INTERNATIONAL CONFERENCE ON
INTELLECTUAL CAPITAL MANAGEMENT
7, 8 OCTOBER 2009
ZANJAN SCIENCE AND TECHNOLOGY PARK - INSTITUTE FOR ADVANCED STUDIES IN BASIC SCIENCES (IASBS)

# Interlanguage normalization of some groups of word forms in Tajik language

Shoira Pulatova
Technological University Of Tajikistan, Dushanbe, Tajikistan

**Abstract:** This article is about mathematical models which allow automatically transforming the class of Tajik words with the complex structure (inflexional forms) into the fragments of sentences in Tajik, Russian and English languages. Received fragments of Tajik sentences consist of the words witch have a simple structure (analytical form). This permits to simplify translation of these fragments into languages with analytical structure, particularly in English language. Russian language is used as the meta-language (intermediate auxiliary language), and mainly for commenting.

**Keywords:** transforming of words complex/simple structure of words

One of peculiarities of Tajik language is that the most of Tajik words are semantically equivalent to a fragment (or whole) of sentence. Such words are formed [1] through combination a number of postfixes to word stem; each of them gives it certain shade of meaning. The translation of such words according to a scheme

*word of input language -> word of output language*

is not always possible.

Therefore it is sometimes reasonable to replace such words by their equivalent expressions, thereby minimizing the number of postfixes attached to the words in the derived fragment of sentence.

This process is named INTRALINGUISTIC (INTRALANGUAGE) NORMALISATION. The expression, obtained after normalization, consists of words of simpler construction that facilitates its translation into another language. Intralinguistic normalization is the material for translation to another language as well as a sense comment of the word within a comment of the Tajik language simultaneously.

Intralinguistic normalization of Tajik words was studied in [2, 3]. Research [3] also gives algorithms for the normalization of Tajik words formed by word stems of numerals and their translation into English language.

Along with Intralinguistic normalization also there is an operation of INTERLANGUAGE NORMALIZATION, when the wordform of input language is in compliance with the grammar construction of words of output language.

It should be noted that the automatic translation of the Tajik words into another language, are preceded by phases, which are executed sequentially:

- Morphological analysis of a translated word;
- Intralinguistic normalization of the word;
- Postediting of the normalized expression;

Articles [2, 6-9] completely described the mathematical model of morphological analysis of Tajik words formed from word stems of all embracing meaningful parts of speech (noun parts of speech, verb and adverb).

It should be noted that the Intralinguistic normalization of Tajik words done on a group of postfixes, which are listed below.

**Admissible group of postfixes for Intralinguistic normalization of words**

In Tajik language there are six groups of postfixes, which allow to perform process of words Intralinguistic normalization [2]:

1. Group of possessive (pronominal) suffixes:

Table No.1

| number / person | singular | plural |
|---|---|---|
| 1 | *-ам, -ям* | *-амон, -ямон* |
| 2 | *-ат, -ят* | *-атон, -ятон* |
| 3 | *-аш, -яш* | *-ашон, -яшон* |

2. Group of predicative links (the personal or verbal endings).

a) Aorist (inflective verb form, consisting of word stem of the present time and the personal endings):

Table No.2

| number / person | singular | plural |
|---|---|---|
| 1 | *-ам, -ям* | *-ем* |

| | | |
|---|---|---|
| 2 | *-ӣ(и), -ӣӣ(ӣи)* | *-ед* |
| 3 | *-ад, -яд* | *-анд, -янд* |

b) The Imperative mood has following endings:

Table No.3

| number / person | singular | plural |
|---|---|---|
| 1 | *-ам, -ям* | *-ем* |
| 2 | - | *-ед(-етон)* |
| 3 | *-ад, -яд* | *-анд, -янд* |

c) Verbs of Simple past tense end as follows:

Table No.4

| number / person | singular | plural |
|---|---|---|
| 1 | *-ам* | *-ем* |
| 2 | *-ӣ* | *-ед(-етон)* |
| 3 | - | *-анд* |

3. Suffixes *-анги, -ангӣ, -янги, -янгӣ,* forming an adverb.
4. The conjunction *-у, -ю, -ву*.
5. Suffix of superlative degree *-тарин*
6. Diminutive-endearment suffixes–*ча, -ича, -ак*.

Here the function [3] determines the presence in the Tajik word *S* postfix $x$ to allow Intralinguistic normalization:

$$\delta(x) = \begin{cases} x, if\ \ x \in S \\ \wedge, if\ \ x \notin S \end{cases} \qquad (1)$$

Then function [3]

$$\delta_1(x, y) = \begin{cases} y, if\ \delta(x) = x \\ \wedge, if\ \delta(x) = \wedge \end{cases} \qquad (2)$$

allows to replace postfix $x$ by semantically equivalent expression $y$. Formulas (1) and (2) define the algorithm of Intralinguistic normalization of Tajik words.

Here the function [3] determines the occurrence of word $\alpha$ in the sentence fragment *R (N)*:

$$\delta'(\alpha) = \begin{cases} \alpha, if\ \alpha \in R(N) \\ \wedge, if\ \alpha \notin R(N) \end{cases} \qquad (3)$$

Then function [3]

$$\delta_1'(\alpha, \beta) = \begin{cases} \beta, if\ \delta'(\alpha) = \alpha \\ \wedge, if\ \delta'(\alpha) = \wedge \end{cases} \qquad (4)$$

allows to replace word $\alpha$ by the semantically equivalent word $\beta$.

Here $R(N)$ is translation of word $S$ in Russian language. Russian language is used as an intermediate-auxiliary language. By means of Russian language you can present an expression *R (N)* as a sentence fragment *E (R)* of English language.

Formulas (3) and (4) determine the Interalanguage normalization algorithm of Tajik words into another language, in particular, into English language. The following models are constructed on the basis of formulas (3) and (4):

1) Mathematical model of Interalanguage normalization of Tajik words formed from the word stems of the noun parts of speech,

2) Mathematical model of Interalanguage normalization of Tajik words formed from the word stems of verb;

Description of these models is specified in [4, 5].

Stated below formulas are the main results of Interalanguage normalization of Tajik words, formed from the word stems of the noun parts of speech and verb.

**Model of Interlanguage normalization of Tajik words formed from the word stems of the noun parts of speech**

Any Tajik word *S* with generalized word stem which belonging to a noun, an adjective, a numeral and a pronoun in the general case is represented in the form [2]:

$$S = O_1 \oplus \delta(q_1) \oplus \delta(m_i) \oplus \delta(n) \oplus \delta(q_2) \oplus \delta(m_k) \oplus \delta(l_j) \oplus \delta(q_3) \qquad (5)$$

This wordform structure appropriates the following Interalanguage normalize expression [2]:

$$\begin{aligned} N(S) = {} & \delta_1(l_j, M_j\_) \oplus \delta_1(n, \text{дар}\_) \oplus O_1 \oplus \delta(q_1) \oplus \\ & \delta_1(m_i, u\_M_i) \oplus \delta_1(n, \_\text{будагӣ}) \oplus \delta(q_2) \oplus \\ & \delta_1(m_k, u\_M_k) \oplus \delta_1(l_j, \_\text{мебош} \oplus l_j) \oplus \delta(q_3) \end{aligned} \qquad (6)$$

Then Interlanguage normalization of formula (6) at the English language has the form [4]:

$$\begin{aligned} E(R) = {} & \delta_1(y_j, y_j'\_) \oplus w_{tj}\_ \oplus \delta_1(d, d'\_) \oplus \delta_1(u, u'\_) \oplus w_{tj}\_ \oplus \\ & \delta_1(v, v'\_) \oplus \delta_1(h_i, h_i'\_) \oplus the\ most\ (least) \oplus O_1' \oplus \delta(q_1') \oplus \\ & \delta(q_2') \oplus \delta(q_3') \oplus \delta_1(v, v'\_) \oplus \delta_1(g_i, g_i') \end{aligned} \qquad (7)$$

**Model of Interalanguage normalization of Tajik words formed from the word stems of the present/past tense verb**

As it is known [1], the verb in the Tajik language has two word stems:

1) the word stem of present tense (imperative for the 2 nd person singular);

2) the word stem of past tense (simple past tense for the 2 nd person singular);

On this basis, we construct mathematical models of Intralanguage and Interlanguage normalization.

Let *S* be the word, which is the word stem of the verb present tense $\Gamma_н$. Then any Tajik word with the generalized word stem $\Gamma_н$ fits in the formula [2]:

$$S = \delta(p) \oplus \Gamma_н \oplus \delta(l_j) \oplus \delta(m_i) \oplus \delta(q) \qquad (8)$$

Word *S*, defined by the formula (8), has Intralanguage normalization of type [2]:

$$N(S) = \delta_1(l_j, M_j \_) \oplus \delta_1(m_i, M_i \oplus po \_) \oplus \delta(p) \oplus \Gamma_n \oplus \delta(l_j) \oplus \delta(q) \qquad (9)$$

Displaying the formula (9) in English (Interlanguage normalization) is given by [5]:

$$E(R) = w_{tj} \oplus \delta_1'(y_j, y_j' \_) \oplus w_{tj} \oplus \delta_1'(h_i, h_i' \_) \oplus V_{pr} \oplus \delta'(e) \_ \oplus \delta_1'(z_i, z_i' \_) \oplus \delta_1'(g_i, g_i' \_) \oplus \delta_1'(q, q') \qquad (10)$$

Now we consider the word *S*, which is the word stem of verb past tense $\Gamma_n$. Similar words in the general case can be presented by the formula [2]:

$$S = \delta(p) \oplus \Gamma_n \oplus \delta(r) \oplus l_j \oplus m_i \oplus \delta(q) \qquad (11)$$

The Intralanguage normalization of formula (11) has the form [2]:

$$N(S) = \delta_1(l_j, M_j \_) \oplus \delta_1(m_i, M_i \oplus po \_) \oplus \delta(p) \oplus \Gamma_n \oplus \delta(l_j) \oplus \delta(q) \qquad (12)$$

The Interlanguage normalization of formula (12) cam be presented in the form of:

$$E(R) = w_{tj} \_ \oplus \delta_1'(y_j, y_j' \_) \oplus w_{tj} \_ \oplus \delta_1'(h_i, h_i' \_) \oplus after \_ \oplus V_{ps} \oplus \delta'(e) \_ \oplus \delta_1'(z_i, z_i' \_) \oplus \delta_1'(g_i, g_i' \_) \oplus \delta_1'(q, q') \qquad (13)$$

[1] Sergey Arzumanov, Orif Jalolov (1969). “Tajik Language”. Dushanbe: Irfon.
[2] Mizrob Ismoilov. (1994). “Fundamentals of automatic morphological analysis of Tajik words”. Dushanbe: PIO NPICenter.
[3] Rano Ismoilova.(1998). “Simulation of automatic translation from the Tajik language to the English language word formed of bases of numerals”. Dushanbe: Academy of Sciences of Tajikistan.
[4] Mizrob Ismoilov, Shoira Pulatova. (2007). “Interlanguage and Intralanguage normalization groups of Tajik words, formed on basis of all-embracing meaningful parts of the speech. Dushanbe: NPICenter, No. 5 (1762).
[5] Shoira Pulatova. (2008). “Interlanguage and Intralanguage normalization groups of Tajik words formed on bases of the verb”. Dushanbe: NPICenter, No. 05 (1773).
[6] Mizrob Ismoilov. (1998). “Mathematical model of morphological analysis and synthesis of Tajik words”. Dushanbe: Academy of Sciences of Tajikistan, Vol.41, No. 9.
[7] Mizrob Ismoilov, Faruh Abdulloev. (1998). “Mathematical model of morphological analysis and synthesis of Tajik words formed on basis of nouns”. Dushanber: NPICenter, Issue No. 1, No. 57 (1201).
[8] Mizrob Ismoilov, Faruh Abdulloev. (1999) “Mathematical model of morphological analysis and synthesis of Tajik words, formed on basis of adjectives”. Dushanbe: NPICenter, Issue No. 1, No. 12 (1254).
[9] Mizrob Ismoilov, Faruh Abdulloev. (1999). “Formal grammar Tajik word-formation on basis of verb past tense”. Dushanbe: NPICenter, Issue No. 2, No.34 (1276).