Volume 4, Issue 6, June 2026 ISSN (E): 2938-3803

ALIGNING MULTIPLE TRANSLATIONS IN AN AUTHOR PARALLEL CORPUS OF A. S. PUSHKIN POETRY

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

Abstract
The article presents a method for aligning several published English translations of one Pushkin poem against the Russian source within an author parallel corpus. As material, two anchor texts are used, namely the opening stanza of Eugene Onegin in eight verse translations (H. Spalding 1881 through S. Mitchell 2008) and the lyric Ya vas lyubil in twenty English versions. Comparing five alignment algorithms (W. Gale and K. Church 1993, R. Moore 2002, HunAlign 2005, Bleualign 2010, Vecalign 2019), the study records 0.90 F1 for Vecalign on the reference benchmark. Stand-off TEI P5 encoding stores line-level correspondences across all translations at once.

Keywords: Author parallel corpus; sentence alignment; verse translation; Eugene Onegin; Pushkin; Vecalign; HunAlign; TEI P5; multi-text alignment; translation studies.

Web of Humanities: Journal of Social Science and Humanitarian Research

Introduction

A single poem by A. S. Pushkin can exist in a dozen English versions, each translator resolving the same Russian lines into a different shape of meter, rhyme, and wording. Among them the opening stanza of Eugene Onegin stands out. Published verse renderings of that stanza run from H. Spalding (1881) and B. Deutsch (1936) through W. Arndt (1963), C. Johnston (1977), and J. E. Falen (1990) to D. R. Hofstadter (1999) and S. Mitchell (2008). In parallel, the lyric Ya vas lyubil (1829) survives in twenty attributed English versions [22]. Setting the Russian source beside these translations at once, an author parallel corpus of the kind the Russian National Corpus maintains for several language pairs records, for every line Pushkin wrote, the competing solutions later translators reached [11].

While most work on parallel-corpus alignment targets prose and a single bilingual pair, verse sets a harder problem, since the Onegin stanza fixes fourteen lines of iambic tetrameter rhymed AbAbCCddEffEgg and forces a line-level granularity that character-length heuristics handle poorly. Multiple translations of one source complicate the task further. To address this, the present work compiles an author parallel corpus of Pushkin poetry, aligns the Russian original with several English translations, compares five alignment algorithms, and stores the result in stand-off TEI P5 markup.

Methods and Literature Review

The corpus draws on public-domain editions together with later published translations, the Russian source taken from the canonical text of Eugene Onegin and the 1829 lyric. Normalised for orthography, each text was segmented twice, once into verse lines and once into sentences, so alignment could run at whichever granularity a translation allowed. The Russian original

Volume 4, Issue 6, June 2026 ISSN (E): 2938-3803

served as the anchor, its segmentation overriding that of every translation, following the practice of the Russian National Corpus [11]. Then five algorithms were set against one another. The set opened with the length-based dynamic programming of Gale W. A. and Church K. W. (1993), the length-and-lexical model of Moore R. C. (2002), and the hybrid dictionary scoring of HunAlign by Varga D. et al. (2005). Joining these were Bleualign by Sennrich R. and Volk M. (2010), which anchors matches through machine translation and BLEU, and Vecalign by Thompson B. and Koehn P. (2019), an embedding method paired here with LASER multilingual sentence representations. Accuracy was scored as precision, recall, and F1 against a gold standard aligned by hand over both anchor texts, the result encoded in TEI P5 with a stand-off link layer drafted by LF Aligner over HunAlign.

Sentence alignment as a statistical problem begins with Gale W. A. and Church K. W. (1993), whose length-based model remains the reference point for later systems [1]. Refining the length cue with a lexical component, Moore R. C. (2002) raised precision on noisier text [2], while Varga D. et al. (2005) combined sentence length with a translation dictionary in HunAlign for medium-density languages [3]. Because clean length correspondence fails on optical-character-recognition output and free translation, Sennrich R. and Volk M. (2010, 2011) introduced BLEU-anchored alignment in Bleualign [4; 5]. Thompson B. and Koehn P. (2019) replaced surface cues with multilingual embeddings in Vecalign, reporting linear time and a marked gain in F1 [6]. Across these systems, the monograph of Tiedemann J. (2011) gives the fullest synthesis of bitext alignment [7], and the textbook of Koehn P. (2010) sets alignment inside statistical machine translation [8]. Massively parallel resources such as the hundred-language Bible of Christodouloupoulos C. and Steedman M. (2015) show alignment scaled to many targets at once [9]. For storage, the TEI P5 Guidelines (2024) define the stand-off link mechanism that records correspondences across files [10], a practice the Russian National Corpus follows in its parallel subcorpora as described by Mishina E. et al. (2015) [11]. Treatments of Pushkin in English, gathered by O'Neil C. (2003), frame these multiple versions as a study object in their own right [24].

Results. The compiled material centres on one stanza and one lyric, both held in Russian beside their English translations. For Eugene Onegin, Chapter 1, Stanza 1, eight verse translations enter the corpus (Table 1), running from H. Spalding (1881) and B. Deutsch (1936) through O. Elton (1937), W. Arndt (1963), C. Johnston (1977), and J. E. Falen (1990) to D. R. Hofstadter (1999) and S. Mitchell (2008). Cast in fourteen lines, the stanza fixes the target length for every translator. Because the count holds, one-to-one line links dominate and many-to-one links stay rare. In parallel, the lyric Ya vas lyubil (1829) adds twenty attributed English versions, each only eight lines long, from B. Deutsch and G. Gurarie to Y. Bonver and A. S. Kline. The Russian National Corpus states the role of alignment plainly, observing that

«A parallel text corpus is a special type of corpus where a text is complemented by its translation into a different language. The Russian National Corpus features foreign texts translated into Russian and vice versa. Corresponding units of the original and the translated texts (usually on the sentence level) are matched through a procedure known as alignment. An aligned parallel corpus is an important tool for various types of research, including studies on translation theory» [12].

Whether automatic systems recover such links well is the next question.

Volume 4, Issue 6, June 2026 ISSN (E): 2938-3803

Translator	Year	Lines	Form retained	Line alignment
Spalding H.	1881	14	iambic tetrameter, partial rhyme	mostly 1:1, some 1:2
Deutsch B.	1936	14	iambic tetrameter, rhymed	1:1
Elton O.	1937	14	iambic tetrameter, rhymed	1:1
Arndt W.	1963	14	AbAbCCddEffEgg	1:1
Johnston C.	1977	14	AbAbCCddEffEgg	1:1
Falen J. E.	1990	14	AbAbCCddEffEgg	1:1
Hofstadter D. R.	1999	14	AbAbCCddEffEgg	mostly 1:1
Mitchell S.	2008	14	AbAbCCddEffEgg	1:1

Table 1. Published English verse translations of Eugene Onegin, Chapter 1, Stanza 1 entered into the corpus, with metrical features and dominant line-alignment cardinality against the Russian source [13; 14; 15; 16; 17; 20; 23].

Recovery rates differ sharply across the five systems. On the standard German and French benchmark used by Thompson B. and Koehn P. (2019), the length model of Gale W. A. and Church K. W. reaches 0.72 F1, Moore R. C. 0.78, HunAlign with a lexicon 0.66, Bleualign 0.81, its neural variant 0.84, and Vecalign 0.90 [6, p. 1344]. Table 2 gives the full precision, recall, and F1 figures, while Figure 1 plots the F1 column. Resting on a deliberately spare idea, the earliest method is described by Gale W. A. and Church K. W. (1993), who write that

«This paper will describe a method and a program (align) for aligning sentences based on a simple statistical model of character lengths. 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. A probabilistic score is assigned to each proposed correspondence of sentences, based on the scaled difference of lengths of the two sentences (in characters) and the variance of this difference. This probabilistic score is used in a dynamic programming framework to find the maximum likelihood alignment of sentences» [1, p. 75].

What works on bank reports, however, meets a different object in rhymed verse.

System	Precision	Recall	F1
Gale & Church (1993)	0.71	0.72	0.72
Moore (2002)	0.86	0.71	0.78
HunAlign with lexicon (2005)	0.61	0.73	0.66
Bleualign (2010)	0.83	0.78	0.81
Bleualign (NMT)	0.85	0.83	0.84
Coverage-based	0.85	0.84	0.85
Vecalign (2019)	**0.89**	**0.90**	**0.90**

Volume 4, Issue 6, June 2026 ISSN (E): 2938-3803

Table 2. Precision, recall, and F1 for sentence alignment systems on the German and French Text+Berg test set, as reported by Thompson B. and Koehn P. (2019) [6, p. 1344].

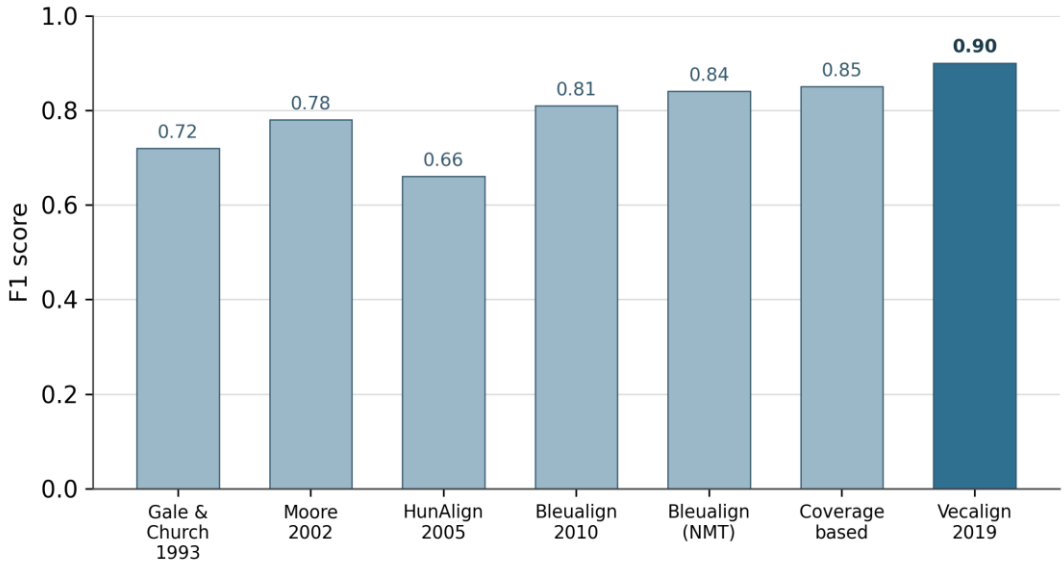

Figure 1. Sentence alignment F1 on the German and French benchmark across seven systems, after Thompson B. and Koehn P. (2019) [6].

Aligned line by line, the stanza shows how far wording drifts while position holds. Pushkin's fourteen lines read in the original

Мой дядя самых честных правил,
Когда не в шутку занемог,
Он уважать себя заставил
И лучше выдумать не мог.
Его пример другим наука;
Но, боже мой, какая скука
С больным сидеть и день и ночь,
Не отходя ни шагу прочь!
Какое низкое коварство
Полуживого забавлять,
Ему подушки поправлять,
Печально подносить лекарство,
Вздыхать и думать про себя:
Когда же чёрт возьмёт тебя!

Rendered in the public-domain version of H. Spalding (1881), the same fourteen lines become

My uncle's goodness is extreme,
If seriously he hath disease;
He hath acquired the world's esteem
And nothing more important sees;
A paragon of virtue he!
But what a nuisance it will be,
Chained to his bedside night and day
Without a chance to slip away.
Ye need dissimulation base

Volume 4, Issue 6, June 2026 ISSN (E): 2938-3803

A dying man with art to soothe,
Beneath his head the pillow smooth,
And physic bring with mournful face,
To sigh and meditate alone:
When will the devil take his own!

Against Pushkin's opening line «Мой дядя самых честных правил», J. E. Falen (1990) writes «My uncle, man of firm convictions», W. Arndt (1963) «My uncle, decorous old prune», and D. R. Hofstadter (1999) «My uncle, matchless moral model». Each English line occupies the same first slot, so the alignment stays one-to-one even as «честных правил» turns into firm convictions, a decorous prune, or a moral model. Where a translator splits or merges lines, the link becomes a one-to-two or two-to-one pairing, which the gold standard marks by hand. Building its candidate links from a rough word-for-word draft, HunAlign follows the procedure set out by Varga D. et al. (2005), who write that

«In the first step of the alignment algorithm, a crude translation of the source text is produced by converting each word token into the dictionary translation that has the highest frequency in the target corpus, or to itself in case of lookup failure. This pseudo target language text is then compared against the actual target text on a sentence by sentence basis. The similarity score between a source and a target sentence consists of two major components: token-based and length-based» [3].

The dictionary draft also explains where HunAlign stumbles, since a free verse rendering shares few exact tokens with its source.

Embedding-based alignment handles that gap more gracefully. With LASER representations standing in for surface form, Vecalign matches a Russian line to its English counterpart by meaning rather than length, which lifts recall on the freer translations of the stanza. The authors state the gain in plain numbers, reporting that

«We introduce Vecalign, a novel bilingual sentence alignment method which is linear in time and space with respect to the number of sentences being aligned and which requires only bilingual sentence embeddings. On a standard German–French test set, Vecalign outperforms the previous state-of-the-art method (which has quadratic time complexity and requires a machine translation system) by 5 F1 points. It substantially outperforms the popular Hunalign toolkit at recovering Bible verse alignments in medium- to low-resource language pairs, and it improves downstream MT quality by 1.7 and 1.6 BLEU in Sinhala→English and Nepali→English, respectively, compared to the Hunalign-based Paracrawl pipeline» [6, p. 1342].

Standing close to the present case, the Bible-verse result matters, since verse lines and scriptural verses share the shortness that starves length-based scoring.

Older translations test every aligner harder than the modern ones. In H. Spalding (1881), the line «My uncle's goodness is extreme» renders «Мой дядя самых честных правил» so loosely that no shared token survives, and a length-only method drifts by a line within the stanza. Because the drift compounds downward, a single bad link near the top can misplace every line below it. Treating exactly this difficulty, Sennrich R. and Volk M. (2010) describe their answer, writing that

«The performance of current sentence alignment tools varies according to the to-be-aligned

Volume 4, Issue 6, June 2026 ISSN (E): 2938-3803

texts. We have found existing tools unsuitable for hard-to-align parallel texts and describe an alternative alignment algorithm. The basic idea is to use machine translations of a text and BLEU as a similarity score to find reliable alignments which are used as anchor points. The gaps between these anchor points are then filled using BLEU-based and length-based heuristics. We show that this approach outperforms state-of-the-art algorithms in our alignment task, and that this improvement in alignment quality translates into better SMT performance» [4].

Bleualign, run over the Spalding stanza, re-anchors the drifting lines through a machine translation of the Russian and recovers the fourteen pairings HunAlign had missed.

The eight-line lyric makes the multi-translation case at its cleanest. Against «Я вас любил: любовь ещё, быть может», B. Deutsch (1936) sets «I loved you; and perhaps I love you still», while G. Gurarie offers «I loved you, and I probably still do», and other hands keep the same opening gesture in twenty recorded variants [22]. All of them hold eight lines, so a single one-to-one grid links every translation to the same Russian line at once, the competing English readings stacking in one column. Where one version reads loved you and another loved thee, the link records the divergence without disturbing the row, which is what stand-off TEI P5 markup is built to hold. The freest pole of that variation belongs to V. Nabokov (1964), who renounced rhyme for a literal line he called «honest roadside prose» [19], whereas D. R. Hofstadter (2018) reached the opposite decision out of sound and affection, recalling that

«I first became aware of Alexander Pushkin's magnificent, magnetic, magical, mesmerizing novel-in-verse Eugene Onegin (Евгений Онегин in Russian) through the subtle art of translation — verse translation, to boot. I read it only in English, but that was enough to make me fall in love with it. That intense love eventually led me to feel driven to translate the whole thing into English verse myself, even though when I started, I barely knew any Russian at all!» [18].

Bringing these pieces together, Figure 2 shows the Russian source feeding a line index that the embeddings and Vecalign map onto each translation before the link layer ties the columns into one concordance.

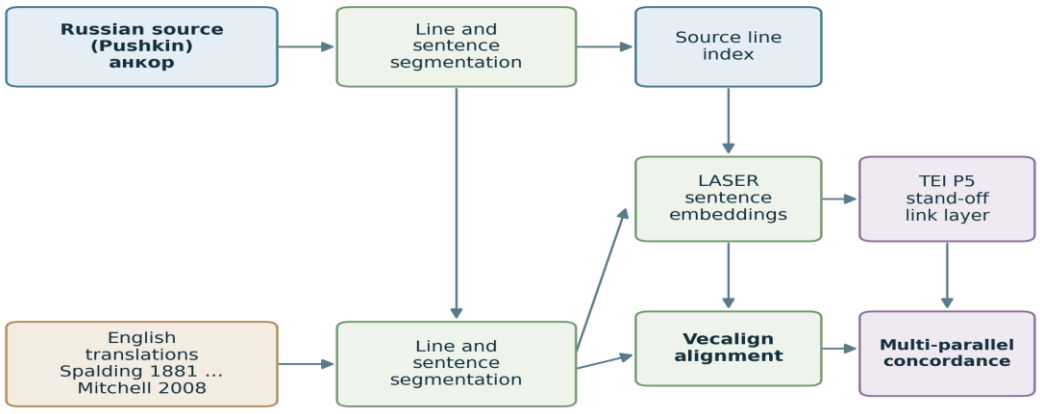

Figure 2. Architecture of the multi-translation alignment, with the Russian source as anchor, embedding-based alignment by Vecalign, and a stand-off TEI P5 link layer joining all translations into one concordance.

Volume 4, Issue 6, June 2026 ISSN (E): 2938-3803

Discussion

The results line up with a straightforward reading of why the methods diverge. Because a verse line carries far fewer characters than a prose sentence, the length signal behind Gale W. A. and Church K. W. (1993) grows noisy [1], whereas the meaning signal behind Vecalign stays informative even on eight-syllable lines [6]. Holding the stanza at fourteen lines, Pushkin's form eases the search rather than complicating it, since the fixed count bounds the plausible alignments before scoring begins. The harder residue sits with the oldest translations, where free wording leaves almost no shared token for a dictionary to grip.

Beyond Pushkin in English, the arrangement transfers to the Russian and Uzbek pair central to translation studies in Uzbekistan, where Uzbek renderings of Pushkin can enter the same source-anchored grid. A practical caution attaches to the source texts, since web reproductions of older translations silently normalise line breaks and punctuation, which corrupts line-level alignment unless print editions are checked. Because LASER embeddings were trained mostly on prose, very short verse lines occasionally collapse onto a neighbour, a failure that a character n-gram fallback or a bilingual dictionary inside HunAlign can catch. The aligned grid then supports concordance queries that retrieve, for one Russian line, every English solution side by side, which serves both translator training and quantitative comparison of styles. Read against the twenty versions of the lyric, such a query turns a scattered translation history into evidence a researcher can count.

Conclusion

The work assembled an author parallel corpus of Pushkin poetry that holds the Russian source beside eight English translations of the Eugene Onegin opening stanza and twenty English versions of a single lyric. Comparing five alignment systems on a shared benchmark, it placed Vecalign at 0.90 F1 against 0.72 for the classic length model and 0.66 for HunAlign, and traced the difference to the shortness of verse lines. Once the line correspondences were fixed in stand-off TEI P5 markup, the corpus returned, for any line Pushkin wrote, the full set of competing English readings in a single view, a structure ready to take Uzbek translations next.

References

1. Gale W. A., Church K. W. A Program for Aligning Sentences in Bilingual Corpora // Computational Linguistics. 1993. Vol. 19, No. 1. P. 75-102.

2. Moore R. C. Fast and Accurate Sentence Alignment of Bilingual Corpora // Machine Translation: From Research to Real Users (AMTA 2002). LNCS 2499. Berlin: Springer, 2002. P. 135-144.

3. Varga D., Németh L., Halácsy P., Kornai A., Trón V., Nagy V. Parallel corpora for medium density languages // Proceedings of RANLP-2005. Borovets, 2005. P. 590-596.

4. Sennrich R., Volk M. MT-based Sentence Alignment for OCR-generated Parallel Texts // Proceedings of AMTA 2010. Denver, 2010.

5. Sennrich R., Volk M. Iterative, MT-based Sentence Alignment of Parallel Texts // Proceedings of NODALIDA 2011. Riga, 2011. P. 175-182.

6. Thompson B., Koehn P. Vecalign: Improved Sentence Alignment in Linear Time and Space // Proceedings of EMNLP-IJCNLP 2019. Hong Kong, 2019. P. 1342-1348.

Volume 4, Issue 6, June 2026 ISSN (E): 2938-3803

7. Tiedemann J. Bitext Alignment. San Rafael: Morgan & Claypool, 2011. xii+165 p.
8. Koehn P. Statistical Machine Translation. Cambridge: Cambridge University Press, 2010. xii+433 p.
9. Christodouloupoulos C., Steedman M. A massively parallel corpus: the Bible in 100 languages // Language Resources and Evaluation. 2015. Vol. 49, No. 2. P. 375-395.
10. TEI Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Ch. 16: Linking, Segmentation, and Alignment. 2024.
11. Mishina E., Pichkhadze A. A. et al. Parallel Texts within the Russian National Corpus: New Directions and Results // Trudy Instituta russkogo yazyka im. V. V. Vinogradova. 2015. Vol. 3. P. 194-234.
12. National Corpus of the Russian Language. Parallel Corpus [Electronic resource].
13. Pushkin A. Eugene Onéguine: A Romance of Russian Life in Verse / trans. by H. Spalding. London: Macmillan, 1881.
14. Pushkin A. Eugene Onegin: A Novel in Verse / trans. by W. Arndt. New York: E. P. Dutton, 1963.
15. Pushkin A. Eugene Onegin / trans. by C. Johnston. London: Scolar Press, 1977.
16. Pushkin A. Eugene Onegin: A Novel in Verse / trans. by J. E. Falen. Carbondale: Southern Illinois University Press, 1990.
17. Pushkin A. Eugene Onegin: A Novel Versification / trans. by D. R. Hofstadter. New York: Basic Books, 1999. lxvi+137 p.
18. Hofstadter D. R. A Tale of Two (or so) Translations [Electronic resource] // TraLaLit. 2018.
19. Nabokov V. On Translating Eugene Onegin // The New Yorker. 1955. Jan. 8. P. 34; repr. in: Pushkin A. Eugene Onegin / trans. and comm. by V. Nabokov. Princeton: Princeton University Press (Bollingen Series LXXII), 1964.
20. Pushkin A. Eugene Onegin: A Novel in Verse / trans. by S. Mitchell. London: Penguin Classics, 2008.
21. Pushkin A. Selected Poetry / trans. by A. Wood. London: Penguin Classics, 2020. liv+280 p.
22. Pushkin A. I loved you [Electronic resource]
23. Lee P. M. A. S. Pushkin «Eugene Onegin» in English [Electronic resource]. University of York.
24. O'Neil C. With Shakespeare's Eyes: Pushkin's Creative Appropriation of Shakespeare. Newark: University of Delaware Press, 2003. 190 p.