[*Mechanical Translation*, Vol.7, no.2, August 1963]

Coding the Russian Alphabet for the Purpose of Mechanical Translation

by John Lyons,† School of Oriental and African Studies, University of London

If we take advantage of our knowledge of the phonological characteristics of Russian and their orthographic representation, it is possible to introduce a number of simple transformations operating on the text at input, the effect of which is to reduce the number of affixes and simplify the morphological analysis.

It is well known that there is in Russian a phonological opposition between palatalized and non-palatalized consonants (or, in the traditional terminology, between "soft" and "hard" consonants). This palatalization is marked in the Russian orthography by the use of one of the set of "soft" vowels or by the special "soft sign" according to whether the palatalized consonant is followed by a vowel or not. This immediately suggests the possibility of replacing the "soft" vowels by the "soft sign" + the corresponding "hard" vowels. Thus "Я" would be transformed into *ЬА, "Ю" into *ЬУ, etc.[1] Furthermore, the "soft sign" and the letter Й are in complementary distribution, the "soft sign" being written after a consonant and Й being written after a vowel.
They may therefore be regarded as "allographs" of the same "grapheme" and represented by the same symbol, Ь. The transformations suggested so far are listed here for convenience:

Я → *ЬА
Е → *ЬО
Ю → *ЬУ (1)
И → *Ь ЫI
Й → *Ь

The effect of these transformations operating on the text at input is not merely to reduce the number of symbols required by five but, more important, to reveal identities in the "hard" and "soft" declensions and conjugations which the Russian orthography tends to conceal. This will be clear from Table 1.

In certain positions in Russian there is what some linguists would call "neutralization" of the palatal non-palatal opposition. That is to say that certain consonants are necessarily either hard or soft. The orthographical conventions of Russian reflect this phonological neutralization, although, for historical reasons, they are no longer in complete accord with contemporary phonetic realization in their prescription of the particular vowels permitted after these consonants. A great simplification is effected in the declensions and conjugations of those lexemes whose stems end in one of these consonants if we introduce the following transformations to operate before the transformations in (1):

(a) After Ш, Ж, Щ, Ч, Ц; { А → *Я ; У → *Ю }
(b) Final Ш, Ж, Щ, Ч, Ц → *Ш6, *ЖЬ, etc. (3)
(c) After Ц; Ы → И (4)
(d) After К, Г, Х; И → *Ы (5)

The effect of these transformations will be clear from Table 2.

The letter О appears after the letters Ш, Ж, Щ, Ц, Ч only when the syllable in which it occurs is under stress and not consistently then (since Е [i.e. Ё] may be written). We thus have marked orthographically the distinction between БОЛЬШЕЙ ("greater") and БОЛЬШОЙ ("great"), though in other cases of these same words, which differ similarly in stress, the distinction is not marked: cf. БОЛЬШИМ and БОЛЬШИМ. It is evident that the effect of the transformations so far mentioned will be to preserve the distinction between these words when the orthography recognizes the distinction, but only at the price of creating two stems for the finally-stressed word: cf.

*БОЛЬШЬ—ОЬ, *БОЛЬШ—ЫМ
but *БОЛЬШ—ОЬ, БОЛЬШЬ—ЫМ (1)

We are now faced with the necessity of deciding among several more or less undesirable solutions to this problem.

Since the number of pairs of words in which there will be minimal contrast consisting in the opposition between Е and О after Ш, Щ, Ж, Ч and Ц, is very small (but exactly how small it is impossible to say in advance) we could introduce a transformation.

† The ideas described in this paper were developed while the author was working as linguistic consultant to the group engaged on mechanical translation at the National Physical Laboratory, Teddington, Middlesex, England, in August, 1939. Although it was decided not to make use of them at the time, it has seemed worthwhile putting them forward for discussion.

[1] The asterisk is used throughout this paper to distinguish the transformed spellings assumed by words inside the computer from the orthographic forms in which they arc met in the text to be translated.

TABLE 1

СТОЛ	СЛОВАРЬ	СЛОЙ	→	*СТОЛ—Ф	*СЛОВАРЬ—Ф	*СЛОЬ—Ф
СТОЛА	СЛОВАРЯ	СЛОЯ	→	—А	—А	—А
СТОЛУ	СЛОВАРЮ	СЛОЮ	→	—У	—У	—У
СТОЛОМ	СЛОВАРЕМ	СЛОЕМ	→	—ОМ	—ОМ	—ОМ
СТОЛЕ	СЛОВАРЕ	СЛОЕ	→	—ЬО	—О	—О
СТОЛЫ	СЛОВАРИ	СЛОИ	→	—Ы	—Ы	—Ы
СТОЛАМ	СЛОВАРЯМ	СЛОЯМ	→	—АМ	—АМ	—АМ

Note that the symbol "ф" stands for the *zero-affix.*

After Ш, Ж, Ч, Щ and Ц; О → *Е (6)

The effect of this would be, for example, to change БОЛЬШОЙ into *БОЛЬШЕЙ (whence ultimately by (1) to *БОЛЬШЬОЬ) and thus to destroy the orthographical difference which exists in the text between certain forms of the comparative and the positive of this adjective. It is worth noting, in this connection, that those forms of the comparative and positive which differ, in stress but not in orthography (cf. БОЛЬШИМ: БОЛЬШИМ) are frequently distinguished in Russian typographical practice by printing an acute over the stressed syllable in the comparative. This suggests that even the native Russian might be momentarily in doubt about the interpretation and unable to decide from the immediate environment of the word whether it is the positive or comparative. It certainly seems gratuitous to throw away information when we have it, if the lack of this information is going to cause difficulties of interpretation later.[2] We should, therefore, be reluctant to introduce a transformation of the form (6) until we are perfectly certain that the information thus lost is of no further use to us.

Another possibility which suggests itself is that of increasing the number of affixes. Such would be the effect, for example, of introducing a transformation of the form:

After Ш, Ж, Ч, Щ and Ц; О → *ЬЕ (7)

Under this rule, БОЛЬШОЙ would become *БОЛЬШЬЕЙ and ultimately *БОЛЬШЬ—ЬОЬ. The result would be satisfactory in that it yields one stem without loss of information, but unsatisfactory in that it would lead to a considerable increase in the list of affixes.

It is now worth enquiring whether having to code two stems in the dictionary is such a bad thing after all. It would seem to be desirable, from many points of view, to have two kinds of stems in a Russian automatic dictionary: "false stems" and "true stems". With the "false stems" will be coded an indication of what addition must be made to arrive at the morphologically acceptable or "true" stem; with the "true stems" there will be given in the dictionary the grammatical and lexical information required for translation. With the techniques available for the treatment of "false stems" in the dictionary it is possible to enter the stem *БОЛЬШ which results from the splitting off of the affix *ОЬ as one among a number of "false stems" in the dictionary. And the possibility of doing this would make the application of the orthographic transformations suggested here more satisfactory.

It will be evident from the list of affixes given in Table 3 that whenever there is a pair of affixes one of which includes the other as a right-hand subpart of itself any automatic splitting routine is liable to produce what is, linguistically speaking, a false split. Take, for example, the affixes *А and *ЬА, the first of which we should wish to regard as the genitival desinence in the word "СЛОЯ" (→ *СЛОЬ—А) and the second of which we would regard as the gerundival desinence in the word "ДЕЛАЯ" (→ *ДЕЛА—ЬА). It is probably more economical to arrange that the largest right-hand segment of the word which matches one of the list of affixes is always automatically split off and to

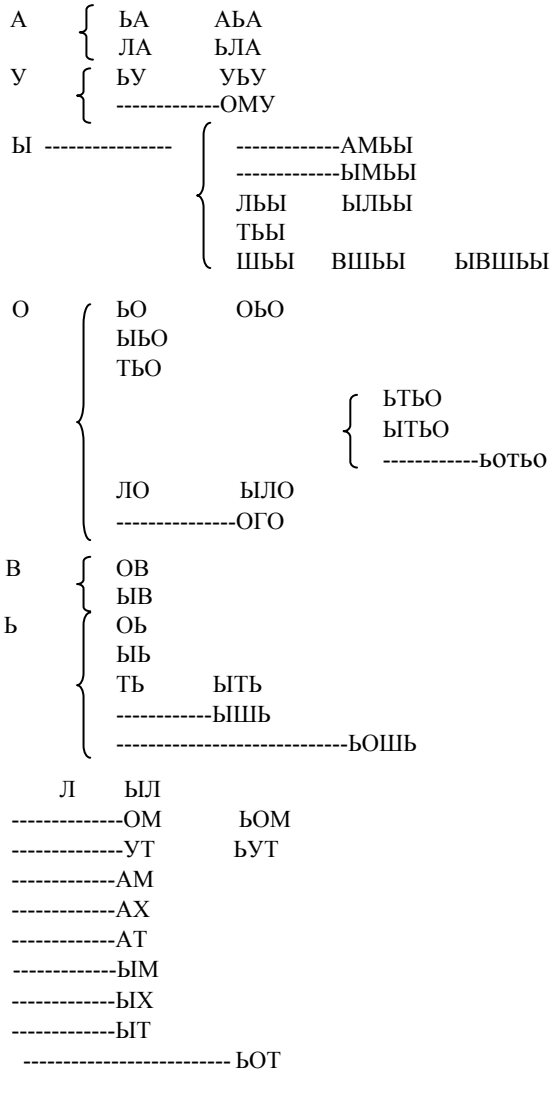
[2] It seems to be widely assumed by MT groups working on Russian that they will not have to have techniques available for coding stress. Although a stress mark is printed only exceptionally in Russian, it is precisely because the orthography is ambiguous and the ambiguity is not easily resolved from context that the diacritic is printed. This would seem to indicate that a technique should be at hand for encoding the information given. From this point of view the Ё when printed should be regarded as Е + diacritic since it may have been printed in order to avoid possible ambiguity, e.g., a confusion between ВСЁ and ВСЕ.

TABLE 2

НОЖ	→	*НОЖЬ—
НОЖА	→	*НОЖЬ—А
НОЖЕМ	→	*НОЖЬ—ОМ
НОЖИ	→	*НОЖЬ—Ы
ТАБЛИЦА	→	*ТАБЛИЦЬ—А
ТАБЛИЦУ	→	*ТАБЛИЦЬ—У
ТАБЛИЦЕЙ	→	*ТАБЛИЦЬ—ЬОЬ
ТАБЛИЦЫ	→	*ТАБЛИЦЬ—Ы
ДЕЛАЮЩИЙ	→	*ДЕЛАЬЫЩЬ—ЫЬ
СДЕЛАННЫЙ	→	*СДЕЛАНН—ЫЬ
ДЕЛАЮЩЕГО	→	*ДЕЛАЬУЩЬ—ОГО
СДЕЛАННОГО	→	*СДЕЛАНН—ОГО

TABLE 3

LIST OF RUSSIAN AFFIXES SHOWING THE POSSIBILITY OF FALSE SPLITS



enter the resultant "stem" in the dictionary with an indication of the addition which must be made to arrive at the "true" stem.[3] The fact that the proposed orthographic transformations will increase the number of stems in some cases should not weigh heavily against their acceptance; for it is equally a fact that these transformations will reduce the number of paradigms for the different word-classes and the number of formally distinct, but functionally equivalent, affixes, and coupled with a more refined splitting-procedure and the technique for handling "false stems", will effect a much greater reduction in the total number of stems, as well as making for a more elegant and satisfactory morphological analysis. And it is the present writer's conviction that the more linguistically appropriate the analysis at the morphological level the simpler will be the subsequent syntactic and semantic analysis.

It remains to be considered whether the proposed transformations are in all instances reversible, in the sense that when they are set to operate in reverse they will yield uniquely the input word. They were based on our knowledge that there is in Russian neutralization of the palatal/non-palatal opposition in certain positions and on the orthographical reflection of this neutralization. In the case of native Russian words the neutralization is absolute. It is well-known, of course, that a number of words of foreign origin "break the rules" and that the transcription of foreign proper names may attempt to approximate to their un-Russian pronunciation by writing combinations of Russian letters which otherwise do not occur. Take, for instance, the word "ПАРАШЮТ" ("parachute"). This would be transformed at input into *ПАРАШЬУТ [by (1)]. Now, if there were also a word "ПАРАШУТ", this would likewise be transformed into *ПАРАШЬУТ [by (2) and (1)]. It would be a laborious task to investigate all the possibilities of false internal homography that might arise from the existence of loan-words in the language that "break the rules"; and it is probable that, if any exist, they would be solved by whatever techniques are developed to deal with real homographs and polysemantic words.

The most likely source of difficulty would seem to be the transformations introduced under (3), by which, for example, НОЖ ("knife") would be changed into *НОЖЬ. It is a matter of orthographic convention that the nominative singular of masculine nouns and the genitive plural of feminines and neuters with stems in Ш, Щ, Ж, Ц and Ч are written without the "soft sign", whereas the nominative and accusative singular of feminine nouns, the imperative singular, the second person of the present indicative and the infinitive take the "soft sign" after these consonants. Thus, "ПЛАЧ" (nom. sing. "weeping"): but "ПЛАЧЬ" (imperative: "weep"); or, "ЛОЖ" (gen. plur. "couch"): but "ЛОЖЬ" (nom. sing. "lie, falsehood"). The effect of (3) would be to destroy the orthographic difference between these pairs. It is probable that all such instances of false homography would be soluble at the syntactic level. Should there exist, however, in the dictionary two stems ending (in their transformed spelling) in *ШЬ, *ЩЬ, *ЖЬ, *ЦЬ, *ЧЬ, one of which was the stem of a masculine noun and the other the stem of a feminine noun and should one of the two words occur in the text in the nominative singular without any adjectival concord or other syntactic feature to relate it to one or the other stem, the problem created would be identical with that presented by a pair of nouns which in their normal orthography have partially isomorphic paradigms. If, however, it is felt that the principle of not throwing away potentially

[3] For an alternative approach, see A.G. Oettinger, *Automatic Language Translation*, pp. 138 ff., (Harvard University Press, 1960).

distinctive features should be followed, it is possible to reject the transformation proposed under (3) and put two entries in the dictionary for all nouns (like "НОЖ") whose stems end in one of the five consonants in question and which do not have the "soft sign" in the nominative singular. The stem without the "soft sign" (in the transformed spelling) would be a short entry on the pattern of the entries for "false stems", while the stem with the "soft sign" would have coded with it in the dictionary all the necessary grammatical and lexical information. It would be the latter stem which would appear in those forms of the words to which the rules of 2 and 4 [and hence also of (1)] would apply. In this paper it has seemed better merely to give a brief general outline of the orthographical transformations proposed and their effect on the morphological analysis. Further refinements will suggest themselves immediately to the reader with some knowledge of Russian.

Received December 10, 1960