GALAXY INTERNATIONAL INTERDISCIPLINARY RESEARCH JOURNAL (GIIRJ)
ISSN (E): 2347-6915
Vol. 14, Issue 6 June (2026)

METHODOLOGICAL PRINCIPLES FOR AN AUTHOR PARALLEL CORPUS OF PUSHKIN WORKS

Jumaeva Zarnigor Zokirovna
PhD Researcher, Department of Russian Language and Literature
Bukhara State University

ABSTRACT

The paper sets out the methodological foundations of an author parallel corpus of A. S. Pushkin's works, with Russian as the source pole and Uzbek as the target pole. The source base comprises the academic Pushkin edition, Oybek's 1937/1956 rendering of Eugene Onegin and A. Qahhor's 1939 rendering of The Captain's Daughter. Markup follows the TEI P5 standard, sentence alignment relies on the W. A. Gale and K. W. Church algorithm in its HunAlign implementation, and morphological analysis of the Uzbek texts draws on UzMorphAnalyser. The three translation universals identified by M. Baker (explicitation, simplification, normalisation) are tested on the aligned material. The Russian National Corpus carries no Russian–Uzbek pair in its parallel section, which makes the project a green-field undertaking.

Keywords: Author parallel corpus, A. S. Pushkin, translation studies, sentence alignment, TEI P5, Uzbek language, translation universals, corpus linguistics, Eugene Onegin, machine translation.

INTRODUCTION

An author corpus differs from a balanced national corpus in one decisive respect. It fixes the writer and varies almost nothing else. For Pushkin this principle has a long pedigree, since the four-volume *Dictionary of the Language of Pushkin*, edited by V. V. Vinogradov and printed in Moscow between 1956 and 1961, already lemmatised more than twenty thousand word-forms drawn from his verse, prose, letters and chancery papers [2]. That dictionary was, in effect, a concordanced author corpus assembled by hand. What it could not do was place the Russian original beside its translations and let a researcher read the two in step. A parallel author corpus adds exactly that second pole, and the present paper specifies how such a resource can be built for the pair Russian and Uzbek, a pair that carries unusual historical weight because Pushkin entered Uzbek letters during the 1937 centenary through Oybek's verse translation of *Eugene Onegin*.

Why this pair, and why now. The parallel section of the Russian National Corpus, documented by V. A. Plungian and developed for translation research by D. V. Sichinava, operates thirty-two bilingual pairs and exceeds two hundred million tokens, yet Uzbek appears in none of them [2]. English contributes roughly fifty-two million tokens to that section, German thirty-two, Chuvash twenty-four, Swedish sixteen. Russian and Uzbek contribute nothing. A philologist who wants to compare how Tatyana's letter survives in Uzbek, or how Pushkin's free indirect speech is redistributed by his translator, has at present no aligned resource to query and must work page by page through paper editions. Closing that gap calls for a method rather than a single dataset, because the texts span verse and prose, two writing systems, and

an agglutinative target language whose morphology resists the tools designed for Russian. The sections below describe that method and the evidence that justifies each choice.

METHODS AND REVIEW OF THE LITERATURE

The study combines four procedures that operate in sequence. Source texts are first encoded in TEI P5 XML, with a *teiHeader* carrying bibliographic and language metadata and with verse marked by *lg* and *l* elements so that the Onegin stanza keeps its shape. Sentence and stanza correspondences are then established by the length-based dynamic-programming aligner of W. A. Gale and K. W. Church, run through HunAlign and LF Aligner and checked by hand wherever the confidence score falls. Uzbek word-forms receive morphological annotation from UzMorphAnalyser, while the Russian side inherits tagging compatible with the Russian National Corpus, after which lemma-to-lemma links make concordance queries possible across the two languages. The aligned material finally supports a quantitative test of M. Baker's three translation universals, measured against a reference corpus of non-translated Uzbek prose of the same decades and read together with the sociocultural account of Soviet translation supervision given by B. Quénu.

Four bodies of scholarship converge on this task. The lexicographic tradition begins with the Pushkin dictionary of V. V. Vinogradov [2] and continues in the corpus programme that V. A. Plungian outlined when he asked, in 2005, why a national corpus is worth building and answered that it must reach far beyond literary fiction [7]; the parallel architecture itself was specified by D. V. Sichinava [9]. Alignment rests on the algorithm of W. A. Gale and K. W. Church, whose 1993 paper remains the reference point for length-based methods [15, 16], and encoding rests on the TEI P5 guidelines maintained by the TEI Consortium [22]. Corpus-based translation research was opened by M. Baker, who proposed explicitation, simplification and normalisation as testable regularities [13, 14], was given its first synthesis by S. Laviosa [17], and connects to the Russian school through the lexical-equivalence theory of V. N. Komissarov [5]. The Uzbek dimension draws on B. Quénu's study of Russian-to-Uzbek transfer under Stalinist rule [20], on the UzMorphAnalyser model of U. Salaev [21], and on the annotated morphological dataset published by N. Abdurakhmonova and colleagues, which supplies the reference material an Uzbek side of the corpus requires [11].

RESULTS

Design comes before data. The corpus is a translation-driven parallel corpus of literary fiction, sampled by whole works rather than by balanced excerpts, with one fixed source author and a target pole that can hold several translations of the same text. F. Zanettin's manual on translation-driven corpora treats author, publisher and date as the decisive sampling criteria for fiction, and that guidance maps cleanly onto Pushkin, whose Uzbek reception is datable to specific editions and named translators. Oybek's *Yevgeniy Onegin* carries the date 1937 for its first journal appearance and 1956 for its revised book edition; A. Qahhor's *Kapitan qizi* carries 1939. Each work therefore enters the corpus with a translator, an edition and a year attached, which lets later analysis separate a translator's habits from the period style around him. The full sequence of construction stages is shown in Figure 1.

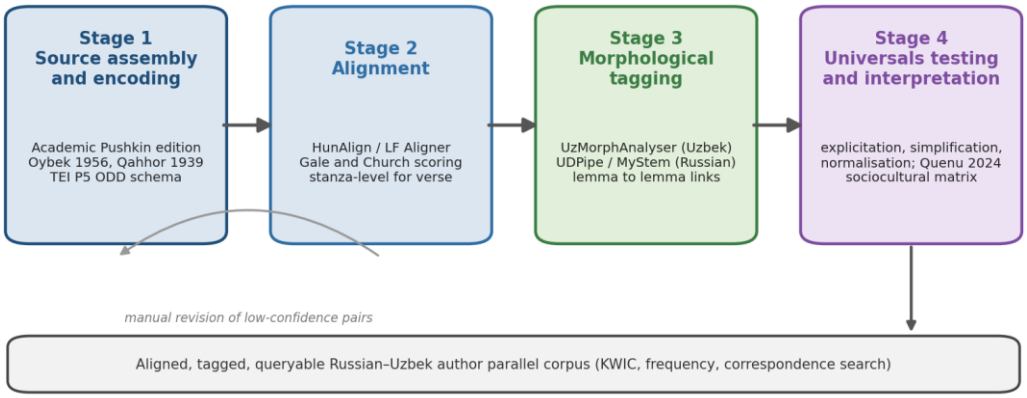

Figure 1. Construction pipeline of the Russian–Uzbek author parallel corpus, from source encoding to universals testing.

Encoding decisions shape everything downstream. The corpus adopts TEI P5, the markup standard that the digital-humanities community has revised on a fixed schedule for nearly two decades, which matters for a resource meant to outlive any single project. The TEI Consortium states the maintenance regime plainly.

"It was initially released in November 2007 and has been updated since then on a six-month cycle, with point releases incorporating maintenance fixes and minor feature enhancements."

TEI Consortium, TEI P5 Guidelines, 2025 (CC BY 3.0)

Working inside that standard, the Pushkin corpus encodes verse with stanza and line elements and records inter-language correspondence through stand-off link groups rather than inline tags, so that the Russian source and the Uzbek target remain separate documents joined by pointers. Such separation pays off when a single Russian stanza answers to a reshaped Uzbek stanza, because the link can span unequal units without distorting either text. The header of every file names its language with an ISO code, which is what lets a query engine keep *ru* and *uz* material apart.

Alignment is the technical core of any parallel corpus, and for Russian and Uzbek it is also the hardest step. The method chosen exploits a simple statistical regularity between sentence lengths, stated by W. A. Gale and K. W. Church in the formulation that has anchored the field since the early 1990s.

"The program uses the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences."

W. A. Gale, K. W. Church, ACL 1991 (open access, ACL Anthology P91-1023)

Their journal version explains how that regularity becomes a working procedure, scoring each candidate correspondence and selecting the best global path through the text.

"This probabilistic score is used in a dynamic programming framework to find the maximum likelihood alignment of sentences. It is remarkable that such a simple approach works as well as it does."

W. A. Gale, K. W. Church, Computational Linguistics 19(1), 1993, p. 75 (open access)

For Pushkin's prose this length-based aligner, augmented with lexical anchors in HunAlign, handles the bulk of *Kapitan qizi* with little manual repair. For the verse it does not, because Oybek frequently keeps the fourteen-line stanza while moving an image from one line to another, which breaks any line-by-line correspondence. The corpus therefore aligns verse at the stanza level and reserves line-level links for the cases where they survive. Table 1 records the units, tools and expected manual-correction load at each step.

Stage	Unit of processing	Tool	Manual correction
Encoding	work, stanza, line, sentence	TEI P5 ODD schema	schema validation only
Alignment, prose	sentence	HunAlign, LF Aligner	about 3 to 5 per cent
Alignment, verse	stanza, sometimes line	Gale and Church score, then hand	about every fifth stanza
Tagging, Uzbek	word-form	UzMorphAnalyser	spot check of affixation
Tagging, Russian	word-form	UDPipe or MyStem	homonymy resolution

Table 1. Processing units, tools and expected manual-correction load by stage.

Tagging the Uzbek side raises problems the Russian side does not. Uzbek is agglutinative, so a single orthographic word can carry a stem and a chain of suffixes that a Russian-oriented tagger will misread. UzMorphAnalyser was built for precisely this, and its author reports the accuracy on a controlled test set.

"The proposed model was evaluated using a curated test set comprising 5.3K words. Through manual verification of stemming, lemmatizing, and morphological feature corrections carried out by linguistic specialists, it obtained a word-level accuracy of over 91%."

U. Salaev, UzMorphAnalyser, AIP Conf. Proc. 3244, 2024 (open access)

Accuracy near ninety-one per cent is workable for a research corpus, though it still leaves roughly one word in eleven for human review, which is why Table 1 keeps a spot check at this stage. A second resource supplies the gold material against which such taggers are trained and tested. N. Abdurakhmonova and colleagues describe it.

"The dataset contains 3022 manually annotated word forms, each annotated with root, affix, and part-of-speech information. Two morphological analysis approaches were implemented and compared, a user-defined rule-based stemming algorithm and a conditional random fields based machine learning model."

N. Abdurakhmonova et al., Data in Brief 61, 2025 (CC BY)

These two resources together let the Uzbek pole of a Pushkin corpus reach the same query depth as the Russian pole, so that a search for a Russian lemma returns its aligned Uzbek correspondences with their morphological analysis attached rather than as raw strings.

One empirical fact frames the whole project and deserves to be seen rather than asserted. The parallel section of the Russian National Corpus, for all its breadth, contains no Russian and Uzbek pair, while it does contain pairs with far smaller speech communities. Figure 2 sets the operational pairs side by side and marks the absence.

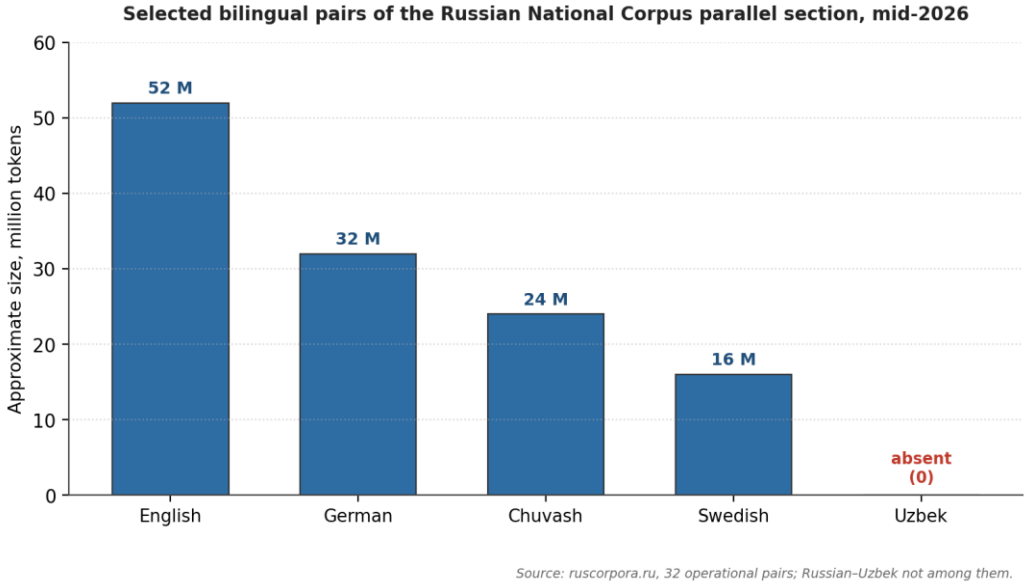

Figure 2. Selected bilingual pairs of the Russian National Corpus parallel section by approximate size, with the Russian–Uzbek pair absent.

The contrast is instructive. Chuvash, with roughly a million speakers, holds about twenty-four million tokens in the parallel section, whereas Uzbek, with more than thirty million speakers across Central Asia, holds none. The reason is institutional rather than linguistic, and it points to the kind of focused, author-anchored resource this paper proposes as a realistic first step toward filling the gap.

The payoff of all this preparation is the ability to test claims about translation itself. M. Baker's programme treats certain features as recurrent across translated texts regardless of language pair, and an aligned Pushkin corpus turns those features into measurable quantities. Table 2 shows how each universal becomes an indicator and what it is measured against.

Universal	Quantitative indicator in the corpus	Comparison base
Explicitation	sentence-length ratio, frequency of optional connectives, pronoun expansion	Russian source against Uzbek target, segment by segment
Simplification	type-token ratio, lexical density of the Uzbek translation	non-translated Uzbek prose of 1937 to 1960
Normalisation	over-use of canonical Uzbek collocations and set phrases	reference corpus of original Uzbek literary prose

Table 2. Operationalisation of M. Baker's three translation universals for the Russian–Uzbek pair.

Any pattern the corpus reveals must then be read against history, not only against statistics. Soviet-era translation from Russian into Uzbek was supervised, edited and shaped by institutions, which means a measured rise in explicitation may reflect editorial policy as much as a translator's instinct. B. Quénu sets out the conditions under which this work was done.

"This chapter focuses on Stalin-era literary translations from Russian to Uzbek in the Soviet Republic of Uzbekistan. Highlighting the different steps for the increasing supervision of the translators' activity within the Soviet Writers' Union of Uzbekistan, it sheds light on the material conditions of the professionalization of the translation industry, including career benefits, risks and opportunities, gender inequality, and strategies of institutional control."

B. Quénu, in Translating Russian Literature in the Global Context, 2024 (CC BY-NC-ND)

Reading the corpus through that account guards against a tempting error, the assumption that every quantitative regularity in Oybek's Onegin springs from his individual choice. Some of it does. Some of it is the residue of an editorial apparatus that a Pushkin-into-French corpus would never register, and a method that ignores this would mistake politics for poetics.

DISCUSSION

Three commitments separate the design proposed here from a generic parallel-corpus build. The source pole is anchored to a historical authority file, the Pushkin dictionary of V. V. Vinogradov, so that orthographic variants and lemma identity are resolved against a recognised reference rather than ad hoc. The alignment unit is allowed to differ between prose and verse, which respects the fact that Oybek's stanza is a unit of meaning even when its lines do not match Pushkin's one for one. The interpretive layer is explicitly sociocultural, following B. Quénu, because the Uzbek translations were produced under conditions that leave measurable traces. Each commitment answers a documented difficulty rather than a hypothetical one, and each can be checked by another researcher against the cited sources.

Limits remain, and naming them is part of the method. The accuracy of UzMorphAnalyser, near ninety-one per cent on its own test set, sets a ceiling on automatic tagging quality that only manual review can raise, and review is expensive at corpus scale. The translation-universals framework of M. Baker has itself drawn criticism since 2010 for treating language-independent tendencies as established before they are confirmed, so the Pushkin study should fix its hypotheses and effect-size thresholds in advance. Verse alignment will demand hand work on roughly every fifth stanza, a figure consistent with the error rates W. A. Gale and K. W. Church reported for harder, many-to-many cases. None of these limits blocks the project, yet each one shapes how its results may honestly be stated.

CONCLUSION

A Russian and Uzbek author parallel corpus of Pushkin is feasible with existing tools and is, at the same time, genuinely new, since no such pair exists in the Russian National Corpus. Its method joins TEI P5 encoding, length-based sentence alignment after W. A. Gale and K. W. Church, morphological tagging through UzMorphAnalyser, and a test of M. Baker's universals read against the historical record assembled by B. Quénu. Built this way, the resource would let researchers ask, for the first time on aligned evidence, how Pushkin's verse and prose were remade in Uzbek and how much of that remaking belongs to the translator. The next task is the schema and a first aligned work, most plausibly *Kapitan qizi*, whose prose will test the pipeline before the harder verse of *Onegin* is attempted.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Бархударов Л. С. Язык и перевод (Вопросы общей и частной теории перевода). – М.: Международные отношения, 1975. – 240 с.
2. Виноградов В. В. (отв. ред.). Словарь языка Пушкина : в 4 т. / Ин-т рус. яз. АН СССР. – М.: Гос. изд-во иностр. и нац. словарей, 1956-1961. – 3232 с.
3. Виноградов В. С. Введение в переводоведение (общие и лексические вопросы). – М.: Изд-во ИОСО РАО, 2001. - 224 с.
4. Гарбовский Н. К. Теория перевода: учебник. – М.: Изд-во МГУ, 2004. – 544 с.
5. Комиссаров В. Н. Теория перевода (лингвистические аспекты): учеб. для ин-тов и фак. иностр. яз. – М.: Высшая школа, 1990. – 253 с.
6. Национальный корпус русского языка: 2006-2008. Новые результаты и перспективы / отв. ред. В. А. Плунгян. – СПб.: Нестор-История, 2009. – 502 с.
7. Плунгян В. А. Зачем мы делаем Национальный корпус русского языка? // Отечественные записки. – 2005. – № 2.
8. Рецкер Я. И. Теория перевода и переводческая практика: Очерки лингвистической теории перевода. – М.: Международные отношения, 1974. – 216 с.
9. Сичинава Д. В. Параллельные тексты в составе Национального корпуса русского языка: новые направления развития и результаты // Труды Института русского языка им. В. В. Виноградова. – 2015. – Вып. 6. – С. 194-235.
10. Фёдоров А. В. Основы общей теории перевода (лингвистические проблемы). – 4-е изд., перераб. и доп. – М.: Высшая школа, 1983. – 303 с.
11. Abdurakhmonova N., Shirinova R., Sayfullayeva R., Mengliev D., Ibragimov B., Ernazarova M. An annotated morphological dataset for Uzbek word forms: towards rule-based and machine learning approaches // Data in Brief. – 2025. – Vol. 61. – Art. 111702.
12. Allaberdiev B., Matlatipov G., Kuriyozov E., Rakhmonov Z. Parallel texts dataset for Uzbek-Kazakh machine translation // Data in Brief. – 2024. – Vol. 53. – Art. 110194.
13. Baker M. Corpus Linguistics and Translation Studies: Implications and Applications // Text and Technology: In Honour of John Sinclair / eds. M. Baker, G. Francis, E. Tognini-Bonelli. – Amsterdam; Philadelphia: John Benjamins, 1993. – P. 233-250.
14. Baker M. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research // Target. – 1995. – Vol. 7, no. 2. – P. 223-243.
15. Gale W. A., Church K. W. A Program for Aligning Sentences in Bilingual Corpora // 29th Annual Meeting of the Association for Computational Linguistics. – Berkeley, CA: ACL, 1991. – P. 177-184.
16. Gale W. A., Church K. W. A Program for Aligning Sentences in Bilingual Corpora // Computational Linguistics. – 1993. – Vol. 19, no. 1. – P. 75-102.
17. Laviosa S. Corpus-based Translation Studies: Theory, Findings, Applications. – Amsterdam; New York: Rodopi, 2002. – 138 p.
18. Olohan M. Introducing Corpora in Translation Studies. – London; New York: Routledge, 2004. – 232 p.
19. Pushkin A. S. Yevgeniy Onegin / tarjimon Oybek. – Toshkent : O'zdavnashr, 1956.

20. Quénu B. From Russian to Uzbek (1928-53): Unequal Cultural Transfers and Institutional Supervision under Stalinist Rule // Translating Russian Literature in the Global Context. – Cambridge: Open Book Publishers, 2024. – Ch. 34. – P. 525-554.
21. Salaev U. UzMorphAnalyser: A Morphological Analysis Model for the Uzbek Language Using Inflectional Endings // AIP Conference Proceedings. – 2024. — Vol. 3244. – Art. 030058.
22. TEI Consortium (eds.). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.10.2. – Text Encoding Initiative Consortium, 2025.
23. Zanettin F. Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies. – Manchester: St Jerome, 2012. – xiii + 244 p.