

ERROR ANALYSIS OF SYSTRAN OUTPUT - A SUGGESTED CRITERION FOR  
THE 'INTERNAL' EVALUATION OF TRANSLATION QUALITY AND A POSSIBLE  
CORRECTIVE FOR SYSTEM DESIGN

Professor F. Knowles  
Department of Modern Languages,  
University of Aston in Birmingham.

' "The only way!" It is my personal conviction that the machine translation system developed under your guidance and improved under government contract represents the only viable translation option available to the nation for servicing bulk translation requirements in the interests of national defense.'

'We feel that computerised machine translation (SIC) is not only feasible and economical but the only way to provide bilingual technical manuals. It provides us with a text which is completely acceptable for use by technicians.'

'...present users include the government of the USA, the government of Canada, the European Economic Community, and a range of industrial companies including some of North America's largest multinational corporations.'

The above statements are excerpts - slightly adapted, but not substantially altered - from SYSTRAN sales literature distributed by WTC of Canada Ltd.; the first statement comes from the United States Air Force, the second from General Motors of Canada Ltd., and the third is WTC advertising copy.

Against this background - indeed in spite of it - I should like to give some details of an "error analysis" which I have recently conducted on Russian texts translated into English by the SYSTRAN system. As I have said in a recent article, I cannot prove statistically that I was dealing with average texts or indeed with an average machine translation system. I will repeat myself by saying that I hope you will not be too scandalised if I say that I am intuitively satisfied that the texts I dealt with were not out of the ordinary and that the machine translation system which was responsible for the translation of those texts deserves attention firstly because it is one of the very few commercially operational machine translation systems in existence and secondly because the E.E.C. has made a considerable financial investment in it and is continuing to optimise the SYSTRAN system until such time as the proposed new EUROTRA system comes on line.

The particular texts that I analysed were made available to me by Professor F. Krückeberg and Mr D. Hoppe of the "Gesellschaft für Mathematik und Datenverarbeitung" in Bonn, Germany, and I should like to record my considerable thanks to these gentlemen for their generous assistance. This institution carried out during 1976, on the basis of a licenced agreement with the WTC of La Jolla (California), U.S.A., a number of tests in order to assess the performance and general capabilities of the

WTC-developed SYSTRAN machine translation system; these tests were carried out on the SYSTRAN software dealing with translation from Russian into English. My remarks in this paper are confined to SYSTRAN'S performance with this particular language pair. The corpus of sentences which I used in my researches was made up of two distinct categories of language data. On the other hand I investigated SYSTRAN'S performance at translating from Russian into English just over 2,000 sentences which had been used as examples in a pedagogical grammar of Russian written for German students.<sup>2</sup> Secondly, I had available four Russian technical texts containing subject matter on i) scales and weighing, ii) airports, iii) helicopters and iv) eyesight. The total number of sentences involved in this textual corpus was slightly under 500. These texts were part of a sample of texts chosen from the Great Soviet Encyclopaedia ("Bol'shaja Sovetskaja Enciklopedija"). It was felt by the GMD experts who initiated the original SYSTRAN tests that this source of textual material offered two advantages in a test situation. Firstly, the texts represent technical material of sufficient but not excessive terminological difficulty and they are, furthermore, adequately representative in so far as their grammatical complexities are concerned. Secondly, the original texts were carefully edited and have also been made available to readers of English by professional translators.

I present in Figure 1 what I have called 'raw' corpus characteristics. In Figure 2 on the other hand I show what I have chosen to call 'edited' corpus characteristics. The discrepancies in the number of tokens in the corpus are the result of editing procedures designed to eliminate number strings and to expand abbreviations. I generated from the 'edited' corpus alphabetic, reverse-alphabetic and descending frequency lists of the lexis occurring in the four technical texts. I also produced concordances to them and used all these materials to aid my error analysis. I have also included in Figure 2 three statistical indices common in the world of statistical linguistics, namely the logarithmic type-token ratio, the logarithmic lemma-type ratio and the index of vocabulary richness.

I carried out an analysis of variance test on the homogeneity of the mean sentence length across the four technical texts. The resulting F value was 1.48 and it is less than the tabular value of 2.60 for the 0.05% confidence level with the given degrees of freedom. On this basis I 'pooled' these thematically differing texts to form one 'technical text' corpus, statistically homogeneous at least from the point of view of sentence length and, in my subjective opinion, homogeneous in a number of other respects as well. By comparison an F test comparing the 'grammar' corpus with the 'technical text' corpus yielded the highly significant ratio of 22.71. I therefore draw a distinction between the two corpora in what follows.

My strategy in this particular piece of research was to scan all the sentences at my disposal, simultaneously noting and aggregating errors under various categories. The next step was to try to correlate the errors which were evident with the sequence of events in the processing of texts by the SYSTRAN MT system. I present in Figure 3 a simplified so-called 'sequence of events' on the basis of what I have been able to glean by reading published literature relating to the SYSTRAN software<sup>3-3</sup> and by discussing these matters with colleagues who have had the opportunity to probe more deeply into these matters or whose actual job it is to run SYSTRAN jobs and service the SYSTRAN system. I append a further figure, Figure 4, relating to SYSTRAN dictionary structure, given its crucial importance within the framework of the SYSTRAN software. In fact, the view is tenable that the success that SYSTRAN has achieved is due in larger

measure to the size and comprehensiveness of its data-base rather than to the inherent power of its processing algorithms.

I must emphasise that my 'results' are inferences. I have never seen - not for want of trying - the bank of SYSTRAN'S Russian dictionaries, neither have I ever seen the suite of SYSTRAN'S Russian-English analysis programs. I have seen - cursorily - SYSTRAN's English-French documentation but notwithstanding this I viewed SYSTRAN for the purposes of this investigation as a 'black box' for which I had input and output and about the inner workings of which I had but meagre information. I would claim, however, that this does enable me to make certain reasonable deductions about SYSTRAN's successes and, more importantly, its failures. One side-effect of this is, incidentally, to highlight the urgent need to overcome barriers standing in the way of collaboration, barriers which effectively prevent, so it seems, new ideas arising without from penetrating inwards, so to speak.

In Figure 5 I present my inventory of translation errors evident in the 'grammar' corpus, that is in the corpus of sentences excerpted from the above-mentioned pedagogical grammar. I must emphasise that the numbers and their accompanying percentages represent error tokens and not error types. In other words, they represent the total number of errors in that category. In Figure 6 I suggest some possible improvements to SYSTRAN software which in my view would have a pronounced enhancing effect; I classify this remark as a sort of 'long distance speculation'. I concede that the mechanisms required to implement suggestions of this sort have to be very finely balanced so as to lessen the risk of combinatorial explosion. Figure 7 presents a similar inventory of the errors which occurred in the 'technical text' corpus. I adhere to the same format as previously and likewise suggest in Figure 8 a small number of modifications which might eliminate, or at least drastically reduce the incidence of 'theoretically' avoidable errors. I am the first to admit, however, that there is what might be called a 'rump' involving in this case over 200 cruces which would be extremely difficult or costly, if not impossible, to solve programmatically. Investigation of these particular sentences by means of the hexadecimal print diagnostic facility would have helped enormously here but I did not have access to this information. I suspect, however, that the homograph disambiguation routines may be largely to blame.

Figure 9 gives details of what I call sentence success rate and it is apparent from this that the success rate is indeed extremely low, with errors occurring every four or five words. Note, however, that this last statement is somewhat misleading because a number of the errors occur at what might be called the supra-word level. I give in Figure 10 a checklist of the problems SYSTRAN appears to be suffering from in the realm of Russian morphological analysis and phrase-structure handling.

Most of SYSTRAN's errors catalogued above derive from a failure to implement functioning routines in a global and consistent fashion. It seems as if in many cases one salient or typical example has indeed been incorporated but that it stands alone like a prototype which never entered mass production. The answer to a given problem is often available yet the data needed by the problem-solving routines or by the dictionaries is either missing or is inadequate. To put it briefly: SYSTRAN, although deficient - in my opinion - at present, has in prospect a chance of giving a much better account of itself. Total consistency and utter perseverance would, I believe, go a long way and given SYSTRAN's modularity and open-endedness - and these are two of SYSTRAN's greatest attributes - it is of course possible to incorporate enhancements without undue process.

I should now like to present (see Appendix I) a number of instances of the SYSTRAN errors which I have categorised and briefly described in order to comment on them specifically. The format of this material is as follows. The first record is the original Russian text in SYSTRAN transliteration code. This code is:

а б в г д е ё ж з и й к л м н о п р с т у ф х ц ч ш щ ъ ы ь э ю я  
 а b v g d e 2 j z i l k l m n o p r s t u f x q c w 5 7 y 6 3 h 4

The second record is SYSTRAN's attempted translation and this is followed by the third record which is a correct English translation of the sentence involved. The fourth record - present only in the examples drawn from the 'technical text' corpus - is a gloss on the mistakes highlighted. It will be obvious that the sentences quoted often also contain other errors which are not commented on.

I take first errors encountered in the 'grammar' corpus; these remain without comment. I concede readily that the automatic translation of this material from Russian into English is a very tough test indeed because by definition the sentences involve the total range of, in this case Russian, grammar. A number of SYSTRAN's facilities such as the topical glossary system cannot be put to use in this case. Use might also be made, in dictionary refinement, of Zsorina's new and major Russian frequency dictionary.

Turning next to the 'technical text' corpus and addressing myself in a sense to the 'real world' of the professional translator. I give (in Appendix II) a further selection of SYSTRAN-generated translation errors, accompanied by explanatory notes.

Turning next to the corpus of technical texts and addressing myself in a sense to the real world I give a further short selection of what schoolmasters generally refer to as 'howlers'.

I should like to close this paper by saying that I requested four subject specialists to give me their expert, albeit subjective, assessment - as a percentage - of SYSTRAN's success in translating materials in their specialist field. I also asked them to give me a subjective percentage assessment of how much their own specialist knowledge had in fact helped them to comprehend the machine translated text. The text on 'scales and weighing' was assessed by Mr O.S. Nicholson, a lecturer in metallurgy specialising in precious metal assay. The text on airports was assessed by Professor E. Edwards, an expert on Applied Psychology and a person involved in investigating psychological factors affecting the performance of airline pilots. More to the point, however, he is a former RAF officer and is a very active amateur aviator of considerable experience. The passage on helicopters was assessed by Lt.C. Wrighton RN, a serving Royal Navy helicopter pilot. The text on eyesight was assessed by Mr D. Farrall, a lecturer specialising in and carrying out clinical work in ophthalmic optics. The results of these inquiries were:

	Intelligibility	Contribution of specialist knowledge
'Scales'	50%	75%
'Airports'	60%	75%
'Helicopters'	30%	65%
'Eyesight'	25%	70%

I am sure that we would all agree that we cannot rely on the reader's knowledge or on his good will to this extent. At one point the translation of the 'scales' text was stated by the expert reader to be seriously misleading and potentially dangerous if the reader should attempt to carry out one of the procedures in the way it is described by the 'machine' !

If this is so, then there ought to be no question of letting SYSTRAN loose, that is to say, letting it off its leash. Rather it appears necessary to have a human reviser holding the leash tightly. However, this of course knocks out one of the system's main pit-props and vitiates many of the claims made about SYSTRAN's translation performance and its throughput.

I must now make a statement that might well appear paradoxical. I state that SYSTRAN does appear to have achieved a performance level which is better than any other MT system has attained, and that I therefore owe it my respect on this account. I cannot in fact conclude this paper without revealing my admiration for SYSTRAN's data-processing sophistication. I hazard the guess that SYSTRAN may be suffering because a lot of the linguistics 'know-how' in the system was put there by 'linguist-programmers' who are now at one remove, having been replaced by 'systems programmers'.

I believe, as I said above, that a good many - but not all - of the pieces in the jig-saw puzzle of the overall strategy for computerised language analysis are already in their correct places.

In summary, what I have been trying to say in this paper is that SYSTRAN's best efforts are being in part frustrated, firstly, by deficient language data in its data-base and secondly, by the fact that some areas of potentially crucial, or - at the very least - promising 'know-how' in semantics have not found their way into SYSTRAN's 'architecture', or have not made their presence felt. I refer to 'tools' such as statistically weighted sub-language glossaries<sup>12</sup>, thesaurus methods for disambiguation<sup>13</sup> or lexeme coding techniques<sup>14</sup>, for instance.

I should like to thank Margaret Masterman for discussing with me many of the issues touched upon in this paper. Her comments were always both willingly given and illuminating and I am indebted to her; I accept responsibility, of course, for all the shortcomings of this paper which I hope, nonetheless, may be of some interest and use.

Figures 1-10 and Appendices I and II

Error Analysis of SYSTRAN output - a suggested criterion for the 'internal' evaluation of translation quality and a possible corrective for system design.

## 'RAW' CORPUS CHARACTERISTICS

	Sentences	Words	Mean Sentence Length	Standard Deviation
'Scales'	124	2476	19.968	10.314
'Airports'	105	2121	20.200	10.886
'Helicopters'	76	1744	22.947	14.603
'Eyesight'	162	3230	19.938	9.802
Text Corpus	467	9571	20.495	11.100
'Grammar'	2034	10868	5.348	2.329
Total Corpus	2501	20439	8.176	7.889

Figure 1.

'EDITED' CORPUS CHARACTERISTICS

	Tokens	Types	LTTR <sup>1</sup>	Lemmata	LTTR <sup>2</sup>	VR <sup>3</sup>
'Scales'	2383	1133	0.904	449	0.868	0.653
'Airports'	2069	1116	0.919	448	0.870	0.702
'Helicopters'	1720	886	0.911	355	0.865	0.663
'Eyesight'	3251	1571	0.910	627	0.875	0.705
Total 'Technical Text' Corpus	9423	4010	0.907	1879	0.909	0.856

Comparative Table for 'Grammar' corpus

	Tokens	Types	LTTR <sup>1</sup>	Lemmata	LTTR <sup>2</sup>	VR <sup>3</sup>
'Grammar' Corpus	10868	2034	0.820	950 <sup>§</sup>	0.900 <sup>§</sup>	0.511 <sup>§</sup>

§ = estimate

1 LTTR = logarithmic type-token ratio

2 LTTR = logarithmic lemma -type ratio

3 VR = vocabulary richness\*

$$VR = (25.0 - Tk + L + - \alpha) / 15.0,$$

where : Tk = tokens, L = lemmata and  $\alpha$  is a constant  
( value : - 0.172 )

10

\* Note: This is a modification of Muller's 'index of lexical richness'<sup>10</sup>.

Figure 2.

SYSTRAN EXECUTIVESimplified 'sequence of events'

- A Input text;
- B Scan for high frequency words  
and append information from HF dictionary;
- C Scan for idioms  
and append information from ID;
- D Alphabetise, access Master Stem Dictionary,  
and append information from MSD;
- E Return 'words' to original text order;
- F Disambiguate homographs;
- G Invoke Limited Semantics disambiguation routine, using Topical  
Glossary parameters as necessary;
- H Invoke Structural Pass 0  
- i.e., establish primary, 'global' syntactic dependencies, 'immediate  
environments' etc.;
- I Invoke Structural Pass 1 ('right-to-left')  
-i.e., recognise syntactic structures such as congruence, agreement,  
apposition, partly by using punctuation marks as 'clues';
- J Invoke Structural Pass 2  
-i.e., establish categorisation of subordinate clauses according to  
'initiators';
- K Invoke Structural Pass 3  
-i.e., establish Subject-Predicate relationships and embeddings;
- L Invoke Structural Pass 4  
-i.e., clear out 'rag-bag' of current unsolved problems  
(Human interaction possible, if needed.)
- M Insert 'articles', wherever necessary;
- N Resolve details of prepositional dependencies;
- O Access Limited Semantics File, if needed;
- P Translate prepositions;
- Q Access Lexical Programs, and switch on routines for semantic  
disambiguation via conceptual calculus;
- R 'Invert' appended information records to produce target language units;  
and execute accompanying instructions for TL Synthesis
- S Re-arrange elements to suit TL structuration;
- T Produce preliminary printout (only if desired for immediate analysis);
- U Printout results and channel to photo-composer.

Figure 3.



SYSTRAN DICTIONARY STRUCTURE**A Simplified Part Record Structure in Master Stem Dictionary  
(144,000+ stem entries reported at present in Russian MSD)**

- I First Grammar Subpart
  - i) syntagmatic information, frequency characteristics, homograph flag etc.
  - ii) Part of speech codes
  - iii) Basic part of speech codes
  - iv) Paradigmatic set codes
  - v) Gender, number, case, person, tense info.
- II Second Grammar Subpart ('Fillmore-type' case information/'valency')
  - i) case requirements
  - ii) prepositional requirements
  - iii) syntactic function codes
- III Third Grammar Subpart (case/preposition codes)
  - i) instrumental codes
  - ii) dative case codes
  - iii) prepositional translation control bytes
- IV Fourth Grammar Subpart
  - i) Topical glossary code
  - ii) English synthesis codes
  - iii) English meaning(s)

**B The Russian Idiom/Limited Semantics (LS) Dictionary is reported to contain at present:**

- a) 3,300+ 'declinable' idioms and (technical) collocations.
- b) 160,000+ 'declinable' idioms and (technical) collocations.

Figure 4.

'GRAMMAR' Corpus Errors

(errors sums refer to error 'tokens', not to error 'types')

A. Input errors:		
i) instances of incorrect or 'unusual Russian'	- 13	(0.43%)
ii) data capture errors	- 215	(7.18%)
B. High-frequency Dictionary Failure:		
i) item missing from the HF Dictionary	- 20	(0.67%)
C. ID/LSD failure:		
i) item missing from the Idioms Dictionary	- 158	(5.28%)
ii) item missing from list of general collocations	- 110	(3.67%)
D. Master Stem Dictionary failure:		
i) 'routine word' not in dictionary	- 258	(8.61%)
ii) abbreviations not in dictionary	- 1	(0.03%)
iii) name not in dictionary	- 20	(0.67%)
iv) morphological analysis data not in dictionary	- 193	(6.44%)
v) exhaustive valency information not in dictionary	- 333	(11.12%)
F. Homograph disambiguation		
i) homograph not disambiguated	- 36	(1.20%)
G. 'LSD'		
i) item not disambiguated by conceptual calculus	- 53	(1.77%)
H. 'STRPASS0'		
i) anaphora failure	- 415	(13.86%)
ii) Russian ellipsis not 'compensated'	- 20	(0.67%)
I. 'STRPASS1'		
i) number agreement failure	- 17	(0.57%)
ii) 'negation' not translated	- 3	(0.10%)
J. 'STRPASS2'		
i) word class not recognised	- 9	(0.30%)
K. 'STRPASS3'		
i) copula not 'found'	- 83	(2.77%)
L. 'STRPASS4'		
i) parse abort	- 262	(8.75%)
M. Articles		
i) article usage incorrect	- 215	(7.18%)
N. Prepositional Dependencies		
i) syntagma unrecognised	- 51	(1.70%)

Figure 5. (a)

P. Preposition Translation	
i) incorrect English preposition	- 99 (3.31%)
R. English synthesis	
i) incorrect English tense form	- 48 (1.60%)
ii) incorrect English voice form	- 31 (1.04%)
S. Element order	
i) incorrect English element order	- 292 (9.75%)
I. Number of sentences rating as 'garbage'	- 218 (7.28%)
II. Number of 'howlers'	- 50 (1.67%)
III. Errors made by 'competing' human translator	- NA

Figure 5. (b)

Possible Improvements

- A. Implementation of 'Wilks-type' general semantic calculus <sup>11</sup> would have inhibited 259 errors under Hi) above, (a 62% improvement).
- B. Insertion of full valency information in MSD would have averted 333 out of 805 dictionary failures, (a 41% improvement).
- C. A 'theme-rheme' structure recogniser, based on fairly simple Russian syntactical patterns, would have eliminated 162 of the 215 errors in article usage, (a 75% improvement).
- D. Altering the transliteration system to avoid difficulties with the 'word' '4' (meaning 'I') would have by-passed some 200 data capture errors arising from confusion of this 'word' with the number 'four'.
- E. Insertion of missing items into dictionaries.

Figure 6.

'TECHNICAL TEXT' CORPUS ERRORS

(errors sums refer to error 'tokens', not to error 'types')

A. Input errors:		
i) instances of incorrect or 'unusual' Russian	- 8	(0.39%)
ii) data capture errors	- 2	(0.10%)
B. High-frequency Dictionary failure:		
i) item missing from the HF dictionary	- 20	(0.98%)
C. ID/LSD failure:		
i) item missing from the Idioms Dictionary	- 5	(0.25%)
ii) item missing from list of technical collocations	- 323	(15.85%)
iii) item missing from list of general collocations	- 6	(0.29%)
D. Master Stem Dictionary failure		
i) 'routine' word not in dictionary	- 210	(10.30%)
ii) abbreviation not in dictionary	- 14	(0.69%)
iii) name not in dictionary	- 16	(0.79%)
iv) exhaustive valency information not in dictionary	- 69	(3.39%)
G. 'LSD'		
i) Topical Glossary information not available	- 221	(10.84%)
H. 'STRPASS0'		
i) anaphora failure	- 19	(0.93%)
ii) Russian ellipsis not compensated	- 29	(1.42%)
I. 'STRPASS1'		
i) number agreement failure	- 34	(0.63%)
J. 'STRPASS2'		
i) word class not recognised	- 33	(0.61%)
K. 'STRPASS3'		
i) copula not 'found'	- 21	(0.39%)
L. 'STRPASS4'		
i) parse abort	- 85	(1.56%)
M. Articles		
i) article usage incorrect	- 557	(10.24%)
N. Prepositional dependencies		
i) syntagma unrecognised	- 3	(0.06%)
P. Preposition translation		
i) incorrect English preposition	- 161	(2.96%)
R. English synthesis		
i) incorrect English tense form	- 7	(0.13%)
ii) incorrect English voice form	- 29	(0.53%)
S. Element order		
i) incorrect English element order	- 165	(3.03%)

Figure 7.

continued...

Figure 7. (continued)

I. Number of sentences rating as 'garbage'	- 63 (1.16%)
II. Number of 'howlers'	- 44 (0.81%)
II. Errors made by 'competing' human translator	- 16 (0.29%)

Possible Improvements

- A. Topical Glossary ( including weighted hierarchy of topical glossaries, a sort of 'dynamic thesaurus') could give an estimated 80% improvement under G.
- B. Insertion of correct and exhaustive valency information could give an estimated 50% improvement under D.
- C. 'Theme-rheme' routine could give an estimated 60% improvement under M.
- D. 'Wilks-type' semantic calculus<sup>11</sup> could give an estimated combined 25% enhancement under H, K and N.
- E. Insertion of missing dictionary data.
- F. Insertion of an 'a/an' option routine.
- G. Note: There were 231 cases involving an 'awkward' choice of English synonyms. Automatic procedures to produce the 'right' answer would appear either to require massive, unproductive overheads, or to be largely elusive.

Figure 8.

SENTENCE 'SUCCESS RATE'

	Sentences Correct	Errors	Errors/Sentence	Error every n'th word
'Scales'	3/124 (2.42%)	384	3.10	6
'Airports'	1/105 (0.95%)	523	4.98	4
'Helicopters'	0/ 76 (0.00%)	374	4.92	5
'Eyesight'	0/162 (0.00%)	757	4.67	4
'Technical Text' Corpus	4/467 (0.86%)	2038	4.36	5
'Grammar' Corpus	150/2034 (7.37%)	2955	1.45	4
Total Corpus	154/2501 (6.16%)	4993	2.00	4

Figure 9.

CHECK-LIST OF RUSSIAN MORPHOLOGY DETAILSCAUSING ANALYSIS PROBLEMS IN SYSTRAN

1. use of non-animate accusative plural for animate nouns
2. use of predicative instrumental
3. 'postponed' adjectives
4. adjectival participles
5. comparative forms in '-ej'
6. use of the prefix 'po-' in comparatives
7. ambiguity of comparative adjective/adverb
8. use of 'sam/samyj'
9. use of compound numerals
10. imperatives
11. use of imperative infinitive (with/without 'by')
12. lack of 'sequence of tenses'
13. separable function words ('ničto' etc.)
14. use of '-ka'
15. uses of 'kak'
16. uses of 'čtoby'
17. use of interrogatives
18. inversion after direct speech
19. 'telling the time'
20. expressing dates

Figure 10.



APPENDIX ISample of translation errors from the 'Grammar' Corpus

A.M. Gor6ki1 dolgoe vrem4 projil Na solnecnom skalistom Kapri.  
A.M. Gor'kiy long time lived on solar rocky Kapri.  
A.M. Gorky lived on the sunny, rocky isle of Capri for a long time.

My trebuem mira.  
We require world.  
We demand peace.

Injener vstupil v cleny naucno-texniceskogo ob5estva.  
An engineer entered the members of a scientific and technical society.  
The engineer became a member of the Science and Technology Society.

Novye zdani4 MGU raspolojeny Na Leninskiz gorax.  
New buildings MGU - Moscow State University are arranged on Leninist mountains.  
The new buildings of the MGU (Moscow State University) are in the Lenin Hills district of Moscow.

Turisty zoteli podn4t6s4 Na samuh vysokuh verwinu.  
Tourists wanted to rise to the highest apex.  
The tourists wanted to ascend the tallest peak.

Nam nujno mnogo ugl4, jeleza, 3lektro3nergii.  
To us much coal is necessary, gland, electric power.  
We need a lot of coal, iron and electricity.

Pionerski1 lager6 Drujba naxodits4 Na samom beregu nebol6wogo ozera.  
A pioneer camp friendship is located on the very shore of a small lake.  
The 'Friendship' Pioneer Camp is on the very shore of a small lake

On dopisal straniqu i otlojil rucku v storonu.  
It wrote a page and put off a knob to the side.  
He finished writing the page and laid his pen aside.

Raki plavaht.  
Cancers sail.  
Crabs can swim.

Letom Vera vseгда nosit svetlye plat64.  
In summer faith always bears bright clothes.  
In summer Vera always wears bright clothes.

Na drugo1 den6 cleny delegaqii poleteli v Kiev.  
Next day the terms of delegation flew to Kiev.  
On the following day the members of the delegation flew to Kiev.

Vcera my qelyl cas katalis6 na lodke.  
Yesterday we the entire hour rolled themselves on a boat.  
Yesterday we went out boating for a whole hour.

Rabocie, stro45ie gidro3lektrostanqih, selcas otdyxahť v klube.  
Workers, that construct a hydroelectric power plant, rest now in a cloud.  
The workers who are building the hydroelectric power station are now  
relaxing in the club.

Prosypa4s6, on selcas je soscacival s krovati i vklhcal radiopriemnik.  
Being spilled, it now soscachival from bed included a radio receiver.  
On waking up he used to immediately jump out of bed and turn on the  
wireless.

K zavtraku nam podali 4lqa vsm4tku i kofe po-tureqki.  
To breakfast to us they fed an egg soft-boiled and a coffee in Turkish.  
For breakfast they gave us soft-boiled eggs and Turkish coffee.

V univermag nepreryvno vxod4ť pokupateli.  
In a department store continuously enter buyers.  
Customers are constantly entering the department store.

Zavtra my poedem v Kiev.  
Tomorrow we will go into Kiev.  
Tomorrow we shall go (travel) to Kiev.

Na uliqe bol6woe dvijenie.  
On a street large motion.  
There is a lot of traffic in the street.

Ona navarila 5el na neskol5ko dne1.  
It welded on cabbage soups on several days.  
She cooked enough cabbage soup for several days.

Ne mewalte nam!  
Do not interfere with us!  
Don't disturb us!

On uxajivaet za nel.  
It handles it.  
He is courting her.

On razvels4 s jenol.  
It was bred with a wife.  
He has divorced his wife.

U devocki dlinnye kosy.  
The girl has long scythes.  
The girl has long plaits.

Spasibo, bol6we ne xocu.  
I no longer want thanks.  
No more for me, thank you.

Byt6 dojdh; byt6 bure.  
To be to rain; to be an auger.  
It looks like rain; it looks as if there's going to be a storm.

Davalte zajjem svet!  
Give let us light up the world!  
Let's put the light on.

Cert voz6mi!  
Features take!  
Damn!

APPENDIX IISample of translation errors from the 'technical text' corpus.

Turbovintovoi dvigatel6 (firmy "Djeneral 3lektrik \*) mo5nost6h 2500 kvv (3400 l.s.) pozvolil dostic6 maksimal6no1 skorosti pol2ta u zemli 408 km/c pri maksimal6no1 pol2tno1 masse 7700 kg.

A TURBOPROP ENGINE (FIRMS "GENERAL ELECTRIC") WITH A POWER 2500 KW (3400 L.S.\*) MADE IT POSSIBLE TO REACH THE MAXIMUM FLIGHT SPEED OF \*\* THE EARTH 408 KM/H WITH THE MAXIMUM FLIGHT MASS 7700 KG.

The turboprop engine ( from the General Electric company) with a power of 2,500 kw (3,400 hp) made it possible to achieve a maximum flying speed of 408 km/hr near the ground, with a maximum flying weight of 7,700 kg.

\* Russian abbreviation not recognised.

\*\* Preposition translated 'disastrously'.

(The computer does not possess 'knowledge of the world' and cannot detect the falsity of its 'statement' about the Earth's velocity of flight.)

Stroitel6stvo a3rovokzalov polucilo znacitel6noe razvitie posle 2-1 mirovol volny v sv4zi s soverwenstvovaniem, obnovleniem parka passajirskix samol2tov, a v SSSR - osobenno posle 1958, s vvodom v 3kspluatacih skorostnyx mnogomestnyx samol2tov Tu-104, Il-18, An-10, Ty-114.

BUILDING AIR TERMINALS OBTAINED A SIGNIFICANT DEVELOPMENT AFTER THE 2ND WORLD WAR IN CONNECTION WITH IMPROVEMENT AND RENOVATION OF THE PARK OF PASSENGER AIRCRAFT, BUT IN THE USSR - ESPECIALLY AFTER 1958, WITH THE INPUT INTO OPERATION OF HIGH-SPEED MULTIPLACE AIRCRAFT THAT-104\*. SILT-18,\* AN-10, THAT-114.\*

The construction of airport terminals greatly expanded after World War II with the improvement in the quality and the increase in the size of passenger aircraft; in the USSR this took place especially after 1958, with the commissioning of the high-speed and large-capacity planes Tu-104, Il-18, An-10, and Tu-114.

\* Russian acronyms (for aircraft designations) not recognised - the computer attempts to translate these 'stumps'.

A3roport imeet bazy mexanizaqii, transporta, texniceskie i xoz4lstvennyye sklady, razlicnyye slujebnyye zdani4, kompleksy injenernyx setel, soorujenil, obespecivah5ix ego vodosnabjenje, kanalizaqih, teplo-, gaxo - i 3lektrosnabjenje.

AN AIRPORT HAS BASES OF MECHANIZATION AND TRANSPORT, TECHNICAL AND ECONOMIC STORAGES, DIFFERENT OFFICIAL BUILDINGS AND COMPLEXES OF UTILITY NETWORKS AND CONSTRUCTION, WHICH ENSURE ITS WATER SUPPLY, CHANNELIZATION,\* THERMO-, GAS- AND A POWER SUPPLY.

An airport has mechanization and transportation bases; technical and other warehouses; and a variety of service buildings, engineering networks, and facilities for water supply, sewerage, heat, gas, and electric power.

\* False information resident in dictionary.

U presmykah5ixs4, isklhca4 gatterih i cerepax, ptiq, isklhca4 kivi, v steklovidnoe telo. vda2ts4 ot mesta vxoda zritel6nogo nerva xarakterny vyrost, obil6no snabj2nnyl sosudami, - greben6.

OF THOSE WHICH ARE CREEPING \*, EXCLUDING GATTERIYU AND TORTOISES, AND BIRDS, EXCLUDING A KIWI, INTO A VITREOUS BODY THE CHARACTERISTIC APOPHYSIS, ABUNDANTLY EQUIPPED WITH VESSELS, SUBMERGES FROM THE PLACE OF THE ENTRANCE OF OPTICAL NERVE A CREST.

In reptiles, except for the tuatara and tortoises, and in birds, except for the kiwi, a characteristic process abundantly supplied with blood vessels - the pecten - protrudes into the vitreous body from the point of entry of the optic nerve.

\* The technical term for 'reptiles' not in the dictionary. The computer has parsed the participial form of the appropriate Russian verb for which the dictionary contains incorrect information. The verb involved now means not 'to creep/crawl' (the etymology of the word 'reptile'), but 'to grovel'.

S rostom prot4j2nnosti vozduwnyx trass i osvoeniem novyx tipov samol2tov povywahts4 trebovani4 k oborudovanih i osna5enih a3roportov.

THE REQUIREMENTS FOR EQUIPMENT AND EQUIPMENT\* OF AIRPORTS RISE WITH AN INCREASE OF THE EXTENT OF AIR ROUTES AND THE MASTERY OF NEW AIRCRAFT TYPES.

As the length of air routes increases and as new types of aircraft are put into operation, the requirements for the equipment of airports increase.

\* The computer is 'unaware' that it has used the same English word twice as equivalents of two different Russian words.

Sovremennyy a3roport predstav4et sobol slojny1 kompleks inženernyx soorujen1, texniceskix sredstv, dl4 razme5eni4 kotorogo trebuets4 territori4, izmer4ema4 v otdel6nyx sluca4x tys4cami gektarov (naprimer moskovskil a3roport Domodedovo, H6h-lorckil a3roport Kennedi).

A CONTEMPORARY AIRPORT IS THE INVOLVED COMPLEX OF ENGINEER CONSTRUCTIONS AND TECHNIQUES, FOR ARRANGEMENT OF WHICH THE TERRITORY, MEASURED SOMETIMES IS REQUIRED BY THOUSANDS OF HECTARES (FOR EXAMPLE THE MOSCOW AIRPORT DOMODEDOVO, KENNEDY'S NEW YORK AIRPORT).\*

The modern airport is an elaborate complex of engineering structures and technical devices requiring a large territory, which, in some cases, measures thousands of hectares (for instance, Domodedovo Airport in Moscow or Kennedy Airport in New York).

\* Collocation or 'hyperword' not in dictionary.

Razlicaht monokul4rnoe zrenie (odnim glazom), binokul4noe, kogda pol4 zreni4 dvux glaz casticno perekryvahts4.

THEY DISTINGUISH MONOCULAR SIGHT (WITH HALF AN EYE)\* AND BINOCULAR, WHEN THE FIELDS OF VIEW OF TWO EYES OVERLAP PARTIALLY.

One may distinguish monocular and binocular vision. In monocular vision, one eye functions; in binocular vision, the visual fields of the two eyes partially overlap.

\* Item translated as idiom instead of literally.

Razrabotka arxitekturno-planirovocnyx sxem a3roportov predusmatrivaet naibolee raqional6noe socetanie zon-l2tnol, slujebnol i jilol; pri 3tom kompozicijonnyx qentrom 4vl4ets4 a3rovokzal vmeste s drugimi ucastkami slujebnol zony, neprosredstvenno sv4zannymi s obslujivaniem passajirov.

THE DEVELOPMENT OF THE ARCHITECTURAL-PLANNING DIAGRAMS OF AIRPORTS PROVIDES FOR THE MOST RATIONAL COMBINATION OF ZONES - FLYING, OFFICIAL AND BY A VEIN;\* IN THIS CASE IS A COMPOSITION CENTER THE AIR TERMINAL TOGETHER WITH OTHER SECTIONS OF AN OFFICIAL ZONE, DIRECT-CONNECTED WITH CATERING OF PASSENGERS.

The architectural and planning schemes of airports provide for the most rational combination of flight, service, and residential zones. The central element of its composition is the air terminal and other service zones directly connected with the servicing of passengers.

\* Grammatical class of word not recognised. False parse route leads to wrong dictionary entry.

Obslujivanie a3roporta vklhcaht v ob5uh sxemu dvijemi4 transporta v gorode i prigorodnom ralone.

THE MAINTENANCE OF AN AIRPORT IS CONNECTED IN THE COMMON FLOW CHART\* OF TRANSPORT IN A CITY AND SUBURBAN AREA.

The servicing of the airport is included in the general plan for urban and suburban traffic.

\* Topical glossary items not differentiated.

The Russian word involved means 'flow chart' only in mathematics or computer science texts - its 'non-sublanguage' meaning is simply 'plan' or 'scheme'.

Apparat zreni4 vklhcaet perifericeskil otdel, razpolojennyi v glaze (setcatka, soderja5a4 fotoreqptory i nervnye kletki), i sv4zannye s nim gentral6nye otdely (nekotorye ucastki srednego i mejutocnogo mozga, a takje zritel6na4 oblast6 kory bol6wix poluwiril).

THE APPARATUS OF SIGHT SWITCHES ON THE PERIPHERAL DIVISION, ARRANGED IN AN EYE (RETINA, WHICH CONTAINS PHOTORECEPTORS AND THE NERVE CELLS), AND CENTRAL DIVISIONS ( SOME SECTIONS OF THE AVERAGE AND INTERSTITIAL\*BRAIN, AND ALSO THE VISUAL REGION OF THE CRUST\*\* OF LARGE HEMISPHERES) THE CONNECTED WITH IT.

The visual apparatus includes a peripheral part located in the eye (the retina, which has photoreceptors and nerve cells) and the central parts connected to the peripheral part (some areas of the mesencephalon and diencephalon and also the visual area of the cerebral cortex).

\* Russian terminological collocates not identified but translated disparately. Topical glossary control would have obviated this error. 'Average' is a 'universal' word - 'interstitial' is a term used in crystallography.

\*\* Topical glossary failure - 'crust' belongs to the realm of geology/geophysics, not to anatomy.

Na periferii setcatki preobladaht palocki, bol6wie gruppy kotoryx sv4zany kajda4 s odno1 nervno1 kletko1; ostrota zreni4 zdes6 znacitel6no nije.

ON THE PERIPHERY OF A RETINA, THE BACILLI\* PREDOMINATE, LARGE GROUPS OF WHICH ARE CONNECTED EACH WITH ONE NERVE CELL; A VISUAL ACUITY HERE IS CONSIDERABLE BELOW.

On the periphery of the retina rods predominate, large groups of which are each connected to one nerve cell; visual acuity is significantly lower here.

\* Topical glossary failure. The computer needs more delicate topical glossary control or a technique for 'thesaurusisation'. The etymology

132

#### F. KNOWLES

of the English word 'bacillus' is, of course, 'little rod', the exact counterpart of the Russian 'palocka'. However, one usage belongs to virology, the other to ophthalmology.

Svet poglo5aets4 fotoreqptorami glaza, soderja5imi zritel6ny1 pigment, preobrazuh5i1 3nergih kvantov sveta v nervnye signaly; ot spektra poglo5eni4 pigmentov zavisit diapazon vosprinimaemogo sveta.

THE WORLD\* IS ABSORBED BY THE PHOTORECEPTORS OF AN EYE, WHICH CONTAIN THE VISUAL PIGMENT, WHICH CONVERTS QUANTUM ENERGY OF THE WORLD INTO THE NERVE SIGNALS; ON THE SPECTRUM OF THE ABSORPTION OF PIGMENTS, THE RANGE OF THE RECEIVED WORLD\* DEPENDS!

Light is absorbed by the eye's photoreceptors, which have a visual pigment that converts the energy of light quanta into nerve signals; the range of light perceived depends on the absorption spectrum of the pigments.

\* A 'Wilks-type' semantic calculus would have disambiguated this polysemic word. Alternatively, the correct choice could have been made on the differing frequency characteristics of this word in 'universal' use on

the one hand and in sublanguage contexts on the other.

Konecny1 3tap posadki samol2tov osu5estvl4ets4 s pomo5h sistemy ogne1  
vysoko1 intensivnosti.

THE FINAL STAGE OF LANDING IS ACHIEVED WITH THE AID OF THE FIRE\* SYSTEM  
OF HIGH INTENSITY.

The final stage of aircraft landings is accomplished with the help of  
high intensity lights.

\*A 'Wilks-type' semantic calculus could have deduced that the 'fires'  
involved are not harmful but helpful. This deduction could then have  
prompted the choice of English words such as 'lights' or 'beacons'.



References

- 1 F. Knowles : A brief "error analysis" of SYSTRAN output  
(*'Der GMD-Spiegel'*, 1978/1)
- 2 E. Tauscher/E.G. Kirschbaum (1962) : Grammatik der russischen Sprache  
(*'Volk und Wissen'* Verlag, Berlin)
- 3 P. Toma et al. (1970) : SYSTRAN machine translation system  
(RADC-TR-70-168, Griffiss AFB, New York)
- 4 P. Toma et al. (1972) : Optimisation of SYSTRAN system (1)  
(RADC-TR-72-73, Griffiss AFB, New York)
- 5 P. Toma et al. (1974) : Optimisation of SYSTRAN system (2)  
(RADC-TR-73-155-REV, Griffiss AFB, New York)
- 6 P. Toma et al. (1974) : Some semantic considerations in Russian-English  
machine translation  
(RADC-TR-74-189, Griffiss AFB, New York)
- 7 J.W. Janda et al. : Dokumentation eines Programms zur Analyse  
russischer Sätze  
(*'Linguistik und Informationsverarbeitung'*,  
1970/19-20)
- 8 P. Toma : SYSTRAN - Ein maschinelles Übersetzungssystem der dritten  
Generation  
(*'Sprache und Datenverarbeitung'*, 1977/1)
- 9 L.N. Zazorina (1977) : Častotnyj slovar' russkogo jazyka  
(*'Russkij Jazyk'*, Moscow)
- 10 Ch. Muller (1977) : Principes et méthodes de statistique lexicale  
(Hachette-Université, Paris)
- 11 Y. Wilks : An intelligent analyzer and understander for English  
(Comm. ACM, 1975)
- 12 Statistika reči i avtomatičeskij analiz teksta (1971/2/4)  
(*'Nauka'*, Leningrad)
- 13 A. Kent/H. Lancour (1970) : Encyclopaedia of Library and Information  
Science  
(see Vol.4 for article by M. Masterman : Cambridge Language Research  
Unit)  
(Dekker, New York)
- 14 G. Wahrig (1973) : Anleitung zur grammatisch-semantischen  
Beschreibung lexikalischer Einheiten  
(Niemeyer, Tübingen)