Gouvernement du Canada
Ministère des Communications

Government of Canada
Department of Communications

Le Centre canadien de recherche
sur l'informatisation du travail

Canadian Workplace
Automation Research Centre

MACHINE TRANSLATABILITY OF COMPUTER SYSTEM MANUALS

by

John Lehrberger

Canada

2. MACHINE TRANSLATABILITY OF COMPUTER SYSTEM MANUALS

by

11 John Lehrberger

Canadian Workplace Automation
Research Centre

Department of Communications

Laval
March 1987

Cat. no. Co 28-1/27-1988E

ISBN 0-662-16587-X

The views expressed in this report are those of the author only.

* Rapport complet en français aussi disponible.

TABLE OF CONTENTS

PAGE

1. INTRODUCTION 1

1.1 Purpose of this study 1
1.2 The corpus 1
1.3 Methodology 2

2. FINDINGS 5

2.1 Size of vocabulary 5
2.2 Semantic range 11
2.3 Homography 12
2.4 Syntactic properties 12

3. CONCLUSION 19

APPENDIX

Corpus in original format

1. INTRODUCTION

1.1 Purpose of this study

The purpose of this study is to examine the linguistic complexity of certain texts currently translated at the Translation Bureau (Government of Canada) with respect to their "translatability by computer." The texts consist of descriptions of particular computer systems, their installation and their use. The aim is to determine, on the basis of a sampling of such texts, the feasibility of translating these documents by MT systems employing current technology; speculation about future developments that may overcome today's outstanding linguistic obstacles is not taken into account here.

1.2 The corpus

The overall volume of computer related texts to be translated is estimated at more than four million words annually, with 50 to 60 percent of that volume being of the type of document represented here. Translation turnaround time is normally two to four weeks, or possibly longer, depending on workload, delays in the typing pool and other priorities. The texts used in this study were provided by the Translation Bureau and the sample selected for detailed examination (the corpus) consisted of 9914 words from three manuals, as indicated below.

CEIC Systems Mainframe Manual: Burroughs B3900	3069 words
Rivera Hartling Systems: How to install WIN using the WINSTALL diskette	3929 words
The Bureau of Management Consulting (BMC) Skills Inventory User's Manual	2916 words

These documents will be referred to hereafter as CEIC, RH and BMC respectively, and the field they represent will be called the COMPUT domain.

1.3 Methodology

Linguistic complexity is not a sharply defined concept. However, an important fact for this study is that certain elements, which clearly contribute to the linguistic complexity of texts, are also well known as obstacles to successful machine translation (MT). This study examined the extent such elements were present in the corpus and used this information to analyze the feasibility of MT for the larger body of texts represented by the corpus. The following specific factors were considered.

A) Size of vocabulary

The rate at which new lexical items occur as more text is added (the lexical growth rate) was used to estimate the number of lexical items for the entire COMPUT domain. The various inflected forms of a particular word were not counted as separate lexical items but were represented by a single base form. Thus "go," "goes," "going," "went," "gone" were considered simply as variants of the base form "go." Alternatively, each such set of related forms can be referred to as a lexical lemma. Intercategorial homographs were not counted as separate base forms; for example, "program" occurs as both noun and verb, but is counted as just one base form. Two numerals with the same number of digits were counted as a single form (and as a single base form), since the number of numerals is not a significant factor in linguistic complexity.

B) Semantic range

Semantic range affects vocabulary size and the incidence of homography as well as the number of semantic classes needed to express selectional restrictions and other co-occurrences in the texts. Although the notion of semantic range itself is not easy to quantify, certain aspects of it are fairly accessible to investigation: i) the variety of subdomains included in a given text domain; ii) the extent of hierarchies of subdomains, sub-subdomains, etc.; and iii) how much knowledge from outside the domain is needed to understand texts within it. Such information provides at least a basis for comparison of the given text domain with others that may be of interest.

C) Homography

Only homographs belonging to different parts of speech have been considered. This should not be taken to imply that homographs within the same grammatical category (for example, facility -- abstract vs. concrete) are unimportant; on the contrary, they contribute significantly to linguistic complexity. But the time required for such a study did not permit its inclusion here.

D) Syntactic properties

The main consideration here was occurrence rate of those syntactic constructions known to present difficulties for MT. Sentence length was also taken into account. The following list was the basis for this study of syntactic complexity as it affects machine translatability.

i) Presence of interrogatives and imperatives as well as declarative sentences.

ii) Topicalization: passive, cleft and pseudo-cleft sentences, extraposition.

iii) Subordination: relative clause, clause introduced by subordinate conjunction, sentential complement of verb or adjective, infinitival and gerundive clauses functioning adverbially.

iv) Co-ordination by "and" and "or."

v) Noun stacking: strings of nouns, possibly interspersed with adjectives, and possibly including co-ordinate conjunctions.

vi) Parentheticals: distribution of parentheticals within the sentence and the kinds of expressions included.

vii) Ellipsis: contexts in which ellipses occur, type of material ellipted, frequency of occurence.

viii) Tense and auxiliaries.

ix) Sentence length: only sentences not containing lists were used to compute average length, since those lists may be extremely long without contributing significantly to syntactic complexity.

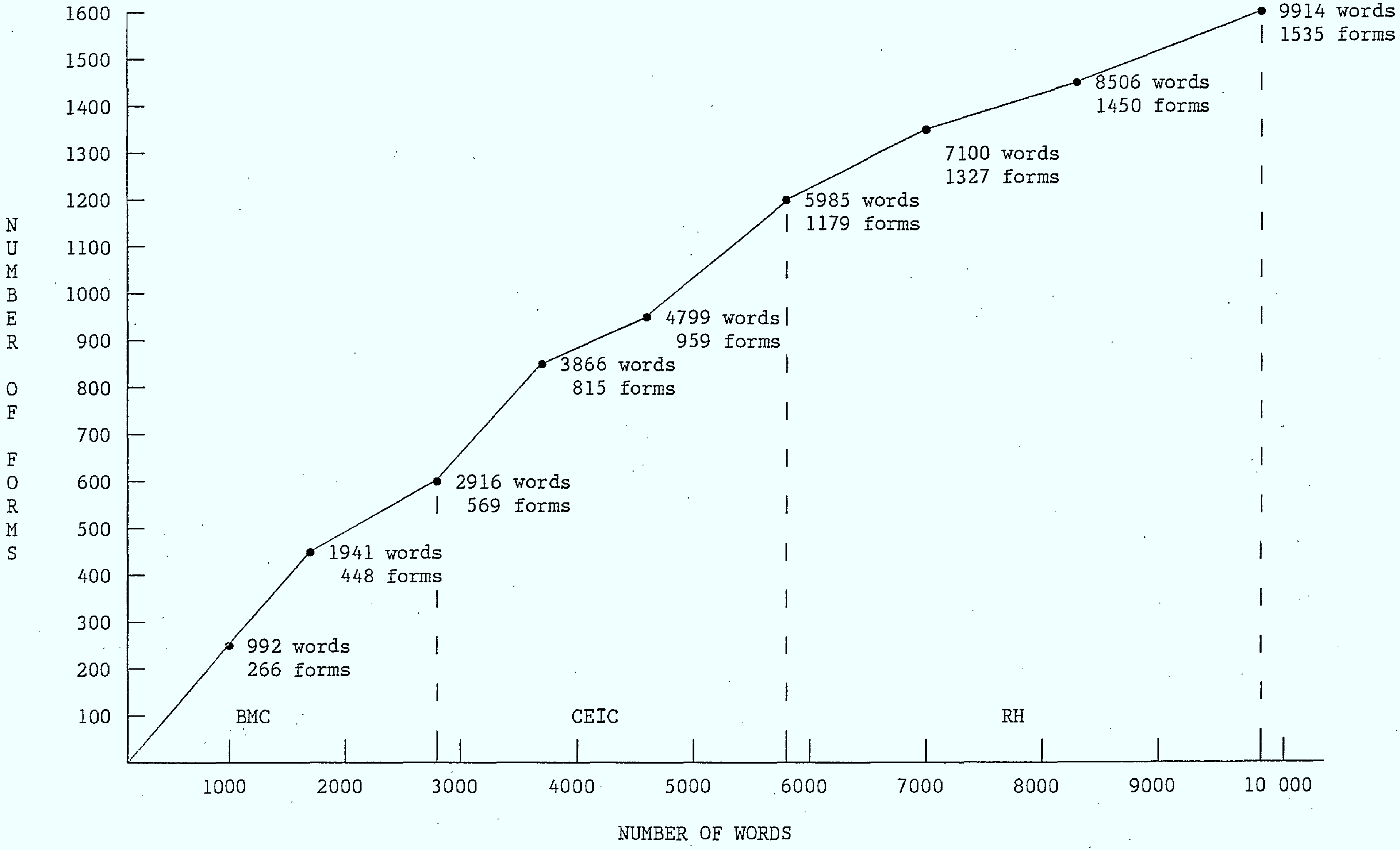
x) Text structure: intersentential pronouns, text formatting.

2. FINDINGS

2.1 Size of vocabulary

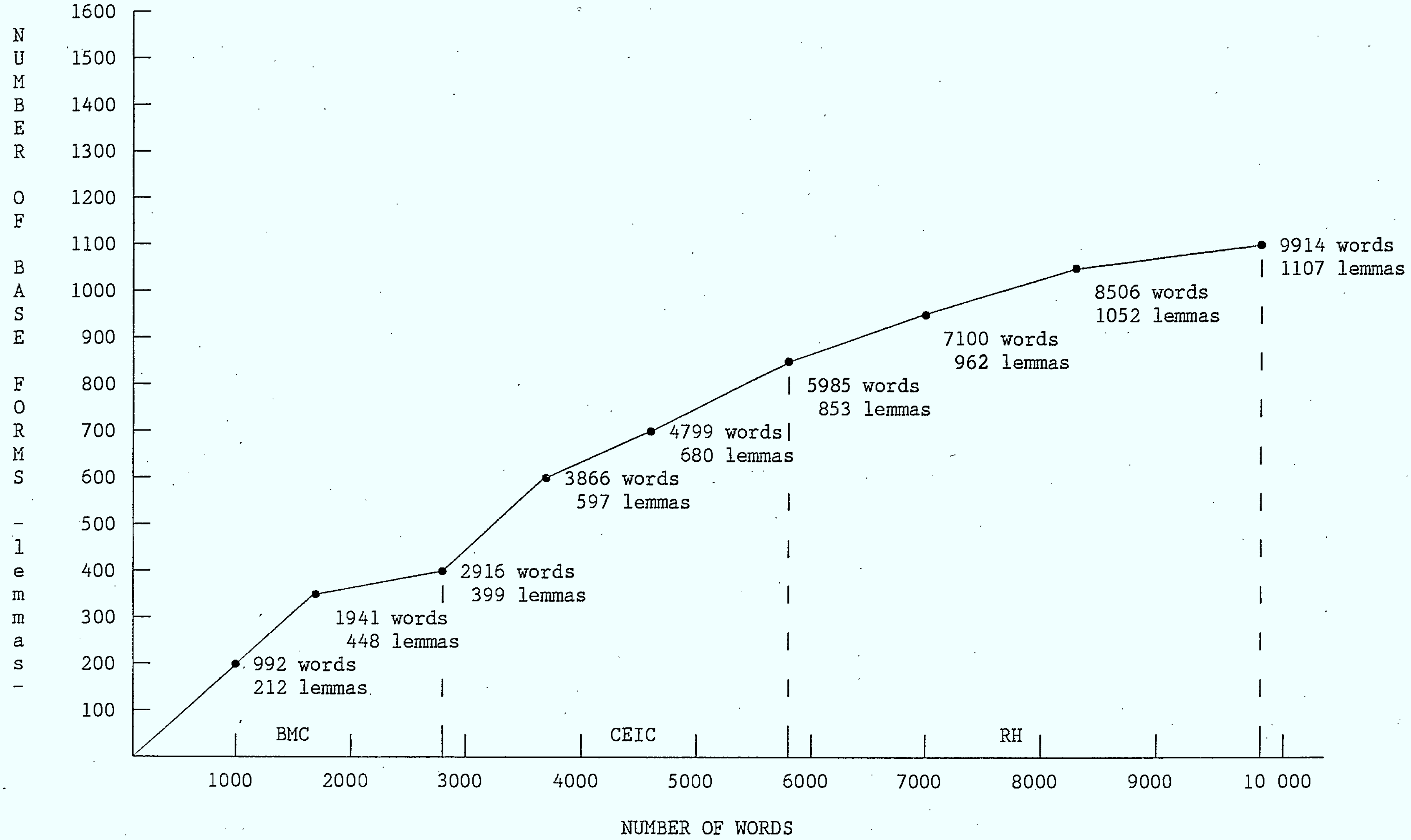
Using FATRAS (Full-Text Retrieval System) it was determined that the corpus of 9914 words taken from the BMC, CEIC and RH manuals contained 1535 different forms. From the alphabetical listing of these forms, the base forms (or lemmas) were obtained by inspection. The number of base forms represented in the corpus was 1107. The vocabulary growth rate is shown in Figure 1 in terms of forms, and in Figure 2 in terms of base forms. Hereafter, the term vocabulary will be used to refer only to the set of base forms; these correspond more closely to the dictionary entries of an MT system, assuming it has a morphological component that recognizes at least the inflectional variants of verbs and nouns. Figure 2 then forms the basis for the following discussion of lexical growth rate in the COMPUT domain.

FIGURE 1



1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
NUMBER OF FORMS
992 words
266 forms
1941 words
448 forms
2916 words
569 forms
3866 words
815 forms
4799 words
959 forms
5985 words
1179 forms
7100 words
1327 forms
8506 words
1450 forms
9914 words
1535 forms
BMC
CEIC
RH
1000
2000
3000
4000
5000
6000
7000
8000
9000
10 000
NUMBER OF WORDS

FIGURE 2


NUMBER OF BASE FORMS - lemmas -
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
992 words
212 lemmas
1941 words
448 lemmas
2916 words
399 lemmas
3866 words
597 lemmas
4799 words
680 lemmas
5985 words
853 lemmas
7100 words
962 lemmas
8506 words
1052 lemmas
9914 words
1107 lemmas
BMC
CEIC
RH
1000
2000
3000
4000
5000
6000
7000
8000
9000
10 000
NUMBER OF WORDS

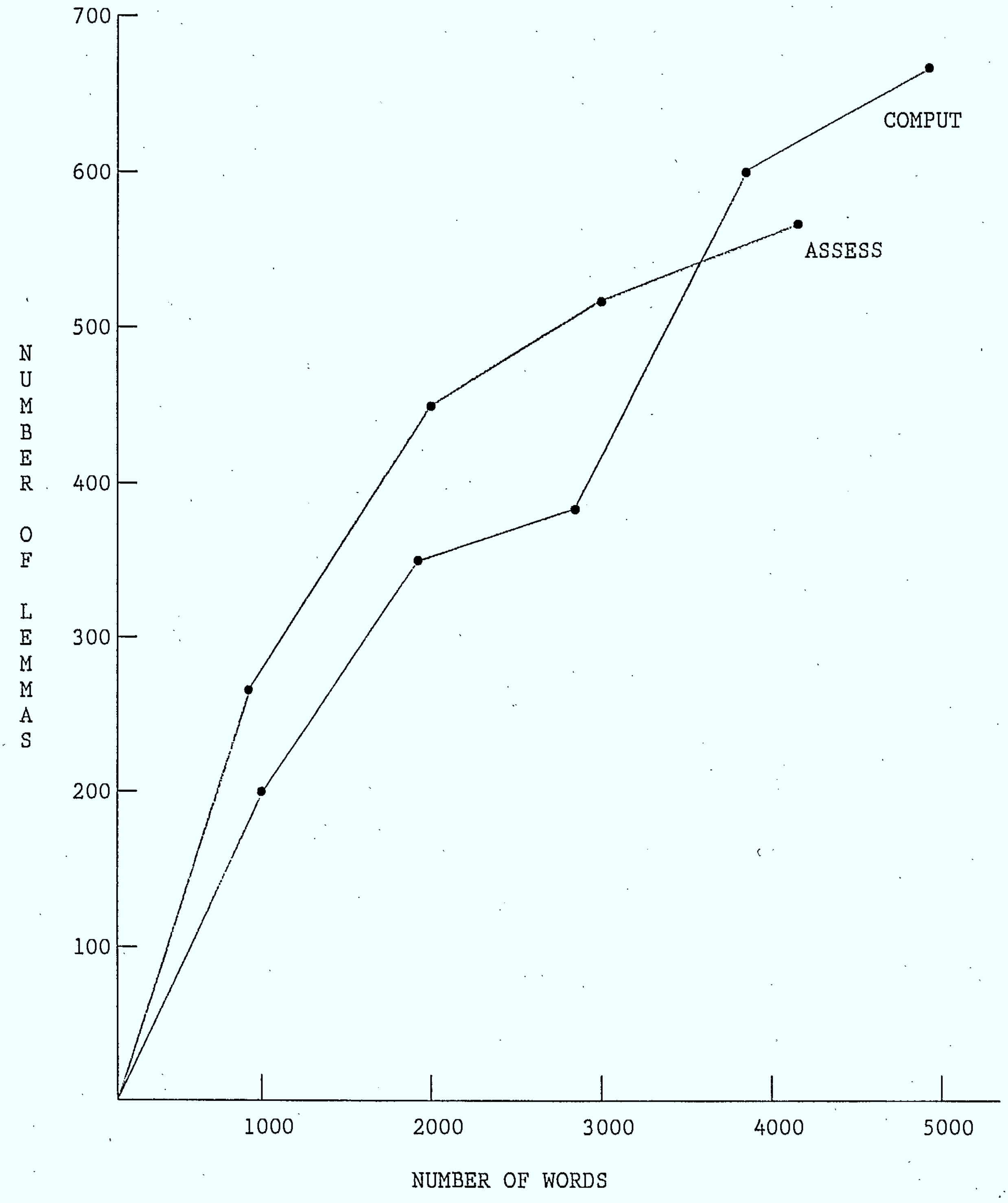
As a sample is built up from a homogeneous set of texts the resulting vocabulary is expected to increase, but at a decreasing rate. That is, as more text is added to a sample from a given domain, the rate at which new base forms are added to the vocabulary is expected to fall off steadily after a few thousand words of text -- or a few hundred words, for some very small domains. The slope of the lexical growth rate curve ultimately approaches zero and the curve is said to "level off."

An examination of the first three points on the curve in Figure 2 reveals a trend toward levelling off, but then a sharp increase occurs in the next segment. That increase corresponds to a change from BMC to CEIC texts. Changing from one manual to another can be expected to affect the curve somewhat, although in this case the effect seems rather large. The final segment of the BMC section shows a lexical growth rate of $\frac{41}{2916 - 1941} \times 1000$ (=42) new base forms per thousand words of text, while the initial segment of the CEIC section shows a rate of $\frac{198}{3866 - 2916} \times 1000$ (=208.4) new base forms per thousand words of text. This five-fold increase in lexical growth rate is more indicative of change in domain than of a change in manuals within the same domain. In fact, the CEIC manual does differ substantially from the BMC, and these differences will be discussed later.

Figure 2 shows a second increase in the growth rate, this time within the CEIC section of the curve. It occurs in the final segment of CEIC, following an apparent levelling off in the middle segment. Specifically, the growth rates for the three segments of CEIC are 208.4, 89 and 145.9 new base forms per thousand words of text. The increase from 89 to 145.9, a factor of about 1.6, is considerably less than the five-fold increase between manuals, but it is significant in that it represents a departure from the trend toward a steadily decreasing rate throughout the rest of the sample. Such changes in the shape of the curve make it difficult to predict just where the final levelling off will occur.

In Figure 3 the COMPUT corpus is compared with the ASSESS corpus studied by Elliott Macklovitch ("Machine Translation of the TOM Manuals," December 1985). Up to about 3000 words of text, which included the BMC part of COMPUT, the ASSESS vocabulary grows faster than that of COMPUT. Then a crossover occurs as the CEIC text is taken into account. Although there is no data on ASSESS beyond 4069 words of text, the vocabulary of the COMPUT corpus can be expected to remain higher because of the variety of material it contains. The total vocabulary for all manuals represented by COMPUT can be expected to contain more than 5000 base forms. The data from the present corpus do not warrant setting an upper bound on the size of vocabulary.

FIGURE 3


700
600
500
400
300
200
100
NUMBER OF LEMMAS
1000
2000
3000
4000
5000
NUMBER OF WORDS
COMPUT
ASSESS

2.2 Semantic range

The COMPUT domain has three distinct subdomains:

i) descriptions of particular computer systems,

ii) instructions for their installation, and

iii) instructions for the use of computer systems for specific purposes.

These are illustrated in the corpus by CEIC, RH and BMC respectively. Given the number and variety of systems and their possible uses by the Government of Canada, the domain is rather large. Although there is a great deal of overlap in the vocabularies of the subdomains, Figure 2 shows that there are also considerable differences. Thus a large infusion of new vocabulary accompanies the change from BMC to CEIC. The existence of significant sub-subdomains is indicated by the increase in vocabulary growth rate in the third segment of the CEIC part of the corpus. This increase coincides with the change from Section 1 ("System Description") of Chapter 1 to Section 2 ("Disk Pack Subsystem"). A great variety of topics within the CEIC manual that are not included in the corpus suggests other large infusions of new vocabulary from CEIC type texts.

The inclusion of BMC type texts in the COMPUT domain brings in a good deal of material that does not have a direct bearing on computers. For example, the BMC manual explains "how to identify registered consultants or firms with specific characteristics and skills," and includes a section titled "Preliminary Contact and Interview by BMC Consultant." Apparently the boundaries of the COMPUT domain are not as precisely drawn as one would like; it is not clear how much knowledge from outside the computer field proper is incorporated within the COMPUT domain, nor is it obvious whether the two types of knowledge could be easily separated in the texts.

On the whole, it appears that a fairly large number of semantic classes would be required for the description of the entire domain.

2.3 Homography

There are many homographs in the corpus, most of them noun/verb pairs: call, command, control, file, drive, form, function, key, input, interface, line, number, process, program, return, run, screen, search, time, type, use, etc. These homographs are among the more frequently occurring non-function words, which creates a problem for parsing.

A number of important verbs in this domain occur as nouns or noun modifiers, although they would normally occur only as verbs elsewhere: execute, read, write, enter, fetch, manipulate. For example, "by pressing the enter key" (BMC), "so that read and write can be transferred" (CEIC), "FETCH MODULE" and "MANIPULATE MODULE" (CEIC), "the write enable switch" (CEIC).

The prevalence of intercategorial homography contributes significantly to the linguistic complexity of these texts. Time did not permit an investigation of the extent of inner-categorial homography (different meanings of a word within the same part of speech) although it is clearly present and should be taken into account.

2.4 Syntactic properties

A) Interrogative and imperative sentences

The corpus contains interrogatives of both the yes/no type ("Were there any errors in the process?") and the wh- type ("What is the total number of workstations...?"). They occur in the RH text and are mainly yes/no questions. Imperatives abound, especially in RH, which consists for the most part of instructions rather than descriptions.

B) Topicalization

Passive sentences occur very frequently, nearly always with the agent omitted. Extraposition, cleft and pseudo-cleft sentences do not occur.

C) Subordination

The prevalence of subordinate clauses contributes heavily to the syntactic complexity of these texts. Relative clauses are numerous, sometimes with wh- deletion and more than 50 percent of the time with wh + BE deletion. Clauses introduced by subordinate conjunctions are equally numerous. Clauses occur frequently as complements of verbs (such as: agree, allow, ask, avoid, cause, choose, ensure, explain, fail, help, indicate, insist, intend, make, mean, note, permit, prefer, remember, request, require, serve, state, try, verify, wish, etc.) and occasionally as complements of adjectives (able, sure, ready) or nouns (ability, chance). There are also many infinitival and gerundive clauses functioning adverbially.

D) Co-ordination

There were 227 occurrences of "and" in the corpus, exceeded in number only by "the," and 118 occurrences of "or." They are accompanied by the usual problems regarding scope, for example:

Special((Command and Editing)Keys)
Special (Command and(Editing Keys))
(Special Command) and (Editing Keys)
(Special (Command and Editing)) Keys

And there are many longer examples involving combinations of "and" and "or."

Ellipsis is common in association with co-ordinate conjunction: "The processor reads an instruction from memory, checks it for validity, decodes and resolves its contents." Co-ordination of noun phrases is most frequent in the corpus, verbs or verb phrases are quite frequent too, and sentences somewhat less so; co-ordination of other constituents occurs sparingly in BMC and RH, but more frequently than sentence co-ordination in CEIC. On the whole, co-ordination is a major factor in the linguistic complexity of these texts.

E) Noun stacking

The frequency of occurence of complex nominals that exhibit noun stacking is very high in the corpus, especially in CEIC. This is characteristic of technical descriptions of complex systems. Many of these compounds contain more than two nouns (for example, "WIN Exports data base files," "Consultants Information Bank Registration Code Book," "remote applications batch processing," etc.). Of course, some nominal compounds could be entered in the dictionary as idioms, in particular those that are capitalized. However, even in such cases they are formed by productive processes which can yield many more compounds of a similar nature in texts throughout the domain. Noun stacking is a complex problem which is not likely to be solved by stacking the dictionary with idioms.

F) Parentheticals

Parenthetical expressions in the corpus are enclosed by parentheses, rectangular brackets and by dashes as well. As an example of the latter, in RH we find: "If you are installing more than one workstation, you will now take -- IN ORDER (Go to station B, then C, then D, etc.) -- the winstall diskette to the next machine and follow the routines which appear on the screen." Here we have one parenthetical within another and the entire insertion occurs between the verb "take" and its direct object. In the following example the insertion separates a relative clause from its antecedent: "...the different application programs (WIN is an example of an application program) that you and your co-workers will use..." Parentheticals occur in almost any position here, and a parser would often find it difficult to place them correctly in the constituent structure of the sentence.

As for the kinds of expressions that are enclosed, they vary widely, including some which do not belong to any generally recognized constituent type, and may simply enclose part of a normal sentence without really constituting an insertion in the usual sense: "Enter the letter corresponding to one of the above functions (or strike F9)." Of course, many occurrences are simply acronyms or other abbreviations immediately following a complex name: "ADDRESS STORE AND MANIPULATE MODULE (ASAM)," "Metal Oxide Semi-conductor (MOS) memory."

G) Ellipsis

The regular omission of agents in passives and the deletion of wh + BE in relative clauses (very high frequency of occurrence) have been mentioned earlier, as well as ellipsis in the presence of co-ordinate conjunctions. It is also common with subordinate conjunctions: "Press either T for tape or D for diskettes. <u>If diskettes</u>, you will have to...," "press any key <u>when ready</u> to continue."

Subject NP deletion occurs in other contexts too. For example, in tables describing the use of the IBM PC keyboard we find: "Moves the cursor one line up...", " "Deletes" the character where the cursor is positioned," etc. In this case the subject (the name of the key) is in a column to the left of the description, separated from it by a vertical line in the table. Within the same table the subject NP does sometimes appear in the description. The same phenomenon occurs in the CEIC manual, but without any vertical separating line: "Indicates the state of the write enable switch..."

Other forms of ellipsis include object deletion: "Press again and the numeric pad is cancelled" and verb deletion, especially BE: "Temperature critical, sensed in the linear motor." Again there is a lack of consistency, as the sentence immediately following the last example is: "The RPM low is sensed." Article deletion likewise occurs, inconsistently, in tables of descriptions. Thus we find two occurences of "...is used with numeric key pad" in the same table.

H) Tense and auxiliaries

Simple past tense occurs in the corpus, though infrequently: "The carriage hit the end stop.", "...the disk is up to full speed or failed to move out of home position." Perfect aspect is used somewhat more frequently: "the name that you have given", "you will have received either a single tape or...", "after a complete command or answer has been typed". Modals of all sorts are used frequently. "Will" occurs 106 times in the corpus; most often it is not used to indicate future time, which creates opportunities for mistranslation into French. Semi-modals also occur (be to, have to, need to).

I) Sentence length

Counting only non-enumerative sentences, the average sentence length for each section of the corpus is as follows:

CEIC	19.5	(11.5% of these have 30 or more words)
BMC	16.6	(8.3% of these have 30 or more words)
RH	14.5	(5% of these have 30 or more words)

J) Text structure

i) Sentence linking. There were about 40 occurrences of intersentential pronominalization in this 46 page corpus.

ii) "Information structure." This refers to the correlation between information and syntactic structure (that is, "information formats") used with some success by Naomi Sager and her colleagues at New York University for information retrieval with hospital records, etc. The text structure, or discourse structure, in the COMPUT domain does not lend itself to this approach; neither does it seem feasible for CEIC, BMC or RH individually.

iii) Text formatting. There are tables, diagrams and illustrations in the corpus. Aside from the problem of delivering the translation in the same format as the original text, these features also create a problem in identifying the unit of translation in some cases. For example, consider a segment of the FETCH/EXECUTE OVERLAP DIAGRAM in CEIC:

FETCH INSTRUCTION FROM MEMORY	DECODE INSTRUCTION	EXECUTE INSTRUCTION	FETCH THE NEXT INSTRUCTION

Here the parser cannot simply sweep across the page from left to right or it will fail to identify the sentence units ("Fetch instruction from memory," etc.). On the other hand, in some of the tables in RH we find examples such as:

8	Moves the cursor one line up for each key stroke

In this example "8" is actually the subject of "moves," and so completes the sentence, continuing to the right. Unfortunately, this is not uniform, even within that series of tables. Also, an examination of the various enumerative structures used in the corpus shows that the expressions in a particular enumeration are not always of the same category. Thus in the same column of an enumeration in CEIC we find "Indicator lights when ..." followed by "Indicates the state of the..."; and in BMC, " "Deletes" the character where the cursor is positioned" followed by " "Insert" sets the keyboard to insert mode so that ...."

In summary, the text structure within this corpus indicates that the structure within the entire set of manuals in the COMPUT domain will be quite varied and, in certain respects, complex.

3. CONCLUSION

The annual volume of texts in the COMPUT domain is high, exceeding two million words per year. The quality of translation is to be "accurate and straightforward, not flowery French." This indicates a tolerance of rigid style, as would be expected since the main concern of technical manuals in general is accuracy and clarity.

The findings in Section 2 indicate that texts in the COMPUT domain rate high in syntactic complexity, that the semantic domain is rather extensive (and, considering BMC-type manuals, somewhat indefinite), and that the vocabulary is sufficiently large to make homography a serious problem.

Although COMPUT was treated as a single domain for the purpose of this study, it is apparent that BMC, CEIC and RH differ substantially from one another. Much of the text in BMC is not about the computer system, but rather the Skills Inventory and "how to propose a consultant or consulting firm for registration into the Skills Inventory." This type of user's manual stretches the semantic domain a bit far. RH is strictly concerned with installation procedures, which is evidenced by a relatively large number of imperatives -- two and a half times as many as in BMC. The CEIC text, a technical description of a data processing system, is syntactically very complex. It has the greatest average sentence length and the highest incidence of noun stacking, the latter constituting one of the most serious problems for parsing.

On the whole, these texts rate high in linguistic complexity; they contain a full complement of structures that are known obstacles to successful MT. In spite of the high annual volume of texts and tolerance of rigid style that make the COMPUT domain a tempting prospect, the author believes it is not a good candidate for machine translation, given the present state of the art.

LEHRBERGER, JOHN
--Machine translatability
of computer system manuals

Pour plus de détails,
veuillez communiquer avec :

Le Centre canadien de recherche
sur l'informatisation du travail
1575, boulevard Chomedey
Laval (Québec)
H7V 2X2
(514) 682-3400

For more information,
please contact:

Canadian Workplace
Automation Research Centre
1575 Chomedey Blvd.
Laval, Quebec
H7V 2X2
(514) 682-3400