Embedding-Assisted Construction of a Japanese-Chinese Parallel Corpus of Contemporary Novels

LIAN Zeqi
Graduate School of Intercultural Studies, Kobe University
1-2-1 Tsurukabuto, Nada-ku, Kobe, Hyogo 657-8501, Japan
lzq980628@outlook.com

Abstract

This paper presents JCFPC, the Japanese-Chinese Fiction Parallel Corpus, a research-oriented corpus linking five contemporary Japanese fiction volumes to their published Chinese translations. JCFPC contains 35,189 Japanese sentences, 32,967 Chinese sentences, 561,597 Japanese morphological tokens, and 23,871 alignment units. Japanese text is indexed with MeCab and UniDic, while sentence groups are aligned by a two-level monotonic dynamic-programming procedure in which multilingual embeddings supply semantic evidence for variable groups up to 8:1 and 1:8, including explicit omissions. Quality control compares independently refined BGE-M3 and Harrier pipelines by their Chinese index sets for each Japanese sentence, then directs manual review to informative disagreements. The final pipelines have the same or overlapping candidates for 98.85% of Japanese sentences, passed regression and structural checks, and achieved 94.0% strict and 97.5% boundary-inclusive acceptability in a stratified single-author audit. A pilot analysis of 1,958 valid mimetic occurrences shows that 70.12% are reorganized through general Chinese vocabulary, constructions, or clause structure. JCFPC demonstrates an auditable workflow for small literary parallel corpora; public access currently consists of a fixed-list mimetic interface rather than redistribution of the copyrighted texts.

Keywords: parallel corpus; sentence alignment; Japanese-Chinese; literary translation; sentence embeddings; corpus construction

1. Introduction

Parallel corpora make translation choices observable at scale, but their usefulness depends on how readers can enter the data and what the returned unit represents. For Japanese-Chinese contrastive and translation research, an especially useful path begins with a Japanese surface form or lemma and returns the corresponding Chinese translation context. Contemporary fiction is valuable here because it combines narration, dialogue, expressive vocabulary, and stylistic variation within published translations. Yet a searchable resource needs more than digitized text: it requires a morphological layer on the Japanese side, reliable links between source and target passages, and enough provenance to explain how uncertain links were reviewed. This paper reports the construction and evaluation of JCFPC, the Japanese-Chinese Fiction Parallel Corpus, which is designed around those requirements.

Literary alignment is difficult because a typographic sentence is not necessarily a translation unit. A translator may combine several short Japanese sentences, divide one sentence into several Chinese sentences, move information within a local passage, or omit material. A fixed 1:1 assumption therefore produces systematic boundary errors. Conversely, unrestricted semantic nearest-neighbour retrieval can ignore narrative order and attach a sentence to a thematically similar but incorrect passage. Manual alignment can resolve such cases, but reviewing every possible link is costly and difficult to reproduce. The central design problem is thus not to eliminate human judgment, but to create a constrained automatic proposal process and direct judgment to cases where it is most informative.

JCFPC makes three contributions. First, it provides a documented contemporary-fiction dataset with Japanese lemma/surface retrieval and native many-to-many alignment units. Second, it combines wide-group monotonic alignment with independent embedding pipelines, anomaly-driven repair, and coverage-closed manual patches. Third, it evaluates the resulting corpus through regression cases, cross-pipeline agreement, structural checks, and a stratified single-author audit, while demonstrating research use through an exploratory analysis of Japanese mimetic expressions. The paper asks how variable translation units can be represented without sacrificing order, how independent models can reduce review effort without being treated as truth, and whether the resulting corpus supports a substantive translation analysis rather than only a concordance demonstration.

2. Background

2.1 Japanese-Chinese Parallel Resources

Japanese-Chinese parallel resources serve several distinct domains. Among the resources reviewed by Miyamoto (2023), the Beijing Center's Chinese-Japanese Parallel Corpus established an important literary foundation, while others include news-oriented material developed for machine translation, the scientific-paper ASPEC-JC corpus, and interpreting resources such as JNPC. These resources are complements rather than deficient predecessors: each reflects a different purpose, access model, and unit of analysis. Retrieval-oriented prototypes in the same review also show that KWIC and document access are not merely interface choices; they shape what copyrighted or controlled corpus data can be used for.

For the present study, the relevant gap is narrower than a claim that Japanese-Chinese resources are generally unavailable. What remains scarce is the combination of contemporary published fiction, Japanese morphological retrieval, variable-sized sentence groups, and a revision history that connects observed anomalies to algorithmic changes. JCFPC addresses that combination in a small, purposively selected corpus. Its contribution is consequently methodological and documentary as well as quantitative: it records how a resource can be built and audited under the constraints of literary translation and restricted source texts.

2.2 Sentence Alignment for Literary Translation

Classical sentence alignment exploited length and order, with Gale and Church (1993) providing a foundational probabilistic formulation. More recent methods use multilingual representations while retaining monotonic search. Vecalign, for example, obtains scalable monotonic alignment from sentence embeddings (Thompson and Koehn, 2019). Bertalign applies embedding-based alignment to Chinese-English literary texts and is particularly close to the domain problem considered here (Liu and Zhu, 2023). These approaches show that semantic vectors can improve candidate selection without discarding document order.

Literary Japanese-Chinese data nevertheless impose four practical requirements that are not simultaneously supplied as defaults by a single off-the-shelf aligner. First, punctuation boundaries may be wrong, as when a detached closing quotation mark becomes a false sentence. Second, genuine compression and splitting require groups considerably wider than 2:1 or 1:2. Third, omissions must remain representable as one-sided units rather than being forced into a neighbouring translation. Fourth, an accurate literary paraphrase can receive a low embedding similarity, whereas a long group can receive a high score because only its first sentence matches. JCFPC therefore uses embeddings as a replaceable evidence layer inside a project-controlled two-level dynamic-programming procedure. The framework adds wide moves, explicit gaps, symmetric pipeline comparison, local repair, and provenance-bearing manual write-back. Scores rank evidence; they do not define correctness.

3. Corpus Data and Design

The source frame was purposive and transparent. Screening covered 24 works that received the 155th through 174th Naoki Prizes in 2016-2025. Eight had a located published Chinese translation. Three historical novels were excluded because deliberately archaizing vocabulary and style could confound a corpus intended to foreground contemporary Japanese usage; all five remaining translated works were retained. The sample is therefore constrained by prize recognition, translation availability, and a register decision, not statistically representative of contemporary fiction. Table 1 gives the editions and final corpus counts.

Japanese and Chinese texts were segmented with language-specific punctuation rules while preserving work, line, and sentence order. The Japanese side was then analysed with MeCab (Kudo et al., 2004) and UniDic (Den et al., 2008). Line-aware token records retain surface form, lemma, part-of-speech information, sentence index, and source position. This separation permits tokenization or retrieval logic to be revised without silently changing the underlying sentence layer.

The SQLite design distinguishes sentences, Japanese tokens, alignment units, and unit items. An alignment unit can therefore contain any supported combination of Japanese and Chinese sentence indices, including a one-sided omission. Retrieval follows Japanese token → Japanese sentence → alignment unit → Chinese sentence group. Both lemma and surface paths are retained because they recover partly different candidates. For example, the lemma query かっと retrieves the orthographic surface カッと and a 1:2 aligned Chinese group containing 给激怒了. Raw texts are treated as read-only inputs, while repair candidates and accepted overrides remain separately versioned so that the final database can be rebuilt from recorded decisions.

Work and editions	Final corpus counts
book01. Ogiwara Hiroshi, *Umi no mieru rihatsuten* (Shueisha, 2016-03-25); Chinese *Haibian lifadian*, trans. Cao Yibing (Nanhai, 2019-03)	JA 4,645; ZH 4,348; tokens 69,820; units 3,027
book02. Onda Riku, *Mitsubachi to enrai* (Gentosha, 2016-09-23); Chinese *Mifeng yu yuanlei*, trans. An Su (China Friendship, 2018-11-01)	JA 12,941; ZH 12,533; tokens 214,034; units 9,090
book03. Hase Seishu, *Shonen to inu* (Bungeishunju, 2020-05-15); Chinese *Shaonian yu quan*, trans. Wen Xueliang (Beijing United, 2022-04)	JA 7,540; ZH 6,172; tokens 95,349; units 4,852
book04. Kubo Misumi, *Yoru ni hoshi o hanatsu* (Bungeishunju, 2022-05-24); Chinese *Zai yewan fangfei xingxing*, trans. Dong Shuhan (CITIC, 2023-09-20)	JA 4,251; ZH 3,972; tokens 72,988; units 2,777
book05. Sato Shogo, *Tsuki no michikake* (Iwanami Shoten, 2017-04-06); Chinese *Yueyuan yueque*, trans. Lu Wanxia (People's Literature, 2020-05)	JA 5,812; ZH 5,942; tokens 109,406; units 4,125
Total	**JA 35,189; ZH 32,967; Japanese morphological tokens 561,597; alignment units 23,871**

Table 1: Corpus composition and editions

Note: JA/ZH are sentence counts. Japanese names and titles are romanized in the compact table; the corpus retains original-script metadata.

4. Variable-Group Monotonic Alignment

4.1 Two-Level Search

Alignment proceeds at two levels. A chunk-level dynamic-programming pass first establishes a coarse monotonic path and limits the search region for detailed alignment. Mutual nearest-neighbour anchors are accepted only when they satisfy minimum text length, semantic similarity, relative-position, span-size, and character-ratio constraints. These filters prevent a short generic sentence from becoming a hard anchor that blocks a correct wider passage. Within each resulting region, a sentence-group dynamic-programming pass selects a monotonic sequence of moves. Narrative order is therefore a hard constraint, while the number of sentences consumed on either side may vary.

For a candidate move containing Japanese group (J) and Chinese group (C), the implemented cost combines length evidence, a move penalty, and an embedding adjustment. In simplified form,

cost(J,C) = |log(observed_length_ratio / expected_ratio)| + move_penalty + semantic_adjustment,

where the semantic adjustment decreases cost when cosine similarity exceeds a baseline and is floored so that it cannot produce an unbounded reward. The final local build uses BGE-M3 representations (Chen et al., 2024), with separate weights for chunk and sentence-group evidence. Empty-side moves receive explicit penalties. The framework thus remains interpretable as a monotonic alignment model rather than becoming unconstrained vector retrieval.

4.2 Wide Groups, Segmentation, and Conservative Repair

The move inventory was expanded through observed cases rather than selected only in advance. It supports sentence groups through 8:1 and 1:8, plus gaps. Example (1) shows five short Japanese sentences from book02 that form one coherent inner monologue; narrow moves split this group and pushed fragments into neighbouring units, whereas the final corpus retains the 5:1 unit.

(1) JA: あたしはこうする。こうしたい。こんなこともできる──ああ、こんなふうにやっていいんだ。呼吸するように当たり前に、音楽を作っていってもいいんだ。
ZH: 我会这么弹，我想这么弹，还可以这样——啊，这样太好了，可以像呼吸一样，极其自然地做音乐。
'This is how I would do it. This is how I want to do it. I can even do this — ah, so it is fine to do it like this: to make music as naturally as breathing.' (book02, 5:1)

Another passage approached 1:8 when one Japanese sentence was distributed across several Chinese sentences. Group embeddings are computed from the whole span and re-normalized, preventing the first sentence from standing in for the group. The database records the complete span rather than a nominal head sentence.

Preprocessing and repair are part of the same quality problem. Detached Japanese closing quotation marks were reattached through delayed sentence finalization because a false 1:0 sentence can shift every subsequent candidate. After the initial dynamic-programming pass, local gap repair considers adjacent mergers only when vector support and score margins are sufficiently strong, with bounded group sizes. It can improve boundary choices but cannot force every gap to disappear. A low-scoring unit is placed in a review queue; it is not labelled erroneous, since literary paraphrase may have weak surface similarity, as in (2), where the published translation reorders cause and consequence yet remains fully correct.

(2) JA: ツケが溜まって断られると、自分の女の店に立ち寄って、こうなった責任を取れと無茶を言って荒れた。
ZH: 赊账太多店里不让喝了，他就跑到自己女人的店里大吵大闹要人家对他落到这种地步来负责。
'Once his tab had piled up and he was refused, he would turn up at his woman's bar and rage that she should take responsibility for what he had become.' (book05, 1:1, low confidence)

Figure 1 summarizes how this principle connects corpus construction, independent quality assurance, and retrievable output.

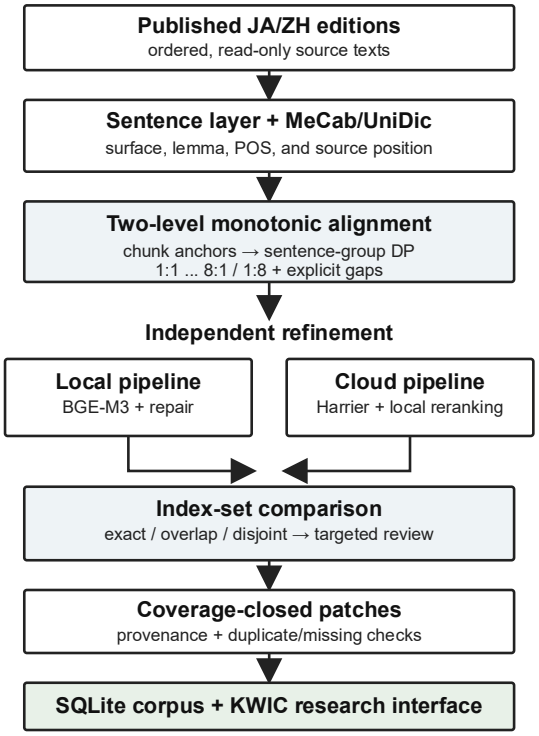

Figure 1: End-to-end JCFPC construction and quality-control workflow. Embeddings provide evidence inside monotonic variable-group alignment; independent refined pipelines create an informative disagreement queue, and accepted patches must preserve sentence coverage before the SQLite corpus and retrieval interface are rebuilt.

5. Independent Pipelines, Anomaly-Driven Repair, and Audited Correction

5.1 Symmetric Pipeline Comparison

A single embedding model can reproduce its own blind spots when it is used both to build and to check a corpus. JCFPC therefore maintains two comparable refined pipelines. The Local pipeline uses BGE-M3, the two-level dynamic-programming procedure, grouping, and local gap repair. The Cloud pipeline applies the same structural framework with Microsoft Harrier embeddings (Microsoft, n.d.); Qwen3 reranking (Zhang et al., 2025) is used only for selected local candidates and is not a third complete alignment pipeline. The cloud refinement used one H100 SXM instance (80 GB VRAM) for approximately three hours, with a dashboard charge of USD 9.025.

Comparison is performed after both pipelines are refined. Alignment-unit identifiers cannot be compared directly because grouping changes their boundaries and counts. Instead, for each of the shared 35,189 Japanese sentence indices, the procedure retrieves the Chinese index set assigned by each pipeline. Relations are classified as exact, overlapping, disjoint bilingual, or missing. This symmetric index-set representation permits the same question to be asked of both outputs and avoids treating a raw candidate list as equivalent to a completed corpus.

The initial comparison produced approximately 903 raw priority candidates. Requiring bilingual units, complete target separation, a Harrier-vector advantage of at least 0.20, and de-duplication by local unit reduced the first-round queue to 27 cases, followed by a targeted residual

queue. Human review did not simply accept the higher-scoring side. In one region, the Cloud pipeline found the right neighbourhood but over-merged a Chinese sentence that expressed two propositions. Accepting only the seed candidate would have duplicated or dropped neighbouring sentences. The correct review object was therefore the complete affected patch group.

5.2 Representative Challenges and Patch Closure

The construction casebook contains 137 numbered anomaly headings across preprocessing, gap repair, model comparison, residual review, and final overrides. Table 2 selects six distinct failure classes rather than a frequency sample. Together they show why successful alignment cannot be represented by one polished example and one omission alone.

Challenge and observable failure	Design response and final treatment
Detached closing quotation mark: Japanese 」 became an independent 1:0 sentence and shifted later links.	Delay sentence finalization and reattach closing marks; rebuild the affected region.
Legitimate 5:1 compression: narrow moves and short hard anchors split a coherent translation.	Add wide moves, whole-span embeddings, and stricter anchor filters; retain the final 5:1 unit; see (1).
Low score, correct paraphrase: a literary causal reformulation was semantically complete despite weak similarity.	Treat low confidence as a review signal, not an error label; retain the verified unit; see (2).
High score, incomplete coverage: a raw Harrier 6:1 candidate was favoured because its first Japanese sentence matched strongly.	Inspect complete group coverage; reject the raw candidate and compare completed refined pipelines; see (3).
Local cascade and unsafe seed replacement: Local drifted, Cloud over-merged, and a seed-only patch broke neighbours.	Expand to the whole local patch group and require duplicate/missing-sentence closure before write-back.
Genuine omission: book02: J12393-J12394 → omission (2:0, manually verified).	Preserve a one-sided manual_omission unit instead of forcing attachment to an adjacent Chinese sentence; see (4).

Table 2: Representative alignment challenges and design responses

Note: Cases are purposively selected to illustrate distinct failure classes, not a frequency sample.

The high-score case is especially instructive. During Cloud-pipeline quality assurance, a raw Harrier 6:1 candidate placed six Japanese sentences against one Chinese sentence and received a Qwen reranker score of 0.97265625. As (3) shows, the score was driven by the first Japanese sentence, which was a strong match; translations of the remaining five occurred only in later units.

(3) JA: レンタルビデオ店はビルの地下一階にある。
[+ five further Japanese sentences in the same 6:1 candidate]
ZH: 录像带出租店在一座写字楼的地下一层。
'The rental video shop is in the basement of the building.' (book05, rejected raw candidate)

The candidate was not a final-corpus error, because it was rejected before final write-back. It instead demonstrates that relevance between concatenated spans does not establish internal coverage. Whole-span representation helps proposal generation, but only group-aware review can determine whether every included sentence belongs there.

Accepted corrections are stored as manual overrides with provenance and expanded to the complete local set of affected units. Before rebuilding, a closure check verifies that the patch consumes exactly its declared Japanese and Chinese sentence sets. Whole-corpus checks then require zero duplicated sentence items, zero missing sentence items, and successful SQLite integrity. This discipline also allows a real omission to remain: structural completeness means that every source sentence is represented, not that every sentence must have target text. Example (4) shows the omission behind this rule: two Japanese sentences in a musical description have no counterpart in the published translation and remain as the corpus's single 2:0 unit.

(4) JA context: 静かにさざなみのような弦楽器のトリルが入る。密やかな導入部。そして、風間塵の弾くメロディが加わる。その最初の一音だけで、あまりのクリアさに観客の耳が覚醒するのが分かった。
'A ripple-like trill from the strings enters quietly. A hushed introduction. Then the melody played by Jin Kazama joins in. From the very first note, I could tell that its extraordinary clarity awakened the audience's ears.'
ZH context: 安静如同微波一般的弦乐器的颤音加入进来。[Ø] 开头的一个音，清澈得让观众的耳朵都清醒了过来。
'A quiet, ripple-like trill from a stringed instrument joins in. [Ø] The opening note was so clear that it awakened the audience's ears.'
Underlining marks the Japanese passage omitted between the two adjacent Chinese sentences (book02, 2:0, manual omission).

6. Results and Exploratory Application

6.1 Corpus and Alignment Validation

The final corpus contains 23,871 alignment units: 3,027, 9,090, 4,852, 2,777, and 4,125 across the five works. Of these, 23,870 are bilingual and one is the manually verified 2:0 omission. Local, Cloud, and patched databases are retained as separate stages, while accepted decisions are replayable from versioned patch files.

Validation separates four kinds of evidence. First, 52 preselected regression cases spanning the five works are aligned or continuously covered. These are development cases, not a random estimate of corpus accuracy. Second, the completed pipelines assign exactly the same Chinese index set to 77.27% of Japanese sentences and the same or an overlapping set to 98.85%; 401 sentences (1.14%) have disjoint bilingual candidates. Agreement is evidence about pipeline stability, not semantic truth, and the disjoint rate is not an error rate. Third, the final database has zero duplicated sentence items, zero missing sentence items, and an OK SQLite integrity result. These are structural rather than semantic checks.

Fourth, a stratified sample of 200 bilingual alignment units was judged manually by one author, with the only one-sided unit inspected separately. The audit found 188 correct, seven boundary-defensible, and five wrong units. Strict correctness was 94.00% (Wilson 95% CI 89.81%-96.53%); including boundary-defensible units, acceptability was

97.50% (94.28%-98.93%). The one-sided unit was justified. Because the audit has one judge and no second annotation, these are author-judged sample estimates rather than gold-standard accuracy or inter-annotator agreement. The five confirmed errors remain in a versioned post-audit maintenance queue so that the frozen evaluation is not retrospectively changed.

The review process changes the structure of labour rather than eliminating it. Approximately 903 raw priority candidates were filtered to 27 first-round cases and a smaller targeted residual queue. Candidate rows, sibling expansions, regression cases, and overrides have different denominators and are not summed as a count of independent human decisions. Table 3 keeps validation, review workload, and pilot results in separate panels for this reason.

Panel and measure	Result / interpretation
A. Regression coverage	52/52 selected cases aligned or continuously covered; not full-corpus accuracy.
A. Cross-pipeline exact	27,191/35,189 Japanese sentences = 77.27%.
A. Same or overlapping	98.85%; disjoint bilingual 401/35,189 = 1.14%.
A. Structural checks	Duplicate items 0; missing items 0; SQLite integrity OK.
A. Stratified single-author audit	188/200 strict correct = 94.00% (89.81%-96.53%); 195/200 correct or boundary-defensible = 97.50% (94.28%-98.93%); one-sided unit separately justified.
B. Targeted review	About 903 raw priority candidates → 27 first-round cases → targeted residual review; stages are not a precision/recall funnel.
B. Stored residual patches	38 accepted rows (28 seeds + 10 siblings); unresolved rows 0.
C. Pilot retrieval and cleaning	2,041 union hits → remove 80 false positives and merge 6 repeated tokens into 3 occurrences → N = 1,958; 141/200 target types attested.
C. Incidence	0.404 per 100 non-symbol Japanese word tokens (denominator 484,968).
C. Main translation outcomes	Meaning retained/partially shifted 89.84%; formally salient Chinese rendering 19.71%; lexical/constructional/clause-level reorganization 70.12%.
C. Seven realization strategies	Sound-symbolic 57 (2.91%); reduplicative 329 (16.80%); general lexical 1,304 (66.60%); constructional 64 (3.27%); clause fusion 5 (0.26%); omission 145 (7.41%); alteration 54 (2.76%).

Table 3: Validation, review workload, and pilot application

Note: Panel C uses 484,968 non-symbol Japanese word tokens, whereas Table 1 reports 561,597 total morphological tokens. The seven strategy percentages sum to 100.01% because of independent two-decimal rounding.

6.2 Pilot Analysis of Japanese Mimetic Expressions

The pilot used a fixed inventory of 200 Japanese mimetic items: the most frequent candidates from a pattern-matching extraction of onomatopoeic forms over BCCWJ and CEJC (LIAN, 2024). Searching each item through both lemma and surface channels produced 2,041 union hits: 1,965 lemma hits, 1,784 surface hits, and 1,708 retrieved by both. Every hit was inspected in bilingual context. Eighty false positives were removed, including 27 conjugation fragments of -garu matching がっ and 18 sequences in which a personal name plus diminutive ちゃん and particle と was mis-segmented as the adverb ちゃんと. Six repeated tokens of the quoted sound す、す、す were merged into three sentence-level occurrences. The final analytical dataset therefore contains 1,958 valid occurrences from 141 of the 200 target types; channel counts are retrieval diagnostics, not alternative analytical denominators.

Each occurrence was coded by the author on separate axes for semantic outcome, Chinese realization, and syntactic function. Every valid occurrence was inspected individually in its bilingual context, and borderline cases were revisited in a second pass, with all coding decisions kept in versioned files. Because the coding was carried out by a single author, no inter-annotator agreement could be calculated, and the results are exploratory.

Most source meanings were recoverable: 89.84% were retained or partially shifted. Formal resemblance was much less common. Only 19.71% received a salient Chinese sound-symbolic or reduplicative rendering, while 70.12% were reorganized through general vocabulary, constructions, or clause-level structure. The examples clarify the distinction. Japanese そっと撫でた 'stroked it gently' becomes Chinese 轻轻抚摸 'gently stroke', where the conventional AA form makes manner visible. In がっかりだ '(I am) disappointed' → 大失所望 'bitterly disappointed', the mimetic contribution is recast as an idiomatic predicate. The same item can also behave differently by context: どんどん大きくなる 'keeps growing rapidly' → 越来越大 'bigger and bigger' preserves progressive change constructionally, whereas どんどん消えていく 'rapidly fades away' → 缓缓沉到... 'slowly sinks into...' changes the direction and event construal. Formal salience was most strongly associated with semantic domain (Cramer's V = .335), but this remains a pooled-sample association rather than a claim about translator intention.

As a scale check under the same fixed inventory and each corpus's reported search/tokenization conditions, incidence per 100 words was 0.404 in JCFPC, 0.164 in the BCCWJ book subset, and 0.282 in the Corpus of Everyday Japanese Conversation (Maekawa et al., 2014; Koiso et al., 2022). Pairwise G-tests were all $p < .001$, and the direction persisted after removing the ten most frequent JCFPC items. Because the inventory was selected for frequency in these reference corpora themselves, the higher JCFPC incidence is unlikely to be an artefact of item selection. This comparison still does not generalize to all mimetics, all fiction, or all conversation. Its value is to show that JCFPC connects Japanese morphological retrieval to variable-length Chinese contexts and supports reviewable analyses of how expressive meaning is redistributed in translation.

7. Discussion and Limitations

The corpus confirms that a translation unit cannot be equated with a sentence-final punctuation mark. Wide variable groups permit compression and splitting to be represented without abandoning monotonic narrative order, while explicit one-sided units preserve genuine omission. The most important efficiency gain is not automatic

perfection, but a change in review structure: a researcher can judge a small set of high-information disagreements and locally close their sentence coverage instead of manually aligning the entire corpus from scratch.

The quality-control results also suggest a cautious role for AI in corpus construction. A similarity or reranker score is evidence, not a truth label. Correct literary paraphrases can score poorly, and incomplete concatenated groups can score highly. Multiple models are useful when they provide partly independent errors within structurally comparable pipelines, not simply because votes can be counted. The unit of comparison must therefore be a completed alignment output, and the unit of correction must be the full local patch affected by a decision. Versioned patches and rejected candidates are part of the research record because they explain why the final algorithm has its present constraints.

Several limitations bound the claims. The Naoki Prize and translation-availability frame is reproducible but not statistically representative; five works cannot separate author, genre, translator, or publisher effects. The 52 regression cases were used during development. The stratified audit has one judge and no inter-annotator agreement, while the 1.14% disjoint set was not exhaustively adjudicated. Thresholds and move ranges were tuned for this language pair and domain. The Chinese side has no symmetric morphological layer, and a monotonic hard constraint cannot encode genuine crossing correspondences. JCFPC has not been benchmarked against Vecalign or Bertalign on a shared Japanese-Chinese literary gold standard. The mimetic pilot further uses a fixed 200-item inventory, pools all five works, and relies on single-author coding. Its distributions should therefore be read as evidence of research potential, not general laws of Japanese-Chinese mimetic translation.

7.1 Availability and Access Conditions

A public pilot interface provides fixed-list access to approximately 200 Japanese mimetic lemmas and selected bilingual examples derived from JCFPC (LIAN, 2026; repository: https://github.com/lzq628/jcfpc-onomatopoeia-interface; demo: https://lzq628.github.io/jcfpc-onomatopoeia-interface/). It does not accept arbitrary full-corpus queries or redistribute the complete Japanese novels, Chinese translations, or text-bearing alignment database. Corrections and interface updates will be reported through the repository. Future work will investigate a server-mediated search service that returns bounded query-dependent snippets without full-text browsing, bulk export, or unrestricted API access, subject to institutional and legal review.

8. Conclusion

JCFPC links contemporary Japanese fiction to published Chinese translations through a morphological retrieval layer and variable-sized alignment units. Wide monotonic groups represent literary compression and splitting; independent refined pipelines identify informative disagreements; and coverage-closed, provenance-bearing patches convert human judgment into reproducible corpus updates. The resulting resource supports both transparent alignment evaluation and substantive translation analysis, as shown by the mimetic pilot. This workflow is potentially transferable to other small, copyright-restricted parallel corpora where research value depends less on raw scale than on retrievability, explicit uncertainty, and an auditable path from anomaly to correction.

Acknowledgements

This work was supported by JST SPRING, Grant Number JPMJSP2148.

Bibliographical References

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. (2024). M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318-2335. https://doi.org/10.18653/v1/2024.findings-acl.137.

Den, Y., Nakamura, J., Ogiso, T., and Ogura, H. (2008). A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. *Proceedings of LREC 2008*.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75-102.

Koiso, H., Amatani, H., Den, Y., Iseki, Y., Ishimoto, Y., Kashino, W., Kawabata, Y., Nishikawa, K., Tanaka, Y., Usuda, Y., and Watanabe, Y. (2022). Design and evaluation of the Corpus of Everyday Japanese Conversation. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 5587-5594. European Language Resources Association.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proceedings of EMNLP 2004*, pp. 230-237.

LIAN, Z. (2024). パタンマッチングによるオノマトペ候補語抽出の試み—オノマトペ形態変換プログラムを用いて— [An attempt to extract all the possible onomatopoeic words by using a pattern matching technique: Utilizing an onomatopoeic morphological transformation program]. *Proceedings of Language Resources Workshop*, 1:255-267. https://doi.org/10.15084/0002000369.

Liu, L. and Zhu, M. (2023). Bertalign: Improved word embedding-based sentence alignment for Chinese-English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621-634. https://doi.org/10.1093/llc/fqac089.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48:345-371. https://doi.org/10.1007/s10579-013-9261-0.

Miyamoto, H. (2023). 日中対訳コーパスの構築と公開に向けて [Toward the construction and publication of a Japanese-Chinese bilingual corpus]. *Proceedings of Language Resources Workshop*, 1:40-51. https://doi.org/10.15084/0002000113.

Thompson, B. and Koehn, P. (2019). Vecalign: Improved sentence alignment in linear time and space. *Proceedings of EMNLP-IJCNLP 2019*, pp. 1342-1348. https://doi.org/10.18653/v1/D19-1136.

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., and Zhou,

J. (2025). Qwen3 Embedding: Advancing text embedding and reranking through foundation models. arXiv:2506.05176.

Language Resource References

LIAN, Z. (2026). JCFPC Onomatopoeia Interface. GitHub repository and public demonstration. https://github.com/lzq628/jcfpc-onomatopoeia-interface.

Microsoft. (n.d.). Harrier OSS v1 and microsoft/harrier-oss-v1-27b model card. Microsoft Foundry Labs and Hugging Face. https://labs.ai.azure.com/innovations/harrier-oss-v1/.