[*International Conference on Machine Translation of Languages and Applied Language Analysis*, National Physical Laboratory, Teddington, UK, 5-8 September 1961]

A FOURTH LEVEL OF LINGUISTIC ANALYSIS *

by

MICHAEL ZARECHNAK
(Georgetown University, U.S.A.)

INTRODUCTION

THE GAT (Georgetown Automatic Translation) programs for Russian/English Machine Translation have, up to the present time, provided for three levels of linguistic analysis (morphological, syntagmatic, syntactic).# The machine translation output produced by these programmes has been subjected to further structural analysis in order to ascertain its strengths and weaknesses.

The first result of this analysis was reported in Los Angeles at the National Symposium on Machine Translation, Session 6, on February 4th, 1960.

The purpose of this paper is to present structural data in order to show why it is necessary to introduce a fourth level into the analysis of the input language to significantly improve the output in the target language. The improvements would affect the following:

1. The Russian case endings would be transferred into English predominantly on the basis of the kernel structures within which they occur, rather than on the present basis of syntagmatically related words. Thus the span of the linear search to select a proper equivalent for the Russian case endings would be increased.

2. The rearrangement of the English output would be based on generalised structural patterns, reducing reliance upon the specific lists. The result will be fewer exceptions to the rearrangement rules.

The routines, which would be worked out according to these conditions, would facilitate the introduction of the analysis of semantic components within a kernel structure on the operational level.

---

* Note the article: it is 'a', not 'the'.

\# Three Levels of Linguistic Analysis in Machine Translation, Michael Zarechnak, Georgetown University, Washington, D.C., Journal of the Association for Computing Machinery, Volume 6, Number 1, January 1959.

I wish to acknowledge the following members of the Georgetown Project for their help in collecting and classifying the data connected with this analysis, and for the editing of the paper: Professor L.E. Dostert, Dr. Milos Pacak, Mrs. Marjorie Richman, Mrs. Irene Thompson, and Dr. Melrad Mellen.

In our experimental approach to MT, we found that certain assumptions had to be modified in the light of experience. As an example, I refer to the structure of a genitive noun-noun government string.

In translating the genitive case from Russian into English, the following rules served as a basis for the algorithm:*

The substance of the genitive transfer routine is as follows:

1. If a word in the genitive case is the first one in the government structure, the translation of the genitive case is zeroed;
2. If not, and if the word is not listed as an exception, the genitive case is translated by the preposition "of".

An analysis of a translated corpus recently brought to our attention problems which make it necessary for us to initiate not only quantitative changes, such as increasing the list of complex prepositions, but also qualitative changes which will replace the given routine by a new one.

THE REASONS FOR QUALITATIVE CHANGES

The genitive transfer routine of the noun in the genitive case (N_{c2}) was based on computer generated codes. It was assumed that in a string of two or more nouns, the second (or third, etc.) noun in the genitive case belonged semantically to the first. This assumption proved inadequate in practice.

Structurally, it became apparent that two or more nouns in the genitive case do not automatically signal a semantic relationship. The conditions which prevented two nouns in the genitive case from being considered as a noun phrase were the following:

1. The second noun belongs to a nested structure;

Example 1: Все скопившиеся за день тучки
All the small clouds which gathered during the day.

2. The second noun (or the third, etc.) is governed by the predicate of the sentence.

Example 2: Дураки нанесли лесу ущерба не меньше хищников
Vandals have done as much harm to the forests as commercial exploiters.

The above examples indicate that the phrase structure exists within the sentence structure. Therefore, the problem of the hierarchy of government structures is introduced.

It is our belief that the sentence type has to be determined before the subsentence units (phrases) are determined.

This in turn raises the perennial problem of the relation of meaning and form.

---

* See *Appendix 1.*

In order to determine the grammatical function of a given form, one has to know its ontological meaning. Similarly, to select its ontological meaning, one has to know its grammatical function. Theoretically this seems to be a vicious circle. However, experimentally, in any given sentence, if one knows the subject matter and the sentence nuclei in Russian, there is little or no problem in determining both the function of the form and the ontological meaning of the word.

The above-mentioned problem is illustrated by translation samples of the nouns in the genitive case. If a genitive Russian string is translated only slightly differently (for example, as to the order of words, or the suppression of the ending of a noun in the genitive case) the translator would be tempted to think of ad hoc solutions.

Example 3: Разрушение части производительных сил

Destroying a part of the productive forces.

On the other hand, if the given genitive Russian structure is transferred by a sentence the difference is more apparent.

Example 4: Перед наступлением кризиса

Before the crisis occurs.

It is suggested that:

1. The genitive string might have been formed from a sentence; and
2. The information conveyed in such a genitive string could be usefully analyzed to discern the semantic components of the genitive string as well as of the sentence.

To summarize:

Transformation of the genitive string into:

1. The sentence kernel facilitates the analysis of structural genitive relations;
2. The genitive string aids in analyzing the semantic components of the sentence structure.

Therefore, if the binary genitive structure (i.e., in terms of each successive pair of nouns) is reduced to the sentence kernels from which the genitive string was formed, any genitive structure could be operationally classified by sub-classes based on the type of kernel into which the genitive string is transformable. Each kernelized sub-class of the genitive string could be again operationally subdivided into sub-classes of governing and governed nouns.

Experimentally, kernels created from genitive strings were observed as follows:

We start with a two-positional string in which only one of the nouns must be in the genitive case (usually the second). We call the first noun N_1 and the second N_2.

a) $(N_1\ \ N_2\ ) \rightarrow (N_2\ \ Vx_{N1})$*

Example обсуждение тезисов ⟶ тезисы обсуждаются

b) $(N_1\ \ N_2\ ) \rightarrow (N_2\ \ V_{N1})$

Example постановление пленума ⟶ пленум постановил

C) $(N_1\ \ N_2) \rightarrow (N_2$ is $A_{N1})$

Example возможность реализации ⟶ реализация возможна

d) $(N_1\ \ N_2) \rightarrow (N_1\ P\ N_2)$

Example программа подъема ⟶ программа по подъему

e) $(N_1\ \ N_2) \rightarrow (N_1\ \ N_2)$

Example в ряде районов ⟶ в ряде районов

From the kernelization procedure it is obvious that the noun in the genitive case occupied the subject position and the other noun the predicate position. This correlation is operationally important for the selection of the translation for the genitive morpheme. Once the subject-predicate positions are established, the remaining positions would be distributed among the identity sub-classes such as adverbs, adjectives, particles, and conjunctions.

If the genitive string exceeds two positions (a position is defined as that which is occupied by a noun-like word), a test is conducted to determine:

1. Whether more than one kernel formed the genitive string;
2. Whether one kernel had identity sub-classes;
3. Whether the multiple genitive string is not a kernelizable unit.

1. $(N_1\ N_2\ N_3) \rightarrow (N_1\ P\ N_2) + (N_3\ Vx_{N1})$

   $(N_1\ N_2) \rightarrow (N_1\ P\ N_2)$ система для организации

   $(N_2\ N_3) \rightarrow (N_3\ Vx_{N1})$ ТРУД ОРГАНИЗУЕСЯ

Example система организации труда → система для организации
→ труд организуется

2. $(N_1\ \ N_2\ N_3) \rightarrow (N_1\ N_2) + (N_3\ Vx_{N2})$

   $(N_1\ \ N_2) \rightarrow (N_1\ \ N_2)$ число самоубийств

   $(N_2\ N_3) \rightarrow (N_3\ V\ N_2)$ люди кончают самоубийством

Example число самоубийств людей → люди кончают самоубийством

* Vx = verb reflexive; P = preposition; "is" = any auxiliary verb.

3. $(N_1\ N_2\ N_3) \rightarrow (N_1\ P\ N_2) + (N_2\ P\ N_3)$

$(N_1\ N_2\ ) \rightarrow (N_2\ P\ N_2)$ волна от банкротств

$(N_2\ N_3\ ) \rightarrow (N_2\ P\ N_3\ )$ банкротства на предприятиях

Example прокатывается волна банкротств
промышленных предприятий

The patterns of combinatory kernelizations are listed in *Appendix 2.*

SEQUENCE OF SEMANTIC COMPONENTS WITHIN THE GENITIVE STRUCTURE

It has been found that in certain instances the genitive case cannot be translated solely on the basis of a pair of nouns co-occurring in a genitive string.

Example Затрата многих десятков дней труда большого числа рабочих

The expenditure of many weeks of labour by a large number of workers.

The English preposition "by" is not conditioned by the words "labour" and "number" but rather by the word "expenditure".

Since the translation of a noun in the genitive case may depend on more than two co-occurring nouns, it can be concluded that the entire genitive string should be analyzed before the English translation is selected.

Analysis of the genitive structure as a unit means that the sequence of the semantic components in a structure occupies two or more positions from which the genitive string is formed.

The preliminary analysis of approximately 10,000 genitive structures demonstrated that it is the sequence of the sub-classes of the nouns rather than the class of the nouns itself which determines the classification of the semantic components within the genitive structure. Furthermore, it was shown that the sequence of semantic components is rigorously structured.

The sub-class of inanimate concrete nouns (discernable by the human senses; for example СТОЛ) shows certain patterns of predictable sequences. These are listed in *Appendix 3* with accompanying examples.

TRANSLATION INTO ENGLISH

The following five problems were considered as relevant for the transfer into English (the translation by "of" and "ϕ" has been previously mentioned: see p.4).

1. The translation of the genitive morpheme by the following set of English prepositions: "for", "in", "by", "on". Examples:

	RUSSIAN	AS TRANSLATED BY THE HUMAN TRANSLATOR
1.	Расстройство торговли	Disorder <u>in</u> trade
	Потрясения хозяйственной жизни	Upheavals <u>in</u> economic life
2.	Основа экономических кризисов.	Foundation <u>for</u> the economic crises
	Пункт нового цикла	Point <u>for</u> a new cycle
3.	Затрата многих десятков дней труда большого числа рабочих	Expenditure of many weeks of labour <u>by</u> a large number of workers
4.	Монополистов денежного рынка	Monopolists <u>on</u> the money market
	Войны мирового масштаба	Wars <u>on</u> a world-wide scale.

2. The noun in the genitive case is zeroed between two nouns. Previously the genitive case was zeroed only if it occurred with the first word in a prepositional structure. Example:

путем сокращения N_1
времени N_2
обращения N_3
капитала N_4

If kernelized, the structure would break down as follows:

(1) $N_1\ N_2 \rightarrow N_2\ Vx_{N1}$ — время сокращается

(2) $N_2\ N_3 \rightarrow N_1\ P\ N_2 \rightarrow N_3$ is A_{N2} — время для обращения, "обращающееся время", "обращаемое время"

(3) $N_s\ N_4 \rightarrow N_4\ Vx_{N3}$ — капитал обращается

It is clear that (1) and (3) are kernels.

Note that the verb is used transitively in (1) and intransitively in (2). This suggests the rule:

(1) If the predicate equivalent in the genitive string is formed from a transitive verb, the genitive case of the following governing noun <u>would be zeroed.</u>

(2) If the predicate equivalent is used intransitively, the following governing noun would receive the preposition "of";
Thus the above genitive string would be translated:
"by curtailing the circulation time of capital"
The reverse was effected by transformation (2) which resulted in $N_2\ P\ N_s$ and $N_2\ N_3 \rightarrow N_3$ is A_{N2}.

3. The noun in the genitive case is transformed into an adjective and

its governing noun is rearranged into the second position. This constitutes a simple reverse. Examples:

потогонная система организации труда

sweatshop system of work organization

If an adjective precedes such a noun in the genitive case, this would be a multiple reverse. Example:

двигатели внутреннего сгорания

internal combustion engines

4. There may be a number of problems within a single genitive structure of more than two positions. In such cases, the order of testing solutions becomes important.* This constitutes zeroing plus reverse. Example:

при сохранении капиталистической системы хозяйства

while retaining the capitalistic economic system

5. The genitive structure could be replaced by an English sentence. Example:

до того как кризис наступил - до наступления кризиса

before the crisis occurs

CODING OF NOUN ENTRIES

The additional coding of nouns will include markers indicating each stem's derivational capacity, i.e. whether or not the given noun-stem is transformable into V or A. This code will be utilized in kernelization formulas (algorithms).

The semantic sub-classes of nouns will also be coded. This code is operationally produced as is apparent from *Appendix 4.*

This code will be used for the generalization of preposition selection in translating the genitive case in such cases where a pair of nouns is not kernelizable or the kernelization is insufficient.

---

* Idioms and nestings are not discussed.

APPENDIX 1

Step 1: If the item is C_2 and it carries the code 5122 or 512x and it is first in the string, transfer to ZERO.

Step 2: If the C_2 does not carry the code 5122, but carries the code 1122 and at i-1 there is ":", or "," or U-6, transfer by ZERO.

Step 3: If the item is C_2 and it does not carry the code 5122 and it carries the code 1122 and there is no ":", or "," or U-6 at i-1, but carries the code 3112 and there is ":" or "," or U-6 at i-n (before the first item carrying the code 3112), transfer by ZERO.

Step 4: If the item does not carry the code 5122 or 1122, but it does carry the code 2122 or 4122, transfer by ZERO, if the item is the first noun in the stretch.

Step 5: If the C_2 and i-1 is два or три or четыре or a number smaller than 1, 2, transfer by ZERO.

Step 6: If the C_2 carries the code 3112 and the item before the 3112 stretch is целью transfer by ZERO.

Step 7: If the C_2 and the item at i-1 is вследствие transfer by ZERO.

Step 8: if the C_2 and i-1 is кривые transfer by "FOR".

Step 9: If the C_2 and i-1 or i-2 is отношении transfer by "TO".

Step 10: If the item is C_2 and it carries the code 1122 and it does not carry the code 5122 and it does not carry the code 3112 and there is no ":", or "," or U-6 at i-1, transfer it by "OF", and insert it immediately before the i - item.

Step 11: If the item is C_2 and it carries the code 1122 and there is no ":", or "," or U-6 at i-1 and there is code 3112 and there is no ":", or "," or U-6 at i-n (before the first item carrying the code 3112), transfer by "OF" and insert it immediately before the first item carrying the code 3112.

APPENDIX 2

N_2 is A_{N1}	N_2 is A_{N1}
N_2 P N_3	N_3 V_{N2}
N_3 P N_4	N_3 P N_4
N_4 P N_5	

N_2 V_{N1}

N_2 P N_3

N_4 V_{N3}

N_4 P N_3

N_2 V_{N1}

N_3 is A_{N2}

N_4 V_{N3}

N_2 V_{N1}

N_2 P N_3

N_4 is A_{N3}

N_2 V_{N1}

N_3 Vx_{N2}

N_3 P N_4

N_5 V_{N4}

N_4 P N_6

N_2 V_{N1}

N_2 P N_3

N_3 P N_4

N_1 P N_2

N_3 Vx_{N2}

N_3 P N_4

N_4 P N_5

N_1 P N_2

N_3 Vx_{N2}

N_4 is A_{N3}

N_1 P N_2

N_3 V_{N2}

N_3 N_4 $-\phi$

QNT

N_1 P N_2

N_3 Vx_{N2}

N_4 Vx_{N3}

N_1 P N_2

N_2 P N_3

N_4 is A_{N3}

N_1 P N_2

N_2 P N_3

N_4 V_{N3}

N_4 P N_5

N_1 P N_2

N_1 N_2 $-\phi$

QNT

N_1 P N_3

N_3 P N_4

N_4 N_5 $-\phi$

N_3 B N_5

N_1 N_2 $-\phi$

QNT

N_2 P N_3

N_4 V_{N3}

N_4 P N_5

N_1 N_2 $-\phi$

QNT

N_3 V_{N2}

N_4 is A_{N3}

APPENDIX 3

Legend to *Appendix 3*

ϕ —- position of the concrete noun
QNT -- quantifier
PART -- portion of the whole
STR -- structured
UNSTR -- non-structured
QLT -- qualifier
PRI —- process intransitive (deverblal noun)
PRTR —- process transitive (deverbial noun)

THE SEMANTIC COMPONENT SEQUENCE

If an inanimate concrete noun is preceded by another noun(s), the following sequence pattern of semantic sub-classes is observed:

If the noun is singular, the sequence on the left side applies; if the noun is plural, or "Massive", the right sequence applies.

The zero stands for the position occupied by the given noun.

The rest of the numbers indicate the expected positional sequences.

If some of the indicated positions are zeroed, the higher position "shifts" accordingly, i.e. relates directly to the lower position (if present) or to the noun itself if there are no lower positions. Arrows indicate this possibility.

The minus sign indicates that the designated positions of semantic sub-classes precede the zero position. The plus sign indicates the opposite.

Number / Position	SINGULAR	PLURAL
ϕ	inanimate concrete noun	inanimate concrete noun
-1	QNT, PART, STR.	QNT, PART, STR.
-2	QNT, PART, UNSTR.	QNT, PART, UNSTR.
-3	QLT, Reverse order	QLT, Reverse order
-4	Sentence kernel Reverse order	Sentence kernel Reverse Order
-5	ϕ	QNT, GROUP, STR.
-6	ϕ	QNT, NUMBER, STR.
-7	Number, any	Number, any
-8	Process ^{I}	Process ^{I}
-9	Process ^{T}	Process ^{T}
-10	Space	Space
-11	Affirmation	Affirmation
-12	Negation	Negation
+1	Colour	Colour
+2	Name of colour	Name of colour
+3	Coloured	Coloured
+4	QNT, Reverse order	QNT, Reverse order
+5	QLT	QLT

1. нитка жемчуга цвета охотничьей картечи
2. груда неразобранных обломков намерзшего льда
3 четыре рубля восемьдесят пять копеек
4. два ломтя черного хлеба
5. две небесного цвета цистерны
6. факт получения ордена
7. тысяча кубометров ивы, вяза, липы
8. две банки только что полученных пайковых консервов
9. колыбель целой сотни молодых елочек
10. первая попытка воспитания холодоустойчивого, быстрорастущего дуба
11. в адрес главы дома
12. в сумме полутора миллиона рублей
13. досадный факт получения двадцати пяти рублей
14. две кисти винограда
15. несколько пудов караморы
16. пара банок мазута
17. перевозка восьмидесяти процентов опилок
18. экспедиция заготовления государственных бумаг
19. при наличии обильных запасов лесного бурелома
20. армия нерассуждающих и безупречной стали топоров
21. те же, два, казалось, ломтя проржавевшей селедки
22. витаминов у меня было собрано кило два
23. подобрали одних лимонов штук не меньше сорока
24. стайку десятка в два немецких двухмоторных самолетов

APPENDIX 4

	книга	слово	рука	часть	сын	кусок	ветка	приказ	убийство	подарок	стадо	деньги	два	бутылка
	отца	отца	отца	отца	отца	железа	березы	отца	отца	отца	коров	отца	стола	вина
N_1 "у" N_2	+	+	+	–	+	–	+	–	–	–	–	+	–	–
N_1 C_1, 2 N_2	+	–	–	–	+ C5	–	+ C5	–	–	+ C5	–	+ C6	–	+ C5
N_1'от' N_2	+	+	–	–	+	+	–	+	–	+	–	+	–	+
N_1'из' N_2	–	–	–	–	–	+	–	–	–	–	+	–	–	–
N_1'для' N_2	+	+	–	–	+	–	–	–	–	–	–	–	–	+
N_1'при' N_2	+	–	–	–	+	–	–	+	–	+	–	+	–	–
N_1'на' N_2	–	–	–	–	–	–	+	–	–	–	–	–	–	–
N_2 O N_1 C5	+	–	+	–	+	–	+	+	–	+	–	+	–	–
N_2 B N_1 C6	–	–	–	–	–	–	–	–	–	–	+	–	–	+
N_2 Ven N_1	–	–	–	–	–	–	–	–	+	–	–	–	–	–
N_2 V_{N1}	–	+	–	–	+	–	–	+	–	+	–	–	–	–
N_2 1s A_{N1}	–	–	–	–	–	–	+	–	–	–	–	–	–	–