AMERICAN JOURNALS
Open Access, Peer Reviewed Journals

American Journal of Technology and Applied Sciences
ISSN (E): 2832-1766 Volume 36, May - 2025

ADDITIONAL NEW MATHEMATICAL MODELS OF VERB PHRASEOLOGY FOR COMPUTER TRANSLATION FROM ENGLISH TO UZBEK

Muftakh Khakimov 1,
Vazira Bekova 1,
Ziyoviddin Sirojiddinov 1,
Akbar Omonov 2.
1National University of Uzbekistan of name Mirzo Ulugbek,
University Street 4, 100104
Tashkent, Republic of Uzbekistan
2Tashkent Pediatric Medical Institute,
Department of "Biophysics, Medical Informatics"
vazira.bekova16@gmail.com
orcid.org/ 0009000865181312.

ABSTRACT

It is well known that the process of any computer translation system involves encoding the meaning of the input text in a natural language and re-encoding this meaning in the target language while maintaining semantic consistency with the source text. One approach to achieving this goal is the formalization of the grammatical structures of the natural languages involved in translation systems. More precisely, grammatical structures are based on linguistic rules that determine word structures and their classification according to parts of speech. Analyzing word structures enables the construction of logically sound mathematical models. This article examines how verbs in English and Uzbek form syntactic relationships and how verb categories are classified during the computer translation process. The study analyzes the key aspects of a morpho-analyzer that ensures accurate and meaningful automatic translation. Additionally, it classifies verb-related words and verb-forming affixes in both languages, along with their patterns and mathematical models.

KEYWORDS

Natural language, expandable input language, computer translation, database, English verb category, Uzbek verb category, additional new mathematical models, word weight.

Introduction

Each natural language (NL) is a complex system consisting of components that are neither mathematically structured nor formalized. However, through the processing of NL, it is possible to identify unstructured elements in the language and formalize them using a linear methodology. This process includes determining word structures, constructing logical-linguistic models based on word and sentence types, and developing mathematical models using a specialized meta-language. This methodology is referred to as the **degree of formalization of a language**. The degree of formalization, in turn, determines the level of semantic formalization of NL and the accuracy of the algorithm. A superficial understanding of NL formalization—i.e., perceiving a formalized language as an abstract, content-independent structure with a simple logical framework—leads to low efficiency in machine translation [1]. Formalization allows for the segmentation of a language into different components, the analysis of their interrelations, and the characterization of its semantic structure.

There are numerous shortcomings in automatic translation systems (Google Translate, Microsoft Translator, DeepL Translator, Yandex Translate, Amazon Translate) when translating between English and Uzbek:

Since Uzbek is an **agglutinative language**, word formation and the semantic structure of sentences depend on the affixes attached to words in each sentence. This difference between English and Uzbek has not been resolved in the automatic translation systems mentioned above. There is no tool for formalizing natural languages to address this issue. For high-quality translation between languages, particularly between English and Uzbek, natural languages have not been formalized, and their various models have not been developed.To address the shortcomings mentioned above, the **first author of this paper has developed a specialized expandable input language** to formalize natural languages [2, 3]. As a result of creating this expandable input language, linguistic and mathematical models have been developed for parts of speech and general sentence structures in English, Russian, and Uzbek [4, 5].

This study adopts the first approach, focusing on the efficiency of transferring words and sentences from language A to language B, classified as class 0 according to Chomsky's classification [6].

1 Methods

To date, various models for translation programs have been developed by mathematicians and linguists. However, due to the lack of high-quality text translation between English and Uzbek in both directions, this research aims to represent the linguistic capabilities of both languages using an **expandable input language**. This approach allows for the development of additional new mathematical models by analyzing the grammatical and morphological structures of natural languages belonging to different classes.

The transformation from formalizing linguistic capabilities to modeling consists of the following stages:

Analysis of parts of speech in natural languages: This stage involves studying the morphological, syntactic, and semantic aspects of word categories such as nouns, adjectives, verbs, pronouns, adverbs, and numerals in both languages.

Formation of word chains: By identifying which prefixes, affixes, and suffixes are added to a root to form words in both languages, multiple word chains (A^1, A^2, ..., A^3) are constructed based on word formation.

Selection of logically and semantically valid chains: Only root and affix chains that are logically and semantically compatible in both languages are selected, leading to the construction of mathematical models.

The **functionality of words in natural languages (NL)** is reflected in their polysemy. Each word has a specific meaning in a given context, although some words may cause exceptions due to their inherent polysemy. This phenomenon leads to two main approaches when creating mathematical models of natural language:

Developing a unified system for linear processing of words and sentences.

Considering each word and sentence as an independent structure and processing them accordingly.

The following steps are also included in this research:

Studying natural language (NL);

Implementing the developed expandable input language;

Developing a semantic database;

Creating a bilingual database of terms and phrases specific to particular scientific fields;

Modeling NL;

Designing translation program algorithms and developing its software environment.

2.1 Mathematical model (Problem statement)

Using an expandable input language is the most effective approach for developing a bilingual translation program for computer translation. Therefore, in this article, new terminal symbols related to the verb category of both languages have been introduced into the terminal symbols of the expandable input language, as presented in Table 1. Based on these terminal symbols, we construct new mathematical models for a computer translation program by analyzing and synthesizing the verb categories of both languages.

New terminal symbols added to the expandable input language Table 1

Verb Types	Terminal characters		Verb Types	Terminal characters
Notional Verbs	G		Gerund	G13
Action verbs	G3		Modal verbs	G14
Mental verbs	G31		Present tense	G15
Perception verbs	G32		Future tense	G16
Speech verbs	G33		Past tense	G17
Gesture verbs	G34		Active voicei	G18
visual perception verbs	G35		Reciprocal voice	G181
Stative verbs	G4		Reciprocal voice	G182
Auxiliary verbs	G5		Causative voice	G183
Defective verbs	G6		Reflexive voice	G184
Light verbs	G7		Passive voice	G185
Linking verbs	G8		To be" verbs	G (D1) G(D2) G (D3)
Simple verbs	G9		Transitive verbs	G19
Compound verbs	G10		Intransitive verbs	G20
Paired (dual) verbs	G11		Regular verbs	G21
Infinitive	G12		Irregular verbs	G22

2.2. **Initial data (Descriptions of datasets)**

Developing an algorithm for the **verb category** in computer translation software by analyzing the verb categories of two natural languages and incorporating additional new terminologies into the **expandable input language**.

Creating additional new **mathematical models** for computer translation between **English and Uzbek** and vice versa using the **expandable input language**.

2.3 Computational algorithm (Solution method)

Algorithm for Translating from English to Uzbek

1. A specialized word database consisting of multiple tables has been built for computer translation between the two languages. Using this specialized database, the following sequential steps are performed.
2. The input word in English is analyzed using data retrieved from the database.
3. The output word in Uzbek is synthesized based on the corresponding data in the database.
4. The input word in English is segmented into its root (K) and affixes, and new mathematical models are constructed accordingly.
5. The output word in Uzbek is synthesized using the newly constructed mathematical models.
6. Based on the newly constructed mathematical models for both languages, weight coefficients for words and affixes are calculated.
7. Translation is carried out using weight coefficients that are either equal or very close to each other in both languages.

8. If no exactly matching weight coefficients are found between the two languages, translation is performed by selecting words with the closest weight coefficients from the specialized database built for computer translation.

2 Results

Based on the above-discussed points, we will conduct the following analysis for both languages. **(G) (national verbs)** — independent verbs — are categorized into **action verbs** and **state verbs** in both English and Uzbek, based on what they denote. In English, **(G)** verbs do not differentiate between action and state verbs in terms of lexical meaning. However, in Uzbek, **(G)** verbs are classified into **six types of action verbs** and **three types of state verbs**, depending on whether they indicate an action or a state in a sentence. This difference highlights the necessity of comparative analysis between the two languages. Although both natural languages have **verbs as an independent part of speech**, their grammatical and morphological structures differ due to belonging to **different language families**. Therefore, in this study, a **new set of terminological symbols** was introduced into the **Expandable Input Language (EIL)** developed by the first author of the paper, designed to formalize the operation of **TT (Translation Tool)** for computer translation. Based on these newly introduced terminological symbols, an **algorithm** was created for each **verb category** in both languages. In this paper, new **mathematical models** and **numerical coefficients** for **notional verbs** (independent verbs) and **auxiliary verbs** in both languages are introduced. The weight values of words belonging to each verb type are provided in the tables below. The **(G) (notional verbs)** in English and Uzbek are similar because, in both languages, **(G) verbs** perform the main **semantic function** in a sentence with their inherent meaning. In the field of **Information Technology (IT)**, such verbs are used to describe processes like **running programs, managing networks, compiling code, and processing data**.

The weight values of parts of speech in any natural language are assigned as defined in **[7]**.

Words associated with noun (C) – 0.1;
- Words associated with adjective (P) – 0.2;
- Words associated with verb (G) – 0.3;
- Words associated with adverb (N) – 0.4;
- Words associated with pronoun (M) – 0.5;
- Words associated with numeral (F) – 0.6.
- Words associated with dependent closed word classes (U, D, Y, L) - 0.07.

The numbering of word classes shown above helps to calculate the weight of the two languages. For example,

tables below indicate that the weight of the declarative, interrogative and negative sentences in the English language,

which include pronoun differentiates from those in Uzbek. The mathematical models of English interrogative and

negative sentences, the weight of the word-forming affixes also differ from each other in both languages.

(MM) - Mathematical model; V1 - word root weight; V3 - weight of affixes in a word. Signs mean: - $\oplus$ joining operation, $\downarrow$- operation of possible “connection” or “not connections” a component following it. $ selection operation, syntax is $[<i>,<1/h>] [3].

In our previous studies on **computer translation systems [8]**, we analyzed **noun categories** in both languages, developing **new mathematical models** and computing **weight coefficients**. In this paper, we conduct **analysis and synthesis** of the **verb category (G)** in English and Uzbek, deriving new **mathematical models** and calculating their **weight coefficients** for each verb type. For instance, the **English verb** <store> translates to **Uzbek** as <saqlamoq>. The **mathematical model** for this verb in English is represented as:

$$G = \$_{[i,1\text{-}h]}G_{[i]} \text{ with a weight coefficient of } \mathbf{0.3}.$$

G The corresponding **Uzbek translation** of this English verb is **G10**, where: <saqla> is the **verb root (G)** <moq> is the **verbal noun-forming suffix** To derive the **mathematical model** for the **Uzbek translation**, we first analyze the translated word based on **[4]**, formulating its model as:

$$\mathbf{(G, G\ (A1))} = \$_{[i,1\text{-}h]}\, G[i] \oplus \downarrow \$_{[j,1\text{-}h[]]}G(A1_{[j]}).$$

This model helps the **computer translation system** process **input text (EVX)** and generate the **translated output text (EVIX)**. The system **first analyzes** the input word using the **mathematical model**, then **synthesizes** the translation by comparing weight coefficients across both languages.

$$\mathbf{(G, G\ (A1))} = \$_{[i,1\text{-}h]}\, G[i] \oplus \downarrow \$_{[j,1\text{-}h[]]}G(A1_{[j]})$$

For the translated Uzbek word **G10**: The **root weight** for <saqla> is **0.3** The **suffix weight coefficient** for <moq> is **0.10221** The **results** are presented as an example in **Table 2**.

New mathematical model and weight coefficient of the action verb of an independent verb in two languages Table 2

№	Ingliz tilida	(MM)	V1	V3	O'zbek tilida	(MM)	V1	V3
1.	store	**(G)**	**0.3**	**0**	**Saqla+moq**	**G10**	**0.2**	**0.10221**

The translations of **English (G) verbs** into **Uzbek** are presented in **Table 3**. For example, the **English verb** <execute> is a **compound verb** in Uzbek: **Amal+ga osh+ir+moq = (amal) asos = K (C), (-ga) jo'nalish kelisshigi = (X3), (osh) asos = K (C), (ir) G183 (A1)** where: **(Amal)** is the **root noun K (C) (-ga)** is the **dative case suffix (X3) (osh)** is another **root K (C) (-ir)** is the **causative verb-forming affix G183 (A1) (-moq)** is the **verbal noun-forming affix G3 (A1)** When translating from **English to Uzbek**, the structure follows **composition rules**: **(C+G) = G10** Thus, <execute> becomes a **compound verb (G19)** in **Uzbek**. The **mathematical model** for the English verb **remains (G)**. However, when translated into **Uzbek**, it follows the formula:

$$\mathbf{K\ (C) \oplus X3 \oplus K\ (C) \oplus G183\ (A1) \oplus G\ (A1) = G10}\ A$$

New **mathematical model** for **computer translation** was developed by analyzing and synthesizing translations. Weight coefficients for **word categories and affixes** were previously established by the **first author** in [4]. This paper utilizes **[4]** for computing **weights** across **both languages**. The English verb <execute> has: A **weight coefficient** of **0.3**. Since the **Uzbek translation** is a **compound verb**, it contains **two root words**, requiring separate weight calculations: **K (C) - 0.1 + K (C) - 0.1 = 0.2;** For the suffixes: **X3- 0.10201 + ir - 0.10217 + moq - 0.10221 = 0.3063** Thus, the **computer translation system** must **automatically generate weight coefficients** for **words and affixes** in both languages. Since **English and Uzbek verbs** are categorized into **action and state verbs**, weight coefficients for **auxiliary verbs** in both languages were calculated using the same methodology. To perform **English-to-Uzbek translation**, the proposed algorithm was extended to include **(G3), (G4) and (G5)** verb types. The final **results** are provided in the **table below**.

Algorithmic mathematical models and weight coefficients of words belonging to the verb phrase class in two languages Table 3

№		Ingliz tilida	(MM)	V1	V3	O'zbek tilida	**(MM)**	**V1**	V3
1.	(G)	store	**(G)**	**0.3**	**0**	**Saqlamoq**	**G19**	**0.3**	**0.10221**
2.	**(G)**	**execute**	**(G)**	**0.3**	**0**	**Amalga oshirmoq**	**(G19)**	**0.2**	**0.3063**
3.	(G3)	Install	(G3)	0.3	0	O'rnat+moq	**(G3)**	**0.3**	**0.10221**
4.	G4	Exist	(G4)	0.3	0	Mavjud bo'lmoq	**(G4)**	**0.6**	**0.10221**
5.	G5	I am	**(G51)**	**0.5**	**0.3**	**Men**	**(G51)**	**0.5**	**0**
6	G5	He is	**(G52)**	**0.5**	**0.3**	**U**	**(G52)**	**0.5**	**0**
7.	G5	You	**(G53)**	**0.5**	**0.3**	**Siz**	**(G53)**	**0.5**	**0**
8.	G5	has	**(G54)**	**0.3**	**0**	**bor**	**(G54)**	**0.3**	**0**
9.	G5	have	**(G55)**	**0.3**	**0**	**bor**	**(G55)**	**0.3**	**0**
10.	G5	do	**(G56)**	**0.3**	**0**	**qilmoq**	**(G56)**	**0.3**	**0**
11.	G5	does	**(G57)**	**0.3**	**0**	**qilmoq**	**(G57)**	**0.3**	**0**

3 Discussion

Unlike other Turkic languages, **Uzbek** is considered a **low-resource language** and has a highly **agglutinative** structure. A single word can form an entire sentence. There are **insufficient rule-based machine translation resources** for **Uzbek**. However, **significant progress** has been made in **Turkic languages** such as **Turkish** and **Kazakh** in this field. For example, **sentiment analysis** has been conducted in [10]. In the era of **globalization**, despite all challenges, **Uzbek** must become an **active participant** in the **information community**. Several studies have been conducted on **Uzbek morphology** and **word stem identification**, such as [11, 12]. There are **very few** studies on **formalizing natural languages**, but numerous papers focus on different aspects of **Turkic languages**. For example, **sentiment analysis** in **Kazakh and Russian** has been conducted in [13], and **ontology-based sentiment analysis of Kazakh sentences** has been performed.

4 Conclusion

In this paper, **new additional mathematical models** have been developed for **verbs** using the **Expandable Input Language**. The **significance** of this study lies in **achieving high translation accuracy** in **automatic translation** between **English** and **Uzbek**. For **future research**, we plan to **analyze the grammar and morphology** of other **parts of speech** in **English** and **Uzbek** to develop **additional mathematical models** for each category. With these **new mathematical models** based on the **Expandable Input Language**, we can **precisely develop** the **translation algorithm**. These models enable the **prediction of word and sentence alignment probabilities** between **English** and **Uzbek**. From this perspective, developing **new additional mathematical models** is **crucial** for achieving **high-quality machine translation** between **English and Uzbek**. The **remaining parts** of the **algorithm** presented in the upper section of this paper will be **fully developed** by conducting **further analysis** in both languages and **constructing additional mathematical models**.

References

1. Khakimov, M.Kh. Formal machine translation systems in a multi-languages situation. In: Materials of Republican – Scientific Conference «Modern Problems of Mathematics, Mechanics and Information Technologies», NUUz, Institute of Mathematics and IT AS RUz, pp. 297–301, Tashkent (2008)
2. Khakimov, M.Kh Expandable input language of mathematical modeling of natural language for multilingual situation of machine translation// - UzMU xabarlari No. 1, 2009, p 75-80
3. Khakimov, M.Kh The Ministry of Justice of the Republic of Uzbekistan. Official Bulletin Journal. No. 10 (238), 2022, pp. 110-113. Patent No. IAP 07121
4. Khakimov, M.Kh. Technology of Multilingual Modeled Computer Translator. Monograph // LAP LAMBERT Academic Publishing, Riga, 2019, 174 p
5. Mersaid Aripov, Muftakh Khakimov, Sanatbek Matlatipov, and Ziyoviddin Sirojiddinov Analysis and Processing of the Uzbek Language on the Multi-language Modelled Computer Translator Technology. In: Vetulani, Z., Paroubek, P., Kubis, M. (eds) Human Language Technology. Challenges for Computer Science and Linguistics. pp 81–95 LTC 2019. Lecture Notes in Computer Science, vol 13212. Springer, Cham. https://doi.org/10.1007/978-3-031-05328-3_6
6. Chomshy, N.: Formal properties of grammars. In: Handbook of Mathematical Psychology, vol. 2, pp. 323–418. Wiley, New York (1963)
7. Khakimov M. Kh., Sirojiddinov Z. Sh. Computer algorithmization in multi-language modelled translator technology // Modern problems of applied mathematics and information technologies al-Khwarizmi. – 2021, P.2.
8. Khakimov M.Kh., Bekova V.G. Morphological analysis of nouns in english to uzbek machine translation // problems of computational and applied mathematics No. 4(58) 2024
9. Khakimov, M.Kh. The extensible source language of mathematical modelling of a natural language for a multi-languages situation of machine translation. In: News NUUz, vol. 1, pp. 80–85, Tashkent (2009)
10. Marciniak, M., Mykowiecka, A. Representation of Uzbek morphology in prolog. In: Marciniak, M., Mykowiecka, A. (eds.) Aspects of Natural Language Processing. LNCS, vol. 5070. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04735-0_4
11. Matlatipov, S., Tukeyev, U., Aripov, M. Towards the Uzbek language endings as a language resource. In: Hernes, M., Wojtkiewicz, K., Szczerbicki, E. (eds.) ICCCI 2020. CCIS, vol. 1287, pp. 729–740. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63119-2_59
12. Sakenovich, N.S., Zharmagambetov, A.S. On one approach of solving sentiment analysis task for Kazakh and Russian languages using deep learning. In: Nguyen, N.-T., Manolopoulos, Y., Iliadis, L., Trawi ´nski, B. (eds.) ICCCI 2016. LNCS (LNAI), vol. 9876, pp. 537–545. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45246-3_51
13. Tuzov, V.A. Matematicheskaja's ases language model, p. 176. LGU (1984)