Lingvisticæ Investigationes Supplementa

Studies in French & General Linguistics / Études en Linguistique Française et Générale

LIS

Volume 15

John Lehrberger and Laurent Bourbeau

MACHINE TRANSLATION

Linguistic characteristics of MT systems and general methodology of evaluation

JOHN BENJAMINS PUBLISHING COMPANY

MACHINE TRANSLATION

Linguistic characteristics of MT systems and general methodology of evaluation

by

John Lehrberger
Laurent Bourbeau

JOHN BENJAMINS PUBLISHING COMPANY
Amsterdam/Philadelphia

1988

Library of Congress Cataloging in Publication Data

Lehrberger, John.
Machine translation.

(Lingvisticæ investigationes. Supplementa, ISSN 0165-7569; v. 15)
Bibliography: p.
1. Machine translating. I. Bourbeau, laurent. II. Title. III. Series.
P308.L44 1988 418'.02 87-17441
ISBN 90 272 3124 9 (alk. paper)

MACHINE TRANSLATION

LINGVISTICÆ INVESTIGATIONES: SUPPLEMENTA

Studies in French & General Linguistics /
Etudes en Linguistique Française et Générale

This series has been established as a companion series to the periodical "LINGVISTICÆ INVESTIGATIONES", which started publication in 1977. It is published jointly by the Linguistic Department of the University of Paris-VIII and the Laboratoire d'Automatique Documentaire et Linguistique du C.N.R.S. (Paris 7).

* * * * *

Volume 15

John Lehrberger and Laurent Bourbeau

Machine Translation

ACKNOWLEDGEMENTS

The content of this book on machine translation owes much to the experience acquired by the authors as members of the group for research in automatic translation at the University of Montreal (TAUM). We would therefore like to express our gratitude to the Department of the Secretary of State of Canada, which financed R & D at TAUM from 1973 to 1981, and to all our colleagues who took part in that research during those years.

The first sketch of the present volume began in December 1981 and the final corrections were finished in December 1986. Most of the work was financed by the Translation Bureau of the Secretary of State. Our thanks go to Alain Landry, Assistant Under Secretary of State, for his support during this long period, and to Gregory Gauld and Fernand Gobeil in the Planning, Management and Technology Branch for putting the project into operation and making the necessary financial resources available.

We are grateful to Françoise Dompierre and Lise Girard for their help in mobilizing human and technical resources involved in word processing for the numerous versions of our text, and to Lucie Legault, Louise Shpak, Monique Grenier, Francine Joanisse, Joanne Derouin and Diane David-Bergeron for their valuable assistance.

Finally, we wish to thank Pierre Isabelle, Marcel Paré and Gilles Stewart for their critical comments on our manuscript.

The opinions expressed and the scientific positions taken in this book are those of the authors; in particular, the authors do not speak in the name of the Canadian Government.

John Lehrberger
Laurent Bourbeau

PREFACE

Mechanical translation is perhaps the first attempt to apply computors to the simulation of a (nonnumerical) human activity. The amount of interest and support for this idea, which was developed in the 1950s has varied according to times and countries, but it has always been closely tied to political interests. The Cold War was the motivation for Russian to English translation in the early sixties; Canada had linguistic problems in the seventies; the Japanese language is a linguistic barrier to communication with America; and the European Economic Community has placed its different languages on the same footing for the communication of reports.

All these national or international patterns have caused a surge in the amounts of translation felt to be necessary by governments. In each cited case, mechanical translation has been seen as providing a solution, regardless of the state of advancement of the various scientific and technological domains involved.

Early research on machine translation suffered from a structural ambiguity. On the one hand, there were many basic problems that should have been studied:

- the construction of electronic dictionaries,
- the construction of electronic grammars.

It was then assumed, in many research centers, that the nonformalized dictionaries (monolingual and bilingual) and grammars available in bookstores and libraries were sufficient for computer applications, provided that they were transferred to some magnetic support in the proper format. A lot of superficial studies were then produced, mainly on the morphology of words. No serious effort was then brought to bear on the deeper linguistic aspects of the problems, and this aroused criticism from the community of theoreticians (e.g. Y. Bar-Hillel 1960: The Present Status of Automatic Translation of Languages, in F.L. Alt ed.: *Advances in Computers*, Vol. 1, New York: Academic Press, pp. 1-163).

From the viewpoint of computer technology, many fundamental problems were approached:
- construction of large memories (G. King's photoscopic disk), access to large data bases by hash-code like techniques (T. Ziehe at the Rand Corporation),
- a variety of models of natural language flourished, and parsing algorithms were developed for them.

On the other hand, the amount of support given to these research projects was motivated by the production of a final program which was to be evaluated on some economical basis. In 1966, the Peirce report (John R. Peirce ed. *Language and Machines*, Washington D.C.: National Academy of Sciences, National Research Council, publication 1416, 124p.) provided this evaluation of the field, which resulted in the ending of massive financial support in the United States, and in some other countries.

In the past five years, mechanical translation has once more raised the interest of potential users, mainly in Europe and Japan. As already mentioned, the wave of the 1960s covered a variety of research topics which were aimed at high-quality translation. As such, they involved many fundamental aspects of linguistics and computer science. Today, these questions are no longer seen as prerequisites, and on the contrary, the present movement is concerned with building cost effective systems that make no claim about quality, but that stress the increase of productivity (1) that organizations or individuals willing to use them would benefit from.

Whereas aspects of early experiments and of their failures seem to be remembered, the Canadian experiment is only rarely referred to. The Canadian Government supported the TAUM project at the University of Montreal consistently for about 8 years. A large amount of work on English and on French has been accomplished, both fundamental and practical, aimed at the translation of texts of a particular technical domain. When in

(1) Productivity appears to be due more to the improvement of text processing systems, including desk top printing, than to the linguistic tools.

1981 the project came to an end, the results obtained went through a remarkable process of evaluation, both from the Government and from private interests.

I think that there is a lot to learn from this experience for both ongoing and future projects, and I am particularly happy to preface this book by John Lehrberger and Laurent Bourbeau which goes systematically into the theoretical steps and the economics of the main approaches to machine translation.

Few specialists are in the position of having made substantial contributions to a project and of being able to follow it up to the end, through an assessment of its merits and deficiencies. Thus, the two authors present us with the first handbook of the field. They describe all the basic components of MT systems, and they review the main approaches from a user's point of view, not from the naive buyer's point of view who would only be interested in the return provided by his investment. They do this from the view point of specialists who will have to improve a system by extending both its vocabulary and grammar, and by customizing and maintaining them. Above all, the authors never forget the finality of MT systems: their ergonomy. This book should be read carefully.

Maurice Gross

TABLE OF CONTENTS

to supply information at the discretion of the machine. This is, in fact, the situation described in the next section (Human-Aided Machine Translation).

2.1.2 HUMAN-AIDED MACHINE TRANSLATION (HAMT)

In the case of HAMT the human translator supplies limited information to "fill out" the machine translation. After being supplied with the necessary data by the translator, the machine completes the translation, producing a raw output suitable for human revision. This can be accomplished in several ways. The required human assistance may take place before machine processing begins, during the translation process, or afterward. The machine may pause in mid-sentence to query the operator and then resume its processing of the remainder of the sentence, or it may make more than one pass through the whole sentence, with the operator inserting the appropriate information between passes.

The need for some human assistance arises primarily from the fact that certain linguistic structures have proven extremely difficult to parse automatically and words with multiple meanings add to the difficulty. Thus the machine may call on the translator:

- to decide on the scope of a conjunction (i.e., "what groups of words are connected by 'and', 'or', 'but'?");
- to bracket or translate a string of nouns in a sentence;
- to decide whether an occurrence of a preposition is part of a verb-particle combination, or whether it introduces a prepositional phrase modifying some noun in the sentence, or whether it introduces a prepositional phrase that functions as a sentence adverbial;
- to resolve homography problems;
  etc.

The boundary between HAMT and MAHT is difficult to draw. The designers of an interactive system may refer to it as "machine translation", but if the machine requires too much assistance, the translator may be effectively providing the translation. In that case, regardless of any claims made by the system's designers, it may be classed as MAHT rather than HAMT. Furthermore, phenomena such as those mentioned in the preceding paragraph are so prevalent

2.1.1 MACHINE-AIDED HUMAN TRANSLATION (MAHT)

MAHT is basically human translation with only limited assistance from the machine. At the lower end of the scale of what might be called "computerized translation" the machine may consist simply of a word processor with provision for looking up translation equivalents of source language words. This may be faster than writing out the translation by hand (or typing it with an ordinary typewriter) and thumbing through a dictionary for unfamiliar terms, but it does not remove from the translator the burden of actually performing the translation. Following are some features that may be included in an MAHT system.

(i) Word processor with provision for dictionary lookup (translation equivalents).

(ii) KWIC facility. The KWIC (Key Word In Context) can be used to show the contexts in which a word occurs in the texts under translation or in texts from the same domain. This helps the translator to understand how a word is used in that domain and may therefore help in the resolution of homographs.

(iii) Grammatical information. In addition to providing translation equivalents, the machine might also supply, for each word in its dictionary, grammatical categories (i.e., parts of speech), subcategories, and various syntactic and semantic properties of the word. The structure of the dictionary (or dictionaries) will be discussed in section 3.1; for the moment we simply note that in an MAHT system the availability of such information to the translator-operator does not imply that the machine itself uses the information to produce a translation of the text.

(iv) Morphological analysis.

(v) Corpus of translated texts. The translator can be provided with easy access to previously translated texts for reference in the current task.

(vi) Spelling and grammar correction.

We might think of MAHT as a system in which the human translator has control; the machine is simply a tool to be used at the discretion of the translator. On the other hand, computers can be made to seem quite human by sufficiently sophisticated programming. Thus we can have a man-machine system in which the computer has control while the human translator is used

2. IDENTIFICATION OF SYSTEM CHARACTERISTICS

The following classification of system characteristics is intended to provide a framework for discussing the features of particular systems and their capabilities. The parameters involved in this classification are: (1) the degree of automation of the translation process, (2) the depth of analysis of the sentences processed, (3) the type of transfer from source to target language, (4) the relation between phases in the translation process, and (5) the extent to which the system is limited to translation of texts from particular domains. The nature of the individual linguistic components of a system (lexical, morphological, syntactic and semantic) will be discussed in Chapter 3, and the relation between the approach used in building a system (corpus-based VS standard grammar) and its domain of application will be examined in Chapter 4.

2.1 DEGREE OF AUTOMATION

The degree of automation expresses the relative contribution of the machine and the human translator to the translation process. If a system is not fully automatic, there is some intervention by the human translator before obtaining the "raw output" (unrevised translation). In such an interactive system there are various ways in which the interaction can take place, resulting in different degrees of automation for the system as a whole. A rough idea of the degree of automation can be obtained by measuring the time spent by the human translator interacting with the machine to produce the raw output; this measurement forms part of a cost-effectiveness study.

But here we shall examine interactivity from the point of view of linguistic evaluation: which aspects of linguistic analysis are performed by the machine alone, and which require human intervention. This will help provide information needed to determine the limitations and improvability of the system.

CHAPTER 4 discusses two diametrically opposed approaches to designing a system: the corpus-based approach and the standard grammar approach. The advantages and disadvantages of each are explained. These two approaches have a direct effect on determining the content of the linguistic information present in the dictionaries and grammars of a system. Knowing which approach the system designer has chosen also gives us some idea of the extendability of the system to different domains.

CHAPTER 5 deals with the methodology for linguistic evaluation: identifying the needs and constraints of translation, evaluating the performance of the linguistic components of the system and evaluating the potential of a system. In addition, because of the importance of taking into account the man/machine relation in computerized translation, i.e. the effect on the human translators and revisers who must use the machine, the evaluation of the user environment is also discussed. The authors suggest steps to be followed in deciding on the acceptability of a system and then summarize the fundamental aspects and limitations of the proposed methodology for evaluating translation systems. They conclude with a discussion of the viability of MT, its future prospects and the impact of evaluation methodology on those prospects.

A preliminary study of evaluation methodology is contained in **Appendix A** (written in 1981) and a detailed flowchart of a typical second generation MT system in **Appendix B.**

editing). Many examples of English/French translation are used to illustrate the principles involved.

Any evaluation of an MT system is made up directly or indirectly of three parts: an evaluation of the quality of the translation produced by the system, an evaluation of the underlying linguistic model for the actual descriptions that constitute the system's dictionaries and grammars, and an evaluation of the computational model used to implement these grammars and dictionaries. For each of these three aspects the evaluation should determine not only the actual performance of the system with particular texts as input, but its **potential** as well. Of course, if we are aware of the potential of a system we are also in a position to understand its limitations.

This study provides a certain amount of technical information which serves to complement a strict cost/benefit evaluation: identification of the main characteristics of a system, classification on the basis of degree of automation, description of the various linguistic components, determination of the potential and limitations of a system, and insight into a too-often neglected area - the requirements and constraints of translation itself as well as the working environment of the translator.

In **CHAPTER 2** systems are classified along a number of dimensions: the degree of automation inherent in the system, the depth of linguistic analysis of the source language, the type of information transfer between source and target languages, the organization of processing phases in the translation chain, and the lexical and syntactic dependence of the system on the domain of application. This classification forms the basis for a multi-dimensional comparison between a system being considered for acquisition and others that are available. It also furnishes information basic to understanding the potential and limitations of a system.

CHAPTER 3 looks more closely at the characteristics of a system by giving an idea of its internal organization in terms of the major linguistic components: lexical (dictionary or dictionaries), morphological, syntactic and semantic. In order to understand the function and scope of these components, relevant linguistic phenomena are defined for each and illustrated with examples. Of course, the components are not isolated and independent of one another, but are interrelated. We must therefore take these relations into account as we examine each component to determine what it does, how it does it, and the nature and structure of its specific linguistic information.