[*Proceedings of the National Symposium on Machine Translation*, UCLA February 1960]

Session 9: SEMANTIC RESOLUTION

AN EXPERIMENT IN THE AUTOMATIC SELECTION OR REJECTION OF TECHNICAL TERMS[1]

Lew R. Micklesen
IBM Research Center

One of the products of the early stages of machine translation work at the University of Washington was the so-called Synoptic Chart for Fields of Science and Technology reproduced in Exhibit 1 of this paper. This chart was discussed in the first report[2] of the University of Washington Project. Whenever a given Russian semantic unit seemed to belong to a specific subfield (i. e. , to one and only one of the rectangles on the chart) or to a specific field (i. e. , to an entire vertical column in the chart) it was given an appropriate number from the chart. This number was to appear as a subscript numeral of the English alternative concerned and was to guide the reader of the output in his selection or rejection of technical terms on the basis of an awareness of the field of science represented by the subject matter of the text. The assignment of various alternatives to fields and subfields of science and technology and the classification itself of fields and subfields could not be checked until the simulated machine translations were produced. Once these translations became available the checking was fairly simple but extremely time-consuming. The first and critical step in the procedure was the perusal of every so-called text passage in the original corpus of the University of Washington MT Project and the subsequent assignment of the contents to one or more of the fields and subfields of science and technology. Next, the simulated machine translation for every text passage was scanned for equivalents containing alternatives bearing subscript numbers. If the subscript number coincided exactly with the number or one of the

---

[1] For the collection of this material I am indebted to Mr. Fredrich Lackmann.

[2] Linguistic and Engineering Studies in Automatic Language Translation of Scientific Russian into English, Department of Far Eastern Slavic Languages and Literature and Department of Electrical Engineering, University of Washington, University of Washington Press, 1959.

Editor's Note: Dr. Micklesen was formerly a member of the faculty of the University of Washington.

	1	2	3	4	5	6	7	8	9
0	Math	Physics	Chemistry	Biology	Medicine	Social Sciences	Integrated Sciences	Applied Sciences	Technology
1	Algebra	Classical & Fluid Mechanics	Physical	Botany	Structure	Anthropology	Astronomy	Mechanics & Mechanical Structure	Machinery Mechanism Tools
2	Geometry	Statistical Mechanics & Thermo	Inorganic	Zoology	Function	Linguistics	Geophysics	Thermo & Heat Engines	Production & Manufacturing Methods
3	Analysis	Electricity & Magnetism	Analytical	Micro-	Diagnosis	Philosophy	Geology	Electrical Engineering	Transportation
4	Statistics	Optics Spectra	Organic	Biophysics	Therapy	Sociology	Geography	Aeronautical Engineering Acoustics	Structures Architecture
5	Numerical Analysis	Quantum Mechanics	Bio-	Psychology	Pharmacy	Politcal Science Diplomacy	Meteorology	Nuclear Engineering	Mining Metals Ceramics
6	Relativity	Solid State	Photo-	Agriculture & Forestry	Public Health Sanitation	Social Planning	Oceanography	Control Engineering	Marine & Naval
7		Nucleonics	Electro-	Animal Husbandry	Psychiatry	Economics Theoretical & Applied		Optics Photography	Military Science Tactics
8		Metrology	Engineering	Fisheries	Veterinary	Law		Materials	Textiles Paper

Exhibit 1: SYNOPTIC CHART FOR SCIENCE AND TECHNOLOGY

numbers assigned the text passage or if the one-digit subscript number corresponded to the first digit of the number or one of the numbers of the text passage, then the alternative associated with the subscript number was treated as if it had been automatically selected as the correct one. All alternatives bearing subscript numbers that did not correspond to numbers assigned to text passages in either of the ways described above were considered to be automatically rejected as if they were incorrect.

Obviously, the selection and rejection of alternatives on this basis had to be evaluated; therefore each selection or rejection was simultaneously evaluated against the context for correctness or incorrectness. Thus, each alternative with a subscript was tested against two sets of oppositions: selection-rejection and correct-incorrect. All these data were recorded in great detail for easy reference and subsequent evaluation. A convenient summary of the data listing the number of the text passage, the field-of-science number or numbers assigned to the text passage, and the numbers of correct and incorrect selections and rejections is presented in Exhibit 2. The total number of decisions is 2, 944 (2, 588 correct decisions + 356 incorrect decisions). This means that 88% of the decisions were correct. The procedure was not entirely automatic since the initial classification of a text passage according to its field of science was made by a human being.

After the data had been properly recorded, the incorrect decisions were subjected to analysis in the hope that they might provide information leading to the improvement in the form and application of the synoptic chart of fields of science and technology. The dichotomy of the incorrect decisions into incorrect rejections and incorrect selections proved to be very significant in the analysis; so the results of the analysis will be discussed first in terms of incorrect rejections, later in terms of incorrect selections.

The analysis of the incorrect rejections revealed that in the vast majority of cases no adjustment was necessary in the assignment of a text passage to a field of science. The analysis also revealed that the assignment of subscript numbers to alternatives, insofar as it allowed only one number per alternative, was correct. Nevertheless the assignment was inadequate in that it was too specific. The incorrect

Text Passage	Field of Science Number(s) Assigned Text Passage	Correctly Rejected	Correctly Selected	Incorrectly Rejected	Incorrectly Selected	Text Passage	Field-of-Science Number(s) Assigned Text Passage	Correctly Rejected	Correctly Selected	Incorrectly Rejected	Incorrectly Selected
1	28	23		4		36	46	7			4
2	13	18			2	37	46	6			
3	13	20	11			38	46	14			
4	15	13	9		3	39	46	16	4	1	1
5	21/22	15				40	73	9		1	
6	21	18		1		41	73	23		12	2
7	11	8	3		4	42	73	35			
8	12	16			14	43	73	12	3	2	1
9	43	20	3	1	1	44	73	14			
10	51	103	18	5	1	45	74	10			
11	51	19	1	14		46	73	19			
12	54	17	7	1		47	72	4		1	
13	43	14	15			48	74	15			
14	54	18	4			49	75	30			
15	53	61	13	2		50	75	20		4	
16	54	14		1		51	38/58	14		4	
17	54	18		10		52	35	23			
18	54	5	1	1		53	33/73	4	1		
19	52	19	3			54	34	26	4	1	2
20	57	21				55	31/72	9			1
21	41	35	6	2		56	34	17			
22	41	27	6	4		57	88	12		1	
23	41	37	8		3	58	35	12		1	
24	41	23				59	23/72	10			
25	41	27		1		60	28	6			1
26	42	10	1		5	61	22/72	10		3	1
27	42	10	1			62	84	11		2	
28	58	16	2	3		63	27	17			
29	42	21				64	84/24	38	3		
30	42	41	5			65	25	22			
31	42	38	7	9		66	25	36		3	
32	42/52/58	20	10		3	67	23	13			
33	42	14			5	68	71	14	2	3	
34	42	43	2	5		69	46	11	1		2
35	42	31	5	2		70	91	18	2		19

Exhibit 2: SUMMARY OF CORRECT AND INCORRECT SELECTIONS AND REJECTIONS ACCORDING TO TEXT PASSAGES

Text Passage	Field of Science Number(s) Assigned Text Passage	Correctly Rejected	Correctly Selected	Incorrectly Rejected	Incorrectly Selected
71	91	7			
72	91	25		7	3
73	91	22	7	1	5
74	91	27	5	1	3
75	91	1	2		2
76	91	22	2	3	
77	72	3		2	
78	91/86	22	1	10	6
79	95/88	11			4
80	95/88	20	5	2	11
81	92/88	25		2	4
82	95	9	3		
83	92	13		1	
84	92	10		12	2
85	6				2
86	92	14	1		3
87	94	13	1		
88	97	19	16		1
89	92	8			
90	92	29			
91	92	19			
92	84	20		1	
93	84	49	21	2	
94	84	22		5	
95	83	58	8	2	2
96	83	27	18	2	2
97	83	21	4		1
98	83	17	2	1	1
99	62	15	10		2
100	62	17			3
101	65	41		5	
102	65	20			
103	97	10	2		2
104	97	19	10		9
105	96	21		11	1
106	96	59	2		4
107	96	58	1	9	1
108	97	23	1	1	16
109	97	21	1		2
110	97	30	2		3
111	84	6		3	
Totals		2288	300	180	176

Exhibit 2 (Continued)

rejections resulting from too high specificity may be classified into two groups on the basis of whether the adjustment necessary to correct the fault involved merely a reduction of the degree of specificity or the complete removal of specificity.

There were two primary areas on the Synoptic Chart where a reduction in degree of specificity of subscript assignments seemed to be particularly effective in eliminating incorrect rejections. Vertical columns 4 and 5 (biology and medicine, respectively) constitute the first area. More than one-fourth of all the incorrect rejections are apparently due to the fact that the subfields of biology and medicine are not as distinct from each other as those in column 6 (social sciences) and column 7 (integrated sciences). A large common vocabulary is shared by most or all branches of medicine and biology, and it seems inadvisable to restrict most words to one specific field or subfield. The following is a partial list of Russian semantic units with alternatives bearing too specific subscript numbers from columns 4 and 5:

зрительный	= optic$_5$	связка	= ligament$_{51}$
лоскут	= graft$_{54}$	слезным	= lachrymal$_5$
оболочка	= membrane$_{51}$	слой	= lamella$_4$
пинцет	= forceps$_5$	срез	= section$_4$
проток	= duct$_{51}$	узел	= ganglion$_{51}$
		чувствительный	= sensory$_4$

Vertical columns 8 and 9 (applied sciences and technology) constitute the second area. About one-ninth of all the incorrect rejections seem due to the fact that these two technical areas are not always distinct. Their vocabularies frequently overlap. The following list includes a partial representation of Russian semantic units with alternatives bearing too specific subscript numbers from columns 8 and 9:

ввод	= lead-in$_{83}$	простои	= demurrage$_{93}$
вкладыш	= bushing$_9$	расчетным	= rated$_{83}$
муфта	= clutch$_{81}$/coupling$_{81}$/socket$_{81}$		
пояс	= flange$_9$	устройство	= working-principle$_{91}$

Two types of solutions suggest themselves for remedying these incorrect rejections. The most obvious method of reducing specificity is to increase the area on the synoptic chart to which given alternatives apply. This can be done by adding the number for another entire vertical column or part of another column, or it can be accomplished by adding the number for part or the rest of the same column. The number of subscript numbers employed certainly constitutes a factor; it would not seem advisable to use more than two such numbers. In the case of the two areas discussed above, columns 4 - 5 and 8 - 9, the decisions do not seem too difficult. Here it seems feasible to give words common to both 4 and 5 or 8 and 9 double subscript numbers, e. g., клетка = "cell$_4$" and ткань = "tissue$_4$" could have number 5 as well as number 4, and вскрытие = "dissection$_5$" and покров = "integumen$_{51}$" could have number 4 in addition to numbers 5 and 51. In instances where alternatives with two-digit subscripts are not shared by columns 4-5 or 8-9 but have wide currency within a single column, they should be re-evaluated to determine whether or not they might be provided with single-digit (columnar) subscripts.

A number of alternatives in these technical areas are undoubtedly specific enough to be permitted either one-digit or two-digit subscript numbers. For example, угнетение = "depression$_5$" and опухоль = "tumor$_{53}$" can certainly be considered medical terms rather than biological; and most of the terms with subscript numbers 96 (marine and naval) and 97 (military science and tactics) are certainly distinct from the terms associated with other subfields in columns 8 and 9.

The other method of attacking the problem of too high a degree of specificity of present subscript numbers is to re-evaluate the synoptic chart itself. Again the columns 4-5 and 8-9 provide an excellent case in point. It is very possible that a careful reconsideration of these columns of the chart might counsel that the fields of medicine and biology, on one hand, and the fields of applied sciences and technology, on the other hand, could subsequently be classified into appropriate subfields which, in turn, would reflect more accurately the distribution of technical terms. Parts of columns may also require consideration and reclassification. Two text passages discussing naval science contained a number of incorrect rejections

because the alternatives concerned bore subscript number 97 (military science and tactics). This situation suggests that one solution may lie in the re-evaluation of these two subfields with the possible creation of another subfield where the two overlap.

In all there were 180 incorrect rejections. If the above suggestions for reduction of specificity were to prove successful, 51% (92 out of 180) of the incorrect rejections would be avoided.

As might be expected, incorrect rejections remedied apparently by the complete removal of specificity were not confined to any particular areas on the Synoptic Chart. Incorrect rejections in this category are all alternatives denoting concepts used extensively in science and technology--more or less general scientific terminology. The following is a list of Russian words with alternatives which had been erroneously limited to one field or subfield of science.

уравнение	= equation$_1$	осаждать	= precipitate$_3$
отношение	= ratio$_1$	переменный	= alternating$_{83}$
значение	= value$_1$	отклонение	= deflection$_2$
сплав	= alloy$_{95}$	вооружение	= arms$_{97}$
вид	= species$_4$	напряжение	= voltage$_{83}$

Undoubtedly the only recourse in these cases is to remove the subscript numbers and treat such alternatives as non-technical terms. The selection or rejection of such alternatives would necessarily be based on a much more sophisticated semantic classification than a synoptic chart of fields of science and technology. Sixty-six incorrect rejections out of the total 180 (37%) could be avoided by removing the subscript numbers. This procedure would not relieve the original ambiguity, but it would prevent loss of essential alternatives.

It should be observed that the largest number of examples were originally assigned the subscript number for mathematics. The necessity of removing this number in many cases is entirely in keeping with the widespread use of basic mathematical terms in the other sciences.

Faulty assignment of the contents of text passages to fields of science occurred in only two instances and gave rise to a very limited number of incorrect rejections. In the first instance, text passage 78, discussing the emergency release of landing gear, landing flaps, and other assemblies, was assigned to 91 (machinery, mechanisms, tools)

and to 86 (control). In retrospect it seems obvious that 84 (aeronautical, acoustic) should have been used instead of 86 because appropriate alternatives for four semantic units, посадочный = "landing$_{84}$", шасси = "landing-gear$_{84}$", щитки = "flaps$_{64}$", and полетный = "gross$_{84}$", (10 occurrences in all) were incorrectly rejected. No new incorrect rejections would have been created by the substitution of 84 for 86. In the second instance, text passage 84 containing a discussion of silver-lap machines was assigned only to 92 (production and manufacturing methods). Limiting the assignment in this way and not including 98 (textiles and paper) caused the incorrect rejection of лента = "silver-lap$_{98}$" and утолщенный = "slubbed$_{98}$" (12 occurrences in all). The addition of number 98, while removing 12 incorrect rejections, would give rise to an incorrect selection, viz., гладкой = "plain$_{98}$". The special problem of incorrect selections will be discussed below. Twenty-two out of 180 (12%) incorrect rejections would thus be removed by adjusting the field to which the article was assigned.

The other half of the dichotomy of incorrect decisions is made up of incorrect selections. There were 176 incorrect selections; so the incorrect decisions were practically equally shared by rejections and selections. All incorrect selections have one particularly interesting feature: this is the only case where there is competition between alternatives with and without subscript numbers that cannot be solved. A few examples will illustrate this feature. The semantic unit вид has the equivalent "view/shape/species$_4$/aspect$_{62}$". Obviously, the third alternative, "species", is found most often in articles on biology, and the fourth alternative in articles on linguistics. The first two alternatives, "view" and "shape", may appear in all kinds of articles, however, including those on biology and linguistics. If вид were to appear in an article on biology, "view" and "shape" would be automatically rejected and "species" selected even though "view" or "shape" might be the correct alternative. The imperfective infinitive приводить has the equivalent "(to)bring/cite/reduce$_1$". The alternative "reduce" is very likely to occur in any article on mathematics; but the other alternatives, "bring" and "cite", have a wide distribution in general technical and non-technical literature and may easily be correct choices in the field of mathematics. The target-language equivalent of the adjective-substantive рабочую is

"working/worker/operating9". The alternative "operating" will frequently be appropriate in the area of technology, but even here the more generally applicable alternatives "working" and "worker" will be required. They must be retained, therefore.

Because of the nature of the equivalents, alternatives with subscripts are always in competition with alternatives without subscripts. It is conceivable that alternatives with subscripts might also compete against other alternatives with subscripts. This could happen if a given target-language equivalent had either (a) two or more alternatives bearing the same subscript number or (b) two or more alternatives bearing two or more subscript numbers also assigned to the article being translated. Neither one of these two conditions obtained among the incorrect selections under discussion.

The only suggestion for remedying such competition between alternatives with and without subscript numbers is to eliminate the competition. That is, the subscript numbers should constitute a basis for selection or rejection of those alternatives that have subscript numbers. For example, in an article on linguistics "aspect" would be selected and "species" rejected, while "view" and "shape" would be retained. These latter alternatives still compete with "aspect" but not actively. Such treatment will remove all incorrect selections. The final resolution of semantic ambiguity would have to be made by more sophisticated procedures.

The results of this experiment in the automatic rejection or selection of technical terms are definitely encouraging. It is undoubtedly true, however, that another set of articles might reveal an almost entirely different set of incorrect selections and rejections. The process of first matching subscript numbers against the fields represented by different sets of articles and then evaluating the incorrect selections and rejections could possibly be repeated until a very high degree of refinement of subscript numbers is attained. Or better still, this procedure should be carried out in only one field of science at a time until the yield of incorrect selections and rejections is almost negligible. The Synoptic Chart of Fields and Subfields of Science and Technology is in no way regarded as a panacea for all the ills of semantic ambiguities among technical terms. A careful re-evaluation and reconstruction of the chart may be indicated, but even

this would not solve all problems of ambiguity. This chart, however, may be useful in the semantics of science and technology if something less than a thorough-going semantic analysis proves feasible for MT.