[*Terminologie et Traduction*, no.1, 1986]

PROVISION AND USE OF RAW MACHINE TRANSLATION

F.W.A. Habermann
Kernforschungszentrum Karlsruhe, FRG

Abstract

Since 1979, efforts have been made to develop the French-English Systran system for providing raw machine translations of research papers concerned with fast breeder technology. Now that most of the necessary terminology for this field has been entered, considerable improvement in quality has been noted. Although further improvements are desirable, the system can already be successfully used to inform German and English scientists about the work of their French colleagues.

1. Introduction

The deployment of nuclear energy for electricity generation was considered such an important issue after the second world war that national research centers were set up in all industrial countries to develop large nuclear reactors for power generation on a commercial scale. The Nuclear Center of Karlsruhe (KfK) was founded in the late fifties and employs more than 3000 people. The main research project in the center is devoted to the development of the fast breeder reactor, which can extract 50 times more energy out of uranium than the light water reactors, which presently are used world wide for electricity generation. Not only Germany, but also the US, England, France and Japan have, for many years, been working on the development of the fast breeder reactor and *a* large amount of information is being exchanged between these countries. Particularly close cooperation has existed since 1977 between France and Germany with the result that we receive in Karlsruhe many technical reports written in French: up till now about 3000. It is not possible to have this number of reports routinely translated by our translators and for this reason we have investigated the possibility of applying Machine Translation.

In 1979 Systran München GmbH translated some texts for us from French into English after addition of 1500 technical words to the Systran dictionary. At the end of 1979, Systran gave a demonstration at Karlsruhe and the result was considered promising. In our opinion, however, the quality of the translations was not yet good enough to be utilized without postediting by the researchers who request translation. In order to reach this goal, further adaptation of Systran to nuclear technology texts was considered necessary. The Luxembourg branch of Systran made some effort in this direction in 1980 and since the beginning of 1982 systematic further development work has been being done at Karlsruhe in the frame of a cooperation agreement with the European Commission at Luxembourg. KfK has received the Systran program for translation from French into English from Luxembourg in exchange for feedback of technical vocabulary and general suggestions for improvement of the Systran program. Since then about 10,000 technical words have been entered in the dictionary in parallel with grammatical improvements of the Systran program, which correspondingly has been updated many times. These efforts have resulted in a substantial improvement of the quality of translated texts.

2. Objective

The aim of our pilot project is to demonstrate that fully automatic translation can be utilized for the transmission of scientific information. Linguistically this is a limited goal, because we do not request translation into perfect English.

From the viewpoint of information transfer, however, our objective is ambitious, because we intend to rely on machine translated texts without human corrections. Our restrictions and difficulties are summed up in the following and in figures 1 and 2 together with the chances which we see in MT.

Restrictions

- In the first place we have limited the application of MT to informative scientific texts
- Secondly we are concentrating on French-English translations with Systran
- Thirdly we do not offer post-editing of the raw MT output.

We leave it to the users to cope with the problem of checking raw MT output. Preferably they should seek assistance of a colleague in their field who knows some French and English.

Difficulties

Regardless of these restrictions we still have enough problems:

- Manual input of texts can be done without errors and in the proper format, but it is too slow if rapid translation of large amounts of text is wanted.
- Use of an Optical Character Reader is limited to a restricted number of typewriter characters and clean text copies. Formating of the output can be performed automatically with a computer program, correction of errors has to be done manually.
- Very few texts are available in machine readable form.
- Electronic transmission of texts is possible in principle, but in practice compatibility problems of equipment occur.
- Machine translation has not yet reached such a good standard that the transfer of the information contained in the source text is completely reliable.

Chances

One of the main requirements for machine translation to become successful is that it should be much faster than conventional human translation. To obtain insight into the speed of MT, the time needed for the various successive operations is listed in Figure 2.

- Manual input can be done at a rate of 3-5 pages per hour.
- Optical character reading by machine is much faster, but it is not free of errors.
- Correction of the input text might cost much more time; 10-20 pages/hour may be optimistic.
- Translation by the Systran program with 1,000 pages per hour is more than 100 times faster than human translation.
- Printing of the raw MT output is also very rapid.
- Post-editing on the contrary is a tedious and difficult task.

It appears that human operations are by orders of magnitude slower than machine operations. To obtain the full benefit of MT, it is therefore mandatory to limit human work to a minimum.

Another inherent advantage of translating with the aid of a computer is the instantaneous availability of the complete terminology needed for a specific field. Without this aid only professional translators with a long experience in the field are able to translate relevant texts correctly.

3. Translation Quality

We have from the very beginning of the application of MT in our field felt the need for criteria to judge the translation quality, our preoccupation being in particular to determine the rate of progress made by further development of the system. This would enable us to estimate if our goal to reach a quality of translated text sufficient for the personal use of scientific researchers can be reached within a reasonable time. We believe now that this is possible and it is shown in the following on which data this assertion is based.

Parameters for a quantitative determination of translation quality

The evaluation of the quality of a translation is very personal. A true linguist will appreciate a good style and is hurt if a language is maltreated. A scientist would accept all kinds of insufficiencies in style, both in respect of syntax and vocabulary, as long as the content of the original text is reproduced understandably in the other language. This transfer of information is our aim at Karlsruhe and this objective automatically leads to a first semiquantitative criterion of translation quality: determine in a text sample how many sentences are translated understandably. This criterion can be differentiated somewhat further by distinguishing between understandable, partly understandable and not at all comprehensible. This kind of analysis of a text is quite crude, but it is quantitative. Subsequent analyses of a text sample after updating of the translation program will indicate if and how much improvement was reached each time.

A somewhat more refined quantitative determination of translation quality can be achieved by counting errors in the text. In order to get at the same time some insight into where the main insufficiencies in the translation are situated I have distinguished in a statistical analysis between the following kinds of errors:

1. Input errors: a trivial, but very necessary distinction. In my opinion not more than one input error can be tolerated per page.

2. Dictionary errors, distinguished between general and technical vocabulary. Initially only technical vocabulary was added to the Systran dictionary, but it appeared soon that also general vocabulary was missing and needed to be added.

3. Syntactical errors, distinguished between:

* errors which do not influence the comprehensibility of the sentence
* sequence of words wrong, which may change the meaning of a sentence
* serious syntactical errors, such as

. words of the source text are deleted
. the meaning of the sentence has been altered
. the sentence becomes incomprehensible.

This definition of errors is not very sophisticated linguistically, in particular not the subdivision of syntactical errors. Nevertheless, it allows one to draw important conclusions concerning the effect of subsequent updates of the computer program on the translation quality.

Statistical analysis of translated text samples

We have used some typical French texts in the nuclear technology field as samples for statistical analysis (Figure 3). The first sample contains 163 sentences and 4870 words and was translated in May 1980. The next sample contains 142 sentences and 3480 words. It was translated five times with different versions of Systran in the course of 5 years. The first translation was made in September 1980, the second in December 1981, the third in February 1983, the fourth in January 1984 and the fifth in June 1985. The last sample contains 270 sentences and 8000 words. It was also translated five times at the same dates as the second sample. It must therefore be expected that the translation quality of these two samples evolves in a similar way, which enables one to check the validity and accuracy of the statistical method. The numerical results of the statistical analyses are given in the table of Figure 3. For easier comparison the data are plotted in the following figures in a number of histograms after normalisation to numbers of errors per 100 sentences.

The first histogram (Figure 4) shows the comprehensibility of subsequent translations. The number of understandable sentences is indicated by the lower blocks of the histogram. Initially 75 to 80% of the translated sentences were understandable. In 1985, more than 95% has become understandable. The dotted blocks indicate partly understandable sentences. The rest until 100% is incomprehensible. The number of incomprehensible sentences has drastically diminished from 6% in 1980 to 1% in 1985. We might hope to be able to eliminate the occurrence of total incomprehensibility completely in the future.

The second plot (Figure 5) shows the number of input errors per 100 sentences. This was reduced markedly after 1980 and then stayed constant, because the same input tape was used each time. The number of 6 errors per 100 sentences is interesting as an order of magnitude: it corresponds to only one input error per page, if we have 15 sentences of 3 lines per page. Such a high quality is required, otherwise the input errors would impair the quality of the translated text significantly. On the other hand it is relatively easy to locate input errors, because these usually give rise to "not found words" in the translation. After correction of the errors the translation can be repeated with a faultless input.

Figure 6 shows errors due to lack of general vocabulary. The number of not known common words has diminished steadily from 1980 till 1985. Apparently the entries made in the Systran dictionary have been very effective in improving the translation quality. The same conclusion can be drawn from the plot in Figure 7 which shows how many technical words were not known in the samples. Here the statistics indicate that this source of errors may disappear completely in the near future. The order of magnitude of 2 technical and 4 common words not known per 100 sentences corresponds to one word per page.

The situation concerning grammatical errors does not look quite as favourable. Figure 8 shows the tolerable grammatical errors. Although these do not influence the understandability it is at least irritating to note that they have not diminished significantly in number in the course of 5 years. According to the histogram of figure 9 the number of word order errors has at first increased and then diminished. Allegedly this error source can be reduced by further improvements in the Systran program. The amount of 8 errors per 100 sentences corresponds to somewhat more than 1 error per page.

Figure 10 shows the most serious grammatical errors. Their number has fortunately diminished significantly; it seems to approach asymptotically a level of about one error in three pages of text. Although this is a very good score no error of this kind can be neglected because it distorts the information contained in the source text.

The last histogram in Figure 11 shows all errors added together. The dotted lines indicate the total number of syntax errors; the vocabulary plus input errors are given by the top parts of the blocks. The tendency is clearly downward for both types of errors. This number is not approaching an asymptotic lowest value.

As a general conclusion from the statistical analysis of the sample texts made until mid 1985, I would in the first place state that a large improvement in the quality of the translations was reached within 5 years both in respect of missing vocabulary and in respect of syntactical insufficiencies. Secondly the statistical data do not indicate that a limit in quality has been reached beyond which the ongoing efforts would have little effect. It appears on the contrary quite meaningful to continue the work in the expectation of further significant improvements in translation quality.

5. Implementation

Handling

Figure 12 shows how we presently handle machine translation of reports at Karlsruhe.

- The French text is normally received in the form of a photocopied typewritten document.
- This document is read page by page by an Optical Character Reader.
- The output of the OCR is received by a Personal Computer and its format is converted into Systran input format. The text is scanned for spelling errors.

- The information is then transferred to our mainframe IBM, where the words occurring in the text are compared with the Systran dictionaries. If the text still contains mutilated words due to reading errors of the OCR these will show up in the "Not Found Word List" and can be corrected.
- Subsequently the corrected French text is translated into English by the Systran program and printed out in the same format as the source text.

For the convenience of the user the French original and the English translation are combined into one volume in such a way that the source text and the translated text appear side by side on adjoining pages. One page of a technical report is added after the figures as an example. The translation contains several typical errors.

Users Comments

Whenever a translation is requested we inform the users beforehand that we produce raw machine translations which are not 100% reliable. At the same time we ask them to return comments on the translations for further improvement of the MT program. Although the readers indeed point at incomprehensible sentences and serious errors they generally find the translations at least helpful for understanding the text. For checking purposes most readers find it indispensable to have the source text side by side with the translation.

Some typical general responses are the following:

> "As I do not know French well the presented translations in English are of great importance to me. The quality is so good that the information can be understood without any problems. I am very much in favour that these translations be continued".
>
> "Sufficient for a non French-speaking reader to get acquainted with the problems under discussion".

Several readers have returned the translations with detailed comments. They do expect that the detected serious errors will no longer occur in translations made with future updates of the Systran program.

FIG 1: MT WITH SYSTRAN AT KARLSRUHE NUCLEAR CENTER

OBJECTIVE: FULLY AUTOMATIC MACHINE TRANSLATION

LIMITATIONS:

- . INPUT: INFORMATIVE SCIENTIFIC TEXTS
  PROBLEM: CONVERSION OF TYPEWRITTEN TEXTS IN
  MACHINE READABLE FORM
- . TRANSLATION: NON-INTERACTIVE FRENCH-ENGLISH MT WITH SYSTRAN
  PROBLEM: RELIABILITY OF TRANSLATED INFORMATION
- . OUTPUT: WITHOUT POSTEDITING
  PROBLEM: CHECK IF INFORMATION OF SOURCE TEXT
  IS TRANSMITTED CORRECTLY

FIG 2: SPEED OF MT IMPLEMENTATION

		PAGES/HOUR
INPUT	. TYPIST	3-5
	. OCR	100
	CORRECTION OF OCR	10-20?
TRANSLATION		
	. SYSTRAN	1000
OUTPUT	. LASER-PRINTER	500
POSTEDITING		
	. TRANSLATOR	2-5

Figure 3 : Statistical Analysis of Systran Translation Samples

Statistical data of text samples			Comprehensibility sentences			Errors (number in text sample)							
S: number of sentences W: number of words			(+) understandable (o) partly underst. (-) incomprehensible			Input	Vocabulary: C: common word T: technical word			Syntax: G: tolerable O: word order GG: serious			
S	W	Date of transl.	(+)	(o)	(-)	I	C	T	total errors vocab.	G	O	GG	total gramm. errors
163	487o	May 8o	123	32	8	3o	62	69	131	68	44	61	173
142	348o	Sep 8o	118	16	8	21	34	14	48	13	16	36	65
142	348o	Dec 81	123	14	5	1o	28	14	42	12	21	17	5o
142	348o	Feb 83	128	1o	4	1o	12	5	17	9	28	1o	47
142	348o	Jan 84	13o	9	3	1o	3	3	6	16	21	1	38
142	3480	June 85	138	3	1	10	3	2	5	5	1o	1	16
27o	8ooo	Sep 8o	22o	32	18	53	89	38	127	16	39	37	92
27o	8ooo	Dec 81	24o	2o	1o	15	62	36	98	13	38	17	68
27o	8ooo	Feb 83	249	14	7	15	33	13	46	1o	43	12	65
27o	8ooo	Jan 84	251	13	6	15	2o	9	29	15	43	6	64
27o	8ooo	June 85	258	7	5	15	15	6	21	8	27	6	41

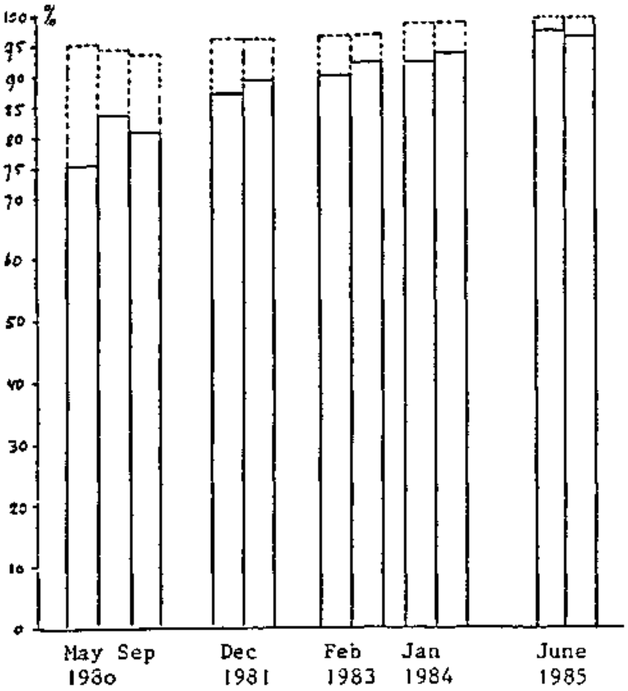


Fig. 4 : Comprehensibility

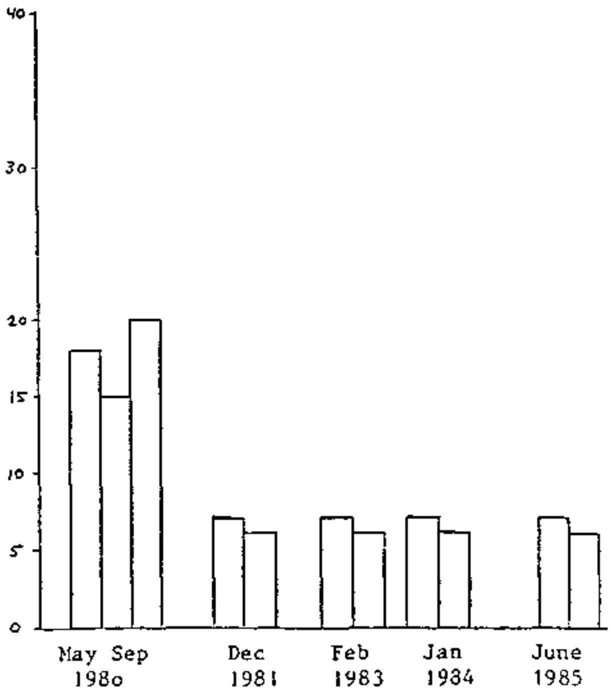


Fig. 5 : Input Errors

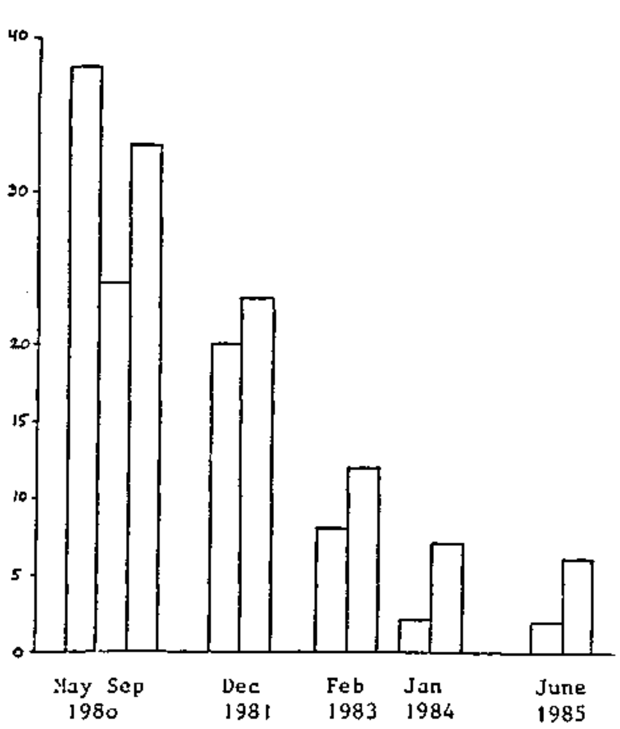


Fig. 6 : Common Word Errors

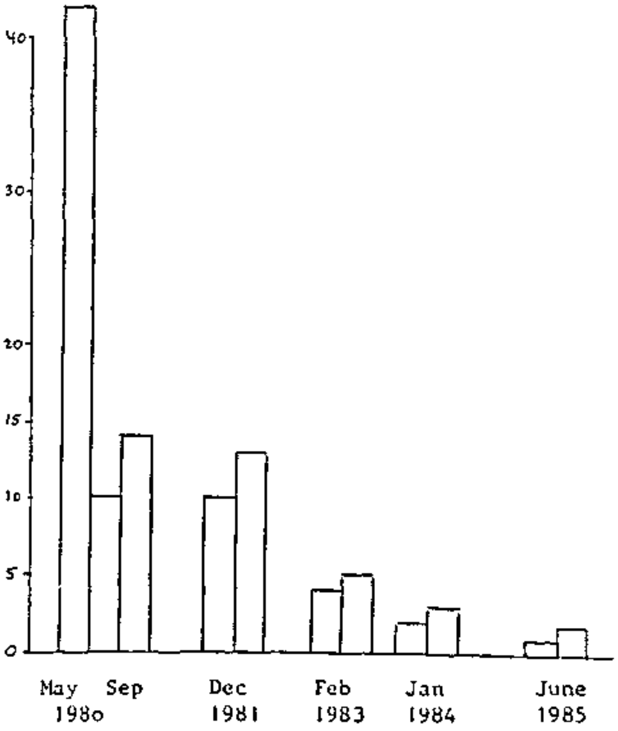


Fig. 7 : Technical Word Errors

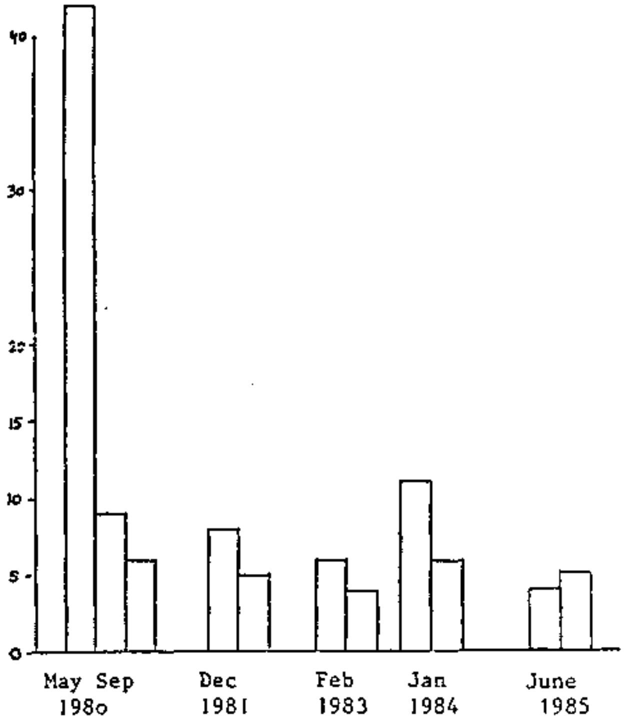


Fig.8 : Grammatical Errors

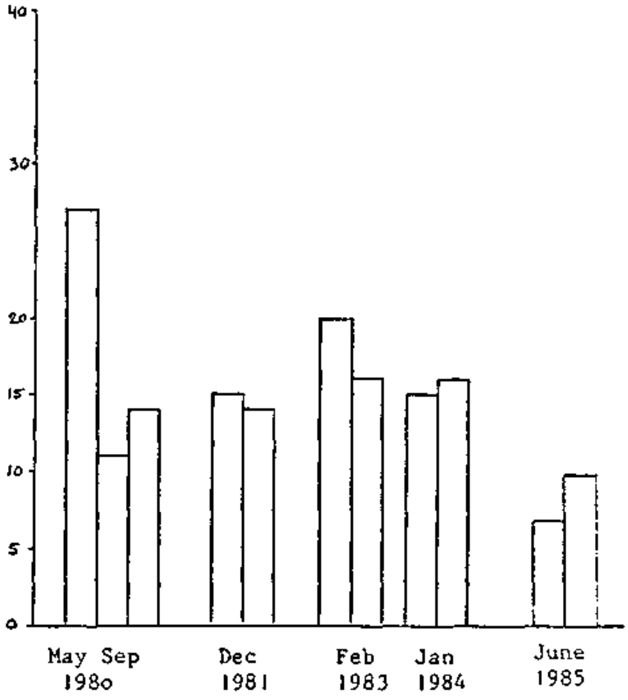


Fig.9 : Word Order Errors

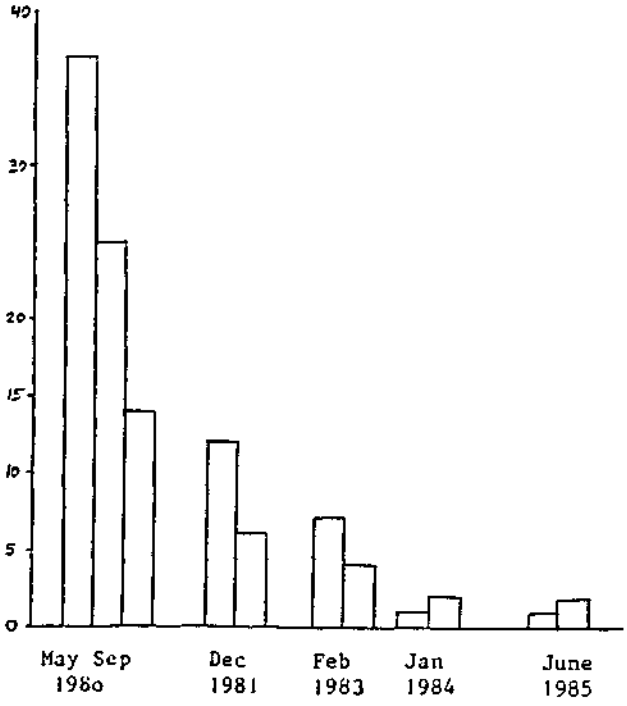


Fig. 10: Serious Grammatical Errors

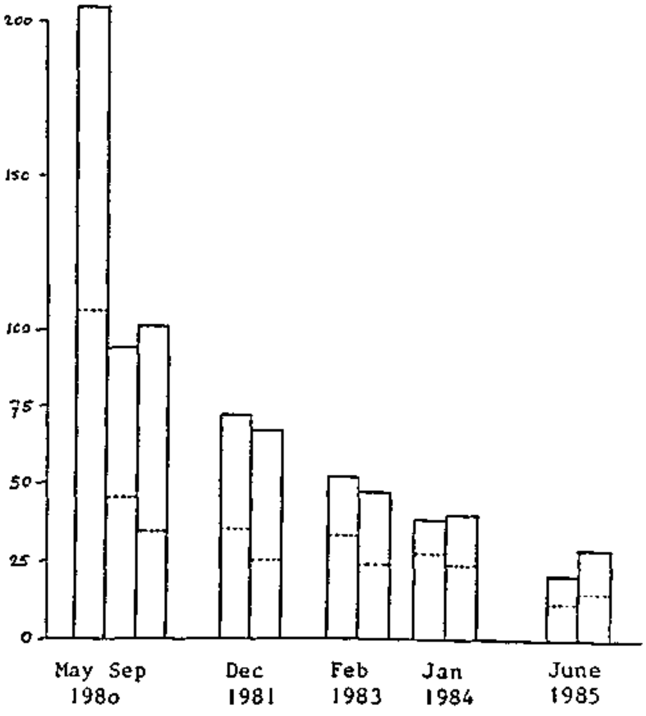
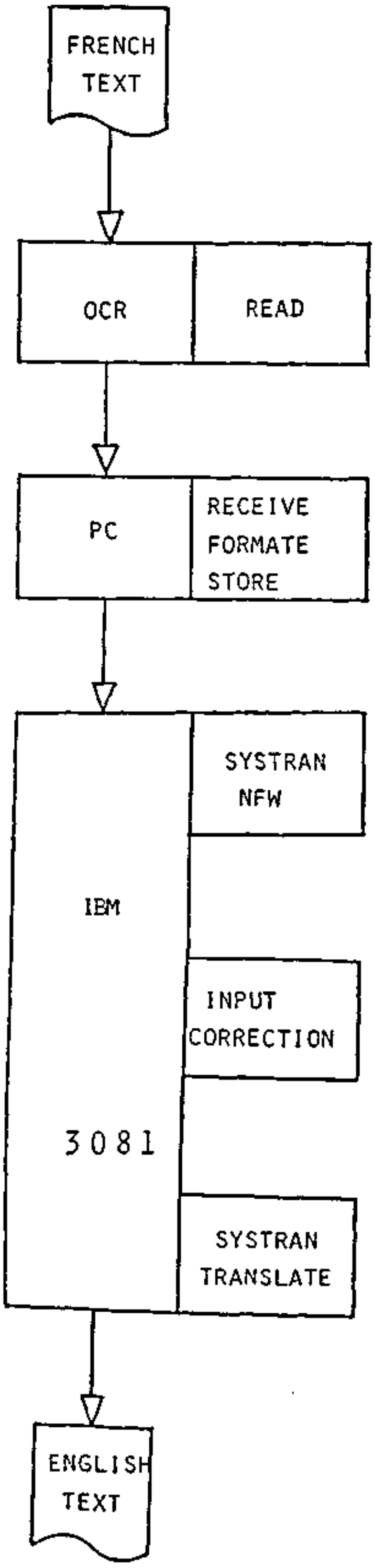


Fig.11 : Total Number of Errors per 100 Sentences

FIG 12: FLOW DIAGRAM OF MACHINE TRANSLATION

- INPUT WITH OCR
- TRANSLATION BY SYSTRAN



- TYPEWRITTEN SCIENTIFIC REPORT

- OCR = OPTICAL CHARACTER READER

- RECEIVE OCR OUTPUT
- CONVERT INTO SYSTRAN INPUT FORMAT
- STORE

- SYSTRAN DICTIONARY LOOK UP AND LISTING OF NOT FOUND WORDS

- MANUAL INPUT CORRECTION AT TERMINAL WITH HELP OF NOT FOUND WORD LIST

- FULL RUN OF SYSTRAN TRANSLATION PROGRAM

- SIDE BY SIDE PRINT OF FRENCH SOURCE TEXT AND ENGLISH TRANSLATED TEXT
- TRANSMISSION VIA DATA LINE POSSIBLE

This zone is in very strong correlation with the visible black ring by autoradiography, evidence of impoverishment in plutonium, which translates a regression of the melt front hanging the power bearing (/7/).

- On this sample is also observed a liquid oxide incursion through a crack, until the clad, without damage of it (Board 13).

- On the sample F, in longitudinal cut (235-265 mm / bcf) (boards 14 and 15) three tunnels were observed micrographiquement (enlargement 200):

- At the mm level 237 / bcf (A1A2), board 16, the zone which borders the cavity is porous although more dense than in periphery of the fuel. Side matter discontinuity is observed clearly probably reflecting the in core liquid oxide level.

- At the mm level 249 / bcf (B1B2), board 17, out of 200 microns starting from the edge of the cavity, the zone is very dense. Its size corresponds to that of the zone of the coarse grains revealed by chemical etching, on macrography. On some 700 microns broad ring starting from the external edge, the fuel does not seem not restructured. Between these two zones lenticular pores are observed on 500 microns at least.

- On the sample K, longitudinal cut (305-335 mm / bcf) of the pin 19 (board 23), in the low part a zone sequence is observed similar to the previous observations. In the high part, to the top of the bubble, one observe micrographiquement a localised very dense zone completely in fissile, and bordered top of column of lenticular pores, showing that the liquid oxide having reached the high part of the fissile column cooled and resolidifi3rd behaving then like a solid oxide. On this sample figure also the lower part of the upper U02 hold, perfectly intact contrary to that of the MELTED pin 6 of.

constellès de porosités allongées.Cette zone est en très bonne corrélation avec l'anneau noir visible par autoradiographie, indice d'un appauvrissement en Plutonium, qui traduit une régression du front de fusion pendant le palier de puissance (/7/).

-Sur cet échantillon est également observée une incursion d'oxyde liquide à travers une fissure, jusqu'à la gaine, sans endommagement de celle-ci (Planche 13).

-Sur l'èchantillon F, en coupe longitudinale (235-265 mm/bcf) (Planches 14 et 15) trois traversées ont été observées micrographiquement (grossissement 200) :

-Au niveau 237 mm/bcf (A_1A_2), Planche16, la zone qui borde la cavité est poreuse bien que plus dense qu'en périphérie du combustible.On observe nettement une discontinuité latérale de matière reflétant vraisemblablement le niveau d'oxyde liquide en pile.

-Au niveau 249 mm/bcf (B_1B_2), Planche 17, sur 200 microns à partir du bord de la cavité, la zone est très dense.Sa taille correspond à celle de la zone des gros grains révélée par attaque chimique, sur la macrographie.Sur une couronne de 700 microns de large à partir du bord externe, le combustible ne semble pas restructuré.Entre ces deux zones on observe des pores lenticulaires sur 500 microns au moins.

-Sur l'èchantillon K, coupe longitudinale (305-335 mm/bcf) de l'aiguille 19 (Planche 23), on observe dans la partie basse une séquence de zones semblables aux observations précédentes.Dans la partie haute, au dessus de la bulle, on observe micrographiquement une zone très dense localisée tout à fait en haut de colonne fissile, et bordée de pores lenticulaires, montrant que l'oxyde liquide ayant atteint la partie haute de la colonne fissile s'est refroidi et resolidifiè se comportant ensuite comme un oxyde solide.Sur cet èchantillon figure aussi la partie inférieure de la cale UO_2 supérieure parfaitement intacte contrairement à celle de l'aiguille 6 de FONDUE.

constellés de porosités allongées.Cette zone est en très bonne corrélation avec l'anneau noir visible par autoradiographie, indice d'un appauvrissement en Plutonium, qui traduit une régression du front de fusion pendant le palier de puissance (/7/).

-Sur cet échantillon est également observée une incursion d'oxyde liquide à travers une fissure, jusqu'à la gaine, sans endommagement de celle-ci (Planche 13).

-Sur l'échantillon F, en coupe longitudinale (235-265 mm/bcf) (Planches 14 et 15) trois traversées ont été observées micrographiquement (grossissement 200) :

-Au niveau 237 mm/bcf (A_1A_2), Planche16, la zone qui borde la cavité est poreuse bien que plus dense qu'en périphérie du combustible.On observe nettement une discontinuité latérale de matière reflétant vraisemblablement le niveau d'oxyde liquide en pile.

-Au niveau 249 mm/bcf (B_1B_2), Planche 17, sur 200 microns à partir du bord de la cavité, la zone est très dense.Sa taille correspond à celle de la zone des gros grains révélée par attaque chimique, sur la macrographie.Sur une couronne de 700 microns de large à partir du bord externe, le combustible ne semble pas restructuré.Entre ces deux zones on observe des pores lenticulaires sur 500 microns au moins.

-Sur l'échantillon K, coupe longitudinale (305-335 mm/bcf) de l'aiguille 19 (Planche 23), on observe dans la partie basse une séquence de zones semblables aux observations précédentes.Dans la partie haute, au dessus de la bulle, on observe micrographiquement une zone très dense localisée tout à fait en haut de colonne fissile, et bordée de pores lenticulaires, montrant que l'oxyde liquide ayant atteint la partie haute de la colonne fissile s'est refroidi et resolidifié se comportant ensuite comme un oxyde solide.Sur cet échantillon figure aussi la partie inférieure de la cale UO_2 supérieure parfaitement intacte contrairement à celle de l'aiguille 6 de FONDUE.

This zone is in very strong correlation with the visible black ring by autoradiography, evidence of impoverishment in plutonium, which translates a regression of the melt front hanging the power bearing (/7/).

- On this sample is also observed a liquid oxide incursion through a crack, until the clad, without damage of it
(Board 13).

- On the sample F, in longitudinal cut (235-265 mm / bcf) (boards 14 and 15) three tunnels were observed micrographiquement
(enlargement 200):

- At the mm level 237 / bcf (A1A2), board 16, the zone which borders the cavity is porous although more dense than in periphery of the fuel. Side matter discontinuity is observed clearly probably reflecting the in core liquid oxide level.

- At the mm level 249 / bcf (B1B2), board 17, out of 200 microns starting from the edge of the cavity, the zone is very dense. Its size corresponds to that of the zone of the coarse grains revealed by chemical etching, on macrography. On some 700 microns broad ring starting from the external edge, the fuel does not seem not restructured. Between these two zones lenticular pores are observed on 500 microns at least.

- On the sample K, longitudinal cut (305-335 mm / bcf) of the pin 19 (board 23), in the low part a zone sequence is observed similar to the previous observations. In the high part, to the top of the bubble, one observe micrographiquement a localised very dense zone completely in fissile, and bordered top of column of lenticular pores, showing that the liquid oxide having reached the high part of the fissile column cooled and resolidifi3rd behaving then like a solid oxide. On this sample figure also the lower part of the upper UO2 hold, perfectly intact contrary to that of the MELTED pin 6 of.