TEXTUAL COHESION WITHIN SUBLANGUAGES:
IMPLICATIONS FOR AUTOMATIC ANALYSIS AND SYNTHESIS*

Richard Kittredge

Université de Montréal

---

* Paper presented at the Seventh International Conference on Computational Linguistics (COLING 78), Bergen, Norway on August 17, 1978.

**This work was supported in part by a grant from the Social Sciences and Humanities Research Council of Canada.

1. Introduction

Much recent work in artificial intelligence is concerned with simulating human linguistic behavior in restricted semantic domains or micro-worlds. The current concern with knowledge representation and inferencing mechanisms has often tended to downplay the importance of purely linguistic description.

Despite the need for a full knowledge representation in building intelligent language processors, very promising intermediate results in text processing are obtainable when the linguistic analysis of a corpus of domain-related texts is used to design the processing system. For years it has been known that the language used in a narrow scientific or technical field has special grammatical and lexical properties which justify calling it a "sublanguage". A few research groups are now using sublanguage analysis with surprisingly good results in machine translation (MT) and information retrieval (IR). Distributional analysis of a representative set of texts in the subfield allows a redefinition of the syntactic classes and semantic subclasses which are active in the structural patterning and lexical selection of the sublanguage sentences. In the few sublanguages analyzed so far the specialized grammars have differed sharply, but have given, in each case, results which are a good approximation of intelligent processing. It is natural to ask just how general these results may be.

This paper is, first of all, a report on a linguistic study of the structural diversity of sublanguages. In a survey of textual cohesion within several sublanguages of English and the corresponding sublanguages of French a clearer view emerged of (1) the variation between sublanguages Of the same language, (2) the relationship between the grammar of a language and the grammars of its sublanguages and (3) the striking similarities between languages like English and French when the comparison is made within parallel sublanguages.

The remainder of the paper 1s a discussion of some of the implications of this sublanguage survey for automatic language processing. Although much more detailed analyses are needed for individual applications, these findings clearly indicate that the variation between dissimilar sublanguages of one language is quantitatively and qualitatively greater than the variation across language boundaries for the same sublanguage. This fact, and the details of the comparison, have implications for the design of intelligent analyzers and synthesizers and for the choice of promising sublanguages. Furthermore, they allow us to take a more optimistic view of the prospects for fairly good (i.e., revisable) MT 1n a number of areas, even without a representation of extra-linguistic knowledge.

2. A Study of Textual Cohesion within Sublanguages

2.1. Motivation for the study

This research was undertaken to answer some specific questions which arose in my previous work on machine translation. During 1974-1976 I was involved with designing two experimental systems to translate from English to French, one for weather bulletins and the other for aircraft maintenance manuals. In parsing such special varieties of English one is faced with grammatical constructions which seem to break or bend the "normal" rules of English syntax. Sentences (l.a.) and (l.b.) are grammatical in the sublanguages of weather bulletins and aviation hydraulics respectively, but not in general English. .

(l.a.) Clear and cooler this evening with chance of snow flurries.

(l.b.) Press filler cap down and hold until internal pressure is relieved.

In both cases the sentences of the sublanguages can be derived from more normal English by certain deletions, as indicated by (2.a.) and (2.b.).

(2.a.) (The weather will be) clear and cooler this evening with (a) chance of snow flurries.

(2.b.) Press (the) filler cap down and hold (it down) until (the) internal pressure is relieved.

Although it might be possible to first recover the deleted material and then subject the expanded text to a general parsing grammar, it seemed more economical to write a specialized grammar stating the allowable combinations of word classes directly. This was all the more reasonable since it appeared that both sublanguages could be described (at least in the first approximation) by using a much smaller number of rules than are necessary for general English.

And even if most sublanguage sentences could be derived by deletion from general English, there were occasional structure types which could not be so derived.

(3) Winds southwesterly 15 to 25 becoming light this evening.

In addition to the deletions typical for the sublanguage, (3) contains _becoming_ in a construction which cannot be related to an expanded source. A plausible source such as (4) is ungrammatical.

(4) *(The) winds ... (will be) becoming light ...

In a number of such cases, there is no alternative to writing special grammatical rules.

What was even more striking than the existence of such patterns in English was the fact that French texts from the same sublanguages showed very similar structures. The stylistic parallels between English and French technical manuals were so strong that translation was often possible on the level of phrase structure. English passives could almost always be translated by passives in French, even though the use of passive is much more restricted in general French. Was it just lucky coincidence that two sublanguages had been chosen where MT seemed easier than in general text (say, newspaper editorials)? A linguistic study of parallel sublanguages seemed in order.

In the fall of 1977,with the support of the Canada Council, colleague Rajendra Singh and I set up a small project in the linguistics department to study textual cohesion in a number of English and French sublanguages.

2.2. Previous work on sublanguage

Informal descriptions of the styles or "varieties" of English have been available for some time.[1] Perhaps the first rigorous definition of "sublanguage" and formal characterization of typical linguistic subsystems was given by Harris (1968) as an outgrowth of his work on discourse analysis and transformations. For Harris a sublanguage is a subset of the sentences of a language closed under the transformational operations. In this formulation transformations map sets of sentences onto sets of sentences. The negation, clefting, conjunction, etc. of sentences in organic chemistry gives other sentences of organic chemistry. Strictly speaking, the set of sentences that constitute a particular style can be a sublanguage. More interesting sublanguages are those which have lexical co-occurrences, structural types and even transformations not found elsewhere in the language.

The first major application of Harris' methods to an extensive corpus of sublanguage texts was made by Naomi Sager (1972) and her colleagues for scientific reports on the pharmacology of cardiac glycocides. Distributional techniques have allowed identification of the word classes which enter into the important structural relations for the sublanguage. Applying a Harris-ian transformational decomposition in terms of these classes gives the elementary information-bearing propositions for this scientific subfield.

This early work on sublanguage left two important questions unresolved.

First, since most apparent sublanguages exhibit some degree of "seepage"[2] from the general language, how does this affect the theoretical definition of sublanguage and the applicability of sublanguage models to practical problems? To some extent, by separating out the meta-science component of in science sublanguage, as Harris suggests, one can clarify and delimit the kinds and points of seepage. Presumably, the usefulness of sublanguage models for translation and information retrieval is **directly** related to the preciseness of statement of word class co-occurrence, so some perspective on the overall typology of sublanguages is called for to evaluate the fruitfulness of the SL approach.

Secondly, none of the early studies investigated the way in which different cohesive links are used in various sublanguages. Little was said about the cohesive role of surface structure. A study of the typology and frequency of various "linking devices" should contribute to a characterization of the sublanguage style. The distribution of the links should be related to the structures and word classes needed for intra-sentential description.

2.3. A Corpus for Eleven Sublanguages of English and French

In order to make a survey of textual cohesion devices across a spectrum of sublanguages, eleven varieties of English were chosen along with the corresponding eleven varieties of French (see figure 1). These ranged from the tightly structured sublanguages of aviation hydraulics and stock market reports to the much more open varieties of economics texts and even children's stories. The aim was not to make a detailed analysis of any given SL, but to get an impression of the range of variation between sublanguages and between English and French. For such a broad survey, the question of defining "sublanguage" or of distinguishing lexically and semantically neat sublanguages from fuzzier "varieties" was considered secondary. Rather, a comparison of several sharp and fuzzy varieties with respect to textual linking could help make the sublanguage notion clearer.

Each SL was represented by one or more texts with a total length of approximately 100 sentences. Since linking between consecutive sentences was to be studied, the length of each sample was standardized at 100 adjacency pairs, meaning 101 sentences in the case of one text and 100 + n sentences when n texts had to be strung together to give the desired length. In

SUBLANGUAGES STUDIED

(for written, edited English & French)

ECONOMICS:

MICRO-ECONOMICS

MACRO-ECONOMICS

STOCK MARKET REPORTS

TECHNICAL MANUALS:

AVIATION HYDRAULICS

COOKING RECIPES

ADMINISTRATION:

UNIVERSITY CATALOGS

METEOROLOGY:

TELEGRAPHIC FORECASTS

SYNOPSES OF REGIONAL CONDITIONS

EXPERIMENTAL SCIENCE:

PHARMACOLOGY - CARDIAC GLYCOCIDES

"LITERATURE":

CHILDREN'S STORIES

LITERARY CRITICISM

Criteria for choice:

(1) variety of linking devices

(2) variety of intra-sentential forms

(3) availability of parallel texts for E & F

(4) interest for automatic processing

(5) availability in machine readable form

Figure 1.

several cases the English and French **texts** were actually "translation twins" where one was the high-quality translation of the other.[3]

2.4. Linking devices

Cohesion was studied first in terms of links between consecutive sentences. A taxonomy of some twenty linking devices (LD's) was drawn up with appropriate groupings (figure 2). Text grammarians might wish to distinguish *grammatical* from *lexical* LD's. Anaphoric devices such as pronominalization, definitization and comparatives and other degree terms can link two separate sentences, and their occurrence can be described in terms of derivation from some canonically simpler (perhaps semantically more transparent form). This is not the case when two sentences are linked by word repetition, synonymy or other semantic relations which can exist between two lexical forms where neither can be taken as "simpler", "more primitive" or "semantically more transparent". Deletion and a number of "topicalization devices" are also listed as grammatical in that they can be described as resulting from simpler canonical sentence forms. Computational linguists might prefer to see any distinctions in terms of *explicit* versus *implicit* linking. Imagining an intelligent language processor which reads a text from left to right, carrying out all possible processing at read time, one can see the need for a list of explicit signals that a link is being made with (usually) preceding text. Most lexical linking (but not conjunctive adverbs or words like *firstly*) is implicit, in that one would have to compare semantic representations of each word with all preceding words to detect a link (unless, as with definitization, the explicit the is accompanied by lexical linking).

The first and most tedious quantitative analysis involved simply counting occurrences of each LD in each sublanguage.[4] The results are discussed in 2.6 below. When a non-initial sentence contained a topicalized

TAXONOMY OF LINKING DEVICES

GRAMMATICAL LINKS:

- Anaphora
  - pronominalization
  - definitization
  - degree terms
  - special indices
- Topica1ization devices
  - adverb preposing
  - subordinate clause preposing
  - passive
  - clefting
  - **reflexive and pronominal verbs**
  - dislocation
- Deletion

LEXICAL LINKS:

- Conjunctive adverbs
- Enumeration
- Repetition
- Morphologically derived forms
- Synonymy
- Hyponymy
- Part-whole
- Contraries and contrastives
  - converse terms
  - antonyms
  - process reversal
  - etc.
- Semantic field (residual relations)

Figure 2 . The major types of linking between consecutive sentences can be divided into grammatical and lexical links depending on whether the link is signaled by a derived structure or by the semantic representation of one or more lexical items. The first two lexical types are also explicit signals in that each occurrence of such a word or locution signals a link with (usually preceding) textual segments outside the same sentence.

structure, it was sometimes difficult to determine whether the inverted order was due to cohesive linking. Other factors, such as length of constituents, can also affect order. Therefore, all occurrences of topicalized structures in main clauses were counted.

Another weak textual link (not counted in the above survey) is uniformity of tense or time reference. Since each main clause carries a tense, it makes no sense to count this as a positive link, but rather only to note how changes in temporal reference are signaled, and some of the global properties of temporal reference in sublanguage texts.

2.5. Some sample texts

Among the eleven sublanguages studied, several showed the kind of structural patterning and restricted lexical co-occurrence that set them apart from the general language and make them particularly interesting as linguistic systems. Consider the sample stock market report given as figure 3. Typical of such texts is the metaphorical use of motion verbs to describe changes in price for certain commercial papers. Words like stocks, issues, securities, golds, etc. belong to a special lexical class for the sublanguage which plays a central role in the information structure of such texts. The subject of an intransitive motion verb, or the object of a transitive motion verb is a member of the set {stocks,...} or of another very restricted set. When the distribution of motion verbs is analyzed more closely, it becomes possible to state the restrictions more precisely. Another characteristic of this sublanguage is the co-occurrence or collocation of adverbs with motion verbs in ways which would be deviant outside the sublanguage, such as move up ... strongly. The terse style of stock market reports also favors deletion, both intrasententially as a characteristic of the SL and intersententially as a linking device. The third para-

SUBLANGUAGE OF STOCK MARKET REPORTS
(Sample Text Fragment)

Stocks move up

Stocks moved up fairly strongly on Canadian exchanges yesterday as the rally on Wall Street ran into some profit-taking.

The MSE industrial index closed with a small gain while the TSE composite zipped up nearly 5 points. All but one of the 14 major groups within the composite index advanced.

Biggest gainers were consumer product, transportation, pipeline, utility, communication, bank and merchandising issues. Golds eased as bullion prices dipped below the $165-an-ounce level.

•
•
*

Figure 3. The sublanguage of stock market reports is characterized by the frequent usage of motion verbs in collocation with the important set of nouns including stocks, issues, securities, etc. This metaphorical usage is marked by adverbial collocations which would be uncommon with motion verbs used literally, e.g., move up ... strongly. The verb ease in the last sentence takes the same noun set as subject and is used only intransitively. Characteristic deletions for the sublanguage include deletion of definite article the as at beginning of the third paragraph. Use of superlative in *biggest* is cohesive.

graph of this text begins with *biggest,* where the deleted *the* is dropped because of sublanguage style and not because it repeats material present in a preceding sentence. The use of the superlative form in *biggest* is cohesive, referring to *all but one of the 14 major groups* in the preceding sentence.

Another sublanguage which exhibits neat structural and lexical organization is that of aviation hydraulics. A sample fragment from an aircraft maintenance manual is given as figure 4. These manuals contain at least two important textual subtypes, both dealing with the same semantic domain. In this sample, paragraph 22 (PRESSURE SWITCH) gives a general description of the operation and main features of a particular aircraft system. Here there is little deletion and the sentences have the familiar declarative structures of general English. Paragraph 23, however, gives a completely different subtype containing procedures for maintenance of the system under discussion. Deletion, although not necessarily cohesive deletion, is quite typical. Definite articles and definite object noun phrases are regularly deleted when this does not introduce ambiguity. The imperative sentences (a)-(d) all show *the*-deletion (indicated by Ø). But sentence (e) retains *the* in *the two mounting bolts,* presumably since deletion would introduce a serious ambiguity.

A detailed analysis of this particular sublanguage has recently been carried out by the TAUM machine translation group and in particular by Lehrberger (1978). Despite the structural and semantic complexities of texts in this sublanguage, the rigid style and relatively precise lexical groupings make translation within this sublanguage feasible. The TAUM-AVIATION system currently under development is detailed in Isabelle et al. (1978).

SUBLANGUAGE OF AVIATION HYDRAULICS
(Sample Text Fragment)

PRESSURE SWITCH

22 Two identical pressure switches, one in each system, arc 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 pressure port.

(d) Disconnect Ø line at drain port elbow.

(e) Loosen the two mounting bolts and remove Ø switch.

Figure 4. The aviation hydraulics sublanguage (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 of definite articles and repeated object noun phrase 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.

A technical manual of a different sort is the well-known cooking recipe. This sublanguage is particularly interesting because of the wide availability and tendency toward spontaneous generation in most languages with written traditions. Figure 5 gives a French soup recipe showing the typical division into a list of ingredients followed by a sequence of procedural "assembly instructions". Some informal recipes begin with a general discussion and this tripartite organization is exhibited in the American bread recipe given as figure 7. (Aircraft manuals often insert a list of components before the procedural section, giving them a similar three-part subdivision). These recipes show a pattern of deletion for definite object noun phrases when in a non-initial procedural sentence. French recipes typically show a progression from full NP in the first sentence, through pronominalization, to deletion of the entire NP. English recipes seem to move to full deletion more rapidly. The same general pattern can be observed in other languages of the Indo-European family. Additional types of deletion occur in English recipes, particularly on definite the, although this is not cohesive.

Another sublanguage of particular interest is that of regional weather forecasting. Unlike the telegraphic weather bulletins which use sentences without tensed verb, the synopses are sequences of coherent paragraphs dealing with the general movement of air masses over the continent with some discussion of their effect on local conditions. The sample in figure 6 illustrates the semantic and linguistic restrictions found in these texts. The domain of reference is quite restricted in terms of properties. Observations about physical description and cause-effect predominate. One particular property of these texts is their shifting temporal reference. Whereas technical manuals are quite uniform in tense

Soupe des «Halles>>

250 g. d'oignons, 95 g. de beurre,
30 g. de farine, 1 litre ½ d'eau chaude
2 cuillères à café de sel, 10 tours de moulin à poivre,
une douzaine de tranches de pain,
100 g. de gruyère râpé.

Epluchez les oignons et hachez-les finement.
Dans une casserole à fond épais, faites-les cuire
avec 75 g 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
laisscz blondir. Mouillez Ø avec de l'eau chaude et
assaisonnez Ø .

•
•
•
•

Figure 5. Cooking recipes typically show deletion of definite object noun phrases, which occurs when the same NP is repeated in successive sentences. This French sample shows a gradual progression from full NP, through pronominal forms, to deletion of entire NP. English recipes typically delete object NP's after the initial occurrence. A wide variety of other languages show similar object deletion. English recipes often delete definite articles in addition. Notice that the definite les in the first sentence of the above instructions indicates cohesion with the list of ingredients.

SUBLANGUAGE OF METEOROLOGICAL SYNOPSES

(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 those amounts.

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

Figure 6. Meteorological synopses consist of one or more cohesive paragraphs made up of full sentence forms. Unlike meteo bulletins, which are telegraphic and lack tensed verbs, synopses show cohesion in the tense repetition on verbs. *Should*, which cannot signify obligation in this sublanguage, is analyzable as *will probably* and *could* as *will possibly*. Other samples exhibit a progression from past tense (in opening sentences describing the recent meteorological events) to future time reference during the later paragraphs (giving the prognosis). Note that predicates which make an observational statement involving human evaluation (as opposed to a purely descriptive statement) may be in present tense. Thus the repetition of semantic tense is more uniform within the sequence of description statements. This and other facts support an analysis which separates the two types of predicate.

and therefore uninteresting in this regard, synopses show either a progression of time reference (past to future) or uniform future reference. However, it becomes clear that there are two levels of text. The first is the sequence of observations about physical phenomena which constitute the essence of the synopsis, and it is this sequence which shows tense uniformity. Operating on these propositions there may be predicates describing the act of observation itself. These are typically in the present tense (in agreement with the time of reporting as opposed to the time of the phenomena). Separation of the two levels of description is important for automatic processing of texts, since shift of time reference can be used to detect shift of phenomena description only when the existence of these two levels (science and metascience) is taken into consideration. For the text of figure 6 the meta-predicates and their tenses are underlined. When the remaining text is given the proper modal analysis (where *should* means *wil1 probably* and *could* means *will possibly),* the sequence of propositions about the physical world is seen to have uniform temporal reference.

2.6. Results for adjacent sentences

When frequency counts were made for each linking device within each of the eleven sublanguages of each language, the results were rather surprising. There appears to be wide variation in the use of the various linking devices when one compares sublanguages of one language. For example, English children's stories had 46 occurrences of pronominalization over 100 adjacency pairs, whereas the stock market reports, aviation manuals and meteorological synopses were all completely lacking in pronominalization. French children's stories likewise came out at the high end of the spectrum with a similar discrepancy. As for deletion, the recipes are relatively full (empty?) of it. English recipes had 172 occurrences in the sample, whereas the samples from pharmacology, literary criticism, university calendars and children's stories all had less than five. Results for French were quite similar except that recipes had far less (29), the difference being made up for by pronominalization to some extent. Although lexical linking was less frequent in absolute terms, here, too, there was a wide spread between sublanguages, and a fairly close correspondence across language boundaries. Synonymy ranged from zero to 53 occurrences in English and from zero to 24 in French, with most sublanguages preserving their approximate position in the scale across language boundary.

2.7. Scope of linking

As indicated above, initial analysis of the corpus was limited to consecutive sentences or adjacency pairs. For both theoretical and practical reasons it is also important to know over what distance in the text (measured in number of sentences, words or morphemes, etc.) each linking device may be used. Some quantitative measure of LD scope for a specific sublanguage would obviously be of use in designing parsing strategies for that sublanguage.

At the outset of the study my hypothesis was that the scope of grammatical LD's (such as pronominalization or deletion) should be significantly less than the scope of lexico-semantic LD's. This would be in keeping with a general presumption that grammatical linking requires reference to the superficial syntactic structure of sentences, which is not retained in memory as long as are the deeper semantic relations between elements on which lexico-semantic linking depends.

Results in this area are quite preliminary. In general, however, they confirm the hypothesis. Counts were made of the scope of linking

SCOPE OF PRONOMINALIZATION

S$_1$ **Basic Wholewheat Bread**

INFORMAL INTRODUCTION

S$_2$ The best flour you can buy is stone-ground wholewheat flour.

....

INGREDIENTS

S$_6$ 9 cups warm water
1 cup honey
1/2 cup vegetable oil
5 tablespoons granule yeast
2 tablespoons salt
20 cups ... flour

INSTRUCTIONS

S$_7$ Allow ∅ yeast to soften in ∅ 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 ∅ onto ∅ floured board.

S$_{12}$ Add more flour if necessary.

S$_{13}$ Knead ∅.

S$_{14}$ Knead ∅ and knead ∅.

S$_{15}$ Knead ∅ until it feels good -- not sticky but warm and elastic.

....

Figure 7.

SCOPE OF LINKING

	ENGLISH			FRENCH		
	S_1	S_2	S_3	S_1	S_2	S_3
SYNONYMY						
Hydraulics	1	-	-	-	-	-
Pharmacology	12	4	1	2	4	1
Children's stroies	4	4	1	9	11	3
Meteo synopses	11	9	3	10	7	3
Totals	28	17	5	21	22	7
HYPONYMY						
Hydraulics	-	-	-	-	-	-
Pharmacology	1	-	-	-	-	-
Children's stories	-	-	-	1	-	-
Meteo synopses	4	2	6	5	1	2
Totals	5	2	6	6	1	2
DEFINITIZATION						
Hydraulics	9	-	1	7	1	-
Pharmacology	9	1	2	20	6	2
Children's stories	2	1	1	9	3	1
Meteo synopses	9	9	-	8	-	-
Totals	29	11	4	44	10	3
ENUMERATION						
Hydraulics	5	-	-	4	-	-
Pharmacology	12	5	4	4	4	3
Children's stories	2	-	-	3	2	2
Meteo synopses	1	4	-	2	4	-
Totals	20	9	4	13	10	5
PRONOMINALIZATION						
Hydraulics	-	-	-	-	-	-
Pharmacology	2	-	-	5	-	-
Children's stories	47	3	1	59	9	2
Meteo synopses	-	-	-	-	-	-
Totals	49	3	1	64	9	2

Fig. 8. Scope of several linking devices in four sublanguages. Frequency counts are given for separation by one, two and three sentence lengths for English and French.

(in terms of number of sentences of separation) for the following five LD's: synonymy, hyponymy, definitizat1on, enumeration and pronominalization. Only four sublanguages were included 1n these counts: hydraulics, pharmacology, children's stories, and meteo synopses. Results are presented as figure 8 for cases of separation by one, two and three sentence boundaries, where linking between consecutive sentences counts as separation by one sentence boundary. A typically semantic LD such as synonymy exhibits roughly the same frequency of separation by one as for separation by two sentence boundaries. In general, all lexico-semantic LD's, where the numbers were significant, showed a gradual decline as scope changed from one to three sentence lengths. Scope of more than three was not rare for lexico-semantic linking.

In contrast, grammatical linking such as pronominalization shows a significantly smaller scope. Pronominal linking with scope of more than one sentence was rare in all the texts studied.[5] More work is needed to confirm these results in other sublanguages. Since pronominalization was entirely lacking in the samples of hydraulics and meteo synopses, it was decided to examine another sublanguage. The microeconomics sample showed twelve occurrences of pronominal links (the same number for the French translation as for the English original) between sentences, all of scope 1. Since pronominalization seems to be relatively infrequent for the most restricted sublanguages of science and technology (for both English and French), further work may show that pronominalization can be related to one or more of the basic parameters for defining sublanguage.

An interesting exception to the preceding generalization about pronominalization occurred 1n the sample of English recipes (see figure 7). On superficial analysis, it appears that the antecedent of it in S_{15} is dough in S_{10}, a scope of 5. Note, however, that sentences S_{11}, S_{12}, S_{14} and S_{15}

have missing object noun phrase. Under an analysis where dough has been zeroed as object NP (and is available in underlying representation), the cohesive links can be established between the adjacency pairs: S_{10}-S_{11} (deletion), S_{11}-S_{12}(deletion) if one assumes that to the dough has been zeroed, S_{12}-S_{13}(deletion), and S_{14}-S_{15}(deletion). In fact, the object NP of knead is S_{15}, which is deleted by virtue of the preceding S_{14} and the it refers to this last occurrence, hence the syntactic antecedent of the pronoun is the deleted NP in the same sentence. The link is in fact intra-sentential. Clearly the phenomena of pronominalization and deletion are related and a full account of scope properties of linking devices of this type raises a large number of questions for the kind of linguistic analysis being used. But it is clear from this example that scope of pronominalization follows a more regular pattern when the cohesive deletions are reconstituted.

3. Some implications for Automatic Language Processing

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 ungrammatical in the language as a whole indicates that no single parsing grammar will be adequate for all types of text (dictionary problems aside).

On the contrary, parsers should be designed which exploit the special characteristics of the sublanguage in question. In an ATN parser where backtracking is to be minimized, for example, it is possible to order the arcs according to the frequency of structural types in the sublanguage. When enough is known about a variety of sublanguages, it may be possible to design a "tunable parser" which could allow recognition of a wide variety of structural types, but allow only certain ones to be activated at any given time, in accordance with the known characteristics of the sublanguage in question. Since it appears that there are dependencies between linking devices in any given sublanguage, knowledge of these relations may allow activating sets of grammatical rules together, or exclude certain combinations of rules from being activated together.

Information about scope of LD's in a sublanguage is clearly needed to help set up parsing and interpretation strategies. Resolution of anaphora and lexical ambiguity generally require a search through preceding text or some representation of that text. If the scope of anaphoric pronouns is highly restricted in the sublanguage, this will allow giving high priority to those possible antecedents which are within one or two sentence lengths of the anaphoric pronoun. Similarly, knowledge about the relative frequency and scope of lexico-semantic links will allow calculating the optimal trade-off point between possible lexical matches. To take a concrete example, suppose that a text in the stock market report sublanguage is being analyzed with a semi-intelligent parser. If the n-th sentence contains an ambiguous noun, say _issues_, and each sense can be linked to a preceding word in the text, say _golds_ and _problems_, then the choice of word sense for _issues_ (all other things being equal) will depend on what we know about the frequency and scope for synonymy and hyponymy.

One of the most obvious applications of the kind of linguistic description cited above is to the generation of text from a semantic base. It seems quite probable that most applications of such systems will require that the text be generated as well-formed in some sublanguage. A text generated by the most straightforward techniques will have a great deal of lexical repetition. Stylistic improvements will require use of some synonyms and perhaps occasional hyponyms. The pattern of usage, the frequency (relative and absolute) and the typical and maximal scope of each LD in such textual linking must be determined for the sublanguage in question. Although figures for the language as a whole may give reasonable texts in some situations, the sublanguage study shows that individual sublanguages may differ sharply from that norm. A text in a serious scientific or technical subfield which has the wrong kind and frequency of LD's will certainly sound "unprofessional".

The application of sublanguage analysis to machine translation is another promising area. Early work on the METEO and AVIATION systems clearly indicated the advantages for those sublanguages of using specialized grammars, setting up lexical classes based on co-occurrence patterns for the sublanguage, etc. Translation in these sublanguages proved feasible because of highly similar stylistic conventions in English and French, and because the limited semantic domain greatly reduced the possibilities of word-sense ambiguity. Because of the common semantic domain English and French technical manuals in aviation hydraulics set up very similar semantic subclasses, exhibit similar co-occurrence restrictions, and have the same descriptive and procedural organization. A question of considerable interest is how many of these similarities are due to close contact between the technical communities in these areas and how much is imposed by the subject matter itself. Since the French aviation manuals in our corpus were high-quality translations of English originals, (no spontaneous French text of the same kind being available), it is too early even to speculate on the reasons for the close similarities. Nevertheless, the view of a variety of sublanguages which results from this study is quite striking. When linking devices were counted in a variety of sublanguages of English and French, these was a strong similarity of frequency and distribution of LD's, even when texts were not translation twins but both spontaneously produced, native texts. Greater similarity was generally found among the sublanguages of a highly technical nature or where texts are oriented to a particular purpose (e.g., cooking, recipes, meteo bulletins, pharmacology experimental reports). This great similarity (greater than that found between dissimilar sublanguages in the same language) across language boundaries obviously means that the chances of successful MT are much greater when systems are designed around specific grammars (and of course lexicons) for the sublanguage. It also means that it is unrealistic to expect that a translation system designed for one sublanguage can be redirected with only a few alterations to texts in a new subject area. Of course, long-range progress in the understanding of the structure of sublanguages and their sources will enable us to "tune" whole translation systems to the appropriate subject area by adjusting sets of rules which turn out to be inter-dependent. It is quite unlikely, however, that some subcomponents will ever be anything but sublanguage-specific.

4. Further questions

A number of questions have been raised in the course of the past year's research which seem worth pursuing.

Texts which are known to be translations show cross-language similarities to the originals (in their LD's and other facets of structure) in varying degrees. It seems likely that quality of translation is a factor in this variation (when it is not simply due to sublanguage differences in the distance between E & F style). Presumably, translated texts have to meet some threshold criteria for naturalness. Spontaneously composed texts (without influence from texts of another language) may be significantly different. It will be quite interesting to track down these differences. That is, what aspects of text structure (or other sublanguage characteristics) are not critical to the perception of the text as being "natural"? Perhaps no generalizations can be made, but comparisons of different translations and natural texts are being considered. In such a case, generalizations may help to determine stylistic criteria for very high-quality translations.

During examination of pronominalization scope and examples of deletion it became obvious that the two LD's are related in a number of ways. Since they both can operate to reduce the form of noun phrases, they may "compete" in some environments. In this way, one would expect to see an inverse proportional relation in their frequencies. Texts which contain a lot of deletion may have less pronominalization and conversely. However, the relationship may be more subtle and have an effect on the overall text structure. For example, in both English and French recipes (but more strongly in French, 1t appears) there is a progression from full

NP objects for verbs to pronominalization of NP object to outright deletion. Wide variation in the rate of progression has been observed in different French recipes (originating from different French-speaking regions). Closer analysis of LD's and their interplay 1s expected to reveal and clarify more dependencies of this type.

An important area for further research concerns the relationship between LD's and intra-sentential links for a given sublanguage. Many of the linking devices studied (e.g., pronominalization, deletion, synonymy, etc.) can just as well occur between clauses of the same sentence or even between elements of the same clause. To some extent, scope measurements for these LD's can be expected to be extended inside the sentence. The sentence boundary may or may not be an important point in determining the mechanism and frequency of each type of linking. In the aviation hydraulics texts, deletion of object NP 1s quite frequent under conjunction intrasententially, but rare as a LD between sentences. Sublanguages will probably behave quite idiosyncratically as far as interaction between LD's and intrasentential linking and structural types (and word classes) is concerned.

Another interesting area with definite applications to automatic language processing concerns the density of textual links in typical texts. Although there may well be considerable variation here and a more uniform statement will certainly depend on being able to classify all kinds of semantic links (including chains of deduction linking two lexical items), it appears that texts within a given sublanguage show a fairly homogeneous density of links. It may therefore be possible to use information about such characteristic frequencies to detect automatically breaks in a text (as when a page is missing) or to use density of linking as a parameter of stylistic well-formedness when generating text from a semantic base. If such measurement of density can be made sufficiently precise, it might be possible to use density levels to help decide how to interpret lexical items or grammatical structure during parsing (particularly when full-semantic approaches to well-formedness turn out to be too complex for the type of text in question).

NOTES

1. A recent study representative of the informal approach is found in Crystal & Davy (1969).

2. I am indebted to Jim Munz for this term.

3. The study of translated texts as opposed to spontaneously generated texts in the sublanguaqe, should make it possible to eliminate this variable, and eventually, to provide a set of criteria for determining degrees of naturalness in translations.

4. The lengthy analysis of English and French texts and counting of LD's was carried out by Susanne Carroll, who also contributed a number of valuable insights into the qualitative analysis of sublanguages.

5. Included in the pronominalization counts are cases of cohesive co-reference between _you_ and _me_ in nearby sentences of dialogues which were embedded in children's stories. This accounts for most of the cases where scope exceeded one sentence boundary.