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

Automatic Translation and the Concept of Sublanguage

John Lehrberger

Contents

1. Introduction

It is common to speak of the language of biophysics, the language of pharmacology, etc. as though there were certain well defined languages used by specialists in various fields. But a glance at technical or scientific writing reveals that the language used is basically a language such as English or French. Even a layman can recognize the language although the presence of special terminology and mathematical formulas may prevent him from understanding the subject matter. If we can recognize that a text is "in English" and yet feel that it is distinct enough to be described as being "in the language of X" (physics, aeronautics, electronics, etc.) then we may be justified in saying that the language of X is a "sublanguage" of English. In fact, the term *sublanguage* is now used by many linguists investigating texts in specialized fields. And it is within the domain of sublanguages that automatic translation appears to be practicable. A system for translating weather reports from English to French is already in use in Canada (TAUM-METEO, [1]) and a system for translating aviation maintenance manuals is under development at the Université de Montréal [4]. This paper will examine the notion of sublanguage, its role within the "whole" language, and its importance in the development of automatic translation.

2. *Description of a Particular Sublanguage*

2.1. The Corpus

Researchers at TAUM (Traduction Automatique Université de Montreal) have made a detailed study of the properties of texts consisting of instructions for aircraft maintenance. The study was based on a corpus of 70,000 words of running text in English. There were 3548 different words in the analysis dictionary distributed among the various categories as follows:[1]

(1)

nouns	1714	prepositions	134
verbs	667	coordinate conjunctions	13
adjectives	664	subordinate conjunctions	29
adverbs	168	pronouns	35
quantifiers	46	articles	15
numerals	63		

Only base forms are listed in the dictionary (e.g., *adjust* is included, but not *adjusted* or *adjusting).* There are 571 idioms included in these figures, 443 of which are "technical" idioms specific to the subject matter. Examples of these idioms and a discussion of the criteria for listing an expression as an idiom will be given in section 2.5. Further study is expected to result in a reduction in the number of idioms at a later stage.

---

[1] These figures represent the stage of development at the end of 1978.

The categories in (1) are traditional; words (and idioms) are assigned to these categories on the basis of their use in the corpus. E.g., "cool-skin", "gear-driven", "loadcarrying", "following" are all listed as adjectives since they occur only as modifiers of nouns. Further subcategorization, essential for parsing, is obtained by associating syntactic and semantic features with the words in the analysis dictionary. Thus with each dictionary entry there is associated a category, features and complementation (a detailed description of the format of the analysis dictionary is given in [4]).

2.2. Restrictions

2.2.1. Lexical Restrictions

Although the corpus contains only 4876 different lexical items it is estimated that in the set of texts which the corpus represents there may be something like 40,000. Comparing this number with the number of entries in Webster's Third (about 450,000), it is obvious that the vocabulary of this sublanguage is highly restricted. One needs to describe the parts of the air-craft, the maintenance of hydraulic systems, electrical systems, turbines, etc., and the tools and test equipment required for such maintenance. Certain words are characteristic of this subject matter: aileron, motor, compressor, jack, filter, check, axial, quick-disconnect. Other words do not occur at all: parsley, meson, seduce, endocrine, hope, think, believe. None of the personal pronouns *I*, *me, we, us, be, she* are used here. The sets of words which characterize different sublanguages are not mutually exclusive however. "Filter", which occurs frequently in this corpus, is also typical of the language of pharmacology. It is not the vocabulary alone which determines a sublanguage, as we shall see in the following sections, although it is certainly an important factor.

Vocabulary restrictions do not apply to the same extent in all categories. The categories noun, verb, adjective and adverb are most limited while nearly all members of the remaining categories may be found in most sublanguages. E.g., all articles and coordinate conjunctions occur in this corpus. About 70% of one-word subordinate conjunctions occur (we do not expect to find "whilst" or "whereupon") and about 80 % of one-word prepositions (nor do we expect "apropos" or "notwithstanding"). This result conforms to the ubiquitousness of "grammatical" words and the fact that the main semantic burden is borne by nouns and verbs. On the other hand there are sublanguages which are characterized by the use of certain archaic or formal grammatical words ("whereupon the Lord commanded") as well as typical nouns and verbs.

2.2.2. Syntactic Restrictions

Since the sentences of this corpus are used either to describe the aircraft and related equipment or to give instructions for their maintenance, direct

questions do not occur at all (*Do you have your tool kit? *Is the motor turned off?). And tag questions indicate an attitude toward the user of the manual which is unacceptable (*Check the batteries, won't you? *The switch should not be on, should it?), hence they do not occur.[2]

There is no use of the simple past tense in the corpus (*The engine stopped. *High temperatures caused buckling.)

There are no exclamatory sentences (*How powerful the engine is! *What a complex hydraulic system this plane has!)

Other sentence types show the full range of syntactic structures in the corpus: passives, restrictive and non-restrictive relative clauses, extraposition, nominalizations of various types, etc. Long and complicated sentences are common in spite of the "telegraphic style" which characterizes most of the text and the internal structure of the noun phrase is often quite complex:

(2) "This unit contains the fuel metering section, shutoff valve, and a mechanical governor that functions as either an over speed governor for the high pressure rotor or provides manual control when the electronic computer section of the fuel control system is deactivated."

(3) ". . . a lightweight, two-spool geared transonic-stage, front-fan, jet-propulsion engine."

One of the most difficult problems for automatic parsing involves conjunction, with its associated reductions and ambiguities. E.g.,

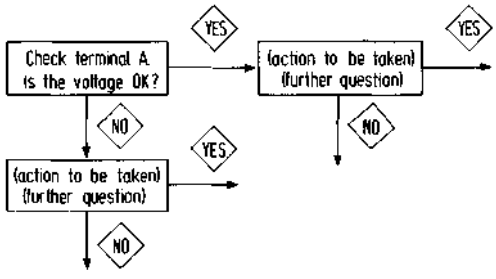
(4) "Disconnect pressure and return lines from pump." (ambiguous)

Another very difficult problem is the proper bracketing of long sequences of nouns:

(5) "The stability augmentor pitch axis actuator housing support" (see 2.4.3).

The corpus is generously endowed with such features so that parsing is by no means simple in spite of the restrictions mentioned above.

[2] In certain related texts direct questions occur, but in well marked environments such as flow charts describing troubleshooting procedures:



2.2.3. Semantic Restrictions

2.2.3.1. Categorization and Subcategorization

We have seen in 2.2.1 and 2.2.2 that the restricted subject matter and the attitudinal relation between text and reader limit the vocabulary and the inventory of syntactic structures in the sublanguage. But more important than the limitation in size of vocabulary is the reduction in polysemy. In some cases this results in a word occurring in only one category in the sublanguage whereas it may occur in several categories in the language as a whole. E.g., in this corpus the words in (6) occur only in the categories indicated in parentheses.

(6) case (N) * *Case* the joint.
lug (N) * They *lugged* the equipment from the plane.
cake (V) * The pilot likes banana *cake.*
jerky (ADJ) * Carry a pound of *jerky* on long flights.
just (ADV) * This is a *just* test procedure.
fine (ADJ) * *Fine* them for smoking. *There is a *fine* for smoking.
cable (N) * *Cable* the forward compartment.

In other cases the range of meanings of a word within a given category is restricted:

(7) eccentric (Adj) Cannot apply to animate objects (*an *eccentric* pilot)
ball (N) Can only be a spherical physical object (*the annual *ball)*
check (N) Abstract only (*Cash this *check.)*
bore (V) Cannot take human object (*Inaction may *bore* the crew.)
bore (N) Cylindrical hole or inside diameter of cylinder (*The pilot is a *bore)*

Since the parser explores the possibility of assigning a structure to a given string of words for each category in which the words occur, a reduction in the number of categories to which the individual words belong results in fewer combinations and less ambiguity. E.g.,

(8) Check pump case drain fitting.
N N N N N
V V V V V

In general English each word of (8) can occur in either category N or V, resulting in thirty-two paths to be explored. Of course all but one of these should be rejected by the parser (the combination in which "check" is a verb and the remaining words are all nouns). But since "case" is not used as a verb in the corpus it is listed in the analysis dictionary only as a noun. This alone reduces the total number of combinations to be tested in (8) from thirty-two to sixteen.

Consider the ambiguity in (9):

(9) Case ejection door locks immediately.

N	N	N	N	ADV
V			V	

In general, either "case" or "locks" may be taken as the verb. However, in this corpus "case" occurs only as a noun. "Locks" is therefore the only candidate for a verb and "case ejection door" is the subject noun phrase. The parser is then relieved of the responsibility for deciding that maintenance personnel are not instructed to case the ejection door locks. Restriction of the semantic range of a lexical item, even when it does not reduce the number of categories to which the item is assigned, is extremely useful in parsing. E.g., in (10) "cooling" may be taken either as a modifier of "purposes" or as the gerundive form of the verb "cool" whose object is "purposes".

(10) (A small heat exchanger) uses engine fuel for cooling purposes.

It will be obvious to the reader that "purposes" is not the object of "cooling", but how does the parsing machine know it? In these texts only concrete things are cooled (not tempers, etc.), hence we need only specify in the dictionary entry for the verb "cool" that its direct object must have the feature CONCRETE. If we were designing a parser for all English this would not suffice. The subcategorization required to establish all necessary cooccurrence restrictions for the whole language would be very fine indeed. Even in a sublanguage the elimination of ambiguities is a serious problem.

2.2.3.2. Specificity

These texts are characterized by the absence of generic reference of the form "the + N". In the language of biology we have *"The dolphin* is a mammal." In a history text we may find "The invention of *the wheel* was a crucial step." But in these aircraft maintenance manuals the sequence "the + N" is specific. E.g.,

(11) *The oil tank* is not a component of *the engine.*

(12) *The computer* provides increased fuel scheduling.

"The wing", "the radio", "the engine", "the wheel", etc. are all specific references. The manual differs from a textbook which may be concerned with theoretical concepts and general definitions. Whereas a text book on motors and generators may contain a statement like (13):

(13) The motor is a machine that converts electrical into mechanical energy.

an aircraft maintenance manual contains statements like (14):

(14) The motor is a constant-displacement piston type.

Thus there is no ambiguity in this corpus involving generic versus specific reference. A further consequence of this fact is illustrated by the sentence.

(15) Clean reservoir system.

We may assume deletion of the definite article has taken place if we wish to compare (15) with the corresponding sentence in "standard English":

(16) Clean the reservoir system.

Instructions for maintenance and repair must be specific; one does not expect to find "Clean *a* reservoir system". Of course, we do not really have to recover deleted articles to *understand* sentences such as (15). We merely need to recognize the general principle concerning specific reference and then accept the fact that (15) is a normal acceptable sentence in this sub-language (see section 2.3.1). (However, the French translation does require a definite article, hence we must recover it for the purpose of automatic translation.)

2.2.3.3. Semantic Features

The semantic restrictions imposed by the subject matter are reflected in both the number and kinds of semantic features needed for parsing. Many nouns which designate either concrete or abstract objects in the language as a whole are used only concretely in this sublanguage; e.g.,

(17) air, battery, dirt, machine, flap, flash, post, rod, solution, speed, spring, tool, net, web, race.

The same is true of words that may be used for either human or non-human objects. None of the following words which appear in the corpus designate human beings or parts thereof:

(18) agent, body, boss, buffer, crank, elbow, governor, joint, nut, page, selector, starter

Verbs are likewise restricted in the kinds of subjects and objects they can take:

(19)

charge	object [+ CONCRETE]
circulate	subject [+ FLUID] (intransitive)
divert	object [+ FLUID]
function	subject [+ PART] (i.e., part of the aircraft or related equipment)
top	object [+ CONCRETE]
die	subject [- ANIMATE]

The features MALE, FEMALE are not relevant in the corpus.

The feature HUMAN has been used on only a few nouns in the parser although many verbs are marked as taking HUMAN subjects since this is implied by the use of imperatives throughout the text:

(20) Check fan blade clearance.
Adjust pump pressure control valve.
Remove and discard gasket.

Thus the feature HUMAN is used mainly in signalling implied subjects rather than in testing nouns in the text as possible subjects of nearby verbs.

The degree of semantic restriction in the sublanguage has a bearing on the manner of representing semantic features. In fact, two types of representation have been considered for the parser, which we may call unary (*F) and binary (+F or −F). The criterion for admitting a noun having the set of unary features {*F1, *F2, . . . , *F_n} as the k^{th} argument of a verb whose k^{th} argument position is assigned the set of unary features {*G_1, *G_2, . . . , *G_n} is that the two sets have a non null intersection. This means that if the n^{th} argument of a verb can be either *CONCRETE or *ABSTRACT then both of these features must be listed in the n^{th} argument position of the verb in the dictionary. And if a noun may be either *CONCRETE or *ABSTRACT then both features must be entered with that noun in the dictionary. This would seem to result in a great deal of redundancy since there are so many nouns which may designate either concrete or abstract objects and so many verbs whose arguments may be either *CONCRETE or *ABSTRACT. The same is true for many other features as well.

The alternative is to use binary features along with the following conventions:

(21)(i) A noun is marked +F if it *always* has the feature F and −F if it *never* has that feature; otherwise it is not marked for F in the dictionary.
(ii) If the n^{th} argument position of a verb is marked αF it can only take arguments marked αF, where α is either + or -; the n^{th} argument position of the verb is not marked for F if it can take either +F or −F arguments.
(iii) A noun is admitted as n^{th} argument of a verb provided there is no feature F such that the n^{th} argument position of the verb is marked αF and the noun is marked −αF.

At first sight it appears that such use of binary features would result in overall economy. However, the semantic restrictions in this sublanguage result in many nouns being marked only +F (or −F) and many verb argument positions being marked only +F (or −F), as was illustrated in (17)—(19). Consequently the advantage of not having to mark a large number of nouns and verb argument positions for certain features is lost. At

the same time, any feature which is rarely used will have to be entered on all those nouns and verbs where it is not relevant. E.g., the small set of nouns including *air, oil, water,* etc. would be marked +FLUID and all other nouns in the dictionary would have to be marked −Fluid as well as all verb argument positions which do not accept +FLUID arguments. However, if unary features were used then *air, oil, water,* etc. would be marked *FLUID and the majority of nouns would not be marked at all for this feature. Likewise for verb argument positions.

Presently the unary method of representing features is used in the parser. At any rate, it is interesting to note the effect of semantic restrictions within a sublanguage on the kind of semantic representation which is used.

2.3. Reductions

2.3.1. Omission of Definite Article

One of the most frequent reductions found in this corpus is the omission of the definite article:

(22) Check indicator rod extension.
One system provides air for bearing compartment sealing.

But such reduction does not always take place as can be seen in the following sentences which are also found in the corpus.

(22') Check *the* ground test system.
Check *the* control stick breakout.
All controls for *the* air conditioning system are located in *the* front cockpit.
Separate outlets are provided for *the* engine and handpump.

It does not seem to be the case that in some contexts the definite article is always omitted while in others it is not. We can only say that it may be omitted and very often is. Yet, in spite of this, *the* is the most frequently occurring word in the corpus (2,925 occurrences). No definitive study has been made of the environments where its omission is most likely to take place. From the point of view of the parser, allowance is made for the fact that it may not be present where it is expected in standard English, but no attempt is made to predict its presence or absence.

2.3.2. Omission of Copula

In standard English the copula BE may or may not be used in certain contexts:

(23) (i) The book (which is) on the desk.
(ii) We considered it (to *be)* unreliable.

We may question whether the shorter forms are "reductions" or simply paraphrastic alternatives, but in the corpus there is another type of construction which seems clearly to be a reduction involving omission of BE:

(24) (i) Check reservoir full. (Check that the reservoir is full.)
(ii) Check fluid level above REFILL mark. (Check that the fluid level is above REFILL mark.)

There is a class of verbs (believe, consider, find, etc.) which can take a noun phrase + *to be . . .*as complement (as in 23ii). When *to be* is not present the complement may consist of a noun phrase followed by an adjective phrase. But *check* does not belong to this class of verbs, hence the construction in (24i) is peculiar to these texts. Both (24i) and (24ii) occur frequently in the corpus. However, there are similar sentences which do contain the copula:

(25) (i) Check that fuel systems are full.
(ii) Check fluid level indicator is registering correctly,
(iii) Check that fuel pressure is between 45 and 55 PSI.

As with the definite article we see that the omission of BE is not obligatory in sentences like (24). It does happen often enough to be considered characteristic of these texts, but it is optional in those contexts where it occurs. The copula is also omitted from progressive forms, as in (26):

(26) Pump not delivering fluid.

2.3.3. Omission of That-Complementizer

A comparison of (25ii) with (25i) and (25iii) shows that the omission of *that* as a sentence nominalizer is optional. This is common in standard English with verbs like *know, suppose, hope* (I $\left\{\begin{matrix}\text{suppose}\\ \text{know}\\ \text{hope}\end{matrix}\right\}$ the fluid level indicator is registering correctly), but not with the verb *check* (*we are checking the indicator is working).

2.4. Frequently Occurring Forms

2.4.1. Imperative

Imperative sentences abound in the corpus. This is to be expected since a maintenance manual, like a cook book, is primarily concerned with instructing the user in the performance of certain actions (Check . . . , Adjust . . .,Turn . . ., Remove . . ., Insert . . ., etc.). Were it not for the fact that these manuals also describe parts of the plane and how these parts function, nearly every sentence would be in the imperative. The significance of the imperative in characterizing this sublanguage is not simply that it occurs, but that it occurs so often.

2.4.2. Non Predicative Adjectives

There are many adjectives in the corpus which never occur in predicate position. They are marked with a feature ATRIB in the parsing dictionary and constitute 25 % of all adjective entries. Some examples are given in (27).

(27)

A.	B.
actual	nickel-cadmium
chief	piston-type
consequent	pressure-regulating
entire	anti-stall
respective	single-point
	non-priority

Those listed in column (B) are particularly important in characterizing this sublanguage. They deal specifically with the subject matter of aircraft maintenance whereas those in column (A) are of a more general nature. There are a number of productive types involved in (B):

(28) X-type
X-Ving
anti-X
X_{num}-Y
non-X

Presently all the adjectives in (27) are listed in the parsing dictionary. However, since those in (B) are productive types it might be preferable to separate the components at pre-edition (see [4]) and analyze the resulting string in the parser. This matter is under study.

In addition to being non-predicative these adjectives are not inflected for comparative or superlative (*chiefer, *pressure-regulatingest). We might question whether they should be considered as forming a subcategory of the adjective class or as simply a separate class of prenominal modifiers in this sublanguage (see 4.2).

The corpus contains many compounds consisting of a numerical expression followed by either a measure unit (29 A.) or a certain kind of noun which may be considered a measure unit in the proper context (29B.)

(29)

A.	B.
115/200-volt	3-phase
0.0045-inch	19-cell
10-micron	2-stage
1000-hour	eighth-stage
15-ounce	two-lobe
11 -ampere-hour	three-way
110-to-infinite HERTZ	two-spool
	three-axis

The nouns following the dashes in (B) are not, strictly speaking, measure units; but in this sublanguage they are used as such: *phase* is a measure unit

with respect to *generator, cell* with respect to *battery, lobe* with respect to *cam*, etc. All compounds of this type should be separated into their components at pre-edition, analyzed by the parser, and assigned a feature MP (measure phrase). They should not be entered as individual lexical items in the dictionary since the numerical portion is of arbitrary size. Just as numerical expressions in general must be parsed, so must these measure compounds.

One of the conventions followed in these texts is to place a hyphen between a numerical expression and following measure unit when the compound is used as a prenominal modifier, and to write the measure unit in the singular; otherwise there is no hyphen and the measure unit may be pluralized (a three-stage turbine, the turbine has three stages). The hyphenated measure compounds behave like the non-predicative adjectives described above, hence they should be treated as adjective phrases by the parser and so labelled. The numerical component might suggest treating them as quantifiers but unlike most quantifiers they occur in the characteristic adjective position between article and noun and do not require pluralization of a following count noun in this context.

2.4.3. Noun Sequences

A major feature of the corpus is the presence of many long strings of nouns, or nouns and adjectives, within nominal groups:

(30) (i) external hydraulic power ground test quick-disconnect fittings
(ii) fuselage aft section flight control and utility hydraulic system filter elements
(iii) fan nozzle discharge static pressure water manometer
(iv) landing gear, flight controls, speed brakes, engine air by-pass flaps, and nose steering systems
(v) stabilizer power control No. 2 system return line check valve failure

This phenomenon is a result of the need to give highly descriptive names to parts of the aircraft in terms of their function in the aircraft and their relation to other parts. It is likely to occur in any texts describing very complex machinery containing a large number of specialized parts.

The segment of such a noun phrase from the first adjective or noun to the last noun is referred to here as an *empilage.* It does not include initial determiners or quantifiers. In the corpus there are about 4400 different empilages, many of them occurring numerous times. They present a major problem in parsing the nominal group.

The proper bracketing of an empilage requires an understanding of the semantic/syntactic relations between the components. Thus in (31)

(31) main fuel system drain valve

we must know that *main* applies to *fuel system,* not to *fuel* or *drain valve,* and that (31) refers to a valve whose function is to drain the main fuel system, whereas in (32)

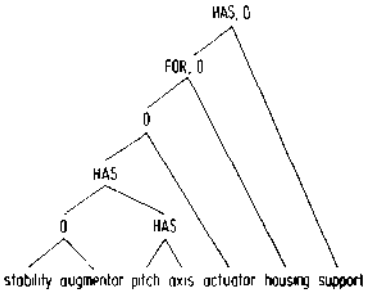
(32) system main drain valve

main applies to *drain valve,* not to *drain,* and (32) refers to the main valve whose function is to drain the system.

The problems involved in parsing empilages are similar to those encountered by linguists attempting to explicate the formation of NOUN + NOUN compounds. Recent investigators have proposed various semantic and syntactic relations between the components as well as "underlying structures" for analyzing these compounds. E.g., Levi [6] proposes a small number of relations (deletable predicates) such as CAUSE (nicotine fit), HAVE (fruit tree), USE (steam engine), FOR (communications system), SUBJECTIVE NOMINALIZATION (cell decomposition), OBJECTIVE NOMINALIZATION (traffic control), etc. in terms of which most NOUN + NOUN compounds may be derived. Downing [2] suggests twelve relations that should be included in any such inventory [WHOLE-PART (duck-foot), COMPARISON (pumpkin bus), TIME (summer dust), PLACE (Eastern Oregon meal), etc.], but she claims that "the semantic relationships that hold between the members of these compounds cannot be characterized in terms of a finite list of 'appropriate compounding relationships'". The study of empilages at TAUM resulted in the definition of about 50 such relations, including some of those mentioned above (HAVE, WHOLE-PART, PLACE, SUBJECT, OBJECT, etc.). This list of semantic/syntactic relations may turn out to be sufficient for analysis of empilages in the sublanguage under investigation, but no claim is made for the whole language. If Downing is correct, success in finding a finite set (at least small enough to be useful in automatic parsing) may depend on just such limitations as are encountered in a sublanguage.

An example of the constituent structure of an empilage showing the relations between constituents is given in (33).

(33)



stability is grammatical OBJECT of *augment; pitch* HAS an *axis; stability augmentor* HAS a *pitch axis; stability augmentor pitch axis* is OBJECT of *actuate;* the *housing* is FOR the *stability augmentor pitch axis actuator* (which is also OBJECT of *housing);* the *stability augmentor pitch axis actuator housing* HAS a *support* (and is also OBJECT of *support).*

We have not been able to define a set of relations which are mutually exclusive, as is evidenced in (33) by FOR, O and HAS, O.

2.5. Idioms

For the purpose of this discussion we will use *idiom* as a technical term for any multi-word expression which is entered in the dictionary (except for left members of normalization rules such as "for comparison of = = to compare"). Several criteria may be used to determine whether a given expression in the sublanguage is to be entered in the parsing dictionary as an idiom.

(I) The meaning of the expression is not predictable from the meanings of its components: with respect to, nose gear, finger tight.

(II) Translation idioms. The corresponding expression in the target language is not predictable by the usual rules for associating structures in the target language with those in the source language. Aspect ratio (Fr: allongement), DC power (Fr: courant continu), buttock line (Fr: section longitudinale).

(III) The expression occurs so often in the sublanguage that it "feels like" a compound word: landing gear, filter element, relief valve. From the point of view of automatic translation it may be more economical to list these in the dictionary than to parse them. That depends on parsing strategies and the desirability of limiting the size of the dictionary.

(IV) The expression occurs very rarely and its parsing would require undesirable changes in strategies developed to handle the majority of cases in the sublanguage. E.g., *right and left of center* occurs only once in the corpus ("right and left of center positions"), while *right* and *left* occur quite often. As things now stand, if the expression is not listed as an idiom the parser will spend a lot of time looking for of-compliments of *right* and *left* when these words occur elsewhere.

(I) is the usual criterion for idioms, (II) is relative to translation into another language, (III) and (IV) are relative to strategies for automatic parsing. The creation of idioms for the purpose of automatic parsing will be discussed further in section 3.4.

2.6. Text Structure

2.6.1. Gross Structure

Texts in this sublanguage are divided into numbered sections each of which deals with a specific part of the aircraft. The occurrence of a polysemous word in a particular section may signal a specific meaning for that word. E.g., *capacity* refers to volume in the hydraulic system and to farads in the electrical system or electronic equipment. In some cases this affects the translation of the word. Thus *valve* is translated into French as *clapet* when it occurs in the section on hydraulics and as *soupape* or *valve* in the section dealing with motors. Since *valve* is one of the most frequently used nouns in the corpus (716 occurrences, including the plural) it would be quite useful at transfer to have available not only the sentence in which the word occurs, but an indication of the section in which the sentence occurs as well.

2.6.2. Linking Devices

Neighboring sentences in the corpus are "linked" in various ways to provide textual cohesion. Repetition of a word or phrase is common:

(34) Install rotor on *shaft*(1), then align index marks on *inner races*(2) of *bearing*(3). Position *bearing*(3) on *shaft*(1) with vendor identification marking on outer race on same side as puller groove on *inner races*(2).

When a noun phrase occurs in one sentence it is often shortened in succeeding sentences rather than repeating it in full:

(35) Remove *jumper hose*(1) from *pressure in line*(2). Cap *line*(2) and *hose*(1). (Note: *in line* is not a prepositional phrase; an in line is a line bringing something in.)

Such reductions complicate the identification of coreferential noun phrases in different sentences even though the head of the NP is retained. And when more than one noun phrase is repeated, as in (34) and (35), the order of occurrence may be different.

Pronouns may also operate across sentence boundaries:

(36) The main system relief valve is located on the left side of the engine compartment, just forward of the hydraulic reservoir. It is adjusted to provide automatic relief when hydraulic pressure in the system exceeds 1225 (+ −25 PSI).

It refers to *the main system relief valve,* not to *the left side of the engine compartment* or *the hydraulic reservoir,* although the latter two are closer to *it.* In this case we know that valves, not reservoirs or compartments are adjusted, and since *it* is the object of *adjust* in the second sentence, the

referend of *it* in the first sentence must be *valve.* In order to make use of this information in a parser nouns could be subcategorized on the basis of their naming things which can be adjusted (perhaps by assigning a feature ADJUSTABLE). This alone would not be sufficient since more than one adjustable item may be mentioned in the first sentence. Semantic analysis of a text, even when restricted to a sublanguage, calls for considerable sub-categorization.

Repetition with "adjustment" of grammatical category is also used as a linking device. Nominalization is one of the most common adjustments:

(37) Vent manifold may be leaking. This leakage will allow . . .

Sometimes there is implicit reference to elements in the preceding sentence without the use of pro-forms or repetition (unless one's theoretical framework makes it more convenient to consider this a case of repetition with reduction to zero):

(38) Remove and inspect the fuselage aft section flight control and utility hydraulic system filter elements. If found to be highly contaminated, clean and reinstall, then remove and inspect all flight control actuator filter elements. If found to be highly contaminated, clean and reinstall, then remove and inspect all hydraulic system restrictors. If restrictors are found to be highly contaminated, clean and reinstall.

The object of find, clean, reinstall in the second sentence is in the first sentence and the object of *find, clean, reinstall* in the third sentence is in the second, but the object of these verbs in the fourth sentence is present in the fourth sentence owing to the repetition of *hydraulic system* restrictors (with reduction), retaining only the head, *restrictors.*

There are many lists and tables in the corpus. References to them (or to particular sentences in them) in other parts of the text often results in direct linking between non-contiguous sentences. The internal structure of a list may disambiguate an expression contained in it. E.g., consider (39) and (40):

(39) Correct wiring.
(40) Bleed fittings on brake assembly.

Since (40) begins with a capital letter and ends with a period we might assume it is a sentence instructing the technician to bleed the fittings. However, it occurs in the second column of a list of components and their bleed points which is headed COMPONENT BLEED POINT and each expression on the right is the name of a location, not an imperative sentence. Now we have to be suspicious of (39) which may be simply an adjective + noun combination in spite of the initial capital and final period. On examining the structure of the list in which (40) occurs we find that it contains three columns:

PROBABLE CAUSE	ISOLATION PROCEDURE	REMEDY

The third column, consisting of imperative sentences (Clean . . ., Install . . ., Remove . . .), includes (39) which is therefore an instruction to correct the wiring (V + N).

These examples show that semantic and grammatical analysis of a text (or even a sentence) requires looking beyond the boundary of the individual sentence. A unit of text larger than the sentence seems to be needed. The use of such a unit was considered in the development of the present system at TAUM but was rejected for reasons of economy. This does not preclude use in future development as it is both desirable and possible on theoretical grounds.

2.7. Odds and Ends

2.7.1. Numerical Expressions and Reference

The corpus makes much use of numerical expressions, either spelled out (secure with *two* attaching bolts) or written with Arabic numerals (gauge should read 1000 PSI). There are certain rules governing the representation of numbers in these texts: spell those from zero to nine except for percentages (5%), numbers in compound adjectives (two 3-phase generators), all numbers in a sequence if one of them exceeds 9 (position clamps 8, 11, 21, 24, 30 on harness), etc. However, all numerical expressions are represented by Arabic numerals after parsing in the present translation system used at TAUM since this is more convenient at the transfer stage.

There are many expressions consisting of a mixture of numerals, letters hyphens and slashes, which are called references ("refer to EO 15-70-5A/2"). These have an internal structure which is semantically significant, but for the purpose of translation they keep the same form.

2.7.2. Labels

Frequently a word in the corpus refers to a label on a part of the aircraft or related test equipment. These words, indicated by spelling with all capitals, are not to be translated.

(41) Set switch to ON.
(42) Ensure that the PITCH CONT switch is ON.

In (41) ON is simply a label, but in (42) it also serves one of its normal grammatical functions as an intransitive preposition (or prepositional adverb). The use of this kind of ambiguity in these texts reflects the general tendency to be as concise as possible. Of course, since the labels are not to be translated this can be troublesome: the switch is ON = 1'interrupteur est sur "ON". The systematic ambiguity does not hold in translation, hence

the restructuring in French with the addition of *sur.* It is as though the English sentence had been "the switch is on ON".

2.7.3. N-Ving, N-Ved

There are many compound words in the corpus consisting of a noun followed by the present or past participle of a verb (gear-driven, air-separating, cockpit-mounted, motor-operated, seat-adjusting, spring-adjusting, spring-loaded, etc.). The noun usually names a part of the aircraft and the verb describes an operation on or by that part. These compounds are entered in the dictionary as adjectives when there is no corresponding verb. Consider the following example involving *gear-driven:*

(43) A spinner hub and an axial flow fan are gear-driven by the low pressure spool.

Since *by the low pressure spool* is agentive, not locative, (43) appears to be a passive and *gear-driven* the past participle of a verb *gear-drive.* But *gear-drive* does not appear as a verb in these texts (*X gear-drives Y). Hence we accept the structure *N be A by N* where *by N* is agentive.

3. Practicability of Automatic Translation

3.1. Formal Grammars for Natural Languages

Perhaps the problem of designing an automatic translation system for a natural language may be viewed more clearly from the perspective of attempts to write formal grammars for natural languages. It is precisely when we try to formalize our knowledge of a language that the difficulties begin. Generative grammarians in particular have put an enormous amount of effort into the formalization of rules of grammar. Their lack of success so far in producing a set of rules that will generate all and only the sentences of a natural language in its entirety hardly seems encouraging to researchers in automatic translation trying to devise a set of rules that will analyze any sentence in one language and generate the corresponding sentences in another. In fact, the prospect may seem even dimmer when we consider that generative grammarians usually aim only for a description of the "standard language" or the language of an "ideal speaker in an ideal community"; presumably a natural language in its entirety includes arbitrary discourse, much of which lies outside these domains.

Is it then realistic to expect success in automatic translation given the difficulty of writing a formal grammar for even one language? One may reply that automatic translation from L_1 to L_2 does not require complete grammars of L_1 and L_2, only context sensitive transfer rules to obtain the proper lexical items in L_2 and some rules for restructuring the resulting

string of lexical items in L_2. Of course, the terms *context sensitive* and *restructuring* in themselves indicate the need to recognize the possible structures in which the lexical items of L_1 and L_2 occur. Experience at TAUM, even with a very limited corpus, has demonstrated that an extremely fine grammatical analysis of both languages is required (especially the source language) in order to translate say 80% of the number of sentences in a text. The system currently in use at TAUM parses the sentences of the source language and puts them in a normalized form indicating their grammatical structure. The "normalized structure" is a tree with labelled nodes and includes semantic as well as syntactic information. Transfer rules map these trees onto other trees containing the proper lexical items of the target language. Rules are then applied which map the trees onto sentences in the target language. Parsing, transfer and generation all require detailed analysis of grammatical structure. The problem of writing rules for a system of automatic translation cannot be separated from the general problem of writing formal grammars for particular languages. The solution in the case of automatic translation seems to lie in restricting one's attention to sublanguages.

3.2. Text Norms

The authors of maintenance manuals, cook books, articles in scientific journals, etc. are generally guided by norms for writing in their particular fields. In some cases guidelines are made explicit. Thus criteria for the texts described in section 2 are given in a booklet titled "Format and Style Guide". These norms do not themselves constitute a grammar — that can only be determined by examining the texts. But they do indicate certain regularities not present throughout the whole language, thus simplifying the task of writing formal grammars for texts in specialized fields.

The existence of norms for texts in certain fields, the reduction in polysemy resulting from semantic restrictions, the limited vocabulary, and the syntactic restrictions generally encountered all combine to make automatic translation practicable for sublanguages. An example of a working system is given in the next section.

3.3. TAUM-METEO

A system for automatic translation of weather reports from English to French is now in use in Canada ([1]). The sublanguage in this case has a very small vocabulary and is characterized by telegraphic style. Because of the telegraphic style verbs appear only in the present participle or past tense forms. These factors make it more economical to include morphological variants in the dictionary instead of listing only the base forms and performing a morphological analysis.

The syntax is highly restricted: no relative clauses or passives, omission of copula, no use of articles, etc. Consequently syntactic analysis depends very much on semantic subcategorization as can be seen by the five sentence types recognized in this system.

(44) (i) place names preceding the forecast
RED RIVER
INTERLAKE

(ii) meteorological conditions for the day
MAINLY SUNNY TODAY
WINDS 25 KM PER HOUR

(iii) statement of maxima and minima
HIGHS TODAY 15 TO 18
LOWS TONIGHT NEAR 3

(iv) outlook for next day
OUTLOOK FOR THURSDAY . . . CONTINUING MAINLY SUNNY

(v) heading of bulletin indicating origin
FORECAST FOR MANITOBA ISSUED BY ENVIRONMENT CANADA AT 6 AM CST APRIL 8TH 1976 FOR TODAY AND FRIDAY

(This is a fixed form; only place names, dates and times change.)

METEO is by no means as complex as the system required for the texts described in section 2, but it does demonstrate the feasibility of automatic translation. A complete description of TAUM-METEO is given in [1].

3.4. Idioms

In 2.5 we examined four criteria for entering strings of words in the dictionary instead of submitting them to analysis by the parser. It might be thought that parsing could be greatly simplified by entering many strings in the dictionary even though they do not meet those criteria, especially, noun sequences. The dictionary would then be rather large, but by removing much of the burden from the parser where theoretical problems in linguistics are still a major stumbling block the translation of arbitrary texts in a language would seem to be a more reasonable goal. However, it is difficult to imagine just how large a dictionary would be required to eliminate major parsing problems. There seems to be hardly any limit on the number and size of noun sequences possible in a language, judging from the corpus described earlier (2.4.3). Furthermore, a string of words which forms a noun phrase in one context may occur in other contexts where the words have different meanings or belong to different categories and are not even within the same constituent. E.g.,

(45) Locate all check points.
(46) Check points for pitting.

Listing *check points* in the dictionary as an idiom of category N would simplify the parsing of (45) but prevent correct analysis of (46) where *check* is a verb and *points* its object. Furthermore, *points* has a different meaning (as well as a different French translation) in (45) and (46).

Few word sequences are idioms in *all* contexts in a sublanguage, and even fewer in all contexts in the language as a whole. When a suspected idiom is encountered it is necessary to check that it really is an idiom in that context (see ([4]) for discussion of treatment of "potential idioms" in TAUM system). An expression which forms an idiom in all contexts in a sublanguage when its components are contiguous may also occur with the same meaning but with its components separated under certain conditions. This is especially true when conjunction reduction is involved, and it poses a problem in parsing. Suppose, e.g., *in spite of* is entered in the dictionary as an idiom and the parser encounters.

(47) He acted without malice in spite and because of her threat.

It is desired to recognize *in spite of* as a unit, but *of* is separated from the rest of the expression. One strategy for putting the pieces back together might be to mark *spite* so that its occurrence immediately after *in* triggers a search for *of* in case a conjunction follows *spite.* More generally, a strategy is needed for reconstituting split idioms.

Problems like these multiply rapidly when a broader range of texts is taken into account and, correspondingly, parsing strategies required to deal with them increase in number and complexity. But when confined to a sublanguage such problems appear to be manageable. E.g., although the idiom *in spite of* could occur in a maintenance manual (the motor may continue operating in spite of fuel leakage), the word *spite* will not occur in the sense of malicious intent and *in spite* will not occur as a prepositional phrase. Thus, having encountered *in spite, of* is sure to follow, immediately or otherwise. There can be no ambiguity involving *in spite* and the parser may proceed with confidence to rejoin the components if they are separated.

Experience at TAUM indicates that even in a very restricted sublanguage potential and split idioms constitute a hazard in parsing. Clearly, the creation of idioms is no cure-all in automatic translation.

3.5. Recognition and Generation

A grammar for a natural language may be constructed for the purpose of generating all and only the acceptable sentences of the language or for the purpose of "recognizing" a given string as a sentence and assigning a structure to it. Grammars of the latter type, which we may call recognition grammars, are used for parsing. Normally the input to the parser in a

system for automatic translation consists not of arbitrary strings whose sentencehood must be determined, but acceptable sentences whose structures are to be determined. It would be nice to have a machine which could decide for an arbitrary string of words whether or not it is a sentence and assign it a structure, but this is not necessary. In order to parse sentences already assumed to be grammatical one needs strategies for locating verbs and their complements, assigning words to various categories depending on context, assigning constituent structure, etc. This goal seems to be within reach in the domain of sublanguages.

Just as parsing begins with sentences of the source language, so generation of sentences in the target language begins with fully analyzed sentences, i.e., the output of the parser. Words have been assigned to categories, constituents determined, semantic features inserted, etc. Lexical items of the source language must now be replaced with those of the target language and many structural changes effected in the process of generating sentences in the target language. But, difficult as this may be, it is by no means as difficult as starting from the semantic representations, deep structures, or other abstract objects currently employed in many generative grammars and generating all and only the sentences of a language. If the source sentences can be parsed, it's a fair bet that the corresponding target sentences can be generated.

4. The Concept of Sublanguage

4.1. Characteristics

It should be clear from the preceding discussion that a sublanguage is not simply an arbitrary subset of the set of sentences of a language. Factors which help to characterize a sublanguage include (i) limited subject matter, (ii) lexical, syntactic and semantic restrictions, (iii) "deviant" rules of grammar, (iv) high frequency of certain constructions, (v) text structure, (vi) use of special symbols.

(iii) refers to rules describing sentences which, though quite normal in a given sublanguage, are considered ungrammatical in the standard language. Such sentences must be considered grammatical in the sublanguage, (iii) also refers to rules describing cooccurrence restrictions within a sublanguage that do not exist in the standard language. E.g., in the sublanguage described in section 2 there is a subclass of adjectives that do not occur with animate nouns (e.g. *eccentric pilot). The rule which in the sublanguage states that "eccentric pilot" is not permitted does not exist in the standard language. It follows that a sublanguage grammar is not a subgrammar of the standard language. Z. Harris states the matter somewhat differently in [3] p. 154: ". . . sublanguages can exist whose grammar contains additional rules not satisfied by the language as a whole". (My reason

for using "standard language" rather than "language as a whole" appears in 4.3). Harris claims that sublanguages are closed under transformations (p. 152): "Certain proper subsets of the sentences of a sublanguage may be closed under some or all of the operations defined in the language, and thus constitute a sublanguage of it." This notion of sublanguage is like that of *subsystem* in mathematics. For example, given an algebra $<A,f_1, \ldots,f_n>$ where A is a set closed under the operations $f_1, \ldots, f_n$, then a subset of A closed under the same operations forms a subalgebra of $<A, f_1, \ldots, f_n>$.

4.2. Cooccurrence and Subcategorization

If a sublanguage has a grammar of its own which is not just a subset of the rules of grammar of the standard language, it follows that the categories and subcategories of the standard language may not suffice for a grammar of the sublanguage. This is particularly true of the subcategories needed to state cooccurrence restrictions.

In the work on noun sequences at TAUM relations between nouns were defined on the basis of their behavior in the sublanguage concerned (2.4.3). (Actually adjectives are included, but the discussion here will be limited to nouns.) Each such relation R defines two subsets of nouns, namely the domain and range of R. E.g., the relation F (xFy iff x is the function of y) has in its domain *access, balance, check, filter, installation, pickup, reduction, safety,* etc., and in its range *aircraft, bar, compound, fixture, installation, lug, pipe, runs,* etc. The two sets need not be mutually exclusive, as *installation* shows *(installation kit, control installation).*

From another point of view, each noun has a left hand relation-set (the set of all relations having the noun in their range) and a right hand relation-set (the set of all relations having the noun in their domain). Thus *kit* has F in its left hand relation-set, *control* has F in its right hand relation-set, and *installation* has F in both relation-sets.

In order to obtain the correct bracketing of a sequence of nouns it is essential to know the relations that each noun in the sequence can bear to other nouns in the texts under consideration. Now suppose a given noun *n* can bear a certain relation R to an immediately following noun. This does not mean that *n* bears that relation to *any* noun that happens to occur immediately following it. For example, although *installation* indicates function in *installation kit* and *installation procedure,* it does not in *installation difficulty* (installation is not the function of difficulty). Thus the subclass of nouns to which *installation* can bear the relation F (in the sublanguage) must be specified, and this is also true for other words in the domain of F *(access, balance, check,* etc.). One way to make such information available to the parser is to indicate in the dictionary entry of a noun all the relations of this type in which the noun participates in the sublanguage as well as the appropriate subclasses in each case. This may not be an unreasonable task if

the number of relations required for the sublanguage is not too great. Of course, noun entries in the dictionary do become fairly complicated and nouns then have a "complementation" similar to that of verbs. The entry for *installation* would specify that the noun can be either abstract or concrete, that when it is abstract it can bear the relation F to any member of a certain subclass of nouns occurring on its right (as in *installation kit, installation procedure,* etc.), that it can bear the grammatical relation OBJECT to any noun of a certain subclass occurring on its left (as in *pump installation, filter installation,* etc.), and that it has certain additional properties when it is concrete rather than abstract.

The whole question of assigning such noun complementation in the dictionary to indicate possible semantic/syntactic relations between nouns (and also noun-like adjectives) is now under study. Clearly, the implementation of such a system depends on a fine subcategorization of the class of nouns, and this subcategorization must be based on a careful study of cooccurrences within noun sequences in the sublanguage concerned. Although the relations in terms of which these subclasses are defined are of a general nature (FUNCTION, PART-OF, SUBJECT, OBJECT, etc.), the subclasses themselves are specific to the sublanguage.

4.3. Sublanguages and the Language as a Whole

It is not known how many sublanguages exist in a given language. They are not determined a priori but emerge gradually through the use of a language in various fields by specialists in those fields. They come to our attention when people begin to refer to "the language of sports-casting", "the language of biophysics", etc. As we have seen, a grammatical sentence in a sublanguage of English may not be grammatical in standard English even though the text in which the sentence occurs is still said to be "in English". When we speak of "the language as a whole" we include all such texts, thus it seems that a grammar of the language as a whole must describe all the sublanguages in it — certainly no mean task.

Many of the sentences of a sublanguage of L are considered "standard L"; the percentage varies within each sublanguage. And those sentences that are not so considered can be paraphrased in standard L (Check reservoir full ↔ Check to ensure that the reservoir is full). This suggests that the standard language may be useful in describing the way a sublanguage fits into the language as a whole. Furthermore, sublanguages overlap and their interrelations form a part of the description of the language as a whole. A language is not simply a union of sublanguages, but a composite including many sublanguages related to varying extents lexically, syntactically and semantically. These relations are implied by statements like the following:

(i) In aeronautics the noun *dope* refers to a chemical compound used to coat fabrics employed in the construction of aircraft, whereas in pharmacology it may refer to narcotics.
(ii) The words *hammer, anvil* and *stirrup* as used in the study of the ear are related metaphorically to these words as used in a smithy.
(iii) Instruction manuals in many fields employ a telegraphic style, often omitting the definite article in contexts where it is required in standard English.
(iv) Expression of emotion may be appropriate in a religious publication, but not in a journal of physics or math.
(v) The philosophy student's thesis was criticized for containing flip comments more appropriate to a term paper in freshman English.
(vi) English texts in various fields share the alphabet a, b, c, . . ., x, y, z but ∃ occurs in those dealing with mathematical logic, ə in phonology texts, etc.

Formalization of such relations may shed light on the role of sublanguages in the language as a whole.

Individual sublanguage grammars are of independent interest for the purpose of information retrieval and automatic translation. A question which stands in need of more investigation is the extent to which corresponding sublanguages in different languages have similar characteristics. E.g., Kittredge claims ([5]) that variation in textual linking devices may be greater between two dissimilar sublanguages in the same language than between two corresponding sublanguages in different languages. If true, this is further evidence of the practicability of automatic translation between corresponding sublanguages in different languages. Of course one may point out individual characteristics of certain sublanguages which do not carry over to other languages. E.g., omission of the definite article in the texts described in section 2 does not occur in the French translation of these texts. Transfer rules are required to insert the appropriate form of the definite article in the French texts.

Within a given language there may be groups of sublanguages that have many characteristics in common. For example much technical writing in English differs more in vocabulary than in syntax. Thus it may be possible to construct a parser whose syntactic rules will suffice for a number of "technical" sublanguages, with only minor variations, even though there are considerable differences in vocabulary and in the semantic ranges of individual lexical items from one sublanguage to another.

Other possibilities for further study of sublanguages not touched on here include phonological traits (e.g., in religious sermons), the growth of sublanguages along with scientific developments and cultural changes, and possible effects of usage within a sublanguage on usage in other parts of the language.

References

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

[2] DOWNING, P. (1977): On the Creation and Use of English Compound Nouns, Language, vol. 53, (4).

[3] HARRIS, Z. (1968): Mathematical Structures of Language, John Wiley and Sons, New York.

[4] ISABELLE, P.,et al (1978): TAUM-AVIATION: description d'un système de traduction automatisée de manuels d'entretien en aéronautique, COLING, Norway.

[5] KITTREDGE, R.: Variation and Homogeneity of Sublanguages, Chapter 4, this volume.

[6] LEVI, J. N. (1978): The Syntax and Semantics of Complex Nominals, Academic Press, New York.