic processing were due to missing Metathesaurus concepts or indicator rules. We briefly comment on these error types and then turn to false negatives due to syntactic processing.

#### 5.1.1 Errors not due to syntactic processing

Unexpanded acronyms, such as *ALE* (*artichoke leaf extract*) and *VAP* (*ventilator-associated pneumonia*), were not recognized by SemRep as potential arguments for syntactic predications, which could not then be interpreted as semantic propositions. Several recent works address acronyms in

medical text (Liu, Aronson, and Friedman 2002; Wren and Garner 2002; Yu, Hripcsak, and Friedman 2002), and MetaMap is being enhanced to expand author-defined abbreviatory devices.

Although the Metathesaurus has extensive coverage of the biomedical domain, a few gaps were noted during our error analysis. "Subject" in the sense of a person undergoing medical treatment is missing, and hence SemRep was not able to extract predications in the sentence *We studied the effect of roxithromycin in subjects with asthma.* Less surprisingly, artichoke leaf extract is not included as a therapeutic agent (*Artichoke leaf extract for treating hypercholesterolaemia*). The Metathesaurus continues to evolve, and concepts for several drugs that did not appear in the 2002 release are found in 2003, including pimecrolimus and ertapenem.

Several false negatives were caused by missing indicator rules, such as the verb *respond* in (44).

(44) Additionally, reports in the literature suggest that pedal onychomycosis caused by Fusarium species may also **respond** to itraconazole and terbinafine.

### 5.1.2 Errors due to syntactic phenomena

The majority of the false negatives due to syntactic processing involve coordination, often comparative structures. A full interpretation of comparatives is ambitious (Ryan 1981; Friedman 1989). We plan to implement a strategy that at least recognizes comparatives as being a type of coordination. That is, we intend to provide an analysis of (45) that asserts that two substances (tacrolimus ointment and hydrocortisone acetate ointment) are being used to treat atopic dermatitis in children, without including the comparative thrust of the interpretation. Currently only hydrocortisone acetate ointment is recognized in the TREATS predication in (45).

(45) This study was undertaken to compare 0.03% and 0.1% **tacrolimus ointment** with 1% **hydrocortisone acetate ointment** in children 2 to 15 years of age with moderate-to-severe atopic dermatitis.

In the sentences from our error analysis, two elements of a comparative structure are cued most commonly by the patterns "compare X with Y" (as in (45)) and "X compared with Y," both of which can be used to identify the (coordinated) elements of a comparative structure.

The remainder of the false negatives due to syntactic processing are caused by SemRep errors on a range of structures, including unrecognized heads of relatives clauses, both present and past participles missed by the tagger, and problems due to incorrect analysis of prepositionally-cued arguments of nominalizations.

## 5.2 False positives

The majority of false positive errors are due to word sense ambiguity and the way this phenomenon is represented in the Metathesaurus. Particularly troublesome are branded drug names. For example, the words in (46) map to drug names in addition to the more expected concepts.

(46) *duration* "Duration brand of oxymetazoline"
*direct* "Direct type of resin cement"
*active* "Active brand of pseudoephedrine-triprolidine"
*correct* "Correct brand of docusate-phenolphthalein"

An analysis of (47) illustrates how such ambiguity causes SemRep to generate false positive predications.

(47) After **correction** for multiple episodes **in the same child** by generalised estimating equations analysis the odds ratio was 0.74, 95% confidence interval 0.56 to 0.99.

The noun *correction* in (47) maps to the drug concept "Correct" in the Metathesaurus, which allows it to be interpreted in the predication "Correct-TREATS-Child" indicated by the preposition *in*.

## 6. Applications

SemRep and programs derived from it have been applied to processing clinical text as well as the research literature (Rindflesch and Aronson 2002). One project addresses the task of identifying arterial branching relations in cardiac catheterization reports, while another investigates data mining of drug-treatment relations from MEDLINE citations. Other research is aimed at extracting molecular biology information from text. One application identifies macromolecular binding relations, while another is concerned with the interaction of genes, drugs, and cells in the research literature on molecular pharmacology for cancer treatment. Recent research concentrates on the genetic basis of disease.

Rindflesch, Bean, and Sneiderman (2000) report on the use of MetaMap and SemRep for processing cardiac catheterization reports. Statements in the arteriography section of these reports describe characteristics of the arteries seen during the catheterization procedure, such as branching configurations and areas of stenosis. This project focused first on identifying the names for the coronary arteries and then on retrieving the branching relations that were asserted to obtain among the arteries observed. The focused nature of this application along with the complexity of the arterial terminology provided a useful context for development of the SKR methodology. Extensive reliance on UMLS domain knowledge contributed significantly to a highly accurate semantic analysis for these reports.

An extension of the MeSHmap technology (Srinivasan 2001) uses SemRep in cooperation with MeSH indexing to provide increased confidence in identifying potentially interesting semantic relationships in large sets of MEDLINE citations (Srinivasan and Rindflesch 2002). The methodology is discussed for citations having MeSH indexing terms in which a drug concept is modified by the subheading "therapeutic use" and a disease concept is modified by "drug therapy." SemRep provided the semantic interpretation of the title for all such citations. Relevant MeSH indexing terms combined with SemRep predications were extracted from more than 15,000 citations, resulting in 7,332 drug-disease pairs identified. Certain drugs occur in disease contexts of varying diversity. For example, pyrithrioxin appears largely in the context of Alzheimer disease and dementia, while pyridazines have been associated with an array of disorders, including congestive heart failure, depressive disorders, and the common cold. It is appealing to suggest that one of the benefits of this research is to compute a "diversity index" for drugs and diseases encountered in the research literature, which may provide useful information to health care practitioners and researchers.

Several recent SKR projects involve the adaptation and extension of MetaMap and SemRep for extracting molecular biology information from the research literature. One such program, Arbiter, identifies macromolecular binding relationships in MEDLINE citations. Arbiter operates in two phases; the first (Rindflesch et al. 1999) identifies all binding entities mentioned in the input text and addresses such phenomena as molecules, genomic structures, cells and cell components, as well as topographic aspects of molecules, cells, and cell components. During the second phase of processing, Arbiter establishes binding relationships using the general SemRep machinery, focused on forms of the verb *bind* (Rindflesch et al. 2000b). Further research on Arbiter extended its application to protein-protein interactions in general (Sarkar and Rindflesch 2002) as a basis for investigating protein function similarities.

Edgar (Rindflesch et al. 2000c) is being designed to address the interaction among genes, cells, and drugs in molecular pharmacology for cancer therapy. The program first identifies drugs, genes, and cells in text and then determines interactions such as "over expresses" that involve these entities. Edgar identifies drugs, genes, and cells in MEDLINE citations using techniques similar to those used by Arbiter. Gene identification is enhanced by calling on several statistical and empirical methods devised by Tanabe and Wilbur (2002). SemRep underpins the identification of semantic relationships in this domain, which is still under development.

Libbus and Rindflesch (2002) report on a project to construct a general tool (called GBD) intended to help researchers manage the literature in molecular biology. GBD is designed to process MEDLINE citations returned by searches to PubMed. A pilot project seeks to identify and extract information regarding the genetic basis of disease. GBD calls on MetaMap to identify diseases and associated clinical findings in the citations retrieved, while the methods of Tanabe and Wilbur (2002) are used to tag genomic phenomena such as genes, alleles, mutations, polymorphism, and chromosomes. Once such information has been identified in the group of citations returned by PubMed, further processing determines distributional and cooccurrence patterns for user-selected categories.

SemGen (Rindflesch et al. 2003) is a modification of SemRep for identifying and extracting semantic propositions on the causal interaction of genes and diseases from MEDLINE citations. In order to increase accuracy, SemGen employs a statistically-based labeled categorizer (Humphrey 1999) for identifying text in the molecular genetics domain before natural language processing begins. Gene name identification uses the methods developed for Arbiter and Edgar, and disease names are supplied by MetaMap from the Metathesaurus. SemRep argument identification techniques are supplemented with indicator rules for three disease etiology relations: CAUSE (indicated by *cause, determine, and underlie*, for example), PREDISPOSE (*predispose, lead to, susceptibility*), and ASSOCIATED_WITH (*associated with, involve, related, in*). Extracted predications are clustered and displayed in graphical form (Batagelj et al. 1999).

## 7. Project Plans

We intend to continue marking up text for evaluation and performing error analysis on a wider range of sentences. This forms the basis for ongoing improvement of SemRep. Current analysis indicates that the majority of false negative errors are due to deficiencies in processing coordinate structures, including comparatives. We are developing mechanisms for recognizing the coordinated arguments of comparative structures and the proper treatment of coordinated modifiers in

the simple noun phrase. A lesser number of errors are due to inadequate recognition of the heads of relative clauses and prepositionally-cued arguments of nominalizations, which we are also addressing. The Xerox tagger makes a considerable number of mistakes in recognizing participles (both past and present), and we are considering the use of support vector machine technology for correcting these errors.

The majority of the errors we have so far encountered (both false positives and false negatives) are ultimately due to word sense ambiguity and the way this phenomenon is represented in the Metathesaurus. We are cooperating with the Indexing Initiative project in exploring and implementing methods for enhancing MetaMap to resolve at least certain classes of such ambiguity.

In developing extensions to SemRep, we are concentrating on methods for automatically determining textual context that can be exploited by subsequent linguistic processing. These efforts are concentrated in two areas: a) document structure and content and b) the referential vocabulary in specified subdomains of medicine. We intend to use methods for accommodating document format being developed by the Lexical Systems project. Further analysis seeks to exploit document content structure, including headings and subsections to guide semantic interpretation.

Semantic interpretation of the referential vocabulary can be improved by developing methods that are specific to definable subdomains of medicine, such as cardiology, gastroenterology, or molecular genetics. Initial efforts in using automatic methods (Humphrey 1999) to determine the subdomain of text being processed were promising in the molecular genetics domain, and we intend to generalize this methodology to other areas. We also intend to take advantage of efforts in the Medical Ontology Research project that isolate a terminology for a specific subdomain.

We recently began a project to investigate the use of SKR methods for automatically suggesting appropriate images as illustrations for anatomically oriented text. We intend to cooperate with the AnatQuest project in using the methods we are developing to link medical text to images from the Visible Human data.

## 8. Summary

The Semantic Knowledge Representation project is pursuing research aimed at investigating principles that might underpin general progress in natural language processing, including identification of semantic concepts and relationships in biomedical text. This research serves as the basis for implementing applications in biomedical information management and includes projects for extracting medical and molecular biology information from text as well as processing clinical data in patient records. Current efforts concentrate on addressing deficiencies in the underlying linguistic processing and extending the methodology by automatically determining textual context that can be exploited by subsequent linguistic processing.

## References

Ait-Mokhtar S, Chanod J-P. 1997. Incremental finite-state parsing. Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 72-79.

Allen JF. 1995. Natural language understanding. Benjamin Cummings. 2nd edition.

Alshawi H (ed.) 1992. The core language engine. The MIT Press.

Amaral MB, Roberts A, Rector AL. 2000. NLP techniques associated with the OpenGALEN ontology for semi-automatic textual extraction of medical knowledge: abstracting and mapping equivalent linguistic and logistic constructs. Proc AMIA Symp, pp. 76-80.

Aronson AR. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp, pp. 17-21.

Bangalore A, Thorn KE, Tilley C, Peters L. 2003. The UMLS Knowledge Source Server: An object model for delivering UMLS data. Proc AMIA Symp.

Baud RH, Lovis C, Rassinoux AM, Scherrer JR. 1998. Alternative ways for knowledge collection, indexing and robust language retrieval. Methods Inf Med 37(4-5):315-26.

Batagelj V, Mrvar A, Zaversnik M. 1999. Partitioning approach to visualization of large graphs. Lect. Notes Comp. Sc. 1731:90-97.

Booth, AD, Brandwood L, Cleave JP. 1958. Mechanical resolution of linguistic problems. Butterworths Scientific Publications.

Christensen L, Haug PJ, Fiszman M. MPLUS: A probabilistic medical language understanding system. 2002. Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, pp. 29-36. Association for Computational Linguistics.

Church, KW. 1988. A stochastic parts program and noun phrase parser for unrestricted text. Proceedings of the Second Conference on Applied Natural Language Processing, pp. 136-143.

Cutting D, Kupiec J, Pedersen J, Sibun P. 1992. A practical part-of-speech tagger. Proceedings of the Third Conference on Applied Natural Language Processing, pp. 133-40.

Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. 2000. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 7(6):593-604.

Fiszman M, Haug PJ. 2000. Using medical language processing to support real-time evaluation of pneumonia guidelines. Proc AMIA Symp, pp. 235-9.

Fiszman M, Rindflesch TC, Kilicoglu H. 2003. Integrating a hypernymic proposition interpreter into a semantic processor for biomedical text. Proc AMIA Symp.

Friedman C. 1989. A computational treatment of the comparative. New York University doctoral dissertation.

Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. 1994. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1(2):161-74.

Friedman C, Hripcsak G. 1999. Natural language processing and its future in medicine. Acad Med 74(8):890-5.

Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. 2001. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 17 Suppl 1:S74-82.

Gaizauskas R, Demetriou G, Artymiuk PJ, Willett P. 2003. Protein structures and information extraction from biological texts: The PASTA System. Bioinformatics 19(1):135-43.

Grishman, R., and R. Kittredge (eds.). 1986. Analyzing language in restricted domains: Sublanguage description and processing. Lawrence Erlbaum Assoc.

Grishman R, Huttunen S, Yangarber R. 2002. Information extraction for enhanced access to disease outbreak reports. J Biomed Inform 35(4):236-46.

Hahn U. 1989. Making understanders out of parsers: Semantically driven parsing as a key concept for realistic text understanding applications. International Journal of Intelligent Systems 4(3):345-93.

Hahn U, Romacker M, Schulz S. 1999. How knowledge drives understanding--matching medical ontologies with the needs of medical language processing. Artif Intell Med 15(1):25-51.

Hahn U, Romacker M, Schulz S. 2000. MEDSYNDIKATE--design considerations for an ontology-based medical text understanding system. Proc AMIA Symp, pp. 330-4.

Haug PJ, Koehler S, Lau LM, Wang P, Rocha R, Huff S. 1994. A natural language understanding system combining syntactic and semantic techniques. Proc Annu Symp Comput Appl Med Care, pp. 247-51.

Hindle D. 1983. Deterministic parsing of syntactic non-fluencies. Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics, pp. 123-128.

Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. 1995. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med 122(9):681-8.

Humphrey S. 1999. Automatic indexing of documents from journal descriptors: A preliminary investigation. JASIS 50(8):661-674.

Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. 1998. The Unified Medical language System: An informatics research collaboration. J Am Med Inform Assoc 5(1):1-11.

Jackendoff R. 1997. The architecture of the language faculty. The MIT Press.

Jain NL Friedman C.1997. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. Proc AMIA Symp, pp. 829-33.

Johnson SB, Aguirre A, Peng P, Cimino J. 1993. Interpreting natural language queries using the UMLS. Proc Annu Symp Comput Appl Med Care, pp. 294-8.

Leroy G, Chen H, Martinez JD. 2003. A shallow parser based on closed-class words to capture relations in biomedical text. J Biomed Inform. In press.

Libbus B, Rindflesch TC. 2002. NLP-based information extraction for managing the molecular biology literature. Proc AMIA Symp, p. 445-9.

Liu H, Aronson AR, Friedman C. 2002. A Study of Abbreviations in MEDLINE Abstracts. Proc AMIA Symp, pp. 464-9.

Manning C D, Schütze H. 1999. Foundations of statistical natural language processing. The MIT Press.

McCawley JD. 1988. The syntactic phenomena of English. University of Chicago Press. (2 volumes).

McCray AT. 1993. Representing biomedical knowledge in the UMLS Semantic Network. High-Performance Medical Libraries: Advances in Information Management for the Virtual Era. Meckler Publishing, pp. 45-55.

McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A and Srinivasan S. 1993. UMLS knowledge for biomedical language processing. Bulletin of the Medical Library Association 81:184-194.

McCray AT, Srinivasan S, Browne AC. 1994. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care, pp. 235-9.

McCray AT, Burgun A, Bodenreider O. 2001. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo 10(Pt 1):216-20.

McDonald DD. 1992. Robust partial parsing through incremental, multi-algorithm processing. In Paul S. Jacobs (ed.) Text-Based Intelligent Systems, pp. 83-99.

Palmer MS, Dahl DA, Schiffman RJ, Hirschman L, Linebarger M, Dowding J. 1986. Recovering implicit information. 24th Annual Meeting of the Association for Computational Linguistics, pp. 10-19.

Pustejovsky J, Castano J, Zhang J, Kotecki M, Cochran B. 2002. Robust relational parsing over biomedical literature: extracting inhibit relations. Pac Symp Biocomput, pp. 362-73.

Riesbeck, Christopher K. 1981. Perspectives on parsing issues. Proceedings of the Nineteenth Annual Meeting of the Association for Computational Linguistics, 105-6.

Rosario B, Hearst M, Fillmore C. 2002. The descent of hierarchy, and selection in relational semantics. Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, pp. 247-54. Association for Computational Linguistics.

Rassinoux AM, Wagner JC, Lovis C, Baud RH, Rector A, Scherrer JR. 1995. Analysis of medical texts based on a sound medical model. Proc Annu Symp Comput Appl Med Care, pp. 27-31.

Rindflesch TC. 1995. Integrating natural language processing and biomedical domain knowledge for increased information retrieval effectiveness. Proceedings of the 5th Annual Dual-use Technologies and Applications Conference, 260-5.

Rindflesch TC, Hunter L, Aronson AR. 1999. Mining molecular binding terms from biomedical text. Proc AMIA Symp, pp. 127-31.

Rindflesch TC, Bean CA, Sneiderman CA. 2000a. Argument identification for arterial branching predications asserted in cardiac catheterization reports. Proc AMIA Symp, pp. 704-8.

Rindflesch TC, Rajan J, Hunter L. 2000b. Extracting molecular binding relationships from biomedical text. Appl. Nat. Lang. Proc, pp. 188-95.

Rindflesch TC, Tanabe L, Weinstein JW, Hunter L. 2000c. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput, pp. 517-28.

Rindflesch TC, Aronson AR. 2002. Semantic processing for enhanced access to biomedical knowledge. Vipul Kashyap and Leon Shklar (eds.) Real World Semantic Web Applications, 157-72. IOS Press.

Rindflesch TC, Fiszman M. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics [submitted].

Rindflesch TC, Libbus B, Hristovski D, Aronson AR, Kilicoglu H. 2003. Semantic relations asserting the etiology of genetic diseases. Proc AMIA Symp.

Ryan K. 1981. Corepresentational grammar and parsing English comparatives. 19th Annual Meeting of the Association for Computational Linguistics, pp. 13-18.

Sager N. 1981. Natural language information processing: A computer grammar of English and its application. Addison-Wesley.

Sarkar IN, Rindflesch TC. 2002. Discovering protein similarity using natural language processing. Proc AMIA Symp, pp. 677-81.

Schank RC. 1975. Conceptual information processing. Amsterdam: North-Holland Publishing Co.

Spyns P. 1996. Natural language processing in medicine: an overview. Methods Inf Med 35(4-5):285-301.

Srinivasan P. 2001. A text mining tool for MEDLINE. Proc AMIA Symp, pp. 642-6.

Srinivasan P, Rindflesch TC. 2002. Exploring text mining from MEDLINE. Proc AMIA Symp, pp. 722-6.

Stockwell RP, Schachter P, Partee BH. 1973. The major syntactic structures of English. Holt, Rinehart and Winston.

Tanabe L, Wilbur WJ. 2002. Tagging gene and protein names in biomedical text. Bioinformatics. 18(8):1124-32.

Tersmette KWF, Scott AF, Moore GW, Matheson NW, and Miller RE. 1988. Barrier word method for detecting molecular biology multiple word terms. In Greenes RA (ed.) Proceedings of the 12th Annual Symposium on Computer Applications in Medical Care, pp. 207-211.

Vourtilainen A, Padro L. 1997. Developing a hybrid NP parser. Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 80-87.

Weischedel R, Meteer M, Schwartz R, Ramshaw L, Palmucci J. (1993). Coping with ambiguity and unknown words through probabilistic models. Computational Linguistics 19(2):359-382.

Wilks YA. 1976. Parsing English II. In Eugene Charniak and Yorick Wilks (eds.) Computational semantics: An introduction to artificial intelligence and natural language comprehension. North-Holland Publishing Co.

Wren JD, Garner HR. 2002. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. Methods Inf Med 41(5):426-34.

Yu H, Hripcsak G, Friedman C. 2002. Mapping abbreviations to full forms in electronic articles. J Am Med Inform Assoc 9(3):262-72.

Zhai C. (1997). Fast statistical parsing of noun phrases for document indexing. Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 312-319

Zweigenbaum P, Bachimont B, Bouaud J, Charlet J, Boisvieux JF. 1995. A multi-lingual architecture for building a normalised conceptual representation from medical language. Proc Annu Symp Comput Appl Med Care, pp. 357-61.

**Curriculum Vitae**

Thomas C. Rindflesch

Information Research Specialist

**Education and Training:**

| | | | | |
|---|---|---|---|---|
| University of Minnesota | BA | *summa cum laude* | 1973 | Arabic |
| University of Minnesota | MA | | 1984 | Linguistics |
| University of Minnesota | PhD | | 1990 | Linguistics |

**Research and Professional Experience:**

Computational linguist, Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD, 1991-

Instructor, Department of Linguistics, University of Minnesota, Minneapolis, 1990-1991

Special Projects Supervisor, Academic Computing Services and Systems (formerly University Computer Center), University of Minnesota, Minneapolis, 1985-1989

Project Assistant, University Computer Center, University of Minnesota, Minneapolis, 1978-1985

Teaching Assistant, Department of Linguistics, University of Minnesota, Minneapolis, 1975-1978

**Selected Publications:**

Rindflesch, Thomas C.; Bisharah Libbus; Dimitar Hristovski; Alan R. Aronson; and Halil Kilicoglu. 2003. Semantic relations asserting the etiology of genetic diseases. Proceedings of the AMIA Annual Symposium.

Fiszman, Marcelo; Thomas C. Rindflesch; and Halil Kilicoglu. 2003. Integrating a hypernymic proposition interpreter into a semantic processor for biomedical text. Proceedings of the AMIA Annual Symposium.

Rindflesch, Thomas C., and Marcelo Fiszman. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics. [submitted]

Rindflesch, Thomas C., and Alan R. Aronson. 2002. Semantic processing for enhanced access to biomedical knowledge. Vipul Kashyap and Leon Shklar (eds.) Real World Semantic Web Applications, 157-72. IOS Press.

Srinivasan, Suresh; Thomas C. Rindflesch; William T. Hole; and Alan R. Aronson. 2002. Finding UMLS Metathesaurus concepts in MEDLINE. Isaac Kohane (ed.) Proceedings of the AMIA Annual Symposium, 727-31.

Srinivasan, Padmini, and Thomas C. Rindflesch. 2002. Exploring text mining from MEDLINE. Isaac Kohane (ed.) Proceedings of the AMIA Annual Symposium, 722-6.

Sarkar, Indra Neil, and Thomas C. Rindflesch. 2002. Discovering protein similarity using natural language processing. Isaac Kohane (ed.) Proceedings of the AMIA Annual Symposium, 677-81.

Libbus, Bisharah, and Thomas C. Rindflesch. 2002. NLP-based information extraction for managing the molecular biology literature. Isaac Kohane (ed.) Proceedings of the AMIA Annual Symposium, 445-9.

Bodenreider, Olivier; Thomas C. Rindflesch; and Anita Burgun. 2002. Unsupervised corpus-based method for extending a biomedical terminology. Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, 53-60. Association for Computational Linguistics.

Humphrey, Susanne M.; Thomas C. Rindflesch; and Alan R. Aronson. 2000. Automatic indexing by discipline and high-level category: Methodology and potential applications. Proceedings of the 11th SIG/CR Classification Research Workshop, 103-16.

Rindflesch, Thomas C.; Carol A. Bean; and Charles A. Sneiderman. 2000. Argument identification for arterial branching predications asserted in cardiac catheterization reports, Proceedings of the AMIA Annual Symposium.

Rindflesch, Thomas C.; Jayant V. Rajan; and Lawrence Hunter. 2000. Extracting molecular binding relationships from biomedical text. Proceedings of the 6th Applied Natural Language Processing Conference, 188-95. Association for Computational Linguistics.

Rindflesch, Thomas C.; Lorraine Tanabe; John N. Weinstein; and Lawrence Hunter. 2000. EDGAR: Extraction of drugs, genes, and relations from the biomedical literature. Pacific Symposium on Biocomputing (PSB) 5:514-25

Rindflesch, Thomas C.; Lawrence Hunter; and Alan R. Aronson. 1999. Mining molecular binding terminology from biomedical text. Proceedings of the AMIA Annual Symposium, 127-31.

Wright, Lawrence W.; Holly K. Gorsetta Nardini; Alan R. Aronson; and Thomas C. Rindflesch. 1999. Hierarchical concept indexing of full-text documents in the Unified Medical Language System Information Sources Map. Journal of the American Society for Information Retrieval 50(6):514-23.

Divita, Guy; Allen C. Browne; and Thomas C. Rindflesch. 1998. Evaluating lexical variant generation to improve information retrieval. Proceedings of the AMIA Annual Symposium, 775-9.

Aronson, Alan R., and Thomas C. Rindflesch. 1997. Query expansion using the UMLS Metathesaurus. Daniel R. Masys (ed.) Proceedings of the AMIA Annual Fall Symposium, 485-9.

Sneiderman, Charles A.; Thomas C. Rindflesch; and Alan R. Aronson. 1996. Finding the findings: Identification of findings in medical literature using restricted natural language processing. James J. Cimino (ed.) Proceedings of the AMIA Annual Fall Symposium, 239-43.

Rindflesch, Thomas C. 1996. Natural language processing. William Grabe (ed.) *Annual Review of Applied Linguistics* 16:71-85. Cambridge: Cambridge University Press.

Rindflesch, Thomas C. 1995. Integrating natural language processing and biomedical domain knowledge for increased information retrieval effectiveness. Proceedings of the 5th Annual Dual-use Technologies and Applications Conference, 260-5.

## Curriculum Vitae

Marcelo Fiszman

Medical Informatics Postdoctoral Fellow

**Education and Training:**

| | | | |
|---|---|---|---|
| State University of Rio de Janeiro | MD | 1991 | Medicine |
| University of Sao Paulo | Residency | 1994 | Internal Medicine |
| University of Sao Paulo | Residency | 1996 | Medical Informatics |
| The University of Utah | PhD | 2002 | Medical Informatics |
| National Library of Medicine | Fellow | Present | Medical Informatics |

**Research and Professional Experience:**

Research Assistant, The University of Utah, Dept. of Medical Informatics, 1999-2002.

Teaching Assistant, The University of Utah, Dept. of Medical Informatics, 2000-2001.

**Scientific and Review Committees:**

AMIA – American Medical Informatics Association, member of the review committee in 2000, 2003.

**Awards:**

**Bruce Houtchens Award** in Medical Informatics from the School of Medicine at the University of Utah in 2002 for the paper: Utilization review of head CT scans: value of a medical language processing system.

**Homer Warner Award** at AMIA 2000 fall meeting for the paper: Using Medical Language Processing to Support Real-Time Evaluation of Pneumonia Guidelines.

**Second Prize Student Paper Competition** at AMIA 2000 fall meeting for the paper: Using Medical Language Processing to Support Real-time Evaluation of Pneumonia Guidelines.

**Grants and Scholarships:**

(FAPESP) Brazilian Scholarship. The University of Utah. Department of Medical Informatics.

**Selected Publications:**

Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. (Recommended for Publication), Jan 2003.

Fiszman M, Rindflesch TC, Kilicoglu H. Integrating a Hypernymic Proposition Interpreter into a Semantic Processor for Biomedical Texts. Proc AMIA Symp. 2003 (Accepted).

Fiszman M, Blatter DD, Christensen LM, Oderich G, Haug PJ. Utilization review of head CT scans: value of a medical language processing system. Radiology. August 2003 (submitted).

Christensen, L. Haug PJ, Fiszman M. MPLUS: a probabilistic medical language understanding system. Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain of the Association of Computational Linguistics. July 2002,:29-36.

Chapman WW, Fiszman M, Chapman E, Haug PJ. A comparison of classification algorithms to automatically identify chest x-ray reports that support pneumonia. J Biomed Inform 2001 Feb;34(1):4-14.

Chapman WW, Fiszman M, Frederick P, Chapman B, Haug PJ. Quantifying the characteristics of unambiguous chest radiography reports in the context of pneumonia. Acad Radiol 2001 Jan;8(1):57-66.

Aronsky D, Fiszman M, Chapman WW, Haug PJ. Combining Decision Support Methodologies to Diagnose Pneumonia. Proc AMIA Symp. 2001;:12-6

Lagor C, Aronsky D, Fiszman M, Haug PJ. Automatic identification of patients eligible for a pneumonia guideline: comparing the diagnostic accuracy of two decision support models. Medinfo 2001;10(Pt 1):493-7.

Fiszman M, Chapman WW, Aronsky D, Evans SR, Haug PJ. Automatic detection of acute bacterial pneumonia from chest x-ray reports. J Am Med Inform Assoc 2000 Nov-Dec;7(6):593-604.

Fiszman M, Haug PJ. Using medical language processing to support real-time evaluation of pneumonia guidelines. Proc AMIA Symp. 2000;:235-9.

Chapman WW, Aronsky D, Fiszman M, Haug PJ. Contribution of a Speech Recognition System to Computerized Pneumonia Guidelines in the Emergency Department. Proc AMIA Symp 2000;:131-5.

Fiszman M, Chapman WW, Evans SR, Haug PJ. Automatic identification of pneumonia related concepts on chest x-ray reports. Proc AMIA Symp. 1999;:67-71.

Fiszman M; Haug PJ; Frederick, P. Automatic extraction of PIOPED Interpretations from Ventilation/Perfusion (V/Q) Lung Scan Reports. Proc AMIA Symp 1998;:201-205.

**Posters:**

Fiszman M, Rindflesch TC, Kilicoglu H. Interpreting Hypernymic Propositions in an Online Medical Encyclopedia. Proc AMIA Symp. 2003. (Accepted).

Tringali M, Fiszman M, Rindflesch TC, Bodenreider O. Automatic Extraction of Concepts from Gastrointestinal Endoscopy Reports. MIE 2002.

Aronsky D, Fiszman M, Chapman WW, Davis AF, Reichert J, Sachet MR, Haug PJ, Dean NC. Unlocking Pneumonia Related Findings from Narrative Chest X-Ray Reports for Computerizing Clinical Guidelines. Am J Respir Crit Care Med 2000;161:A307.

Lagor C, Aronsky D, Fiszman M, Haug PJ. Comparing diagnostic decision support systems for pneumonia. Proc AMIA Symp 2000: 960.

Chapman WW, Fiszman M, Haug PJ. Correct vs. Parsed for inferring pneumonia in chest x-ray reports. Proc AMIA Symp. 1999:1038.

## Curriculum Vitae

Halil H. Kilicoglu

Software Engineer, Aquilent, Inc.

**Education and Training:**

| | | | |
|---|---|---|---|
| Istanbul Technical University | BS | 1997 | Computer Engineering |
| George Washington University | MS | 2000 | Computer Science |

**Research and Professional Experience:**

2000-present, Software Engineer, Aquilent Inc., Laurel, MD

Mr. Kilicoglu worked on ABPP (Agency BRIO Pilot Program) project for NASA IFMP Program from July 2000 to September 2001. ABPP project involved the implementation of an agency-wide data warehouse and web-based reporting environment with a limited set of legacy Financial and Contractual Status (FACS) System financial (FACS B) data.

Mr. Kilicoglu was also involved in AVATAR project for NASA Code 588 from July 2001 to September 2001. This project involved researching and prototyping the application of visualization technology to spacecraft operations and science engineering data analysis in order to increase operator performance in multi-mission, constellation, and lights-out environments.

Mr. Kilicoglu worked for Small-Scale Vocabulary Testing (SSVT) project at National Library of Medicine between October 2001 and December 2001. This project involved the implementation of a Web-based application, which allows medical experts from different medical fields to determine the extent to which a combination of existing health terminologies define existing medical images in ImageMed database.

Mr. Kilicoglu currently supports Semantic Knowledge Representation (SKR) project at National Library of Medicine as a software engineer. He developed a client/server application to address the word sense ambiguity problem using different methodologies. He is currently involved with NLP research and application development.

1999-2000, Teaching Assistant and Database Administrator, Department of Computer Science, George Washington University, Washington, DC

1997-1998, Software Engineer, Sistek Yazilim, Istanbul, Turkey

**Honors:**

NASA Award for Quality and Process Improvement and Customer Service Excellence for the contributions to Agency BRIO Pilot Program (ABPP) project.

**Publications:**

Rindflesch, Thomas C.; Bisharah Libbus; Dimitar Hristovski; Alan R. Aronson; and Halil Kilicoglu. 2003. Semantic relations asserting the etiology of genetic diseases. Proceedings of the AMIA Annual Symposium.

Fiszman, Marcelo; Thomas C. Rindflesch; and Halil Kilicoglu. 2003. Integrating a hypernymic proposition interpreter into a semantic processor for biomedical text. Proceedings of the AMIA Annual Symposium.

**Curriculum Vitae**

Bisharah Libbus

Visiting Scientist

**Education and Training:**

| | | | |
|---|---|---|---|
| American University of Beirut | B.S. | 1967 | Biology |
| American University of Beirut | M.S. | 1971 | Biology |
| University of Missouri, Columbia, | Ph.D. | 1976 | Genetics |

**Research & Professional Experience:**

Visiting Scientist, Lister Hill National Center for Biomedical Communication, National Library of Medicine, National Institutes of Health, Bethesda, MD, 2002-

Fellow, Bioinformatics, University of North Carolina, Chapel Hill, 2002-2002

President, Genetic Research, Inc., Research Triangle Park, NC, 1995-2000, 1988-1993

Senior Scientist, Integrated laboratory Systems, Research Triangle Park, NC

Research Associate Professor, Department of Pathology, University of Vermont; Director, Clinical Cytogenetics Laboratory, Medical Center Hospital of Vermont, Burlington, Vermont, 1986-1988

Visiting Scientist, Animal Reproduction Laboratory, U.S. Department of Agriculture Animal Research Center, Beltsville, Maryland, 1985-1986

Visiting Associate Professor, Department of Pediatrics, American University of Beirut; Director, Clinical Cytogenetics Laboratory, National Unit of Human Genetics, American University Medical Center, Beirut, Lebanon, 1984-1985

Associate Professor, Chairman of the Science Division, Haigazian College, Beirut, Lebanon, 1980-1984

Senior Research Scientist, Division of Reproductive Biology, The Johns Hopkins University, Baltimore, MD, 1978-1980

**Selected Publications:**

Libbus, B.L., and Y-C. Hsu. 1980a. Sequential development and tissue organization in whole mouse embryos cultured from blastocyst to early somite stage. Anat. Rec. 159: 317-329.

Libbus, B.L., and Y-C. Hsu. 1980b. Changes in S-phase associated with differentiation of mouse embryos cultured from blastocyst to early somite stage. Anat. Embryol. 159: 235-244.

Libbus, B.L., B.Y. Barakat, and J.M. Schneider. 1980. Failure of estrogen administered to adult rats to suppress meiotic DNA synthesis and differentiation of spermatocytes. Biol. Reprod. 22: 619-627.

Libbus, B.L., and A.B. Burdick. 1980. Ultrastructural changes induced by steroids and gonadotropins in cultured rat Sertoli cells. Cytobios 30: 173-187.

Libbus. B.L. 1985. The ordered arrangement of chromosomes in the Chinese hamster spermatocyte nucleus. Hum. Genet. 70: 130-135.

Libbus, B.L., S.D. Perreault, L.A. Johnson, and D. Pinkel. 1987. Incidence of chromosome aberrations in mammalian sperm stained with Hoechst 33342 and UV-laser-irradiated during flow sorting. Mut. Res. 182: 265-274.

Libbus, B.L., and L.A. Johnson. 1988. The creeping vole (Microtus oregonii): Karyotype and sex chromosome differences between two geographical populations. Cytogenet. Cell Genet. 47: 181-184.

Libbus, B.L., and J.E. Craighead. 1988. Chromosomal translocations with specific breakpoints in asbestos-induced rat mesotheliomas. Cancer Res. 48: 6455-6461.

Libbus, B.L., S.A. Illenye, and J.E. Craighead. 1989. Induction of DNA strand breaks by crocidolite asbestos as assessed by nick translation. Cancer Res. 49: 5713-5718.

Libbus, B.L., L.S. Borman, C.H. Ventrone, and R.C. Branda. 1990. Nutritional folate-deficiency in CHO cells. II. Chromosomal abnormalities associated with perturbations in nucleic acid precursors. Cancer Genet. Cytogenet. 46: 231-242.

French, J.E., B.L. Libbus, L. Hansen, J. Spalding, R.T. Tice, J. Mahler, and R.W. Tennant. 1994. Malignant skin tumor cells from chemically treated TG-AC transgenic and FVB/N wildtype mice exhibit trisomy 6 or 15. Mol. Carcinog. 11: 215-226.

Jeffay, S.C., B.L. Libbus, R.B. Barbee, and S.D. Perreault. 1996. Acute exposure of female hamsters to carbendazim (MBC) during meiosis results in aneuploid oocytes with subsequent arrest of embryonic cleavage and implantation. Reprod. Toxicol. 10: 183-189.

Bermudez, E., B. Libbus, and J.B. Magnum. 1998. Rat pleural mesothelial cells adapted to serum-free medium as a model for the study of growth factor effects. Cell Biol. Toxicol. 14: 243-251.

Robbins, W.A., K.L. Witt, J.K. Haseman, D.B.Dunson, L. Troiani, M.S. Cohen, C.D. Hamilton, S.D. Perreault, B. Libbus, S.A. Beyler, D.J. Raburn, S.T. Tedder, M.D. Shelby, and J.B. Bishop. 2001. Antiretroviral therapy effects on genetic and morphologic end points in lymphocytes and sperm of men with human immunodeficiency virus infection. J. Inf. Dis. 184: 127-135.

Libbus, B., and T. C. Rindflesch. NLP-Based Information Extraction for Managing the Molecular Biology Literature. Proc. AMIA Symp. 2002:445-9.

Rindflesch, T.C., B. Libbus, D. Hristovski, A.R. Aronson, and H. Kilicoglu. Semantic Relations Asserting the Etiology of Genetic Diseases. Proc. AMIA Symp. 2003: In Press.