[From: *Information retrieval and machine translation*, ed. Allen Kent. International conference… September 6-12, 1959, Cleveland, Ohio; part 2 (New York: Interscience Publishers, 1961)]

CHAPTER 36

Morphology in Terms of Mechanical Translation

MILOS PACAK

Georgetown University, Washington, D. C.

One of the most crucial questions concerning mechanical translation is the size of an adequate and well-organized glossary. Since the input language for the time being is Russian which is known as a highly inflected language—according to data collected by Josselson's group at Wayne State University, 86 percent of the running words in Russian are inflected—the listing of each item in all its paradigmatic forms would increase unnecessarily the glossary storage and slow down the dictionary look-up. According to our estimation, the total number of entries required for one noun averages form 6 to 10, for adjectivals 9, and one verb—including participial forms—might require 59 full-form dictionary entries.

Is there any short cut in reducing the size of the glossary and any procedure of a more efficient and economic look-up? There is definitely one, i.e., split-glossary.

There have been discussions for a long time about split versus nonsplit glossary and many problems have been brought up.

After long research considering all pluses and minuses of both approaches, I definitely came to the conclusion that the split technique, based on the morphological analysis of the input language, i.e., Russian, brings positive results.

Let me describe the split method in some more detail.

As I mentioned above, one reason for the split method is the reduction in the size of the glossary. The second reason is the goal (achievement) of an exhaustive identification of input language which is necessary for any type of syntactic analysis and for transformation into the output language—whichever it might be.

In all Slavic languages the recognition key is in the suffix of inflected items which carries the grammatical information within itself. We have decided to exploit this information nucleus for getting all the grammatical features which are pertinent for the given suffix.

The logical system of morphological analysis is based on establishing classes and subclasses of nominal and pronominal items according to their inflection. To date 58 distinct types of nouns and 27 distinct types of adjectivals are established. The amount of grammatical information which can be derived from the stem itself depends on the part of speech to which the given stem belongs. If it is a noun-stem the three-digital code is added to the stem in the following sequence.

The first digit is used as part of speech indicator ("1" stands for noun).

The second digit indicates the gender ("1" masc. "2" feminine, "3" neuter).

The third digit stands for animation.

The next five digits have been used for matching procedure as declension-type markers.

The adjectivals are coded in a similar way. "3" in the first position indicates the class of adjectivals, the second digit is used for the indication of the possibility of comparative inflection and for the identification of pronouns and numerals. The following seven positions have been employed in a similar way as for nouns. It is obvious to me that this nine-digital code is redundant and can be decreased to a five-digit code which will contain all the information which is pertinent for the identification of the analyzed input item. This approach is closer to my original scheme of Russian Morphology which is described in Georgetown University Seminar Work Paper MT-74.

Verbs have been entered in the dictionary as "simple split bases" and "multiple split bases." It should be understood that "single split base" refers to a single verb base which takes the complete set of conjugational paradigmatic endings.

Example: CITA which is entered as a single base form takes the following set of paradigms:

-T6 for infinitive
H, EW6, ET, EM, ETE, HT for nonpast tense,
L, LA, LO, LI for past tense,
1 ITE for imperative,
4 for present gerund,
V, VWI for past gerund,
N, NA, NO, NY for past participle passive, short form.

Those verbs the stem of which is subject to alternation have been listed in as many modified forms as necessary ("multiple split bases").

Example: The verb "PISAT6" had to be entered in two forms: PIS- and PIW because of S W alternation.

The frequency of verbs which undergo the process of morphemic alternations is relatively high. Therefore, it seemed feasible to develop a routine which would permit handling this type of verbal base as a single entry instead of listing it as a multiple entry. This procedure has been described in my recent paper "The Morphological Abstraction of Russian Verbs." I am going to review briefly the main points of this paper.

Thirty-nine different patterns of morphemic alternations have been established and coded. They fall into three major classes, 1 1, 1 2, and 1 3 alternations. The four digit alternation code is alphabetic because it is mnemonic and easier to use. The first digit designates the verb form and the 2nd, 3rd, and 4th digits indicate the type of alternation.

The dictionary research for a verb alternant is performed on two levels:

Level A—search for zero-alternant type. If the verb stem carries 000 in the 2nd, 3rd, and 4th position, in other words, if the stem belongs to the zero-alternant type the suffix operation goes into effect, and the search for an alternant is skipped.

Level B—search for alternant "2." If the identified base carries an alternant code the program checks for the base final. If the stored base-final (alternant 1) is identical with the input base-final, the suffix operation continues.

If the compared base is not identical the program checks for alternant "2."

Example: Input stem is "PISAT6" (write). The stem which is listed in the dictionary is PIS with the code 2S^{V}. The code 2S^{V} indicates that the final "S" (alternant 1) alternates with "S^{v}" (alternant "2" - S^{v} modified "S"). In the case mentioned above the stem PIS is sensed and the AT6 suffix operation proceeds. The search for the alternant "2" does not go into effect.

Now let's suppose that the input item is "PIS^{V}ET." No base "PIS^{V}" is found in the dictionary. The program checks for the only possible alternant of "S" and locates "S^{v}." When "S^{v}" (alternant "2") is located the ET suffix operation proceeds.

I am not going to explain here the further technical details because it would require too much time. I would like to mention only that the proposed procedure is flexible. The addition of new patterns of alternations or the modification of existing patterns would be possible without any substantial change in the logical structure.

The size of the dictionary will be reduced so only one base will be required, for multiple verb stems. The described analytic scheme can be used for input as well as for output. If it is used for the output certain small modifications of the suffix operations will be necessary. In general, the system which has been developed for Russian verbs can be easily applied to other Slavic languages. It will be of greater value for Czech and Polish because of the high frequency of morphemic alternations in these languages.

In our present system—which has already been tested on the computer—participle forms have been entered into the split dictionary as individual items for each participle type. The idea of using the infixes (around 15) has been temporarily abandoned for programming reasons. Because the participles are functionally the verb derivatives, they carry "2" in the first position (the same as verbs); "3" in the second position is the participle indicator and "l," "2," "3," or "4" in the third position describes the type of the given participle stem. The suffix operation follows the same rules which are used for adjectives.

Our morphological system is based on the formal logical interrelations among classes of stems and a class of suffices. The output, i.e., the obtained grammatical information, is in fact the logical product of stem plus suffix logical combination. By combining the given suffix with stems of different values (these different values are expressed by distinct codes) we obtain the different outputs as I am going to demonstrate on one example.

Example: Let us consider ϕ suffix operation.

If ϕ suffix is matched with the type of stem 1A (STOL), 1B (OSTROV), 1C (POL-), etc., then the output obtained is (C1 • C L1 • N1) [nom., ace., sing.]. If the same suffix is matched with the type 4A (SLON) or 4B (BRAT), etc., the result is C1 • N1 nom. sing. . By matching ϕ with 7A (KNIGA), 7B (STENA), etc., the output (C2 • N2) [gen. pl.] is obtained.

The following different types of morphological output are obtained
by matching ϕ suffix with some other types of matchable stems:
(C1 C2 C3 C4 C5 C6) (N1 N2)
(C2 C4 N2)
(C1 C4 N1) (C2 N2)
(G1 C1 N1)
(G2 C1 N1)
(G1 N1 F2)
(G1 A2 C2 C4 N1) (G1 A1 C1 N1)

Note: C stands for case, N stands for number, G stands for gender, A stands for animateness, and F stands for form (short or long)

We have seen that ten possible types of morphological output were obtained by matching the suffix ϕ with different types of stems, each with different truth-value. These distinct truth-values of stems are cues for different types of output. But it has become necessary to point out that in some instances the suffix itself carried the grammatical information disregarding the matching value of the stem. For example the suffix "4M1"—the output, value of which is unambiguously C5 • N2 [instr. pl.]—indicates at the same time that the given stem can be only a noun stem. This information is complete because the recognition of gender and animateness is redundant.

The same is valid for most verbal suffixes which are distinct from noun and adjectival suffixes.

Each listed suffix is matched with the stem in the same way as suffix ϕ. The total number of distinct outputs is 110. About one-half of these (56) are unambiguous and 54 are ambiguous.

The programming technique which has been developed for morphological analysis and dictionary look-up can be briefly described as follows:

The program has been written for 705 II computer. The read-while-write (RWW) instruction has been used; while 25 new words are read into the memory, 25 looked-up and processed text words are written on tape. Two dictionaries—the dictionary of full forms and the dictionary of split forms—are read into the memory in blocks of 50 split and 10 unsplit records. Each input item is first compared with the dictionary of full forms and if it has not been found there, it is compared against the split glossary. In this stage there is a check for "S4" or "S6" particles (post suffixes). If one of the post suffixes is sensed, the computer stores the appropriate code in the appropriate location and the suffix operation proceeds. During the removal of suffixes, the dictionary may move forward into a position beyond that required by the next item, and backspacing becomes necessary. When a stem equivalent has been located, the suffix operation goes into effect, and the derived grammatical information is added by the computer to the nonshifting stem information.

The morphological program contains about 4,000 instructions. But according to the view of our programmers it might be possible to cut down this program to about one-half. Notwithstanding the backspacing, the look-up and morphological analysis are relatively fast. A text of 30,000 words was looked-up and analyzed in 25 minutes. Because the dictionary passes through the memory quite fast, the size of the dictionary can be substantially increased without reducing the efficiency of the system.

The system of morphological analysis is not limited to a specific kind of text. It is intended to cover scientific text as well as general. It is flexible, in that the number of stem types can be expanded and the necessary logical operations supplemented without any essential difficulty. It should be mentioned that the counterpart of Morphological analysis, i. e., the English synthesis, has been worked out by Mr. Philip H. Smith from Georgetown University. I refer here to his research report No. 17, called "English Synthesis Codes."

In my opinion there is no doubt that the analogous morphological analysis could be applied to all Slavic languages because of their inflectional similarity. I have gone even one step further and started working on comparative multislavic morphology, which is the first step of the multislavic translation research.

I am well aware of the fact that there are many more details which should be mentioned in my report. But is has not been my intention to present here the exhaustive explanation of our morphological system which is described in more detail in corresponding papers, namely, "Scheme of Russian Morphology," Seminar Work Paper MT 74; "Morphological Analysis," Research Report No. 5; "The Morphological Abstraction of Russian Verbs," Research Report No. 22.