American Journal of Computational Linguistics *Microfiche 46*

F B I S SEMINAR ON

M A C H I N E T R A N S L A T I O N

Edited by

DAVID G. HAYS AND J. MATHIAS

Department of Linguistics
State University of New York
Buffalo 14261

M R M Incorporated
9811 Connecticut Avenue
Kensington, Maryland 20795

Summary proceedings of a Seminar held at Rosslyn, Virginia, on 8-9 March 1976, organized by MRM Inc for the U. S. Government Foreign Broadcast Information Service. Attendance was limited to about 100 persons. Mathias co-ordinated the Seminar; Hays and Richard See presided.

SUMMARY

More than a decade after the ALPAC report, an agency of the U.S. Government called for a review of machine (aided) translation: What operations are in regular use, and with what success? What developments are coming? What research has been completed in the decade, is in progress now, should be stimulated? The Seminar fell far short of such a vast objective.

But it brought in several kinds of persons, whose expertise or established position in the field made their opinions important. For certain expositions, the organizers of the Seminar sought the best they could find; for others, quality was to be determined by hearing the presentation, not by prior judgment. A promise to be in another place on the same day prevented a few from joining us; unwillingness to speak before an open audience stopped one or two others.

In general, the spirit that we found in the field was excellent. Our colleagues made the effort to prepare their expositions and bring them to Washington; the audience listened attentively. The Seminar was more successful than this terse report can show. Successful, that is to say, as an act of communication. Future publication of longer reports, as contributors write them and the Editorial Board of AJCL accepts them, will communicate more. Future support of research and of MT installations will show whether the Seminar succeeded as an act of persuasion. -- David G. Hays

TABLE OF CONTENTS

*Presented at the Seminar in summary form, because the principal designer was not available.

**No summary available.

KEYNOTE ADDRESS

John Yeo, FBIS

On behalf of FBIS, I welcome you to this two-day seminar on machine translation. I would like to point out first of all that there is no political or social significance to the name tags we are wearing as far as color goes. The most offensive color we picked for the FBIS participants. Most of you wearing white tags are representatives of the United States Government and other agencies who are interested in the subject of machine translation or who have responsibility for translation problems. There are some exceptions, however, such as a few people from private industry and a few people from the academic community who are not on our speaker list, but nevertheless are interested in the subject, and whom we are happy to have here today.

The original concept of this conference was to have a relatively small round table consisting of myself, a few aides from FBIS, and several people from our speaker list. Thanks to Jim Mathias, our conference coordinator, we have a much more expansive conference. It now includes most of the institutions in government who are facing translation problems, and particularly those with an interest in discovering what has happened in recent years to move us forward in the area of machine aids to translation. Obviously because of this more expansive participation, we will end up with a thorough airing of problems of man machine, and translations.

I should point out that as far as the Foreign Broadcast Information Service is concerned, we are rapidly approaching the translation of 100 million words a

year, that our need at the present time is to keep abreast of the latest developments which can assist us because we feel the 100 million mark will be only a bench mark and that the demand on us for translation services will continue to grow. At the present time, all of our translation is done by humans, some in-house and a good bit of it by independent contractors. We find a good deal of customer satisfaction with our product despite occasional criticism from the academic community on the quality of translations. There are a minimum of complaints; however, we are not complacent because of this, and feel that it is necessary to be aware of aids that could be incorporated to help with problems now and as the load grows heavier.

We hope to reassess the state of the art during this conference and to find out what there is in it that we ought to be thinking about. We wish to turn out the very best translation at the very least cost. Many of the guests from other govern ment agencies face problems similar to ours. They are also being besieged for more and more translation. I recently sat with a government agency dealing with a new U. S. Joint Commission for Foreign Countries whose first act was to talk about an exchange of information, the result of which is a flow of innumerable documents into Washington. We understand that one agency in Washington has six file cases filled with foreign documents. They have no capacity to translate them. It is this sort of problem the conference will point toward and we hope those of yo with translation responsibilities will carry away new insights into the problem and its possible solutions.

We would ask that any who prepare assessments of this conference for your own agency kindly make a copy of that report available to us. It can be sent to me at FBIS, P. O. Box 2604 Washington, D. C. 20013.

I would point out that our conference will be tape recorded. This is to provide a record of the conference. To accommodate the recording, I would like to ask that those of you who ask questions please precede them with your name. I would also like to point out that on your agenda is a note that the evening session will include demonstrations by commercial representatives to this conference. Any and all of you are invited to come back at seven o'clock and stay as late as you care, to watch the demonstrations and to talk with the commercial representatives.

FOUNDATIONS OF MACHINE TRANSLATION

LINGUISTICS

Wallace L. Chafe

There is presently a theoretical opening in linguistics. Computers have been unfashionable; the party line has been against them, except in phonetics. Linguistics has suffered a real lag in manipulating large amounts of data. Linguists consider MT an impossible dream: The dreamer does not know what kind of thing a language is.

Devices for machine-aided translation do not define a basic area for the linguist; the real interest is in simulating the processes of a human translator.

Framework for MT: Surface structure (what is directly represented) vs. deep structure: ambiguity, idioms. Translation via conceptual representation, which may or may not be the same in all languages. Nature of the conceptual representation is the basic question for many fields. Two views: Logical net, easy to compute, a great discovery if correct; analogic form, not easy to compute, a mental image.

At what point does one make the image-language conversion? Different plans in different languages: in Southeast Asia, the image is more spatial than temporal. Japanese does not open a discourse with a summary of what is to follow.

Years of hard work and creative insight are needed for MT.

Real MT takes such deep knowledge it is utopian.

Intermediate goals: Stepwise simulation. (Notes by DGH)

WALLACE L. CHAFE

Professor of Linguistics

University of California

Berkeley 94720

Chafe was born in Cambridge, Massachusetts, in 1927. He did his graduate work at Yale, majoring in German. He was then employed for four years by the Department of State, principally at the American Embassy in Bern, Switzerland. In 1954 he returned to Yale to do graduate work in linguistics, and received the Ph.D. in 1958. He taught for one year at the University of Buffalo, and was then employed for three years as a specialist in American Indian languages in the Bureau of American Ethnology of the Smithsonian Institution. He joined the faculty of the Department of Linguistics at Berkeley in 1962. From 1969 to 1974 he was chairman of that department.

Chafe's principal research has been in American Indian languages and semantics, and most recently in the cognitive aspects of language use. His publications on linguistic theory include various articles and the book Meaning and the Structure of Language. He is presently the director of a project sponsored by the National Institute of Mental Health to investigate various processes involved in the verbalization of recalled experience. From 1972 to 1974 he directed a project funded by the U.S. Air Force dealing with the semantic prerequisites to machine translation.

OPERATIONS

Martin Kay

Compare translation with transportation: Hannibal could not have conquered Rome if he had waited for development of jet aircraft. Do what you can do; MT is the one thing we cannot do with present knowledge.

Consider a system, one of 100 that might be built. We have a problem, can do something about it; but choose only what we know can be done.

First, a display, keyboard, and pointer.

Next, an editor (program) and dictionary lookup

Then, morphological analysis, which is linguistically easy.

A program to take an advance look and offer a list of interesting words to the translator before the text begins to flow would be possible.

Call the translator's attention to specific difficulties. Avoid cascades of decisions, all following an initial error. System allows translator to develop a history. (This method is dangerous for pure MT--ultimate error is irreversible.) But this is the only way to make the posteditor's job easier than the translator's.

Small details of a man-machine system determine its actual usability. (Notes by DGH)

Erhard O. Lippmann*

Experimental On-line Computer Aids for the Human Translator

Experimental computer aids for the human translator are being developed which basically consist of storage, retrieval, editing and formatting operations carried out on line with a computer by an experienced human translator during the time in which a translation is produced. The system is not programmed to simulate the human translator by producing automatic translations. Rather, the user can call upon the computer's resources as needed in the translation process to shorten the delay between the initiation of a translation and production of a finished version. A combination of display terminals, computer hardware, and software is used to perform functions which have habitual human counterparts of a mechanical nature, e.g., dictionary look-up, dictionary updating, creating of text-related glossaries, editing and layout, collection of text statistics, combination, insertion, and deletion of text. An essential aspect of this system of computer aids is that, while assuming the burden of much of the mechanical drudgery associated with production of a translation, it leaves to the translator those tasks whose successful completion is most heavily dependent on characteristics that are uniquely human, in particular, the ability to produce grammatical output in which appropriate target translations have been selected on the basis of understanding of text content rather than through heuristics or brute force.

The system is designed to make it maximally simple for inexperienced computer users such as translators, terminologists, lexicographers, editors, and typists to work in an on-line environment. The translation aids are

*IBM Thomas J. Watson Research Center
Yorktown Heights, N. Y. 10598

implemented as modules which are compatible with existing text processing systems. As such they can either be integrated into such systems or isolated and put to other language processing uses with minimal modification.

The goal of the experimental computer-aided translation system is to streamline the entire translation production process from the reception of a source text to the printing of the finished version of the translation, thereby significantly increasing the productivity of the translator. In this connection, the user can perform the following tasks on line:

1) Enter and/or edit a text, e.g., a translation or a dictionary.
2) Look up dictionary entries and browse through dictionaries and other reference files.
3) Update dictionaries or other text files.
4) Print text in formatted or unformatted layout.
5) Obtain text-related glossaries in textual word order or alphabetically sorted.
6) Obtain statistical information and concordances on translations and/or (machine-readable) source langauge texts.
7) Delete, merge, and duplicate text files or text portions.
8) Permit other users to share texts and dictionaries on-line and/or off-line.
9) Obtain instructions on how to use the system.

Expected advantages include:

(a) increased productivity through accelerated dictionary and terminology lookup, rapid and convenient revision of successive translation drafts, and high-speed layout and printing of translations;

(b) easily activated production of text-related glossaries, which can be saved for future work;

(c) maintenance of consistency in terminology through immediate accessibility of standardized terminological digests;

(d) easily activated automatic insertion of previously-translated text portions and boiler-plate information;

(e) reduced handling and consumption of paper through emphasis on the use of visual displays rather than printed output during all but the final processing phase.

ERHARD O. LIPPMANN

Erhard O. Lippmann received the B. B. A. degree from the Free University of Berlin, Berlin, Germany, in 1956, and the M. A. degree in economics from the University of Michigan, Ann Arbor, in 1958.

After joining IBM World Trade Corporation in 1959, he was engaged in the conversion of manual business systems to automated data processing operations. At various times during his work in systems engineering, he was responsible for the translation of company product literature into the German language, and for the design, programming, and testing of software for automatic processing of textual material. Currently at IBM Thomas J. Watson Research Center, Yorktown Heights, N. Y., he is concentrating his efforts on the development of terminal-oriented programs specifically for non-numerical information processing. He has taught information processing at universities in the U. S. and Europe, most recently as a visiting professor at the University of Exeter, England, in 1972/1973. Since 1974, he has been serving as Chairman of the Committee on Computer-assisted Translation of the American Translators Association.

AUTOMATIC LANGUAGE PROCESSING PROJECT

BRIGHAM YOUNG UNIVERSITY

PROVO, UTAH 84602

Eldon G. Lytle

The Project emphasizes the refinement of computer-assisted translation, as opposed to fully automatic translation, and has devised for this purpose techniques of man-machine interaction which utilize the human for those aspects of the translation task requiring human intelligence and the computer for those aspects of the translation task which can be managed mechanically. Junction Grammar, a new theory of language structure which captures linguistic universals hitherto unknown, serves as the basis for the system.

Phase I of the development (now operational) provides computer editing, file management, and dictionary lookup. Phase II of the development provides computerized analysis, transfer, and synthesis of sentence structure (implementation 1978-79). Proto-type systems are designed for translation from English to Spanish, French, German, and Portuguese, but the method is equally adaptable to any combination of source and target languages.

The primary sponsor of BYU ALP is the Church of Jesus Christ of Latter-day Saints (Mormon), which annually translates approximately 17,000 pages of material into more than fifty (50) languages. It is planned that dictionary lookup and linguistic processing will initially be accomplished at a large central installation. The output of this processing will then be forwarded on "floppy" disks to regional translation centers around the world where residual aspects of the translation and printing task will be accomplished with the aid of mini-computer work stations.

The Project has a staff of 12 full-time and 18 part-time researchers.

ELDON G. LYTLE

Associate Professor of Linguistics

Brigham Young University Provo

Eldon Grey Lytle was born June 6, 1936, in Cedar City, Utah. He received his elementary and secondary schooling in the public schools of southern Nevada, graduating from Lincoln County High School in 1954 as valedictorian of his class. From 1954 to 1956 Mr. Lytle attended Brigham Young University at Provo, Utah. In 1956 he accepted a call to serve in Mexico as a missionary for the LDS (Mormon) Church. Upon returning from Mexico (1959), Mr. Lytle resumed his studies at BYU, specializing in Spanish. As a student he received tuition scholarships for academic excellence. In 1961 he graduated with high honors, receiving a B.A. in Spanish and a commission in the United States Air Force. Mr. Lytle completed requirements for the M.A. in Spanish (Russian minor) at BYU in 1962, prior to his tour of duty with the Air Force (1962-65).

From 1965 to 1968 he attended the University of Illinois at Urbana-Champaign as an NDFL Title VI fellow, and in 1968 he accepted a position with the Linguistics Department at BYU. In 1969 Mr. Lytle issued his first monograph on the theory of Junction Grammar and initiated a project in automatic language processing at BYU. In 1971 he received his Ph.D. degree in Slavic Linguistics from the University of Illinois. Between 1971 and 1976 Dr. Lytle initiated a series of courses in Junction Grammar at BYU and authored instructional materials for them. He currently divides his time between teaching, research, and the administration of the BYU Automated Language Processing Project.

CHINESE-ENGLISH MACHINE TRANSLATION

PROJECT ON LINGUISTIC ANALYSIS

UNIVERSITY OF CALIFORNIA

BERKELEY 94720

Research on machine translation from Chinese to English under the direction of William S-Y Wang was carried on at the project on Linguistic Analysis (University of California, Berkeley) during the period 1967 to 1975. During the early part of the effort, System I was developed which includes: a) CHIDIC: A Chinese to English machine dictionary of about 80,000 entries (60 percent physics, 30 percent biochemistry, and 10 percent general), and b) Monolithic grammar of about 4,000 rules (context-3, phrase-structure rules). In 1973, two factors caused redesign of the approach toward the development of System II. One, the grammar had become so cumbersome and ad hoc that its effectiveness as well as its potential for improvement were curtailed. Second, the sponsor requested conversion of the system from CDC machines to IBM machines. In response to these factors, System II is designed along the lines of "structured programming" (i.e., it is built on self-contained program modules). It is also designed to be machine-independent, so that it can be implemented at different computer installations.

Efforts in research and development have been aimed at an operational system. We have experimented with numerous trial sentences as well as several "live" texts (from articles of 3,000 characters in length) and have accumulated machine texts of over 560,000 characters. System II is incomplete, lacking especially the machine-editing of output to conform to those morphological features absent in Chinese but required in English.

WILLIAM S-Y. WANG

Professor of Linguistics, University of California, Berkeley

Wang received his Ph.D in Linguistics at the University of Michigan in 1960, and was appointed Professor of Linguistics at the University of California (Berkeley) in 1967. He is interested in the structure and function of language, including the processes whereby one language is translated into another. Some of his work have been on system simulation of linguistic processes humans do easily, such as speech recognition and machine text analysis. He is the editor of a bilingual journal, Journal of Chinese Linguistics.

LEIBNITZ--A MULTILINGUAL SYSTEM

John Chandioux

Leibnitz is an international cooperation between computer translation centers interested in a multilingual system. Several european groups, the TAUM project from the Universite de Montreal and a Brazilian group are presently working on this project. Most parts of the system are being written in one of the three languages made available by the CETA in Grenoble. The first one is the ATEF language, a string tree transducer for dictiona yrlook-up and morphological analysis. The second one is CETA and is a tree manipulating language for both transfer and generation. The last one is a tree/string transducer to be completed sometime in summer of 76.

Each group is either working on the design of an analyzer or generator for a specific language or on the transportability of the available formalisms. Research is presently under way on French, German, English, Italian, Portugese and Russian. English analysis is done by the TAUM team which is presently experimenting with a parser written in REZO its own version of Wood's Augmented Transition Networks. All particpating groups have agreed on a normalized tree representation for the output of analyzers and input of generators in order to minimize problems in the design of transfer components. The first part of the system is expected to be operational within two years.

CHINESE-ENGLISH TRANSLATION ASSISTANCE GROUP

J. Mathias

The U. S. based, intergovernment/academic CETA (Chinese-English Translation Assistance) Group is building a machine-readable dictionary file for use in on-line retrieval and for development of dictionaries and indexes for use of human translators. The experimental on-line retrieval system can store an unlimited number of entries. The current file of 640,000 machine-readable entries is divided into approximately 110,000 general entries; 10,000 colloquial entries; and 500,000 scientific and technical Chinese-English entries. The experimental system designed for an IBM 360 illustrates the facility of computer storage, retrieval, and display of Chinese characters and Roman alphabet as well as other scripts. It also illustrates the facility of computer techniques for indexing Chinese characters and special adaptability for synthesizing Chinese queries to search telecode-sorted files.

METEO, an operational system for the translation of public weather forecasts

Introduction

The TAUM project, from the University of Montreal, has been engaged for more than six years in the development of experimental models for the fully automatic translation of general texts from English into French. The first of these models has become known as TAUM 7I [I] and the latest one will be presented at the COLING 76 Congress. Because of the huge amount of data which needs to be compiled in order to make such a system, not to mention the introduction of a truly semantic component, no such system will be available for years to come. It is however possible to consider immediate limited applications for computer translation. METEO is an example derived from the TAUM 73 model. TAUM has also tried to demonstrate that computer translation could be successfully applied to the translation of technical manuals in a two-month experiment with the Canadian Translation Bureau and will concentrate in the next two years on the design of more appropriate parsing techniques and procedures for the treatment of idiomatic expressions.

General description

METEO is a fully automatic system for the translation of public weather forecasts from English into French covering the whole of Canada. It has been operating on an experimental basis since last December and due to be fully operational on the I5th of May I976.

Public forecasts for Canada are prepared in several regional centers from data sent by measuring stations throughout the country and centralized on a computer via the CN/CP communications network. Forecasts to be translated are retrieved, placed in a special file and processed one by one by the translation system. There is no human intervention prior to translation other than the actual typing of the text by a communicator at the corresponding regional office. The output of the translation system is handled by a specially designed editor (GERANT) which displays rejected sentences on a screen terminal at the local Translation Bureau. These sentences are taken care of by a human translator and as soon as a communication is completed it is redistributed to radio stations and newspapers using the same communications network as before. There is no

revision of the sentences accepted and translated by the system and to our knowledge this is the first time the product of a computer translation system will be distributed directly to the public. The sentences rejected by the sytem represent less than 20% of the total input, the main causes being: mispelled words, characters blurred in transmission, words not in dictionary, poor English, syntactic structures unknown to the parser, etc. The estimated load of the system is 30.000 words per day at the rate of over 1000 words per minute and the all-inclusive cost is about one third that of human translation.

The program

The program is divided in two main parts, the translation program and the editing program. The translation program is a succession of grammars of rewriting rules written in Q-System (2) Interpreters for this language are available in ALGOL, FORTRAN and COMPASS; the FORTRAN version was implemented on a CDC 7600 computer because it was judged to be the most transportable by both parties concerned. The editing program was written in FORTRAN for that particular application and also performs automatic preediting and formatting before and after translation.

The linguistic approach

The grammars are four in number:

The idiom dictionary
The main dictionary
The parser
The generator

The idiom dictionary

The idiom dictionary contains about 300 entries which can be divided into three types:

a) Several true idiom-like expressions such as:

clear period → eclaircie

b) A few strings of words which are not parsed for reasons of performance because they are compulsory elements of all communications:

"forecast issued by the atmospheric environment service"

c) A majority of place names which need to be translated or have an unpredictable translation:

Lake St Claire → lac Ste Claire

The main dictionary

the main dictionary contains all the lexical information necessary for parsing and generation and gives for each word the possible syntactical categories, for each category the possible translation or translations and for each category/translation pair the corresponding semantic features. Morphological variations of words also appear in the dictionary because there is only a very small number of them; in some cases the root form has even been omitted altogether, the infinitive of verbs for example. the present dictionary contains about I200 entries in all.

The parser

The originality of the METEO system lies mainly in the parsing techniques used. The world of weather forecasts is not unlike Winograd's blocks' world: lexicon, syntax and semantics are all restricted and make up a well-defined microworld. From the syntactic point of view, sentences are short and structurally simple, no relative clauses or passives for instance. The main problem is the delimitation of syntagms owing to the essentially telegraphic style of weather forecasts and the abundance of conjunctions. It was evident from the start that a conventional syntactic parser would be of little use because of the frequent omissions of function words and that it would be necessary to rely on some sort of semantic information. The ground work was laid out by. Richard Kittredge, Director of the TAUM project, in a preliminary study and a multiple-pass parser relying both on syntactical and semantic information was designed.

The aim of the parser is to give for each input string a single description giving the categories and translations realized in that particular string:

In a first pass substrings containing numerals are identified as dates, hours or temperatures.

In a second pass substrings expressing time or location are recognized as such. In the case of time, a distinction between durative and more punctual expressions is necessary because of the associated variations in French:

in the morning → dans la matinee

this morning → ce matin;

also, the distibution of determiners is often necessary when there is a conjunction:

this arternoon or evening →

cet apres-midi ou ce soir.

As far as locatives are concerned, the most difficult part is the identification of words not in dictionary as place names on the basis of context for place names which do not need to be translated were not entered in the dictionary because of their very high number.

In a third pass the remaining substrings are analyzed. The corresponding rules rely heavily on the semantic subcategorizations introduced in the dictionary to chose the proper translation for a given word:

heavy fog → brouillard généralisé

heavy rain → forte pluie

or to determine the scope of conjunctions:

snowflurries or rainshowers becoming intermittent tonight

(snowflurries or rainshowers) becoming ... .

For instance, it was necessary to divide weather conditions according to whether they were stationary, wind-like or precipitations in order to parse properly at this stage.

In a fourth pass the sequences of conditions, time references and locatives are tested for ambiguity and well-formedness and

each time a single structure can be built for a given input string, the parsing is retained and later processed by the generator.

In a fifth and final pass incomplete parsings are rejected and "styllistic" adjustments are made. An interesting example of this is the treatment of the word "occasional" which is entered in the dictionary as meaning "passager" because the predominant interpretation is the repetition in time, yet surely this is not the case for:

occasional cloudy periods -

*passages nuageux passagers

where one must assume that the meteorologist meant repetition in space, hence:

passages nuageux isolés

The generator

The task of the generator is to decompose the structure built by the parser, introducing articles where necessary, taking into account the word order of French:

gusty westerly winds →

vents d'ouest soufflant en rafales

and taking care of agreement.

Conclusion

The METEO system could not be used for the translation of texts other than meteorology because it is based on the semantics of that particular microworld but the strategy described here could certainly be adapted to limited fields where the amount of text to be translated largly compensates for the cost of designing a specific system. Neither do we wish to claim that our system is foolproof as demonstrated by a recent output:

apercu pour demain: faible possibilte

but then again, the communicator did type:

outlook for tomorrow: little chance.

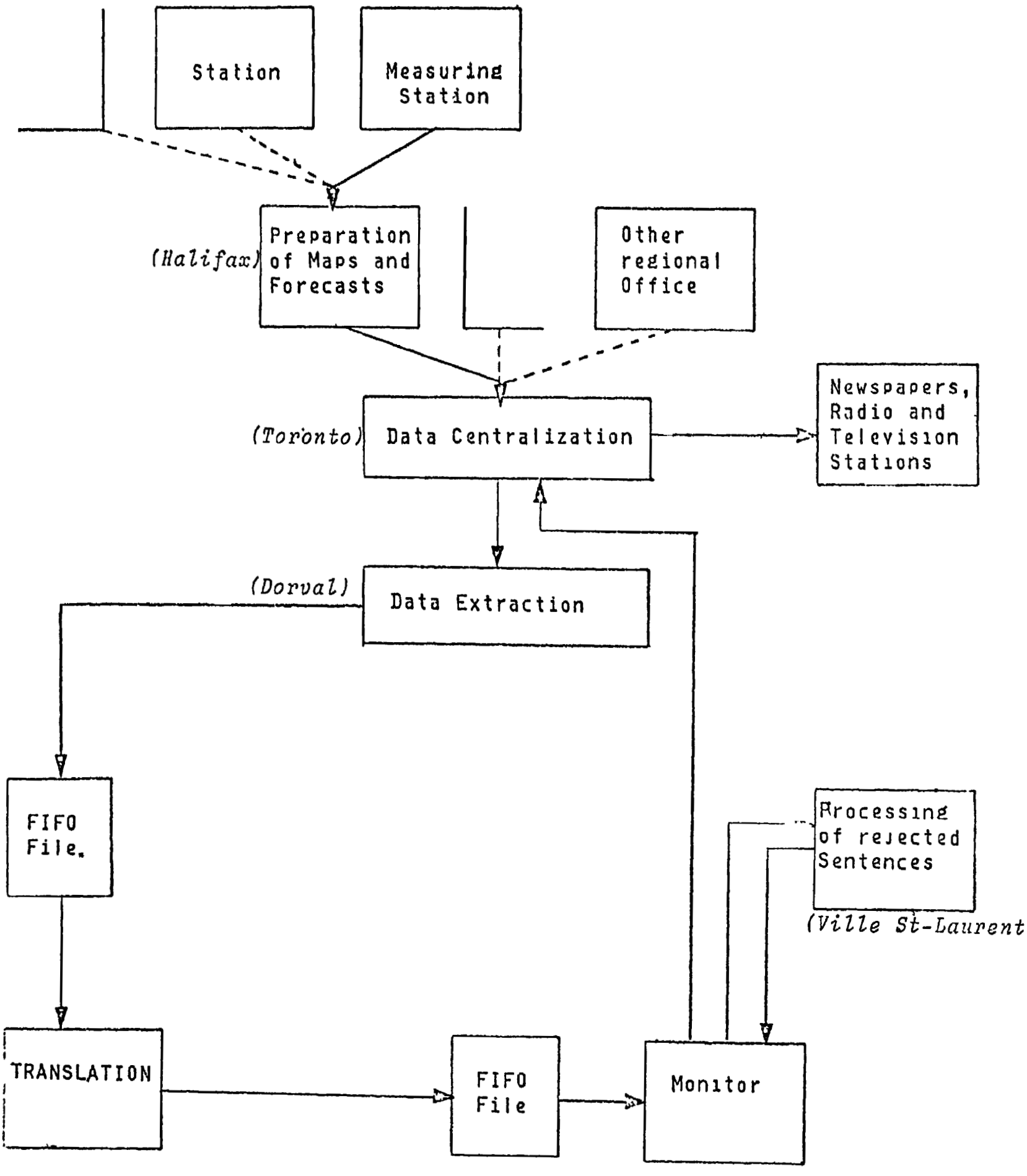
Nevertheless, we have found it to be a most entertaining project and our TAUM 76 model will benefit from it.

John CHANDIOUX
Head of the METEO team
TAUM project
University of Montreal

Chandioux received his Licence in English teaching in 1971 and in Linguistics in 1972, when he also received a Masters in English teaching; in 1973, he took a Masters in Linguistics. He is presently doing a Ph.D. in applied linguistics. Before joining TAUM he worked in France and Canada, teaching English as a second language and teaching contrastive linguistics.

(I) A. Colmerauer, Les Systemes-Q, université de montréal.

(2) TAUM 7I, université de Montréal.


Station
Measuring Station
(Halifax)
Preparation of Maps and Forecasts
Other regional Office
(Toronto)
Data Centralization
Newspapers, Radio and Television Stations
(Dorval)
Data Extraction
FIFO File.
Processing of rejected Sentences
(Ville St-Laurent
TRANSLATION
FIFO File
Monitor

HIGH LEVEL

WOOD BUFFALO REGIONS

MOSTLY CLEAR AND COLD WITH PERIODS OF VERY LIGHT SNOW TODAY AND WEDNESDAY. HIGHS NEAR MINUS 10 BOTH DAYS. LOWS TONIGHT MINUS 20 TO MINUS 22.

HIGH LEVEL

WOOD BUFFALO

AUJOURD HUI ET MERCREDI GENERALEMENT CLAIR ET FROID AVEC TRES FAIBLES CHUTES DE NEIGE PASSAGERES. MAXIMUM POUR LES DEUX JOURS ENVIRON MOINS 10. MINIMUM CE SOIR MOINS 20 A MOINS 22.

System Specifications

Languages Field Purpose	English to Fench Meteorology General Public
Pre-editing Post-editing Interactive editing	none none human translation of rejected sentences
GRAMMARS	Written in Q-Systems, a high-level programming language specifically designed for linguistic applications. Available in ALGOL, COMPASS, FORTRAN.
Dictionary	1200 rewriting rules loaded in central memory
Parser	300 rewriting rules bottom-up context sensitive
Generator (including Morphology)	300 rewriting rules
Maximum memory capacity required for loading and execution	60K words
Translation speed using the FORTRAN version on a CDC 7600 computer	1000 words per minute

Estimated operating cost (including human correction)	3.5 cents per word
Failures (including transmission errors, spelling mistakes and poor English)	Less than 20% of the input sentences.
Load	30,000 words per day
Operation	Has been operating 24 hours a day for 3 months on an experimental basis.
Delivery	May 76

EXCERPT FROM MR. BEDRICH CHALOUPKA PRESENTATION ON XONICS MT SYSTEM

The system known as the Xonics MT System was developed in the last six years from private sources. Those responsible for its development are Dr. Giuliano Gnugnoli, Dr. Allen Tucker, and Mr. Bedrich Chaloupka.

It is coded entirely in the PL1 programming language. It runs in a 100K memory region and may be executed on any IBM 360/370 computer in either a DOS or a OS environment.

The program may be executed in three different modes.

1. The Batch mode for translation of large volumes of text.
2. The sentence-by-sentence mode for translation of articles, abstracts, and titles.
3. The interactive mode, which allows translations and dictionary maintenance to be performed at a terminal. In this mode the dictionary update and the translation program may be executed simultaneously.

The system consist out of two programs.

1. The dictionary maintenance program.
2. The translation program.

The dictionary maintenance program allows the user to enter new items into the dictionary, to delete items from the dictionary, to change any field of items in the dictionary, and to enter semantic units.

The Dictionaries

The dictionaries are residing on direct access storage devices. The organization of the dictionary is indexed. This gives the possibility to open the dictionary files in either sequential or indexed sequential mode, depending on the mode of translation.

There are separate source and target dictionaries. Presently the source dictionary is 160 characters long. The target dictionary is 80 characters long. This organization is undergoing changes, so that a given dictionary may be used interchangeably as a source or a target dictionary.

The grammatical information in the dictionary is very rudimentary. There is no special skill or linguistic training required to work with the dictionaries.

The system is using both stem and full form dictionaries. The dictionary contains approximately 25,000 items in physics and chemistry.

The Translation Program

The translation program is small, consisting of approximately 650 PL1 statements. The translation algorithm simulates the mental processes of a human translator, and is not styled on any specific linguistic theories. The translation program is modularly designed.

In addition to proper recognition of grammatical properties the system eliminates case prepositions after conjunctions and punctuation marks, properly translates prepositions and semantic units, and rearranges participle and nested structures.

The system was designed for translation from Russian to English, but other languages with similar structures as Russian, such as Czech, may be translated. Even German sentences can be handled.

The system was demonstrated on a terminal. The demonstration consisted of translation of sentences in Russian, Czech and Serbian into English. The dictionary update, as well as semantic unit insertion and deletion, was demonstrated. The attached illustration shows some sample translations that were done by the system.

BEDRICH CHALOUPKA

Senior Analyst.
Xónics, Inc., McLean, Virginia

Mr. Chaloupka has worked on machine translation projects since 1956. He first became involved with machine translation at Georgetown University in 1956 where he worked on the GAT, SLC and Code Matching Techniques. He has done studies on major efforts in machine translation both in the United States and abroad. Mr. Chaloupka is the principle researcher in the development of the machine translation project at Xonics, Inc.

Mr. Chaloupka taught courses in computers and systems analysis. He is an accomplished systems analyst and computer programmer.

Mr. Chaloupka received his B.S. in Business Administration and M.S. in Political Science from the Charles University, Prague, Czechoslovakia. He received a B.S. in Languages and an M.S. in Linguistics from Georgetown University.

GIULIANO GNUGNOLI

Systems Consultant
Xonics, Inc., McLean, Virginia

Dr. Gnugnoli is currently Professor of Computer Science at Georgetown University and Systems Consultant to Xonics, Inc. He has over ten years experience in the development of computer translation systems. He is responsible for the design and implementation of the Xonics computer systems for machine translation.

Dr. Gnugnoli has developed and taught courses on the undergraduate and graduate level in data structures, PL/1, operating systems, file management and information processing. He is an expert in systems programming and computer communications.

Dr. Gnugnoli received his A.B. in Mathematics from Harvard University and Ph.D. in Mathematics from Georgetown University. He is author of the book "Simulation of Discrete Stochastic Systems", published by Science Research Associates, 1972.

SYSTRAN

The following presentation excerpts and paraphrases the highlights of the oral presentation given at the FBIS Seminar on Machine Translation, Monday, March 8, 1976, at Rosslyn, Virginia.

The major claim made for SYSTRAN is that it works — reliably, economically, and to the satisfaction of its users. It has continued to satisfy old and new users because it cannot become obsolete. It is in no way a black box. SYSTRAN has a very strong and flexible software framework enabling

1) immediate glossary expansion;

2) immediate implementation and testing of new or additional lexicographic, semantic and syntactic rules; and

3) universality in natural language translation.

The SYSTRAN system is "universal" in that it allows incorporation of additional translation capabilities (translation between new language pairs) without requiring modification of the existing software. Moreover, the addition of new translation capabilities requires only the implementation of additional source language analysis or target language synthesis programs. Everything else — all the parts that make the system work — remains the same. Thus, for example, since the system was already capable of translating from Russian to English, when the pilot Chinese to English capability was developed, only the development of a set of rules for analyzing Chinese as a source language was

necessary. Everything else, from the dictionary lookup and update programs to the English synthesis (generation) module, remained unchanged.

The SYSTRAN linguistic macro language is a great aid to the efficient development of these source language analysis and target language synthesis modules. These macros were developed to allow linguists to program their own rules. The formulation of the macros reflects types of operations (questions or tests, etc.) conceptualized by linguistic researchers as opposed to straight data processing-type programmers. The existence of these macros allows our linguists to modify existing programs quickly and with minimum effort and, of course, to write and check out new programs or even parsing or synthesis modules within relatively short periods of time.

The SYSTRAN translation system can run on either a 360 or a 370 with a minimum of 450K core storage available for application programs and dictionaries. Additional random access space is required for intermediate and sort work files. Input Russian text is accepted on 9-track tape or random access from either an ATS print file or MT/ST converted file. An alternative input file is accepted on punched cards which is normally used for system test. Output English translation can be printed on-line, via the SYSOUT printer, or offline, utilizing magnetic tape.

The system is programmed in direct assembler language and in SYSTRAN macros.

The computer processes batches of text at a rate of 300,000 words per CPU (Central Processing Unit) hour during an elapsed

time of 3 to 5 hours. Processor time per sentence is 1.2 seconds; for 1,000 words 18 seconds is average. Since the majority of refinements are additions of dictionary items and codes, rather than major additions to the programs, this speed will not lessen. It will increase, however, as the next generation of computers will further decrease cycles on the nanosecond level.

While SYSTRAN requires no human intervention in performing its translation tasks (other than the initial mounting of a disk pack or system tapes), it allows a maximum amount of interaction with its human components. First of all, because its linguists are its programmers, they know the system inside and out. On top of that, it produces hexadecimal displays with each sentence translated at the option of the user. Our linguists evaluate these records of the computer memory to identify translation problems and to identify precisely what program or routine is at fault. Having identified the problem, they then request SYSTRAN to produce concordance listings of a sufficient number of sentences containing exactly the same problems. After the linguist analyzes the resultant corpus, he designs, programs, implements and tests the necessary modifications. Modifications to the system do not always require such extensive research. Sometimes they are self-evident and require only a change in a single line of coding. The SYSTRAN macro instructions used by the linguist are automatically converted to assembler language during processing.

Since the Government has sponsored the refinement of this system, LATSEC, Inc. feels that any Government agency has the

right to have the system installed at minimal cost. (Expenses incurred when staff members train the user's staff to run the system should be covered.)

Maintenance costs, i.e., those costs involved in simply running the system to achieve raw output, can be directly calculated by any potential user by just finding out the per-hour cost of machine time at his installation. Any cost for improvement after installation depends on the user's requirements.

Our average keypunch or MT/ST input rate is about 1,500 words per hour. You can use this figure, along with how much your agency pays its keyers, to determine input costs. Of course, these costs would be virtually done away with if we could use optical character recognition devices. There is no pre-editing. Post-editing varies according to the user. Costs will vary according to the type of post-editing desired. According to FTD representatives, they are increasingly favoring the use of either un-edited, raw output or minimally edited output. (At a Bidder's Conference last September, Mr. Robert Wallace, the FTD SYSTRAN system monitor, said that nearly half of the 15 million words of text translated were distributed without post-editing.) NASA routinely used raw output of translations of working papers for the Apollo-Soyuz project. Yet, even when post-editing was performed, NASA found it both cheaper and faster to use machine translation rather than human translation.

At present, the system translates from Russian to English, from English to Russian, from English to French, and it has

lesser abilities in German to English and Chinese to English. Each capability is achieved by source language analysis and target language generation modules which fit interchangeably in the basic SYSTRAN frame.

As a final note. SYSTRAN works; it has proved itself useful as an operational system for the past six years. At this point, we are not interested in theoretical models of syntax; we are interested in making SYSTRAN the best possible machine translation system. It incorporates many aspects of modern linguistic thought. In doing so, it has transformed hypotheses about language into actual rules or descriptions of the behavior of language.

— END —

PETER TOMA

President and Chairman of the Board

Latsec, Inc.

La Jolla, California

Dr. Toma studied at the Universities of Budapest, Basel, Geneva, and Bonn, and at the Graduate Institute of International Studies in Geneva. He holds a Ph.D. in Communications Sciences, Slavistics, and Computer Sciences. He first developed machine translation algorithms in 1956 and joined the Georgetown (GAT) project in early 1956. As head of programming, he demonstrated that system at the Pentagon 6 June 1959. It was this system which was eventually converted for use at Oakridge and Euratom. (See The Serna System, Peter Toma, Georgetown Press, 1959.)

As a guest lecturer, Dr. Toma taught about machine translation at the Universities of Frankfurt, Bonn, and Cologne, the Institute of Technology in Darmstadt, and at the European Atomic Energy Commission (EURATOM) in 1960 and 1961.

In order to achieve, as early as possible, an operational system which would prove economical and reliable for the Government, Dr. Toma spent several years working in a private environment. The results were, first, Autotran and then Technotran.

In 1964, while the ALPAC hearings were in progress, Dr. Toma, working abroad, had a new system on the drawing board: a fully automatic, universal machine translation system. This system was SYSTRAN. Unde contract with the German Science Foundation, he implemented the system. Later, in July 1967, Air Force sponsorship supported further SYSTRAN development. In 1968, LATSEC, Inc. was formed. LATSEC, Inc.'s staff expanded SYSTRAN's translation capabilities to include English-to-Russian, English-to-French, German-to-English, and Chinese-to-English. In 1973, the formation of World Translation Center, Inc. furthered the development of the English-to-French system which has received significant recognition from the Canadian government. It was recently installed for the Commission of the European Communities and will be the first machine translation system to be used by the Common Market.

CULT

Chinese University Language Translator

Research into machine translation at the Chinese University takes a different approach than the others in that the Chinese University of Hong Kong places a heavy emphasis on pre-editing the source text instead of post-editing the target text. It is the only group taking this approach of computer-pre-editor partnership. All the other groups, who realized the FAHQT is not really attainable in the near future, have adopted a tendency to compromise in finding some computer-post-editor partnership.

A fixed set of pre-editing rules must be formulated to enable inexperienced and even mono-lingual people to transform quickly the input into machine-translatable form. With this arrangement, post-editing can be kept to a minimum, if not all together eliminated. Given time and better programming techniques, these pre-editing rules will gradually be reduced so that the computer will eventually take up this routine work. Pre-editing can therefore solve many of the present linguistic problems that are otherwise dependent on further research in natural language, computational linguistics, and transformation mathematics. In the present stage of development, very complex sentences can be translated with the aid of pre-editing. *

CULT (Chinese University Language Translator) was developed based on the principle mentioned above and has been rigorously examined and tested. Since the beginning of 1975, the CULT System has been used on a regular basis to translate two Chinese scientific journals, ACTA Mathematica Sinica

*An average of 5% of text is pre-edited by computer or editor.

and ACTA Physica Sinica, which are published by the Peking Academy of Science. This accomplishment indicates the correctness of our approach and the potential capability of CULT.

THE NEW LANGUAGE TRANSLATOR

Initially, CULT (Chinese University Language Translator) was designed as a special natural language translator with Chinese as the source language and English as the target language. Of course, a separate language translator will be required if English is to be used as the source language and Chinese as the target language.

The present translator consists of four modules, namely: 1) Dictionary look-up procedures employing the largest matching principle, 2) Syntatic analyzer, 3) Semantic analyzer, and 4) Output procedures including re-arrangement of word-order for the output sentences.

1. The Dictionary Look-Up Module

The basic dictionary look-up algorithm employs the "largest match" principle, designed for Chinese input (i.e., five digits numbers) and can readily be used for other non-alphabetic language input. However, an additional procedure for languages with alphabets (i.e., English, Malay, etc.) may be required to convert the alphabetic characters into numerical form by forming a "hash" before performing the look-up.

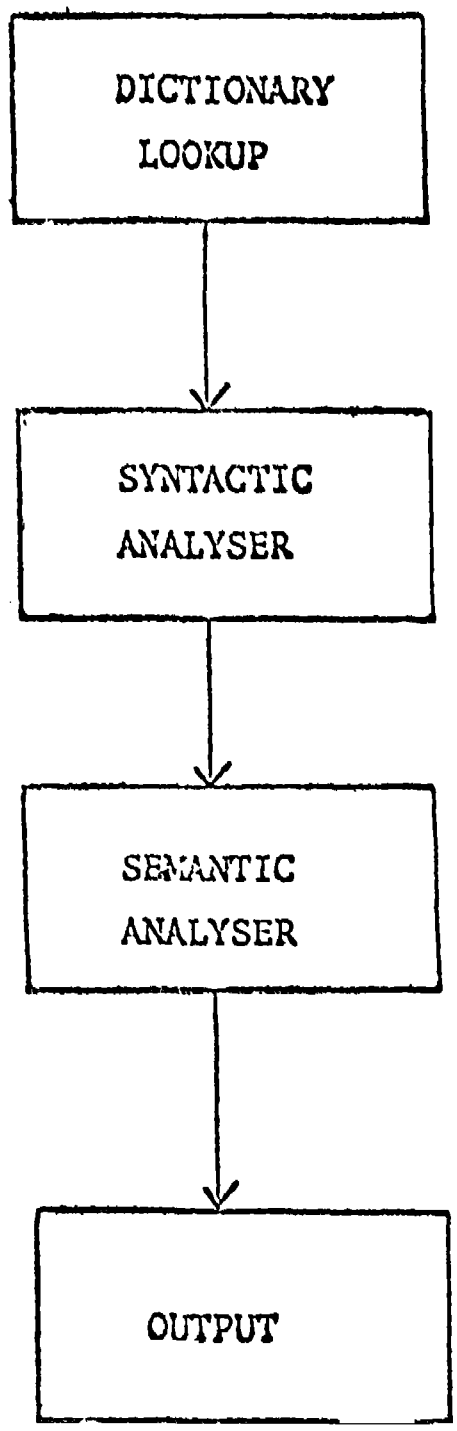


FIG. 1 TRANSLATION ALGORITHM

2. Syntactic Analyzer Module

The main function of the syntactic analyzer is to determine precisely the role that the individual words play in the sentence (i.e., to which parts of speech the words belong, whether noun, verb, etc.). The process is accomplished by means of a rather sophisticated true-false table.

While working on the machine translation of the Chinese scientific journals, a number of interesting linguistic difficulties experienced have been identified and defined. Previously, such structures would have to be pre-edited or post-edited in order to obtain the correct translation, but now they can be readily translated without any pre-editing.

3. Semantic Analyzer Module

At present, the semantic analyzer is able to offer only limited facilities, and the problem of semantic ambiguities is essentially resolved by: 1) a dictionary with specialized subject matter and 2) by pre-editing.

4. Output Module

The function of the output module is simply to rearrange word-order of the output sentence structure appropriate to the target language.

CONCLUSIONS

The successful translation of Chinese scientific journals, as well as non-scientific articles, by means of CULT has amply demonstrated the capability and the potential usefulness of the machine translation system in overcoming language barriers.

Though a number of linguistic problems are still to be defined and solved, the present machine translation system developed so far, if used with care and understanding, may contribute in some small measure in easing the desperate translation needs facing us today.

Automatic translation cannot be perfect. Whether it could even be high quality or not is dependent on how high the standards are set. The immediate goal is not to design a perfect automatic translation system or to achieve high quality machine translation, but to design a machine translation system that is better and more efficient than the ones we have today.

SHIU-CHANG LOH
Professor of Computer Science
Director, Machine Translation Project
United College Chinese University
Shatin, New Territories Hong Kong

RUSSIAN-ENGLISH SYSTEM

GEORGETOWN UNIVERSITY

Michael Zarechnak

The Georgetown University Russian-English System is running on IBM 360/70 .CPU time for 2000 words @ 9 seconds. The texts translated include scientific,technological,and economic materials.
M.Zarechnak in close cooperation with the linguistic research staff.
The linguistic statements are coded in symbolic language designed by Dr. A.Brown ('SLC'-Programming Language). Input/output is in Assembler language.
A dictionary entry contains a split or unsplit Russian stem, grammatical coding, lexical number, and English part. The clustered entries are recognized through special local operations when the calling signals occur within the sentence under processing.
Syntactic analysis is partly based on morphosyntactic markings and partly on semantic coding.
Users: Primarily scientists at ORNL. Users' comments essentially favorable.
The unddited translation is used primarily for information purposes, although in a few instances, the translations were post-edited when the user requested it.
The quality of the present translation is the same as it was in 1964.
No linguistic improvements were inserted in the system although there are some linguistic programs ready to be inserted.
The semantic level will be added. Its underlying procedured are based on the semantic collocational and colligational distributional patterns as observed in the real corpora, with such generalization as these corpora would suggest. It is hoped that after large corpora will be described both semantically and analytically, then some theories might be developed and tested deductively for the improvement of the next MT cycle. Each sentence.is scanned from the left to the right, and from right to left at least forty times, following a path of certain priority-based strategies. All these scannings in both directions are grouped into four levels: word recognition, syntagmatic,syntactic, and synthesis of English. Some parts of the synthesis are independent of the Russian input.
Size of the dictionary: 50,000 stems.

MICHAEL ZARECHNAK

Associate Professor of Linguistics

School of Languages and Linguistics

Georgetown University

Born November 18,1920, Czechoslovakia.
Education: PhD Harvard University in the field of Slavic Languages and Literature.
Experience related to the seminar: Teaching Russian to American students on introductory,intermediate,and advanced levels of proficiency.
Doctoral thesis:"Application of A.A.Kholodovich's theory of subclasses to Russian Temporal Nouns" (1967)
1956-64: While teaching at GU, participated in the Machine Translation Project at GU and had a significant role in the development of General Analysis Techniques (GAT), a system for computer translation from Russian to English in various scientific fields.
1964-66: Conducted research at Computer Concepts,Inc., in Silver Springs,Md.,in the field of automatic analysis of Russian, English, and German semantics, and automatic abstracting and indexing.
1966-67: Worked as programming specialist in Oak Ridge,Tenn., at the Union Carbide Computer Technology Center. Programmed in Cobol,Pl/I on the IBM 7090, and IBM 360 computers. Also conducted research on Russian to English Machine Translation GAT-SLC field tested at Oak Ridge jointly by CTC andORNL.
1967-68: Worked as research associate at GU on the GU MT Project in coordination with the MT Project of the University of Texas, under the general direction of Prof. W. Lehmann.
1968-Present: Have taught various computational courses at GU and a course on the theory of translation and its application. Worked also as a consultant for Union Carbide at Oak Ridge, updating the existing MT dictionary and doing semantic and syntactic research for Russian-English MT system used at ORNL by the scientists.
Published articles in the field of MT.

REGULAR USE OF MACHINE TRANSLATION OF RUSSIAN AT OAK RIDGE NATIONAL LABORATORY

Martha W. Gerrard

Oak Ridge National Laboratory*

Oak Ridge, Tenn. 37830

Abstract

User reaction has been favorable to routine computer translation of Russian scientific articles at Oak Ridge National Laboratory. Speed is the chief advantage of the machine translation, illegible copy being one of the greatest problems. Costs are comparable with those of human translation. Training of key punchers is not difficult. The machine dictionary is updated frequently, and fields other than the original chemistry of the dictionary are being included.

User Reaction

The ORNL program is aimed at machine translation, not machine-aided translation. We see no prospects of eliminating either pre- or postediting in the immediate future. Our earlier work was usually post-edited, and we have eliminated some of the necessary post-editing by judicious pre-editing.

A report, "User's Evaluation of Machine Translation," prepared by Bozena Dostert, was issued in August 1973. Dr. Dostert's study was based on 10 years of use of the Georgetown machine translation system at Oak Ridge National Laboratory and at the Euratom Research Center in Italy. The results of the study indicated that 92% of the persons responding to the questions judged machine translation to be "good" or "acceptable," i.e., to be generally informative and readable. Learning to read "machinese" seemed not to present any particular problems. Ninety-six percent of these users have recommended or would recommend machine translation to their colleagues. Eight-seven percent even expressed a preference for machine translation over human.

*Operated by Union Carbide Corporation, Nuclear Division, for the Energy Research and Development Administration

Since this evaluation was made, the program in use at ORNL, formerly operated on the IBM 7090 computer, has been converted for operation on the 360. During this conversion our regular use of machine translation was suspended, but recently we have started using it again on a regular basis. Again, we are finding a generally favorable reaction. When the reaction is unfavorable, we usually can elicit favorable comment after we explain the limitations of machine translation. For example, a user was rather disturbed because a Russian word (I forget what it is) was translated "descendents" instead of "progeny." When I explained to him that meanings are selected for the dictionary which are most generally applicable rather than using meanings that apply to a specific field and that we felt that "descendents" would be meaningful even if not specific for his field, he sent us another article to translate.

Before the conversion from the 7090 to the 360 some 75% of our research scientists' needs for Russian translation were met with computer output. We expect to be operating at that high level again very soon. The 25% that we do not translate on the computer, except for requests from a few die-hards, are articles that are too badly printed or copied to be readily legible and articles with a very high proportion of mathematical equations and symbols. We encourage our users to send us the best available copy. Even so, a key puncher may not be able to distinguish among the Russian letters и, и, и, ш, and п. A human translator, because of knowledge of the language, can usually decide, maybe with a hand lens, which letter is present. But the key punchers know no Russian and can only guess. Parenthetically, I might mention that not knowing Russian typography sometimes has its advantages. A key puncher does not tend, as I do, to type "sh" for ш, but uses the "w" as required by our system.

Articles with a high proportion of mathematical expressions and symbols are not too well suited for computer translation because the material that cannot be key punched must be inserted by hand on the finished copy, a time-consuming procedure. A recent development in our program has, however, facilitated such insertion. Formerly the key puncher typed in the word(s) "long equation" or "symbol" whenever something occurred that could not be key punched. Now, a means is provided for leaving a space so that the omitted material can be written or pasted in conveniently. With

this development we are less reluctant to use the machine for translating mathematical articles than we were formerly.

Advantages of Machine Translation

Of course, the greatest advantage of a machine translation is speed. A 10-page article can be translated on the computer in minutes. Key punching requires 3 to 4 hours, and pre- and/or post-editing maybe half an hour. Thus a requester could have the translation back the day after he asks for it.

Costs

A recent estimate indicates that costs of key punching, computer operation, and the small amount of pre- and/or post-editing that we do are comparable with those of human translation.

We do a minimum of pre-editing. I usually go through an article and mark, the first time it occurs, letters or words that are not to be translated. The key puncher then punches so that "Vitamin A" is translated as such rather than as "Vitamin and". Manual post-editing consists in indicating, at the first appearance, the meaning of a work that is not in the dictionary. The time for this is really chargeable to research rather than to routine translation because we then code such words and enter them in the dictionary. We have a program for post-editing which can be used to change the meaning of a word that is obviously wrong in the context; for example, we can instruct the machine to change the word "floor" as a translation of the Russian "pol" to "sex" in a biological article.

Training of Key-Punchers

Training of key punchers is not difficult, even though, naturally, some of the letters look alike, e.g., ь, ы, Ь, and Ь. However, I have been pleasantly surprised at how quickly the operators learn to distinguish these letters.

Updating

We have a program developed by Fred Hutton for updating the dictionary readily. Formerly we had to have around 2000 words to enter before the adding of new words became economic. Now we can add a few or large number whenever we have a list coded.

The original dictionary was primarily for chemistry, with some physics and nuclear energy terms. We are expanding the dictionary with nuclear energy terms, obviously because of our particular interest, and other energy terms and are adding biology and other fields. We are working on a means for indicating that a word has one meaning in the field of chemistry and another in biology, for example.

Future

We are able to take on a few customers from outside the Laboratory if arrangements can be made for transfer of funds. We are now charging $3.00 per hundred words for this service, which is about what it costs us. We are currently paying $3.00 per hundred for human translation. We request feedback from our customers on meaning of words or possible misinterpretations of grammatical construction and are using this feedback to improve our system.

Persons interested in our services are invited to call on us for more information.

Acknowledgment

Oak Ridge National Laboratory uses the Georgetown MT system. The system was brought to ORNL by Dr. François Kertesz and its use and further development were supervisied by him until his recent retirement. The project is now monitored by the Office of Language Services, a division of the ORNL library, and we have been fortunate in having as consultants Drs. Anthony Brown and Michael Zarechnak from the original project and Mr. Fred Hutton of the Computing Technology Center in Oak Ridge. The maturity of the program is indicated by its now being supervised by the library and used for routine translations.

Office of Language Services
Oak Ridge National Laboratory
P. O. Box X
Oak Ridge, TN. 37830

GEORGETOWN UNIVERSITY MT SYSTEM USAGE

NUCLEAR DIVISION, UNION CARBIDE CORP.

P. O. BOX X, OAK RIDGE, TENN. 37830

Fred C. Hutton

Ten years' experience in running the programs on the IBM 7090 is described. The present system, reprogrammed for the IBM 360, is described and capabilities of the system are set forth. An example of the use of the language invented by A. F. R. Brown (SLC for "Simulated Linguistic Computer") used in the preparation of the dictionary and linguistic routines, will be presented.

FRED C. HUTTON

I have worked as a computer programmer since 1957. Primary interest has been information storage and retrieval. I have been responsible for operation of the Georgetown University Russian-to-English Machine Translation almost from the day it arrived in Oak Ridge in 1964. With consultant A. F. R. Brown, programmer of the system as used on the IBM 7090, I participated in the reprogramming for the IBM 360.

Papers include:

Analysis and Automated Handling of Technical Information at the Nuclear Safety Information Center. American Documentation 18, 4 (October 1967). (Joel R. Buchanan, co-author)

PEEKABIT, Computer Offspring of Punched Card PEEKABOO for Natural Language Searching. Communications of the ACM 11, 9 (September 1968), 595-598.

RESPONSA--A Computer Search of a Subject Index. Proceedings of the American Society for Information Science (Vol. 5, 1968), 121-124.

Translation Aids

Federal Republic of Germany

Friedrich Krollmann

Germany's Federal Bureau Computer Translation Aids System contains over 700,000 foreign language (English, French, Russian, and Portuguese)-German entries of a technical and scientific nature. These entries can be accessed in a number of different ways depending on the needs of the user. Thus, the programming of the system allows for more specialized foreign language-German glossaries and lexical concordances, as well as linguistic analysis and frequency counts on the technical vocabulary of a given language.

Optical Character Recognition

based on phenomenal attributes

Robert J. Shillman

Research Laboratory of Electronics
MIT Cambridge, Massachusetts 02139

A theory of character recognition has been proposed and a methodology has been developed which is expected to yield a machine algorithm that will equal human performance in the recognition of isolated, unconstrained, handprinted characters. The methodology is based on the study of ambiguous characters, characters that can be assigned two letter labels with equal probability, rather than on letter ardhetypes. A description of the underlying representation of each of the 26 upper case letters of the English alphabet was obtained through analysis of ambiguous characters which were generated for this purpose. The descriptions are in terms of an abstract set of invariants, called functional attributes, and their modifiers. The relationship between the physical attributes, derived from physical measurements upon a character, and the functional attributes is given by a set of rules called Physical to Functional Rules. Three different techniques for determining these rules through psychophysical experimentation have been tested, and the particular rule for the attribute LEG has been determined. The remaining rules can be obtained in a similar fashion, and the combined results are expected to provide the basis for a machine algorithm. We are currently investigating the Physical to Functional Rules for the remaining attributes and are also interested in the way in which the rules are to be combined.

As a staff member of the Research Laboratory of Electronics at M.I.T., Dr. Shillman has been involved in research on visual physiology and the perceptual processes involved in vision. His doctoral dissertation, "Character Recognition Bases on Phenomenological Attributes" (M.I.T., 1974), proposes a new methodology for optical character recognition; the proposed technique is based on the incorporation of relevant psychological features into OCR algorithms.

Dr. Shillman has published numerous papers in the field of automatic character recognition and is a member of the IEEE, AAAS, Eta Kappa Nu, Tau Beta Pi, Phi Kappa Phi and Sigma Chi.

MACHINE PROCESSING OF CHINESE CHARACTERS

William Stallings
Center for Naval Analyses
Arlington, VA

Chinese Characters

Chinese characters, used to encode all the dialects spoken in China as well as the historically unrelated Japanese language, present a unique machine processing and optical character recognition (OCR) problem. Written Chinese is a pictorial and symbolic system which differs markedly from written Western language systems. Chinese characters are not alphabetic; they are of uniform dimension, generally square, and are composed of strokes, each one a line that can be drawn without lifting the pen. In these highly structured characters, many regularities of stroke configuration occur. Quite frequently, a character is simply a two-dimensional arrangement of two or more simpler characters. Nevertheless, because strokes and collections of strokes are combined in many different ways to produce thousands of different character patterns, the system is rich.

Written Chinese is very difficult to learn: there are over 40,000 characters, each corresponding roughly to a word in Western languages, of which an educated person would be expected to know about five to ten thousand. The meaning of each character and its fixed monosyllabic pronunciation must be learned by rote. Usually, these two tasks are eased somewhat because one component of a character gives a clue to its meaning and the rest gives a clue to its pronunciation. But since there is no alphabetic order to Chinese characters, another difficulty is dictionary lookup; a number of special systems have been devised to impose an ordering, none of them terribly convenient. Finally, a student of Chinese must learn to draw the strokes of each character in a particular order; a character may have from one to thirty strokes with eight to twelve being typical.

Of direct relevance to the use of OCR for Chinese is the desire of the Peoples' Republic of China to simplify the written language through a series of language reforms. The first is that the government has recommended the general use of only 2000 characters. Publishers, being government-controlled, are under instructions to stay within the total of 2000 as far as possible. Secondly, the government has simplified a large number of characters with the result that the average number of strokes per character has been reduced by about a factor of two to an average of about 6 to 8 strokes per character. This continuing policy of language simplification will ease the difficulty of Chinese OCR.

Data Processing Requirements

The requirements for machine-processing of Chinese characters, whether for machine translation or other applications, are four:

- input
- storage
- data processing
- output

The requirements in the latter three areas are formidable compared to those imposed by the Latin and Cyrillic alphabets. For example, a machine translation device might be required to print out the Chinese text together with its translation. An adequate representation of each character would require a 32X32 black/white matrix. Hence the storage of the image alone of each of 5000 to 10,000 characters would require 1000 bits. Nevertheless, because of the vast improvements made in memory density and processing speed of computers, these requirements no longer present a problem.

The only remaining bottleneck is input. Because of the many thousands of characters in common use, a keyboard for Chinese (for typesetting, typewriting, keypunching, on-line computer entry, etc.) is an ungainly affair. One common model has 192 keys with 13 shifts, another simply has 2300 keys! Among their disadvantages:

Slow speed - a rate of 40 characters/minute is typical of experienced operators, compared with 70-75 words per minute for an English-language typist. It should be remembered, though, that a Chinese character corresponds roughly to a word in English, so the discrepancy is not so great.

- High error rate - error rates on Chinese typewriters are much higher than Latin letter typewriters - as high as several percent. Considering the high information content of each character, this is a serious problem.

- Training requirement - efficient use of a Chinese keyboard requires a great deal of training and is almost unattainable by those who do not know the language well.

In recent years, a number of approaches to reducing the keyboard complexity, all of which exploit some structural characteristic or the stroke order of Chinese characters, have been taken [1]. It is safe to say that none of these devices has produced an improvement in any of the problem areas listed above.

OCR for Chinese

The only alternative to keyboard entry of Chinese characters is OCR. While there is much optimism about developing satisfactory OCR devices for Latin or Cyrillic letters, the prospect for Chinese OCR is dim. Three problems arise:

- Size - to be useful, a Chinese OCR device would need to be able to recognize 5000-10,000 characters. This is two orders of magnitude greater than the number of images a Latin OCR device would have to handle.

- Complexity - a Chinese character may have as many as 27 strokes. There is so much detail that it is difficult to develop a set of features for distinguishing among characters.

- Density - compounding the complexity problem is the density of printed Chinese characters. On a given document, all characters will occupy the same amount of space, from the simplest to the most complex. The result is that the space occupied by a character is, on average, 50% black. This causes the smudging and overlap of features and strokes, even for the highest-quality printing.

Not much progress has been made in solving these problems, although a number of attempts have been made [2]. The most promising attempt currently underway is at Hitachi Ltd. in Tokyo. A group there has reported an error rate of one in a thousand with a reject rate of one in a million for a set of 1000 characters under rather ideal experimental conditions. It remains to be seen if they can go further with their approach.

Incurring the usual risks associated with such predictions, one might set the following as reasonable goals for a Chinese OCR device given a vigorous short-range development project:

- error rate: 1.5%
- rejection rate: 0.5%
- cost: 5 times the cost of a practical Latin OCR device, whatever that might be.

That these goals can be attained is questionable. If they can, then the choice between OCR and keyboard input of Chinese will be based on a tradeoff between cost, speed, and accuracy.

REFERENCES

[1] Stallings, W., "The Morphology of Chinese Characters: A Survey of Models and Applications", Computers and the Humanities, 9, 1975.

[2] Stallings, W., "Approaches to Chinese Character Recognition" Pattern Recognition, 8, 1976.

WILLIAM STALLINGS

Analyst

Center for Naval Analyses

Arlington, Virginia

William Stallings received a BS desgree in Electrical Engineering from the University of Notre Dame in 1967 and MS and PhD degrees in Computer Science from MIT in 1968 and 1971. His doctoral thesis was on Chinese character recognition.

From 1971 to 1974, he was with the Advanced Systems and Technology Operation of Honeywell Information Systems, Inc., where he worked on the development of interactive computer systems and Chinese character input/output systems. Dr. Stallings is currently on the staff of the Naval Warfare Analysis Group of the Center for Naval Analyses in Arlington, Virginia, where his principal interests are discrete-event simulation, systems analysis, and decision theory.

PROGRAMS TO UNDERSTAND STORIES

Roger C. Schank

Research at Yale centers around the building of computer programs that will understand stories. Two program are currently being developed, SAM and PAM.

SAM is composed of the following
1) an analyzer that maps English into a deep conceptual representation.
2) a script applier that uses its knowledge of contexts to supply missing or or implicit inferences about a situation.
3) a memory that finds references for things that it knows about in a text so as to bring its knowledge to bear on the text.
4) a generator that reads information provided to it by (1), (2), and (3) and states that information in English, Chinese, Russian, Dutch or Spanish.
5) a question answerer that interacts with the script applier to answer questions about an input text.

SAM is capable of mechanical translation, automatic summary and paraphrase and question-answering about texts in domains that it has knowledge about.

PAM is like SAM except that it does not have a script applier but instead has a more general mechanism that to infer the goals and intentions of the actors in the stories it hears.

Both of these programs are beginning approaches to the problem of computer understanding.

ROGER C. SCHANK

Associate Professor of Computer Science and Psychology

Yale University *New Haven, Connecticut 06511*

Roger Schank has a B.S. in mathematics from Carnegie-Mellon University (1966) and a M.A. and Ph.D. in Linguistics from the University of'Texas at Austin (1969). His thesis was concerned with designing a language-free representation of meaning that could be used as the basis of machine translation. He was Assitant Professor of Linguistics & Computer Science at Stanford University from 1968-1973. His research at the Stanford Artificial Intelligence Project resulted in the MARGIE system that did sentence to sentence paraphrase and inference using a language-free base. He spent one year (1973-1974) at the Institute for Semantics & Cognition in Switzerland working on representations of text. He is currently at Yale University where he is director of the Yale Artificial Intelligence Project. His research is currently focussed on the use of knowledge about context and human planning to aid in the building of a computer understanding program.

CONTEXT/TOPIC SPECIFIC KNOWLEDGE

(No summary available)

CHARLES JOSEPH RIEGER, III

Assistant Professor of Computer Science

University of Maryland College Park 20742

EDUCATION:

B. S. Purdue University, 1970, Mathematics/Computer Science (dual major), summa cum laude with honors, Physics minor

Ph. D. Stanford University 1974, Artificial Intelligence (Computer Science)

CURRENT RESEARCH INTEREST:

Representating commonsense algorithmic world knowledge on the computer and using such knowledge 1) in understanding children's stories and 2) in problem solving

EXPERIENCE:

1967-68	Teaching Assistant, Computer Science Department, Purdue University
1972-73	Research Assistant, Computer Science Department, Stanford University
1974-	Assistant Professor, Computer Science Department, University of Maryland
Fall 1974	Invited Visiting Assistant Professor, M. I. T.

JOURNAL PUBLICATIONS AND BOOKS:

Conceptual Memory and Inference, in Conceptual Information Processing, R. Schank (ed.) (forthcoming)

Understanding by Conceptual Inference, American Journal of Computational Linguistics, Microfiche 13, 1974.

Everyman's LISP: The LIST Family of Computer Languages (How to use Them, Their Implementations) Book in progress, 1974

Inference and the Computer Understanding of Natural Language, Artificial Intelligence, v. 6, no. 1, Spring 1975 (coauthor R. Schank)

Semantics and World Knowledge in M T

Yorick Wilks

I presented very simple and straightforward paragraphs from recent newspapers to show that even the most congenial real texts require, for their translation, some notions of inference, knowledge, and what I call "preference rules", over and above those found in standard approaches to the problem of MT.

I argued that the MT problem has not been solved in any sense even though there have been real improvements in the performance of large commercial systems, yet, contrary to some impressions given at the seminar, we are by no means exactly where we were twenty years ago and about to go through the same agonizing cycle of optimism and disillusion again. That is because the lesson of the 'first MT cycle' has been appreciated within Artificial Intelligence (AI), or at least some parts of it, and solid attempts have been made to produce small-scale intelligent systems directed towards tackling the great problems thrown up, but not solved, by MT research: ambiguity, of word sense, case structure and pronoun reference.

I then presented a sketch of a small research English-French MT system that takes in paragraphs on-line and translates them via an interlingua of deep meaning structures and inference rules. This system was described in the course of a recent survey (AJCL Microfiche 40, 1976).

I argued finally that systems of this sort can play an important role in advancing MT, by occupying a space, as it were, between three better-known positions: (i) that we can just go on as before with "brute force" systems (ii) that we can only get advance by devoting ourselves here and now to purely theoretical AI systems that "represent all knowledge" and (iii) that we should make do with techniques that are simple but fully understood, such as on-line editors.

YORICK WILKS

Senior Visiting Fellow
Department of Artificial Intelligence
University of Edinburgh Scotland

My doctoral work was done at Cambridge, after which I worked with a small group who was attempting to apply semantic methods to the processing of natural language with the aim of Machine Translation. In 1967 at SDC (Sta. Monica, Calif.) I constructed a system in LISP that input paragraphs of text, converted them to deep semantic structures, from which the resolved ambiguities of the word sense were then read off (i.e., output was still in English). Later, while at Stanford University (artificial intelligence lab.) I constructed an on-line system that would input paragraphs of English and produce French translation, via a representation in a semantic interlingua that could be suitably massaged with inference rules representing "real world knowledge." On leaving Stanford in 1974, I went for a year to the Institute for Semantic and Cognitive Studies in Switzerland and then to the University of Edinburgh, where I have worked on theoretical defects in that Stanford model and ways of overcoming them in a later implementation.

Formal Representation

Robert F. Simmons

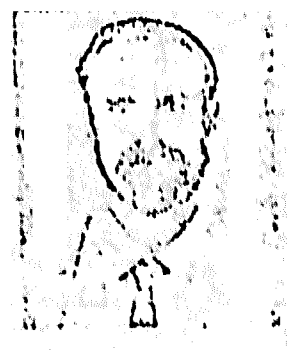
A developmental program is proposed to create a socially useful system that will integrate several existing natural language processing procedures into a robust, transportable, General Text Understanding System for eventual use in applied information centers. The proposal is comprised of seven tasks: 1. Continued development of quantified case predicate forms of conceptual memory structure. 2. Integration of question answering and problem solving procedures. 3. Development of a human-aided, multi-pass, text-to-memory compiler. 4. Generation of natural language outputs for summaries, abstracts, expansions, translations, etc. 5. Generation of special purpose text teaching materials. 6. Implementation of natural language dialogue capabilities. 7. Development of a textword management system for linguistic analysis, retrieval and lexicon development.

The work will be accomplished on a DEC10 to enhance the transportability and communication of documentation for the resulting system.

ROBERT F. SIMMONS

Professor of Computer Sciences and Psychology

University of Texas Austin 78712



Robert F. Simmons

Dr. Simmons received the B.S., M.S., and Ph.D. degrees in psychology from the University of Southern California, Los Angeles, in 1949, 1950, and 1954, respectively. He worked as a Managerial Staff Researcher at Douglas Aircraft from 1953 to 1955. At that time he joined the RAND Corporation as an Associate Social Scientist on the System Development Project, which later became the System Development Corporation. By 1960 he had instituted the Synthex Research Project, which led to the development of a number of early natural language retrieval and understanding systems, and eventually to the Natural Language Research Group which he headed for some years at SDC. From 1960 to 1968 he published numerous articles, the most influential of which were 1964 and 1970 surveys of question-answering systems and a series of papers in 1967 and 1968 that described Protosynthex III, a deductive question-answering system that parsed natural language sentences into a semantic structure and answered questions and generated paraphrases in English from that structure. In 1966 he and Harry Silberman began a project to study potential applications of language processing technology to computer-aided instruction. This has been his main line of effort since that time.

Since 1968 he has been a Professor of Computer Sciences at the University of Texas, Austin, and currently teaches computational linguistics and supervises eight graduate students in their research in this area.

He is a member of the American Psychological Association, the Association for Computing Machinery, the Association for the Advancement of Science, and is Past President of the Association for Computational Linguistics. He was Sector Editor for *Computing Reviews*.

Summary of FBIS Seminar on Machine-Aided Translation

S. R. Petrick
IBM T. J. Watson Research Center
Yorktown Heights, N. Y. 10598

In the first paper delivered at this conference Wallace Chafe presented the following model of translation: a source language sentence is first parsed to produce a surface structure. This is converted by some process of comprehension to a deeper, conceptual structure that reflects the meaning of the sentence in a more direct way. This conceptual structure may or may not be a language-independent universal structure. In those models where it is not universal but instead is tailored to the source language, it must be converted to a corresponding conceptual structure that is similarly specific to the target language. In any case, conceptual structures must be mapped by a verbalization process into corresponding target language surface structures whose debracketizations yield the required target sentence output.

Other speakers suggested extensions to this model, for example, to provide for context beyond isolated sentences. Basically, however, Chafe's model provides a good basis for discussing the translation efforts which were described by the other speakers at this conference. For example, one way in which different systems roughly based on Chafe's model can vary is in the relative depth of their conceptual structures. Actual systems that were discussed varied in this respect all the way from rather abstract structures that directly represented meaning to shallow structures whose relationship to corresponding sentence meanings was, at best, tenuous. All of the commercially intended MT systems which were described appeared to rely upon such shallow structures, in some cases on surface structure itself. In most cases this was explicitly stated by speakers at this conference, and in other cases it could be inferred from outright errors in exhibited sample output where intended meaning was not correctly determined. All of these MT systems, however, exhibited what might be

called extensive coverage of the source language, i.e., output was produced for every source language sentence (undoubtedly also for ungrammatical source language utterances).

In contrast to the commercially intended MT systems stand the Artificial Intelligence systems for natural language understanding, which in most cases have yet to be applied to MT. Their advocates point out the necessity for deeper conceptual structures as well as supplementary information and inference in order to adequately translate certain sentences. They pay a price, however, for their insistence on more adequate conceptual structures, because those structures are not easily obtained for unrestricted text input. It is equally true of the AI and Computational Linguistics systems, whether based on formal grammars or procedurally defined, that the coverage of the source language provided is currently very sparse. Due to the fact that most source language sentences are not processed by these systems, they are unsuitable for unrestricted text and have been applied only to question answering systems and to restricted toy-world domains. The amount of effort required to extend the coverage of, say English, to a state useful for MT while maintaining the adequacy of assigned conceptual structures might be variously estimated by different authorities, but it is my opinion that it is very large indeed, large enough to make such applications as question answering systems more attractive candidates for consideration in the next few years.

Another point to note in conjunction with all of the systems discussed at this conference is that their treatment of the process Chafe referred to as verbalization is rather primitive. Thus in spite of the fact that this aspect of a computational linguistic system is often referred to as uninteresting or trivial compared to the task of understanding an input utterance, and in spite of the fact that many normally difficult facets of verbalization do not present

a problem in MT, the current output of language processing systems is very unnatural and rough. This is true of AI systems as well as operational MT systems.

If, in fact, we examine the specific realizations of the components in Chafe's model which were reported to be included in the MT systems described at this conference, we find very few changes over the situation that prevailed ten years ago. The comprehension component is realized by such means as a context free grammar, a Q-System, or an analysis-based ad hoc procedural specification. Difficulties and shortcomings related to conceptual structures have already been noted. These have changed very little over the past few years. Similarly, we have already commented on that portion of the target language output inadequacy which is attributable to shortcomings in the treatment of verbalization. In summary then, currently operational or projected MT systems are only marginally different in their underlying organization and design than their predecessors.

If, then, there is little that is novel about the underlying models of current and projected MT systems, it is natural to ask how many hardware and software improvements have been made. Several claims were made about improvements in procedural programming languages. Although I am fully aware of the benefits which follow from the use of a well suited programming language, I don't think the improvements which are claimed are very significant. For one thing, many language processing tasks are still very difficult to program using the best programming languages. And for another, convenient programming is no substitute for the absence of satisfactory models and algorithms. Recent advances in editors and time sharing systems might, however, be significant factors in making the development of machine-aided human translation more attractive.

Hardware developments of the past decade include time sharing hardware, automatic photocomposition devices, larger primary and secondary storage, faster processing speeds, and lower costs. Optical character recognition was reported not to have advanced significantly in the past few years. There is still a limitation to a fixed set of fonts, and the only large scale applications at this point involve fonts carefully designed for OCR.

We have seen increases in computational power per unit cost and can expect to see more such increases. The question which arises, however, is what their effect is likely to be on MT. The key issue is how much of the total effort can be handled by a computer and how much must still be done by human labor. Text input, pre-editing, and postediting can take as much human time and effort as complete human translation.

Of critical importance is the evaluation of current MT systems to determine the quality of their unedited output, the uses for which such output is acceptable, and the amount of postediting that is required to meet well defined higher standards. No clear results of this type were provided at the conference and careful study is necessary to resolve certain seemingly contradictory claims. Thus, there were reports of translation output which was not postedited, other output which was only lightly postedited, and still other output that was extensively postedited. The implication was given that no more editing was required than was given, and, although there is a sense in which that claim is undoubtedly true, it fails to take into consideration the quality of the output, the purpose for which the translation was requested, and the degree of requestor satisfaction. Although I did not systematically examine large quantities of source language input and corresponding unedited target language output, the examples which I did examine

suggested a rather low level of performance with respect to both fidelity of meaning output and to smoothness and naturalness of the output. The overall quality of output produced strikes me as comparable to that of ten years ago, and a colleague of mine with more experience in MT than my own assessed the output I showed him as more ambitious in its attempt to achieve natural output than past systems but probably not any more successful. Attempts to produce natural target language word order and correct insertion of articles helped in some cases but just as often made the translation worse. Clearly, it is no simple task to evaluate the quality of output achievable through the use of a particular MT system, to determine the amount of post-editing necessary to bring it up to required standards of quality, and to estimate the likely cost of achieving that quality. Each prospective user of an MT system must carefully do this, but from what was presented at this conference I would not expect any current MT systems to compete economically with human translation except in those few cases where requirements for quality and accuracy are so low that unedited or very lightly edited output suffices.

In addition to postedited MT, this conference also discussed the use of hardware and software aids to human translation. There seemed to be a consensus that well-engineered systems can be produced now, that their use looks promising, and that they probably are limited to increasing the productivity of human translators by a factor of 2-1 or 3-1. Opinion was divided as to whether they might evolve into human-aided MT systems. It did appear clear that existing systems have not yet been carefully field tested, and that they do not contain all the aids to translation that have been suggested.

STANLEY R. PETRICK is President, 1976, of ACL. For biography, see AJCL Microfiche 37:3.

Summary Remarks for Machine Translation Conference

Sally Yeates Sedelow

An issue which emerged early in the conference and recurred either explicitly or, more often, implicitly during subsequent sessions concerned the relative values of pragmatic solutions and more basic research. An additional factor was the often presumed relationship between more basic research and science, and between the pragmatic and its synonym, 'ad hocness.'

I suspect we would all agree that there is no necessary progression from what some here have termed 'engineering' solutions to theory from which one can generalize. On the other hand, neither is there any necessary relationship between science (in a strict definition) and what is sometimes called basic research by linguists, computational linguists, psycholinguists or whoever among us is dealing with natural language. In my judgment, for any major leap forward in machine translation or in natural language understanding in general, more classical science is badly needed. Science is needed not only for its rigor, which implies well-articulated models and thorough and extensive predictive-type testing (including efforts to reproduce results in a number of 'laboratories'), but also for cumulativeness. In the situation under consideration, I am struck by the number of isolated hypotheses and experiments which don't seem to lead anywhere, and upon which others seem unable to build.

By way of elaboration upon the point I'm making, it may be helpful to note that in the humanities, there is precious little difference

between the pragmatics (for example, writing a poem) and basic reseach. I would argue, for example, that much research on and criticism about a poem is, simply, in effect another poem or set of poems, even though couched in prose. When a literary scholar cites other relevant work at the beginning of an article or book, he sometimes does so to create an illusion of cumulativeness, but often to disagree with much of what others have said because it is through such divergence that creativity as a critic is demonstrated.

In my opinion, much social science is closer to the humanities than it is to physical sciences when it comes to the pragmatics/basic research distinction. Such is the case in part because in the social sciences--notably in linguistics--we are studying our own artifacts, and it is all too easy at (one hopes) the unconscious level to manipulate those artifacts (in the case of linguistics, symbol systems) to demonstrate a particular notion or theory. Although sometimes a problem, this kind of manipulation is much less likely to occur in the physical sciences, where some natural phenomenon is being studied. In the social sciences and in natural language research, a much greater openness to testing is needed. Lacking, as it does, an "unconscious level," the computer is in many ways ideal for such testing. For example, Joyce Friedman's programs have been used to test the consistency of grammars based upon a particular model of transformational grammar.

On the other hand, with reference to the value of the computer, we should be wary of constructing very elaborate, computer-based systems which do some one or two things very nicely, but which have no generality and make no contribution to the cumulativeness which we must have if

we're going to move toward any "ut›ia" (to use a word employed yesterday) re natural language understanding applications, such as machine translation. In other words, whe build computer systems we should think less about ad hoc demonstra s of notions or theories, and more about testable, generalizable syst s.

At present, as to machine translation, pragmatists should be encouraged to continue to blend together known technologies and techniques from which useful feedback into theory may evolve, while theorists should be encouraged always to do more than build elaborate demonstrations lacking general signdfi-ance (elephants which will never fly, to draw on yesterday's popular image).

****

Now I'd like briefly to turn to a couple of human factors issues relating to discussions in this meeting. The first concerns the consumer, or reader, of machine translations and the second involves the translation process in a computer-aided environment.

As to the first, I'd simply like to applaud the response to a suggestion that tran.lation of weather broadcasts into French would be much easier if only a few formats and phrases were permitted. The response: "That would be boring to read," shows laudable recognition of the importance of stylistic variety for readability and, more generally for communication; also, presumably for those of us gathered here, some grace in the use of language is one of life's pleasures and we would not care to be a party to its abandonment.

The second factor relates to the first, and concerns the suggested use of computer-based editing systems as an aid to translation. The

point I want to make may seem trivial or obvious but since a show of hands indicated that few if any of the professional translators at these sessions have used editing systems and I know that linguists who might advise on such systems have tended to concentrate on language strings no longer than a sentence, I think the point is worth making. That is, cathode ray screens which form the interface between man and machine in editing systems really can't display much text at a time. As someone whose professional concern for years was extended discourse, I find a cathode ray tube very confining; when reading and writing I like to be able to look backward at strings of at least a medium-sized paragraph's length. An ability to see that much text enables me to correct the kind of lapses one makes when writing--frequent repetition of a word or phrase, repetitive patterning in sentence length or structure, and so on. Although I've never been a professional translator, I assume that they have analogous requirements. Therefore, I'd urge that a system to be used in machine translation either provide larger screens or keep a kind of running summary which could be used to alert the translator through underlining, a warning message, or whatever, that, for example, a given word or phrase was being used too often. As you see, I am again speaking ot the issue of readability for, insofar, as possible, translations should be readable.

SALLY SEDELOW

Director of Techniques and Systems Program
Division of Computer Research
National Science Foundation
Washington. D. C. 20550

Dr. Sally Sedelow is the program director for the Techniques and Systems Program at the National Science Foundation. She funds projects directed toward providing machines with the capacity to understand natural language. Dr. Sedelow is on leave from the University of Kansas where she is Professor of Computer Science and Linguistics.

SUMMARY REPORT ON THE FBIS CONFERENCE
BY Richard See

1. The conference was very successful in bringing together experts on systems and techniques which one day may be useful in aiding the translation process or which are already available.

2. The overall impression I carried away from the conference was that none of the approaches presented were ready for immediate application by an agency now engaged in manual translation, without a great deal of preliminary preparation.

3. The machine translation systems demonstrated were clearly not yet able to completely replace human translation.

4. Whether or not presently available machine translation systems would be useful to an agency's translators in the preparation of human translations would have to be determined by each agency, based on the kind of text, the MT system available, and the type and quality of translation desired as a finished product.

5. General multifont OCR is not yet available and manual input is quite expensive with the techniques described.

6. In some special instances, text may already be available in machine-readable form.

7. The various possible benefits or advantages below would have to be examined by anyone interested in mechanizing some phase of the translation process:

a. lower cost
b. more rapid response (shorter lag)
c. higher quality, thru consistency of technical terms, for example
d. flexible capacity (possibility of handling larger volume than normally
e. byproducts of value (text-based dictionaries, concordances, IR)
f. training aids (using Chinese dictionary indices, gaining familiarity with the state-of-the-art)

8. The technology relevant to machine-aided translation is advancing and many costs are coming down. The conclusion is, that in order to be prepared for future developments, any agency seriously involved with translation should begin to be involved with this technology, if only on a small scale.

9. Conversely, because of the unlikelihood of immediate substantial payoff from investment in this technology and uncertainty as to the exact direction it should take, a cautious and evolutionary approach is recommended.

10. In addition to in-house experimentation with some of the elements discussed above, the support by contract of carefully designed comparative experiments involving two or more competitive approaches would aid in evaluating prospective techniques.

RICHARD SEE

Head, Data Processing Department

U.S. Naval Medical Research Unit

Taiwan

Educational Background:
Harvard College, Cambridge, Mass. (B.A.)
University of Oslo
University of California, Berkeley (M. A.)

PROFESSIONAL POSITIONS:

1958-1962	Professional Assistant, National Science Foundation, Washington, D. C.
1959-1963	Chairman, Interagency Committee on Mechanical Translation Research.
1959-1964	Chinese Translator, U. S. Joint Publications Research Service, Washington, D. C.
1961-present	Reviewer, Mathematical Reviews.
1962-1964	Deputy Program Director, Documentation Research Program, National Science Foundation, Washington, D. C.
1962-1965	Chinese Translator, American Mathematical Society, Providence, R. I.
1964-1966	Program Director, Information Systems Program, National Science Foundation, Washington, D. C.
1965-1967	Member, Panel 2, Committee on Scientific and Technical Information (COSATI), Federal Council on Science and Technology.
1966-1967	Program Director, Research and Studies Program, National Science Foundation, Washington, D. C.
1967-1970	Chief, Research and Development Branch, National Library of Medicine, Bethesda, Maryland.
1970-present	Head, Data Processing Department, U. S. Naval Medical Research Unit No. 2, Taipei, Taiwan.

Related Publications

See, R.: Mechanical Translation and Related Language Research. Science 144: 621-626, 1964.

See, R.: La Traduzione Automatica. Sapere 657, 1964.

See, R.: Machine-Aided Translation and Information Retrieval. Chapter 8 of Electronic Handling of Information, Thompson Book Co., 1967.

See, R., Editor: The Information Programs of the National Library of Medicine, 1969.

See, R.: Finite State Representation of Interactive Languages, FDT, ACM Special Interest Committee on File Description and Translation, 1: 44-46, 1969.

MACHINE (AIDED) TRANSLATION: GENERALITIES AND GUIDES TO ACTION

David G. Hays

Machine translation is Golem astride the Tower of Babel. Golem the automaton is the symbol of man's horror of the thing that straddles the line between spirit and flesh. The crumbling tower symbolizes ethnocentricity and xenophobia. Combined, these irrational feelings can influence national policy and retard progress toward important goals. To move too fast is as much an error as not to move at all. The principles of the first section summarize my reaction to the contributions presented at the conference; the guides of the second section express my opinion about the making of decisions in a fairly broad area.

GENERALITIES

1. Almost everyone hates computers, including most computer scientists. In "Information Handling" (Current Trends in Linguistics, ed. T. A. Sebeok et al., volume 12, pp. 2719-2740), I noted that professors who give their students clever tricks for skimming technical articles refuse to permit their computer programs to use the same tricks; the computer must work the hard way, in accordance with general theories of the structure of information. A friend suggests that hatred of the machine must be responsible. Anyone who hates computers is likely to design cumbersome systems.

2. The more programmers there are, the lower their average skill. In the early days of computation, the few programmers were brilliant; as the number has increased, the number of brilliant programmers has gone up, but the number of adequate or inadequate programmers has gone up faster. The buyer of a system must ask which kind will make it.

3. The best in computing is vastly better than ever before, but almost everything is worse. Tasks that required senior professionals long hours ten years ago can now be accomplished by students in courses, because the software is more powerful. Yet systems that cost too much for each transaction are in general use, thwarting their customers' hopes, and the public is led to believe that inflexibility and intolerance are characteristic of machines.

4. Scientists care how a system works; engineers care only how well it works. The buyer of a system for use is with the engineer, but the buyer of development is with the scientist. The claim that a system works "as a human does" needs to be checked by psychologists; but the claim has nothing to do with operating effectiveness, and not much to do with developmental promise.

5. A computer system is like zuppa inglese. English soup is an Italian dessert, made in a large hemispherical bowl. Layers of cake, soaked in liqueurs, are separated with thin layers of jam and covered with a thick layer of whipped

cream. The layers of a system are hardware, software, application programs, data base formats, data base contents, and so on. Claims of universality, simplicity, and the like are often no more than the assertion that a layer of whipped cream can cover anything. Deep probes are necessary to evaluate such claims.

6. If everyone optimizes his own cost effectiveness, the system goes to pieces. The classic example is the war against German submarines in the Mediterranean. It was so successful that the Germans moved into the North Atlantic and nearly starved the British. Translation is not the end of the whole system; to raise internal costs can make the system at large much more effective if done right.

7. Brevity counts. The time of the reader has to be reckoned into the cost of the system; translations of key points can be more suitable than full translations. The machine may be more useful in finding passages than in translating them.

8. You cannot make a jumbo jet out of an elephant by pulling its ears. Martin Kay suggested that Hannibal was wiser to buy elephants to cross the Alps than he would have been if he had let a development contract for jet transport. Contrariwise, suitability as a chassis for the future jet is no criterion for selection of a first-stage machine; sooner or later it will be necessary to scrap the whole system and start over. What counts in the first installation is whether or not it works as installed, for however limited a purpose has been selected.

9. Almost everyone hates translators. They arouse our xenophobia by bringing the enemy into our camp. To give them help in their task, or credit for doing it, is loathsome.

10. Big ideas are easier to understand than little ones. Some examples of big ideas mentioned in the conference are words (as opposed to characters) as objects for optical recognition; syntactic patterns (as opposed to diagnostic contexts) in language processing; and scripts or frames (as opposed to grammatical and syntactic structures) as objects for computers to seek in texts. It might be easier to find that a news story is about a certain frame (detente), and that the source is Sadat. than to translate the whole; and the summary ("Sadat endorses detente") might be more helpful to the user than the translation would be.

GUIDES

1. A prima facie case has been made for gradual introduction of language-processing capacity into intelligence facilities.

2. System design and cost analysis remain the essential prerequisites to procurement.

3. The design should take into account as fully as possible the needs.of users of translations.

4. No adequate reason for selecting a single system and excluding the rest has come to light thus far.

5. The main developmental track for a few years ahead is from character processing (editing systems) to word processing (dictionaries).

6. A plausible further development for the three to seven year prospect is automatic recognition of topic (for example, of requirements), and the matching of new text against old for partial identification of redundant, and therefore omittable, information.

7. The operational suitability of language-processing systems depends crucially on the smallest details of their design. As yet, only those of clearly superior knowledge, taste, and judgment can be entrusted with the work.

8. Several classes of systems are fundamentally different and cannot usefully be intermingled. Current commercial MT systems, which make no provision for editorial intervention between the earliest and latest stages of processing, are not suitable bases for machine-aided (editorial) systems; and the latter are not necessarily suitable bases for full-scale language-processing systems that may reach installability in as little as ten years if research and development are well supported.

SUMMARY NOTES

FBIS Seminar

Jim Mathias

Mr. Mathias concluded the summary presentation by restating some common threads running throughout.

The moderators and commentators participated in the conference in order to assist the sponsor in arriving at reasoned decisions on planning and budgeting for possible application of computer technology where it would increase the cost effectiveness of performance. The summary panelists did not address themselves to the users of FBIS material since the user is unknown but to the translation services as described by the sponsor. This omits the important element alluded to by Mr. Hays when he suggested that the sponsor should look beyond the function of translation and consider the purposes for which the work is done.

The nature of human motivation is critical in the translation process and in the undesirable effects that can result from unwise division of tasks between the human translator and the computer. It was said that too often the human translator is asked to do the difficult tasks while the system designers assign the simpler tasks to the computer. This relegates the translator to second-class citizen and can seriously affect his motivation and his production. The obvious preference is to assign to the computer functions which it can perform well without imposing added undesirable tasks on the human translator in order to compensate for computer shortcomings.

There was a general consensus that the computer should be introduced into FBIS translation process wherever it is possible to maximize current capabilities for current needs. This would imply use of off-the-shelf items, research and

development where off-the-shelf items were not really adequate to the tasks, or establish a holding pattern for those functions which have been developed in the research community and not yet applied to off-the-shelf hardware.

It was suggested that the sponsor should develop a means of verifying usefulness of existing technology and systems. The verification of existing technology might be best achieved by establishing an in-house awareness through maximum exposure to research and development in the commercial and academic community. This might require the establishment of one or more high-level slots for personnel assigned specifically to monitoring developments and capabilities, or it might require establishment of a series of seminars for intensive familiarization of sponsor personnel. The verification of systems, however, might be far better undertaken through the application of dependable objective scientific tests. These tests should be conducted by the sponsor or an independent agent for the sponsor and not by designers, developers, or promoters of candidate systems. The need for experimental methodology was emphasized.

It was generally concluded that during the process of selecting systems or hardware, for application to sponsor tasks, that maximum flexibility be one of the principal criteria applied in order to assure long term usefulness and avoid costly replacement. The approach taken should not be set in concrete but should reflect the ability to cut off one method of approach if it appears unfruitful and shift to another effort or another direction. Avoid the forced choice of any single system by avoiding reliance on any one approach.

APPENDIX

Conceptual Outlines of Machine Translation. Systems and Experience

In preparation for the Seminar, we prepared outlines and distributed them to contributors. The first draft was prepared by members of the staff of the Foreign Broadcast Information Service. The Co-ordinator and Commentators then added their suggestions. Hays edited them, and Mathias re-edited them.

Since few reports on machine translation, and very few design proposals, cover every point, yet every system put into even the most tentative operation soon encounters at least all of the problems indicated by these outlines, we have included our Outlines in the hope of stimulating fuller planning and reporting. They might even suggest areas in which linguistics and artificial intelligence still have theory to build and research to complete.

With a dozen more contributors, the outlines would have grown; they are not presented as complete. -- DGH

OUTLINE FOR SYSTEM BUILDERS-

Applicability

Language (s)
Field (s): science, technology; international relations, and so; narrowly specified
Purpose: trained or untrained readers, skimming or detailed understanding, other

Operational configuration

Pre-editing: nature and extent
Postediting: nature and extent
Interactive editing: nature (kinds of interrupt, control structure) and extent

Hardware

Equipment: required, optional
Input mode (s): punched cards, display terminal, teletype, light pen, OCR
Output mode (s): lineprinter, display screen, teletype, photocomposition; best currently available quality, cheapest currently available quality
Processing mode (s): batch, remote batch, interactive

Software

Programming philosophy: system sketch
Modules: dictionary, grammar, semantics, real-world knowledge, or other scheme
Control Flow

Linguistics

Underlying model: general characterization
Lexicon: format, size
Syntax: agreement of number, tense, person; conjunction of words, phrases, clauses; relativization; complementation; size
Semantics: control of field and domain-specific terminology; choice of equivalent by part of speech, syntactic function, semantic agreement; handling of idioms; size of semantic component
Discourse: anaphora, cataphora; consistency of universe of discourse, tense, person, number; figures of speech (metaphor, simile); paragraph linking and transition; size of discourse component
Style: variation among synonymous words or grammatical constructions, of sentence length, of paragraph order; control of tone (lexical and grammatical); size of style component

OUTLINE FOR SYSTEM BUILDERS

Extendability

Feasibility of revising or extending each linguistic component: adding rules to the grammar, adding words to the dictionary, adding conditions to a rule for selection of an equivalent, etc., according to the linguistic model used
Standard procedures for feedback from user to system that result in permanent changes

Evidence

Evaluations: date, name of evaluator, extent, method, results
Failures: frequency and method of handling words not in the dictionary, sentences not parsed, other failures

Speed

Turnaround time for a batch of text; batch size
Input rate: words per operator per working day
Pre-editing rate: words per editor per day
Postediting rate: words per editor per day
Interactive editing rate: ratio of editor's time to system output
Processor time: per sentence, according to length, syntactic complexity, etc.; per 1000 words, average
Rates for any other operations

Cost

Dollar cost for installation of the existing system
Dollar cost for recommended immediate development
Dollar cost for operation by component: input, editing, etc.
Dollar cost for improvement after installation
Do operating options permit modes with different costs?

Status

Is the system ready for immediate installation, for development, or for research?
Are the remaining R & D questions factual or theoretical? The answer to this question requires data and argument.

SUPPLEMENT TO SYSTEM OUTLINE

Output

Form of output: similarity to polished translation
Inclusion of source language: complete, partial, none
Commentary: remarks, diagnostics

Documentation

User tools: manuals, dictionaries
Operator documentation
System documentation

User Role

Linkage to system: input, dialog

Extendibility

Means of quality control for changes
Preventing oscillation: changing back and forth between alternatives

Software

Portability: can system be transferred to different hardware?

OUTLINE FOR SYSTEM OPERATORS

System description

Source and date of the current installation

Application

Annual output
Language (s): percentage distribution
Field (s): percentage distribution
Users: percentage distribution by level of training in foreign language and technical field; total number
Purpose (s): skimming for selection; keeping up with a field; background for research; state-of-the-art reviews; evaluation of progress in a field

Operational configuration

Pre-editing: nature and extent
Postediting: nature and extent
Interactive editing: nature (kinds of interrupt, control structure) and extent

Hardware

Equipment: required, optional
Input mode (s): punched cards, display terminal, teletype, light pen, OCR
Output mode (s): lineprinter, display screen, teletype, photocomposition; best currently available quality, cheapest currently available quality
Processing mode (s): batch, remote batch, interactive

Software

Programming philosophy: system sketch
Modules: dictionary, grammar, semantics, real-world knowledge, or other scheme

Linguistics

Underlying model: general characterization
Lexicon: format, size
Syntax: agreement of number tense, person; conjunction of words, phrases, clauses relativization, complementation; size
Semantics: control of field and domain-specific terminology; choice of equivalent by part of speech, syntactic function, semantic agreement; handling of idioms; size of semantic component
Discourse: anaphora, cataphora; consistency of university of discourse, tense, person, number; figures of speech (metaphor, simile); paragraph linking and transition; size of discourse component
Style: variation among synonymous words or grammatical constructions, of sentence length, of paragraph order; control of tone (lexical and grammatical size of style component

OUTLINE FOR SYSTEM OPERATORS (2)

Extendibility

Feasibility of revising or extending each linguistic component: adding rules to the grammar, adding words to the dictionary, adding conditions to the rule for selection of an equivalent, etc., according to the linguistic model used
Standard procedures for feedback from user to system that result in permanent changes

Evidence

Evaluations: frequency, names of evaluators, extent, method, results
Trends in quality since installation

Speed

Turnaround time for a batch of text; batch size
Input rate: words per operator per working day
Pre-editing rate: words per editor per day
Postediting rate: words per editor per day
Interactive editing rate: ratio of editor's time to system output
Processor time: per sentence, according to length, syntactic complexity, etc.; per 1000 words, average
Rates for any other operations

Cost

Dollar cost for installation of the existing system
Dollar cost for recommended immediate development
Dollar cost for operation by component: input, editing, etc.
Dollar cost for improvement after installation
Do operating options permit modes with different costs?

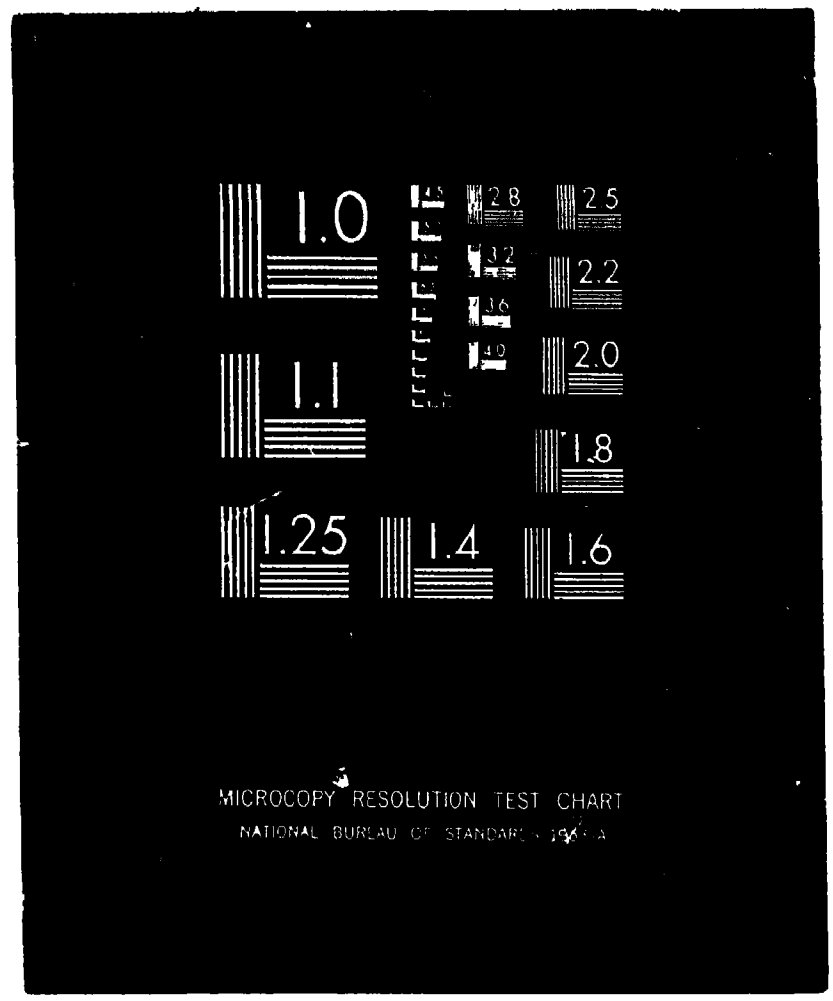
User response

Errors in MT: detected or missed? inaccuracies in resulting analyses? corrected by user or referred to translator?
Requests for translation or verification by human specialist: before or after seeing MT? reason given? frequency, quantity
Morale, producitivity, effectiveness of users

Improvement

Effectiveness and cost of improvement program

END


1.0
1.1
1.25
1.4
1.6
2.0
1.8
2.2
2.5
2.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS