Creating a Knowledge Base of Biological Research Papers*

Carole D. Hafner, Kenneth Baclawski, Robert P. Futrelle, Natalya Fridman, Shobana Sampath
College of Computer Science, Northeastern University, Boston, MA 02115
{hafner, kenb, futrelle, natasha, shobanas}@ccs.neu.edu
Tel. 617-373-2462 FAX 617-373-5121

Keywords: knowledge representation, natural language, text retrieval, semantic nets, taxonomy, frames, parsing, object oriented databases.

Abstract

Intelligent text-oriented tools for representing and searching the biological research literature are being developed, which combine object-oriented databases with artificial intelligence techniques to create a richly structured knowledge base of Materials and Methods sections of biological research papers. A knowledge model of experimental processes, biological and chemical substances, and analytical techniques is described, based on the representation techniques of taxonomic semantic nets and knowledge frames. Two approaches to populating the knowledge base with the contents of biological research papers are described: natural language processing and an interactive knowledge definition tool.

1. Introduction

Biological data and research results are rapidly becoming electronically accessible on CD-ROM or through computer networks such as Internet. Since published papers represent the primary output of biological research - about 600,000 are published each year - the prospect of a "digital library" presents an opportunity for computer scientists and biologists to move beyond exact reproduction of hard-copy resources to create intelligent text-oriented tools for representing and searching the biological research literature.

Our project is investigating the potential for using artificial intelligence techniques in combination with object oriented databases to create a richly structured knowledge base of biological research papers. Several electronic text and knowledge resources are being utilized:

a. A corpus of 132 papers in Bacterial Chemotaxis, annotated using the Standard Generalized Markup Language [Bryan 1988]. This is the primary corpus around which we are building our prototype tools and knowledge base.

b. The Unified Medical Language System, a large taxonomy of medical concepts created by the National Library of Medicine [UMLS 1993]. The UMLS provides a valuable point of comparison for our knowledge model.

Initially we are dealing only with papers in the field of bacterial chemotaxis, but the techniques and tools we develop will be applicable to other branches of molecular biology. We are focusing on the Materials and Methods sections of these papers, as being both typical of texts in experimental biology and sufficiently narrow and patterned to be amenable to knowledge engineering techniques.

This report describes research aimed at creating a knowledge base of the Materials and Methods sections of the 132 bacterial chemotaxis papers, including both the text and associated knowledge frames in an integrated object-oriented structure. This knowledge base will be used to create a prototype of an intelligent retrieval system for biological research, and to experiment with a variety of information retrieval techniques.

The major challenges we face are: first, to create a knowledge model capable of expressing a significant range of biological concepts (Section 2); and second, to overcome the "knowledge bottleneck" by creating automated or semi-automated tools to populate the knowledge base with frames for a corpus of papers (Section 3). Although 132 documents is a very small corpus which might be represented without automated tools (although this is still a non-trivial effort), the aim of our research is to develop techniques and tools that will help us "scale up" to larger knowledge bases in the future.

We are also investigating concept-based retrieval algorithms for large document collections [Baclawski 1994] and developing an interactive query system for the knowledge base described in this report [Baclawski 1993b]. Software is being developed on the Apple Macintosh computer,using the WOOD object-oriented database system [St. Clair 1993].

2. Knowledge Model

Intelligent processing of language requires background knowledge, which permits an agent (whether computer or human) to make connections between a current input and other objects and events that have been or are being observed. In the sample text (Figure 1)[Kuo 1986], an instance of a complex method called **Immunoblots** is described, and details are provided for a large number of specific sub-processes, as indicated in the following quotations:

Electrophoretic transfer of proteins from the gel to nitrocellulose

Efficiency of protein transfer was *determined*
Ponceau S *staining*
Washes were performed
Blocking steps
antibody *incubations*
^{125}I-protein A *incubations*
Rabbit sera were *diluted.*
filters were *rinsed* . . . and *washed* twice
Filters . . .(were) *autoradiographed*
Quantitation was performed

Immunoblots. Polyacrylamide protein gels were assembled and run by the method of Laemmli(16). Electrophoretic transfer of proteins from the gel to nitrocellulose for immunoblots (35) used a buffer containing 25 mM Tris hydrochloride (pH 8.3), 192 mM glycine, 0.01% (wt/vol) sodium dodecyl sulfate, and 20% (vol/vol) methanol methanol at 65 V overnight (12 to 18 h) in a Bio-Rad Transblot System. Efficiency of protein transfer was determined by Ponceau S staining of nitrocellulose filters. . . Washes were performed in a buffer containing 50 mM Tris hydrochloride (pH 8.0), 150 mM NaCl, and 0.05% (wt/vol) Nonidet P-40 (TBSN). Blocking steps, antibody incubations, and ^{125}I-protein A incubations were performed in TBSN buffer containing 5%(wt/vol) instant nonfat dry milk (TBSN-milk). Rabbit sera were diluted with TBSN-milk buffer 1:200 for the a-Che Y antibody serum and 1:500 for the a-CheZ-antibody serum . . After each incubation step, filters were rinsed with TBSN buffer and washed twice in TBSN buffer with 10 min of agitation. Filters were air dried before being autoradiographed with intensifying screens. Quantitation was performed by using a Searle γ--radiation counter to count bands excised from the nitrocellulose filters.

Figure 1. Materials and Methods Text [Kuo 1986]

To interpret such phrases and see how they fit together, we create structured knowledge frames that specify for each type of process, the purposes or effects of the procedure, the materials and equipment required to carry it out, and (where possible) its contribution to more complex processes. Each of the processes described above contributes to the Immunoblot method, whose goal is to measure the concentration of CheY and CheZ protein in a solution. (Note that this goal is not explicitly stated anywhere in the paragraph.)

The creation of an appropriate knowledge model, or ontology, for molecular biology experiments underlies all of the major algorithms being investigated in this project:

A. Intelligent retrieval. A knowledge model provides the ability for retrieval systems to recognize conceptual similarities that affect the relevance of a document to a user's query. For example, two procedures that measure the same kind of thing (e.g., concentration of protein) are more similar than two procedures that measure different things. If the two procedures both involve radioactive labeling, that is another indicator of similarity.

B. Automatic acquisition of knowledge frames from text. In order to extract knowledge automatically from text, a well-defined target model is required, as well as a "grammar" that specifies how words and phrases can be translated to knowledge frames. Experiments in analysis of text from our bacterial chemotaxis corpus are described in Section 3.2.

C. Interactive tools for human definition and correction of knowledge frames. Since current automatic text analysis techniques [Sundheim 1992] have a high error rate, converting scientific text into a high-quality knowledge frame representation will require human intervention for the foreseeable future. A knowledge model provides a framework for presenting human knowledge definers with a series of meaningful templates and choices. A prototype for knowledge definition is described in Section 3.3.

What is an *appropriate* knowledge model for representing texts such as shown in Figure 1? Considering the three tasks our knowledge model is intended to support (listed above), it is clear that the model must reflect, as accurately as possible, the way scientists think and talk about the subject. A model with this characteristic, which [Shortliffe 1981] refers to as a *clinical reasoning* model, is important for two reasons: the task of extracting knowledge from research articles will be more direct, for both automated and human translators; and (most importantly) the interactive retrieval and knowledge defining tools, which are directly based on the knowledge model, will be more intuitive and easier for scientists to use.

On the other hand, it is neither necessary nor possible to represent the complete range of concepts that the scientist understands. (Nor would such a goal be feasible with today's artificial intelligence methods.) By restricting our attention to a very narrow domain (bacterial chemotaxis, Material and Methods), we can make some simplifications. For example, our classification of living organisms, includes only two sub-types: bacteria and viruses (see Figure 3). Plants, animals, and other organisms are not included in the model., and "person" is treated as a fundamental category. Although the experimenters are also living organisms, that fact is not relevant for our purposes.

2.1. Knowledge Structures I: Taxonomy

The simplest kind of knowledge model organizes the concepts of the domain into a taxonomic hierarchy of (more general) superclasses and (more specific) subclasses. Our hierarchy, like that of the UMLS, makes a high level distinction between *entities* and *events* (see Figure 2). We also include *piece of information* as a high level category. As in the case of the UMLS, categories have a relatively small number of subclasses until we reach the most specific level; then there may be hundreds (for example, there are a very large number of different proteins.).

Within the hierarchy of entities, subclasses include: *person, organization, publication, equipment, Methods and Materials text,* and *substance.* Under the substance category, a number of concepts important to the bacterial chemotaxis domain are represented, further divided into

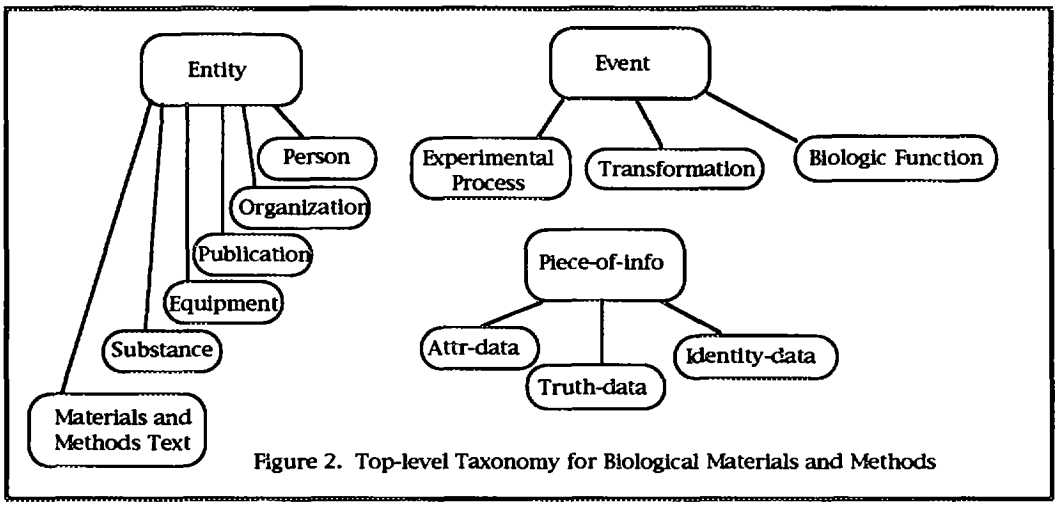

Figure 2. Top-level Taxonomy for Biological Materials and Methods

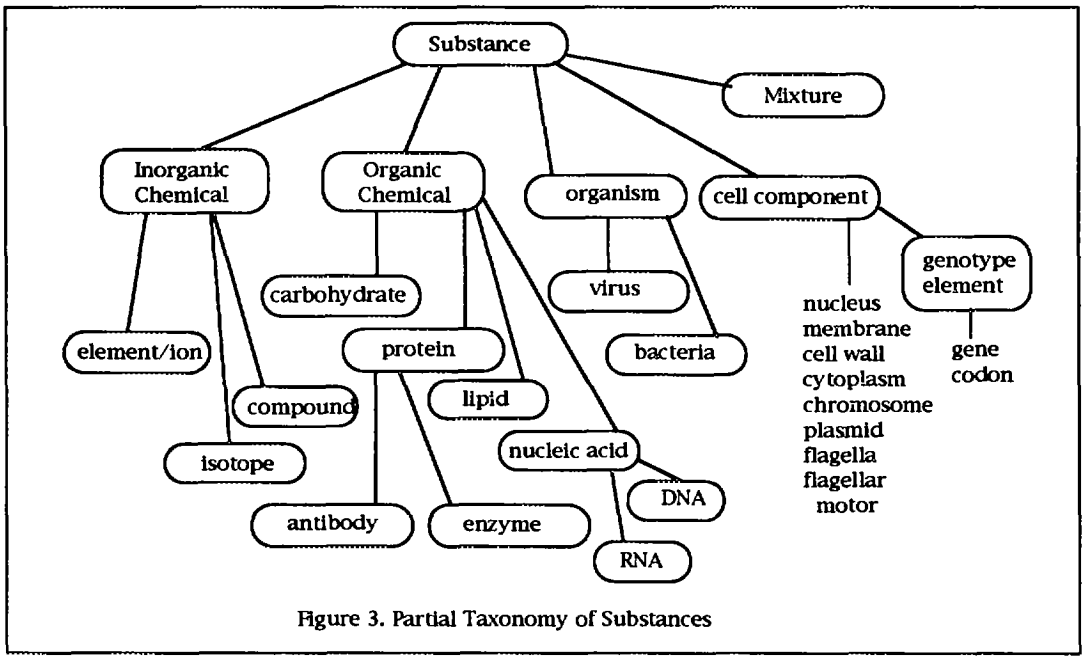

Figure 3. Partial Taxonomy of Substances

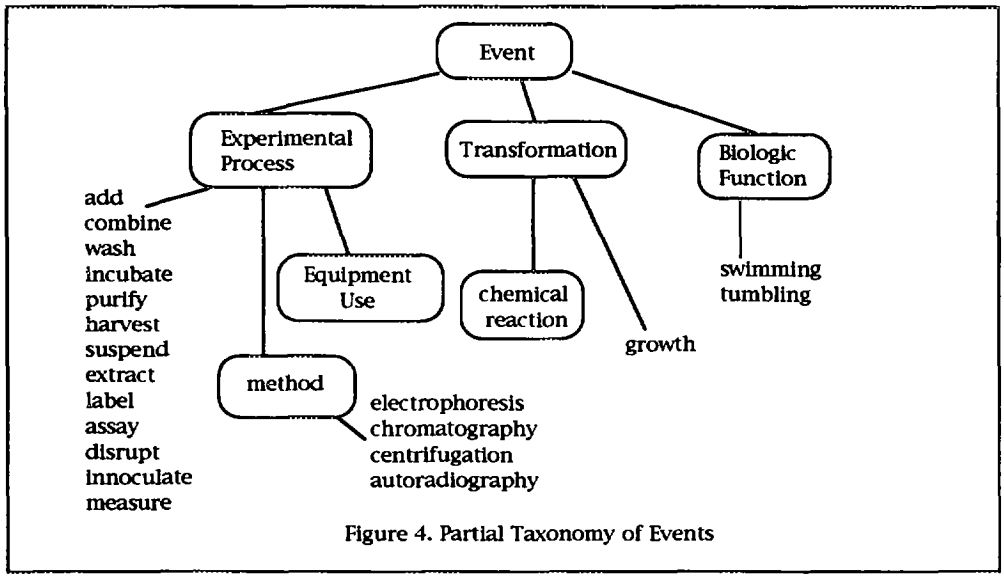

Figure 4. Partial Taxonomy of Events

inorganic chemicals, organic chemicals, organisms, cell components, and *mixtures*. Figure 3 shows a portion of the taxonomy under substance. It is similar to, but simpler than the UMLS taxonomy representing the same categories.

Figure 4 shows the most interesting part of the Materials and Methods taxonomy: the event hierarchy. Events have sub-categories including *experimental process, transformation* and *biologic function*. A biologic function is any activity performed by a biological organism, such as swimming and tumbling (in the case of bacterial chemotaxis), or within the organism (such as DNA replication and metabolism). A transformation is an event that changes the state of some substance, such as methylation of cell membrane, or break-up of intact cells into cell fragments.

A distinguishing characteristic of the Materials and Methods domain is the variety and complexity of experimental processes described. For example, there are a large number of terms used in research papers to describe various methods of combining substances: add, combine, mix, suspend, dilute, inoculate, insert, etc. Other terms such as treat, stain and label, also entail combining of substances. We have identified four basic categories of experimental processes: those that combine substances (described above); those that separate substances (remove, extract, separate, harvest); those that transform substances (incubate, disrupt, break, tether) and those that analyze information (measure, determine, assay, compute).

However, some processes such as "wash" do not fall into this simple categorization. In a wash process, buffer is first added to a substance, and then the buffer (plus some part of the original substance) is removed. Some terms, such as "purify" do not describe any specific experimental process at all, but rather identify the outcome of a process. Other terms describe complex multi-step procedures, which we call *methods*, such as precipitation, electrophoresis, and chromatography. The knowledge engineering enterprise in which we are engaged involves the analysis, for each experimental process in the knowledge model, of the entities and attributes that characterize the process, the transformations of the substances involved, and any new substances that arise from the process.

2.2. Knowledge Structures II: Frames

While taxonomy represents the overall categorization of concepts, frame structures represent the attributes of entities and events, such as the duration and temperature of an incubation process, as well as the related objects (called "role fillers") that make up a complex structure or process. Each entity or event described in the scientific text is represented by a unique "frame instance". The elements of a frame instance are:

a. The category identification (a concept from the taxonomy)

b. The unique ID of the instance.

c. A "species" slot (where appropriate)

d. Other named slots with fillers chosen from a restricted class of objects, according to the frame definition for the category.

Slots representing attributes are filled with symbolic expressions that directly represent information about the object, while slots representing roles are filled with pointers to other objects. It is customary in frame-based representation systems to treat all slots as optional, and when describing an instance to specify only those slots for which information is available. The *frame definition* for each category specifies the superclass of the category in the taxonomy, and range of fillers allowed for each slot. Some example frame definitions are the following:[1]

```
(defclass substance
   (superclass entity)
   (species <string>)
   (source <person> <organization> <reference>
   <process>)
   (attributes <property-expression> . . . ))
```

The frame definition of "substance" includes slots for the species, the source, and other attributes. "Species" is used here, not strictly in the biological sense, but to represent very numerous sub-categories such as the Strain Number of bacteria and plasmids, or the names of specific genes and proteins, without adding them to the taxonomic network. This convention, adopted from the UMLS classification scheme, prevents the taxonomy from becoming too large and difficult to manage. Since new strains and plasmids, as well as other entities such as equipment and chemical compounds, are constantly being introduced, the convention also avoids the necessity of constantly updating the taxonomy.

It is common for papers in bacterial chemotaxis to identify the source of materials used: a researcher, a laboratory, or a reference in the bibliography. Alternatively, the process which created a substance is often described, even within the "Materials" section of a paper, for example:

> The *cheW* overexpression plasmid pCW was created by inserting the *CheW* gene from pJL63 [7] into pHSe5 [12].[2]

The class definition for organism (shown below) adds a new slot, that of genotype. Genotype elements identify particular genes or chromosome sites that have been

---

[1] The notation <category> below means an object pointer or text string that refers to an object of the category mentioned.

[2] [Gegner 1991, p.750]

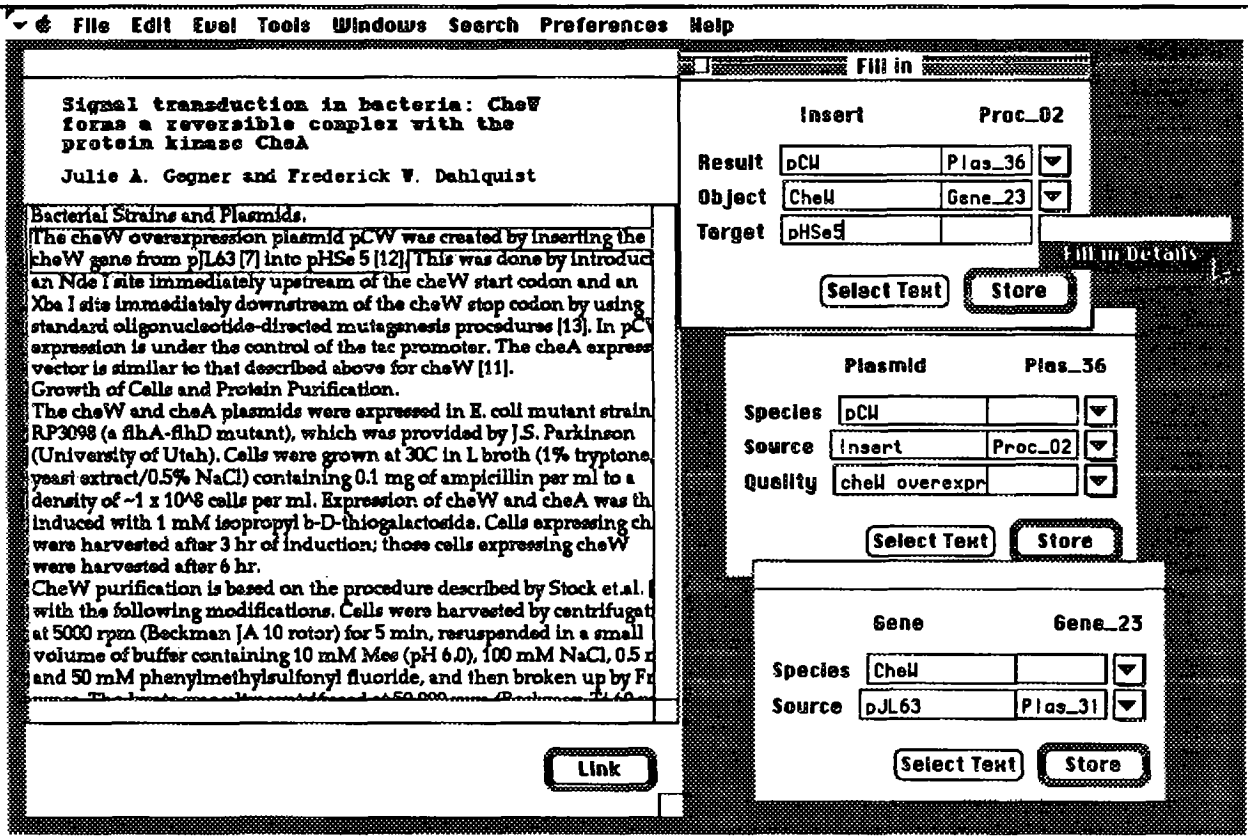

Figure 5. Defining frames for the sentence "The *cheW* overexpression plasmid pCS was created by inserting the *CheW* gene from pJL63 [7] into pHSe5 [12].

modified in the organism. The definition for organism also illustrates the use of *inheritance* in a frame-based taxonomic representation: the *organism* class implicitly includes all the information from its parent, the *substance* class.

(defclass organism
(superclass substance)
(genotype <genotype-element>))

The class definition for process includes the slots common to all processes:

(defclass experimental-process
(superclass event)
(species <string>)
(result <substance> or <transformation> . . .)
(parent <process>)
(end-test <property-expression>)
(substeps <process> . . .)
(sequence <sequence-expression> . . .)
(equipment <string> or <equipment-use>)
(manner <property-expression> . . .))

The species slot identifies specific named procedures, such as the method of Laemmli, which is a subclass of the electrophoresis method. The result of a process is to create or transform substances; the parent of a process is another process in which it is a substep; the end-test of a process describes the time duration or some other condition (such as heating to a particular temperature) that defines when the process is over; the substeps link processes to the particular methods or actions used to accomplish them. The sequence slot describes the temporal order of substeps; this information may be partially specified or omitted, since the sequence of substeps may be unknown or unimportant. Equipment and manner slots provide further specification of the process.

The process frame definition does not specify slots for the materials or substances on which the process is performed, since these vary from one process category to another. The frame definition for "insert" includes role filler slots for the substance that is inserted, and the substance or equipment into which it is inserted:

(defclass insert
(superclass experimental-process)
(object <substance>)
(target <substance> or <equipment>))

In Figure 5, we show the definition of knowledge frames for the "insert" action described in the example sentence shown above[Gegner 1991]. The frames are shown as they

are being created using the Knowledge Definition Tool described in Section 3.2

3. From Text Structures to Knowledge Structures

In order to reap the benefits of intelligent retrieval, we must populate our knowledge base with the contents of a significant body of research literature. Thus a key problem is the translation of text structures into knowledge structures, usually referred to in artificial intelligence as the problem of knowledge acquisition.

3.1 Natural Language Processing

One reason for choosing the Materials and Methods sections for our study is that they exhibit patterns that are amenable to *sublanguage analysis* techniques for natural language processing (described in Section 3.1.2). In an earlier report [Baclawski 1993a] we compared Methods and Material sections to cooking recipes: there is an initial list of materials, followed by a description of what actions were performed to transform the materials in the desired fashion. The Recipe Acquisition System developed at the University of Connecticut [McCartney 1992] applies sublanguage analysis to translate recipes into frame-like descriptions. In the DARPA sponsored Message Understanding Conferences MUC-3 and MUC-4 [Sundheim 1991, Sundheim 1992], more than 20 different research groups created special purpose natural language processors to translate news service stories into frame database structures. The sublanguage approach has also been used for processing the free-text comments written on life insurance applications describing applicants' medical treatment history[Liddy 1992] .

Our research on acquiring knowledge from biology texts is aimed at adapting the techniques used by these systems, and extending them where necessary, to the characteristics of molecular biology Materials and Methods texts. Although the text of biology research papers is much more complex than simple recipes, terrorist news reports, or medical treatment summaries, we can still exploit the patterned features of Materials and Methods sections to perform similar text-to-frame translation.

3.1.1 Lexical and Notational Complexity

Requirements for translating Materials and Methods text to knowledge frames go beyond the ability to parse ordinary English sentences. To process scientific text, specialized software must be developed to handle the complex lexical and notational conventions of each scientific domain [Futrelle 1991]. This can be illustrated by the following excerpt from a Materials and Methods section from a biology paper [Hazelbauer 1989].

> **Bacterial Strains and Plasmids.** CP177 is a derivative of OW1 (14) carrying Δ*trg-100 zdb* ::Tn5; HB789 is CP177 Δ(*cheR- cheB*) 2241; CP362 is OW1 Δ(*tar-tap*) 5201 Δtsr-7028 Δ*trg-100 zdb* ::Tn5; and CP467 is OW1 Δ*trg-100 zdb* ::Tn5 *polA12(ts)* - Lac$^+$. The plasmid pMG2 (15) contains *trg* in pUC13. In pMG1021 *trg* codons 305, 312, 319 and 501 were changed to create *trg* (4A) using procedures as described (15).

Analysis of the above text shows that we need specialized molecular biology knowledge to understand that a notation with a 'p' as in pMG2 refers to a plasmid. A name starting with two upper case letters followed by some numbers (e.g., CP177) refers to a bacterial strain. Any notation such as *cheR* in italics refers to a gene, while the same characters CheR, in plain text beginning with an upper case letter refers to a protein. The notation pMG2 (15) refers to Reference 15 citing another research paper that describes the plasmid pMG2. *trg* (4A) refers to a particular mutation of the *trg* gene.

Knowledge of the domain itself combined with the knowledge about conventions of scientific writing in molecular biology, including specialized notation, is required to understand such complicated text. Even with these notational conventions there is no accepted universal naming scheme for materials, and there are many local variations. Even experts have difficulty in interpreting complex text notations.

3.1.2 Sublanguage Analysis

Our approach, like those of the other projects mentioned above, is based on *sublanguage analysis* techniques, which focus on developing special purpose linguistic models of a particular domain of discourse [Grishman 1986]. This results in some helpful restrictions on the range of the linguistic data that needs to be accounted for in a sublanguage analyzer. At the lexical level, the sublanguage eliminates large parts of the total vocabulary of a language because the number of senses for each word that are actually used is limited and many of the words that can function as more than one part of speech probably will not. At the syntactic level, a sublanguage is characterized by a limited range of sentence forms and makes extensive use of compound nominals such as "polyacrylamide protein gels" that reflect the specialized nature of the subfield.

The most common sentence types we have observed are of the following three "normal" forms, shown here with simplified examples from the text in Figure 1:

N1. <process/transformation noun> was performed <preposition or "using"> <substance or equipment>

- Washes were performed in a buffer containing 50 mM Tris hydrochoride.
- Quantitation was performed by using a Searle γ-radiation counter.

N2. <substance> was <process/transformation verb> <preposition or "using"> <substance or equipment>

- Rabbit sera were diluted with TBSN-milk buffer.
- Filters were rinsed with TBSN buffer.

N3. <piece of information> was <analyze verb> <preposition or "using"> <process or attribute or equipment>

- Efficiency of protein transfer was determined by Ponceau S staining of nitrocellulose filters.

The sublanguage parsing problem is, in effect, to map the input text into one of the recognized normal forms, so that the correct frame structure can be built. In the general case, this mapping can be extremely difficult, requiring information from several sentences to be combined; however, sometimes it is simple and direct, as shown in the following additional examples from Figure 1:

Input:

Filters were air dried before being autoradiographed with intensifying screens.

Normal Form translation (N2):

Filters were air dried.

Filters were autoradiographed with intensifying screens.

Input:

Blocking steps, antibody incubations, and ^{125}I-protein A incubations were performed in TBSN buffer.

Normal Form translation (N1):

Blocking steps were performed in TBSN buffer.

Antibody incubations were performed in TBSN buffer.

^{125}I-protein A incubations were performed in TBSN buffer.

3.1.3 Parsing experiments

Experiments were done by considering a sample of the Materials and Methods sections from the bacterial chemotaxis corpus. A sublanguage grammar was created to parse sentences containing the verbs "measure", "determine", "compute" and "estimate". Some typical sentences from the sample set are:

- The buffering power of the medium was measured by the addition of small aliquots of 10 mM HCl or NaOH.
- The viscosity of 90% D_2O at 32° C was measured using a Cannon-Ubbelohde viscometer (Cannon Instrument Company no. 75 -L321, viscometer constant 0.00813 centistokes per second) and found to be 1.18 times that of pure water.
- Swarming rates were measured at 20° C, 32° C, and 37° C, and these rates, together with the pattern of swarm rings, were used to determine which mutant isolates corresponded to different alleles and which were likely to be duplicates.

For purposes of our experiment we simplified these sentences so that they would correspond to one of the three sentence types (N1, N2, N3) described above, and created a sublanguage grammar that builds frame-based representations using an Augmented Transition Network parser [Allen 1987]. The simplified sentences consist of a single clause containing the verb "measure" or "determine", e.g.:

- The viscosity of 90% D_2O at 32° C was measured using a Cannon-Ubbelohde viscometer.
- Swarming rates were measured at 20° C

Unlike a general-purpose English grammar which has lexical categories such as noun, verb, adjective and phrase categories such as noun phrase, verb phrase, and prepositional phrase, we identified domain specific lexical categories for the sample corpus including the following:

<analyze-verb> - measure, determine, compute, estimate
<process-verb> - wash, incubate, rinse, probe, dilute
<substance-noun> - membranes, D_2O, HCl, medium, buffer, IPTG
<process-noun> - washes, incubation
<event-noun> - rotation, swarming
<attribute-noun> - pH, trajectory, viscosity, buffering power, rate
<equipment-noun> - electrode, swarm plates, Cannon-Ubbelohde viscometer

and domain-specific grammar rules such as the following *phrase structure* rules:[1]

R1. <Sentence> => <process-noun-phrase> was performed <process-modifier>
R2. <Sentence> => <substance-noun-phrase> was <process-verb> <process-modifier>
R3. <process-modifier> => <preposition> <substance-noun-phrase>
R4. <process-noun-phrase> => <process-noun > OR <substance-noun> <process-noun>
R5. <substance-noun-phrase> => <substance-noun> OR <modifiers> <substance-noun>

which can parse sentences such as the following:

Antibody incubations were performed in TBSN buffer. (R1)

Washes were performed in 50 mM Tris hydrochloride. (R1)

Filters were rinsed with TBSN buffer. (R2)

Rabbit sera were diluted with TBSN-milk buffer. (R2)

Cells were grown in 100 mM IPTG. (R2)

Our first grammar was able to interpret about half the sentences from the sample set. This is not surprising, since a sub-language grammar must not only match the syntax of the input, but must describe the semantic relations as well. For example, the sentence "Filters were washed with TBSN buffer" matches sentence type N2, while the similar sentence "Filters were washed with 10 minutes of agitation" does not match. The failure of the sublanguage grammar to interpret the latter sentence is not due to natural language syntax, but is due to the fact that the modifier ("10 minutes of agitation") is neither a substance nor a piece of equipment. The conceptual structure for the second sentence was not defined in our first sublanguage model. Creation of sublanguage models is an iterative process in which the most frequent constructions are identified and tested on sample texts, and the failures of those test guide the extension of the model.

Given these experiments, we are encouraged by the fact that just three sentence types could account for most of the conceptual relations in a complex paragraph such as shown in Figure 1, and the fact that our first grammar could handle a significant fraction of sentences from other

[1] Phrase structure rules define a grammatical category on the left-hand side, and an acceptable composition for that category on the right-hand side. Eventually all sentence forms can be reduced to lexical categories that match specific words.

articles. We are working on "scaling up" the grammar to parse the remaining sentences in the sample set, and also to include other verbs as well.

3.2 An Interactive Knowledge Definition Tool

We would like to depend on natural language processing to automatically create accurate representations of Materials and Methods texts. However, full and accurate natural language processing is many years away. The performance of systems at the MUC conferences is about 50% accuracy, and the terrorist news stories are simpler both linguistically and conceptually than biology research papers. Although we hope to develop systems that perform better than previous efforts, it still will be necessary for scientists to have the ability to examine, correct, and hand-create the knowledge frame translations for biology research texts.

Accordingly, we have developed a simple interactive system that allows manual definition and/or correction of the knowledge frames. This can be used for annotating papers "from scratch" or for updating the frames already in the database. The Knowledge Definition Tool can be viewed as the first prototype for a complete knowledge defining environment that will guide scientists through the process of knowledge base creation and update.

The tool has a convenient menu- and window-based user interface. There are a number of systems on the market which have similar interfaces for defining knowledge frames. A good example is the Object Editor in the Nexpert Object expert-system [Arcidiacono 1988], an interactive system that allows the user to enter the frame data for instances of a specific class. The features that distinguish our system from similar tools are the following (they are explained in detail below):

- There is a list of possible values to choose from for each slot in the frame.
- A sub-frame for a role filler can be brought up immediately for any slot value which requires it.
- Since the frames we are working on are related to the text of the biological papers, each frame can be associated with a particular part of the document being described.

The structure of the frames is defined in Section 2. The knowledge definition tool includes a "template" for defining instances of each category in the ontology, which displays all the slots for that frame type. These slots include those of the specific class as well as the slots it inherits from its parents in the hierarchy.

To define a frame instance, first the category of the instance is selected from a comprehensive list, then the frame template for the class appears. To fill in each of the slots, one can either select from a list of suggested values in a pop-up menu, or type in a new value. The lists of suggested values are derived from the semantic restrictions on the slot. For example, for a slot whose filler is a chemotaxis protein the corresponding list will display those proteins (CheA, CheW, etc.).

The value of the slot can, in turn, be another object. For instance, a substance that is used as a process input, may itself be elaborated in the paper. Then it is, of course, desirable to be able to fill in its frame right away. This recursive frame definition is allowed in the system. The elaboration on the slot is done in a separate window, which itself can have another detail window, and so on.

Each frame instance is represented as an object in an object oriented database system. Each part of a document is also represented as an object in the same object-oriented database. This allows the frames to be viewed as "annotations" of the text, or alternately the text to be viewed as "documentation" of the frames.

Figure 5 illustrates the process of annotating the text of the paper with frames. To do that, the user chooses a sentence in the text and links it to one or more frames that will store the information about materials and methods described in the sentence. In the example in Figure 5, the main process of the highlighted sentence is insert. The result and target of the insertion are plasmids, and the object being inserted is a *cheW* gene. The user has already created frames for the result and the object and is about to create a frame for the target plasmid (pHSe5) by selecting "Fill in Details" from the pop-up menu at the right. When the mouse is released, a frame similar to the one for plasmid pCW will appear with the species slot already filled in (pHSe5). When the frame definitions are complete, the user links the part of the text being annotated to the frames.Then it can be used later for retrieval.

We are also developing an interactive retrieval system, called the M&M Query System [Baclawski 1993b], which uses a similar interface to retrieve documents. In the retrieval system, instead of creating frames to describe a document, users can create a (partial) frame structure to describe the information they are interested in, and the documents whose frame structures match the query are retrieved. Both systems (which share many interface elements) are at the stage of working prototypes on the Apple Macintosh computer using Macintosh Common Lisp and the WOOD object oriented database system.

4. Conclusions

In summary, in order to describe and retrieve documents using concepts instead of or in addition to specific word strings, it is necessary to have a formal, computer-understandable model of the domain knowledge of the scientist. Artificial intelligence techniques at present support the creation of very simplified models, compared to the knowledge of human experts; however, the greater the extent to which the meaning of documents can be captured in a knowledge base, the greater the opportunity to create retrieval tools capable of meaning-based search. In this report, we have described a formal knowledge model of the meaning of Materials and Methods sections of biology research papers. We have described our ongoing efforts to create tools for translating text into a formal representation of its meaning.

5. Acknowledgments.

We would like to thank Dr. Arthur Miller and Dr. Maurice Pescitelli for their generous assistance in helping us understand the biology literature and design the knowledge model. Any errors or omissions are the responsibility of the authors.

6. References

[1]. Arcidiacono, T. (1988). Expert System On-call. *PC Tech Journal* 6(11): 112-128.

[2]. Baclawski, K., N. Fridman, R.P. Futrelle and M.J. Pescitelli Jr. (1993a). Database Techniques for Biological Materials and Methods. In *1st Inter'l Conf. on Intelligent Systems for Molecular Biology*, 21-29. Washington, DC: AAAI Press.

[3]. Baclawski, K., R. Futrelle, C. Hafner, M. Pescitelli, N. Fridman, B. Li and C. Zou (1993b). M&M-Query Materials & Methods Knowledge Base and Query System. NU-CCS-93-06. Northeastern University.

[4]. Baclawski, K. and E. Smith (1994). High Performance, Distributed Information Retrieval. Technical report, NU-CCS-94-05. Northeastern University, Boston.

[5]. Bryan, M. (1988). *SGML: An Author's Guide to the Standard Generalized Markup Language*. Reading, MA: Addison-Wesley.

[6]. Futrelle, R.P., C.C. Dunn, D.S. Ellis and M.J. Pescitelli Jr. (1991). Preprocessing and lexicon design for parsing technical text. In *Proc. 2nd Intern'l Workshop on Parsing Technologies*, 31-40. Morristown, New Jersey: Association for Computational linguistics.

[7]. Gegner, J.A. and F.W. Dahlquist (1991). Signal transduction in bacteria: CheW forms a reversible complex with the protein kinase CheA. *Proceedings National Academy Sciences* 88: 750-754.

[8]. Grishman, R. and R. Kittredge (1986). Analyzing Language in Restricted Domains: Sublanguage Description and Processing. In ed. 246. Hillsdale, NJ: Lawrence Erlbaum Associates.

[9]. Hafner, C.D. (1990). A Linguistically Sound Approach to Content Analysis of Natural Language Text. In *AI Systems in Government*, 142-149. Washington, D.C.: IEEE Computer Society Press.

[10]. Hazelbauer, G.L., C. Park and D.M. Nowlin (1989). Adaptational "crosstalk" and the crucial role of methylation in chemotactic migration by *Escherichia coli*. *Proceedings National Academic Sciences* 86: 1448-1452.

[11]. Kuo, S.C. and D.E. Koshland (1986). Roles of *che* Y and *che* Z Gene Products in Controlling Flagellar Rotation in bacterial Chemotaxis of *Escherichia coli*. *Journal of Bacteriology* 169(3): 1307-1314.

[12]. Liddy, E.D., C.L. Jorgensen, E.E. Sibert and E.S. Yu (1992). A Sublanguage Approach to Natural language Processing for an Expert System. In *Information Processing & Management*, 633-645. Great Britain: Pergamon Press Ltd.

[13]. McCartney, R., B. Moreland and M. Pukinskis (1992). Case acquisition from plain text: reading recipes from a cookbook. Technical Report, TR-CSE-92-20. University of Connecticut.

[14]. Shortliffe, E. (1981). Consultation Systems for Physicians. In *Readings in Arificial Intelligence*, ed. B. Webber and N. Nilsson, 323-333. Palo Alto, CA: Tioga Publishing Co.

[15]. St. Clair, B. (1993). WOOD (William's Object Oriented Database) Documentation. Available through Internet ftp. cambridge.apple.com.

[16]. Sundheim, B.M. (1991). Overview of the third message understanding evaluation and conference(MUC-3). In *Proceedings of the Third Message Understanding Conference*, 3-16. Morgan Kaufmann.

[17]. Sundheim, B.M. (1992). Overview of the fourth message understanding evaluation and conference(MUC-4). In *Proceedings of the Fourth Message Understanding Conference*, 3-21. Morgan Kaufmann.

[18]. U. S. Department of Health and Human Services, National Library of Medicine. (1993). *Unified Medical Language System:4th Experimental Edition* .