[TA Informations 1966, part 1]

COMPUTATIONAL MORPHOLOGY

Abstract Computational Morphology

The purpose of the paper is the description of a generalized method for the morphological analysis of highly inflected languages in terms of inflectional morphemes.

The inflected word-form is defined as a grammatically characterized unit in which the order of productive morphemes is rigid. The inflectional morphemes are analyzed according to their different morphosyntactic function, i. e. as markers of syntactic relations.

Two approaches to the morphological analysis are discussed:

a) *Admissible correlations between a finite set of single inflectional morphemes and classes of stem morphemes, where the allomorphs of the given inflectional morpheme are treated as distinct inflectional morphemes.*

b) *Application of the principle of complementary distribution of classes of inflectional morphemes where the allomorphs of inflectional morphemes were treated as members of one class.*

The advantages and disadvantages of both methods are discussed in the paper.

The formal representation of the preliminary morphological analysis is demonstrated in the form of trees, tables and matrices.

Acknowledgment

The author expresses his grateful appreciation to Dr. R. R. Macdonald for commenting and editing this paper.

COMPUTATIONAL MORPHOLOGY

The purpose of automatic morphological retrieval is the identification of word forms which can be segmented into their basic productive components (morphemes). By 'productive' components are meant those components which convey grammatical or semantic information which is, or could be, relevant in automatic language data processing.

The establishment of the distributional classes and subclasses of productive morphemes will constitute a reasonable basis for automatic morphological retrieval.

It is clear that the very possibility of any kind of language data processing depends, in the main, upon the construction of an adequate rapid-access storage device and of a well-organized dictionary. The recent development of computers with large disc memories will make it possible to store and retrieve the great amounts of linguistic data which it is necessary to have in the dictionary for the automatized analysis of natural languages.

The old discussions of the organization of dictionaries in terms of segmented versus unsegmented word forms are no longer of crucial importance. However, the prevailing opinion is that the segmentation of word forms into stem morphemes and inflectional morphemes in highly inflected languages is still to be preferred over non-segmentation for such practical reasons as the decrease in the size of the dictionary, the faster look-up, and so on.

Since the Georgetown University Machine Translation Project has always focused mainly on translation from Russian into English, it was necessary to design a system of morphological analysis which would be as complete as possible for Russian and which would be applicable to other inflected languages as well.

It is not the purpose of this article to describe this system of morphological analysis in detail; this description can be found elsewhere. The purpose is rather to point out some interesting features of the research.

Description of Morphological System.

The word-form in inflected languages can be defined as a grammatically characterized unit. The order of morphemes within the word-form is rigid, but the order of word-forms relative to each other is usually variable.

It is assumed by some linguists that this fact justifies a contrast between morphology and syntax. On the other hand, it can be argued that the inflectional morphemes in Slavic languages are markers not simply of morphological relationships, but of morphosyntactic relationships, and, for this reason, it does not seem useful to draw a sharp distinction between morphology and syntax.

Any analysis of an inflected language must discuss the inflections in terms of their morphosyntactic functions. As has already been pointed out, inflectional morphemes serve as markers of syntactic relationships and so help to indicate the boundaries of syntactic connections.

The system of morphological analysis which was developed at Georgetown University was focused on the identification of inflectional suffix-forms, on their admissible correlations with classes of stem morphemes, and on their morphosyntactic functions. The stem morpheme is defined here as a segment of an inflected word-form which can be correlated with the maximum set of inflectional morphemes; it is considered to be a constant segment.

The inflectional suffix-form is here defined as all those inflections which have one particular form, whether they represent one particular morpheme or several; it is considered to be a variable segment.

The relationship between the major classes of stem morphemes and inflectional suffix-forms may be represented as the functional dependence of the dependent variables upon the independent constant,

$$f\ (x, y),$$

where x is the distributional class of the stem morpheme (which is the constant), and y is the class of the inflectional suffix-form (which is the variable).

The morphosyntactic value of an inflected form is the logical sum of the class or subclass of the stem morpheme and the class or subclass value of the inflectional suffix-form

$$\Sigma(x_m, y_n),$$

where x is the class of stem morpheme and subscript m denotes a subclass of x, and where y is the class of the inflectional suffix-form and subscript n denotes a subclass of y.

The morphosyntactic value of a given inflected form is shown in:

a. the category of case and number if the stem morpheme belongs to the class of nominals or pronominals, excluding the personal pronouns,

b. the category of case and number, gender, animateness, form (long or short) and degree of comparison if the stem morpheme belongs to the class of adjectivals, including participle forms,

c. the category of gender, number, voice, tense and person if the stem morpheme belongs to the class of verbals,

d. the category of case and number if the stem morpheme belongs to the class of personal pronouns.

Classes of Stem Morphemes

The First Approach:

In one approach to the research, the classes of the stem morphemes were established only on the basis of the distribution of inflectional suffix-forms.

The resultant number of stem classes in the various form classes in Russian is as follows:

a. Nominals	91
b. Adjectivals functioning as modifiers	29
c. Pronominals, including numerals	32
d. Verbals	39

The stem classes were set up by treating every different suffix-form separately (-A is distinct from -JA, U from JU, and so on through the

following list: A/JA; U/JU; I/Y; IM/YM; AM/JAM; AX/JAX; IX/YX; YJ/IJ; OM/EM; YE/IE; AJA/JAJA; UJU/JUJU; AMI/JAMI; YMI/IMI; EMU/OMU; OGO/EGO).

This type of distribution corresponds to a certain extent to the traditional distinction between 'soft' and 'hard' types of declension.

This approach has the advantage that it can be used if Russian is either the source or target language.

The Second Approach:

The second approach used the complementary distribution and mutual exclusiveness of classes of inflectional suffix-forms in relation to classes of stem morphemes (if A, then $\overline{B} + \overline{C}$; if B, then $\overline{A} + \overline{C}$; if C, then $\overline{A} + \overline{B}$).

By this procedure it was possible to combine the soft and hard types of declension into a smaller number of classes and to eliminate, to a certain extent, the traditional distinction between them.

For example, the following fiftreen different declension types of the first approach were combined into one class, or into two classes if the distinction of animateness as against inanimateness is observed: STOL, FLAG, KORABL6, MUZEJ, PALEQ, KARANDAŚ (inanimate types); SLON, BRAT, KRESTJANIN, KNJAZ6, SOSED, AKADEMIK, KON6, GEROJ, TOVARIŠČ (animate types).

The distribution of inflectional suffix-forms within this class (class M-I/A, where M indicates masculine, I indicates inanimateness, and A animateness) is as follows:

a. If Ø, then $\overline{6} + \overline{J}$;
if 6, then $\overline{Ø} + \overline{J}$;
if J, then $\overline{Ø} + \overline{6}$.

The morphosyntactic value of suffixes -Ø/-6/J is nominative and accusative singular if the subclass is M-I, or nominative singular only if the subclass is M-A.

b. If -OV, then $\overline{\text{-EV}} + \overline{\text{-EJ}}$;
if -EV, then $\overline{\text{-OV}} + \overline{\text{-EJ}}$;
if -EJ, then $\overline{\text{-OV}} + \overline{\text{-EV}}$.

The morphosyntactic value of -OV/-EV/-EJ is genitive plural if the subclass is M-I, or genitive and accusative plural if the subclass is M-A.

c. If -Y, then $\overline{\text{-I}}$;
if -I, then $\overline{\text{-Y}}$.

The morphosyntactic value of -Y/-I is nominative and accusative plural if the subclass is M-I, or nominative plural only if the subclass is M-A.

d. If -U, then $\overline{\text{-JU}}$;
if -JU, then $\overline{\text{-U}}$.

The morphosyntactic value of -U/-JU is dative singular.

e. If -OM, then -$\overline{\text{EM}}$;
if -EM, then -$\overline{\text{OM}}$.

The morphosyntactic value of -OM/-EM is instrumental singular.

f. If -E, the morphosyntactic value is prepositional singular.

g. If -AM, then -$\overline{\text{JAM}}$;
if -JAM, then -$\overline{\text{AM}}$.

The morphosyntactic value of -AM/-JAM is dative plural.

h. If -AMI, then -$\overline{\text{JAMI}}$;
if -JAMI, then -$\overline{\text{AMI}}$.

The morphosyntactic value of -AMI/-JAMI is instrumental plural.

i. If -AX, then -$\overline{\text{JAX}}$;
if -JAX, then -$\overline{\text{AX}}$.

The morphosyntactic value of -AX/-JAX is prepositional plural.

The pairs of allomorphs -I/Y, -U/-JU, etc., can be classified as replacive inflectional suffix-forms because their morphosyntactic function and value is identical, and their occurrence is conditioned by the phonological determination of their distribution.

An identical procedure was used for determining the distribution of replacive inflectional suffix-forms with regard to other major classes of stem morphemes.

The effect of this second procedure was to reduce the number of stem classes of nominals from 91 to 59, and to reduce the number of stem classes of pronominals and adjectivals from 61 to 17.

However, this second procedure is one-directional and can be used only if Russian is the source language. But it does have the advantages that the number of classes of stem morphemes is substantially smaller than in the first approach, and that there are more mechanisms for ensuring correct matching of stems and inflectional morphemes. Moreover, accurate encoding by human coders is much simpler and faster, and the number of human errors is consequently smaller.

Distribution of Inflectional Morphemes

The distribution of inflectional morphemes is described in terms of their

a) co-occurence with stem morphemes;
b) morphosyntactic function.

a) Co-occurrence with Stem Morphemes.

The identical inflectional suffix-form can occur with stems of any number of major word-classes, for example:

1) One word-class:

-AX/-JAX occurs only with nominals;
-AJA/-JAJA occurs only with adjectivals;
-ETE/-ITE occurs only with verbals.

2) Two word-classes:
-AM/-JAM occurs with nominal and pronominal stems.

3) Three word-classes:
-E occurs with nominal, pronominal and adjectival stems.

4) Four word-classes:
-Ø occurs with nominals, pronominals, adjectivals and verbal stems.

b) Morphosyntactic function.

Case homonymy is extremely widespread in Russian as well as in other Slavic languages. There is no type of declension in which all six cases have different forms. Consequently, it is appropriate to classify the inflectional suffix-forms according to their morphosyntactic function. This can be done in the following ways:

1) The inflectional suffix-form is monovalent if it has a single morphosyntactic value (*i.e.*, if it refers to one case and one number only).

2) The inflectional suffix-form is polyvalent if it has more than one morphosyntactic value (i.e., if it refers to more than one case in the same number or in different numbers *).

The monovalent inflectional suffix-forms are:

-EMI/-IMI/-YMI	instrumental plural;
-OMU/-EMU	dative singular;
-6JU	instrumental singular;
-AMI/-JAMI	instrumental plural;
-AM/-JAM	dative plural;
-AX/-JAX	prepositional plural
-UJU/-JUJU	accusative singular (adjectival)
-AJA/-JAJA	nominative plural (adjectival).

All other inflectional suffix-forms are polyvalent.

One of the most productive and, at the same time, most ambiguous inflectional suffix-forms is -I and its allomorph -Y.

The total morphosyntactic value of -I/-Y within the major class of nominals is genitive singular (s2), dative singular (s3), prepositional singular (s6), nominative plural (p1) and accusative plural (p4).

In theory, there are 3125 (5^5) potential morphosyntactic combinations possible for -I/-Y within the class of nominals. However, only 9 combinations of morphosyntactic values actually exist.

* The class of verb inflections is not treated in this paper. See Pacak, M.: « MORPHOLOGICAL ABSTRACTION OF RUSSIAN VERBS »; Machine Translation; M.I.T., Vol. VI; November 1961.

The admissible combinations are:

1.	REAKQI-I	s236 p14
2.	MYŠ-I	s236 p1
3.	SANATORI-I	s6 p14
4.	ZELEN-I (singulare tantum)	s236
5.	STEN-Y	s2 p14
6.	PIAT-I (cardinal numeral)	p 236
7.	ŽEN-Y	s2 p1
8.	STOL-Y	p12
9.	PROLETARI-I	s6 p1

The inflectional suffix-form -I/-Y is monovalent if it occurs with the following subclasses of nominals:

10.	AKADEMIK-I	p1
11.	ZNANI-I	s6
12.	VOZN-I (singulare tantum)	s2

Altogether, then, the inflectional suffix-form -I/-Y exhibits 12 different morphosyntactic values.

The range of the morphosyntactic values of -I/-Y within the class of modifiers is more restricted. If -I/-Y occurs with the adjectival subclass of the type OTCOV-, it is the marker of the nominative plural if the noun modified by OTCOVY is animate, or the marker of the nominative and accusative plural if the noun is inanimate.

Otherwise -I/-Y is the marker of the short from of the plural in adjectivals, including the participles. The categories of case and gender are unmarked in the plural of the short forms.

Patterns of Morphosyntactic Polyvalencies

The patterns of morphosyntactic polyvalencies in nominal, pronominal and adjectival forms can theoretically be divided into three major classes:

I. Case and number are both ambiguous;
II. Case is ambiguous but number is unambiguous;
III. Number is ambiguous but case is unambiguous.

Practically, in Russian only Class I and Class II are to be found. The type of polyvalency of Class III is not found.

Class I.

There are 15 subclasses in which both case and number are ambiguous.

1.	PROLETARI-I	(prepositional singular; nominative plural)
2.	ŽEN-Y	(genitive singular; nominative plural)
3.	DOL-EJ	(instrumental singular; genitive plural)
4.	IVANOV-YM	(instrumental singular; dative plural)
5.	VS-EM	(instrumental & prepositional singular; dative plural)
6.	DOKTOR-A	(genitive & accusative singular; nominative plural)

7. OSTROV-A (genitive singular; nominative & accusative plural)
8. SANATORI-I (prepositional singular; nominative & accusative plural)
9. QAPL-E (instrumental singular; genitive & accusative plural)
10. SOLDAT-Ø (nominative singular; genitive & accusative plural)
11. GLAZ-Ø (nominative & accusative singular; genitive plural)
12. VS-E (nominative & accusative singular; nominative & accusative plural)
13. MYŠ-I (genitive, dative & prepositional singular; nominative plural)
14. REAKQI-I (genitive, dative & prepositional singular; nominative & accusative plural)
15. TAKSI (all cases both singular and plural).

Class II.

There are 14 subclasses in which case is ambiguous, but number is not

1. SAXAR-U (genitive & dative singular)
2. STOL-Ø (nominative & accusative singular)
3. SLON-A (genitive & accusative singular)
4. LES-U (dative & prepositional singular)
5. STOL-Y (nominative & accusative plural)
6. ŽEN-Ø (genitive & accusative plural)
7. Č-EM (instrumental & prepositional singular)
8. IVANOV-U (dative & accusative singular)
9. BEDNOST-I (genitive, dative & prepositional singular)
10. POL-E (nominative, accusative & prepositional singular)
11. IVANOV-A (feminine nominative singular; masculine genetive & accusative singular)
12. IVANOV-YX (genitive, accusative & prepositional plural)
13. ST-A (genitive, dative, instrumental & prepositional)
14. PORTN-OJ (masculine nominative & accusative (inanimate) singular; feminine genitive, dative, instrumental & prepositional singular).

Distributional Table of Morphosyntactic Polyvalencies

Class I: Case and number ambiguity

Number of values	Number of subclasses having that number of values
2	4
3	7
4	2
5	1
12	1
Total	15

Class B: Case ambiguity only

Number of values	Number of subclasses having that number of values
2	8
3	4
4	1
6	1
Total	14

This table shows the distribution of morphosyntactic polyvalencies within Classes I and II. For example, within Class I there are four different dyadic subclasses; reference to the preceding list will show that these are:

1. locative singular and nominative plural.
2. genitive singular and nominative plural.
3. instrumental singular and genitive plural.
4. instrumental singular and dative plural.

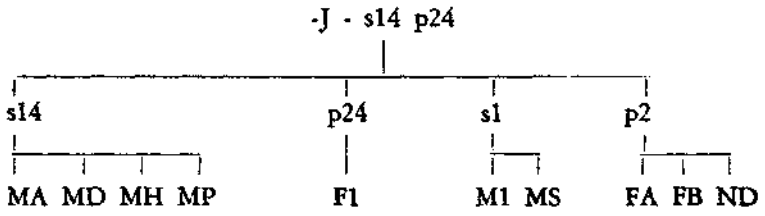
Representation of Data

There are several ways of representing the results of a morphological or of any other linguistic analysis in a computable form.

Algorithmic Tree Representation.

The algorithmic tree representation is demonstrated by means of the inflectional suffix-form -J and its morphosyntactic distribution within the major class of nominals.

The symbolic notation s14 p24 is a coverall for the widest range of morphosyntactic values of -J under all circumstances; these values are nominative singular (s1), accusative singular (s4), genitive plural (p2) and accusative plural (p4). The actual ranges of morphosyntactic values are s14 (nominative, accusative singular), p24 (genitive, accusative plural), s1 (nominative singular), and p2 (genitive plural), depending on the subclass of the nominal stem (the symbols MA, MD, MH, MP, MI, MS, FA, FB, FI, ND, are here used to represent different subclasses of nominal stems).



MA	MUZEJ	MS	PROLETARIJ
MD	KRAJ	FA	LINIJA
MH	SANATORIJ	FB	STAJA
MP	ČAJ	F1	MARIJA
M1	GEROJ	ND	ZNANIE

Logical Formulae.

The same type of morphological analysis for the suffix-form -J can be written as four logical formulae.

Formula 1 : J + [(MA) v (MD) v (MH) v (MP)] = *AD* = s14
Formula 2 : J + [(F1)] = *HJ* = p24
Formula 3 : J + [(M1) v (MS)] = *A* = s1.
Formula 4 : J + [(FA) v (FB) v (ND)] = *H* = p2

Matrix Representation.

The distribution of inflectional morphemes can be represented by a set of matrices which are constructed separately for every inflectional suffix-form with its allomorphs.

The matrix for the inflectional suffix-form -J is given as an example.

Matrix for -J

	MA	MD	MH	MP	M1	MS	FA	FB	F1	ND
s1	+	+	+	+	+	+	—	—	—	—
2	—	—	—	—	—	—	—	—	—	—
3	—	—	—	—	—	—	—	—	—	—
4	+	+	+	+	—	—	—	—	—	—
5	—	—	—	—	—	—	—	—	—	—
6	—	—	—	—	—	—	—	—	—	—
p1	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	+	+	+	+
3	—	—	—	—	—	—	—	—	—	—
4	—	—	—	—	—	—	—	—	+	—
5	—	—	—	—	—	—	—	—	—	—
6	—	—	—	—	—	—	—	—	—	—

The subclasses of the stem morphemes of nominals are listed across the matrix, and the separate morphosyntactic values are listed down the matrix.

A plus sign at the intersection of column and row indicates that -J has the morphosyntactic value to the left of that row, provided that it is matched with the subclass of nominal stems at the top of that column.

Three-dimensional Table.

The construction of a three-dimensional table could also prove useful in morphosyntactic analysis.

The subclasses of the stem morphemes are listed parallel to the x-axis and the particular inflectional suffix-forms are listed parallel to the y-axis. The intersection of the horizontal row and vertical column defines a line parallel to the z-axis can be considered as containing the morphosyntactic values which are the logical sum of the values of the stem and of the inflectional morpheme.

Three-dimensinal Table

	MA	MD	MH	MP	M1	MS	FA	FB	F1	ND
-J	s14	s14	s14	s14	s1	s1	p2	p2	p24	p2
-Ø	—	—	—	—	—	—	—	—	—	—
-A	—	—	—	—	—	—	—	—	—	—
-JA	—	s2 p14		—	s24	s24	s1	s1	s1	—
-E	s6	s6	s6	s6	s6	s6	s36	s36	s36	s14

etc.

The three-dimensional table has the advantage that all admissible combinations can be represented on one table instead of in a series of matrices. In some cases, however, it may be more advantageous to use matrices because they can be more easily replaced and controlled. The preference for using either one large three-dimensional table or a set of smaller matrices will depend on the type of computer which is available.

Both the table and the matrices are flexible; additional subclasses of stem morphemes and additional inflectional morphemes can easily be incorporated into the table without changing the logical basis.

Conclusions

(a) The segmentation of inflected word-forms in stem-morphemes and inflectional morphemes is practical for highly inflected languages, especially if the machine dictionary is of large size.

(b) The system of morphological analysis which is described in this paper was tested on about 5 million running words in Russian on 705 and 7090 IBM computers with positive results. The percentage of errors was 1½%. The number of stem-morphemes listed in the Georgetown dictionary was about 40,000.

(c) The same system of morphological analysis can be applied to other Slavic languages as well. Morever, it would be possible to develop a common system of morphology and syntax for a group of Slavic languages which would facilitate machine translation to and from those languages *.

Milos PACAK
Institute of Languages and Linguistics
Georgetown University Washington, D. C.

* PACAK M., Slavic languages : *Comparative Morphosyntactic analysis ;* Machine Translation ; The Massachusetts Institute of Technology ; Vol. 8, No. 1, August 1964.

LITTERATURE

DOSTERT, L.-E. : « Automatic Translation and Language Data Processing » ; *Vistas in Information Handling,* Vol. 1, Chapter 5 ; Spartan Books, Washington, D. C., 1963.

GARVIN, P.-L. : « A Study of Induction Method in Syntax ; *Word,* Vol. 18 ; August, 1962.

GENTILHOMME, Y. : « Enseignement du Russe aux Scientifiques » ; *Études de Linguistique Appliquée ;* Vol. 3 ; Paris, 1964.

HARRIS, Z.-S. : « Structural Linguistics » ; *The University of Chicago Press,* 1961.

HAYS, D.-G. : « Research Procedures in Machine Translation » ; *Natural Language and the Computer,* Chapter 4 ; McGraw-Hill Book Company, 1963.

LAMB, S.-M. : « Outline of Stratificational Grammar » ; *University of California,* 1962.

KING, G.-W. : « The Requirements of Lexical Storage » ; *Georgetown University Monograph Series on Languages and Linguistics,* No. 10, 1959.

MACDONALD, R.-R. : « General Report 1952-1963 » ; *Georgetown University Machine Translation Research Project,* Paper No. 30 ; June, 1963.

MEL'CHUK, I.-A. : « Morphological Analysis in Machine Translation » ; *Problemy Kibernetiki,* No. 6 ; 1961.

NIKOLAJEVA, T.-M. : « Opyt Algoritmiceskoj Morfologii Russkogo Jazyka ; *Strukturno-Tipologiceskie Issledovanija ;* Akademija Nauk, USSR, Moskva, 1962.

NIDA, E.-A. : « Morphology » : *The University of Michigan Press ;* Ann Arbor, 1963.

LEHMANN, W.-P. ; PENDERGRAFT, E. : « Structural Models for Linguistics Automation » ; *Vistas in Information Handling,* Chapter 4 ; Spartan Books ; Washington, D. C., 1963.

ŒTTINGER, A.-G. : « Automatic Language Translation » ; *Harvard University Press ;* Cambridge, Massachusetts ; 1960.

PACAK, M. : « Morphology in Terms of MT » ; *Advances in Documentation and Library Science,* Vol. 3, Part 2, Chapter 36 ; Interscience Publishers ; New York, 1961.

PACAK, M. : « Syntagmatic Limits of Morphological Sets » ; *Methodos No.* 49-50, Vol. XIII ; Milano, Italy, 1961.

HENISZ-RETMAN, B. : « Morphological Analysis of Polish Nouns » : *Georgetown University Machine Translation Research Project,* June 1962.