[From: *Overcoming the language barrier*, 3-6 May 1977, vol.1
(München: Verlag Dokumentation, 1977)]

THE PRESENT STATE OF MACHINE AND MACHINE-ASSISTED TRANSLATION

H. E. Bruderer

Institut für linguistische Datenverarbeitung
Münsingen, Switzerland

Abstract

The purpose of this contribution is to describe the latest status of research and application in the field of machine and machine-assisted translation.

First the bases of machine translation and its technical and linguistic premises are briefly explained. This is followed by an account of the practical results so far attainpd and a survey of present translation systems, two of which are described in detail.

By means of a calendar of meetings it is demonstrated in conclusion that research has received a new impetus in recent years, not least thanks to the growing demand. If the European Communities are to be able to overcome their difficulties of communication they must in future follow these efforts closely.

The discussion is illustrated by numerous figures.

1. INTRODUCTION

This contribution is intended as a source of information and a survey of the state of the art. Above all it is meant to give the practical man initial assistance in making decisions.

2. FUNDAMENTALS OF MACHINE TRANSLATION

Linguistic data processing is a branch of applied linguistics; it is an interdisciplinary subject
subfields may be distinguished (Fig. 2). The possibilities of computational linguistics are manifold; automatic translation is a central field (Fig. 3). A translation system consists in essence of three components (Fig. 4); as a rule the process of translation is effected in three main stages (Fig. 5).

3. THEORETICAL PREMISES

The technical bases for machine and machine-assisted translation are available today: computers with sufficient main storage and high processing speed. However, text collection is still a bottleneck. Translation costs are often difficult to ascertain. With large quantities of text translation systems ought in general to be economic.

The linguistic premises for fully automatic translation, on the other hand, are still deficient. For up to now it has not proved possible also to include factors of semantic and pragmatic considerations to a sufficient extent. Most systems are syntax-related. Semantically based methods are barely ready for application. Difficulties are caused above by ambiguities and pronominal reference (Figs. 6 and 7). Consequently, the quality of machine translation leaves something to be desired. Perfect automatic translations are not to be expected in the near future.

4. PRACTICAL RESULTS

At present only a few machine translation systems can be used in practice: SYSTRAN (Russian-English), CULT (Chinese-English), METEO (English-French), TITUS (German/English/French/Spanish), and GAT (Russian-English).

For western languages the METEO and TITUS methods, and in part also SYSTRAN (English-French), enter into consideration. In the near future further operational systems will probably join the above ones: those of the universities of Grenoble and Montreal, Brigham Young University and the Logos Development Corporation.

The following machine-assisted translation systems are capable of functioning: LEXIS (Bundessprachenamt), TEAM (Siemens), TERMIUM (Montreal University), EURODICAUTOM (European Communities), EWF (Dresden University of Technology). The Bundessprachenamt has the longest practical experience.

Figs. 8 and 9 show the language combinations at present in use.

5. SURVEY OF THE EXPERIMENTAL AND OPERATIONAL SYSTEMS

Cf. Figs. 1O and 11.

6. DESCRIPTION OF TWO METHODS

The two examples, the fully automatic translations system of the University of the Saarland and the machine-aided translation system of Messrs Siemens, were chosen at random (Figs. 12 and 13). All systems are described in exactly the same form qeneral, linguistic, technical and exonomic aspects being considered. This form of presentation facilities comparison.

7. THE UPSWING OF MACHINE AND MACHINE-ASSISTED TRANSLATION

In recent years the number of congresses on fully automatic and semi-automatic translation in Western Europe, in the countries of the Eastern Bloc and in the United States has greatly increased. In addition, several experiments have been performed, for instance the trial of the SYSTRAN system by the Canadian Government, the Gesellschaft für Mathematik und Datenverarbeitung and the European Communities. In Europe there have been numerous demonstrations of translation systems: Bonn and Zürich (SYSTRAN), Luxemburg (SYSTRAN, EURODICAUTOM, TERMIUM, LEXIS), Saarbrücken (Saarbrücken system), Brussels (METEO), Paris (EURODICAUTOM), among others. New projects are known to exist in the USA, the USSR, Canada, Iran, and the European Communities.

Fig. 14 lists some of the activities in the field of natural language processing.

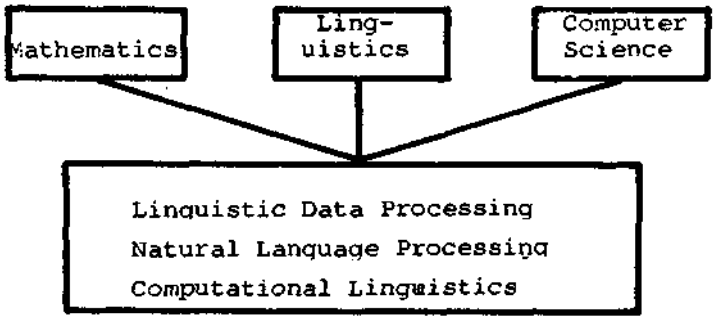
8. CONCLUSION

Unfortunately research into machine translation has made little progress in the last few years. Today there is not a single translation system which can make a perfect translation of any desired technical of scientific text from one natural language into another. However, it is of great importance to the Commission of the European Communities, whose language problems will become even greater in the future, to keep a close watch on the results of research. Experience with the various terminology data banks should also be highly instructive.

Perhaps I may close by referring to my "Handbook of Machine Translation and Machine-Aided Translation - Automatic Translations of Natural Languages and Multilingual Terminology Data Banks", North-Holland Publishing Company, Amsterdam 1977, some 700 pages.

Fig. 1

Building blocks of computational linguistics

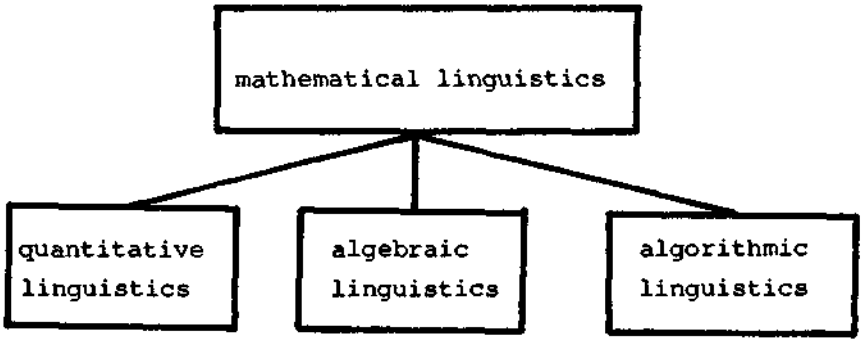


Note

only the most important components are shown here.

Fig. 2

The structure of mathematical linguistics



Note

Quantitative linguistics is also called statistical linguistics, and algorithmic linguistics computational linguistics.

Fig. 3. POSSIBILITIES OF LINGUISTIC DATA PROCESSING

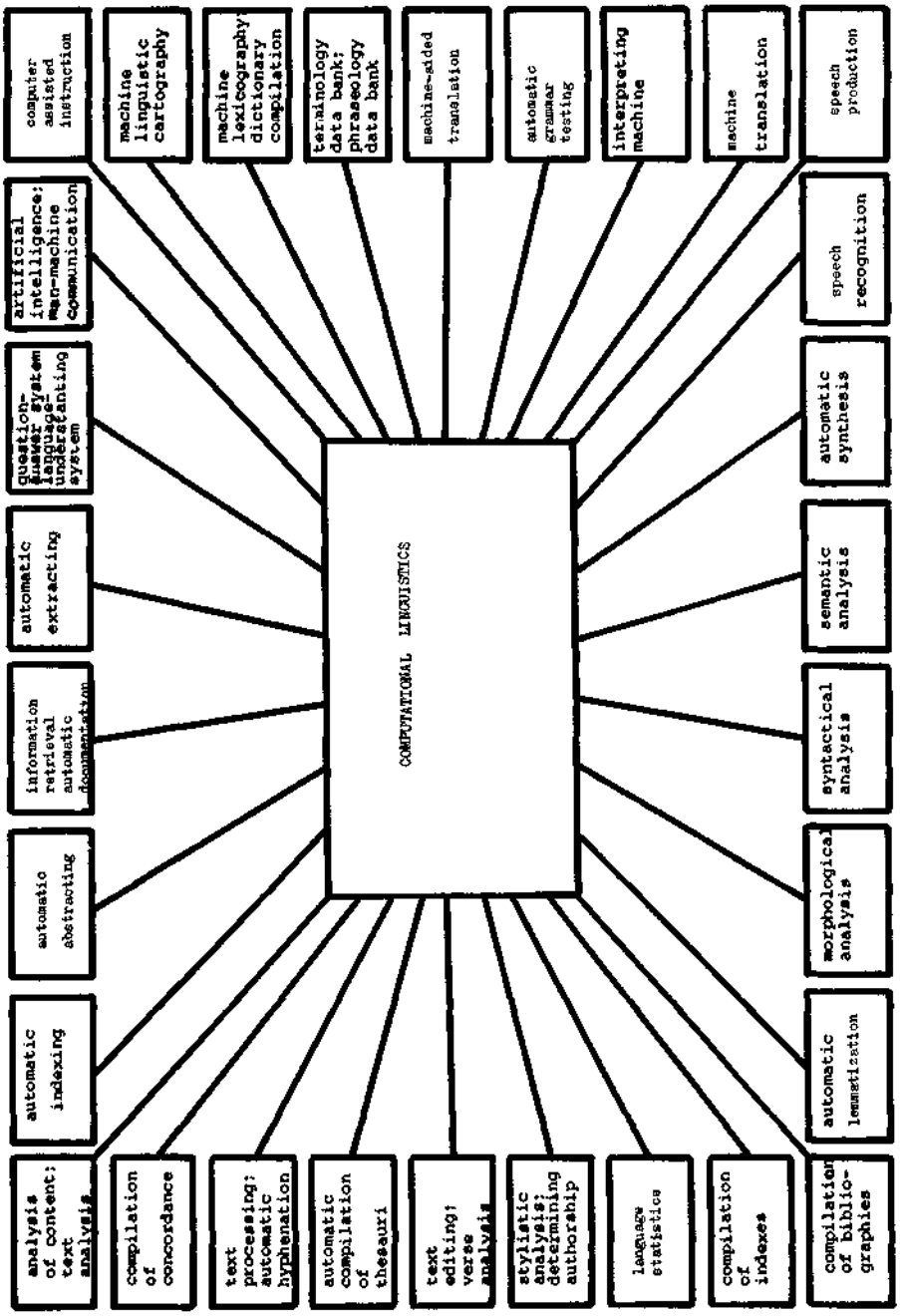
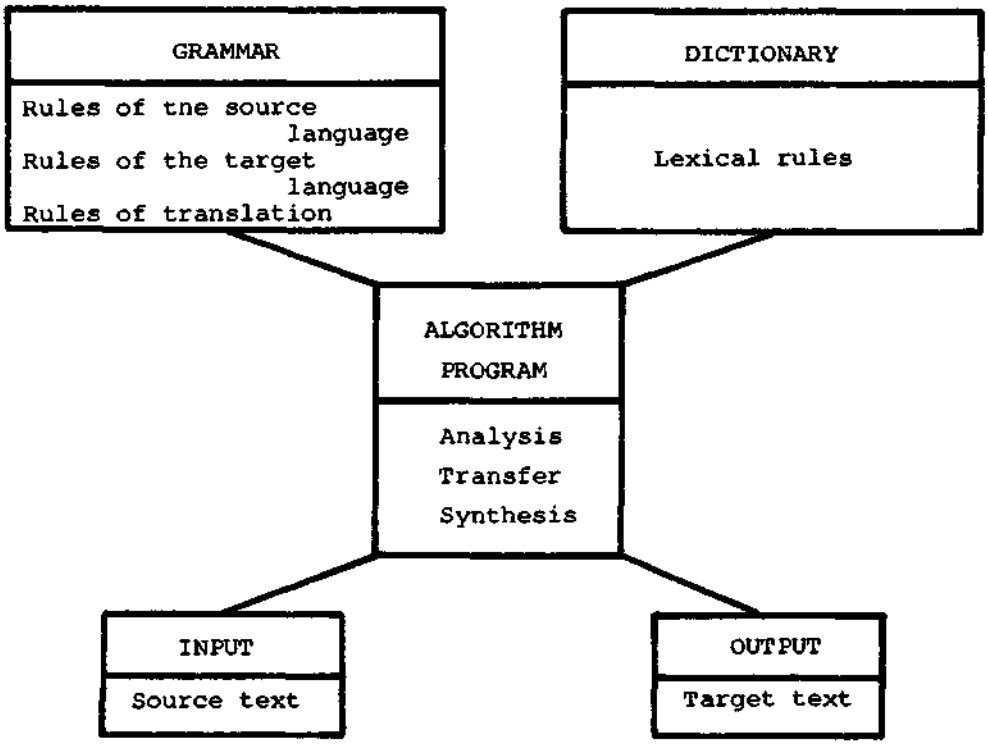


Fig. 4

Components of the translation system



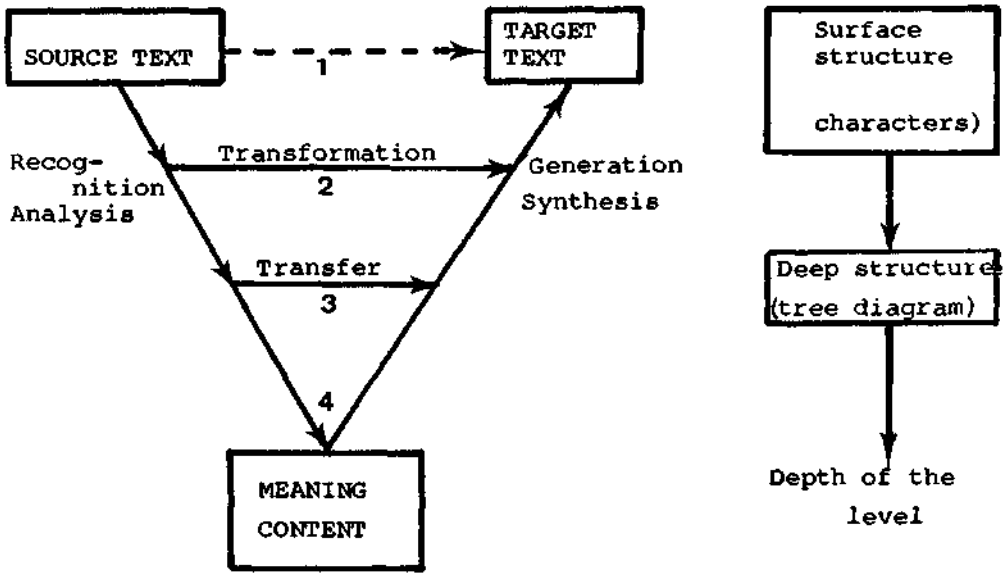
Note

A tripartite translation system comprises a grammar, a dictionary and an algorithm (a program). In a bipartite system the grammar is incorporated in the algorithm. The grammar and the dictionary may also form one unit. In this case the grammar consists of a part devoted to rules and a lexicon.

Advantages of a tripartite system: language-independent programs; it is possible to change the grammar without changing the program. Disadvantages: less efficient (more storage required, longer processing time).

Fig. 5

Stages of the process of translation



universal intermediate language

language-independent representation

of meaning

Note

Method 1: word-for-word translation with morphological analysis

Method 2: translation with morphological and syntactical analysis

Method 3: translation with morphological, syntactical and semantic analysis (possibly in addition pragmatic analysis with the aid of artificial intelligence).

Method 4: translation via a universal intermediate language (Interlingua)

Solutions 1 and 2 are possible today; 3 is partly possible, 4 is still beyond reach.

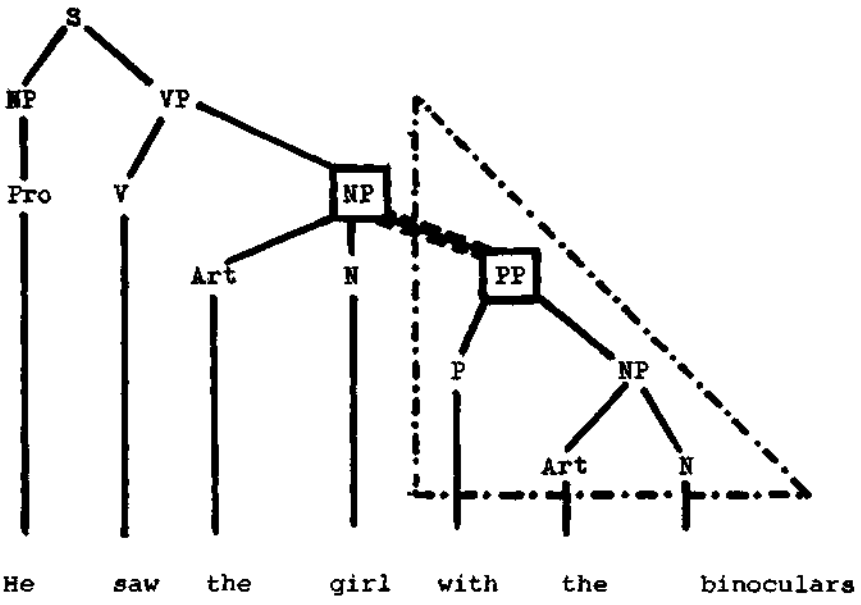
Fig. 6

Interrelation of prepositional phrases

Explanation of abbreviations:

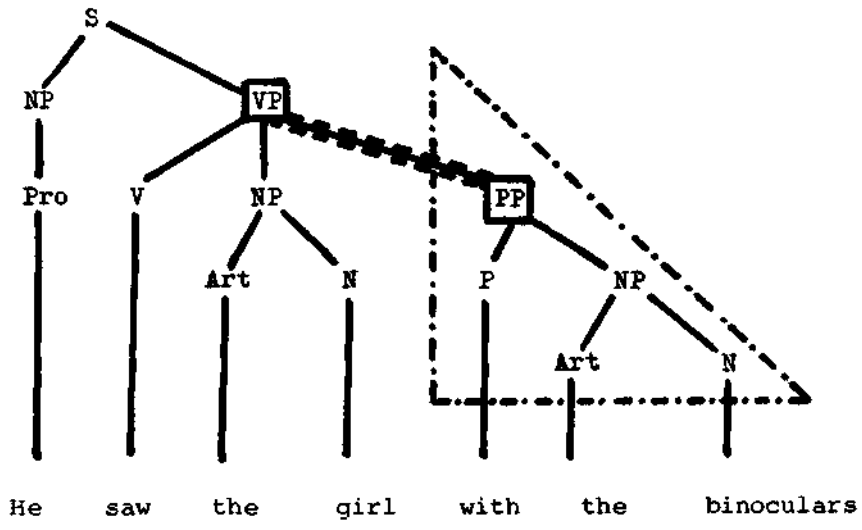
Art	=	article
N	=	noun
NP	=	noun phrase
P	=	preposition
PP	=	prepositional phrase (prepositional group)
Pro	=	pronoun
S	=	sentence
V	=	verb
VP	=	verbal phrase (verbal group)

a) Enlargement of object
What/which girl does he see?



b. Adverbial statement

With what does he see the girl?



Note

In meaning a. the prepositional phrase accompanies the noun phrase (girl); it depends on the direct object.
In meaning b. the prepositional phrase relates to the verb phrase, and in particular to the verb (saw).
The triangle marks the prepositional phrase.

Fig. 7

Syntactical and semantic ambiguities

a) Text

source sentence	Target sentences			
Das ist ein grosses Schloss.	This is a	large / big	castle	C'est un grand château.
	This is a	large / big	lock.	C'est une grande serrure. / C'est une grande culasse.

b) Dictionary search

Form of word	Part of speech			English equivalent					French equivalent			
das	Definite article	Neuter singular	Nominative / Accusative	the					le	la		
	Demonstrative pronoun	neuter singular	Nominative / Accusative	this (one)	that (one)	he	she	it	cela	ceci	c' / ce	
	Relative pronoun	neuter singular	Nominative / Accusative	who / who(m)	which		that		qui / que	lequel	laquelle	
ist	Verb	present indicative singular	3d person	be					être			
	Adverb	Idiom	uninflected	(Idiom)					(Idiom)			
	Verb(al) Prefix	Separable verb	uninflected	-					-			

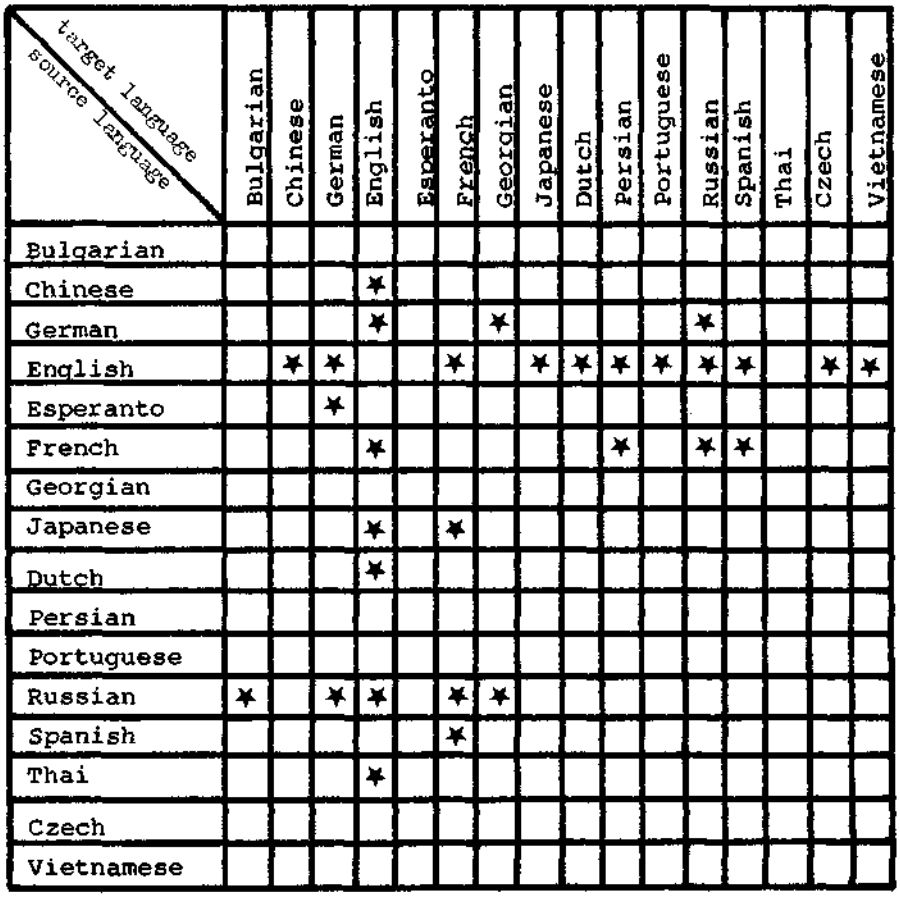
ein	indifinite article	masculine singular	nominative	a(n)				un(e)			
		neuter singular	nominative								
			accusative								
	indifinite pronoun		uninflected	(Idiom)				(Idom)			
	Numeral (cardinal number)	masculine	nominative	one				un(e)			
		neuter	nominative								
			accusative								
			uninflected								
grosses	adjektive	positiv neuter singular	nominative	big	large	great	tall[2]	grand(e)[2]			
			accusative								
schloss[1]	noun	neuter singular	nominative	castle		lock[2]		château	serrure	culasse[2]	
			dative								
			accusative								
	verb	imperfect indicative singular	1st person	shut	close	lock	conclude[2]	fermer		conclure[2]	
			3rd person								

Note

[1] Capitalization has been ignored

[2] Only the most common meanings – in the nominative – have been given

Figure 8 - Source and target language used for machine translation



target language / source language	Bulgarian	Chinese	German	English	Esperanto	French	Georgian	Japanese	Dutch	Persian	Portuguese	Russian	Spanish	Thai	Czech	Vietnamese
Bulgarian																
Chinese				*												
German				*			*					*				
English		*	*			*		*	*	*	*	*	*		*	*
Esperanto			*													
French				*						*		*	*			
Georgian																
Japanese				*		*										
Dutch				*												
Persian																
Portuguese																
Russian	*		*	*		*	*									
Spanish						*										
Thai				*												
Czech																
Vietnamese																

Fig. 9

Source and target languages used for machine-assisted translation

Institution / Language	University of Montreal	IBM New York	Bundes-sprachenamt Hürth	Siemens AG Munich	Tech-nological University of Dresden	European Communities Luxembourg
Danish						*
German		*	*	*	*	*
English	*	*	*	*	*	*
French	*		*	*	*	*
Italian			*	*		*
Dutch				*		*
Portuguese			*	*		
Russian			*	*	*	
Spanish				*		

Fig. 10

Survey of machine translation systems

1 AMERICA

11 Canada

Lakehead University, Thunder Bay
University of Montreal

12 United States
Atomic Energy Commission, Oak Ridge/Georgetown University, Washington D.C.
USAF, Dayton
NASA, Houston
Brigham Young University, Provo
Massachusetts Institute of Technology, Cambridge
University of California, Berkeley
University of Texas, Austin *
University of Texas, Austin/Ramkhamkaeng University, Bangkok
Yale University, New Haven
Latsec,Inc./World Translation Center,Inc., La Jolla
Logos Development Corp., New Hampton
Smart Communications,Inc., New York
Xonics Inc., McLean/Tabor.Inc., Nokesville
Xyzyx Information Corp., Canoga Park

2 ASIA

21 British Crown Colony of Hongkong
Chinese University of Hongkong

22 Japan
University of Kyoto
Kyushu University, Fukuoka
Electrotechnical Laboratory, Tokyo

23 Lebanon
International Language Centre, Beirut

24 Malaysia
University of Science of Penang

3 EUROPE

31 Belgium
University of Antwerp

32 Bulgaria
Bulgarian Academy of Sciences, Sofia

33 Federal Republic of Germany
University of Heidelberg/University of Constance
University of Cologne
Ruhr University, Bochum
University of the Saarland, Saarbrücken
Centre for Textile Documentation and Information, Düsseldorf

34 France
University of Science and Medicine, Grenoble
French Textile Institute, Paris

35 Great Britain
University College, Cardiff
University of Essex, Colchester
Pearl Assurance Co., Ltd., London/Natural Language Translation Specialist Group

36 Italy
Euratom, Ispra *

37 Soviet Union
State University of Leningrad
Group for Language Statistics, Leningrad, Minsk, Kishiniov, Machatshkala, Alma-Ata, Irkutsk
Central Research Institute for Patent Information, Moscow *
Atominform, Information Centre for Nuclear Energy, Moscow
Informelektro, Documentation Centre of the Institute for Electrical Engineering, Moscow
Institute for Applied Mathematics, Moscow
State Pedagogic Institute for Foreign Languages, Moscow
All-Union Centre for Translation of Scientific and Technical Literature and Documentation, Moscow
Institute for Electronics, Automation and Telemechanics, Tiflis

38 Czechoslovakia
Karlovy University, Prague

Note
* Further development or use abandoned or interrupted.

Fig. 11

Survey of machine-assisted translation systems

1 AMERICA

11 Canada
University of Montreal/Department of the Secretary of State, Ottawa

12 United States
IBM, New York

2 EUROPE

21 Federal Republic of Germany
Bundessprachenamt, Hürth
DEMAG A.G., Duisburg
Siemens A.G., Munich

22 German Democratic Republic
Technological University of Dresden

23 Luxembourg
European Communities, Luxembourg

24 Netherlands
Netherlands Ministry of Foreign Affairs, The Hague
N.V. Philips Gloeilampenfabrieken, Eindhoven

Note

This list does not include the numerous multilingual standardization data banks (dictionary-related terminology data banks).

Fig. 12

Description of a machine translation system

UNIVERSITY OF THE SAARLAND, SAARBRUCKEN, FEDERAL REPUBLIC OF GERMANY

1. General	
11 Name of the system	-
12 Characterization	automatic translation
13 Researchers	Hans Eggers, Heinz Dieter Maas
14 Objective	method-oriented
15 State of development	testing phase
16 Start of research work	1967
17 Start of experimental operation	in the latest version 1974 (earlier version 1969-73)
18 Information as at	September 1976
2. Linguistic data	
21 Overall characterisitics	
211 Language pairs	Russian-German, English-German, Esperanto-German
212 Direction of translation	at present not reversible (dependent on the dictionaries
213 Capability for extension	The analysis can be used for any (similar) languages; at present it is utilized both for Russian and for German
22 Corpus	mathematical and linguistic publications; popular scientific texts (several million word forms)
23 Romanization	1:1 conversion (Cyrillic)
24 Dictionary	
241 Basic structure	a. separate dictionaries for analysis and transfer b. frequency dictionary, general dictionary c. stem dictionary
242 Data record	a. variable length of the dictionary entries b. morphological, syntactical and partly semantic data
243 Size	Russian dictionary: approx. 13,500 entries (8500 lemmas, i.e. basic forms) Russian frequency dictionary: 255 word forms

Russian-German dictionary: 8500 entries
Esperanto: 4000 words
English: a few hundred words

244 Fields covered
general vocabulary

245 Dictionary lookup
a. Each text word is split (segmented) from right to left until a possible ending is found. The left-hand part - the possible stem - is sought indexed-sequentially in the dictionary. Splitting is continued even if a solution has already been found.
b. principle of longest possible match for fixed syntagms (idioms)

25 Grammar

251 Language model
a. The system is developed in several phases:
1st phase: surface grammar
(with partial solving of homographs on the basis of distribution)
2nd phase: consideration of transformational structures
(with treatment of syntactical ambiguities)
b. transformational rules
c. syntax-based, related
d. input of the grammar partly as phrase structure rules independent of the algorithm, partly incorporated in the algorithm

252 Translation procedure
a. three-stage: analysis, transfer, translation
b. The analysis is directed towards the source language, and the synthesis towards the target language.
c. analysis form the word to the sentence. For noun groups predictive analysis is partly used.
d. at least three sentence passes form left to right

26 Human intervention
no pre- or post-editing

27 Quality of translations
translations are possible for relatively simple sentence structures (e.g. with a relative connection)

28 Further applications
KWIC (keyword-in-context) indexes

3. Technical data	
31 Computer	Telefunken TR 440
32 Operating system	MV 17 (Maintenance Version)
33 Main memory requirements	52 K words (The translation program is broken down into five main programs that run successively;otherwise approx. 200 K words would be needed)
34 Storage media	magnetic dis s, magnetic tapes
35 Programming languages	Fortran, Telefunken assembler (TAS)
36 Data collection	punched cards, screen
37 Output	printer, screen
38 Character representation	capital letters
4. Economic data	
41 Number of translations	cannot be stated (testing phase)
42 Promoter of research	Deutsche Forschungsgemeinschaft
43 Users	none
44 Speed	15,000-20,000 words an hour
45 Cost	cannot be stated (testing phase)
5. Remarks	The University of the Saarland is a member of the "Leibniz" international research group for automatic translation, founded in 1974. The Saarbrücken translation system was successfully demonstrated in Saarbrücken on 24 September 1976 (language pairs: Russian/English/Esperanto - German)

Fig. 13

Description of a machine-assisted translation system

SIEMENS AKTIENGESELLSCHAFT, MUNICH, FEDERAL REPUBLIC OF GERMANY

1. General	
11 Name of the system	TEAM (Terminologie-Erfassungs- und Auswertungs-Methode)
12 Characterization	terminology data bank for purposes of machine-assisted translation (mechanical aids to translation) and other uses
13 Directors of Research	Karl-Heinz Brinkmann, Joachim Schulz
14 Objective	practice-related: mechanical aids to translation in the widest sense
15 State of development	operational: lexicographical branch, inquiry branch (batch and conversational mode)
16 Start of development work	1967
17 Start of practical use	1970
18 Information as at	October 1976
2. Linguistic data	
21 Overall characteristics	
211 Source and target languages	German, English, French, Italian, Dutch, Portuguese, Russian, Spanish
212 Direction of translation	as desired
213 Capability for extension	capable at present of extension to nine languages
22 Corpus	technical literature of every kind, standards, manuals, specifications etc.
23 Romanization rules	Cyrillic: ISO transliteration
24 Dictionary	
241 Basic structure	basic forms (full forms in phraseological entries)
242 Data record	a. variable field lengths, variable data record (at present up to 3000 bytes) b. 99 information criteria, e.g. single-word and multiword terms and/or phrases in the various

	languages, additional information like subject codes, part of speech (in terms), source data, synonyms, abbreviations, device and system compatibility and the like, definitions, examples of contexts etc. (possible, but not used at present: phonetic transcription, data on language level, references to illustrations etc.)
243 Size	approx. 700,000 entries with some 2 million terms in the languages mentioned (of the entries, some 80% in German, 70% in English, 40% in French,and 30% in Spanish are supported by terms; less than 20% in other languages)
244 Fields covered	mainly electrical engineering and fields of application
245 Dictionary lookup	in batch operation and for lexicographical purposes sequential; in conversational communication direct access, search by term or phrases via "sort criteria" (in this way spelling variants are taken into account); further, searches by any of the criteria stated under 242. In batch operation and conversational communication automatic generation of "secondary questions" when the primary questions are not answered
25 Grammar	
251 Machine data preparation	none
252 Translation procedure	machine-assisted
26 Human intervention	feedback to the basic forms
27 Result of the inquiry	direct answers via screen and/or teleprinter. In batch operation text-related and/or text-related alphabetical lists of technical words via high-speed printer; dictionaries via Digiset photocomposition unit. Output of microfiches via COM
28 Further applications	monolingual register. Through supplementary programs word and text concordances (with variable context length), statistical linguistic investigations (research analysis)

machine-assisted language teaching

3. Technical data

31 Computer	Siemens 4004/35 or larger, Unidata series 7000
32 Operating system	BS (Betriebssystem) 1000
33 Main memory requirements	at least 65 K
34 Storage media	magnetic discs, magnetic tapes
35 Programming language	assembler
36 Data collection	6-channel TTS punched tape and/or OCR-B sheets (character reader)
37 Inquiry	a. text-related inquiry: the translator underlines the technical expressions which he does not know in the text, and these are then input into the system as a list of questions via punched cards or tape. As answers the system supplies a text-related or - at choice - alphabetical list via a high-speed printer. Cf. also 245 b. conversational communication via visual display unit and/or teleprinter. In both cases the output can be recorded as required. c. Inquiry for dictionary compilation cf. 245
38 Character representation	capitals and lower case, special characters Cyrillic in ISO Romanization (37a and b) Via Digiset the ISO Romanization is transliterated back into Cyrillic etc. (37c)

4. Economic data

41 Number of inquiries	at present about 500 inquiries a week (with approx. 150 permanent employees of the translation department, of whom just under 100 translators)
42 Sponsor	Siemens Aktiengesellschaft; assistance by the Federal Ministry for Research and Technology
43 Users	translation departments of Siemens Aktiengesellschaft N.V. Philips Gloeilampenfabrieken and the Dutch Ministry of Foreign Affairs, InTra 1, Fachübersetzergenossenschaft GmbH, several publishers

44 Speed

in conversational communication less than 500 millseconds per inquiry, in batch operation 0.5 -3 seconds per inquiry, depending on the size of the batch (large batches need less time per inquiry)

45 Cost

between 0.15 and 0.08 DM per question in batch operation, depending on the size of the batch.

5. Remarks

Team is a program system for the solution of terminological and lexicographical tasks, in particular for the provision of machine translation assistance. The files contained in its data bank can also be introduced via an interface program into the GOLEM information system (grosspeicher orientierte listenorganisierte Ermitblungsmethode, a bulk memory-orientated, list-organized system). The recall ratio in inquiry operation is between 60 and 90, depending on the subject field.
At present conversational communication is available only to the terminologists. Through text-related lists of technical words TEAM reckons on an increase in productivity of up to 60%. Moreover, in the case of large translation jobs divided among several translators such lists guarantee the uniform use of the desired terminology and thus render a higher quality of translation possible.
As up to 30 subject codes can ge assigned to one entry in the data bank, it is possible to make allowance for the various classification systems.
Inquiries are in natural language.
An automatic inquiry method, i.e. the automatic recognition and assignment of single-word and multi-word terms (also in inflected form) in machine-readable texts, is in preparation.

Fig. 14

Calendar of meetings

The most important meetings on linguistic data processing, machine translation and terminological data banks since the beginning of 1975 have been the following:

- February 1975	International symposium on "Computer-assisted technical lexicography", Dresden
- March 1975	Meeting of the Leibniz Group (international research group for automatic translation), Lugano
- March 1975	Tutorial on Computational Semantics Lugano
- April 1975	Second international conference on computing in the humanities, Los Angeles
- April 1975	First symposium on international co-operation in terminology, Vienna
- May 1975	First national conference on the application of mathematical models and computers in linguistics, Varna
- June 1975	Demonstration of the Systran system in Bonn, Luxembourg and Zurich
- June 1975	Meeting of the Leibniz Group, Bonn
- August 1975	Fourth international congress of applied linguistics, Stuttgart
- October 1975	Meeting of the Leibniz Group, Grenoble
- November 1975	International seminar on machine translation, Moscow
- February 1976	Systran discussion, Bonn
- March 1976	Seminar of the Foreign Broadcast Information Service on machine translation, Washington D.C.
- March 1976	Systran Workshop, Luxembourg
- April 1976	Third European conference on cybernetics and system research, Vienna
- May 1976	Workshop on linguistics and information science, Stockholm
- May 1976	Meeting of the Leibniz Group, Brussels
- June 1976	Seminar on automatic translation, Luxembourg

- June 1976	International terminology seminar, Paris
- June/July 1976	Sixth international conference on computational linguistics, Ottawa
- September 1976	International symposium on "Automatic lexicography, analysis and translation", Saarbrücken
- October 1976	Workshop on "Advances in natural language processing", Amsterdam
- March 1977	Swiss conference on linguistic data processing, Zurich
- April 1977	Discussion of Systran-Titus III, Compiègne
- Spring 1977	Second symposium on international co-operation in terminology, Vienna
- May 1977	Third European congress on information systems and networks, Luxembourg
- June 1977	Eighth world translation congress, Montreal
- August 1977	Third international conference on computing in the humanities, Waterloo
- August-September 1977	Twelfth international congress of linguists, Vienna
- August 1978	Fifth international congress of applied linguistics, Montreal
- 1978	Second international seminar on machine translation, Moscow

Note

This list is anything but exhaustive. For instance, it does not mention the summer schools for computational linguistics the annual meetings of the Association for Literary and Linguistic Computing, the Association for Computational Linguistics, the Natural Language Translation Specialist Group of the British Computer Society, LDV Fittings, Verein zur Föa rderung der wissenschaftlichen linguistischen Datenverarbeitung e.v., the workshops on artificial intelligence etc.