[Proceedings of the *Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Colgate University, Hamilton, New York, August 14-16, 1985]

The Significance of Sublanguage for Automatic Translation

Richard I. Kittredge
University of Montreal

1 Introduction

This paper addresses three questions:

What is sublanguage?

Why is sublanguage analysis important for automatic translation ?

- How can a translation system take advantage of sublanguage properties?

The first of these questions appears to have a simple answer. Natural languages clearly have specialized varieties which are used in reference to restricted subject matter. We speak, for example, of the "language of chemistry" to mean a loosely defined set of sentences or texts dealing with a particular part of reality.

But when we consider the automatic translation of specialized language, we are forced to be more precise. We must describe sublanguages as coherent, rule-based systems. The attempt to write grammars for special-purpose sublanguages raises a number of theoretical and practical problems, which are only now being intensively discussed. But since the only path to high-quality automatic translation seems to lie through sublanguage (at least during the next decade or two), we have no choice but to solve these problems. This paper should therefore be considered as a brief summary and progress report.

2 What is sublanguage?

2.1 Two definitions

In the science of linguistics, one of the most difficult problems has always been the one of definitions. Language can be viewed from so many different perspectives that no single definition of a basic term such as "sentence" or "noun" is likely to suffice to characterize all aspects of the term. To further complicate matters, language is often

a fuzzy phenomenon. One is often unable to say, for example, whether a particular sentence made up *of* English words is or is not "good English".

It is therefore not surprising that the study of sublanguage meets the same definitional problems that arise in general descriptive linguistics. But looking at language in restricted domains gives a much better picture of the relationship between language and information than is the case when we study the "whole" language.

For the purposes of this paper we can informally define a "sublanguage" to be any subsystem of a language which has the following properties:

- the language subsystem is used in reference to a particular domain of discourse, or family of related domains ,

- the set of sentences and texts in the language subsystem reflect the usage of some "community" of speakers, who are normally linked by some common knowledge about the domain (facts, assumptions, etc.) which goes beyond the common knowledge of speakers of the standard language,

- the subsystem has all the "essential" properties of a linguistic system, such as "consistency", "completeness", "economy of expression", and so forth,

- the language subsystem is maximal with respect to the domain, in the mathematical sense that no larger system has the same properties.

This definition is vague on a number of points, but serves to indicate some of the important theoretical dimensions from which sublanguage can be viewed.

A more precise theoretical definition of sublanguage has been given by Harris [1]:

> "certain proper subsets of the sentences of a language may be closed under some or all of the operations defined in the language and thus constitute sublanguages of it"

In Harris' theory, the important grammatical operations are transformations between sets of sentences. Thus, for example, if the sublanguage of analysis in mathematics contains sentence (la), it also contains sentences (1b-1f):

---

1. Harris, 1968.

(1a) This theorem provides the solution to the boundary value problem.

(1b) It is this theorem that provides the solution to the boundary value problem.

(1c) What this theorem does is provide the solution to the boundary value problem.

(1d) The solution to the boundary value problem is provided by this theorem.

(1e) Does this theorem provide the solution to the boundary value problem?

(1f) This theorem does not provide the solution to the boundary value problem.

In essence, Harris' theoretical definition guarantees that a set of sentences will be considered a sublanguage only if it is grammatically complete and maximal with respect to the subject matter. But it does not tell us directly how to identify sublanguages, or how to determine their boundaries.

The search for a better theoretical definition of sublanguage should not overly concern us here [2]. If we are mainly interested in the engineering design of automatic translation systems, we should look at some cases of sublanguages which have proven to be "computationally tractable".

2.2 Sublanguages in The Real World

2.2.1 Weather bulletins

Figure 1 gives a typical weather bulletin of the kind translated by the Canadian METEO system.

FORECASTS FOR YUKON AND NORTHWESTERN BC
ISSUED BY ENVIRONMENT CANADA AT 5:30 AM PDT
FRIDAY JULY 11 1980 FOR TODAY AND SATURDAY

---

2. Theoretical questions are treated in some detail in Harris (1968), Sager (1972), Kittredge (1982) and Lehrberger (1982,1985), among others.

KLONDIKE
BEAVER CREEK
STEWART RIVER
RAIN OCCASIONALLY MIXED WITH SLEET TODAY CHANGING TO SNOW THIS EVENING. HIGHS 2 TO 4. WINDS INCREASING TO STRONG NORTHWESTERLY THIS AFTERNOON. CLOUDY WITH A FEW SHOWERS SATURDAY. HIGHS NEAR 6.

Figure 1. typical text in weather bulletin sublanguage

Weather bulletins are highly "formatted", and written in a telegraphic style. Well-formed bulletin sentences have no tensed verbs, very few articles, etc. In fact standard English grammar is of little use in describing the sentences of weather bulletins. A completely new grammar must be set up. In this specialized bulletin grammar, the "head construction" of the sentence is a string of words describing the primary weather condition, such as "partly cloudy", "rain", "clearing", etc. If we construct a grammar for this sublanguage, we find it necessary to set up a SYNTACTIC class (i.e. , <weather condition>) in which there is great SEMANTIC homogeneity, but no syntactic homogeneity in terms of the standard grammar of English. We are required to put adjective phrases, noun phrases and gerundive phrases into the same SYNTACTIC class as far as weather sentence patterns are concerned. Thus, the syntactic patterning of words and word groups in this sublanguage:

1. does NOT correspond to the syntactic classes of general English,

2. but IS a direct reflection of the important conceptual categories and relations used in the world of meteorological observation.

2.2.2 Market reports

Figure 2 shows a representative text from a second sublanguage, daily stock market reports of the kind published in most North American newspapers.

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

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

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

Figure 2. Stock market summary (Montreal Star, August 9,1979)

Reports from stock exchanges, commodities markets, agricultural markets and the like often belong to relatively well-behaved sublanguages. In stock market summaries, even though a large variety of words may be used to describe changes in the value of stocks, they fall into a very small number of classes. It is possible to write a very precise grammar for the reports in terms of the word classes that can be discovered using distributional analysis [3]. A very good correspondence can be established between the data contained in the reports and the linguistic patterns used to convey that data [4]. The correspondence is in fact so good that some market reports have been generated directly from the data.

2.2.3 Aircraft maintenance manuals

One of the most complex sublanguages which has been described in some detail is that of aircraft hydraulics manuals. Figure 3 gives a text fragment which illustrates the two distinct varieties of text in such maintenance manuals, (1) system description, and (2) maintenance instructions:

PRESSURE SWITCH

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

REMOVAL AND INSTALLATION OF
PRESSURE SWITCH - NO. 1 SYSTEM

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

Figure 3. The aviation hydraulics sublanguage

---

3. Harris, 1963.
4. Kittredge, 1983

The two types of text found in hydraulics manuals share a very large vocabulary (estimated to be on the order of 40.000 words). Despite the lexical size and syntactic complexity of this sublanguage, hydraulics manuals use fairly predictable sentence structures. What is more important, these structures can best be described in terms of sublanguage-specific word classes. Instead of stating sentence patterns in terms of major classes such as "noun phrase", "manner adverbial", etc., they can generally be stated in terms of specific word classes such as <fluid>, <instrument>, <replaceable component>, etc.

3 Why is sublanguage important for automatic translation?

We are now in a better position to see just why and how sublanguage study is useful for automatic translation. Let us first summarize our major points about sublanguage.

First, the notion of sublanguage is a theoretical construct which stresses the systematic nature of specialized language. In a sublanguage, the rules for constructing meaningful sentences can be made much more precise than in the language as a whole. Most importantly, these rules can be made in terms of word classes which are discovered by studying exactly how language is used in the particular domain (i.e., studying the distributional properties of words in texts).

Second, in a sublanguage system the rules for constructing sentences may be quite different from (and even contrary to) the rules for sentences in the "standard" language. The grammar of standard English does not "contain" the grammars of all English sublanguages, because some structures or operations exist only in particular sublanguages and have no role in standard English grammar.

Third, sublanguages may be rather small (e.g., weather bulletins), or very large (e.g., texts in aircraft hydraulics or organic chemistry). What qualifies a variety of language as a sublanguage is not its size or complexity, but its adherence to systematic usage. The "well-behaved" sublanguages of science and technology may use terminology from the everyday world, but this "seepage" from general language is usually possible only in specific grammatical positions. We must admit that some sublanguages appear to be more systematic than others. It is in fact the DEGREE of systematicity which will determine how amenable a

sublanguage is to automatic translation.

3.1 The importance of sublanguage grammar during analysis

It is generally agreed that the most difficult part of automatic translation is that of obtaining the correct analysis structure for each input sentence. If a source language analyzer is based on a sublanguage grammar, instead of (or in addition to) a grammar of the "whole" language, then a significant gain in efficiency is possible[5]. First, the parsing time is reduced, since sublanguage grammars are always smaller than the grammars of whole languages. Second, the problem of structural and lexical ambiguity is greatly reduced, since many interpretations or analyses which are possible in the standard language are not "legal" (i.e., they are meaningless) in the sublanguage, and therefore can be ruled out. In cases where technical or scientific language contains reference to the outside world, a good sublanguage grammar will also state where in the sentences or texts this intrusive language can be expected [6].

3.2 Help from sublanguage grammar during word translation and structural transfer

Even if analysis is the most difficult problem in automatic translation, lexical and structural transfer can pose many thorny problems as well. There is now strong evidence that languages are more similar in sentence and text structure within scientific and technical writing than in non-technical writing (e.g., newspaper editorials). Examination of English and French sublanguages for a variety of structural features shows that corresponding sublanguages of English and French are often structurally more similar

---

5. Slocum, 1985, reports on experiments conducted to this end. Isabelle, 1984, summarizes this approach as used at the TAUM project. Sager, 1981, presents an English analyzer which uses a general grammar, but filters parses with a sublanguage- specific "restriction grammar".

6. Sager, 1972, reports on how metascience predicates embed science predicates. Kittredge, 1983, deals with grammatical subordination used for embedding reference to a secondary domain .

than are two dissimilar sublanguages of the same language[7]. It is thus important to write transfer grammars as mappings between corresponding sublanguage grammars, both on the level of sentence and text. The functional equivalence of sentences can then be computed with respect to the particular sublanguage, and not the whole language. Furthermore, when an analyzed sentence carries the word class labels assigned by the sublanguage grammar, word translation equivalence is much easier to compute. This is because a word is translated as a function of its position in the analysis tree; since the syntactic labels of the tree have semantic import, this means that a word can be translated as a function of its semantic relations to its neighbors.

4 Preparation for Sublanguage-Based Automatic Translation

Building a translation system which depends partly or entirely on a sublanguage grammar is a painstaking process. It can pay off handsomely in terms of translation quality, but only in the long run, and when the volume of texts justifies the development investment. In the case of the Canadian METEO system, the investment has already paid for itself many times over. In the case of the AVIATION system, the development time proved to be too long to meet the practical needs of the user. It is therefore of crucial importance to choose a sublanguage of the right size and complexity. Given the small number of sublanguage-based systems now working, any new system inevitably involves a significant component of basic research, both in linguistic description, and in strategies for optimizing analysis and transfer.

4.1 Comparing candidate sublanguages

Before picking a particular sublanguage for system design, it may be advisable to compare candidate sublanguages and estimate the computational tractability of the most likely choices. Methods for doing this are still experimental, but certain guidelines can be given.

The simplest measure of sublanguage size and complexity involves only its vocabulary. One can plot a curve of vocabulary growth against number of running text words in a

---

7. Kittredge, 1982

corpus which is considered representative of the sublanguage. Figure 4 gives these growth curves for nine separate sublanguages based on a recent study carried out for the Canadian Translations Bureau [8].

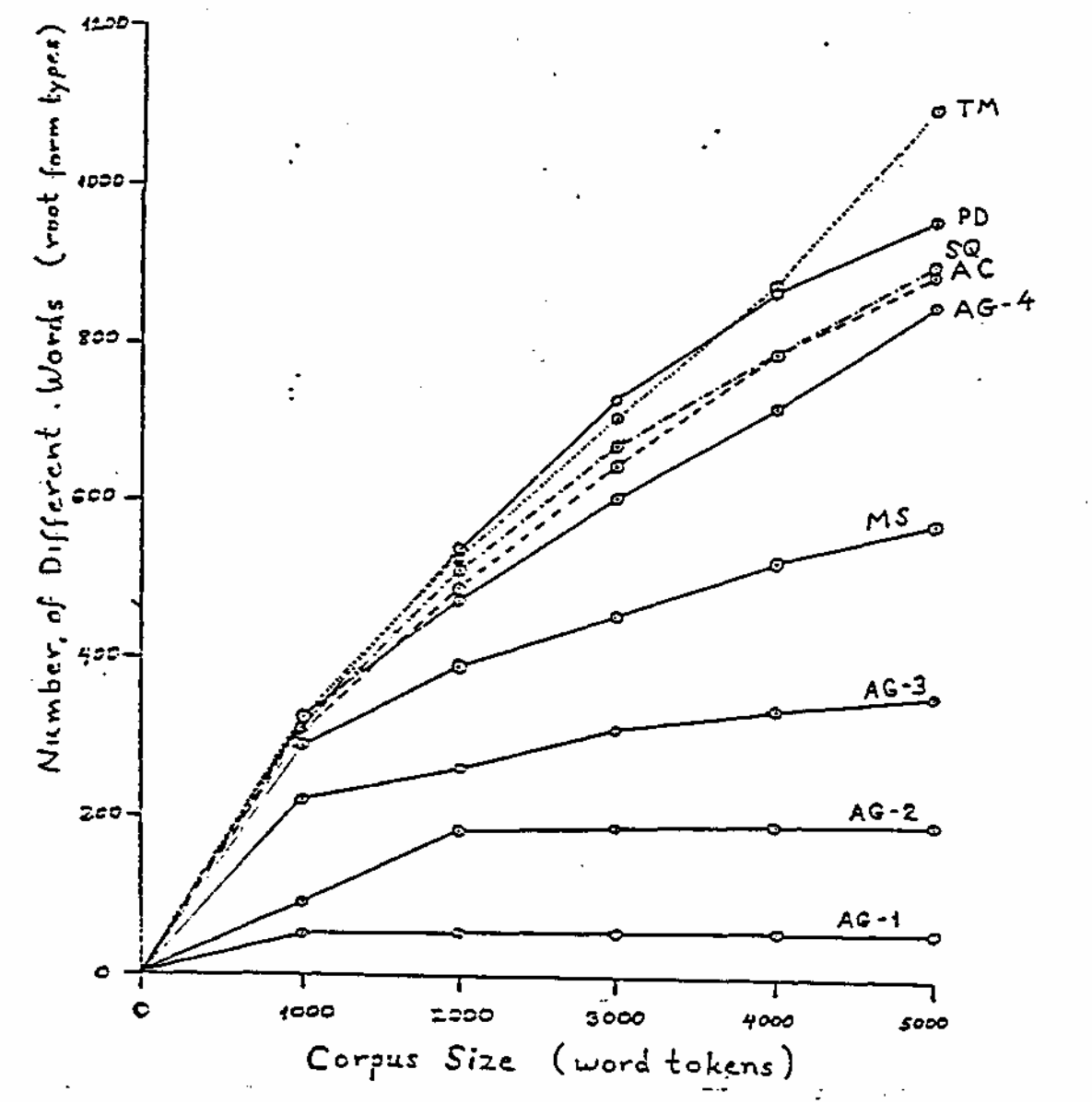

Figure 4 . Comparison of vocabulary growth rate curves.

To the extent that these curves flatten out after a certain point, one may assume that the sublanguage word usage is

---

8. Kittredge, 1983

relatively constrained. From the slope of the curve and the maximal value of different words found in the largest corpus used, one can estimate the total size of the vocabulary. Thus in figure 4, the agricultural market reports marked AG-2 used less than 200 different root words (lexemes) in a 5000-word corpus, and showed a marked tendency to vocabulary closure. In contrast, the technical registry describing foreign trademarks approved for use in Canada, marked TM, showed no sign of lexical convergence, with over 1100 distinct words used in 5000 words of running text. Although several thousand words of representative texts may give a rough indication of closure tendencies, large sublanguages will require many times that sample size for reasonably accurate estimates of convergence and vocabulary size to be made.

Vocabulary growth curves are easy to compute and present only minor problems of methodology, but they do not give the most accurate picture of sublanguage closure. What is more important than vocabulary growth is the degree of closure of the grammar itself. One recent attempt to measure grammatical closure [9] has used the number of grammatical production rules of a general English grammar which were applied in analyzing a corpus of sublanguage texts. Still better than this would be to measure the specific sublanguage grammar rules (assuming that no other grammar exists) needed to account for a growing corpus. This requires rewriting the sublanguage grammar several times for a growing corpus (a lot of work!), but should give the most accurate prediction of sublanguage closure.

4.2 Estimating computational tractability

Estimating the computational tractability of sublanguage texts goes beyond the question of sublanguage closure. In the case of automatic translation, the feasibility of correctly analyzing the source language texts is somewhat separate from the transfer problem.

For predicting the difficulties of analyzing English texts, some of the following questions are relevant:

- Is there ellipsis of articles, copula, object noun phrases, etc.? (this is frequent in many sublanguages and often a factor in sentences which are structurally ambiguous, even within the sublanguage);

---

9. Grishman et al., 1984

- Is there frequent conjunction using "and" (which raises problems in determining the scope of its arguments)?;

- Are there quantifier words and negation (which raises still other scope problems)?;

- Does the sublanguage use long nominal compounds (the bane of the TAUM-AVIATION project)?; if so, these must be analysed for scope of modification and often paraphrased using domain knowledge before translation can be attempted);

- Are there parenthetical expressions (which raise questions concerning points of attachment in the syntactic structure)?;

- Is a text grammar possible for the sublanguage? if so, can it be made precise enough to help in structural and lexical disambiguation?

- Do co-referential pronouns link consecutive sentences? if so, what are the problems of determining co-reference within the sublanguage?

- What sort of cohesion devices does the sublanguage use to link consecutive sentences? are synonyms frequently used (as in stock market reports) or avoided (as in technical manuals)? how much can be inferred from the use of a given cohesion device?

For predicting the problems of making correct translation correspondences, much less is known of a general nature. The experience of U.Montreal's TAUM project, which concentrated entirely on English-to-French . translation, showed that some of the following grammatical and semantic phenomena were generally problematic in establishing correspondence between the two languages. In most cases, the restriction of the correspondence problem to a technical sublanguage allowed fairly good rules to be set up:

- tense and aspect: most English sublanguages use a subset of possible forms, and their functional equivalents in French are both idiosyncratic for the French sublanguage, and simpler than in general French;

- verb modality: translations of English "can", "must", "should", etc. present many problems for general language that can be solved in specific sublanguages;

- passive: although English passive has at least six possible renderings in French, within aircraft hydraulics manuals the correspondence algorithm is far simpler;

- lexical choice: there is a complex interaction between structural transfer rules and the valency (i.e., semantic case slots) of available verbs in the target language; this complexity is usually much reduced within the limits of a given sublanguage

- textual constraints: this is one area where corresponding technical sublanguages of English and French were found to share many of the same features; thus, textual constraints of the target language must be used properly to given a natural-sounding output text

REFERENCES

Grishman,R. & Kittredge,R. (in press) Analyzing Language in Restricted Domains, Erlbaum.

Grishman,R. et al. (198A) "Automatic Determination of Sublanguage Syntactic Usage" in Proceedings of ACL/COLING84, Association for Computational Linguistics.

Harris,Z. (1963) Discourse Analysis Reprints, Mouton.

Harris,Z. (1968) Mathematical Structures of Language, Wiley-Interscience.

Isabelle,P. (1984) "Automatic Translation at the TAUM Project"

Kittredge,R. (1982) "Variation and Homogeneity of Sublanguages" in Kittredge & Lehrberger (1982).

Kittredge,R. (1983) "Semantic Processing of Texts in Restricted Sublanguages" Computers and Mathematics with Applications, vol.9, no.1, pp.45-58.

Kittredge,R. (1983) Sublanguage-Specific Computer Aids to Translation: A Survey of the Most Promising Application Areas, Secretary of State Department, Government of Canada.

Kittredge,R. & Lehrberger,J. (1982) Sublanguage: Studies of Language in Restricted Semantic Domains, deGruyter.

Lehrberger,J. (1982) "Automatic Translation and the Concept of Sublanguage" in Kittredge & Lehrberger (1982).

Lehrberger,J. (1985) "Sublanguage Analysis" in Grishman & Kittredge (1982).

Sager,N. (1972) "Syntactic Formatting of Science Information", AFIPS Proceedings, reprinted in Kittredge & Lehrberger (1982).

Slocum,J. (1985) "How one might Automatically Identify and Adapt to a Sublanguage: an initial exploration" in Grishman & Kittredge (1985).