ИНТЕРНАУКА
internauka.org

СБОРНИК СТАТЕЙ ПО МАТЕРИАЛАМ
CCX МЕЖДУНАРОДНОЙ
НАУЧНО- ПРАКТИЧЕСКОЙ КОНФЕРЕНЦИИ

ИННОВАЦИОННЫЕ ПОДХОДЫ В СОВРЕМЕННОЙ НАУКЕ

№6(210)

ISSN 2587-8603

Москва, 2026

ИНТЕРНАУКА
internauka.org

ИННОВАЦИОННЫЕ ПОДХОДЫ В СОВРЕМЕННОЙ НАУКЕ

Сборник статей по материалам CCX международной научно-практической конференции

№ 6 (210)
Март 2026 г.

Издается с июля 2017 года

Москва
2026

ИНТЕРНАУКА
internauka.org

INNOVATIVE APPROACHES IN THE MODERN SCIENCE

Proceedings of CCX international scientific-practical conference

№ 6 (210)
March 2026

Published since July 2017

Moscow
2026

УДК 08
ББК 94
И66

И66 Инновационные подходы в современной науке. сб. ст. по материалам CCX междунар. науч.-практ. конф. – № 6 (210). – М., Изд. «Интернаука», 2026. – 144 с.

ISSN 2587-8603

Оглавление

CONFERENCE PAPERS IN ENGLISH

SECTION 1.

TECHNICAL SCIENCES

A MODEL AND ALGORITHMIC APPROACH TO RESOLVING LEXICAL AMBIGUITY IN ENGLISH–KARAKALPAK MACHINE TRANSLATION

Allanyazov Rustem Baxavedinovich
Independent PhD Researcher,
Tashkent University of Information Technologies (TUIT),
Lecturer in Tashkent state university of oriental studies,
Uzbekistan, Tashkent

ABSTRACT

This paper presents a formal model of morphological analysis and synthesis aimed at resolving lexical-semantic ambiguity in English–Karakalpak machine translation. The proposed approach takes into account the typological differences between English and Karakalpak and integrates corpus data, Universal Dependencies annotation, WordNet, and semantic models. The formalization of the algorithms ensures accurate sense selection and the generation of grammatically correct target word forms**.**

Keywords: machine translation, lexical-semantic ambiguity, morphological analysis, agglutinative languages, Karakal-pak language, WordNet, Universal Dependencies.

I. INTRODUCTION

One of the key challenges in machine translation is lexical-semantic ambiguity, particularly when translating between languages of different typological structures. English is characterized by an analytic structure, whereas Karakalpak belongs to agglutinative languages, which entails increased complexity in morphological synthesis and the selection of appropriate target word forms.

The correctness of translating polysemous lexical units depends not only on the selection of their appropriate sense but also on the alignment of morphological and syntactic features in the target language. In this regard, there is a need to develop formalized algorithms for morphological analysis and synthesis that take into account the typological characteristics of the languages involved.

The aim of this study is to develop a formalized algorithmic model of morphological analysis and synthesis aimed at resolving lexical-semantic ambiguity in the English–Karakalpak machine translation system.

II. CORPUS AND LINGUISTIC BASIS

Parallel bilingual and multilingual corpora play an important role in the creation and improvement of statistical machine translation systems within the global research environment.

Corpus linguistics studies large collections of electronic texts used for linguistic analysis and the development of natural language processing systems.The importance of corpus-based methods for translation studies was demonstrated by Baker[1]. Particularly important is the definetion of a corpus proposed by John Sinclair: a corpus is “a collection of text excerpts in electronic form designed to represent linguistic reality” [2].

Corpora may differ in structure and purpose; however, for machine translation tasks, bilingual parallel annotated corpora are of the greatest importance. The English–Karakalpak corpus be-longs precisely to this type, which ensures the possibility of aligning translation units and training algorithms for the analysis and synthesis of word forms.

Corpus statistics provide information about word frequency, contextual usage, and collocations, which are essential for probabilistic sense selection in machine translation.

Multilingual corpora are typically divided into aligned (parallel) and non-aligned corpora. Parallel corpora contain explicit correspondences between translation units, usually at the sentence level[3]. Alignment facilitates the training of algorithms for establishing structural correspondences between English and Karakalpak sentences, making it possible to identify patterns of changes in word order, morphological markers, and syntactic dependencies.

The formal basis of the morphological analysis algorithms relies on a word structure model of the form w ::= p + r + a, which consists of a prefix, a root, and an affixal chain, reflecting the agglutinative nature of the Karakalpak language. For syntactic–morphological annotation, the Universal Dependencies framework is employed, in which each token is assigned a set of morphological and syntactic features [4]. These features form a scoring

matrix ,which is subsequently transformed into a weight matrix W determined by a regression model.

To eliminate semantic ambiguity, lexical-semantic networks such as WordNet are employed, where words are organized into synsets representing conceptual relations [5]. Fillmore's theory of case grammar introduces semantic role structures ("agent," "patient," "instrument"), enabling the algorithm to select appropriate word forms according to their semantic roles [6]. Within the Meaning–Text Theory, developed by A.K. Zholkovsky and I.A. Mel'čuk, the transformation of deep semantic representations into surface syntactic structures is formally described [7]. Knuth's attribute grammar formalizes the correspondence between syntactic structures and the semantic attributes of expressions [8].

Thus, the algorithms for word analysis and synthesis integrate:

(1) corpus data,

(2) formal morphological models,

(3) syntactic annotations,

(4) semantic networks and role structures.

Together, these components ensure effective semantic disambiguation and the selection of grammatically and semantically appropriate word forms in English–Karakalpak machine translation.

III. FORMAL MODEL OF MORPHOLOGICAL PROCESSING

The proposed system is implemented as a formal algorithmic pipeline for linguistic data processing.To ensure strict structural organization of the developed procedures, the algorithms for morphological analysis and synthesis in this paper are presented as formalized models integrating linguistic and mathematical components.

Classical machine translation systems include three main stages: morphological analysis, feature transfer, and morphological synthesis [9]. Each of these stages contributes to the formation of a correct output word form and must be strictly aligned with the model of multivalued mappings.

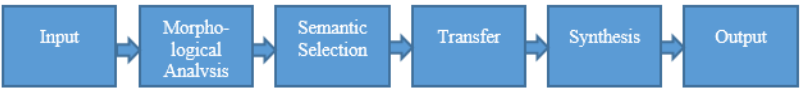

Figure 1 – Algorithmic Pipeline for Processing the Input Sentence

Figure 1.

The morphological analysis algorithm begins with the tokenization of the input English text, followed by lemmatization aimed at identifying the

base form of the word and its set of grammatical features. Formally, the result of the analysis can be represented as a mapping

$$w_{\text{i}} \rightarrow (lemma_{\text{i}}, F_{\text{i}}), \tag{1}$$

where w_i is the input word form, and F_i – is the set of its morphological features (part of speech, number, tense, mood, etc.).For the English language, a particular difficulty is posed by the presence of homonymous and multifunctional forms, which requires taking contextual features into account and using probabilistic estimates derived from the corpus.

The selection of the correct morphological analysis is carried out on the basis of the word's usage context and can be formalized as a problem of maximizing the posterior probability:

$$\widehat{F_i} = \arg\max_{F \in \mathcal{F}} P\,(\,F \mid w_i, c_i\,) \tag{2}$$

where c_i denotes the local and global context of the word within the sentence. The elaboration of the semantic sense selection mechanism is based on a probabilistic model. Such a probabilistic approach is widely applied in statistical and hybrid machine translation systems.

3.1 Algorithm 1 – Morphological Analysis of the English Text

Let the input sentence be:

$$S_{en} = \{w_1 w_2, \ldots., w_n\} \tag{3}$$

For each token w_i the following is performed:

1. Lemmatization:

$$lemma_i = L(w_i) \tag{4}$$

2. Determination of grammatical features:

$$F_I = \{POS_{I,} tense_{i,} number_i, role_i\} \tag{5}$$

3. Construction of the morphological structure:

$$A_i = (lemma_i, F_i) \tag{6}$$

4. Algorithm output:

$$A = \{A_1, A_2, \ldots A_n\} \tag{7}$$

3.2 Algorithm 2 – Semantic Sense Selection

Semantic sense selection for a polysemous lexeme:

$$w_i = \{sense_1, sense_2, \ldots\ldots sense_k\} \tag{8}$$

The sense is selected according to the following formula:

$$sense^* = arg \max_{sense_j} P\ (sense_j|c, F_i) \tag{9}$$

where $sense_j$ – denotes the possible meanings of the word in the lexical-semantic database.

Input:

Word form w_i,

Set of morphological features F_i,

Sentence context c.

Algorithm steps:

1. Generation of a set of possible senses:

$$w_i = \{sense_1, sense_2, \ldots\ldots sense_k\} \tag{10}$$

2. Computation of the probability of each sense:

$$P = (sense_j|c, F_j) \tag{11}$$

3. Selection of the optimal sense:

$$sense^* = arg \max_{sense_j} P\ (sense_j|c, F_i) \tag{12}$$

4. Transfer of the selected sense to the semantic transfer module.

Output:

Semantically interpreted unit

$$(w_i, sense^*) \tag{13}$$

In English–Karakalpak translation, this approach is applied at the stage of English text analysis in order to determine the correct sense of a word prior to transfer and morphological synthesis. The selected sense is then matched with the equivalent lexical unit in the Karakalpak language, taking into account its morphological and semantic characteristics.

3.3 Semantic Transfer

After completing the morphological analysis and selecting the sense of the lexical unit, the stage of semantic transfer is carried out, which consists in mapping the structural and semantic features of the source language onto the corresponding categories of the Karakalpak language.

The transfer is formalized as a mapping:

$$T: (r_{en}, G_{en,} s^*) \rightarrow (r_{kk,} G_{kk}) \quad (14)$$

where

r_{en} – is the root of the source word,

G_{en} – are the grammatical features of the English word,

s^* - is the selected sense,

$r_{kk,} G_{kk}$ - are the corresponding elements of the target language.

The result of morphological analysis is the pair ($\text{lemma}_{\text{en}}, \text{F}_{\text{en}}$), which at the transfer stage is transformed into ($\text{lemma}_{\text{kk}}, \text{F}_{\text{kk}}$) and used as input for the morphological synthesis algorithm.

$$\text{T}: (\text{lemma}_{\text{en}}, \text{F}_{\text{en}}) \rightarrow (\text{lemma}_{\text{kk}}, \text{F}_{\text{kk}}) \quad (15)$$

where F_{en} and F_{kk} represent vectors of morphological features of the source and target languages, respectively. To demonstrate the operation of the proposed algorithmic pipeline, consider the following sentence:

English: He booked the ticket.

The word booked is polysemous and may be interpreted as:

– “reserved”;

– “recorded”;

– “ordered”;

– “registered.”

At the stage of morphological analysis, the lemma book and the set of grammatical features are identified: part of speech – verb; tense – past; number – singular; person – third.

At the stage of semantic selection, the meaning is уточняется using contextual features. The presence of the direct object ticket activates the WordNet synset associated with the action of reservation. The probabilistic model selects the sense corresponding to the act of booking.

As a result of semantic transfer, a grammatical vector of the target language is formed, including features of past tense and agreement with the subject.

The resulting lexico-grammatical set is transmitted to the morphological synthesis module, where the correct word form in the Karakalpak lan-

guage is generated, taking into account its agglutinative structure and morphophonological rules.

The stage of morphological synthesis represents the process of constructing the target word form based on the root and an ordered sequence of affixes. Given the agglutinative nature of the Karakalpak language, synthesis can be represented as follows:

$$w_{kk} = r \oplus a_1 \oplus a_2 \oplus \cdots \oplus a_m \tag{16}$$

This approach corresponds to formal models of morphology based on finite-state automata, which are successfully applied to the processing of Turkic languages.

The selection of specific affixes is carried out using corpus statistics and contextual features. For this purpose, a probabilistic model is employed to determine the most likely affix in a given context:

$$\hat{a} = \arg\max_{a \in A} P\,(\,a \mid stem, F_{kk}, c\,), \tag{17}$$

where is the set of permissible affixes, and c denotes the context of word usage.Such models are employed in statistical and neural machine translation systems for languages with rich morphology.

$$\mathrm{T}: (\mathrm{lemma}_{\mathrm{en}}, \mathrm{F}_{\mathrm{en}}) \rightarrow (\mathrm{lemma}_{\mathrm{kk}}, \mathrm{F}_{\mathrm{kk}}) \tag{18}$$

where F_{kk} includes information about case, number, possession, tense, and other grammatical categories.

3.5 Algorithm 3 – Morphological Synthesis of the Karakalpak Word Form.

The Karakalpak word form is modeled as a path in an affixal graph:

$$w_{kk} = r \rightarrow a_1 \rightarrow a_2 \rightarrow \cdots \rightarrow a_m \tag{19}$$

where:

r – root

a_j – affixes,

The order is determined by the grammatical vector:

$$G_i = \{case, number, person, derivation\} \tag{20}$$

Synthesis is formalized as a function:

$$w_{kk} = G(r_{kk,} G_i) \tag{21}$$

Input:
root r
semantically selected lemmatic equivalent

$$lemma_{kk} \tag{22}$$

Grammatical vector G_i
Output:
Correct word form w_{kk}
Step 1
Determine the permissible grammatical categories from the vector. G_i
Step 2
Construct a valid path in the affixal graph:

$$Path = \{a_j \in A | a_j \; is \; compatible \; with \; \text{c} \; G_i\} \tag{23}$$

Step 3
Arrange the affixes in accordance with the morphological rules of the Karakalpak language:

$$a_1 \prec \; a_2 \; \prec \cdots \prec a_m \tag{24}$$

Step 4
Check phonetic and morphophonological constraints (vowel harmony, assimilation).
Step 5
Generate the final word form:

$$w_{kk} = r \; \oplus a_1 \oplus a_2 \oplus \cdots \oplus a_m \tag{25}$$

Particular importance is given to the consideration of morphophonological rules related to vowel harmony, consonant assimilation, and phoneme alternation characteristic of the Karakalpak language. These rules are applied at the final stage of synthesis and ensure the formation of a normative word form.

IV. HYBRID AND NEURAL EXTENSIONS

Modern machine translation systems often combine rule-based, statistical, and neural approaches.Neural models make it possible to automatically detect morpheme boundaries and correct analysis errors, especially under conditions of limited corpus resources.

It should be noted that errors arising at the stage of morphological analysis may propagate to subsequent levels of processing and lead to distortions in syntactic and semantic interpretation. Therefore, analysis and synthesis algorithms are considered interrelated components of a unified machine translation system operating in an end-to-end data processing mode.

Thus, morphological analysis and synthesis ensure the formal representation and generation of word forms in English–Karakalpak machine translation. However, the correctness of morphological generation directly depends on the accurate selection of the sense of a polysemous lexical unit.

V. SCIENTIFIC NOVELTY

Scientific novelty of the study:

- a formal model for the integration of corpus-based and semantic data is proposed;
- an algorithm for morphological synthesis is developed, taking into account the agglutinative structure of the Karakalpak language;
- a probabilistic mechanism for word form selection is formalized.

VI. PRACTICAL SIGNIFICANCE

The practical significance of the study lies in the possibility of integrating the proposed model into hybrid and neural machine translation systems for low-resource agglutinative languages.

VII. CONCLUSION

This study proposed a formalized model of morphological analysis and synthesis for resolving lexical-semantic ambiguity in English–Karakalpak machine translation. The model integrates corpus statistics, Universal Dependencies annotation, WordNet semantic networks, and probabilistic sense selection. The proposed approach enables accurate generation of Karakalpak word forms and provides a methodological basis for hybrid and neural machine translation systems for low-resource agglutinative languages.

References:

1. M. Baker, “Corpus Linguistics and Translation Studies: Implications and Applica-tions,” in Text and Technology, pp. 233–250, June 1993.
2. J. Sinclair, Corpus, Concordance, Collocation. Oxford: Oxford University Press, p.100. 1991.
3. Jiang Chao, “Studying the Translation of Chinese Political Terms into Russian Using a Parallel Corpus,” Political Linguistics no. 2 (104), 2024, UDC 811.581’373+811.581’42+81’25.
4. M.-C. de Marneffe, C.D. Manning, J. Nivre, and D. Zeman, “Universal Dependencies,” Computational Linguistics, 47(2): pp. 255–308. 2021.

5. G.A. Miller, “WordNet: A Lexical Database for English,” Communications of the ACM, vol. 38, no. 11,pp. 39–41, 1995.
6. C.J. Fillmore, “The Case for Case,” in Universals in Linguistic Theory, E. Bach and R.T. Harms, Eds. New York: Holt, Rinehart and Winston, 1968, pp. 1–88.
7. A.K. Zholkovsky and I.A. Mel’čuk, Language: From Meaning to Text. pp.7-174, 2012.
8. D.E. Knuth, “Semantics of Context-Free Languages,” Mathematical Systems Theo-ry, vol. 2, no. 2, pp. 127–145, 1968.
9. W.J. Hutchins and H.L. Somers, An Introduction to Machine Translation. London: Academic Press, 1992.

ИННОВАЦИОННЫЕ ПОДХОДЫ В СОВРЕМЕННОЙ НАУКЕ

Сборник статей по материалам CCX международной научно-практической конференции

№ 6 (210)
Март 2026 г.

Подписано в печать 20.03.26. Формат бумаги 60x84/16.
Бумага офсет №1. Гарнитура Times. Печать цифровая.
Усл. печ. л. 9. Тираж 550 экз.

Издательство «Интернаука»
123182, г. Москва, ул. Академика Бочвара, д. 5, корпус. 2, к. 115
E-mail: mail@internauka.org
Отпечатано в полном соответствии с качеством предоставленного оригинал-макета в типографии «Allprint»
630004, г. Новосибирск, Вокзальная магистраль, 1

16+

ИНТЕРНАУКА
internauka.org

ООО «Интернаука» (г. Москва) проводит международные заочные научно-практические **конференции по 26 научным направлениям**. Предоставляя возможность опубликовать статьи быстро и качественно, мы помогаем аспирантам, соискателям и докторантам представить на суд научной общественности результаты проведенных исследований, открываем дорогу молодым, привлекаем в научную среду как начинающих ученых, так и профессионалов, имеющих богатый практический опыт в прикладной сфере и упрощаем процесс вхождения в научное сообщество, снижая барьеры расстояния, финансов, языка, статуса, возраста, опыта.

Мы проводим заочные конференции на двух языках: русском и английском, способствуя сближению научных сообществ разных стран.

Нашим изданиям присваиваются коды ISSN, УДК, ББК. Производится их регистрация в Российской книжной палате и рассылка по библиотекам нашей страны.

На сегодняшний день в рамках проекта "Интернаука" было **проведено свыше 250 конференций, в которых приняли участие более 6000 ученых из 15 стран мира**: России, Казахстана, Узбекистана, Азербайджана, Украины, Белоруссии, Польши, Армении, Латвии, Болгарии, Молдовы, Румынии, Эстонии, Греции, Турции.

Конференции по 26 направлениям науки:

Архитектура
Астрономия
Биология
Ветеринария
География
Геология
Информационные технологии
Искусствоведение
История
Культурология
Математика
Медицина
Менеджмент
Педагогика
Политология
Психология
Сельскохозяйственные науки
Социология
Технические науки
Фармацевтические науки
Физика
Филология
Философия
Химия
Экономика
Юриспруденция