Volume 3, Issue 2, February - 2025 **ISSN (E):** 2938-3811

Web of Scientists and Scholars: Journal of Multidisciplinary Research

LINGUISTIC TAGGING OF NATIONAL-CULTURAL UNITS IN THE PARALLEL CORPUS OF THE NOVEL "QIYOMAT"

Aybibi Iskandarova
Associate Professor, PhD in Philology
National University of Uzbekistan
Tel: (99) 877 42 53
E-mail: aybibiiskandarova1962@gmail.com

Abstract
This article provides theoretical information on the internal mechanism of a parallel corpus, the linguistic and cultural adaptation of translation units, the example of a Turkic tagger, the tagging of linguistic and cultural units, phrase models, and structural differences in translation units from unrelated languages, supported by examples. It also explores how these linguistic and cultural units in the novel Qiyomat are formed through metonymy, metaphor, and personification, highlighting the translator's uniqueness-specifically, their stylistic individuality-and how the national color is embedded in the translation process from the original language to the target language.

Keywords: Machine translation, parallel corpus, phrase models, linguistic and cultural adaptation, tagging, Turkic and Uzbek tags, token, lemma, concordance.

Introduction

Corpus linguists suggest that a national corpus should contain at least 100 million words. Today, the database of the Uzbek national corpus is continuously enriched with samples of spoken and written texts from various genres. This is because building a database of parallel texts and studying their linguistic aspects has become one of the key issues for neural or statistical machine translation. If the alternative pairs of texts stored in translation memory are identified and expanded, the quality of automatic translation programs for different styles will improve over time. A parallel corpus serves not only as a linguistic resource for machine translation but also as a valuable source for fields such as bilingual lexicography, comparative translation studies, and contrastive linguistics.

Literature Review and Methodology:

T.O.Dobrovolskiy, in his discussion on parallel text corpora and comparative lexicology, highlights the linguistic uniqueness of certain lexemes, interlingual equivalence phenomena, and their representation in bilingual dictionaries. He asserts that "reliable, empirically validated answers to many questions in lexicology and lexicography can only be obtained when widely available, open parallel corpora exist" [1].

Vera Sibirseva emphasizes that research methods related to parallel subcorpora are increasingly integrated into pedagogy and philology. She notes that tools like semantic text referents, tag

cloud generators, and programs such as LF Aligner play a crucial role in studying different translations of the same work, tracking diachronic and stylistic changes in vocabulary, and understanding translator strategies [2.107].

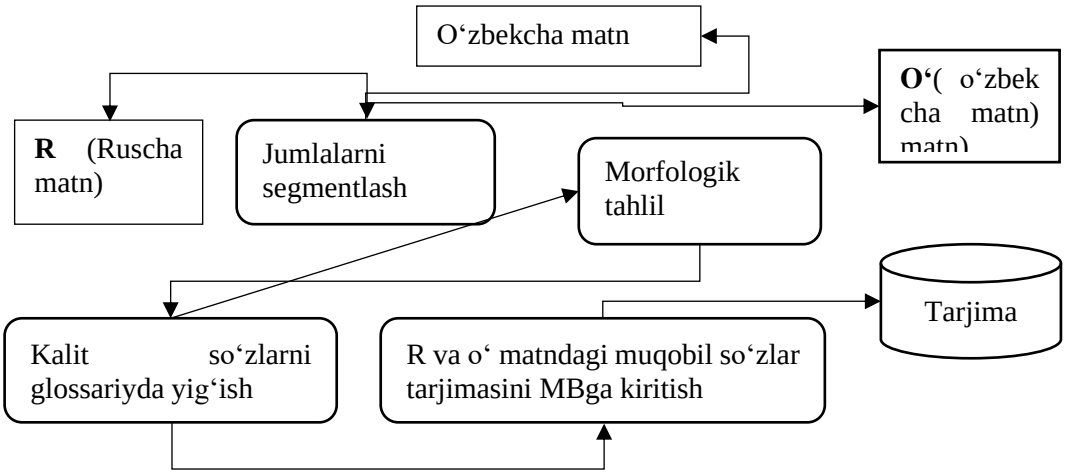
In her monograph Computer Models of the Uzbek Electronic Corpus, Professor N.Abdurakhmonova explains that the morphological and syntactic models of the Uzbek language are tagged using Protégé technology. She states that text annotations are built upon a morphological base: "An electronic corpus consists of morphological features such as word classes' morphotactic states, lemmas, grammatical attributes, and name-expressing units. The morphological database of the Uzbek language (TAG) includes a morphotactic rule system (Morpho Tac), a lexicon (LEXICON), and graphemes (Alphabet). The Uzbek morphological tag set comprises 140 tags representing grammatical categories. For the Uzbek electronic corpus, we use special tags applied to universal Turkic languages" [3.83].

Turkic Morpheme portal provides a tagging system for annotation in Turkic languages, which can be illustrated in the following diagram:

	So'z Turkumlari	**Marfologik teg**
Korpusda nomlanish ID	Fe'l	V
Grammema	Ot	N
Grammema	Olmosh	Pron
Grammema	Son	Num
Grammema	Sifat	Adj
Grammema	Undov so'z	INJ
Grammema	Modal so'z	MOD
Grammema	Taqlid so'z	IMIT
Grammema	Ko'makchi	POST
Grammema	Yuklama	Part
Derivatema	Harakat nomi	VN
Grammema	Buyruq istak + ko'plik	Hor_PL

Y.N.Marchuk's Contribution to Corpus Technology In his work Typology of Texts and Machine Translation, Y.N.Marchuk explains that the AMPAR machine translation system in corpus technology is built on a contextual dictionary framework. He outlines its key functions as follows: "According to this approach, the dictionary content is created based on the concordance of the source language text and its corresponding translation (parallel texts). The dictionary then compares polysemous words or homonyms in the translated text with their linguistic context in the original. Based on this linguistic methodology, phraseological units, syntactic structures, and morphological categories are analyzed during the translation process. Additionally, through analysis of parallel texts, identified concordances are compiled into a database." [4]

The mechanism of the parallel corpus is illustrated in the following appendix:



Tokens lemmas and stemming in corpus linguistics. In corpus linguistics the term token is used instead of word. A lemma refers to a word’s dictionary form For example in the novel the word og‘iroyoqli is a variation of its lemma. Concordance refers to how og‘iroyoqli appears within the surrounding words in the text

The process of adapting a text in a parallel corpus consists of three stages. Tokenization identifying word forms in the text. Lemmatization determining the dictionary form of words. Stemming identifying the root of derivative words.

In Qiyomat the translator rendered the Russian word затяжелела as og‘iroyoq This word is a metaphorically derived unit. In Muslim societies pregnant women are carefully protected to ensure the safe birth and well-being of both mother and child In the novel Toshchaynar’s spouse Akbara is guarded at the entrance of a cave during her pregnancy an act symbolizing a traditional Muslim custom metaphorically personified through the image of a wolf.

Next, we will count the tokens and lemmas from a selected microtext in the novel.

Toshchaynar, who had mostly stayed not in the den but in a quiet spot among the thickets since the she-wolf became pregnant.

In the original text, there are 18 tokens and 13 lemmas.

Since Akbara became heavy-footed, Toshchaynar spent most of his time not in the den but among the dense and peaceful blackcurrant bushes, never straying far from the cave.

In the translation, there are 20 tokens and 20 lemmas.

Changes in the number of tokens and lemmas are natural when translating literary texts from the original language to the target language.

An uncountable herd of steppe antelopes-all of them identical in color since the beginning of time, white-flanked with chestnut backs-grazed unsuspectingly in the wide tamarisk valley, eagerly consuming the feather grass beneath the fresh snow.

In the original text, there are 33 tokens and 27 lemmas.
Since their creation, all of them have looked alike-white-rumped, golden-shouldered herds of countless gazelles roamed the vast tamarisk valley, free from danger, eagerly nibbling on wormwood and feather grass beneath the freshly fallen snow.

In the translation, there are 30 tokens and 30 lemmas.
Names given to individuals, nicknames, and names assigned to animals are also considered linguistic and cultural units of a nation. For example, the word Akbara is one of the linguistic and cultural units specific to Muslim communities. This word is also among the 99 names of Allah. The naming of the she-wolf Akbara in the novel carries an implicit reference to the Turkic myth that claims their lineage descends from wolves. Chingiz Aytmatov's works are distinguished by their multilayered, philosophical depth.
In the parallel corpus, linguistic and cultural units are morphologically tagged as follows:
Акджалы/**Noun**/ - Oqyol/Noun/
Акдалы/**Noun**/ - Oqdil/Noun/
Белохолкой/**Noun**/ - Akbari/Noun/
Акбару – Великую/**Noun**/ - Akbara – Ulug'/Noun/

Волчица прозывалась среди здешних чабанов **Акдалы**, иначе говоря, **Белохолкой**, но вскоре по законам трансформации языка она превратилась в **Акбары**, а потом в **Акбару – Великую**, и между тем никому невдомек было, что в этом был знак провидения.	Шу ерлик чўпонлар бу қанжиқ бўрини **Оқдил** деб юришди. Кейин-кейин бориб **Акбари** дедилар, яна бирмунча вақтлар ўтиб эса **Акбара - Улуғ** деб атадилар.Ким айтса айтгандиру, лекин унга шундай ном берилиши бежиз эмасди.Бунда қазонинг ҳам ҳукми борга ўхшаб кўринарди...

The novel contains many linguistic and cultural units formed through metonymy metaphor and personification the word Xunta is a linguistic unit created through external characteristics meaning it is a metaphor It is used in reference to Ober Kandalov's gang It was terrifying to look at these people drenched in blood from head to toe.
The word xun originates from Persian and means blood. In the Uzbek explanatory dictionary several variations of this word exist xunbor bloodthirsty one who sheds blood xunolud stained with blood xunoba bloody tears xunrezlik a violent clash bloodshed xunxo'r blood-drinker cruel tyrant xunxo'rlik the act of bloodshed.
In the novel metonymy is used as a renaming method resulting in the creation of such polysemous words.

Вторым лицом в этой **хунте** – а хунтой они окрестили свою команду с общего согласия, – единственным, кто слабо возразил, был **Гамлет-Галкин**, бывший артист областного драматического театра: «Ну ее к шутам, хунту, не люблю я, ребята, хунты.	Бу тўп ўзини **хунта** деб атайди. Ҳаммалари шунга келишганлар. Фақат область драма театрининг собиқ артисти **Гамлет-Галкингина** «хунта» деб аташга эътироз билдирган: «Э қўйинглар, шу хунта-пунтасини. Жиним ёқтирмайди шу хунтани, болалар.

The nickname Gamlet-Galkin was also formed through metonymy as he was a former actor at the regional drama theater and thus the gang referred to him as Xunta Gamlet-Galkin. There is a connection between this person in the gang and his nickname. Such polysemous words are formed based on an internal connection between space and time through metonymy meaning shifts to another word and expands. For example **До самого конца зимы** to **qish oyoqlaguncha** and **день клонился к концу kun og‘di** are expressions formed using metonymy The words **oyoqlaguncha** and **og‘di** in these phrases indicate a segment of time.

The linguistic and cultural units in the original and target language are morphologically tagged the same way.

Xunta **Noun** - Xunta **Noun**

Gamlet-Galkin **Noun** – Gamlet Galkin **Noun**

Mishka Shabashnik **Noun** - Mishka Shabashnik **Noun**

Professor Gaybulla Salomov discusses how a translator’s aesthetic belief leaves a mark on the translation and gives the work new life within a national context. He states that although translation involves a comparison of linguistic elements. it is ultimately a complex psychological process of reinterpretation. Unlike original creation literary translation is distinguished by working with an existing object. The translator breathes new social literary and cultural life into the work Artistic translation is not mere copying but an interpretation. The translator’s aesthetic belief inevitably influences the translation 697.

Ibrohim G‘afurov retained certain linguistic units in the novel Qiyomat in their original form such as **xunta chigiri tush** and **chiy** for this very reason.

The phrase Разные зверушки да птицы особенно куропатки was translated as **ilvasin kaklik karkildak kabi qushlar-u jonivorlar** based on national interpretation. In translating from a structurally different language expressions such as крученьми облаками were rendered as **sallador bulutlar душной горячей ночью** as **taffot tun** and **волчица as serka Akbara.** The translator’s aesthetic belief has influenced these words by embedding them with cultural meaning.

Words like **salla cho‘tir ma’raka serka xufton pot-haybat chot** and **kelbat** emerged as national concepts in the translation process. It is important to note that in translation from the original to the target language the semantics and pragmatics of the national language are reflected in these units within phrases and sentences through national concepts.

Web of Scientists and Scholars: Journal of Multidisciplinary Research

1.	Первое совместное лето	**boshlari qovushgan birinchi yoz kunlari;**
2.	прочертится черным земляным шрамом	**qora cho‘tirga aylandi**;
3.	но то будет преславный час	**Zo‘r ma’raka bo‘ladi o‘shanda**;
4.	он выделялся среди мощным загривком и мосластью, тяжеловесностью телосложения	boshqalardan **pot-haybati, kelbatining zo‘ri** bilan ajralib turardi;
5.	и тут вновь ворвались в ее сознание звуки реального мира	shunda yana uning qulog‘i ochildi**, olamning butun qiylu qolini** eshita boshladi;
6.	и в степи стало смеркаться	**cho‘lga xufton kirdi**;
7.	бедовые головы	**sho‘rlik boshlar;**
8.	они всякий раз поджимали хвости	**dumlarini chotlariga qisib**

Ch.Aytmatov in the novel Qiyomat frequently uses complex compound sentences. Ibrohim G‘afurov often translates them as simple sentences. If microtexts are not semantically aligned in Excel the machine will not be able to correctly identify the units it is searching for.

In corpus linguistics the initial process of creating a parallel corpus involves aligning texts in two languages side by side using SmartCat technology. The next stage requires manually adjusting concordances in an Excel format If the concordances are not properly matched the machine translation system will fail to accurately detect equivalent phrases linguistic and cultural units’ historical terms and concepts between the source and target languages.

In Excel the texts are manually aligned in the following manner.

На панические вопли Акбары в нору просунулся ее волк – Ташчайнар, находившийся с тех пор, как волчица **затяжелела**, большей частью не в логове, а в затишке среди зарослей.	Акбара қўрқиб увлаб, ғингшийвергандан кейин ғор ичига унинг жуфти Тошчайнар бошини суқди. Акбара **оғироёқли бўлиб** қолгандан бери Тошчайнар кўп вақтини инида эмас, ўнгирдан узоқ кетмай қалин ва тинч қорағат бутазорлар орасида ўтказарди.

Так, у самого крупного из волчат был широкий, как у Ташчайнара, лоб, и воспринимался он потому как Большеголовый, а средний, тоже крупня- чок, с длиннющими ногами-рычагами, которому быть бы со временем волком-загонщиком, тот воспринимался Быстроногим, а синеглазая, точь-в-точь как сама Акбара, и с белым пятном в паху, как у самой Акбары, игривая любимица Акбары значилась в ее сознании бессловесном Любимицей.	Бўричаларнинг энг каттаси худди Тошчайнарга ўхшаган калладор эди. Шунинг учун уни Хумкалла деб фарқлайди. Ўртанчаси ҳам бўла ва йирик. Унинг оёқлари узун ва кучли эди. Вақти соати келиб у овда ҳайдовчиларга бош бўлади. Шунинг учун уни Илдам деб билади. Учинчи бўрича эса Акбаранинг худди ўзи - кўккўз, човида оқ ямоғи бор, энасининг суюкли ўйинқароғи. Акбара уни ўз назарида Суйгиной деб атаган.

In the parallel corpus, after the text is aligned, the linguistic and cultural units are collected, and corresponding ontological models are created.

At the syntax stage of Uzbek phrase structures, the machine translation has identified the following models. Using their tags, we model the linguistic and cultural units in the literary work:

Материнского лона Noun+Noun – **ona qursog‘i** Noun+Noun
Задвигались в чреве Verb+Noun – **qursoqda harakatga kelmoq** Noun+Verb
Горячей ночью Adj+Noun – **taffot tun** Adj+Noun
Крученными облаками Adj+Noun – **sallador bulutlar** Adj+Noun
После сытной еды Adv+Adj+Noun – **et-moyga to‘yib** Noun+Verb
Прислушиваясь к тому Verb+Noun | PNoun – **batniga quloq tutib** Noun+Verb
Богатый приплод Adj+Noun – **barakali qulunlamoq** Adv+Verb
Вовремя zona Noun+Gerund – **sayg‘oqlarning kuyikkan kezlari** Noun+Gerund+Noun
Изначальный ход вещей Adj+Noun+Noun – **tirikchilikning bu yo‘rig‘i** Noun+Noun | PNoun+Noun

Translation is a process in which two nations and two languages, two perspectives and two material-spiritual worlds, two national arts and literatures, two eras and two authors intertwine and interact. Translation scholars have emphasized that this process follows its own distinct rules.

“A Kazakh’s chanqovuz cannot be translated as a Ukrainian sivizga, nor can a Russian truba be rendered as an Uzbek surnay. In translation, everything is usually called by its own name.” [6.105]

In “Qiyomat”, there are many linguistic and cultural units belonging to other nations, and they are represented by their original names.

Исключением мог считаться разве что самый молодой из них со странным, ветхозаветным именем Авдий – упоминался такой в **Библии в Третьей Книге Царств, – сын дьякона** откуда-то из-под Пскова, поступивший после смерти отца в **духовную семинарию как подающий надежды отпрыск церковного служителя** и через два года изгнанный оттуда **за ересь**	Улар ўртасида ёшгина бир йигит ажралиброқ тураради. Унинг исми ҳам ғалати - Авдий эди. Бу ном қадим китобатда бор: **Инжилнинг учинчи Мулк битигида** айтилган. Ўзи асли **дьяконнинг ўғли** бўлиб, Псков томонлардан эди. Отасининг вафотидан сўнг, **черков ходимининг куртаги деб**, таҳсил учун **диний семинарияга** қабул қилинган, аммо икки йилдан сўнг **куфронага йўл қўйиб** у ердан ҳайдалган эди.

В Библии в Третьей Книге Царств **[Noun+Num+Noun+Noun]** –
Injilning uchinchi Mulk bitigi **[Non+Num+Noun+Noun];**
Сын дьякона [**Noun+Noun] -** dyakonning o‘g‘li[**Noun+Noun]**;
Духовный семинар **[Adj+Noun] -** diniy seminariya **[Adj+Noun]**;
Изгнанный оттуда за ересь **[Verb+ADV+Noun] -** kufronaga yo‘l qo‘yib u yerdan haydalmoq **[Noun+Noun +Verb+Noun|PNoun+Noun+Verb];**
Отцами богословами **[Noun+Adj] -** ruhoniy otalar **[Adj+Noun]**;

Церковной карьеры **[Noun+Noun]** - cherkov mansab shotisi pillapoyalari **[Noun+Noun+ Noun+Noun];**
Преданный церковью анафеме [**Verb+Noun+Noun] –** cherkov tomonidan murtad deb e'lon qilinmoq **[Noun+Noun+Noun+Post+ Noun+Verb];**
Пасхальные концепции **[Adj+Noun]** – diniy e'tiqod **[Adj+Noun];**
Епархия **[Noun]** - eparxiya [**Noun];**
поп **[Noun]** – pop **[Noun].**

Conclusion:
Chingiz Aytmatov and Ibrohim G'afurov are creators who grew up in the Turkic lands of Kyrgyzstan and Uzbekistan. The way of life, customs, and values of these two nations are similar, stemming from a common ancestral root. The writer and translator share a mutual worldview, having been raised in and shaped by comparable traditions, values, and beliefs. This connection has enabled the translator to quickly and deeply grasp the writer's artistic vision, allowing for a nationalized interpretation of the literary text. Building **parallel corpora** for the Uzbek national corpus is crucial in today's world, where intercultural communication is widespread. A parallel corpus based on original and translated texts allows for the identification of methods used by translators and facilitates various translation studies: "Through parallel corpora, it becomes possible to identify linguistic universals across different languages and cultures, as well as the unique mental characteristics, realia, and lacunar units of languages. The parallel text corpus also contributes to the advancement of machine translation and supports the development of computational lexicography." [7.1447] The linguistic and cultural units found in Qiyomat primarily serve as a linguistic resource for machine translation. Additionally, in the future, they will play a significant role in the development of Russian-Uzbek bilingual lexicography, comparative translation studies, and contrastive linguistics.

REFERENCES:

1. Т.О. Добровольский. Корпус параллелных текстов и сопоставительная лексикология. https://www.researchgate.net/publication.
2. Вера Сибирцева. Технология использования параллельного подкорпуса Национального корпуса русского языка и коллекций текстов в обучении иностранным языкам. Rocznik Institutu Polsko – Rosyjskiego Nr 2 (5), 2013. 98-110
3. Abduraxmonova N. O'zbek tili elektron korpusining kompyuter modellari (monografiya). – Toshkent, 2021. 202 b.
4. https://www.nop-dipo.ru/node Марчук Ю. Н. Типология текстов и машинный перевод.
5. Ўзбек тилининг изоҳли луғати. – Тошкент, 2008. 5 жилд. – 608 б.
6. Ғ. Саломов. Таржима назариясига кириш. – Тошкент, 1978. – 218 б.
7. Sobirova N. Korpus lingvistikasi va parallel korpuslar tavsifi. Academic Research in Educational Sciences. Volume.3. Issue 5. 2022 - P 1447.

8. Abduraxmonova N., Iskandarova A., Xolmuradova I.Tarjima texnologiyasini rivojlantirishda parallel korpuslarning o‘rni. “Yangi O‘zbekistonda o‘zbek adabiy tilining rivojlanish tendensiyalari: muammolar, yechimlar va tavsiyalar” mavzusidagi Respublika ilmiy-amaliy anjuman materiallari. 2024-yil. 338-343 b.
9. Айтматов Ч. Плаха. Алма – Ата, 1978. - 302 б.
10. Айтматов Ч. Қиёмат. – Тошкент, 1989. – 334 б.
11. Abdurakhmonova, N., & Urdishev, K. (2019). Corpus based teaching Uzbek as a foreign language. Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL), 6(1-2019), 131-7.
12. Abdurakhmonova, N., Tuliyev, U., & Gatiatullin, A. (2021, November). Linguistic functionality of Uzbek Electron Corpus: uzbekcorpus. uz. In 2021 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-4). IEEE.
13. Iskandarova, A. (2019). THE MAIN CRITERIA OF LITERATURE IS HUMANISM. Inernational Journal Anglistticum Literature, Linguistics, Interdis.
14. Iskandarova, A. (2020). "MIRZO ULUG‘BEK" DRAMASIDA MIRZO ULUG‘BEK - ILM-MA’RIFAT HOMIYSI SIFATIDA. Maqsud Shayxzoda Adabiy Merosi Va Zamonaviylik Mavzusidagi Xalqaro Ilmiy Amaliy Konfrensiya to`plami.