Comput. & Maths with Appls. Vol. 9, No. 1, pp. 45–58, 1983
Printed in Great Britain.

0097-4943/83/010045-14$03.00/0
Pergamon Press Ltd.

SEMANTIC PROCESSING OF TEXTS IN RESTRICTED SUBLANGUAGES

RICHARD I. KITTREDGE

University of Montreal, Montreal, Canada H3C 3J7

Abstract—Practical results in information retrieval and automatic translation have recently been achieved for naturally-occurring texts in certain narrow technical areas. For each application, the processing system must exploit the distinctive linguistic properties of the appropriate sublanguage; in fact, a precise description of these properties, incorporated into a sublanguage grammer and lexicon, is what enables the system to build a representation of the information (meaning) conveyed by the text.

Sublanguages which appear insufficiently closed for semantic processing often carry an important component of information which is encoded in a linguistically well-behaved way and is hence computationally separable. By way of illustration, a procedure is outlined for processing stock market reports into a predicate-argument representation of their content, for that part of the report which refers to the stock exchange activity. The procedure may have applications beyond information retrieval, in particular to the synthesis of informative stock market reports in one or more languages.

1. SEMANTIC PROCESSING OF "REAL" TEXTS

Computational linguistics as a (more or less well-defined) discipline can now be considered about 30 yr old (the first experiments in machine translation were carried out in the early 1950s). But is is only in the last few years that significant advances have occurred in processing the content (or meaning) of texts.

Substantial progress has been made both in constructing theoretical models for the meaning representation of texts and in implementing these models in experimental computer systems. In the early 1970s, impressive semantic capabilities were demonstrated in systems whose input was restricted to examples constructed by the experimenters. But since then it has proved quite difficult to extend those results to large samples of "real" (naturally occurring) texts, such as those which must be processed in many commercial applications. The reason for this seems to be that no powerful semantic model has been worked out in sufficient detail to accommodate the overwhelming variety of words and structures that one typically finds in arbitrary real texts.

In certain application areas, the problem of incomplete semantic modelling can be partially circumvented. For example, in the case of systems for querying restricted data bases, a "semantic grammar"[1, 2] can be set up to describe and interpret a subset of sentences which is adequate for the particular purposes of the system. Each sentence pattern recognized by the system is formulated in terms of semantic word classes, a fact which greatly reduces the possibility of misunderstanding queries. During dialogs with a human user the system provides instant feedback which helps the user to stay within the predefined limits. For example, when the system receives queries which are not formulated in accordance with its grammar or vocabulary, it may guide the user to rephrase his input. Human linguistic performance is therefore constrained in the direction of the system's capacities.

In other application areas, however, there may be no possibility of reformulating the natural language input. This is typically the case in automatic translation and information retrieval from documents, where the wide variety of semantic problems posed by real texts must be tackled head-on. Because of this, there is a growing consensus among researchers in these areas that (a) only texts from highly restricted domains will be amenable to semantic processing in the near future, and that (b) any practical system must be based on a thorough empirical description of the language as it is actually used in the subfield in which the texts originate.

In this paper we set out to do two things. First, we summarize briefly some recent results in the semantic processing of real texts for the purposes of automatic translation and information retrieval. These results illustrate the needs for restricting the domain and for carrying out a detailed linguistic analysis within the appropriate sublanguage. Second, we outline a procedure for automatically deriving semantic representations of texts in certain restricted sublanguages. To illustrate the procedure, we give an example of the analysis of a stock market report into a predicate-argument representation of the data contained in the report. Our illustration

suggests direct application to problems of information retrieval from texts. But since many of the individual steps are reversible in principle, it also suggests how one might approach the problems of automatic translation and automatic synthesis of text from data, at least in such restricted sublanguages.

2. THREE PRACTICAL APPLICATIONS OF SEMANTIC PROCESSING TO REAL TEXTS

Computational linguists are achieving some initial successes in processing the content of technical sublanguages by basing each applied system on the linguistic analysis of a large corpus of representative texts. Before discussing the methodology of this approach, we survey briefly the scope and limitations of sublanguage processing of three kinds of text.

2.1 *Automatic translation of weather bulletins*

Automatic translation may have been the earliest practical goal of computational linguistics but it was not until recently that translation systems began to actually ease the load on human translators. One of the most successful cases has been the TAUM-METEO system, developed at the Université de Montréal, which since 1977 has been translating English weather bulletins into French of 10,000 words/day for the Canadian environment ministry[33].

METEO is designed to translate only those sentences in weather bulletins which are in telegraphic style such as (1).

(1) RAIN OCCASIONALLY MIXED WITH SLEET TODAY CHANGING TO SNOW THIS EVENING.

This is because more than 99% of bulletin sentences conform to this style and those sentences can be translated with virtually no errors. But weather bulletins may occasionally contain non-telegraphic sentences such as (2).

(2) PERSONS IN OR NEAR THIS AREA SHOULD BE ON THE LOOKOUT FOR THESE SEVERE WEATHER CONDITIONS AND WATCH FOR UPDATED WARNINGS.

In the presence of dangerous or unusual weather conditions, forecasters tend to abandon telegraphic style and resort to full sentence forms. The METEO parser rejects such sentences; instead, the system sends them to a terminal where a human translator provides the French equivalents, which are then inserted into the computer-translated text to give the complete French bulletin.

It is no coincidence that METEO, one of the most reliable systems for automatic translation, is limited to one of the most restricted, stereotyped sublanguages known. Successful translation depends on the fact that weather reports normally carry only a few kinds of information, and this information is encoded linguistically in very predictable ways, both in English and in French. The two languages have similar telegraphic styles in their respective sublanguages. Even if words cannot be mapped one-to-one between the two sublanguages, the semantic word classes and relations between classes define structures which are roughly isomorphic. The linguistic predictability which the system exploits in normal texts breaks down only in sentences where unusual kinds of information are being conveyed. In fact, the occasional shift from telegraphic to non-telegraphic style is an unmistakable sign of a shift from normal to abnormal (i.e. less predictable) information type.

2.2 *Automatic translation of aircraft maintenance manuals*

A second, far more difficult type of technical text has been the subject of a 5-yr research and development effort in automatic translation at the Université de Montréal. The TAUM-AVIATION system[4] is designed to translate English aircraft maintenance manuals into French in the field of aviation hydraulics The sublanguage of these manuals is linguistically quite complex, with a vocabulary of over 10,000 words (not counting proper or compound nouns) and a wide variety of problematic syntactic structures.

The domain of reference of hydraulics manuals is more complex than that of weather forecasting by several orders of magnitude. The possible physical objects which must be named

in these texts number in the millions and their possible functions are also quite varied. As a consequence, the system of noun compounding is quite rich. For example, a general grammar of English will permit the "empilage" in (3) to be parsed in many ways, including (3a) and (3b):

(3) wing fold logic tree diagram
(3a) ((wing fold) ((logic tree) diagram))
(3b) (((wing fold) logic) (tree diagram))

where the left member of each parenthesized pair is taken to modify the right member.

A related and equally serious problem, which may intersect the empilage problem, concerns the scope of conjunction. The string of words in (4) may be taken to denote one (4a) or two (4b) separate objects:

(4) swivel joint and door hinge center line
(4a) (((swivel joint) and (door hinge)) (center line))
(4b) (swivel joint) and ((door hinge) (center line))

The proper analysis, and hence translation, of these compounds presupposes that we can establish a small set of semantic noun-noun relations which are at least partially domain-dependent. Although much progress has been made towards discovering and representing such relations, general and complete solutions to these and other problems do not appear imminent.

In recent tests[5] the TAUM-AVIATION system demonstrated the ability to produce an acceptable translation for somewhat more than half of a new 200-page text for which only the vocabulary list had been seen in advance. A small percentage of the remaining sentences were mistranslated, the others failing the parse. Output quality for translated sentences was judged at roughly 80% that of a first-draft human translation.

In view of the complexity of the domain, it is perhaps surprising that these texts should be relatively amenable to automatic translation. That this is so appears attributable to the fact that the domain is quite well-defined. The sets of objects, categories and relations in the domain are viewed from a similar functional perspective by technicians (whatever their language) and this coherent, precise view of a particular subworld is reflected in the structure of the language used. Whether for reasons of logical necessity or professional contact, the style for presenting maintenance procedures and system descriptions is also quite similar in English and French.

2.3 *Information retrieval from medical texts*

Many of the same challenges that impede progress in automatic translation also show up in research aimed at retrieving information from scientific and technical documents. In both cases real texts must be analyzed into content representations which are appropriately structured and sufficiently nuanced for the purpose at hand. Moreover, both automatic translation and information retrieval must deal with the analysis of continuous texts (as opposed to dialogs), and thus face the same set of primary linguistic problems (e.g. scope of conjunction and modification).

In one respect, however, the work in information retrieval faces a problem not encountered as such in automatic translation. Of primary concern for information retrieval is a way of comparing (and contrasting) the information (meaning) of different sentences from one or more documents in a functionally homogeneous set, and in storing together those units of information which have the greatest similarity among them. These requirements have been favorably met by the development, over the past decade, of the notion of INFORMATION FORMAT as a linguistically justified encapsulation of text content. Instrumental in the evolution of this notion has been the work of Sager *et al.*[6–8] on the information formatting of texts in certain narrow sublanguages of pharmacology and medicine.

Basically, an information format is a tabular structure in which each row represents the information contained in a simple sentence or a part of a sentence which corresponds semantically to a simple proposition. A single text sentence may correspond to one or many rows in a format. The theoretical origins of information formats can be found in Harris' work on discourse analysis[9], including the use of grammatical transformations (or their inverses) to

	PATIENT	PT-STATUS								
		V-PT	BODY-PART	FINDING						
				TEST		V-SHOW-NEG	RESULT			
				LAB	EXAM-TEST		NORMALCY	QUANT	QUAL	
									SIGN-SYMPTOM	LAB-RES
1.			[urine]	urin-alysis		showed no			abnormalities	
2.			lungs			revealed			bilateral rhonchi	
3.			abdominal		felt	no			masses	
4.			liver		palpable			4 cm		
5.			right lung		to per-cussion		clear			

[ ] Material in square brackets reconstructed from other entries in the row.

Sentence 1: Urinalysis showed no abnormalities.
Sentence 2: Lungs revealed bilateral rhonchi.
Sentence 3: No abdominal masses felt.
Sentence 4: Liver palpable 4 cm.
Sentence 5: Right lung clear to percussion.

Fig. 1. Partial information format illustrating syntactic variations in the TEST-RESULT relation. (From Hirschman and Sager "Automatic Information Formatting of a Medical Sublanguage".)

decompose one complex sentence into two or more elementary sentences (i.e. format entries). But Sager's work has considerably refined the formatting procedure and developed it for the purposes of retrieval.

Figure 1 gives a simple information format, taken from recent paper by Hirschman and Sager[8]. Note that each word of the five formatted sentences is assigned to a column in the format in such a way that semantically similar words in structurally dissimilar sentences are aligned under the same heading. As a result, the constituents of sentences 3 and 5 are not in original order and some row-column positions are left empty. Columns are grouped together hierarchically under larger headings. What the format does, in a sense, is provide a maximal framework in which to fit all the sentences of a certain class. The class may be defined in terms of distributional regularities, but the members have a semantic unity in terms of underlying relations.

Sager *et al.* have developed a number of techniques for mapping texts into information formats. Texts are first analyzed syntactically using a general English parser[10] which is based on Harris' string grammar[11]. Most sentences receive multiple analyses, but these are then filtered by a "restriction grammar", which embodies a set of word co-occurrence restrictions valid only for the given sublanguage. (The restrictions state which semantic classes of nouns may serve as logical subject of which semantic classes of verbs, which adjectives may modify which nouns, which adverbs may modify which adverbs, etc.). The semantically characterized lexical restrictions for the sublanguage are usually compatible with only one of the syntactic parses. The output of the parse is therefore a grammatical structure for each input sentence, where each word in the structure is tagged with the labels of the semantic subclasses to which it belongs. These word subclasses can then be used to map the sentence onto the information format. Before this can be done, however, the sentence structures identified by the parser (and restriction grammar) must be put into a more canonical form. This essentially requires removing the effect of any grammatical transformations (e.g. passive sentences are converted to active form; nominalized clauses are replaced by the corresponding full sentences).

Experiments have been conducted in automatically mapping sentences of various kinds of medical texts onto information formats[12]. In general more than one format must be used to represent the content of an entire text. Once formatted, a text may serve as a kind of relational data base for purposes of automatic question answering or statistical analyses. It has proved possible to summarize information on hospital patients by formatting doctors' radiology reports and discharge summaries[13].

3. SUBLANGUAGE

3.1 *The importance of word co-occurrence patterns*

Each of the applied systems cited in the preceding section is oriented towards a particular processing goal, and limited to the particular sublanguage associated with a single knowledge domain. In order to process the content of a text, three things are required: (1) representations of meanings which make clear and computationally accessible both the differences and the similarities (e.g. equivalence, consequence, etc.) between meanings required by the processing goal, (2) ways of associating with each input word string a set of possible meanings on the basis of meaning representations for individual words and some combinatory rules which can operate on word representations, and (3) ways of isolating intended meanings from among the possible ones on the basis of axioms of common non-linguistic knowledge, some of which may be particular to the domain.

On the first of these requirements, very little is known of a general nature since only a few practical systems for real texts have incorporated anything near to a satisfactory general solution to the meaning representation problem. More can be said about the second and third requirements, at least in the case of sublanguage processing, since successful real-text semantic processors have relied heavily on a precise grammar of the possible elementary sublanguage sentence patterns, formulated in terms of the classes of words that are actually found in equivalent environments in a corpus of texts.

At the moment, linguists are quite incapable of specifying the semantic restrictions on word co-occurrence for the language as a whole, and it is not even clear that this is a worthwhile goal, for to do so would amount to an attempt to delineate what can be said in the language. But the situation is quite different in relatively fixed scientific and technical sublanguages, where there are fairly sharp restrictions on what is "sayable" (meaningful), at least with respect to the primary subject matter of the technology or science. The members of a given technical community share certain knowledge about sets of objects, their properties, and possible relations between them that consititute the common domain of discourse within the community. These common conceptual categories are directly reflected in the semantic word classes and grammatical configurations of these classes, found in a sample of texts in the sublanguage. A distributional analysis of a corpus of texts puts words into functionally similar equivalence classes that happen to mirror the accepted taxonomy of the associated subworld. A grammar of the sublanguage, when stated in terms of the semantic word classes, reflects the possible relationships between objects. It is important to realize that a precise study of a sublanguage grammar can thus reveal an important part of the structure of knowledge of the subworld.

3.2 *Factors giving rise to sublanguages*

Sublanguages have been characterized in various ways, but there is no widely accepted definition of the term. There is, however, a consensus as to the factors which are usually present when a subset of a natural language is restricted enough for efficient semantic processing[14].

- RESTRICTED DOMAIN OF REFERENCE. The set of objects and relations to which linguistic expressions can refer is relatively small.
- RESTRICTED PURPOSE AND ORIENTATION. The relationships among the participants in the linguistic exchange are of a particular type and the purpose of the exchange is oriented towards certain goals.
- RESTRICTED MODE OF COMMUNICATION. Communication may be spoken or written, but there may be constraints on the form because of "bandwidth" limitations (e.g. telegraphic style).
- COMMUNITY OF PARTICIPANTS SHARING SPECIALIZED KNOWLEDGE. The best, canonical examples of sublanguages are those for which there exists an identifiable community of users who share specialized knowledge and who communicate under restrictions of domain, purpose, and mode by using the sublanguage. These participants enforce the special patterns of usage and ensure the coherence and completeness of the sublanguage as a linguistic system.

3.3 *Sublanguages as infinite subsystems*

Harris has noted that sublanguages resemble mathematical subsystems in that they are sets

closed under certain grammatical transformations [15]. Thus, for example, if the sublanguage of analysis in mathematics contains sentence (5a), it will also contain many others including (5b–f), which differ from (5a) only by grammatical changes which leave invariant the meaning relationship of "content words" (i.e. nouns, verbs, adjectives):

(5a) This theorem provides the solution to the boundary value problem.
(5b) It is this theorem that provides the solution to the boundary value problem.
(5c) What this theorem does is provide the solution to the boundary value problem.
(5d) The solution to the boundary value problem is provided by this theorem.
(5e) Does this theorem provide the solution to the boundary value problem?
(5f) This theorem does not provide the solution to the boundary value problem.

Certain sublanguages may not use all the grammatical transformations of the whole language, but most are closed under one or more recursively applicable operations (such as conjunction or relative clause formation). Since there is no limit in principle to the number of applications of such operations, it follows that most sublanguages are infinite sets of sentences (for the same reasons that whole languages are).

3.4 *Sublanguages as imperfectly homogeneous systems*

The very notion of sublanguage is introduced on the assumption that certain subsets of the language have special characteristics (regularities) that are not discernible in the language as a whole. But this appearance is a matter of degree. As we have already seen, even weather reports are not perfectly homogeneous, showing occasional departures from the familiar telegraphic style. Under ususual conditions the domain of reference can be extended and viewed from a different perspective. As a consequence, the set of linguistic forms used is also expanded to include forms which bear little resemblance to those habitually employed. Fortunately, separating the "habitual" sentences of weather reports from the "emergency" sentences is a simple task for a parser, because telegraphic sentences obey special constituent structure rules.

But sublanguages which are less stereotyped than weather bulletins may also have non-homogeneities of style or grammatical structure which can still present problems during computational treatment. Preliminary indications are that these linguistic singularities can be correlated with a shift of subject matter or viewpoint within the text. To the extent that the non-homogeneities can be detected automatically, we may improve the performance of semantic processing programs by calling up different sub-programs to operate on the separable components of the text. When, as often happens, only one component (i.e. the more homogeneous portion of the text) is computationally tractable, the information carried in that component may still be of interest even without the information of the less accessible remainder. The next two sections are devoted to a particular sublanguage where this is in fact the case.

4. STOCK MARKET REPORTS

4.1 *Two worlds of reference*

The sublanguage of daily stock market summaries affords a simple, yet revealing case study of the relationship between language and information. In the most common variety of these reports, we can distinguish two principal domains of reference:

THE PRIMARY DOMAIN—one or more stock exchanges and the trading activity (price changes, volume of shares traded, halt in trading, etc.) taking place during well-defined business hours (e.g. 10 a.m.–4 p.m. on the Montreal Stock Exchange).

THE SECONDARY DOMAIN—the less clearly defined world of economic and political events in which the causes of market changes can be perceived. Included in the secondary domain are other relatively well-defined sites of economic activity which bear a resemblance to the stock exchanges (e.g. the gold market, bond markets, commodity exchanges, etc.). Also included are wars, strikes, nuclear power plant accidents—in short, nearly any event of interest to investors.

Stock reports come in different varieties, depending on the expertise of the intended reader. Reports from some sources may refer only to the primary domain. In contrast, highly analytical reports may be concerned more with the economy in general than the stock market itself, treating the latter mainly as a barometer of the former. Reports of the kind considered here may be called INFORMATIVE, in that they describe the day's trading activity, interspersed with a certain number of comments about events in the outside world (i.e. the secondary domain). In such reports, the primary domain is normally viewed from the perspective of a single stock market. Reference to other markets is usually made in a way which reveals the relative importance and causal relationships between movements on the separate markets.

4.2 *Grammatical subordination reflects separation of domains*

Stock market reports have an interesting and useful property which can be exploited during semantic processing. The semantic division between the two major domains is reflected in the sublanguage syntax in the way subordination is used. To a large extent, text segments which refer to market activity constitute non-subordinate (independent) clauses. Text segments referring to the outside world usually occur in grammatically subordinate structures.

Grammatical subordination of propositions is usually indicated by one of the following five devices in this sublanguage (in each case, the italicized portion encodes a proposition which is considered to be grammatically subordinate to the remainder):

● CLAUSE INTRODUCED BY SUBORDINATING CONJUNCTION

(6) Seabord World Airlines plunged 4 1/2 to 12 5/8 *after Flexi-Van Corporation disclosed it had abandoned plans to take over the airline for about* $18.25 *a share.*

● COMPLEMENT OF NOUN IN THE CLASS N-news

(7) C.I.T. climbed 9 3/4 points *on rumors of an impending merger offer.*

● NON-RESTRICTIVE RELATIVE CLAUSE

(8) Superior Oil, *which had been hit by profit taking recently*, rocketed ahead 15 to 480.

● SENTENCE OR NOMINALIZATION AS COMPLEMENT OF VERB OR PREPOSITION

(9a) The advance occurred despite *a fairly sharp rise for short-term rates in the credit market.*

(9b) Analysts said *a number of concerns are weighing on the market.*

● NON-INITIAL SENTENCE IN A "COMPANY NEWS" PARAGRAPH (a sublanguage-specific device—certain paragraphs at the end of a report give trading activity in shares of single companies, with explanations)

(10) Reliance Electric held steady at 58. *The Federal Trade Commission has indicated that it will try to block Exxon Corporation's* $1.17 *offer for Reliance.*

Certain subordinate constructions may also serve to downplay one primary domain event to a second such event because of remoteness in time, space, etc. For example:

(11a) The MSE industrial index was down a fraction *while the Toronto composite index held a small gain.*

(11b) *The continuing downturn on Wall Street* pulled Canadian stock markets lower in the early going today...

Such occurrences, which can be distinguished on the basis of their formal properties, obscure an otherwise strong tendency to correlate subordinate constructions with secondary domain reference. In any case, one rarely finds secondary domain references in independent clauses. We are therefore motivated to distinguish a sub-sublanguage within stock market reports. This "core" sublanguage has interesting linguistic properties which can be exploited for computational purposes.

4.3 *Properties of the "core" sublanguage of stock market reports*

If we remove from a typical stock market report the portion which refers to the outside world (plus any subordination connectives such as *on news of*) the remaining portion can still be read as a coherent text. (When excising a nominalization, however, we must leave behind a pronoun). Let us refer to the sublanguage of stock market reports as L_s and to the "core" component which refers to the primary domain, as l_s. It turns out that l_s has a number of properties which make it much more tractable computationally than L_s as a whole. The lexicon of l_s is far simpler and more closed than that of L_s. The number of semantic word classes needed for a grammatical description is smaller, the words fall more neatly into distributional classes, and words in the same class have greater semantic homogeneity (and thus more predictable meanings). This is of course natural in view of the fact that l_s refers to a far more tightly constrained domain than does L_s in general.

The grammar of l_s is simpler and more predictable than that of L_s. Verbs denoting value change (*climb, jump, turn higher*) have corresponding event nominalizations (*climb, jump, upturn*) which are semantically regular. There are relatively few basic sentence patterns, describable in terms of word classes which are semantically homogeneous. To these patterns correspond the few basic kinds of information carried by l_s. What is striking about l_s, however, is the rich variety of vocabulary and locutions used to encode the few basic types of information carried. A major challenge in the computational processing of l_s is therefore a proper model for the syntactic and lexical means of expressing the same meaning through different forms (paraphrases). Although such models are available, their discussion is beyond the scope of this paper. In what follows, we will assume their existence and present only the few details necessary for outlining the computational procedures.

5. AUTOMATIC EXTRACTION OF CONTENT REPRESENTATIONS

In this section, we illustrate a general procedure for automatically deriving semantic representations for texts in the relatively straight-forward sublanguage of informative stock market reports. The content representations which result can be used to constitute a relational data base for a set of reports, an intermediate representation in an automatic translation system, or the starting point for the linguistic component of a text generation system. The procedure has sufficient generality to be applied in a number of sublanguages. The sublanguage l_s has been chosen for illustration purposes because it is linguistically non-trivial, yet amenable to computational treatment in the framework of the proposed procedure. (Whether or not the procedure can be implemented economically in a given application is a separate question which we do not attempt to answer here.)

Figure 2 gives a fragment of the kind of stock market report on which we illustrate our procedure:

Stocks were narrowly mixed in the early going on Canadian exchanges today as the pace-setting New York market slumped on news of a higher-than-expected rise in July's producer prices.

The MSE industrial index after the first hour of trading was down a fraction while the TSE composite index of 300 key stocks held a small gain. Financial service and metal issues sagged while oil, paper and utility stocks edged ahead. ...

Dom Stores edged up 1/4 to 19 after posting higher profits. CP, a recent high flyer, was off 1/8 at 33 5/8. Gaz Metro, which posted lower profits and filed for a rate increase, was unchanged...

Fig. 2. An informative daily stock market report. (Source: *Montreal Star*, 9 August 1979.)

If we are interested in processing "real" texts and in exploiting the special properties of a given sublanguage, we must first manually prepare a grammar and lexicon based on a detailed examination of a large corpus of texts considered to be representative of the field. We apply the techniques of distributional analysis, noting all the environments in which each word occurs. It quickly becomes clear that we can improve the description if, before comparing environments, we remove the effect of certain general grammatical transformations[9]. We may make use of automatic clustering techniques to discover important tendencies of distribution[16]. Since our sublanguage is relatively restricted, we find that classes of words which are equivalent in their distribution have a great deal of semantic homogeneity (e.g. noun classes designate functionally similar objects in the domain, verb classes designate functionally similar actions or states, etc.).

Once the important word classes have been established, at least in the first approximation,

sentence patterns are stated in terms of these classes. Consider the elementary case of stock market sentences of the form $N_{\text{stock}} V_m \Omega$, where $N_{\text{stock}} = \{golds, industrials, IBM, \ldots\}$ and $V_m = \{plunge, add, gain, \ldots\}$ and Ω is an appropriate object string (possibly empty). The sentences of (12) are acceptable in L_s but the very similar sentences of the form $NV\Omega$ in (13), while normal in general English, are unacceptable in L_s.

(12) (a) Golds plunged.
(b) IBM added 1/2 to 64 3/4.
(c) Industrials chalked up a 10-point gain. (Derived from: Industrials gained 10 points.)

(13) (a) ● Analysts plunged (on news of lower brokerage profits).
(b) ● Traders added 1/2 to 64 3/4 (to get 65 1/4).
(c) ● Corporations chalked up substantial gains.

Even though the three nouns *analysts*, *traders* and *corporations* are used in L_s, they are not used as subjects of verbs of the class V_m. Information of the kind contained in the sentences of (13) is simply never communicated in stock market reports.

For the sublanguage l_s within L_s, only a few word classes are required to state the basic sentence patterns. The most important sentence structures cover information on (i) price changes in individual stocks or in group indices, (ii) volumes of shares traded for individual stocks or for the entire daily market, (iii) comparisons in the number of stocks moving up and stocks moving down in price, and (iv) halts and resumptions in trading. The most important part of the grammar of l_s, therefore, will be a "syntactic" statement of the form of the most elementary sentences, in terms of semantic word classes such as N_{stock} and V_m. In a separate part of the grammar will be a statement of the grammatical transformations (including conjunction and relative clause formation) which are allowed to operate on the various patterns. Some transformations will normally be particular to the sublanguage and their scope of application will be defined in terms of the semantic word classes. Even the more general transformations, which may have correlates outside the sublanguage, may be semantically restricted.

The lexicon of l_s will give information about the semantic class and subclass membership of each word. This information, as well as the description of sentence patterns and transformations on those patterns, need not have any validity outside the sublanguage in question, although the grammatical information may resemble that of the whole language or of other sublanguages in important respects.

5.1 *Stage* 1: *automatic separation of the "core" text*

Let us assume that the grammar and lexicon of l_s are described in detail, but that no similar precision can be brought to the description of L_S (this does in fact appear to be the case). The problem in gaining access to the information stored in the l_s component of a text in L_s is first of all that of determining the boundaries between text segments in l_s and those in the complement $L_S - l_s$ (henceforth $-l_s$). Vocabulary alone is not enough, since some words appear both in l_s and in $-l_s$ (e.g. *drop* is an intransitive verb in the class V_m in l_s but also appears as an intransitive in $-l_s$). Fortunately, we know that sentences can be divided grammatically into clauses (we include among clauses the nominalizations of sentences which occur superficially as noun phases and infinitives): clauses encode propositions, and simple propositions are either entirely in l_s or entirely in $-l_s$. Thus the problem of determining boundaries between segments in l_s and $-l_s$ is greatly reduced to the problem of identifying clause boundaries and then finding a way to verify, for each subordinate clause (encoded proposition) whether or not it belongs to l_s. (*Remember that main clauses are normally in* l_s.) Again fortunately, it turns out that the number of clause boundary types is quite small and easily recognizable (the syntactic recognition routine is easy to write).

The problem of determining which clauses belong to l_s is also not difficult. Although there is some lexical overlap between l_s and $-l_s$, no simple clause rule for l_s will "fit" a clause in $-l_s$ because the rules of grammar for l_s are stated in terms of tight semantic subclasses of words. Thus a clause which is successfully parsed with the subgrammar and a sublexicon of l_s MUST be in l_s; otherwise (if our grammar is good), we may assume it is in $-l_s$.

5.2 *Stage 2: mapping core clauses onto entries in an information format*

Given that we have succeeded in extracting from a text in L_s the sub-text consisting of all clauses in l_s, we are in a position to use the grammar of l_s to operate on the form of the subtext sentences in such a way as to lay bare the information structure of that text. Within this stage we can distinguish two steps: (1) segmenting each sentence into its grammatical constituents, and (2) assigning each constituent to a "slot" (column) of a specific kind of information format. Each elementary sentence structure is mapped to a specific format (e.g. there is one format for price-change sentences, one for trading volume sentences, etc.). In principle, as a sentence is segmented, enough structure must be recognized by the parsing program to discriminate sentences whose structure has been altered by grammatical transformations, and separate mapping rules applied to such sentences, or else the transformations must be reversed before mapping applies.

Consider now the stock market report and its resultant mapping onto the information format of Fig. 3. Text segments which refer to the secondary domain (i.e. segments belonging to $-l_s$) are set off in square brackets. The first sentence has the structure of (14):

(14) S1: Stocks were narrowly mixed in the early going on Canadian exchanges
S2: (As the pace-setting New York market slumped
S3: (on news of a higher-than-expected rise in July's producer prices))

S2 is the subordinate part of S1. S3 is the subordinate part of S2. The first line of S1 (i.e. its independent part) belongs to l_s. The independent clause of S2 also belongs to l_s. S3, as the nominalization of a sentence referring to the outside world, is clearly in $-l_s$.

A parser can easily segment the independent clause of S1, using the grammar of l_s, as follows:

(15) Stocks / were narrowly mixed / in the early going / on Canadian exchanges / today

For the purposes of information formatting, we need to extract certain types of modifier (place, degree, time, etc.) which may occur as a part of a larger consitituent. We face exactly this problem with *narrowly*, which occurs inside *were narrowly mixed*. The same grammatical rules

Text (from the Montreal Star, August 9,1979):

Stocks were narrowly mixed in the early going on Canadian exchanges today as the pace-setting New York market slumped [on news of a higher-than-expected rise in July's producer prices]
The MSE industrial index after the first hour of trading was down a fraction while the TSE composite index of 300 key stocks held a small gain. Financial service and metal issues sagged while oil, paper and utility stocks edged ahead. ...
Dom Stores edged up 1/4 to 19 [after posting higher profits]. CP [, a recent high flyer,] was off 1/8 at 33 5/8. Gaz Metro [, which posted lower profits and filed for a rate increase,] was uncharged. ...

CONJ	N-STOCK	EXCHANGE	PRICE TREND			TIME	
			V-CHANGE	AMOUNT	END VALUE	DAY	INTERVAL
	stocks	on Canadian exchanges	were mixed	narrowly		today	in the early going
as	the pacesetting market	(in) New York	slumped				
	the industrial index	(at/on the) MSE	was down	a fraction			after the first hour of trading
while	the composite index of 300 key stocks	(at/on the) TSE	held a gain	small			
	financial service and metal issues		sagged				
while	oil,paper and utility stocks		edged ahead				
	Dom Stores		edged up	1/4	to 19		
	CP		was off	1/8	at 33 5/8		
	Gaz Metro		was unchanged				

Fig. 3. One information format used to represent the "core" component sentences of an informative stock market report. Text segments in square brackets are outside the core component. Degree and location modifiers are separated from the constituents they modify. Row entries can still be read as sentences (allowing for small paraphrastic changes such as addition of prepositions in the EXCHANGE column).

which permit *narrowly* to be recognized in this structure, however, can separate it as a format entry, moving it to the right of its parent constituent, since both actions require the lexical information that *narrowly* belongs to a certain distributional subclass (of degree adverbs). This same general principle, using subclass information to trigger the appropriate rules, is what allows us to write a set of general formatting transformations which recognize constituents and map them to column locations in the format in the same operation. Each clause in l_s can thus be given a canonical form and entered in the format, in essentially the same way as Sager *et al.* have done for medical texts.

By arranging the columns of the format properly, we may give a canonical order to constituents which retains the property of each row entry's being readable as a sentence. Note that at least three permutations are required to construct the first table entry from the clause form, yet this entry retains sentencehood. Although this property is not essential to our procedure, it appears that formats can usually be chosen which have it. We may, however, be required to insert grammatical constants (such as *in* which is added when *the pace-setting New York market* is transformed to *the pacesetting market/in New York*. Such additions are a normal part of linguistic transformations in general.

Since we have not changed the order of clauses in the original text (and we have preserved the conjunctions for the moment), and since the deletion of material from $-l_s$ does not destroy the cohesion of the text, our information format can be viewed as a kind of regularized text in which the recurrence of one basic sentence pattern is emphasized by separating the constituent classes (i.e. columns) and giving them semantic labels. One property of the information format of Fig. 3 is worth noting here. Only two columns have constituent entries for every sentence: *N*-stock and *V*-change. It is these two distributional classes which define, in a sense, the associated sentence pattern. So it is in fact obligatory that each clause have a constituent entered in each of these columns.

5.3 *Stage 3: normalization by means of paraphrase*

The formatted text of Fig. 3 is still not in a form appropriate for efficient semantic processing. In a third stage of our procedure, we must NORMALIZE the text so that entries in the same column have maximum conformity, within the limits of general rules of linguistic and non-linguistic knowledge formalizable for the sublanguage. Some of the most important steps in this stage are:

(i) Replacing semantically complex words by their most regular and semantically transparent paraphrases within the sublanguage (e.g. *sagged* becomes *moved down slightly*).

(ii) Eliminating redundant words or phrases which carry no new information in the context; this can be regarded as a kind of paraphrasing operation also, since meaning is preserved (e.g. *the pacesetting New York market* becomes *the (New York) market* since the two phrases have similar distribution, and their equivalence is confirmed on the basis of non-linguistic knowledge).

(iii) Expanding sentences with conjoined constituents into two or more separate sentences. For example, *financial service and metal issues sagged* becomes *financial service issues sagged and metal issues sagged*. (This is a traditional formatting operation which could have been carried out in creating a more regular version of Fig. 3.)

(iv) Recovering adverbials of time and place for each elementary row entry (normalized sentence) on the basis of rules of text structure. For example, in the third sentence of the text, it may not be clear on which market *financial service and metal issues sagged*... since the preceding sentence refers to both the Montreal (MSE) and Toronto (TSE) exchanges. But since the place adverbial of the main clause takes precedence over that of the subordinate clause, it is the former which is copied onto all following sentences which lack an explicit place adverbial (until the next occurrence of one).

It could be argued that the normalizing operations sketched above should be carried out prior to formatting, since their explicit formalization may occasionally depend on the structure of the original text. Suffice it to say here that (1) normalization and formatting are conceptually quite separate, and that (2) the order proposed here can be maintained in an algorithm with the aid of simple non-*ad hoc* devices.

Figure 4 gives a normalized information format for the text of Fig. 2. One of the most radical operations on the text has been the replacement of *stocks were mixed/narrowly* by a

CONJ	N-STOCK	EXCHANGE	V-CHANGE	AMOUNT	END VALUE	DAY	INTERVAL
	some stocks	(on) Canadian exchanges	moved up	slightly		today	(by) 11:00
and	some stocks	(on) Canadian exchanges	moved down	slightly		today	(by) 11:00
as	the market	(in) New York	moved down	moderately		today	(by) 11:00
	the industria index	(in) Montreal	moved down	a fraction		today	(by) 11:00
while	the composite index	(in) Toronto	moved up	slightly		today	(by) 11:00
	financial service stocks	(in) Montreal	moved down	slightly		today	(by) 11:00
and	metal stocks	(in) Montreal	moved down	slightly		today	(by) 11:00
while	oil stocks	(in) Montreal	moved up	slightly		today	(by) 11:00
and	paper stocks	(in) Montreal	moved up	slightly		today	(by) 11:00
and	utility stocks	(in) Montreal	moved up	slightly		today	(by) 11:00
	Dominion Stores	(in) Montreal	moved up	1/4	(to) 19	today	(by) 11:00
	Canadian Pacific	(in) Montreal	moved up	1/8	(to) 33 5/8	today	(by) 11:00
	Gaz Metro	(in) Montreal	moved	0		today	(by) 11:00

Fig. 4. A normalized information format derived from the format of Fig. 3 by replacing text words by their most freely occurring synonyms *in the sublanguage*. Note that conjoined index names have been separated in otherwise identical sentences. The verb phrase *were mixed* in Fig. 3 has been paraphrased using a conjunction of *moved up* and *moved down* with separate subjects. Values for the place (EXCHANGE) and TIME are filled in on the basis of general rules of text structure. The linguistic value under INTERVAL should be as in Fig. 3, but the absolute value is inserted here to save space (paraphrases based on non-linguistic knowledge should not appear at this point). Values are obligatory in all columns except CONJ and END VALUE. Note that the whole table is still readable as a coherent (but uninteresting) text, although the vocabulary is now substantially reduced to a certain key subset. The CONJ values can be dropped without loss of information except for *as*, which indicates a causal link in this sublanguage: S_1 *as* S_2 means S_1 *as a result of* S_2.

complex paraphase using a conjunction of two sentences in which the generic quantification of *stocks* has been broken into two disjoint subsets: *some stocks moved up slightly and some stocks moved down slightly*. Such paraphrases are uncommon, but fully justified within the method. As a result of the four normalization operations, all rows have entries in each column except for the column labelled END VALUE. There is no basis for reconstituting specific values on the basis of the text given.

5.4 *Stage* 4: *conversion to a relational data base*

As a result of the normalization procedure given above, the format of Fig. 4 now contains a set of sentences still readable as a coherent text. But the vocabulary of these sentences is now quite reduced, and the words used have a very direct and obvious relationship to the central concepts of the domain. From the normalized format it is evident that each of the normalized sentences expresses a relation between entities of very restricted types. If we now give a name to this relation, using the highly regular verb *MOVE*, and fix the order of its arguments (according to their most regular surface order, to stay close to linguistic form), we may represent each basic proposition in the format as a formula of predicate logic, with the general form:

MOVE ((index or stock), (direction), (amount), (final value), (place), (date), (time interval))

In order to make this transition, we must drop the conjunctions *and*, *as* and *while* which appear in the first column of Fig. 4. This amounts to a loss of foregrounding information which would be needed to reconstitute a well-formed text in the stock market sublanguage. But such information is irrelevant to the enterprise of comparing and collecting propositional content in the sum-total of all text clauses in l_s.

Figure 5 gives one possible representation for the thirteen propositions which concern the primary domain and which are derivable from the text of Fig. 2 by means of our procedure. Taken together, the set of propositions constitutes a relational data base for the text. We could

MOVE{some stocks,up,slightly,--,{MSE,TSE},1979/08/06,10:00-11:00}

MOVE{some stocks,down,slightly,--,{MSE,TSE},1979/08/06,10:00-11:00}

MOVE{stocks,down,moderately,--,NYSE,1979/08/06,10:00-11:00}

MOVE{industrial index,down,a fraction,--,MSE,1979/08/06,10:00-11:00}

MOVE{composite index,up,slightly,--,TSE,1979/08/06,10:00-11:00}

MOVE{financial services,down,slightly,--,MSE,1979/08/06,10:00-11:00}

MOVE{metals,down,slightly,--,MSE,1979/08/06,10:00-11:00}

MOVE{oils,up,slightly,--,MSE,1979/08/06,10:00-11:00}

MOVE{papers,up,slightly,--,MSE,1979/08/06,10:00-11:00}

MOVE{utilities,up,slightly,--,MSE,1979/08/06,10:00-11:00}

MOVE{Dominion Stores,up,1/4,19,MSE,1979/08/06,10:00-11:00}

MOVE{Canadian Pacific,down,1/8,33 5/8,MSE,1979/08/06,10:00-11:00}

MOVE{Gaz Metro,--,0,--,MSE,1979/08/06,10:00-11:00}

$$\text{MOVE}\{\langle \overset{\text{index or}}{\text{stock}} \rangle, \langle\text{direction}\rangle, \langle\text{amount}\rangle, \langle \overset{\text{final}}{\text{value}} \rangle, \langle\text{place}\rangle, \langle\text{date}\rangle, \langle \overset{\text{time}}{\text{interval}} \rangle \}$$

Fig. 5. The thirteen sentences of the normalized information format of Fig. 4 are here represented as propositions in predicate-argument form using the seven-place predicate MOVE. Order of arguments may differ from column order of the related format. Taken as a set, these propositions constitute a relational data base for the core sentences of the stock market report of Fig. 1, and can be interrogated in a question-answer system. Argument representations are convenient memonics, but could be coded differently, in particular in a way closer to some more standard data representation at the stock exchange where *metals*, for example, may refer to a specific index for metal stocks (note that in this case what appears to be a quantified plural is actually treated as an individual). Where a plural noun-phrase argument contains a genuine quantification (e.g. *some stocks*), it would be necessary to introduce a more complex propositional formula using an appropriately defined and interpreted quantifier *some* (different from $\exists$) for the purposes of formal theorem proving. Degree adverbs such as *slightly* could be defined as (fuzzy) intervals in a percentage change gradient. Such definitions are discoverable from the textual data by comparing use of each adverb with quantity changes given in the same sentence or accompanying quotations. It is not difficult to imagine how such propositions could be automatically generated from the relevant raw data.

imagine interrogating such a database either in relational or natural language form. Asking "Which stocks moved up in the first hour on 6 August 1979" would amount to asking for the set of all x such that MOVE (x, up, y, z, w, 1979/08/06, 10:00–11:00) is in the database for any values of y, z and w. Many refinements and extensions are of course possible. In particular, degree adverbs such as *slightly* could be given definitions in terms of some fuzzy range of percentage change (defined somewhat differently for indices than for stocks). By processing queries through the same procedure as texts, we could allow queries to exploit the full paraphrastic range of the sublanguage. We would thus process a query such as "Did any issue nosedive today?" by first paraphrasing to "Did any stock move down sharply today?", then normalizing and converting to logical form. We would then reply "no" to the query if we find no x such that MOVE (x, down, y, z, u, v, w) where y/z as a percentage change is defined as being large, u and w are any values, and v is equal to the date of the query.

6. CONCLUDING REMARKS

Sublanguages which can be analyzed semantically according to the procedure outlined above will almost certainly be good candidates for one or more types of automatic processing.

The applications to information retrieval are several. The propositional form of text content can be used for queries of a cumulative data base of texts. Statistical analyses can also be carried out directly on these forms. Or, the forms could serve an intermediate role in a system for automatic abstracting of sublanguage texts. Since we have shown that some sublanguages may have separable components, it is clear that we are not obliged to have a complete grammar and lexicon in order to extract useful data from texts. Indeed, perfect separability may not be a requirement. This approach would seem useful for extracting selected

kinds of information from semantically complex texts if the sublanguage in question is well-defined and the relevant parts of the grammar are known (i.e. the semantic word classes for the sentence patterns of interest).

In the case of stock market reports, the most interesting application might be to generate the reports from data (i.e. the price quotations at different times of the day). This would require some domain-based principles for selecting "interesting" data (a non-linguistic problem). But the problem of sequencing the propositions and integrating more than one proposition in the same sentence is already partly solved on the basis of the linguistic description used in extracting propositional content.

It is obvious that the procedure could serve to design a semantic analyzer which would represent the first stage of an automatic translation system. The propositional form could serve as a kind of intermediate representation between two languages (provided the languages "speak in the same way about the same things" in their respective sublanguages—this often appears to be the case for technical sublanguages [17]).

But if two (or more) languages have the same propositional forms for their text content in some sublanguage (and if the corresponding semantic subclasses are indeed comparable), a further step is immediately suggested: co-synthesizing texts in those languages from the same data representation. In the case of simple texts which are as well-behaved as the "core" of stock market reports, this may indeed be a viable alternative to automatic translation.

Acknowledgements—This research was supported in part by the Social Sciences and Humanities Research Council of Canada under grants 410-79-0070 and 410-81-0249.

The main ideas of Section 4 were first presented at the COLING80 conference (Eighth International Conference on Computational Linguistics, Tokyo, Sept. 1980) as a talk entitled "Embedded sublanguages and natural language processing". The procedure of Section 5 was outlined in a report to a panel on natural language text processing at ASIS81 (American Society of Information Science, Washington, Oct. 1981). Discussions with conference participants led to improvements in both content and presentation Special thanks are due to Igor Mel'čuk for detailed comments on earlier drafts of this paper; his feedback has forced me to be much more precise about things I have always taken for granted.

REFERENCES

1. R. Burton, Semantic Grammar: An Engineering Technique for Constructing Natural Language Understanding Systems. *BBN Rep.* No. 3453 (1976).
2. G. Hendrix, E. Sacerdoti, D. Sagalowicz and J. Slocum, Developing a natural-language interface to complex data. *ACM Trans. Data Base Systems* 3(2) 105–147 (1978).
3. Chevalier, J. Dansereau and G. Poulin, TAUM-METEO. Groupe de Recherches pour la Traduction Automatique, Université de Montréal (1978).
4. L. Bourbeau (Ed.), Linguistic documentation of the computerized translation chain of the TAUM-AVIATION System. Groupe de Recherches pour la Traduction Automatique, Université de Montréal (1981).
5. Final Report—Evaluation of the TAUM-AVIATION machine translation pilot system. Secretary of State, Government of Canada (1980).
6. N. Sager, Syntactic formatting of science information. *AFIPS Conf. Proc.* 41 (1972).
7. N. Sager, Natural language information formatting: the automatic conversion of texts to a structured data base. *Advances in Computers* (Edited by M. Yovits), pp. 17, 89–162. Academic Press, New York (1978).
8. L. Hirschman and N. Sager, Automatic information formatting of a medical sublanguage. In *Sublanguage: Studies of Language in Restricted Semantic Domains* (Edited by R. Kittredge and J. Lehrberger), pp. 27–69. de Gruyter, Berlin (1982).
9. Z. Harris, *Discourse Analysis Reprints*. Mouton, Paris (1963).
10. N. Sager, *Natural Language Information Processing*. Addison-Wesley, New York (1981).
11. Z. Harris, *String Analysis of Sentence Structure*. Mouton, Paris (1962).
12. N. Sager, L. Hirschman and M. Lyman, Computerized language processing for use of narrative discharge summaries. *Proc. 2nd Ann. Symp. Computer Applications in Medical Care* (Edited by F. Orthner), pp. 330–343. New York (1978).
13. L. Hirschman and R. Grishman, Fact retrieval from natural language medical records. *Proc. 2nd World Conf. Medical Informatics* (Edited by D. Shires and H. Wolf), pp. 247–251. North-Holland, Amsterdam (1977).
14. R. Kittredge, J. Bachenko, R. Grishman, D. Walker and R. Weischedel, Sublanguage panel report. NSF/NRL Workshop on Applied Computational Linguistics in Perspective. Palo Alto (1981).
15. Z. Harris, *Mathematical Structures of Language*. Wiley-Interscience, New York (1968).
16. L. Hirschman, R. Grishman and N. Sager, Grammatically-based automatic word class formation. *Inform. Proc. Management* 11, 39–57 (1975).
17. R. Kittredge, Variation and homogeneity of sublanguages. *Sublanguage: Studies of Language in Restricted Semantic Domains* (Edited by R. Kittredge and J. Lehrberger), pp. 107–137. de Gruyter, Berlin (1982).