[From: Richard Kittredge and John Lehrberger (eds.) *Sublanguage: studies of language in restricted semantic domains* (Berlin, New York: Walter de Gruyter, 1982)]

Variation and Homogeneity of Sublanguages

Richard Kittredge

1. Introduction

Since the introduction of the notion of sublanguage in 1968 by Zellig Harris, a small but growing number of linguistic subsystems have been explored in some detail by computational linguists.

First at New York University and then at the Université de Montréal, Harris' transformational and distributional techniques have been applied to set up sublanguage grammars which characterize the structure of specialized texts in a form adequate for computerized information retrieval and machine translation. The work on linguistic description coupled with parsing experiments has given substance to the impression that specialized linguistic subsystems can differ quite sharply, both in complexity and in the particular linguistic features which set them apart from the general or standard language.

Although no sublanguage has been described in all its details, a relatively complete description for the sublanguage of weather reports has led to the design of the first translation system whose output does not have to be revised[1]. The work on more complex sublanguages, as described in chapters 2 and 3 of this volume, is already in an advanced stage and shows every promise of providing the foundation for highly useful practical computational systems.

These encouraging first results have prompted a certain amount of reflection as to the generality and limits of the sublanguage approach for computational systems. In addition, they raise intriguing questions for long-term research into the ways in which various semantic and pragmatic constraints are reflected in the sentence structure and textual organization of language used under these constraints.

As early as 1974, when the first experiments in weather bulletin translation were taking place at Montreal's TAUM project, it became evident that the textual organization of English and French bulletins was virtually identical. Furthermore, the two languages' sentence formulae showed the same kinds of deletion patterns when compared with their

---

[1] In the TAUM-METEO system (Chevalier et al., 1978), all sentences which are parsed into a single syntactic structure are subsequently translated with virtually no errors. A small percentage of sentences fail the parse for linguistic reasons and these are automatically sent to a terminal for human translation.

paraphrases in the standard languages. Their syntactic and semantic categories were virtually isomorphic. And even if lexical items did not always correspond one-to-one, it was possible to use the English analysis output of the sublanguage-specific parser to help make the proper lexical correspondences in French.

When the first experiments on aviation maintenance manuals began in 1976, it became even clearer that the written style of English and French tended to be more similar in specialized technical texts than in general language texts. The stylistic parallels were so strong that translation was often possible on the level of phrase structure. English passives could often be translated by passives in French, even though the use of passive is much more restricted in general French. Clearly, there was a need to verify these early indications of stylistic parallels in other sublanguages in order to evaluate the possibilities of sublanguage translation more generally.

But cross-linguistic comparisons constituted just one of many aspects of sublanguage that seemed ripe for linguistic investigation. In late 1977 a broad study of English and French sublanguages was initiated within the Contrastive Syntax Project of the Université de Montréal. Some of the problems which have been addressed by this study are the following:

(1) What are some of the parameters of sublanguage complexity, particularly as this affects automatic language processing? Weather bulletins and aircraft manuals represent two opposite extremes on the scale of complexity. Are there sublanguages of intermediate complexity which might be natural candidates for automatic processing?

(2) How does the "professionalization" of a sublanguage affect the rigidity of style in texts? It is known that aircraft manuals are composed according to strict norms for organization and non-ambiguity, and the community of "speakers", i. e. technical writers and technical users, is highly trained. To a lesser extent, weather report writing is a professional specialty and some guidelines also exist for composition of texts in this area. In less professionalized sublanguages, lacking recognized norms, is there considerably less consistency of structure and any greater "distance" between English and French style?

(3) Do parallel sublanguages of different languages show resemblances only because of the shared semantic and pragmatic conditions, or is this partly due to stylistic borrowing between technical subcultures in contact?

(4) How does the cohesiveness of text vary from sublanguage to sublanguage? Are the means of cohesion comparable across language boundaries? Do some cohesion devices found in specialized sublanguages not show up at all in the standard language? How does the

lexico-semantic cohesion in sublanguages exploit the particular semantic categories of words which are established by the sublanguage grammar?

(5) How do the constraints of sublanguage semantics and pragmatics influence sentence structure and text structure? Can structural resemblances between semantically different sublanguages (e. g., recipes and aircraft manuels) be related to similarities of text purpose?

(6) How are the boundaries of a sublanguage determined? The sublanguages chosen for computational applications have tended to be as narrow as possible and uniform with respect to the kind of text producer and text user. What complications arise as larger and larger supersets are included in the sublanguage description?

In the first phase of the broad study, eleven varieties of English and the corresponding eleven varieties of French were examined for their sentence and text structure. The frequency of various sentence types and intersentential linking devices was used as a basis for comparison. The results of that study, given in Kittredge (1978) and summarized in section 4. of this paper, indicate in general that parallel sublanguages of English and French are much more similar structurally than are dissimilar sublanguages of the same language. Parallel sublanguages seem to correspond more closely when the domain of reference is a technical one.

A second phase of research, underway since mid-1979, is a deeper investigation of three particular sublanguages of English and French. Some results for two of these areas, English stock market reports and meteorological synopses, are included in the sketches of section 3.

The overall organization of this paper represents an attempt to summarize some perspectives on sublanguage under three headings. First, in section 2 we take up briefly the question of the defining properties of sublanguage. The reader is also referred to chapters 2 and 3 of this volume for further information on sublanguage definitions. In section 3 we consider informally some of the sublanguage varieties that have been studied to date, including four "thumbnail sketches" of English sublanguages by way of illustration. Section 4 is devoted to a more quantitative view of sublanguage variation. In section 5 we take up the problem of sublanguage homogeneity, which is of considerable importance for estimating the complexity of a sublanguage grammar. To what extent can we be sure that the linguistic description of a "representative" set of sublanguage texts can be projected onto the whole (unlimited) set of possible texts pertaining to the same domain? Are some sublanguages more homogeneous than others? What kinds of heterogeneity do we find?

In the final section we look at some implications of the preceding discussion for the automatic processing of natural language texts. In particular, how can we use what is known about the cohesive properties of

individual sublanguages to improve techniques for parsing and synthesizing coherent texts?

2. *Defining Properties of Sublanguage*

In considering which samples of specialized language can be regarded as representing "genuine" sublanguages we are immediately faced with the lack of an empirically adequate definition of the term. As an intuitive concept we have no difficulty accepting the statement that the "language of stock market reports", for example, constitutes a rather separate linguistic subsystem. On closer examination, though, we feel the need for some criteria for deciding what the limits are for a given sublanguage, and whether closely related varieties of language should be considered parts of the same sublanguage or as constituting separate systems.

The closure property proposed by Z. Harris (1968) is not in itself sufficient to resolve these questions. If a sublanguage can be *any* subset of sentences which is closed under the transformational operations, this definition could identify a very large number of linguistic subsets as sublanguages. But closure under the operations is only intended as a necessary condition. Closure only assures us that if we already have a set of sentences which we intuitively consider to be a linguistic subsystem, we must include in it all sentences generated from the candidate set by means of the transformational operations of negation, question formation, clefting, conjunction, etc.[2] Harris' major concern is with the sublanguages of science and technology, where there is a commonly accepted domain of interest and fairly sharp intuitions on the part of the specialized community of "speakers" as to the acceptability of sentences in the subfield. The semantic limitation of the domain of discourse is thus all-important, but only a necessary condition, not sufficient (see also the introduction to this volume). What is required in addition is that there be shared habits of word usage on the part of the speakers. Hierschman and Sager (p. 27, this volume) include this criterion in their working definition of sublanguage. Thus a new term or grammatical construction does not become a true part of the sublanguage until its use has been conventionalized by the community of specialists.

The requirement of conventionalization poses a problem, however, since most dynamic scientific sublanguages are constantly borrowing terms from the standard language, particularly when new concepts are being introduced and analogies are needed. Thus a recent article on subatomic particles contained the following sentence:

---

[2] In fact, as Harris observes, a linguistic subsystem may be closed under only *some* and not all of the operations. This is seen in the sublanguages L_s and L_{ws}, for example, where question forms are absent.

> It is apparent that somehow the naked quark and antiquark "dress" themselves with other quark-antiquark pairs before they emerge to macroscopic distances where they can be detected.

Whether or not the word *dress* settles down to becoming a part of the sublanguage, with accepted distribution, most articles in scientific journals have some degree of "contamination" from the general language used essentially for the first time. This poses a particular problem for automatic processing, since some access to information about the whole language is required if all of the text is to be parsed with some level of understanding.

But the situation may be less than hopeless. As Harris himself notes, the ways in which terms outside the sublanguage proper may be introduced into scientific sublanguages appear to be limited. This may be possible only under certain kinds of conjunction, or in situations which are analyzable as the right-conjoining of a sentence outside the sublanguage to one which is in it. And this kind of controlled contamination does occur quite clearly in many types of stock market report (see section 3.4 and section 5.1 on embedded sublanguages).

For some sublanguages, the relevant community of speakers is not well-defined. This is particularly the case for written sublanguages where access to the texts is relatively free (e. g., stock market reports, recipes, newspaper columns on playing bridge, weather bulletins, etc.). What often occurs in these cases is that different subtypes evolve which are oriented to different categories of user, dependent on their level of expertise. These subtypes involve different degrees of deletion, somewhat different sizes of lexicon and complexity of sentence structure. There is every reason to call them different sublanguages for purposes of computational treatment and linguistic taxonomy.

What seems to set technical sublanguages apart from scientific ones (where the four examples above are considered technical), is their one-directional character. Expert technicians or analysts may form a community which establishes norms, but the text users are often less experienced and give little feedback to the text writers. A scientific community, on the other hand, is based on a much more even exchange, particularly within precise speciality areas.

The insistence on shared patterns of usage is an important part of defining sublanguage for practical applications. From the shared patterns as they are seen in word usage in texts, it is possible to infer some aspects of the shared knowledge of the speakers. When a representative set of texts from a specialized field is subjected to distributional analysis, the words can be grouped into categories and subcategories depending on their similarities of occurrence. These categories are used to state a sublanguage grammar, which expresses the structural patterning and specific lexical co-occurrence restrictions active in the sublanguage sentences. But the sublanguage gram-

mar is more than just a linguistic characterization of the texts. The lexical classes and the hierarchical relations between the classes usually reflect the accepted taxonomy which the specialized field of knowledge imposes on the objects of its limited domain of discourse. And the combinations of lexical classes which are permissible in the sentences of the specialized texts reflect the conceivable relations between these objects (regardless of truth or falsity). Thus the sublanguage grammar can be said to incorporate certain aspects of a knowledge representation for the subfield.

3. *Varieties of Sublanguage: Four Thumbnail Sketches*

In order to give some perspective on the types of variation discussed quantitatively below, we begin by reviewing some of the salient properties of a few sublanguages included in the broad study.

3.1. Technical manuals – aviation hydraulics (L_{ah})

Many technical sublanguages are most easily observed in the form of written manuals.[3] Because of the interesting similarities between technical manuals from very different semantic areas, we include here a text fragment from an aircraft manual followed in the next section by a recipe sample.

Aircraft maintenance manuals of the kind described in detail in chapter 3 of this book can be classified according to the aircraft system referred to. In figure 1 a hydraulics manual sample can be seen subdivided into two parts, a descriptive passage written in a style not unlike that found in many areas of standard English, and a procedural section whose sequence of imperative sentences is the essence of the manual. In this sample, paragraph 22 (PRESSURE SWITCH) describes the operation and main features of a particular component of the hydraulic system. Sentence structures resemble those of general English. Paragraph 23, however, gives a completely different subtype. Definite articles and repeated definite object noun phrases are usually deleted when this does not introduce ambiguity. The imperative sentences (a) – (d) all show *the*-deletion (indicated by ∅). But sentence (e) retains *the* in *the two mounting bolts*, presumably since deletion would introduce a serious ambiguity.

In some cases a maintenance procedure may be preceded by a list of parts.

One particular feature of such manuals is the numbering of sections and indexing (here by letters (a) – (e)) of the procedural subsections. This clearly satisfies an important need for unambiguous anaphora over indefinitely large portions of the text (e. g., the reference to "Paragraph 13, preceding").

---

[3] See Grosz, chapter 5 of this volume, for an interesting sample of a spoken technical sublanguage.

SUBLANGUAGE OF AVIATION HYDRAULICS
(Sample Text Fragment)

PRESSURE SWITCH

22 Two identical pressure switches, one in each system, are electrically connected to lights on the warning light panel. When the system pressure drops to 1250 (0,−150) psi, the switch closes the circuit to the hydraulic pressure warning light.

REMOVAL AND INSTALLATION OF PRESSURE SWITCH – NO. 1 SYSTEM

23 Removal procedure:
(a) Depressurize Ø hydraulic system (refer to Paragraph 13, preceding).
(b) Disconnect Ø electrical connector on Ø pressure switch.
(c) Disconnect Ø line at Ø pressure port.
(d) Disconnect Ø line at Ø drain port elbow.
(e) Loosen the two mounting bolts and remove Ø switch.

Fig. 1. The aviation hydraulics sublanguage (L_{ah}) found in maintenance manuals exhibits two important subtypes: general descriptive material (eg. paragraph 22) in which deletion is rare, and procedural sections (paragraph 23) with imperative sentences in which deletion (ellipsis) of definite articles and repeated object noun phrases is typical. Paragraph numbering allows cohesive anaphora of greater scope than would be possible in non-technical texts. Retention of the definite article in (e) seems due to the following numeral *two*. Deletion would introduce a serious ambiguity. The symbol Ø indicates deletion of an occurrence of the definite article.

3.2. Technical manuals – cooking recipes (L_r)

A different kind of technical manual can be found in any kitchen recipe book. The sublanguage of recipes is particularly interesting because it is accessible to a large community of speakers, even though some distinctions of expertise appear to be reflected in the subtypes of recipe observed. Furthermore, as one of the few sublanguages to be found in every human language, it provides an exellent area for cross-linguistic comparisons. Since written recipes are generated spontaneously in most languages with written traditions, cultural borrowing may be less of an influence on sublanguage style than in highly technical sublanguages.

Figure 2 shows a French soup recipe with the typical division into a list of ingredients followed by a sequence of procedural "assembly instructions". Some informal recipes begin with a general discussion; this three-part organization is exhibited in the American bread recipe given as figure 3. (Other technical manuals, including aircraft maintence manuals, show

Soupe des «Halles»

250g. d'oignons, 95g. de beurre,
30g. de farine, 1 litre $^1/_2$d'eau chaude
2 cuillères à café de sel, 10 tours de moulin à poivre,
une douzaine de tranches de pain,
100g. de gruyère râpé.

Epluchez *les oignons* et hachez-*les* finement. Dans une casserole à fond épais, faites-*les* cuire avec 75g de beurre sur feu doux. Au bout de 15 minutes environ, *ils* doivent être cuits et à peine colorés. Saupoudrez Ø alors avec la farine que vous laissez blondir. Mouillez Ø avec de l'eau chaude et assaisonnez Ø.

Fig. 2. Cooking recipes typically show deletion of definite object noun phrases, which occurs when the same NP is repeated in consecutive procedural sentences. This French sample shows a gradual progression from full NP, through pronominal forms, to deletion of entire NP. English recipes typically delete (Ø) object NPs after the initial occurrence. A wide variety of other languages show similar object deletion. English recipes also delete definite articles in addition as a feature of telegraphic style. Notice that the definite *les* in the first sentence of the above instructions indicates a cohesive link with the list of ingredients.

this same tripartite division when a list of components appears between the descriptive introduction and the procedural section). The procedural sentences of the recipes show a pattern of deletion for repeated definite object noun phrases (NPs) which is typical of many technical manuals.[4] French recipes often show a progression from full NP in the first procedural sentence occurrence, through pronominalization, to deletion of the entire NP after a few sentences. English recipes seem to move to full deletion more rapidly. The same general pattern of deletion (with or without transition via pronominalization) can be observed in other languages of the Indo-European family, and may not be limited to this language family. Other types of deletion are found in English recipes, particularly on definite *the*. The two types have an important difference, however. Object NP deletion can only occur in the context of a preceding co-referential NP, often in a preceding sentence. This type of deletion is an important part of the network of cohesion devices for a typical recipe. Article deletion, on the other hand, is not anaphoric, since it may occur in the first sentence of a section,

[4] In the longer sentences of L_{ah} this deletion occurs mainly under internal sentence conjunction and not as a linking device between sentences.

SUBLANGUAGE OF RECIPES (L_r)
(Sample Text Fragment)

S_1	Basic Wholewheat Bread
S_2	The best flour you can buy is stone-ground wholewheat flour.
	. . . . .
S_6	9 cups warm water 1 cup honey 1/2 cup vegetable oil 5 tablespoons granule yeast 2 tablespoons salt 20 cups . . . flour
S_7	Allow Ø yeast to soften in Ø warm water, for about 5 minutes, along with *the* honey.
	. . . . .
S_{10}	Add enough of the remaining flour to make the dough easy to handle.
S_{11}	Turn out Ø onto Ø floured board.
S_{12}	Add more flour if necessary.
S_{13}	Knead Ø.
S_{14}	Knead Ø and knead Ø.
S_{15}	Knead Ø until *it* feels good – not sticky but warm and elastic.
	. . . . .

Fig. 3. Informal recipes often begin with an introduction in the style of the standard language. Here both cohesive deletion of repeated object NPs and non-cohesive deletion of determiners *the* and *a* are indicated by Ø. The *it* of S_{15} is co-referential to *the dough* of S_{10} if only surface structure is considered, but to the closer deleted occurrence of *the dough* in S_{15} in a representation where deletions are recovered.

provided that the reference is clear from the restrictions on the domain of discourse. Definite articles are deleted in a large number of more or less telegraphic sublanguages of English, but in far fewer sublanguages and situations in French.

3.3. Regional weather synopses (L_{ws})

Another sublanguage of particular interest is that of regional weather forecasting. Unlike local weather bulletins, which are highly telegraphic,

SUBLANGUAGE OF METEOROLOGICAL SYNOPSES (L_{ws})
(A Sample Text)

MARITIMES WEATHER OFFICE
APRIL 9 1974
5:00 A.M.

A storm centred at forecast time over Virginia will move slowly northeastward during the next two days. Precipitation should begin over extreme southwestern Nova-Scotia before noon and spread northeastward later in the day and overnight.

Snow should fall at most localities for a few hours at least before changing to rain. Over northern New-Brunswick however *indications are* that this change will not occur and that a sizeable snowfall could result tonight and on Wednesday. It *is* however too *early* to make a reasonable estimate of these amounts.

For the remainder of the Maritimes rain will be heavy at times and continue Wednesday.

Fig. 4. Meteorological synopses consist of one or more cohesive paragraphs made up of complete sentences. Unlike meteo bulletins (in L_w) which are telegraphic and lack tensed verbs, synopses show cohesion in the repetition of tense. Note that predicates which describe the observer's relation to the phenomena (italicized) may be in present tense. If *should* is analyzed as *will probably* (*should* cannot signify obligation in this sublanguage) and *could* is analyzed as *will possibly*, we see that sentences (or propositions) describing the phenomena themselves have a uniform semantic tense (future). Thus the repetition of semantic tense is more uniform within the sequence of description statements.

regional synopses consist of one or more full paragraphs describing the general movement of air masses over the continent and their effect on local conditions. The sample in figure 4 illustrates some of the semantic restrictions found in these texts. The domain of reference is quite constrained, being limited to a fixed geographic area (for any set of texts produced by a given weather station) and the predicted changes in weather conditions within this area during a period of a few days following the time of forecast. Verbs of location, motion, causality and inchoation serve as the core predicates.

One important property of these texts is their pattern of temporal and modal reference. Whereas technical manuals are quite uniform in tense, synopses show either a progression of time reference (past to future) or uniform future reference. However, in order to see the temporal uniformity

one must isolate two distinct levels of text. The first is a sequence of observations and predictions about physical pheomena which constitute the essence of the synopsis, and it is this sequence which shows uniformity of temporal reference (consistent future in the sample given). The modals *should* and *could* in figure 4 must be replaced by their semantic paraphrases *will probably* and *will possibly* respectively. *Should* cannot have the meaning of obligation in this sublanguage. With this regularization of modals the sequence of statements about the physical phenomena has a uniform future reference. The second level of text is found in the predicates which describe the act of observation itself or relate the observer to the phenomena *(indications are, is too early)*. The second level has uniform present temporal reference. Although the two levels are interwoven syntactically, the proper predicate-argument semantic analysis is that the first level is embedded in the second.

3.4. Stock Market Reports (L_s)

One of the most interesting sublanguages from the linguistic point of view is that of stock market reports. These reports, available in most newspapers, summarize the daily trading activity in corporate shares, and relate the changes in share prices to changes in economic and political conditions. The basic "objects" of concern are the shares of specific corporations. Most of the detail in a typical report concerns the price changes for individual stocks or for groups of stocks in the same economic sector (e. g. banks, utilities, oils, industrials, papers, etc.). Thus a very frequent sentence type in L_s is illustrated by (1).

(1) Rio Algom jumped 1½ to 32.

Typically, it is not the company or the company's stock which is referred to, but the price of the stock. Nevertheless, in the same report, we may see the same proper noun play other roles, as in:

(2) Rio Algom said there aren't any corporate developments to account for . . .
(3) Rio Algom resumed trading after the halt at 32½ . . .
(4) . . . and could affect uranium producers such as Rio Algom.

In (2) the reference is to a spokesman for the company. In (3) the proper noun refers to the shares of the company. And in (4) the reference is to the company itself. Although such multiple reference is avoided in reports from some sources, there is no ambiguity when it does occur since for the most part the verbs have sufficiently precise selectional restrictions to separate the intended meanings.

3.4.1. Verb subclasses

One of the characteristic features of L_s is the metaphorical use of motion verbs to describe changes in stock prices.[5] Although certain intransitive motion verbs like *advance, climb, rise, fall, drop, jump, move up, plunge* and *dip* are used frequently, a large number of others may occur from time to time. What is striking about the basic verb set is its variety. Whereas the semantics of market changes is quite simple, there is a tendency to have large synonymous sets and avoid repeating the same verb too often in one report. Within the two sets of intransitives denoting upward and downward movement, respectively V_{m+} and V_{m-}, we observe that not all members are exact synonyms. A distributional analysis gives the following division into subsets:

		neutral	marked for large degree	marked for small degree
V_m	V_{m+}	move up advance gain rise climb push upward etc.	jump soar surge bounce up spurt shoot up	edge up creep up firm struggle upwards
	V_{m-}	move down fall dip drop decline slide etc.	plunge tumble nosedive	drift down slip (back) sag ease settle down

In the first approximation, verbs in the second column co-occur with large percentage changes in price (e. g. IBM jumped 8½ to 64¼) and those in the third column with small changes. Verbs in the first column are unmarked for amount, co-occurring with the full range of percentage changes as well as with adverbs of the *sharply*-class and with the *slightly*-class. The degree-marked classes select only one of the adverb classes each:

(5a) *The gold index jumped slightly.
(5b) The gold index eased (slightly).

[5] Motion verbs are significantly less used in reports from some British sources, where such verbs as *cheapen* and *strengthen* are prominent. The sublanguage L_s is described on the basis of North-American reports, but reports from other regions are not substantially different.

(6a) Mines plunged sharply.
(6b) *Mines sagged sharply.

The frequent use of exact price quotations, and changes, co-occurring with the V_m class provides a rather unusual opportunity to observe part of the semantic distinctions directly. *Jump* can be established as a hyponym of *advance* or *rise*, since every time a stock can be said to jump it can be said to rise, but not the converse.

On a deeper analysis, the distributional patterns turn out to be somewhat more subtle. First, a given percentage change for an individual stock does not have the same meaning in this subfield as the same percentage change for a sector or market index. We may get:

(7) Massey Furgeson *eased* 1/4 to 9 1/4. (a change of 2.6%)

but also:

(8) The TSE 300 index *plunged* 45 to 1740. (a change of 2.6%)

Thus the semantic distinctions must be set up on the basis of co-occurrences with the same semantic subclass of nouns in subject position. That is, *ease* and *plunge* can be assigned to complementary verb classes provided that their percentage changes are compared when both take a subject from the noun class N_{stock} = {*IBM, Rio Algom, Massey*, etc.} (the individual stocks) or both take a subject from the noun class N_{index} = {*The Dow, golds, utilities, banks, (the oil and gas)index*, etc.}.

3.4.2. Collocations

The specific word co-occurrences which obtain in a sublanguage are a reflection of the special domain and its organization. Between the noun classes and verb classes set up on the basis of a first-order distributional analysis, certain refinements can be introduced. For example, even though *sag* and *slip* are in the same verb subclass on the basis of their percentage change behavior with respect to N_{index}, there is a preference to use *sag* only with N_{index} subjects and not with N_{stock}.

More interesting are the co-occurrence restrictions between verbs and adverbs. Adverbs such as *sharply* or *strongly* are not normally used with motion verbs in their physical sense, although they co-occur quite naturally with the same verbs in L_s:

(9a) *The book dropped sharply.
(9b) The gold index dropped sharply.
(10a) *Bubbles moved up strongly in the test tube.
(10b) Stocks moved up strongly on Canadian exchanges yesterday.

In fact, some adverbs such as *strongly* seem to have the particular meaning of *in great numbers* rather than *to a great extent* or *at a rapid rate*.

3.4.3. Sentence structures

Sentences in L_s resemble sentences of the standard language, but tend to be short, with distinct patterns of subordination. Whereas co-ordinating conjunctions are very limited in distribution, three types of subordination play an important role in relating the two major levels of text: (a) the description of market activity at the various stock exchanges, and (b) the juxtaposed information about the economic and political factors that are causally related to events of the marketplace. The first subordinating device is the right-adjunction of subordinate clauses using *as, while,* or *although* (but not *because, until, unless*) as the typical conjunction. A second important subordination device is the use of non-restrictive relative clauses (restrictive relatives are rare) to provide extra-market information about a stock issue that merits particular attention in the report:

(11) Imperial Oil, which sparked Monday's upsurge on speculation over its NWT oil play, opened up 1/8 at 33 1/4 . . .

A third way of subordinating extra-market information is through use of embedded complement clauses introduced by one of a small set of verbs such as *say, report,* etc.:

(12) Massey said its committment to return to profitability in fiscal 1979 is achievable.

Imperative and interrogative structures are essentially absent from the stock market report sublanguage. Passive sentences are infrequent, unlike many technical languages.

L_s is a slightly telegraphic sublanguage. *The*-deletion is not uncommon (see the first and fifth sentences of the sample text in figure 5). Gapping, or deletion of repeated verb within a conjoined sequence of sentences, is frequent:

(13) Canadian Merrill slipped 1 3/8 to 20 5/8, Prairie Oil 1 1/8 to 15 3/8, Cambell Red Lake 1 to 37 and Asbestos Corp. 1 to 45 1/4.

3.4.4. Other features

Variation of tense is an important phenomenon of L_s. The main clause of a sentence, or the head clause of a paragraph, describes the market activity (change of prices, number of shares traded, halts and resumptions of trading, etc.) and is in the past tense. The subordinated material, which may be a sequence of non-initial sentences for a paragraph in the latter part of a report, is unrestricted for tense, but usually in present:

(14) Calgary Power, which is seeking a rate hike, added 1/2 to close at 39 7/8.

SUBLANGUAGE OF STOCK MARKET REPORTS (L_s)
(Sample Text Fragment)

S_1 Stocks moved higher again on the Canadian markets yesterday in an extension of Tuesday's upturn, with a buoyant oil-gas group leading the way.

S_2 At the close the 300-stock composite index was up a further
S_3 5.28 points at 1280.50 after a rise of 5.13 in the previous session. On Monday the indicator had plunged 13.76 points in a sell-off touched off by the news of a sharp boost in oil prices planned by the Organization of Petroleum Exporting Countries.

S_4 Of the 14 groups included in the composite average, 11
S_5 chalked up index gains yesterday. Firmest sections, along with the oil-gas issues, were the industrial products, managements, consumer products and utilities.

S_6 The golds posted the only sizeable decline.

. . . . . .

Fig. 5. A stock market report of intermediate complexity illustrates the frequent use of motion verbs (*move, plunge, gain, decline*). The latter two verbs are nominalized under verb operators (*chalk up . . . gains* in S_4 and *post . . . decline* in S_6). Key nouns such as *news, reports, rumors* may have complements which are lexically and grammatically outside the language of market transactions. The occurrence of *firmest* in S_5 shows a cohesive link, in the use of superlative, with *11 (groups)* in S_4. Lack of *the* before *firmest* is indicative of the telegraphic style of L_s.

(15) A highlight of the session was Basic Inc., with a jump of 12¾ to 44⅜. Combustion Engineering plans to offer $46 a share for Basic stock.

The shift between past and present tense in a typical report corresponds to the distinction between sentences or sentence fragments of the market "core" of L_s and the extra-market matrix portion of the text. (See section 5.1 below for a discussion of embedded sublanguages.)

An important feature of L_s is the frequent use of verb operators, verbs such as *post, show, sport, chalk up,* etc., which nominalize a typical verb of the V_m class with what appears to be little semantic effect:

(16) post a gain ← gain
show a loss ← lose
sport a gain ← gain
register a decline ← decline

The operators do, however, introduce a semantic change which shows up in the complementation properties of the verb:

(17) Xerox gained ½ to 68.
(18) *Xerox gained.
(19) Xerox posted a gain.

Post and the other operators take as their argument a verb which normally requires a complement of amount, giving a resultant which does not require any such complement. The semantic effect of these operators is aspectual in nature.[6] They turn potentially durative processes (e. g. *Xerox gained for three hours*) into events *(Xerox chalked up a gain (*for three hours))*. The resultant of such a verb operator-operand application is a transitive sentence structure into which quantity modification is insertable only within the syntactic framework of the object noun (a nominalized verb);

(20) The gold index gained $\left\{\begin{matrix}\text{6 points.}\\ \text{substantially.}\end{matrix}\right\}$

(21) The gold index chalked up a $\left\{\begin{matrix}\text{6-point}\\ \text{substantial}\end{matrix}\right\}$ gain.

The verbs of L_s seem to be blissfully ignorant of linguists' remarks on nominalization in the language as a whole. Nearly every V_m verb has a corresponding event nominalization which is semantically regular:

(22) . . . due to the *slump* on Wall Street
. . . which sparked Monday's *upsurge* . . .
The *advance* in stocks occurred . . .
Traders were upset by the *plunge* of over 3.3 percent . . .

Verbs outside the V_m set are also often nominalized as events:

(23) . . . after a *halt* called last week.
. . . which is seeking a rate *hike* . . .
. . . fell 3 to 48 after a delayed *opening*

Agent nominalizations are common, and help to give a certain flavor of animacy to stocks and indices:

(24) Other fractional *gainers* included . . .
. . . Dome Pete was again a *standout* . . .
. . . among the scattering of *climbers* in the Canadian list . . .

The existence of such a broadly based relationship between verbs and these two major classes of nominalizations is an important element in the tend-

---

[6] For a discussion of the aspectual effect of such verb operators, see Kittredge, R. (1970) *Tense, Aspect and Conjunction: Some Inter-relations for English*. Transformations and Discourse Analysis Project papers no. 80. University of Pennsylvania.

ency to include grammatically subordinated descriptions of relevant events in sentences whose main clause (or clause kernel) describes the foregrounded market event:

(25) . . . the pace was beginning to slow *due to the slump on Wall Street.*
. . . the 300-stock composite index was up a further 5.28 points
. . . *after a rise of 5.13 in the previous session.*
A surge of late buy orders helped the Dow Jones industrial average in New York recover some of the ground it lost in the morning . . .

4. *Measures of Sublanguage Diversity*

The number of sublanguages that have been studied from any single point of view is not large. Thus it is somewhat early to attempt to draw up a taxonomy of linguistic subsystems which are semantically and pragmatically determined. Some insight, however, can be gotten from a broad survey of eleven varieties of English and French undertaken by the Contrastive Syntax Project at the Université de Montréal during 1977–78. Of the eleven varieties, three might be considered semantically too diverse to qualify as sublanguages:

(1) macro-economics – introductory textbook
(2) children's stories
(3) literary criticism

The other eight contained samples of three sublanguages which have been studied in detail for computational purposes, but for which frequency of structural types had not been counted ((4)–(6)):

(4) weather bulletins (L_w)
(5) aviation hydraulics manuals (L_{ah})
(6) pharmacology reports – cardiac glycocides (L_p)
(7) weather synopses (L_{ws})
(8) recipes (L_r)
(9) stock market reports (L_s)
(10) micro-economics – section on mathematical foundations
(11) university catalogs – section on requirements for a degree

Using a sample of roughly 100 sentences for each variety represented, counts were made of major sentence structure types and important grammatical and semantic linking devices between adjacent sentences. The discussion below is based on (a) the 100-sentence samples, (b) additional samples of the eleven varieties for surprising results, (c) published and unpublished studies of sublanguages (4)–(6) done in Montreal and New York.

4.1. Lexical Field – size and complexity

One of the most obvious parameters of sublanguage comparison is the size and diversity of the sublanguage lexicon. But comparisons of size are both difficult and misleading. First of all, a precise measure of size is possible only to the extent that the sublanguage is lexically closed, and it appears that few sublanguages are. When a sublanguage is relatively "convergent" (see section 5), it may make sense to use some confidence level in considering a lexicon to be "well-specified". In a sublanguage whose lexicon is specified to the 99.99% confidence level, we would expect to meet one new word per 10,000 words of new text in the sublanguage. For many sublanguages, this may be too high a confidence level to expect. During the early work on the cardiac glycoside subfield of pharmacology at NYU, it was estimated that, even in such a narrow specialty, more *new* words than that could be expected to enter the sublanguage in a typical stretch of a new scientific article. This would be true even if one discounted the contribution of new proper names.[7] (Authors often cite one another, but there is no closure of this set.) Although high confidence levels may be possible in relatively stable technical areas such as meteorology, they seem unrealistic in constantly changing sciences, particularly those with large lexicons. Damaging as this might seem for the prospects of computational treatment of scientific sublanguages, there are some reasons for optimism (see sections 5.1 and 6.4).

Thus far, the sublanguage which seems to have the smallest lexicon is L_w, the sublanguage of weather bulletins. The METEO system, which daily handles some 30,000 words of English text, is based on a lexicon with under 1000 dictionary words. It nevertheless accepts an indefinitely large number of place names, provided that these occur within one of a small number of place name formulae. But this lexicon does not specify the sublanguage to a high confidence level. There are two reasons for this. First, a certain number of lexical items (e. g. *Christmas Day*) appear in reports so rarely (though predictably) that reasons of efficiency favor treating them as other "unfound" words. The second and major reason for incomplete specification is that even in the highly stereotyped milieu of weather reporting, some lexical liberties are taken, particularly in the face of unusual weather conditions. Thus, a small part of the last confidence-percent of the lexicon in L_w is rather unpredictable. This does not prevent the operation of an automated system which translates extremely well when it recognizes all the words. It simply asks for human help in the small percentage of remaining cases.

---

[7] See Sager et al. *String Project Reports*. Linguistic String Project, New York University.

Lehrberger estimates (chapter 3, this volume) that the sublanguage of aircraft maintenance manuals (which subsumes L_{ah}) may approach a lexical size of 40,000 words. This is particularly impressive in that it does not include proper names, which are virtually limitless in the form of alphanumeric labels, etc. Of the other sublanguages for which some estimates may be possible (L_p, L_{ws}, L_r and L_s) the varying rates of convergence make comparisons difficult. Although L_s has a sublanguage core which may be composed of only a few thousand words, the adjunction of statements concerning the world at large carries a corresponding increase of lexicon of an order of magnitude. Both L_{ws} and L_r, though relatively convergent, still may exceed 10^4 lexical items each (excluding proper names, an important part of L_{ws}). L_p is certainly in excess of that figure.

Estimates of lexical size are rather misleading as measures of complexity for computational processing. The language L_r of recipes is lexically extensive by virtue of the long list of possible ingredient names and names of possible prepared dishes. But this portion of the lexicon falls into a few major syntactic subclasses of nouns (where count nouns, e. g. *capon*, are distinguished from non-count nouns, e. g. *flour*) with some semantic subgrouping to account for selectional restrictions. The set of instrument names (e. g. *whisk*, *spatula*, *blender*) has very homogeneous properties and the set of verbs consists of only a few important homogeneous subsets. The lexical complexity of L_r, though not insignificant, is therefore less than the size of the lexicon might indicate. What counts most is the number of lexical categories and subcategories which must be distinguished for the proper grammatical description, and the average complexity of description of a lexical item in terms of these categories. Nouns which play only one role, and can be described adequately by assigning them only a few category labels contribute less to overall lexical complexity than predicate words, whose subject, object and instrument selections must be specified, and which may fall into different semantic subcategories.

4.2. Sentence structures

Telegraphic sublanguages such as L_w may have "sentence" structures which cannot be easily equated with those of standard English. The structural inventories of such sublanguages are difficult to compare with less elliptical varieties. Structural typology is even more difficult when the major categories in one sublanguage do not correspond to those in any other. For example, in L_w, the most important single category is that of "weather conditions", since these may stand alone in a well-formed weather bulletin (e. g. *Rain.* or *Sunny.*). But this category, which is in fact a phrase category, does not show syntactic homogeneity of the kind seen in the standard language:

(26) { Rain / Partly cloudy / Becoming cooler / Blowing snow } { this evening throughout the Lower St. Lawrence Valley. }

This major category contains some, but not all, noun phrases, adjective phrases and gerundive phrases. The only homogeneity is semantic.

Most sublanguages of science and technology show enough similarity in sentence structure to standard English that the standard language can be used as a basis for their comparison. In figure 6 we give a list of some of the major structural features of (English) sentences which have proved useful for the comparison of sublanguages. Although we have no precise measure for the standard language, relative frequency judgements can be obtained by matching absolute frequencies against an approximate norm as represented by the texts which seem subjectively to most closely approximate the standard language: introductory economics texts. The frequency or type of each sentence feature (we include tense and modality as well as structural types) is given for each of the six sublanguages L_{ws}, L_s, L_{ah}, L_r, L_p and L_e, the last being the macroeconomics sample which approximates the English standard in most respects. Relative frequency is based on counts made in the broad-based study cited above, and confirmed over additional texts when deemed appropriate.

The five relative gradings (++, +, A, −, −−) give a clear enough picture of the structural tendencies of a sublanguage to make general comparisons possible. Clearly, the profile of any given sublanguage could be extended (see section 4.3 below for the corresponding profile using intersentential linking devices), and include other features; nevertheless, some comparisons can be made on the basis of this information. In particular, we notice that the similarities between aviation technical manuals and recipes for imperatives and deletions (++ in both) do not carry over to all other characteristics. Whereas the descriptive introduction to each section of a technical manual is an important component, such introductions are not typical except in the case of informal recipes.

4.3. Intersentential Linking Devices

An important dimension of sublanguage variation is in the means of textual organization. Instruction manuals have clear divisions into subsections, with frequent use of indices which allow for anaphora over large stretches of text (and including extra-linguistic material). Weather bulletins have a clear sequence of presentation, with certain elements obligatory, others optional, but always an underlying text template into which the subparts must fit. Some of these macro-structural properties of sublanguage texts are difficult to quantify, or characterize in a sublanguage-independent way.

FREQUENCY PROFILES FOR SIX SUBLANGUAGES OF ENGLISH

Sentence Feature		L_{ws}	L_s	L_p	L_{ah}	L_r	L_e
Tense: dominant (other)		future (pres)	past (pres)	present (past)	pres.	pres. (fut)	present (past)
Modality type	probability	prob.	prob.	prob.	oblg.	oblg.	both
	obligation						
Interrogative		– –	– –	–	– –	– –	A
Imperative		– –	– –	– –	+ +	+ +	–
Subordination	relative cl. (restrictive)	–	–	A	–	A	A
	relative cl. (non-restr.)	– –	+ +	A	–	–	A
	embedded complement	A	A	+	A	–	A
	subordinating (w/conjunctn.)	A	+	+	A	A	A
Deletion	article	A	+	A	+ +	+ +	A
	object NP	– –	– –	A	+	+ +	A
	copula	– –	+	A	+	+	A
Topicalized Structures	passive	A	–	+ +	+	–	A
	clefting	–	–	A	–	–	A
	pre-posed clauses	A	–	+	+	A	A
Proper Nouns		+ +	+ +	A	A	– –	A
Human Nouns		– –	A	A	– –	– –	A

Fig. 6. + + = far more frequent than in standard language: salient
+ = significantly more frequent than in standard language
A = average frequency for the language
– = significantly less frequent than in standard language
– – = not found in a sizeable corpus[8]

L_{ws} = weather synopses L_s = stock market reports L_p = pharmacology of cardiac glycocides
L_{ah} = aviation hydraulics L_r = recipes L_e = economics text (closest to standard)

What is easier to use for sublanguage comparisons are the cohesive links by which texts are "held together" on a local level, sentence by sentence. In a broad survey of textual cohesion in several sublanguages of

[8] The frequency designation "– –" might be interpreted as meaning less than one occurrence per 10,000 words of text. At this level it is often more efficient to exclude the feature from the grammar for computational purposes, since this is at the level of general structural and lexical seepage for most sublanguages.

RELATIVE FREQUENCY OF LINKING DEVICES IN SIX SUBLANGUAGES (ENGLISH)

Cohesion Devices	L_w	L_{ws}	L_s	L_p	L_{ah}	L_r
Pronominalization	− −	− −	−	−	− −	+
Deletion of Object NP	− −	− −	− −	− −	+	+ +
Comparatives and Superlatives	+	+	+	+ +	A	+
Conjunctive adverbs (*nevertheless, however, . . .*)	− −	+	A	+	A	A
Lexical Repetition	A	+	A	+ +	+ +	+
Synonymy	− −	+ +	+ +	+	− −	−
Hyponymy ("classifiers")	A	+	+ +	−	+	+

Fig. 7. Relative frequency of seven types of intersentential linking is given with respect to a norm for English. The norm is based on an average for eleven varieties of English including the six sublanguages here.

+ + = far more frequent than in standard English: salient feature
+ = significantly more frequent than in standard English
A = average frequency for the language composite
− = significantly less frequent than in standard English
− − = rare or not used in the sublanguage: salient feature

English and French, several linking devices were studied, and their frequency noted for each sublanguage. Figure 7 summarizes the results of this study, giving the relative frequency of each type of cohesive link for the six sublanguages L_w, L_{ws}, L_s, L_p, L_{ah} and L_r.

One of the most striking aspects of this comparison is the fact that certain linking devices, such as pronominalization, are apparently not used in a number of technical sublanguages. These sublanguages tend to be those where the purpose of the text is the most rigidly defined (e. g. L_w, L_{ws} and L_{ah}; perhaps also L_s). The interesting counterexample is in recipes, where a moderate amount of pronominalization occurs. Considering the high degree of co-reference in recipes (and the fact that NP-deletion is much more frequent than pronominalization), it is surprising perhaps that L_r doesn't have more.

It should be noted that certain important devices for making texts cohesive do not appear in the list of figure 7. In particular, the word order in a sentence may be due largely to the topic structure of a text segment. Thus many "topicalized structures", such as those listed in figure 6, can indicate textual cohesion. But since word order may also depend on length of constituents and other factors, it is not always easy to decide which occurrences of passive sentence structure, for example, are due to the topic

structure of the text. Thus passives are simply listed with the sentence structures and topicalization is not represented among the linking devices. Refined techniques of substitution testing may eventually allow a direct measurement of the contribution of word order to textual cohesion.

4.4. Comparisons of English and French

The overall inventory of grammatical structures in English is rather vast. Only a part of this inventory is actually used in technical and scientific writing. From this still-large number of structures used in written language, each sublanguage seems to make a rather idiosyncratic selection. But it is particularly striking that the idiosyncratic selection of structures in French sublanguages should resemble those in English so well. In the relatively narrow technical and scientific sublanguages studied so far, when some structural type is absent from an English sublanguage, it is usually absent from its French counterpart. In virtually all the cases where parallel structures for English and French can be identified, we see the same relative frequencies for the structures used in parallel sublanguages. Thus a frequency profile for the six French sublanguages analogous to the profiles of figure 6 looks very much the same. This is all the more important when it is seen how wide the variation between the sublanguages of each language can be. This is not to minimize the remaining differences between English and French, but only to stress that each sublanguage seems to move away from its respective language norm in the same way when cross-linguistic comparisons are made.

Likewise, when textual cohesion devices are used as a means of comparison, we again find striking similarities between English and French within corresponding sublanguages. When an English sublanguage is entirely lacking in some linking device (e. g. pronominalization in technical manuals) the same is true for the French. Linking devices which are salient tend to be the same in parallel sublanguages. Variation from the language norm is the same for virtually all linking devices studied.

5. *Homogeneity of Sublanguages*

The use of sublanguage analysis in existing computational systems has tended to presuppose that texts treated by the system will be rather homogeneous as to domain of reference and purpose. In order for a sublanguage grammar to be drawn up, information is needed on the possibilities of occurrence of each word which will figure in the corresponding lexicon. These patterns of word usage are only discernable from a large corpus of homogeneous texts, although they may be verified and somewhat extended by using the intuitions of a "speaker" of the sublanguage. Word

classes in the sublanguage grammar are established on the basis of similarities of word usage, also requiring a substantial number of occurrences of each word in order to safely postulate class membership. If more sublanguages have not been explored, and wider sublanguages have not been used in processing, it is mainly because of the enormous investment of time and discipline required in investigating even a small portion of language coherently and in such detail.

In undertaking the study of a given sublanguage, it is natural to ask to what extent a given corpus of texts can be called "representative" of that sublanguage. Clearly, the answer to this question will depend on the way in which the boundaries of the sublanguage are defined. Even at this early stage of sublanguage research, some comments may be useful.

The idea of a "representative" corpus is related more to the description of the sublanguage *grammar* than to the lexicon. What is desired is a large enough view of the sublanguage to set up all the necessary categories for its description, and all the admissable sequences of categories which are to be admitted as sentences. Furthermore, all the subcategories necessary for stating the lexical selection between verbs and their arguments, nouns and their modifying adjectives, etc. must be discernible. When this much about a sublanguage is known, any new words encountered will fit into existing classes and subclasses, allowing their total behavior to be inferred.

The idea that it makes sense to look for a representative corpus is due to a feeling that sublanguages are relatively closed systems. On a practical level, this means that if one collects a steadily growing corpus of texts, $C_1 \subset C_2 \subset \ldots \subset C_n \subset \ldots$, then the sequence of linguistic descriptions $G_1, G_2, \ldots, G_n, \ldots$ will eventually converge (remain unchanged) on "the" grammar of the sublanguage, after some n. From the discussion above, we see that it is possible for a sublanguage to be grammatically convergent without being lexically convergent. A simple example is L_w, where the grammar is essentially known, but new place names may be encountered.

Clearly, the question of closure is a relative matter. Imperfect sublanguage descriptions can be tolerated in computational applications; it is up to the system user to specify an acceptable level of system failure. A computational linguist can usually judge, by the rapidity of convergence over a preliminary sample, whether a sublanguage will be sufficiently convergent to warrant practical computational treatment.

5.1. A poorly convergent sublanguage – L_s

Some perspective may be gained on the question of sublanguage convergence by considering the apparently marginal case of stock market reports. There are two particular difficulties in working with these reports.

First, the style and content of stock market reports seems to vary quite sharply depending on the place of publication (and hence level of user) of the reports. Most newspapers publish what may be called a market summary, which may vary in complexity, but where the grammatical organization clearly shows that the major concern is the trading activity and results on the various stock exchanges. A different kind of report, found in more specialized publications, can be called an analytical report. Here, the market activity is seen as both dependent on and an indicator of the general economic health of the society. Although analytical reports would be treated differently in a computational system, they share enough similarities with summary reports as to leave open the question of their classification in or out of L_s (defined on the basis of summary reports).

A second difficulty in examining L_s is the relatively slow convergence of this sublanguage, considering the relatively simple semantic domain. As seen in the sketch of section 3, subordinate clauses may introduce grammatical structures and lexical material which are much less restricted than the main clauses, which refer to the basic activity of the stock exchanges. The grammatical and lexical convergence of the extra-market clauses is much slower than that of the market clauses. This fact suggests that L_s can be described as the embedding of a rather narrow sublanguage, let us call it l_s, within a much broader variety of language which shows substantially less of the semantic and grammatical restrictions that we normally associate with sublanguages. If l_s were not easily indentifiable within L_s, this distinction would not be a useful one.

In a typical market summary report from L_s the juncture between l_s and the matrix portion of L_s is usually limited to one of the following cases:

(a) The matrix material is limited to a non-restrictive relative clause (often in present tense) which is in apposition to a proper noun. The proper noun may play a different semantic role in the main clause than in the restricted relative. The main clause, from l_s, is in past tense:

(27) Calgary Power, which is seeking a rate hike, added ½ to close at 39⅞.

main clause (in l_s)	subordinate clause (extra-market – outside l_s)	main clause (in l_s)

(b) The extra-market material may be adjoined as a complement of a noun from the class N_{news} = {*news*, *report*, *worry*, *rumor*, etc.}. Thus sentence S_3 of the report in figure 5 contains the fragment:

(28) . . . in a sell-off touched off by the news *of a sharp boost in oil prices planned by the Organization of Petroleum Exporting Coutries*.

A *that*-complement clause following the noun is not restricted for tense:

(29) Trading in Maple Leaf was halted on news *that Norin Corp. plans an $18-a-share offer for the shares of the company that it doesn't already own.*

(c) The extra-market material may be grammatically subordinated by the nominalization of a sentence from the matrix for use as prepositional object in a main clause which is otherwise in the restricted l_s:

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

(d) The extra-market material may be introduced in a subordinate clause headed by one of the conjunctions in a small set including *as*:

(31) Stocks fell sharply across the board on Canadian and New York exchanges in the early going today *as investors worried about the effects of the nuclear power plant accident and the lockout against Teamster members by the trucking industry in the U.S.*

These subordination mechanisms are not used exclusively for introducing extra-market portions of the text. They also serve to subordinate other market events of a secondary nature, such as trends of the market on previous days. Nevertheless, all extra-market material seems to be adjoined by one of these subordination devices. It is thus possible to propose a computational treatment in which only the "core" clauses of l_s are used. It is reasonable to say that l_s is an identifiable sublanguage embedded in L_s since the facts of grammatical subordination, the tense restriction patterns, and the different semantic domains all conspire to differentiate a special subpart of L_s.

5.2. Other sublanguage embeddings

The fact that a more regular subpart can be identified within a sublanguage is not an isolated fact about L_s. During the discussion of L_{ws} in section 3, it was observed that the variation of semantic tense within weather synopses could be correlated with the change from statements about meteological events to statements about the process of observing and predicting those events. In this case, a grammatical analysis shows that the event statements are embedded under observation predicates when both types of material appear in the same sentence.

The situation in L_{ws} closely resembles what was found by N. Sager and her colleagues (see chapter 1) in the sublanguage of cardiac glycosides. Some verbs could be ascribed to the process of observing, explaining and predicting the phenomena of pharmacology while others belonged to the statements of fact about the phenomena themselves. This distinction

between science and meta-science predicates correlates with a difference in the grammatical embedding patterns. Science predicates may be embedded under other science predicates or under meta-science predicates, but one never finds a meta-predicate under a science predicate. What makes the distinction easier in pharmacology is the lexical separation of the two classes of predicates. One rarely finds the same predicate in both components of the sublanguage. This lexical division is not so clear-cut in stock market reports, although this may be due to the fact that one rarely finds verb complementation as a way of embedding one component in the other. A few verbs, such as *drop*, may be used both in l_s and in the extra-market portion of L_s, even in reports from the same source:

(32) The oil and gas index, which had *dropped* over 91 points in the three previous sessions, recovered 21½ points.

(33) IT & T, which *dropped* plans for a joint venture in the consumer electronics field, was up ¼ at 28⅜.

Whereas *drop* is only intransitive in l_s (the core), it may easily have different properties outside of l_s.

5.3. L_s derivable from l_s by right-adjunction of subordinate clauses

It should be pointed out that an embedded sublanguage such as l_s (in L_s) can be regarded as the resultant of adjoining certain subordinate clauses from outside l_s to clauses in l_s. In the case of non-restrictive relatives, nominalizations and noun complements (a–c) of 5.1 above, the surface subordinate structure is derived from such an underlying $S_1C_sS_2$, where S_1 is from l_s and S_2 is outside l_s. This is a clear illustration of Harris' statement (1968):

> ". . . If S_1 is in the sublanguage and S_2 is not, $S_1C_sS_2$ retains properties of S_1 and is in the sublanguage. But in some cases this holds only for those conjunctions which require strong similarities; or else it holds if the special grammatical properties of the sublanguage are defined only on the first S of each of its SCS . . . . CS."

Whereas the type (d) structure of 5.1. is already in the form SCS, the other three structures observed are produced from SCS structures by operations readily available in Harris' transformational analysis.

6. *Implications for Automatic Language Processing*

The view of sublanguage variation and homogeneity that comes out of this survey gives cause for guarded optimism as to the prospects of computational treatment of texts from restricted domains.

6.1. Automatic parsing within sublanguages

The fact that sublanguage grammars differ so widely within the same language underscores the need of carrying out a precise linguistic study on each sublanguage for which computational treatment (on the level of syntax or semantics) is planned. Although many of the differences are due to differing frequencies in the usage of structural types, the presence in some sublanguages of structures which are unknown in the standard language indicates that no single parsing grammar will be adequate for all types of text (dictionary problems aside).

A refined sublanguage profile stating the relative frequencies of different sentence and text structures for a given sublanguage could nevertheless act as a kind of probabilistic template which could be added to a general parser for a sublanguage which differs from the standard language only in its choice of standard structures. Information of the kind given in figures 6 and 7 could help to organize the parsing strategy within the sublanguage. In an augmented transition network parser, for example, arcs are sometimes ordered according to likelihood, and this is clearly a function of the sublanguage.

One of the most difficult problems in parsing is the resolution of anaphora. It is encouraging to see that some of the most technical sublanguages, such as L_{ah}, completely avoid pronominal anaphora, and make use of indices as a device to ensure non-ambiguity. In the case of less controlled texts where pronominal anaphora does occur, it may be possible to improve resolution strategies by studying the scope and functioning of anaphora in the particular sublanguage. For example, during the study of linking devices, it was found that pronominalization as a linking device was generally limited to reference in the immediately preceding sentence. The few exceptions to this occurred in the least technical or semantically restricted varieties. Knowledge of the characteristic scope for anaphora within the sublanguage will make the calculation of tradeoff points possible when two or more candidates for a pronoun's antecedent are available.

A second serious problem in text processing has been lexical ambiguity. Although lexical ambiguity is much less of a problem when each word is allowed to take on only the meanings and functions possible in the sublanguage, there are still cases of polysemy in many sublanguages. In many cases an ambiguous word can be matched in one of its meanings with a synonym or hyponym occurring in a nearby sentence. But in order to pick among two or more possible interpretations, one must know whether for that sublanguage such semantic links are frequent, and if so, over what distance in the text. It has been found that semantic links (such as synonymy) can occur with a greater scope in text than can grammatical links (such as pronominalization or deletion used cohesively), and that such linking over three or more sentence boundaries is quite possible.

But the frequency of such linking depends very much on the particular sublanguage.

6.2. Automatic translation

One of the most striking results of the contrastive English–French sublanguage comparisons was the parallels in frequency of sentence type and linking device for corresponding sublanguages. That this should occur in the face of a generally sharp variation for different sublanguages of the same language is all the more evidence that purpose of text and semantic domain have a powerful influence on text and sentence structure. With such strong parallels, particularly for technical sublanguages, it is possible to feel fairly optimistic for the prospects of automatic translation in restricted domains.

When we compare the parallel sublanguages more closely, we begin to get a clearer picture of what it is that accounts for this parallelism. The structural similarities between recipes and aviation manuals are rather strong, and are shared by other types of manual or assembly instruction found in everyday life. Deletion patterns in individual sentences and types of textual linking are quite similar, although the different length of sentences and complexity of the subject matter seems to account for different scope of linking in these texts. This situation is in contrast to what we find if we compare English weather bulletins with weather synopses. Despite the strong semantic similarities (e. g. the words for weather conditions, place names, temporal expressions), there are very few structural similarities in these sentences. A fairly common semantic domain has little to do with structure at the sentence or text level, apparently. The major similarities, apart from partly shared vocabulary, are resemblances of lexical selection (adjectives with nouns, adverbials of time and place with descriptions of condition).

One is therefore drawn to conclude that English and French technical texts show the strongest parallels because the text purpose is more similar here than in descriptive texts. Weather reports, recipes and aviation manuals, which show the strongest parallels, all have very well-defined text purpose. Most of the unexpected structures one finds in a sublanguage text can be associated not so much with a shift in semantic domain as with a shift (usually quite temporary) in the attitude which the text producer takes towards his domain of discourse.

6.3. Generation of sublanguage texts

One of the most obvious applications of the observations given above on sentence and text structure within sublanguages is to the generation of text from a semantic base. It seems quite probable that most applications of text

generation systems will be confined to a single domain, requiring that the text be well-formed in a particular sublanguage. A text generated by the most straightforward techniques will have a great deal of lexical repetition. Stylistic improvements will require use of synonyms and hyponyms, depending on the sublanguage (e. g. stockmarket reports will require that virtually all lexical repetition be replaced). In order to make these improvements properly, information is needed on the frequency and possible scope of synonymy and hyponymy in the sublanguage. Although general language rules may work in some sublanguages, others have sufficiently specialized kinds of linking that only sublanguage-specific tendencies will give natural-sounding texts. The kind of formatted data base that Hirschman & Sager describe in chapter 2 could be used as input to a text-generation program which, using a sublanguage grammar, would produce such a well-formed text. This might even serve as an alternative to translation of texts in the case where a text is produced in a tightly controlled technical situation. Particularly in the cases where texts are produced daily on the basis of numerical or other easily quantifiable information (weather reports, stock market reports), it would be feasible to generate texts in one or more languages by using what is known about a sublanguage's peculiarities of sentence and text structure. The fact that the same information format seems usable for technical languages of English and French would encourage this as an alternative to translation for semantically simple sublanguages.

6.4. Computation in sublanguages with embedded cores

At the moment it is not clear whether there are many sublanguages which are describable in terms of embeddings. But it seems very likely that this is the case. Sublanguages of mathematics, for example, almost certainly can be found with one or more levels of embedding.

It may be turn out to be quite useful to be able to carry out automatic processing of embedded sublanguages, even if the same amount of processing cannot be done on the grammatically subordinated matrix. The embedded core of a stock market report, for example, carries an important component of information, useful even without the matrix information.

If the junctures between embedded sublanguage and matrix are identifiable, the strict sublanguage grammar rules of L_s (for example) can be limited to apply only on those clauses or clause fragments of the embedded sublanguage. When the parser scans matrix material, a more general parsing strategy and lexicon can be called upon. Such a dual processing strategy would make a processing system more powerful and less vulnerable to unfound words and structures, since as a rule these would occur in the loose matrix, which is less closed as a system.

Even in sublanguages where both L_x and l_x can be described to the same degree of detail, it will certainly be more efficient to exploit the two levels. Strategies of anaphora and ambiguity resolution will certainly be sharpened by constraining some aspects of the resolution procedures to operate in only one text component.

Acknowledgment

This work was supported by the Social Sciences and Humanities Research Council of Canada under grant no. 410-79-0070.

References

[1] CHEVALIER, M. et al. (1978): *TAUM-METEO: description du système*. Groupe de recherche en traduction automatique, Université de Montréal.

[2] GROSZ, B. (1978): "Discourse Analysis" chapter 5 of this volume.

[3] HARRIS, Z. (1968): *Mathematical Structures of Language*. Wiley-Interscience.

[4] HIRSCHMAN, L. & SAGER, N. (1980): "Automatic Information Formatting of a Medical Sublanguage" chapter 2 of this volume.

[5] KITTREDGE, R. (1970): *Tense, Aspect and Conjunction: Some Interrelations for English*. T.D.A.P. papers no. 80, University of Pennsylvania.

[6] KITTREDGE, R. (1978): "Textual Cohesion within Sublanguages: Implications for Automatic Analysis and Synthesis" proceedings of the COLING 78 conference on computational linguistics, Bergen, Norway.

[7] LEHRBERGER, J. (1980): "Automatic Translation and the Concept of Sublanguage" chapter 3 of this volume.

[8] SAGER, N. (1972): "Syntactic Formatting of Science Information" chapter 1 of this volume.

[9] SAGER, N. et al. (1970): *String Project Reports*. Linguistic String Project, New York University.