Discovery Procedures for Sublanguage Selectional Patterns: Initial Experiments

Ralph Grishman

Computer Science Department
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012

Lynette Hirschman

Research and Development Division
System Development Corporation – A Burroughs Company
Paoli, PA 19301

Ngo Thanh Nhan

Computer Science Department
Courant Institute of Mathematical Sciences
New York University

Selectional constraints specify, for a particular domain, the combinations of semantic classes acceptable in subject-verb-object relationships and other syntactic structures. These constraints are important in blocking incorrect analyses in natural language processing systems. However, these constraints are domain-specific and hence must be developed anew when a system is ported to a new domain. A discovery procedure for selectional constraints is therefore essential in enhancing the portability of such systems.

This paper describes a semi-automated procedure for collecting the co-occurrence patterns from a sample of texts in a domain, and then using these patterns as the basis for selectional constraints in analyzing further texts. We discuss some of the difficulties in automating the collection process, and describe two experiments that measure the completeness of these patterns and their effectiveness compared with manually-prepared patterns. We then describe and evaluate a procedure for selectional constraint relaxation, intended to compensate for gaps in the set of patterns. Finally, we suggest how these procedures could be combined with a system that queries a domain expert, in order to produce a more efficient discovery procedure.

1 Introduction: The Need for Discovery Procedures

In order to analyze natural language texts reliably, a computer system requires a great deal of information about the syntax of the language, about the structure of the discourse, and about the subject matter with which the text deals. Because of the need for detailed knowledge of the subject matter, natural language systems at present are limited to handling texts within very limited domains of discourse. Once such a system has been developed, the question of **portability** naturally arises: can the system be readily moved to a new domain?

Portability involves two separate issues. The first issue is whether a large portion of the natural language system is domain-independent, so that this "core" can be used in the new application without modification. The second issue is whether the domain-dependent information required by the system can be gathered in a methodical and efficient fashion. Our paper addresses the latter

0362-613X/86/030205-215$03.00

issue. Specifically, we report on some experiments aimed at developing a semi-automated procedure for discovering selectional patterns (the local semantic constraints of language in a particular domain) from the analysis of a sample of text in that domain.

2 Sublanguage and Selection

2.1 Sublanguage

A sublanguage is a specialized form of natural language used to describe a limited subject matter, generally employed by a group of specialists dealing with this subject. Examples of sublanguages that have been studied are weather reports (Chevalier et al. 1978), aircraft maintenance manuals (Lehrberger 1983), medical reports (Hirschman and Sager 1983), and equipment failure reports (Marsh, Hamburger, and Grishman 1984). A sublanguage will generally be much more constrained than the "standard language", but it may also include extensions to the standard language, such as sentence fragments found in telegraphic-style message text.

Zellig Harris, one of the first linguists to study language use in restricted domains, defined sublanguages in terms of one particular constraint: the constraint on what words can co-occur within a particular syntactic pattern, such as a subject-verb-object structure (Harris 1968). Just as speakers of the standard language distinguish between grammatical and ungrammatical sentences, speakers of the sublanguage will distinguish between acceptable and unacceptable (meaningless) sentences, even though the unacceptable sentences may be grammatical sentences of the standard language. For example, in the sublanguage of medical records, a speaker would accept the sentence *The X-ray revealed a tumor.* but not *The tumor revealed an X-ray.*

Harris hypothesized that, for any particular sublanguage, we can define **sublanguage word classes** – sets of words that are acceptable in the same contexts (Harris 1968). For example, in the context __ *revealed a tumor,* we might find words such as *X-ray, film,* and *scan.* Such classes, even though defined on purely distributional grounds, correspond closely to the natural semantic classes that might be identified by an expert in the domain. Thus, we might label the group of words *X-ray, film,* and *scan* as a TEST class, and similarly (for medical reports) identify classes such as FINDING and MEDICATION. See section 3.2 below for discussion of an experiment verifying Harris's hypothesis. The sublanguage co-occurrence constraints, when stated between word classes, are commonly called **selectional** constraints.

2.2 Implementing Selectional Constraints

It is generally recognized that these selectional constraints play an important role in distinguishing between correct and incorrect sentence analyses. Consequently, most natural language systems incorporate some form of selectional constraints. We describe here, briefly, how these constraints are implemented in the Linguistic String Parser; more detailed descriptions are given in Grishman, Hirschman, and Friedman (1982, 1983).

The Linguistic String Parser English grammar (Sager 1981) is an augmented context-free grammar. Its principal components are a context-free grammar (stated in terms of the grammatical categories of English), a set of procedural restrictions (written in Restriction Language: Sager and Grishman 1975), and a lexicon. Adding selectional constraints to this grammar involved specifying the sublanguage classifications of words, specifying the allowed co-occurrence patterns in positive terms, and providing restrictions that check the parse tree for these patterns.

Each word in the domain vocabulary is assigned to one or more sublanguage word classes; these class assignments are recorded as part of the lexical entry for each word. Thus a lexical entry consists of each major syntactic class for a word followed by a list of attributes, including its domain subclass. Words that have two or more meanings and, as a result, occur in different contexts will be assigned to more than one word class; such words are referred to as **homographs.** For example, in the medical domain, *discharge* refers both to a patient's discharge from a hospital and to the excretion of something from the patient's body. It is classified as a noun with attributes H-VMD (medical verb for 'discharge from hospital') and H-BODYPART (for 'bloody discharge'). As a verb, however, *discharge* is classified only as H-VMD. (Only words specific to the domain receive sublanguage classes and participate in selection.)

The allowed co-occurrence patterns are specified by a set of lists in the grammar. There is a separate list for each major syntactic relation: clauses (subject-verb-object structures), prepositional phrases, prenominal adjectives, compound nouns. Each list element may have associated sub-lists; this recursive structure provides a reasonably compact specification of the allowable combination of sublanguage classes. For example, the V-S-O list gives the positive co-occurrence patterns for the VERB-SUBJECT-OBJECT relation, where the list associated with each verb class (e.g., H-SHOW below) has a sub-list of associated subject classes (e.g., H-BODYPART and H-TEST), each of which have associated lists of object types (H-INDIC, H-RESULT, H-DIAG).

```
LIST V-S-O =
  H-SHOW: (H-BODYPART: (H-INDIC, H-RESULT,
                                    H-DIAG),
           H-TEST:      (H-INDIC, H-RESULT,
                                    H-DIAG),
           ...),
  ...
```

This is interpreted as follows:

- H-SHOW verbs with H-BODYPART ("body part" class) as subject take objects of classes H-INDIC ("indicators"), H-RESULT or H-DIAG ("diagnosis"), as in *liver showed no abnormalities*;

- with H-TEST ("test") as subject, H-SHOW verbs also take the objects H-INDIC, H-DIAG, and H-RESULT, as in *test showed metastasis.*

The list imposes selection only for listed verbs, and not all verbs appear on the list. In particular, *be* and related verbs do not, since they obey a different kind of selection (between subject and object). Similarly, not all prepositions participate in selection for prepositional phrases; specifically, *of* has too broad a distribution for the statement of selectional patterns. In this way, selection is applied only to sublanguage-specific constructs, where it is possible to describe the allowed patterns with reasonable conciseness.

The selectional constraints are enforced by a set of restrictions that use the lists of co-occurrence patterns. Whenever a structure involved in selection (e.g., clause, noun phrase, prepositional phrase) is completed during a parse, a restriction is executed that compares the classes assigned to the words in the parse tree with the allowed selectional patterns for this structure. If the word participates in selection, but its associated arguments do not match the patterns on the list, then the analysis is rejected and the parser backs up to seek an alternative analysis. Because it operates on surface structure, the restriction that tests for subject-verb-object selection must take into account all the transforms of this basic structure. For clauses, the restriction checks selection for both active and passive sentences, sentences with intervening aspectuals (as in *patient continued to receive medication*), and relative and reduced relative clauses. It does this by identifying the "transformed" subject, verb, and object; it can then use a single canonical set of subject-verb-object patterns for selection.

2.3 THE VALUE OF SELECTION

Although it is generally agreed that selectional constraints are important in separating correct and incorrect analyses, we are not aware of any *measurements* of the impact of selectional constraints, particularly in text analysis (as contrasted with the analysis of natural language database queries, for example). In order to obtain a more objective measure of the importance of these constraints, we conducted an experiment comparing the effectiveness of grammars with and without selectional constraints.

The test corpus was a set of hospital discharge summaries containing 407 sentences and sentence fragments. We used the NYU Linguistic String Project medical grammar, which is a modification of the Linguistic String Project English grammar including the sentence fragments and other constructs (such as descriptions of medication dosages) that appear in medical reports but not in standard English (Marsh 1983). Each sentence was analyzed twice, once without any selectional constraints and once with selectional constraints (the selectional patterns were developed manually at NYU by linguists from a study of this test corpus and other similar medical reports). The results of each analysis were classified into one of three categories: no parses obtained; one or more parses obtained, **good** first parse;[1] one or more parses obtained, **bad** first parse. These results are summarized in Table 1.

	with selection		without selection	
good parses	308	(76%)	306	(75%)
bad parses	30	(7%)	68	(17%)
no parses	69	(17%)	33	(8%)

Table 1. Parsing results for 407 sentences, run with and without selectional constraints.

We found, in brief, that adding selectional constraints had only a marginal effect on the number of good parses. However, it greatly reduced the number of bad parses. Sentences that previously got bad parses now got no parse at all. This somewhat surprising result can be explained by noting that a certain number of sentences that had previously parsed correctly were blocked by over-constraining due to selection. For example, the phrase *herpes type lesion* was parsed successfully without selection, but failed to parse with selection, because there were no selection patterns for allowing a compound noun of the form *herpes type* (H-DIAG + H-TYPE) + *lesion* (H-INDIC). On the other hand, some sentences that received an incorrect first parse without selection received a correct parse with selection because the incorrect parse is blocked by selection. For example, *the patient had no JVD and no increase in thyroid size* parsed incorrectly without selection due to incorrect distribution of the adjunct *in thyroid size,* but correctly with selection. Overall, 21 sentences (out of a corpus of 400) changed from good to no parse and 23 from bad parse to good or acceptable parse.

Despite the fact that the number of correct parses did not show any significant increase, the use of selection produced a very substantial improvement in reliability of parses. We consider this an important benefit, for two reasons. First, in critical applications, an undetected error (bad parse) may lead to erroneous data in the data base; this is much worse than an error detected by the system (no parse). Second, if the analysis failure can be detected, it is possible to try various recovery techniques, such as employing a different analysis technique or asking the user for additional information.

2.3 THE ROLE OF SELECTION

We recognize that selectional constraints may only be the tip of the iceberg in terms of domain-specific information. Much more detailed knowledge of the domain and the structure of discourse in the sublanguage will doubtless be needed for a high-quality text analysis system. Nonetheless, we believe selection has an important role to play. As shown by the experiment just described, more than half of the analysis errors resulting from syntactic analysis can be detected using selectional constraints alone. In addition, selectional constraints are simple in structure and have been more intensively studied than most other domain knowledge; in particular, their relationship to distributional information in the sublanguage is better understood. It therefore seemed appropriate to focus on selectional constraints in our studies of discovery procedures for domain-specific knowledge.

3 Discovery Procedures

3.1 Expert vs. Text-Based Procedures

Two basic approaches have been proposed for mechanizing (or partially mechanizing) the acquisition of domain-specific information for natural language systems. One of these is based on the systematic interviewing of a domain expert, who provides information on the basic semantic classes and relations of the domain and their linguistic forms and properties. Such an approach has been incorporated into some natural language interfaces for database retrieval, such as TEAM (Grosz 1983) and LDC (Ballard, Lusth, and Tinkham 1984). This approach assumes that the domain expert has some model of the relations in the domain, and a knowledge of the different ways in which these relations can be referenced. This is not unreasonable in the database context, since the database schema can serve as a domain model (divining all the ways in which a relationship can be referenced may still be difficult, however). This approach is more difficult, however, in text analysis applications, particularly because the user may not have such a clear model of the domain semantics from which to work.

An alternative approach is to acquire some of the domain-specific information *from the text itself.* To the extent that this information is reflected in distributional relationships in the text, we can hope to extract this information by automatically analyzing a sample of text in a new domain. We have been pursuing this approach for a number of years, and describe some of our earlier efforts in the next subsection.

Although we present the expert and text-based approaches as alternatives, we do not believe them to be mutually exclusive. It may turn out that the most effective approach is a combination of these two, in which information gleaned from the text "fills in" the skeletal information provided by an expert, and the expert provides generalizations that could not easily be derived directly from the text. We shall return to this point in our concluding section.

3.2 Our Prior Work

Our previous work on discovery procedures has aimed at automating the characterization of syntactic usage and the identification of the principal semantic classes in sublanguages. Both of these procedures, as well as the procedure to be described below, start from a set of parse trees (prepared automatically or manually) for a sample text in the domain. The procedures for determining syntactic usage process the file of parse trees to extract frequency data on the use of various productions from the context-free grammar. Recent tests of this procedure on both medical records and equipment failure records indicate that accurate characterizations can be obtained from a sample of a few hundred sentences, and that (for both sublanguages) the size of the grammar used was roughly one-third the size of the full Linguistic String Parser English grammar (Grishman, Nhan, and Marsh 1984; Grishman, Nhan, Marsh, and Hirschman 1984).

The procedure for discovering sublanguage classes is based on identifying words that occur in the text in similar syntactic contexts, e.g., as subject of a given verb or as object of a given verb or as adjective modifying a given head, etc. We defined a similarity coefficient for pairs of words, based on the number of contexts the words shared. Then, using a statistical clustering procedure, we grouped together words of high mutual similarity. This procedure was successful in identifying classes containing the high frequency words of the domain (Hirschman, Grishman, and Sager 1975); the procedure was not effective for words that occurred only a few times in the sample corpus. Also, a number of false clusters were generated, due to linguistic phenomena we were able to identify, such as the omission of the head in a noun phrase. This produced anomalies in the classification, since the text contained occurrences such as *chest normal* (understood as 'chest X-ray normal') as well as *X-ray normal.* The result was a high similarity between *chest* and *X-ray* and a resulting false cluster containing *chest* with various test words such as *X-ray, film,* and *mammogram.*

4 Automatic Generation of Selectional Patterns

Determining the selectional constraints for a new sublanguage involves both determining the sublanguage word classes and determining the allowed co-occurrence patterns among those classes. In principle, both can be determined by a distributional analysis of a sample corpus. In practice, this is a labor-intensive procedure involving iteration between setting up sublanguage classes and identifying sublanguage patterns. However, in order to simplify the work during our initial experiments, we chose to separate these two tasks. We mentioned just above the experiments we had conducted earlier on discovering sublanguage classes. We describe here a complementary set of experiments to demonstrate our ability to generate co-occurrence patterns from a text sample. These experiments assume a (manual) assignment of words to sublanguage classes and aim at collecting the co-occurrence patterns and evaluating their completeness. These complementary experiments are needed to validate our techniques before we address the more difficult problem of building the selectional patterns for a new domain (see section 7 for a discussion of this issue).

Given our goal of evaluating the completeness of automatically generated patterns, our initial experiments drew on a domain where sublanguage vocabulary had already been classified. Our test corpus consisted of eleven medical reports. Six were patient documents that included patient history, examination, and plan of treatment; five were hospital "discharge summaries" which included patient history, examination, and summaries of the course of treatment in the hospital. The corpus contained about 750 sentences and sentence fragments.

In analyzing these sentences, we used the Linguistic String Project medical grammar, a modification of the LSP English grammar that had been used for processing a number of medical documents (Hirschman et al. 1981), including the discharge summaries in our corpus. The sublanguage word classes, which are recorded in our lexicon, had been developed based on the discharge summaries and other similar medical records. However, neither the grammar nor the word classes had been revised to reflect any new syntactic forms or semantic patterns that appeared in the other six patient documents; these documents were being analyzed for the first time.

The discovery procedure for selectional patterns has five principal steps:

1. generating a set of correct parses;
2. resolving homographs;
3. generating instances of selectional patterns from the parses;
4. collecting the patterns into lists sorted by syntactic construct (e.g., a list for subject-verb-object patterns, a list for head-prepositional modifier patterns, etc.);
5. final review of patterns for correctness.

We began by parsing the entire text with the Linguistic String Parser and the Linguistic String Project medical grammar, and collecting the resulting parse trees. Generation of correct patterns depends critically on having correct parses; therefore, the automatically generated parse trees had to be manually screened to select only correct parses. One possibility would have been to generate (without relying on selection) all parses for each sentence and then to choose the correct parse by hand. This would have required a great deal of work, since without selection there may be many parses for each sentence. Since we were focusing on evaluating the completeness of the set of generated patterns, rather than on the feasibility of acquiring the selectional patterns in a new domain, we chose to use the existing selection mechanism as a short-cut to getting the correct parse. This reduced drastically the number of parses that had to be screened; it did not affect correctness of the chosen parse, since the parse is or is not correct, regardless of how it is generated. However, for sentences blocked by selection, we did parse these sentences without selection and did go through the manual procedure to select the correct parse. For a number of sentences, we failed to obtain an automatically generated correct parse by either parsing method. These sentences were not included in the corpus. This procedure furnished us with good parses for about 520 sentences and sentence fragments (about 2/3 of the initial corpus).

The next step was to resolve homographs. As we noted above, some words may have more than one meaning or be used in more than one way, and thus be assigned to more than one sublanguage class. Within any particular sentence, the word was used in one of these senses and should therefore have been identified with the corresponding sublanguage class in order to produce the correct sublanguage co-occurrence patterns. Much of the homograph resolution was done automatically, by the selection mechanism. However, in certain cases, a word emerged from the processing with multiple sublanguage classes. In some cases, this was due to insufficient context resulting from omission of implicit (zeroed) information, e.g., *discharge on 1/12* would probably refer to discharge of patient from the hospital, but selection could not rule out the SYMPTOM reading of *discharge* from this limited context. A second source of unresolved homographs came from parses generated without selection, in which case there was no automatic mechanism for homograph resolution. Whatever the source, words listed as having multiple subclasses were screened and disambiguated manually: we scanned the parse trees for sentences containing multiply-classified words, and, in each case, selected manually the sublanguage class relevant to its use in that sentence.

We then proceeded to the task of extracting the sublanguage class co-occurrence patterns from the file of correct surface parse trees. Since co-occurrence patterns reflect a regularized or canonical structure (e.g., verb-subject-object relations), it was necessary to map surface structure into the normalized set of relations required for co-occurrence patterns. This involved locating the "logical" subject and object in passive sentences, relative clauses, reduced relatives, and clauses with aspectual verbs. For example, in *patient continued to receive medication,* the verb/subject/object co-occurrence pattern of interest is "receive/patient/medication".

The computation of co-occurrence patterns was done by a set of restrictions that borrowed code from those used for the selection mechanism itself. (This was possible because the selection mechanism also has to find the logical elements involved in co-occurrence, including the cases where these differ from the surface structure). An appropriate restriction (e.g., a subject-verb-object or an adjective-noun or a noun-preposition-noun restriction) identified the structures that participate in selection (subject-verb-object, adjective-noun, or noun-preposition-noun). For each such structure, the restriction located the words participating in the co-occurrence relation and retrieved the sublanguage classes associated with these words. The restriction then wrote the sublanguage class pattern onto a file. The pattern consisted of a tag identifying the pattern type (e.g., PRED-ARG1-ARG2 for verb-subject-object or NVAR-APOS for noun-adjective) and the actual words in that instance of the pattern, followed by a line for each member of that pattern, containing the major class (e.g., noun = N or past participle = VEN), followed by the word, followed by the subclass.

```
* 81A  1C. 1.11
PRED-ARG1-ARG2  EXAMINED   ()        JOINTS
VEN              EXAMINED  (H-VMD)
N               ()         (NIL)
N               JOINTS     (H-AREA)
* 81A  1C. 1.11
NVAR-APOS       JOINTS     OTHER
N               JOINTS     (H-AREA)
ADJ             OTHER      OTHER
```

The final stage involved collecting, counting and reformatting the set of co-occurrence patterns into the selectional lists required by the grammar. This permitted us to use the automatically generated sets of co-occurrence patterns as input to the selectional constraints of the grammar.

Prior to running any parsing experiments, we compared the automatically generated selection lists to the lists produced manually by a linguist. Our expectation was that the automatically generated lists would be a subset of the manually prepared set. It turned out that this was not the case, primarily because a number of human errors had allowed erroneous patterns to enter the file: errors in assigning sublanguage classes to words, errors in resolving homographs, oversights in weeding out bad parses. We therefore found it necessary in practice to make a final manual pass over the file of patterns, discarding bad patterns that had crept in in one way or another. Only then were the patterns suitable for use as data to the selectional restrictions.

Although most of the data manipulation (generation of parse trees, generation and collection of selectional patterns) was automated, considerable manual intervention was still needed to verify the processing at each stage. We shall return to this issue below.

5 Evaluation

We have evaluated the selectional patterns obtained by the procedure just described in two ways. First, we have tried to estimate how complete the set of patterns is. Second, we have compared the effectiveness of these patterns in parsing new material with that of selectional patterns generated by hand.

5.1 Growth Curves

A crucial question we wanted to answer with our experiment was whether the size of our text sample was adequate to obtain a reasonably complete set of selectional patterns. To answer this question, we plotted the growth in our sets of selectional patterns as a function of the size of the sample we have processed (i.e., the number of different patterns encountered in the first X sentences). Figures 1, 2, and 3 show the growth curves for the subject-verb-object, prepositional phrase, and adjective-noun selectional patterns.[2]

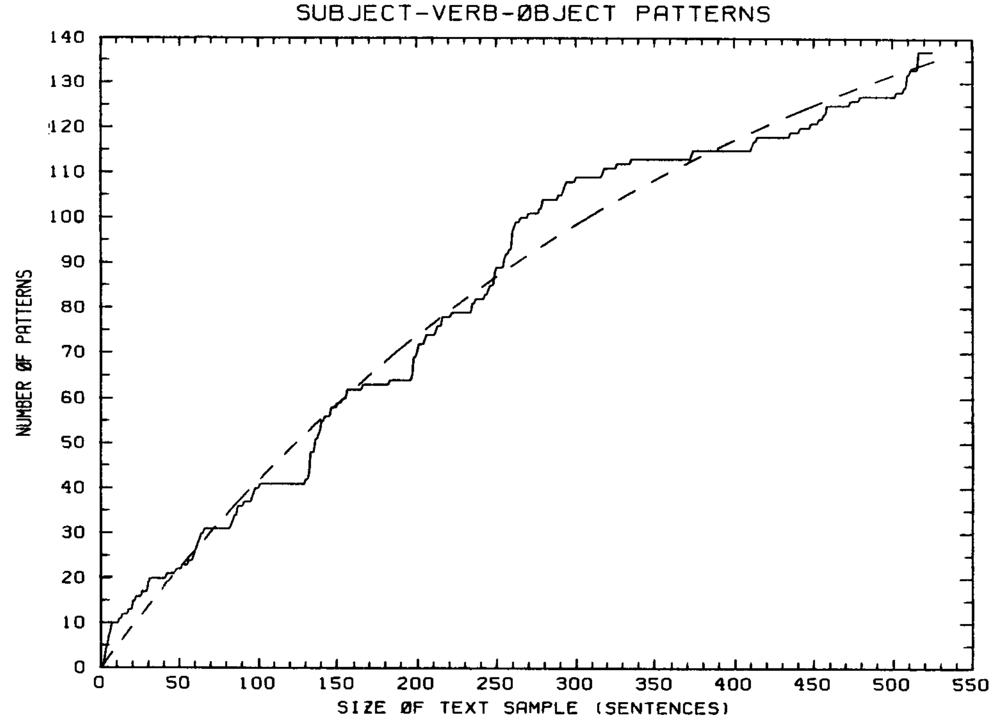

Figure 1. The growth in the number of subject-verb-object selectional patterns as a function of the size of the text sample (in sentences). The solid curve is the actual data, the dashed line the least-squares fit to a function of the form A*(1–exp(–B*x)).

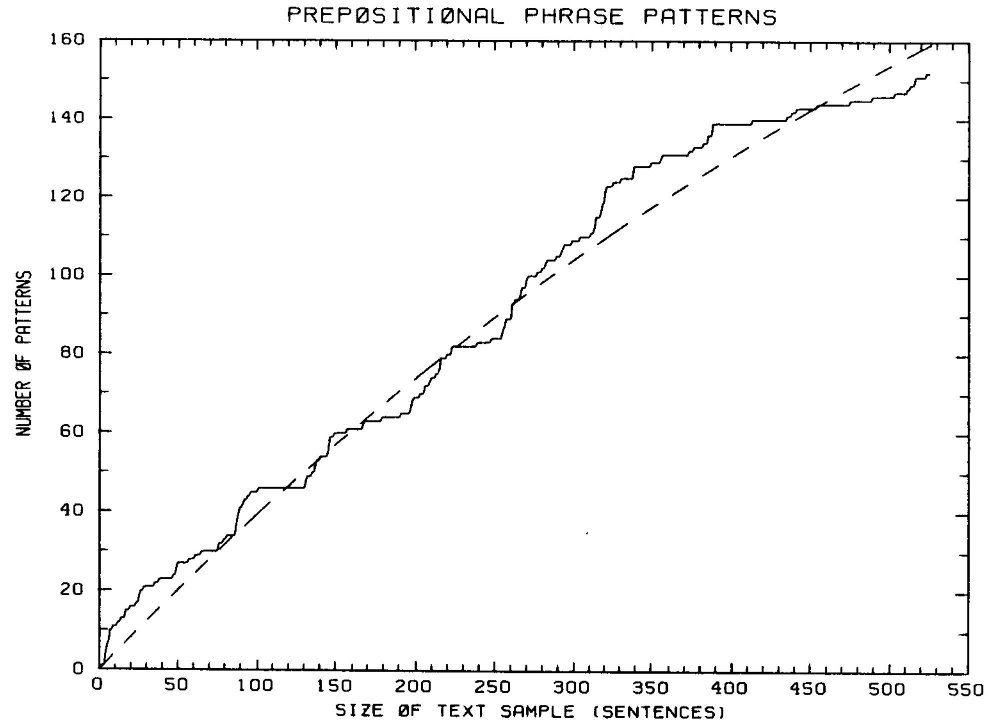

Figure 2. The growth in the number of prepositional phrase selectional patterns as a function of the size of the text sample (in sentences). The solid curve is the actual data, the dashed line the least-squares fit to a function of the form A*(1−exp(−B*x)).

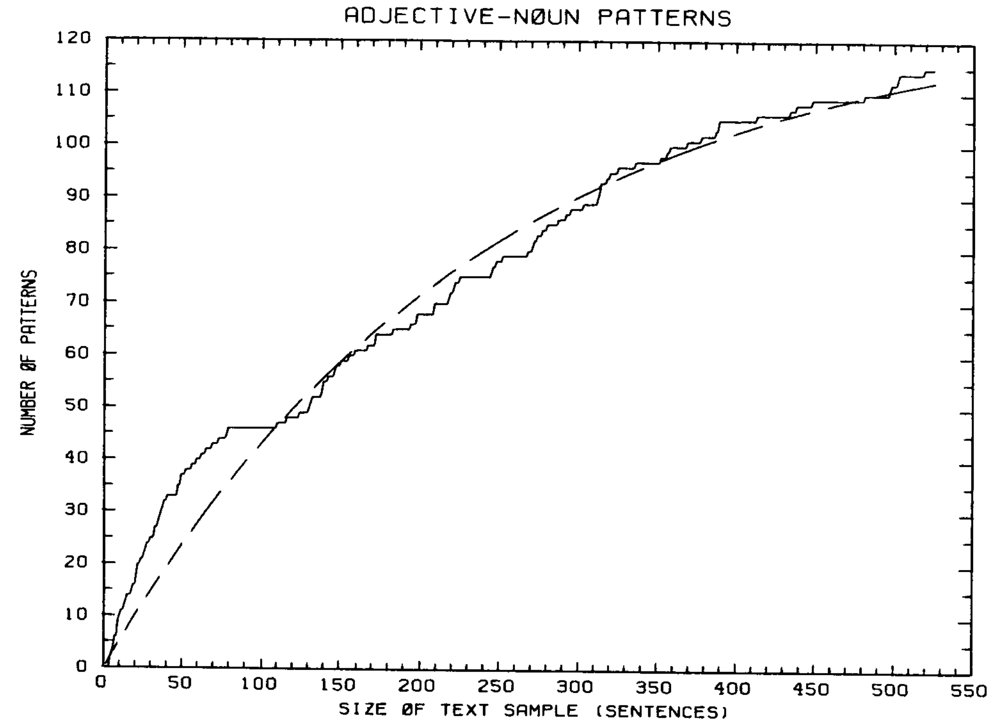

Figure 3. The growth in the number of adjective-noun selectional patterns as a function of the size of the text sample (in sentences). The solid curve is the actual data, the dashed line the least-squares fit to a function of the form A*(1−exp(−B*x)).

If our corpus had yielded a reasonably complete set of patterns, we would expect the growth curves to flatten out by the end (indicating that very few new patterns were being encountered in the text). In our earlier study of syntactic patterns in sublanguages, we had found just such an effect after 200-250 sentences. Unfortunately – as is evident in the figures – this is not the case here even after 500 sentences; the slope of the curve has clearly decreased, but it is by no means flat.

A pessimistic reader might suggest at this point that the set of selectional patterns is not closed, and that the curve will continue rising at a substantial rate until nearly all possible patterns are present. Our experience with sublanguage selection – and that of other linguists – suggests, however, that, to a first approximation, the set of patterns is closed and that, with a text sample several times larger than the present one, the curve will flatten out. In order to get a more quantitative estimate of the size of corpus that will be needed, we can use the following crude model. The successive patterns encountered in processing the sentences represent a random selection (with replacement) from a finite population (the complete set of patterns for the sublanguage). We therefore expect the growth curve to have the form $Y=A*(1-\exp(-Bs))$, where s is the number of sentences processed, A is the size of the complete set of patterns, and B is a parameter related to the rate of growth of the set of patterns. A least-squares fit of this function to the growth curve yields the following values: for subject-verb-object patterns, A=180, B=0.00264; for prepositional phrase patterns, A=276, B=0.00162; for adjective-noun patterns, A=126, B=0.00416. The fitted exponential curves are shown as dashed lines in Figures 1–3. These results can be more meaningfully viewed in terms of the size of the corpus we would need to get 90% complete patterns: for subject-verb-object patterns, about 900 sentences; for prepositional phrase patterns, about 1400 sentences; for adjective-noun patterns, about 550 sentences.

5.2 PARSING TESTS

The primary objective of our discovery procedure is to produce a set of selectional patterns that can be used in parsing further texts in the sublanguage; the ultimate test of the patterns we generate, therefore, is to use them in parsing new text and see how they affect the parsing rates. The prospects for such a test are clouded by the results of the previous subsection, which showed that the set of patterns we had collected was still quite incomplete. Nonetheless we thought it worthwhile to proceed with this second stage of evaluation.

In order to avoid the substantial labor associated with processing a new text (entering the text, entering definitions for the new words, etc.), we proceeded as follows. We took two medical records from our corpus of 11 records (about 20% of the corpus) and treated them as "new text". We reran the programs for collecting selectional patterns, using only the remaining nine records (the "old text"). We then parsed the new text using the selectional patterns thus generated, and classified the results for each sentence: good parse (first parse correct), bad parse (first parse incorrect), or no parse for the sentence.

We compared these results with the results of parsing the text using manually prepared selectional patterns. These patterns had been prepared by a computational linguist, based on a study of various medical records (including five of the records in our current corpus), generalizing from observed patterns where it seemed reasonable.

The results of the comparison are shown in Table 2. The rate of successful analyses was substantially lower with the automatically generated selectional patterns. This is not surprising given our observations in the earlier section about the incompleteness of these patterns. Where a selectional pattern is missing, an analysis will be blocked, thus generally producing no parse for the sentence.

	Selectional patterns generated:	
	manually	automatically
good parses	54	43
bad parses	25	22
no parses	27	41

Table 2. Parsing results for 106 "new" sentences, comparing manually and automatically generated selectional patterns.

The parsing rates shown here are relatively low (about 50% good parses) when compared with the data of Table 1 (about 75% success). This reflects the fact that the six patient summaries in our corpus, including the two treated here as new text, were analyzed "cold": words were added to the lexicon as needed, but otherwise no adjustments were made to the grammar or lexicon. The documents are unedited medical reports with many sentence fragments; they contain a substantial number of sublanguage-specific constructs not previously encountered in processing other types of reports, (for example, prepositions were sometimes omitted before body parts: *Synovial thickening both wrists bilaterally.*). In addition, the experiments revealed a substantial number of errors in the lexicon. Of the 63 failures (bad or no parse) using the automatically generated patterns, 20 were due to syntactic constructs not present in the grammar, 3 to other grammar or parser bugs, 11 to errors in the dictionary, and 23 to missing selectional patterns; 4 sentences got a bad parse on the first parse and a good parse on the second parse; 2 more were considered unanalyzable sentence fragments. The syntactic gaps and lexical errors uniformly depress the success rates for these experiments, but we feel that the data is still valid for comparing different sets of selectional constraints.

5.3 RESTRICTION RELAXATION

The incompleteness of semantic information is a serious and general problem that transcends our particular work on discovery procedures. As the domains with which natural language systems deal become more complex, it becomes more difficult to acquire a complete set of selectional patterns. Furthermore, in many sublanguage texts there are passages that fall outside the sublanguage; for example, in one medical record domain, there is a mention of a vacation a patient took, during which he got sick. These passages will not satisfy the selectional constraints of the sublanguage.

In the manual preparation of selectional patterns, some small measures were taken to compensate for this incompleteness. In preparing the patterns, the linguist generalized from the patterns observed in the text, adding new patterns that seemed equally reasonable based on a knowledge of the domain. For certain very common prepositions (e.g., *of*) for which it would be difficult to collect all the selectional patterns, selection was disabled. Similarly the linguist chose to omit some verbs from the selectional patterns; in these cases, subject-verb-object selection was not applied.

In the automatically generated patterns, no similar measures were taken. This, combined with the limited corpus used to gather the patterns, accentuated the effect of the incompleteness of the patterns. Because absence of a co-occurrence pattern can be interpreted as either negative information (a particular pattern is not allowed in the sublanguage) or as incomplete information (this pattern has not yet been seen), any automatically generated set of patterns will over-constrain the parsing. We therefore sought some way of automatically compensating for this incompleteness.

The approach we chose to try was **restriction relaxation.** If no parse can be obtained satisfying all selectional constraints, the parser tries for an analysis that will satisfy all but one of the selectional constraints.[3] In effect, the parser is willing to relax any one of the selectional constraints in order to get an analysis. Such an approach has been suggested before by several computational linguists (for example, Weischedel and Sondheimer 1983), although primarily to account for ungrammatical input rather than for incompleteness of semantic knowledge.

We originally applied this technique in connection with the manually generated selectional patterns. These results were not very encouraging: about 5% of the sentences in the sample that had previously gotten no parse now got a correct parse, but another 5% got a bad parse. This was not too surprising in retrospect; the various measures mentioned above to compensate for the incompleteness of the patterns resulted in a set of relatively "loose" constraints, and any further loosening (such as restriction relaxation) would let quite a few bad parses through.

Our results using this technique in connection with the automatically generated patterns, which are tighter and less complete, have been more positive (although based to date on an extremely small sample). Within our two-record sample, there were 14 sentences that had previously not gotten a parse and now got one with restriction relaxation. Of these, 10 were correct and 4 were incorrect. The automatically generated patterns, when coupled with the mechanism for restriction relaxation, did about as well as the manually generated patterns (Table 3). Given the small text sample, and the acknowledged incompleteness of the set of patterns, we found this somewhat encouraging. Of course, these experiments are still too small to reach any definite conclusions.

	Selectional patterns generated:	
	manually	automatically
good parses	54	53
bad parses	25	26
no parses	27	27

Table 3. Parsing results for 106 "new" sentences, comparing manually and automatically generated selectional patterns, and using restriction relaxation when parsing with automatically generated patterns.

6 Why Is It So Hard?

When it is first described, the discovery procedure – parse the text, extract certain syntactic structures, collect the sublanguage class patterns – may seem quite simple and straightforward. It has, however, taken us several iterations to achieve even the small success described here. It is worthwhile to reflect briefly on why this is so.

First, there are several sources of human error, each of which contributes some errors to the final set of patterns. There are errors of word classification, where the wrong sublanguage class is recorded in the lexicon. There are errors in weeding out bad parses: a small defect (e.g., incorrect conjunction scope) is easily overlooked. Finally, there are errors due to selecting the wrong subclass for a homograph. We have tried to cope with these errors by repeatedly reviewing the generated set of sublanguage patterns, going back each time to find the source of any unexpected patterns. However, as our text samples grow from thousands of words to tens of thousands (as they must to get a better set of patterns), more systematic control will be needed to minimize such errors.

Second, there are a number of linguistic phenomena that complicate the extraction of the selectional patterns. Specifically, there are cases in which the sublanguage class of a noun phrase cannot be determined from the class of the head alone. In some constructs of the form N1 preposition N2, the head N1 is "transparent", and the phrase has the class of (has the distribution of) N2. Examples are *history of ...*, *increase in ....* In other cases, the class of the phrase depends on both the head and the modifier; thus *throat* has the class BODYPART but *sore throat* the class SYMPTOM. We have incorporated the patterns and procedures for computing such phrasal attributes for the medical domain into our selectional restrictions. In moving to a new domain, we would have to acquire new sets of phrasal attribute patterns as well as selectional patterns. To limit our current experiment, however, our procedure for generating selectional patterns used the phrasal attribute patterns that had been previously developed manually.

None of these difficulties pose insurmountable roadblocks to our goal. Rather they point out that, as in any experiment where a large body of reliable data must be collected, the procedures may be complex and special measures must be taken to assure accuracy.

7 Conclusions

Overall, the experiments we have conducted using our discovery procedure are encouraging but not conclusive. The selectional patterns gathered from a limited text sample – when coupled with a procedure for restriction relaxation – do about as well as manually prepared selectional patterns. Furthermore, the growth curves for the selectional patterns suggest that a corpus several times larger would yield a more complete set of patterns and thus better performance in parsing.

We have learned that such a procedure requires substantial human interaction and we intend, before advancing to a larger corpus, to restructure the system to facilitate this interaction. The present system is basically organized for batch processing; interaction takes place by editing intermediate files. Our next step will be to move to an interactive environment that supports the following capabilities:

- isolating parse ambiguities and homographs and prompting the user to choose the appropriate reading/meaning;
- displaying new selectional patterns the first time they are encountered;
- supporting simultaneous inspection and manipulation of text, parse tree, and selectional patterns.

In all of this interaction, however, the user is still acting only as a monitor of the patterns generated. We are still faced with the difficult issue of how to bootstrap the system into a new domain. In the absence of selectional patterns, choosing the correct parse can become a tedious and time-consuming procedure, requiring extensive interaction with both a domain expert and a linguist. It is clearly not a realistic method of building up a set of patterns sufficient for semi-automated processing of the type described above.

Combining the text-based approach with elicitation procedures offers a more practical method of acquiring an initial set of domain knowledge. An expert could provide some initial word classes and a partial set of relationships, from which to generate selectional patterns. A sample of text would then provide additional examples, with the expert available to elaborate on further patterns. For example, a system being developed at BBN[4] uses a hierarchy of sublanguage classes; given a selectional pattern, it asks the user to generalize it by replacing classes with superclasses where possible. This initial period of intensive interaction with an expert would provide a sufficient pattern base so that the text-driven tools would become effective in filling in the knowledge base. Such an approach would offer the assurance of good coverage provided by a text-based system while requiring a smaller text sample than a purely text-based procedure.

Acknowledgement

This material is based upon work supported by the National Science Foundation under grants MCS-80-02453, DCR-82-02373, and DCR-82-02397.

References

Ballard, B.; Lusth, J.; and Tinkham, N. 1984 LDC-1: A Transportable Knowledge-Based Natural Language Processor for Office Environments. *ACM Transactions on Office Information Systems* 2: 1-25.

Chevalier, M. et al. 1978 TAUM-METEO: Description du Système. Groupe de recherche en traduction automatique, Université de Montréal.

Grishman, R. and Hirschman, L. 1982 Natural Language Interfaces Using Limited Semantic Information. NSF Final Report, New York University, New York, New York.

Grishman, R.; Hirschman, L.; and Friedman, C. 1982 Natural Language Interfaces using Limited Semantic Information. *Proceedings of the 9th International Conference on Computational Linguistics.* Prague, Czechoslovakia: 89-94.

Grishman, R.; Hirschman, L.; and Friedman, C. 1983 Isolating Domain Dependencies in Natural Language Interfaces. *Proceedings of the Conference on Applied Natural Language Processing.* Santa Monica, California: 46-53.

Grishman, R.; Nhan, N. T.; and Marsh, E. 1984 Tuning Natural Language Grammars for New Domains. *Proceedings of the Conference on Intelligent Systems and Machines.* Rochester, Minnesota: 342-346.

Grishman, R.; Nhan, N. T.; Marsh, E.; and Hirschman, L. 1984 Automated Determination of Sublanguage Syntactic Usage. *Proceedings of COLING84 (Tenth International Conference on Computational Linguistics).* Stanford, California: 96-100.

Grosz, B. 1983 TEAM: A Transportable Natural-Language Interface System. *Proceedings of Conference on Applied Natural Language Processing.* Santa Monica, California: 39-45.

Harris, Z. 1968 *Mathematical Structures of Language.* Wiley Interscience, New York, New York.

Hirschman, L.; Grishman, R.; and Sager, N. 1975 Grammatically-based Automatic Word Class Formation. *Information Processing and Management* 11: 39-57.

Hirschman, L.; Story, G.; Marsh, E.; Lyman, M.; and N. Sager. 1981 An Experiment in Automated Health Care Evaluation of Narrative Medical Records. *Computers and Biomedical Research* 14: 447-463.

Hirschman, L. and Sager, N. 1983 Automatic Information Formatting of a Medical Sublanguage. In Kittredge and Lehrberger: 27-80.

Kittredge, R. and Lehrberger, J., Eds. 1983 *Sublanguage: Studies of Language in Restricted Semantic Domains.* Series of Foundations of Communications, Walter de Gruyter, Berlin: 27-80.

Lehrberger, J. 1983 Automatic Translation and the Concept of Sublanguage. In Kittredge and Lehrberger.

Marsh, E. 1983 Utilizing Domain-Specific Information for Processing Compact Text. *Proceedings of Conference on Applied Natural Language Processing.* Santa Monica, California: 99-103.

Marsh, E.; Hamburger, H.; and Grishman, R. 1984 A Production Rule System for Message Summarization. *Proceedings of the 1984 National Conference on Artificial Intelligence.* Austin, Texas.

Sager, N. 1981 *Natural Language Information Processing: A Computer Grammar of English and its Applications.* Addison-Wesley, Reading, Massachusetts.

Sager, N. and Grishman, R. 1975 The Restriction Language for Computer Grammars of Natural Language. *Communications ACM* 18: 390-400.

Weischedel, R. M. and Sondheimer, N. K. 1983 Meta-rules as a Basis for Processing Ill-Formed Input. *Journal of Computational Linguistics* 9: 161-177.

Notes

1. "Good parses" included some parses that were not entirely correct but that were good enough so they did not cause errors in the process that converted the parsed trees into information formats (a structured data base).
2. The curves are rather jagged because the reports are divided into sections containing different types of information; when we begin processing a new section, new patterns are encountered, and there is therefore a sharp rise in the growth curves.
3. If no analysis can be obtained by relaxing one restriction, the parser is able to try for an analysis that relaxes two, three, or more restrictions. Our experiments have indicated, however, that relaxing more than one restriction produced bad parses more often than good ones.
4. Private communication with M. Bates.