jcls. Journal of Computational Literary Studies

OPEN ACCESS

Article

Translation-Based Connotation Visualization for Classical Poetic Japanese Vocabulary of the *Kokin Wakashū* ca. 905

Xudong Chen[1]
Yamamoto Hilofumi[1]
Hodoščček Bor[2]

1. School of Environment and Society, Tokyo Institute of Technology, Tokyo, Japan.
2. Graduate School of Humanities, Osaka University, Osaka, Japan.

Citation
Xudong Chen, Yamamoto Hilofumi, and Hodoščček Bor (2023). "Translation-Based Connotation Visualization for Classical Poetic Japanese Vocabulary of the Kokin Wakashū ca. 905". In: *Journal of Computational Literary Studies* 2 (1). 10.48694/jcls.3596

Date published 2024-02-15
Date accepted 2023-12-12
Date received 2023-01-31

Keywords
classical poetic Japanese text, parallel corpora, connotation, operationalization

Reviewers
Kiyonori Nagasaki, Paul Vierthaler

Note
This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. To offer a visualization of connotation in the classical poetic Japanese vocabulary of the *Kokin Wakashū* as an independent supplement for poetic language dictionaries, this paper presents an operationalization to tackle connotations using non-literal elements which are unveiled during the cross-cultural communication process, i.e., the translation process. Grounded on Schramm's communication model, we suggest calculating the set difference between the *Kokin Wakashū* and its ten contemporary Japanese translations to visualize the lexical explanatory additions (non-literal elements) in the translations. Methodologically, we apply the set difference in two distinct ways and implement the visualization on the six most frequent poetic flora words in the *Kokin Wakashū*, resulting in various depictions of non-literal elements. The set difference-based approaches to non-literal element visualization showed associative images and rhetorical techniques related to the flora words, which are two crucial aspects of connotation. While the other aspects of connotation, such as encyclopedic knowledge, sociolinguistic style, and emotion, are not covered by the proposed visualizations.

1. Introduction

Classical Japanese poetry is a representative type of poetry of classical Japanese literature, with a thematic tendency toward natural elements and a strict length limitation. Ki no Tsurayuki, the principal compiler of the first emperor-ordered anthology, the *Kokin Wakashū* (henceforth *Kokinshū* for short), noted in the *Kana Preface*:

> The seeds of Japanese poetry lie in the human heart and grow into leaves of ten thousand words [...] all that they think and feel is given expression in description of things they see and hear [...]. In the age of the awesome gods, songs did not have a fixed number of syllables [...]. By the time of the age of humans, [...] poems of thirty-one syllables were composed. Since then many poems have been composed when people were attracted by the blossoms or admired the birds, when they were moved by the haze or regretted the swift passage of the dew, and both inspiration and forms of expression have become diverse. (translation by Rodd et al. 1996 35–36)

As shown in the *Kana Preface*, classical Japanese poets used natural elements to hint at inner thoughts rather than expressed their thoughts directly. Moreover, the standard form, *tanka* 短歌 (en. short poem), is limited to 31 syllables (see the example below).

(1) つれづれの/**ながめ**に**まさる**/**なみだがわ**/そでのみぬれ て/あふよしもなし (*Kokinshū* 617)
tsurezureno **nagame**ni **masaru namidagawa** sodenomi nurete au yoshimo nashi
"in idle reverie/I weep tears that overflow like/the long rains of spring/my sleeves are drenched with the stream/that flows when we cannot meet" (translation by Rodd et al. 1996)

Poem (1) employs the ensuing rhetorical stratagems, thereby augmenting the polysemy and conveying rich information within the 31-syllables constraint:

- *kakekotoba* 掛詞, pivot words permitting manifold interpretations;
- *engo* 縁語, affiliated words fabricating parallel imagery.

In poem (1), *kakekotoba* and *engo* are in bold. "Nagame" is *kakekotoba* that has parallel meanings in its string (長雨, en. long rain, or 眺 め, en. reverie); "namidagawa" (en. river of tears) and "masaru" (en. rise/overflow) is *engo* for "nagame", since they are conceptually related to the sense of water. When "nagame" is interpreted as "long rain", "masaru" means the "rising" water in the river because of the rain; when "nagame" is interpreted as "reverie", "masaru" means the "overflowing" tears because of "reverie".

Such words are known as typical words in *uta-kotoba* 歌ことば, the vocabulary used in classical Japanese poetry (for the definition, Kubota 1994, 89). Words in the poetic vocabulary convey indirect and non-literal information densely as shown in the above example, which we view as connotation in this study.

To better understand the vocabulary used in the poetic language, we aim to visualize such lexical connotation in the *Kokinshū*, using its ten contemporary Japanese translations. The connotation visualization is intended to assist in supplementing classical poetic Japanese dictionaries. Technically, we adopt a connotation visualization strategy based on a comparison of textual information between original poems and translations, alluded to by the communication model (Schramm 1954). That is, the visualizations reveal connotation as the non-literal part of the original poems that bear explicit explanatory addition in the translations. To compare visualizations and dictionary descriptions, we experiment with the most frequent poetic flora words, as they are representative natural elements emphasized in the *Kana Preface*.

We structure the study as follows. In section 2, we introduce the background regarding the concept of connotation. Then we provide a brief view of the classical poetic Japanese dictionary (Katagiri 1983) and the translations of Japanese poetry. Finally, we present the theoretical basis of our method, Schramm's (1954) communication model. In section 3, we introduce the materials and two implementations of connotation visualization. In section 4, we provide six examples of flora poetic words. Besides, we also compare the two implementations and examine whether the visualization can reproduce the connotation included in Katagiri (1983), and whether it can reveal some aspects of connotation that do not exist in the dictionary. In section 5, we discuss what aspects of connotation the proposed visualization can present and what the visualization cannot. Also, we will discuss the contributions and the limitations of the current work.

2. Motivations

In section 2, we begin with the definitional inconsistency of connotation and the challenges in the operationalization, clarifying why we use non-literal elements as the basis for connotation visualization. Next, we explain why we use explanatory additions in translations rather than other materials to supplement the connotation description in the dictionary. Finally, we address non-literal elements from the perspective of the communication model by Schramm (1954).

2.1 Aspects of Connotation

In a basic definition, denotation is a word's explicit meaning, whereas connotation encompasses sociocultural and individual affiliations (Chandler 2002, 173–174). Nevertheless, the definition and scope of connotation vary among scholars. The scope pertains to sociolinguistic aspects (e.g., Bloomfield 1933; Hjelmslev 1969), emotional aspects (e.g., Eco 1976), and associative aspects (e.g., Rössler 1979). According to Eco (1976), semantic relationships such as hyponyms, hypernyms, and antonyms also pertain to the realm of connotation. In communication, pragmatic aspects of meaning, a speaker's attitude and a listener's value judgment of the perceived speech is also connotation (Mounin 1976, 159–160). They are independent between the speaker and the listener. From a speech without any intended connotation by the speaker, the listener can perceive unintended connotation (Rössler 1979, 101). Moreover, connotation is understood as inseparable from denotation (Stede 1999, 91; Stubbs 2002, 198; Voloshinov 1986, 105) and connotative meanings are inexhaustible (Chandler 2002, 139). Therefore, operationalizing connotation is challenging.

On the other hand, most of the aspects of connotation mentioned above are non-literal, except for some associations like collocates or phraseology as connotation. The non-literal aspect is a vital aspect we can use, as such non-literal elements can become literal when explicated for a better understanding of cross-cultural communication. Denotation and connotation exist at the conceptual level; they are inseparable and thus nonoperational without access to the target speech community to conduct psychological experiments or questionnaires. Conversely, literal and non-literal elements are physically written/unwritten; they are separable and thus operational when explanatory material is available. Hence, we may view non-literal elements as a physical world projection of the concept of connotation to aid operationalization (Table 1). Although the two are not equivalent, non-literal elements can intuitively reflect certain aspects of connotation. Dictionaries, introductory books, philological annotations, and translations can serve as explanatory materials.

The operationalization of lexical connotation has been attempted within computational linguistics and corpus linguistics. The operationalization mainly focuses on certain aspects such as emotional value, sentiment, valence, impact, and collocates (e.g., Allaway and McKeown 2021; Rashkin et al. 2016; Stubbs 2002), while the intuitive aspect, the

	explicit	implicit
physical (operational)	literal (projection of denotation)	non-literal (projection of connotation)
notional (nonoperational)	↑ denotation	↑ connotation

Table 1: Relationship between denotation/connotation and literal/non-literal: Dashed lines indicate the instability of the division; solid lines indicate the stable division.

non-literal aspect of connotation, is not widely utilized.

2.2 Dictionary and Translation as Explanatory Materials

Many materials can serve to turn non-literal elements into literal ones. Among the materials, dictionaries and translations can provide a systematic approach to classical Japanese poetic vocabulary, whereas the others typically select certain poems or words to offer specific explanations and therefore are selective. Subsection 2.2 discusses a comprehensive dictionary for classical poetic Japanese (Katagiri 1983) and explains why we need to use additional information in translations as an independent supplement for the dictionary.

Katagiri (1983) includes 830 lexical entries. The dictionary does not explicitly distinguish between connotation and denotation, yet covers a broad spectrum of meaning, encompassing the two. Connotations in the dictionary pertain to collocates, phraseological patterns, and figurative/rhetorical usages. For instance, its description of "sakura-bana" (en. cherry blossoms) incorporates collocates such as "chiru" (en. falling) and associations such as the impermanence of life and the passing of spring (Katagiri 1983, 172–173).

However, dictionary compilers inevitably apply their conscious knowledge in the dictionary compilation. For example, compilers decide which words to include, which examples and meanings to emphasize. Whereas, when actually translating Japanese poetry, relying solely on the dictionary proves insufficient for producing comprehensive translations. As Masao Takeoka, one of the translators of the *Kokinshū*, noted:

> The translation is not solely an introduction or explanation of the original text's "plot". It is, after comprehending the author's way of perception and feeling from all aspects, entirely transforming the perceptions and feelings into the same or as similar as possible expressions that exist in contemporary language. (Takeoka 1976a, 11, translation by the authors)

To provide a comprehensive understanding of the "perceptions and feelings" of poets in translations, beyond consciously aligning corresponding equivalent words, translators must simultaneously incorporate those perceptions and feelings about each poetic word into their translations. These incorporations are rarely covered in the dictionary, as they remain unconscious for compilers until they are translated within actual contexts. Let us have a look at the following example of a parallel text.

(2) a. つれづれの/ながめにまさる/なみだがわ/そでのみぬれて/あふよしもなし (*Kokinshū* 617)

tsurezureno nagameni masaru namidagawa sodenomi nurete au yoshimo nashi
"the long rains of reverie drives the river of tears to surge. [I can] only dampen my sleeves and have no means to meet [you]" (literal English translation by the authors)

b. 長雨のみならず私の物思いの眺めによって水かさが増している涙の川、渡ろうと思っても、涙で袖が濡れるばかりで渡ることができず、逢いに行く手段とてありません。(contemporary translation by Katagiri 1983, 318)
nagamenominarazu watashino monoomoino nagameniyotte mizukasaga mashiteiru namidano kawa, wataruto omottemo, namidade sodega nurerubakaride watarukotoga dekizu, aini iku shudantote arimasen.
"Due to not only the long rains but also my gaze of reverie, the river of tears surges. [I] endeavor to cross the river, but can only dampen my sleeves with tears, unable to cross. There is no means to go to meet [you]." (literal English translation by the authors)

In this parallel text, the translation not only renders the two parallel images, long rain, and reverie, of the *kakekotoba* "nagame" by adding corresponding words "monoomoi no nagame" (en. gaze with reverie), but also uses additional words to highlight the aspiration to "cross the river" for the word "namidagawa" (en. river of tears). Although the dictionary covers rhetorical usages like "nagame", an important aspect of connotation, understanding the connotation of the dilemma of longing to cross the river to reach a lover, yet unable to, is challenging through the dictionary alone. We, therefore, pay attention to the explanatory addition in translations. The objective of providing an independent supplement to the dictionary necessitates the visualization of the unconscious knowledge exposed by translators. Through a dictionary-independent visualization system, novice readers can query all lexical entries in the *Kokinshū*, not just the consciously crucial lexical entries and usages pre-selected by compilers.

2.3 Non-Literal Elements in Schramm's Communication Model

Based on the above two sections, the operationalization of the concept connotation will be based on the operationalization of non-literal elements revealed in the translations. This section introduces a specific way of addressing non-literal elements in parallel texts with hints from Schramm's (1954) communication model.

Schramm's model is considered to be more suitable for this study compared to other communication models because it uses fewer concepts and adopts a concept, the field of experience of participants, which is useful for explaining the success or failure of communication (Yamamoto 2005, 26). Stuart Hall's (2018) encoding/decoding model of communication, which is well-established in the cultural studies community, also includes variables that influence communication. However, it serves different analytical purposes less suited to the current study.

Schramm's model features five components: source, encoder, signal, decoder, and destination (Figure 1). The source encodes a message that the destination decodes. In

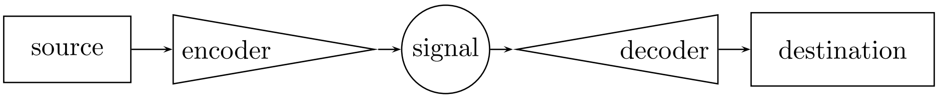
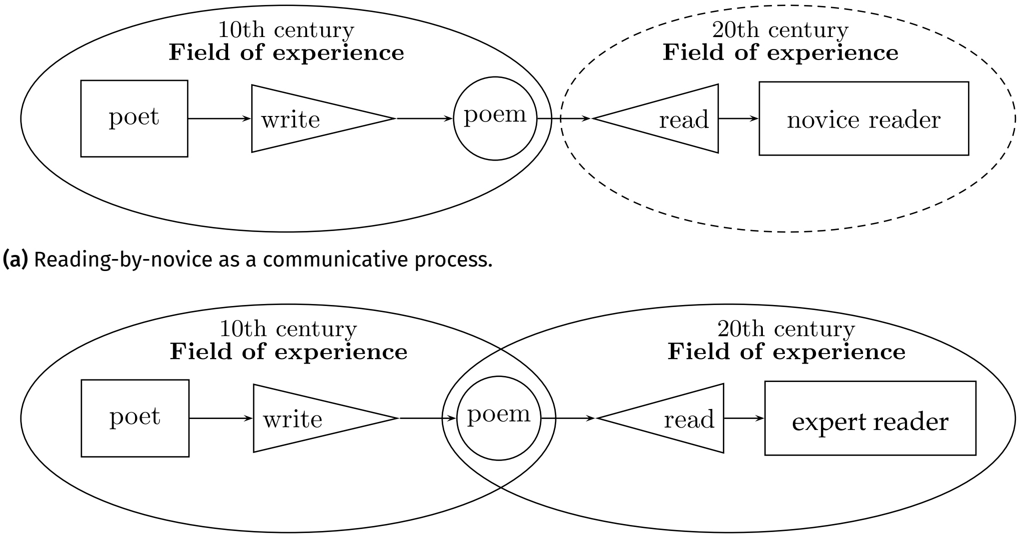


Figure 1: Communication process by Schramm (1954, 4): A communicative process consists of source, encoder, signal, decoder, and destination.



(a) Reading-by-novice as a communicative process.

(b) Reading-by-expert as a communicative process.

Figure 2: Reading poetry as communicative processes based on Yamamoto (2005, 27): (a) reading classical Japanese poetry by a novice reader; (b) reading classical Japanese poetry by an expert; circles indicate the field of experience.

human communication, senders encode thoughts into the language (the signal), and receivers decode it (Schramm 1954, 6).

Yamamoto (2005) proposes that the translation and reading behaviors of classical Japanese poetry be considered communication processes. In the process of reading, the poet is the encoder who transfers thoughts into a poem; the reader (novice/expert reader) is the decoder who processes the information in the poem into their understanding; the poem is a signal in the communication process (Figure 2). The translation process follows the reading process: After processing the information of the poem into understanding as a decoder, the expert turns into an encoder who transfers the understanding of the poem into contemporary translations.

The difference in the reading process between a novice reader and an expert reader lies in the “field of experience” (Schramm 1954, 6), which refers to accumulated experience. When a novice reader attempts to understand a poem as shown in Figure 2a, the novice reader may fail to understand the poem as the novice reader in the twentieth-century shares no field of experience with a poet who lived in the tenth century. Conversely, an expert in classical Japanese poetry, having accumulated relevant knowledge from numerous materials, shares a broader field of experience (Figure 2b). Hence, an expert can decode classical Japanese poetry better than a novice reader.

Sharing a similar field of experience means that most of the information can be understood non-literally; on the other hand, sharing a different field of experience means most information must be explained literally to be understood. Between the two communication processes with distinct fields of experience, the process of translation by experts

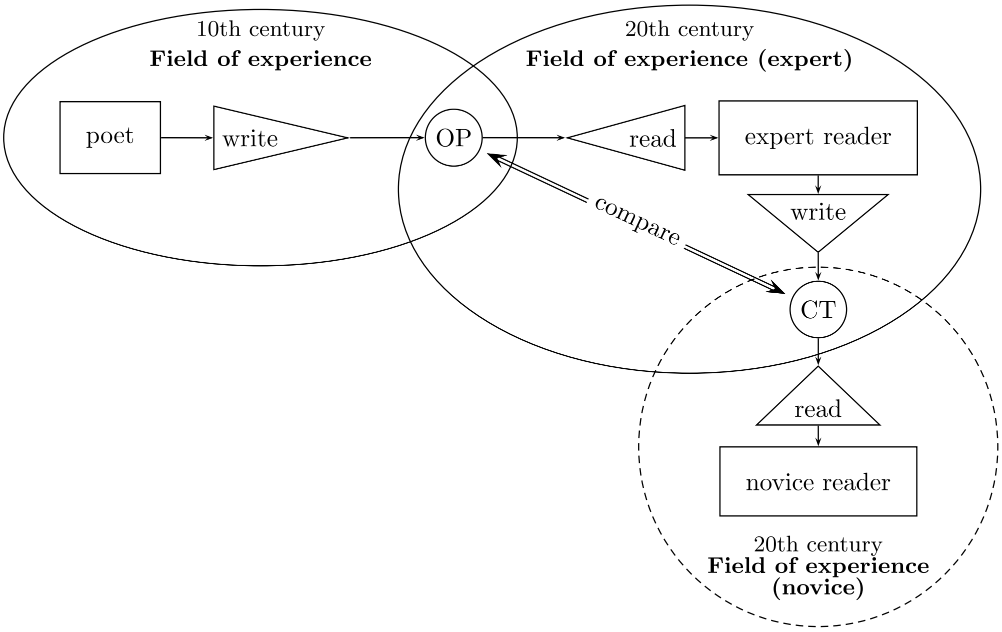


Figure 3: Non-literal elements in the schema of Schramm's communication process (reused from Yamamoto 2005, 28): We simplify non-literal elements as the residual between two kinds of signals – the poem and the translation; CT indicates contemporary translations; OP indicates original poems.

acts as a bridge (Figure 3). In the translation process, to share the information in the original poem smoothly with novice readers, experts translate literal elements and add non-literal elements into translations.

On this basis, we can formalize the non-literal elements (Equation 1), where an original poem (OP) is a set consisting of literal elements and a contemporary translation (CT) is a set consisting of elements in OP and additional non-literal elements. Such formalization turns the extraction of non-literal elements into a well-defined technical problem of calculating the complement set.

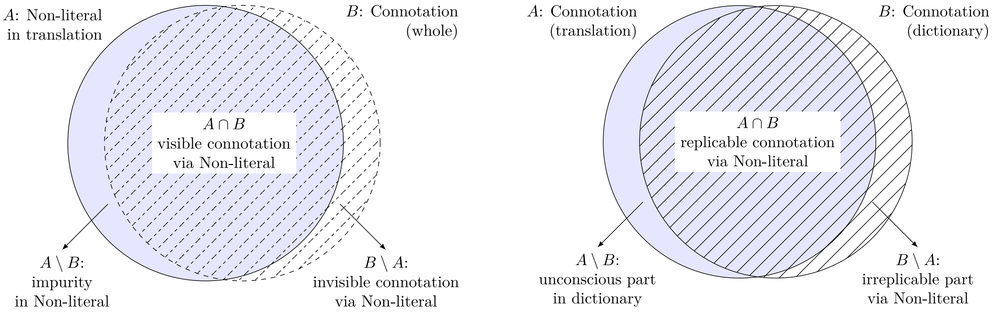
$$\text{Non-literal elements} \quad = \quad \text{CT} \setminus \text{OP} \tag{1}$$

2.4 Summary and Potential Issues

In section 2, we assumed that through non-literal elements revealed in translations, we can visualize a portion of lexical connotation. Moreover, non-literal elements revealed in translations can provide a heuristic for the unconscious aspect of connotation that is not covered by dictionaries. We can extract the non-literal elements by computing the set difference between translations and original poems, guided by the hint from Schramm's communication model.

On the other hand, we must highlight some issues when utilizing non-literal elements for connotation visualization.

The first is the coverage issue, which is unsolvable because "no inventory of the connotative meanings generated by any sign could ever be complete" (Chandler 2002, 139).


A: Non-literal in translation
B: Connotation (whole)
$A \cap B$ visible connotation via Non-literal
$A \setminus B$: impurity in Non-literal
$B \setminus A$: invisible connotation via Non-literal
A: Connotation (translation)
B: Connotation (dictionary)
$A \cap B$ replicable connotation via Non-literal
$A \setminus B$: unconscious part in dictionary
$B \setminus A$: irreplicable part via Non-literal

(a) Relationship of non-literal elements and connotation.

(b) Relationship of connotation in translations and dictionaries.

Figure 4: Relationship between connotation and non-literal elements, and relationship between connotation in translations and dictionaries: (a) non-literal elements can reflect some aspects of connotation, but they can also include non-connotative elements and cannot cover the whole connotation; (b) Connotations revealed in translations cover some of the connotations described in dictionaries and visualize the unconscious aspects not explicitly mentioned in the dictionaries; the dashed line represents an open set; the solid line represents a closed set.

That is, through non-literal elements, we cannot visualize all aspects of connotation and cannot reproduce all connotations described in the dictionary (Figure 4).

The second is the "impurity" issue. The impurity mainly refers to the inclusion of function words used to facilitate the translations in the extracted non-literal elements (Figure 4a), which represent the syntactic differences between target languages and source languages. In the following sections introducing the methodology, we will explain how we handle these "impurities".

3. Methods

Section 3 introduces the materials and two set difference-based implementations for visualizing connotations – one based on word misalignment and one based on salient word co-occurrence.

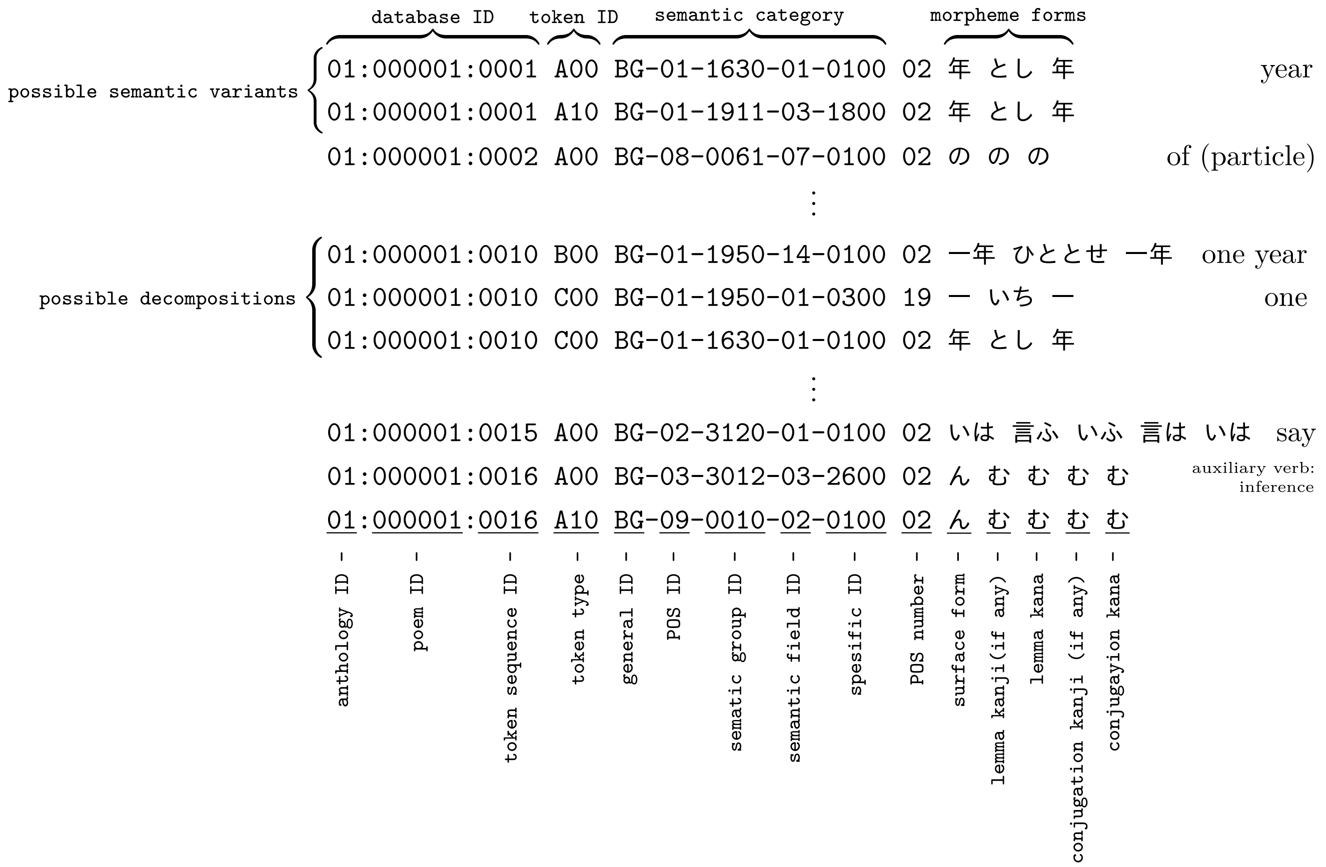
3.1 Materials

3.1.1 The *Kokinshū* Text Data

The *Kokinshū* comprises 1,111 poems. We use the initial 1,000 poems[1], which are classified as *tanka* to maintain uniformity in poem length.

We use the *Kokinshū* text data from the *Hachidaishu* Vocabulary Dataset (Hodošček and Yamamoto 2022).[2] The dataset provides a TEI format version and a space-delimited format version. For compatibility with the translation data format, we choose the space-delimited format (refer to Figure 5 for details). The space-delimited format version is generated by automatically processing using a domain-specific morphological analysis system for classical Japanese poetry (Yamamoto 2007) and a semantic category code

[1] The remaining 111 poems consist of *choka* (long poem), *sedoka* (head-repeated poem), *azuma-uta* (poem written in Eastern dialect), and others. Some vary in length and form from *tanka*, some employ dialects, and some touch upon religious matters.
[2] The *Hachidaishū* consists of the first eight imperial anthologies of classical Japanese poetry, with *Kokinshū* being the initial anthology.



```
01:000001:0001 A00 BG-01-1630-01-0100 02 年 とし 年
01:000001:0001 A10 BG-01-1911-03-1800 02 年 とし 年
01:000001:0002 A00 BG-08-0061-07-0100 02 の の の
                            ⋮
01:000001:0010 B00 BG-01-1950-14-0100 02 一年 ひととせ 一年
01:000001:0010 C00 BG-01-1950-01-0300 19 一 いち 一
01:000001:0010 C00 BG-01-1630-01-0100 02 年 とし 年
                            ⋮
01:000001:0015 A00 BG-02-3120-01-0100 02 いは 言ふ いふ 言は いは
01:000001:0016 A00 BG-03-3012-03-2600 02 ん む む む む
01:000001:0016 A10 BG-09-0010-02-0100 02 ん む む む む
```

Figure 5: Space-delimited format of the *Hachidaishu* Vocabulary Dataset (Yamamoto and Hodošček 2021): A line consists of seven columns separated by spaces. The first column `01:000001:0007` consists of three fields separated by colons: 1) anthology, 2) poem number, and 3) serial ID of the token. The anthology ID 01 indicates the *Kokinshū*. the second column indicates the type of token: `A` is a single token; `B` is a compound token; `C` is a breakdown of `B`. `A00` indicates a single token; `A01` indicates a single token with another meaning; `B00` indicates a compound token; `B01` indicates a compound token that has another meaning; `C00` indicates the first element of the `B00`/`B01`.. breakdown; `C01` indicates the second element of the `B00`/`B01`.. breakdown. the third column `BG-02-1527-01-0102`: Classification ID based on semantic categories; 4–9th column indicates respectively: a Part-of-Speech number, a form that appears in literary works, a lemma in kanji script, a lemma in kana script, conjugated form in kanji writing form, conjugated form in kana writing form.

annotation system (Yamamoto 2009). The semantic category codes are based on an old version of the *Word List by Semantic Principles* (WLSP; Nakano et al. 1994), a collection of words classified and organized by semantic categories. The dataset includes both compound tokens and their constituents, i.e., decomposed simplex tokens, and assigns multiple semantic category codes to each polysemous token.

We use semantic category codes during processing instead of lemmata. For each compound token, we use the token rather than its decompositions. In cases where a token has multiple semantic category codes, we use the first code. We do not exclude any stop words in the preprocessing since function words in classical Japanese poems can function as content words simultaneously.[3]

[3] For example, *"tsuru"* is a conjugated form of *"tsu"*, which is a function word used to express the perfective aspect. The reading of *"tsuru"* also simultaneously carries the additional meaning of *"crane"*.

	abbreviation	references	manuscript	translation style
1	KNK	Kaneko (1933)	Teika	word-for-word
2	KBT	Kubota (1960a,b,c)	Teika	word-for-word
3	MTD	Matsuda (1968a,b)	Teika	not mentioned
4	OZW	Ozawa (1971)	Teika	wording changed
5	TKOK	Takeoka (1976a,b)	Teika	word-for-word
6	OKMR	Okumura (1978)	Teika	intention oriented
7	KSJ	Kyusojin (1979)	Teika	word added
8	KMCY	Komachiya (1982)	Teika	not mentioned
9	K&A	Kojima and Arai (1989)	Teika	not mentioned
10	KTGR	Katagiri (1998a,b,c)	Teika	word-for-word

Table 2: Ten contemporary Japanese translations of the *Kokinshū* and their translation styles, ordered by year.

3.1.2 Translation Text Data

We use the ten contemporary Japanese translations (Table 2), where all translation texts are tokenized using Chasen (Matsumoto et al. 2002), a Japanese morphological analysis system.[4] We use Chasen because it has the same Part-of-Speech schema as that of the *Hachidaishu* Dataset, which allows us to generate consistent and compatible tokenization results with the *Hachidaishu* Dataset. Contemporary Japanese translations are typically rendered in a plain writing style rather than adopting a poetic tone, representative texts of which can be processed by the algorithm of Chasen with an F-score of 98% (Asahara and Matsumoto 2000; Kudo et al. 2004). We furthermore conducted manual revisions of the results, and in cases where certain words were absent from the Chasen system's dictionary, we supplemented the lexicon by consulting authoritative Japanese dictionaries. We annotate each token with the same semantic category code system (i.e., WSLP) as that of the *Kokinshū* data. The annotation allows us to determine semantic equivalence between tokens in the old and contemporary Japanese.

Although many translators claimed to follow a word-for-word translation style, Yamamoto and Hodošček (2019) have demonstrated that regardless of the translation styles, all translations include approximately 50% of tokens that lack semantic equivalence with any token in the corresponding original poems.

The preprocessing for the translation data is similar to the *Kokinshū* data. Besides, we remove all punctuation marks presented in the translation data. Currently, the translation data is not open-sourced due to copyright restrictions.

3.1.3 Data Summary

Table 3 summarizes the post-preprocessed data. The *Kokinshū* and its ten translations form a parallel corpus. The corpus includes 10,000 parallel texts. The corpus is smaller than the so-called big data used in state-of-the-art natural language processing studies. The lack of parallel historical data hinders the application of state-of-the-art learning techniques (Kalouli et al. 2019, 109). Therefore, we opt for traditional computational methods in this study.

[4] Because Chasen can only process contemporary Japanese orthography, we convert the historical *kana* orthography used in the earliest translation (Kaneko 1933), which was widely used before orthographic reforms post World War II.

abbreviation	# of tokens	# of types	# of texts
KNK	42,439	3,356	1,000
KTGR	36,362	2,882	1,000
K&A	33,867	2,955	1,000
KMCY	30,869	2,692	1,000
KBT	32,210	2,701	1,000
KSJ	34,050	2,770	1,000
MTD	31,860	3,007	1,000
OKMR	32,321	3,153	1,000
OZW	36,173	3,384	1,000
TKOK	29,844	2,861	1,000
total	339,995	8,252	10,000
Kokinshū	16,687	1,496	1,000

Table 3: Details of the preprocessed data.

3.2 Implementation A: Misalignment-Based Visualization

Implementation A builds on Chen et al. (2022), which visualizes the non-literal elements with a statistical word alignment model, IBM model 2 (Brown et al. 1993). This method applies the set difference in the form of misalignment extraction.[5] We train two statistical word alignment models: one trained with parallel texts in source-to-target order (Model A); the other trained with parallel texts in target-to-source order (Model B). From the alignments inferred by Model B, we deduct alignments inferred by both Model A and Model B. The remaining pairs form a misalignment of Model B. In each pair in the misalignment, the target word is not a precise translation for the source word, yet statistically and semantically related to the source word. We view the target word as a non-literal element added for the source word to transmit information that requires verbalizing in the poem to novice readers. As observed in subsection 2.4, non-literal elements can be function words, which do not represent connotation. We omit those function words from the extracted non-literal elements with post-processing. Detailed steps are as follows.

Step A: Training of Word Alignment Models. We apply IBM Model 2 as the word alignment algorithm to the parallel corpus.[6] We train two models: a source-to-target model (Model A) and a target-to-source model (Model B). We use the IBM Model 2 from NLTK 3.7.0, a natural language processing toolkit, with Python 3.8. The intersection of alignments by Model A and B results in a 69.85% precision while Model B results in a significantly lower precision of 38.25% when iteration is 8, indicating we can extract numerous misaligned patterns.[7] For the intersection part, the algorithm itself is very

[5] Alignment of a pair of parallel texts is a set of word pairs, each comprising a word in the source text and a word in target text (Koehn 2010, 84). When pairs in an alignment by a model are not replicated in an alignment by another model, we call them misalignment.
[6] The implementation can also utilize any other algorithm if the algorithm does not require additional data annotations.
[7] Detailed precision, recall, and AER (average error rate) of word alignment models are reported at `https://github.com/nehcx/kokinMisalign`, accessed on 20 May 2023. To calculate these statistics, we used the consistency between the WLSP semantic category codes of each word pair in an alignment to judge whether a pair is a sure link, a possible link, or a misaligned link. This led to an extremely strict criterion for sure alignment and possible alignment; therefore, the precision and recall reported are considered much lower than actual, and AER is higher than actual.

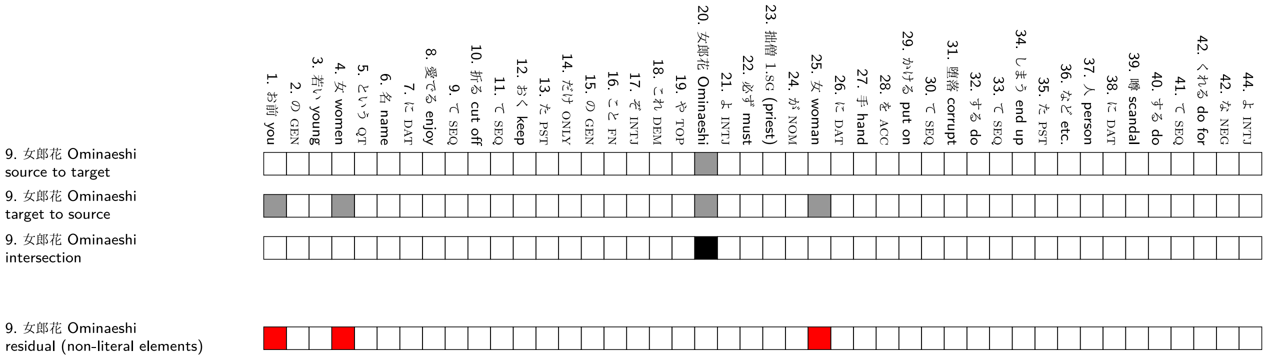


Figure 6: Procedure of the set difference of aligned results of the poetic word "*ominaeshi*" 女郎花 (en. golden valerian) in the 226th *Kokinshū* poem and its translation by Kaneko (1933) (figure based from Chen et al. 2022): The first row is the aligned results of (a); the second row is the aligned results of (b); the third row is (c), the intersection of (a) and (b); the final row is the complementary set of (a) in (b) and hence the non-literal information.

open to improvement to guarantee that most of the sure links are ruled out; for the Model B part, the accuracy criterion should be loosened in the implementation to add more misaligned links.

Step B: Extraction of Misalignment. For each parallel text containing the queried word, we apply the word alignment models trained in Step A to the parallel text in the source-to-target direction (Model A) and the target-to-source direction (Model B). For a queried poetic word, there are three types of alignments with words in translations inferred: (a) an aligned translation word inferred by Model A; (b) one or several aligned translation words inferred by Model B; (c) an intersection of the above two, which indicates the potentially precise alignment. Aligned translation words in (b) can contain incorrect but statistically related translation words for the queried poetic word. To obtain misalignment, we subtract (c) from (b). Figure 6 illustrates the procedure.

Step C: Post-Processing to Filter Function Words. While Chen et al. (2022) do not explicitly include this step, the study implied that without the step, the non-literal elements in the results inevitably contain function words. Therefore, Step C removes those functional elements from the extracted non-literal elements, retaining only the connotative elements. The filtering is based on part of speech. Function words that are not nouns, verbs, adjectives, or adverbs are excluded before visualization.

Step D: Network Visualization in Three Phases. We execute steps B and C on each parallel texts and build an aggregate visualization by accumulating misalignments, following a bottom-up strategy. This means that we can deconstruct the visualization, phase by phase (Figure 7): Phase 1 visualizes the non-literal elements from a single parallel text; Phase 2 visualizes the non-literal elements from ten parallel texts of the same poem (a single poem containing the queried word with its ten translation texts); Phase 3 visualizes the non-literal elements from all parallel texts that include the queried word. Phase 3 is the overall aggregate visualization of the non-literal elements for the queried word. The aggregate visualization shows which non-literal elements of the queried poetic word are commonly added by various translators. On the other hand, Phases 1 and 2 depict non-literal elements for the word at the poem level, aiding in contextual checks.

In summary, Implementation A visualizes non-literal elements for poetic words as misaligned semantic additions in translations. Implementation A can trace non-literal elements of a poetic word in detail with its three-phase visualization: It shows from which translation version a non-literal element originates, and from which poem the non-literal element comes. On the other hand, Implementation A is designed to visualize non-literal information with a strict criterion – if there is no explanatory addition in the translations directly for the queried poetic word, the visualization outputs nothing. This does not mean the poetic word has no connotation; rather, it means all ten translators commonly consider that the poetic word is understandable to novice readers in a poem or multiple poems without any explanation. In other words, the connotation of the word is largely shared between poets and contemporary readers. In this case, visualizable connotation via non-literal elements is limited, which can be a drawback. Implementation B in the next section can address this drawback.

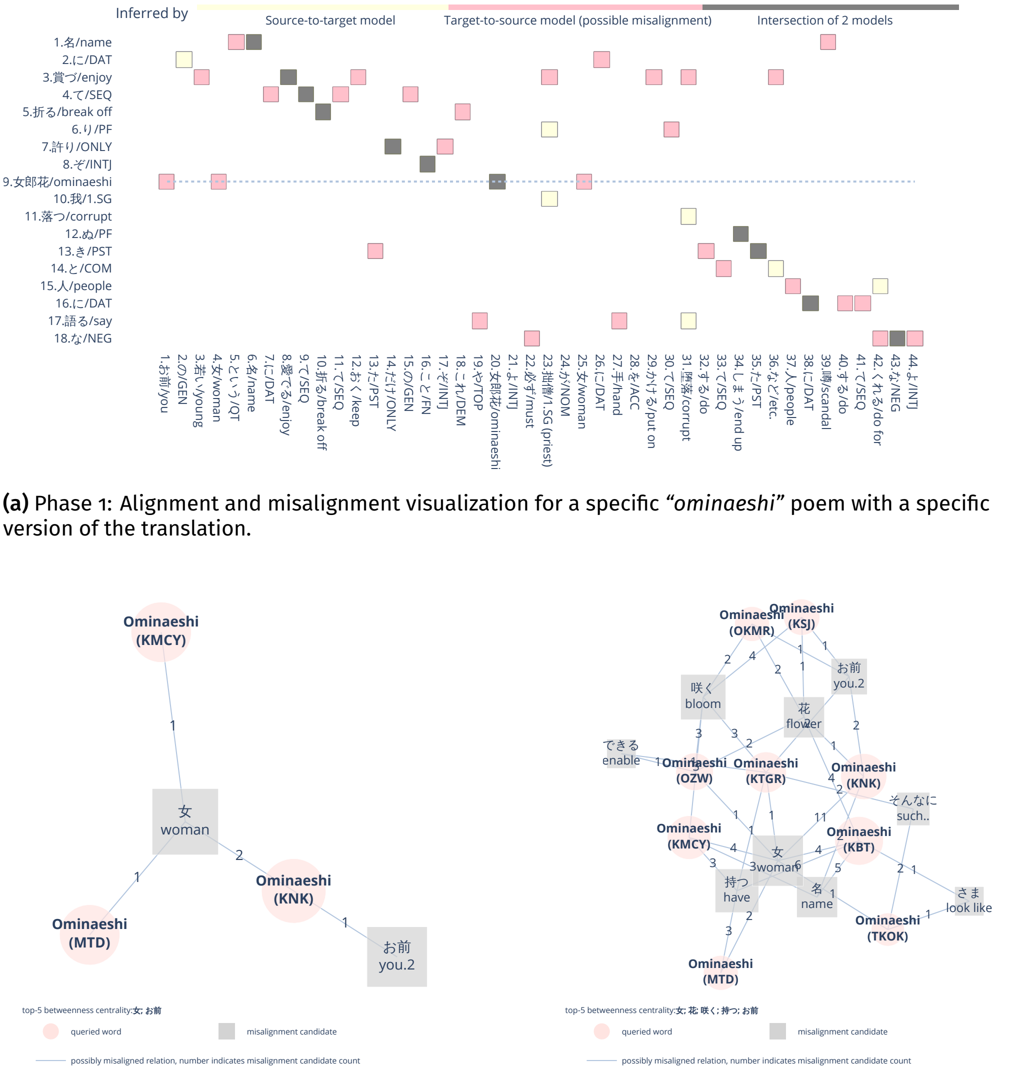
Implementation A is available on Github.[8] The repository provides a dashboard built with Dash 2.7.1 in Python 3.8. This dashboard integrates all the three phases discussed in Implementation A. Upon entering the dashboard, we can first access the query interface to search for words in classical Japanese poetry. The query results in a table composed of parallel texts where the searched word appears. We can select a certain version of the translation (the default version is KNK/Kaneko 1933). By clicking on the word alignment visualization tab, we can see the word alignment outcomes of the selected parallel text (Phase 1 visualization). By clicking on the misalignment network visualization tab, we can see the Phases 2 and 3 visualizations.

3.3 Implementation B: Co-Occurrence Pattern-Based Visualization

Implementation B builds on Yamamoto (2005), which is an information-theoretical-based approach. Unlike Implementation A, the analysis unit of Implementation B is the co-occurrence pattern rather than the word. In Implementation B, both the poem set (OP) and translation set (CT) are sets of co-occurrence patterns. The set difference is, therefore, used to visualize co-occurrence patterns that exist only in CT.

In Implementation B, we use co-occurrence patterns as units for subsequent reasons. Firstly, if we merely calculate the set difference using two bag-of-words, the remaining words in the complementary set for CT cannot keep their minimum contexts in OP. This diminishes the interpretability of the visualized non-literal elements. Secondly, without co-occurrence, we cannot tell whether an extracted non-literal element is added for the queried word. Unlike Implementation A, which associates non-literal elements with queried words as misaligned links, the set difference of two bag-of-words can fail to indicate which elements link to the queried word and how. Thirdly, the co-occurrence-based implementation can tackle the weaknesses of Implementation A: When there is no explanatory addition directly for a queried poetic word, Implementation A may visualize nothing for the word; inversely, Implementation B can still visualize some elements. In the instance of the co-occurrence pattern, for a poetic word p_a, if it shares

[8] See: https://github.com/nehcx/kokinMisalign, accessed on 20 May 2023.



(a) Phase 1: Alignment and misalignment visualization for a specific *"ominaeshi"* poem with a specific version of the translation.

(b) Phase 2: Misalignment network visualization for a specific *"ominaeshi"* poem with multiple versions of translations.

(c) Phase 3: Overall misalignment network visualization for all *"ominaeshi"* poems with all versions of translations.

Figure 7: Phase-by-phase misalignment visualization for *"ominaeshi"* (女郎花, en. golden valerian): (a) the detailed supplementary translation words for golden valerian in a specific parallel text; (b) collective visualization of supplementary translation words by different translators for golden valerian in a specific poem; (c) overall visualization of supplementary translation words for golden valerian in all the poems.

a co-occurrence relation with a word p_b in both OP and CT, the relation (p_a, p_b) will not be visualized directly for p_a; however, when p_b holds a co-occurrence pattern with an additional element c in translations, p_b can remain as the non-literal element in the visualization for poetic word p_a in the form of co-occurrence pattern (p_b, c). That is, Implementation B can visualize not only co-occurred words in poems (e.g., p_b) but also indirectly related textual additions in translations (e.g., c) as non-literal elements[9] for the queried word p_a. Such non-literal patterns are not remnants of the direct co-occurrence with the queried word in original poems, but minimum context indirectly related to the queried word that occurs in translations.

Implementation B also has the issue that non-literal elements may contain non-connotative impurity. To tackle this issue, unlike the post-processing in Implementation A, Implementation B filters co-occurrence patterns with a flexible keyness threshold of patterns. The detailed processes are as follows.

Step A: Construction of Two Sets of Co-Occurrence Patterns for Queried Words. Suppose we query a poetic word. We keep each original poem and each translation that contains the queried poetic word. From the retained poems, we collect all the co-occurrence patterns, regardless of the linear distance between two co-occurring words, and form the queried Poem set (OP) of co-occurrence patterns. The same is carried out on the retained translations to form the queried Translation set (CT).

Step B: Co-Occurrence Pattern Keyness Calculation. We calculate the keyness of each co-occurrence pattern in OP and CT relative to the whole poems/translations respectively. The keyness is for filtering function words and unimportant words. The relative keyness is modeled by cw (Equation 2).[10]

$$cw(t_1, t_2, d) = (1 + \log ctf(t_1, t_2, d)) \cdot \sqrt{idf(t_1) \cdot idf(t_2)} \tag{2}$$

$$idf(t) = \log \frac{N}{df(t)} \tag{3}$$

where $ctf(t_1, t_2, d)$ is the co-occurrence frequency of words t_1 and t_2 in a document d; $df(t)$ is the document frequency of word t. We here opt to use cw because the cw considers not only the keyness of co-occurrence patterns as a whole but also the keyness of each word in each pattern simultaneously. $\sqrt{idf(t_1) \cdot idf(t_2)}$ emphasizes the keyness of t_1 and t_2 should be high simultaneously for document d; otherwise, (t_1, t_2) may not be a key pattern. By doing so, we can avoid unimportant function words from being included in key patterns. Conversely, general keyword extraction methods for single

[9] More precisely, p_b is literal in poems; while c and the minimum context (p_b, c) for p_b in translations is non-literal for the poems.

[10] cw is an extended form of $tfidf$ (Manning et al. 2008, 109), $tfidf(t, d) = tf(t, d) \cdot idf(t)$, where $tf(t, d)$ is the relative frequency of term t within document d; $idf(t)$ (Equation 3) is the inverse document frequency of term t (Spärck Jones 1972). $tfidf$ models the importance of a single word in a specific document. In Equation 2, $(1 + \log ctf(t_1, t_2, d))$ and $\sqrt{idf(t_1) \cdot idf(t_2)}$ correspond to the tf part and the idf part in $tfidf$, respectively.

words may ignore the keyness of each word in patterns, leading to high keyness for patterns like high *idf* content word with low *idf* function word.[11]

Step C: Filtering Co-Occurrence Patterns. We filter co-occurrence patterns where *cw* is lower than a specific threshold to automatically remove "impurity" before the calculation of set difference. According to Yamamoto and Hodošček (2018), *cw* is normally distributed; after standardization, *cw* can sample key patterns consisting only of content words by setting the threshold as one standard deviation of the normal distribution. In this study, however, we do not use the threshold of one standard deviation. To simultaneously reduce the overlap of labels in visualization, we adjusted the threshold flexibly. Because of the different sizes (usually a ratio near 1:3 or 1:4) between the OP and CT, the thresholds for the two sets are set differently. The thresholds for CT should be 3 or 4 times higher than that of OP.

Step D: Set Difference. We subtract the intersection of the two filtered sets of co-occurrence patterns from the filtered CT. The residuals are patterns that saliently occur in translations but not in originals. Figure 8 shows the procedure of the set difference. Details of the example complementary set can be seen in Table 4.[12]

Step E: Network Visualization. The visualization uses the `dot` language to visualize important co-occurrence patterns that are added in translations. These co-occurrence patterns form networks and present a global view of additional information and the relation between the queried word and the additional information.

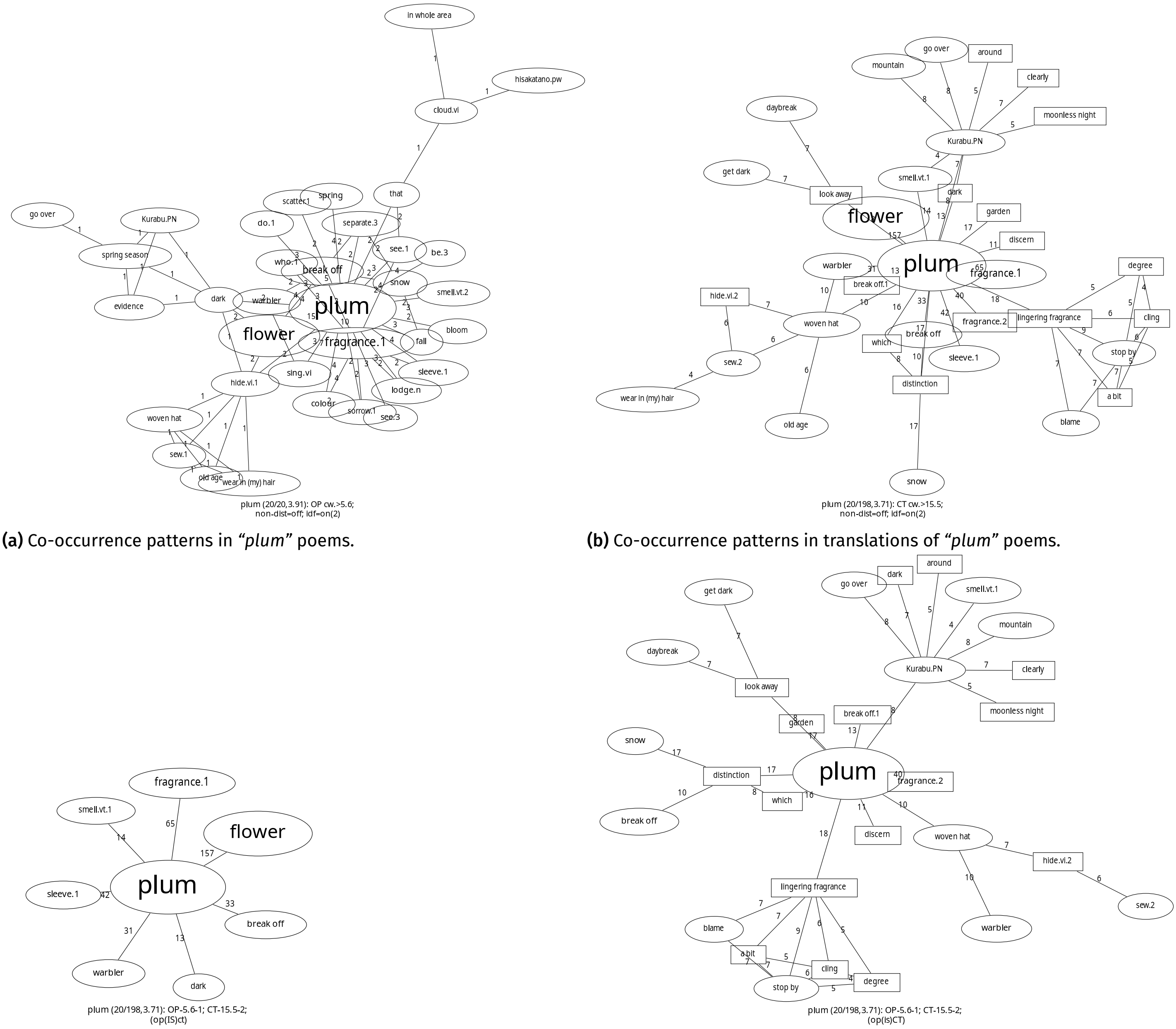
In Implementation B, the complementary set contains two types of non-literal elements: (a) explanatory additions that are added directly for the queried words (e.g., "break off" for "plum" in translations); (b) collocates with the queried words in original poems, which hold explanatory additions in translations (e.g., "woven hat"-"hide" for "plum", in which "hide" is the explanatory addition for the "woven hat" in translations). Different from Implementation A, non-literal elements visualized by Implementation B include not only words but also direct and indirect relations.

We provide a simple web application to visualize key co-occurrence patterns for classical poetic words.[13] The web application can also be applied to translations, but translations are currently protected by copyright. Therefore, the set difference of co-occurrence patterns is not publicly available.

[11] Aside from *cw*, various keyness measures are available (e.g., Burrows 2007; Dunning 1993; Spärck Jones 1972), while comparisons among measures (e.g., Du et al. 2022; Paquot and Bestgen 2009; Schöch et al. 2018) have mainly focused on traditional linguistic units, i.e., words, rather than extended units of meaning such as co-occurrence patterns. The recent systematic comparison (Du et al. 2022) among the measures suggests dispersion-based methods can better differentiate texts in the case of shorter, randomly selected segments. However, because each classical Japanese poem ends within a single sentence, with each word type appearing about 1-2 times, the advantages of dispersion-based methods may not be fully utilized. Comparison of measures specific for such extremely short forms of literary texts and for extended linguistic units is necessary in the future.

[12] The full list, before reducing the translation set by removing the original set, is available on `https://github.com/nehcx/kokinMisalign/tree/master/supplementary_materials`, visited on 27, May 2023.

[13] See: `https://cuckoo.js.ila.titech.ac.jp/~yamagen/waka/poem.cgi`, accessed on 20, May 2023. Currently, the visualization does not provide English translations.



(a) Co-occurrence patterns in "*plum*" poems.

(b) Co-occurrence patterns in translations of "*plum*" poems.

(c) Intersection of (a) and (b).

(d) Set difference of (a) and (b).

Figure 8: Procedure of set difference for the salient co-occurrence network of "*ume*" (梅, en. plum) poems and their translations: Co-occurrence patterns are connected by edges; nodes represent words; Square nodes indicate words that exist only in CT; elliptical nodes indicate poetic words.

	cw	ctf	t_1	$tf(t_1)$	$idf(t_1)$	t_2	$tf(t_2)$	$idf(t_2)$
1	21.15	5	Kurabu.PN	8	9.21	moonless night	7	7.13
2	19.83	8	Kurabu.PN	8	9.21	go over	10	4.50
3	19.83	7	Kurabu.PN	8	9.21	dark	10	4.92
4	19.57	7	Kurabu.PN	8	9.21	clearly	10	4.79
5	19.37	40	plum	198	3.71	fragrance.2	42	4.59
6	18.12	18	lingering fragrance	19	5.84	plum	198	3.71
7	18.01	8	plum	198	3.71	Kurabu.PN	8	9.21
8	17.97	9	lingering fragrance	19	5.84	stop by	10	5.40
9	17.50	6	cling	7	6.72	lingering fragrance	19	5.84
10	17.21	7	lingering fragrance	19	5.84	blame	7	5.84
11	17.07	5	Kurabu.PN	8	9.21	around	5	4.65
12	16.90	16	which	16	5.40	plum	198	3.71
13	16.83	6	stop by	10	5.40	cling	7	6.72
14	16.82	17	plum	198	3.71	distinction	17	5.18
15	16.80	4	smell.vt.1	14	5.38	Kurabu.PN	8	9.21
16	16.76	7	look away	8	7.01	get dark	10	4.62
17	16.55	7	blame	7	5.84	stop by	10	5.40
18	16.33	7	a bit	9	5.26	lingering fragrance	19	5.84
19	16.32	11	discern	11	6.21	plum	198	3.71
20	16.30	8	distinction	17	5.18	which	16	5.40
21	16.26	5	lingering fragrance	19	5.84	degree	5	6.64
22	15.95	4	cling	7	6.72	degree	5	6.64
23	15.91	7	look away	8	7.01	daybreak	9	4.16
24	15.71	7	a bit	9	5.26	stop by	10	5.40
25	15.71	8	plum	198	3.71	look away	8	7.01
26	15.69	6	sew.2	6	6.72	hide.vi.2	14	4.70
27	15.64	5	degree	5	6.64	stop by	10	5.40
28	15.63	10	distinction	17	5.18	break off	33	4.32
29	15.63	7	hide.vi.2	14	4.70	woven hat	10	5.99
30	15.62	10	woven hat	10	5.99	warbler	40	3.73
31	15.61	17	plum	198	3.71	garden	17	4.46
32	15.58	10	plum	198	3.71	woven hat	10	5.99
33	15.57	17	distinction	17	5.18	snow	40	3.18
34	15.54	13	break off.1	14	5.12	plum	198	3.71
35	15.52	8	Kurabu.PN	8	9.21	mountain	8	2.76
36	15.52	5	cling	7	6.72	a bit	9	5.26

Table 4: Set difference of co-occurrence patterns for "plum": Thresholds for poems and translations are 5.6 and 15.5 respectively; 30 nodes (14 poetic words 14 and 16 translation words) and 36 edges in total.

	Implementation A	Implementation B
analysis unit	word	co-occurrence pattern
breakdown visualization	support	not support
equivalence judgment	based on word alignment	based on semantic category code
filtering function word	based on part-of-speech	based on threshold of keyness
requirement of semantic annotation	not require	require
set difference level	parallel text level	corpus level

Table 5: Differences between Implementation A and Implementation B.

3.4 Summary

Implementation A and B are two different strategies to identify non-literal elements based on the same perspective, Schramm's communication model. We summarize the differences in Table 5.

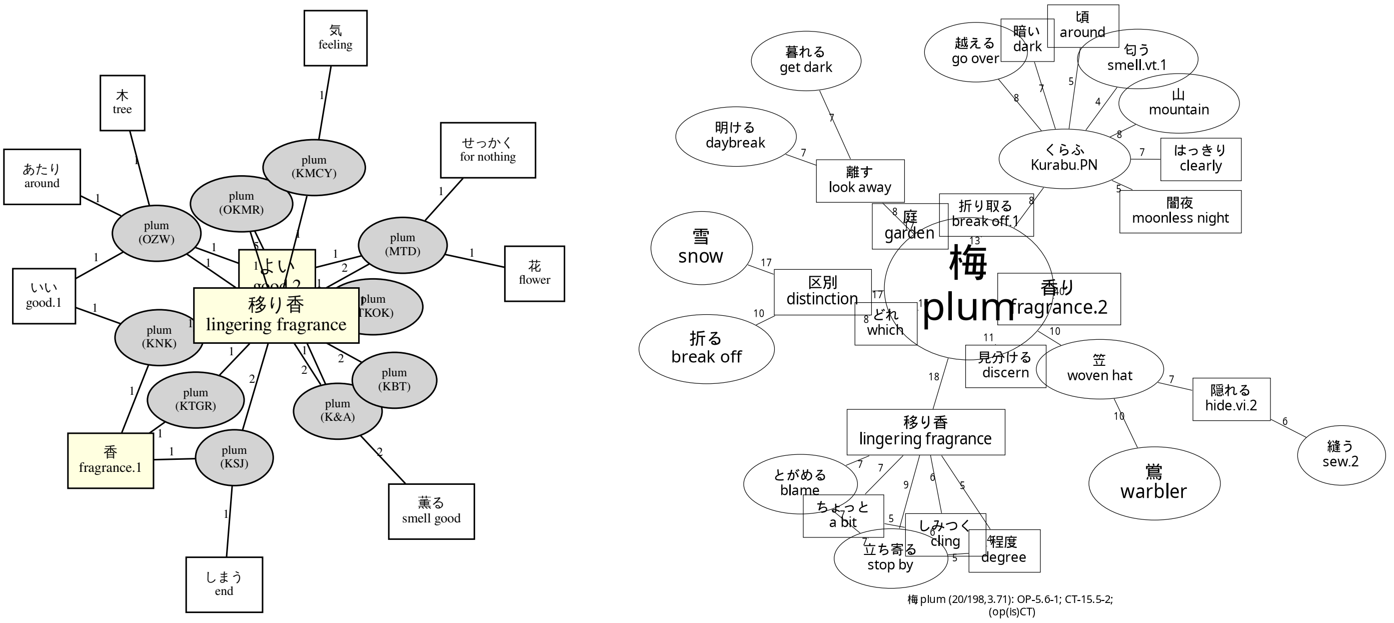
Implementation A uses words as the unit while Implementation B uses co-occurrence patterns. To identify intersections of OP and CT, Implementation A uses potentially proper word alignment and Implementation B uses semantic category code to determine the semantic equivalence between old and contemporary Japanese. Therefore, Implementation A does not require a prior language-specific semantic annotation but Implementation B requires. To filter function words from non-literal elements, Implementation A uses part-of-speech while Implementation B uses a flexible keyness threshold. The two set difference-based implementations are also different in the level of set difference. That is, Implementation A firstly calculates the set difference at the parallel text level and then aggregates the set difference as a whole; Implementation B, on the other hand, calculates the set difference at the corpus level and directly returns the aggregate set difference. Hence, implementation B cannot provide breakdown visualizations as Implementation A.

4. Results

We applied the two implementations to the six most frequently used flora poetic words in the *Kokinshū*: *"ume"* (梅, en. plum), *"ominaeshi"* (女郎花, en. golden valerian), *"kiku"* (菊, en. chrysanthemum), *"sakura"* (桜, en. cherry), *"matsu"* (松, en. pine), and *"yamabuki"* (山吹, en. kerria). This section shows the visualizations of the two implementations. Figure 9 to Figure 14 display the visualizations for each of the flora.[14]

In this study, we did not develop the methods for a specific theoretical assumption. This section, hence, only includes the results of illustrations. However, the examples presented are not selective ones, but rather based on an objective frequency of entities typically observed in classical Japanese poetry. In comparison to the dictionary descriptions, we do not selectively focus on specific points aligned with our visualizations, either. Instead, we showcase consistencies and inconsistencies between the visualization and the descriptions in the dictionary for poetic vocabulary (Katagiri 1983) to demonstrate the strengths and weaknesses of the methods.

[14] For visual consistency with Implementation B and improved readability, we used `dot` to generate graphs for Implementation A, similar to Implementation B, instead of the original visualization using the dashboard.



(a) Visualization A: grey = queried word, yellow = common misalignment.

(b) Visualization B: cw threshold = 5.6, 15.5 for poem/translation.

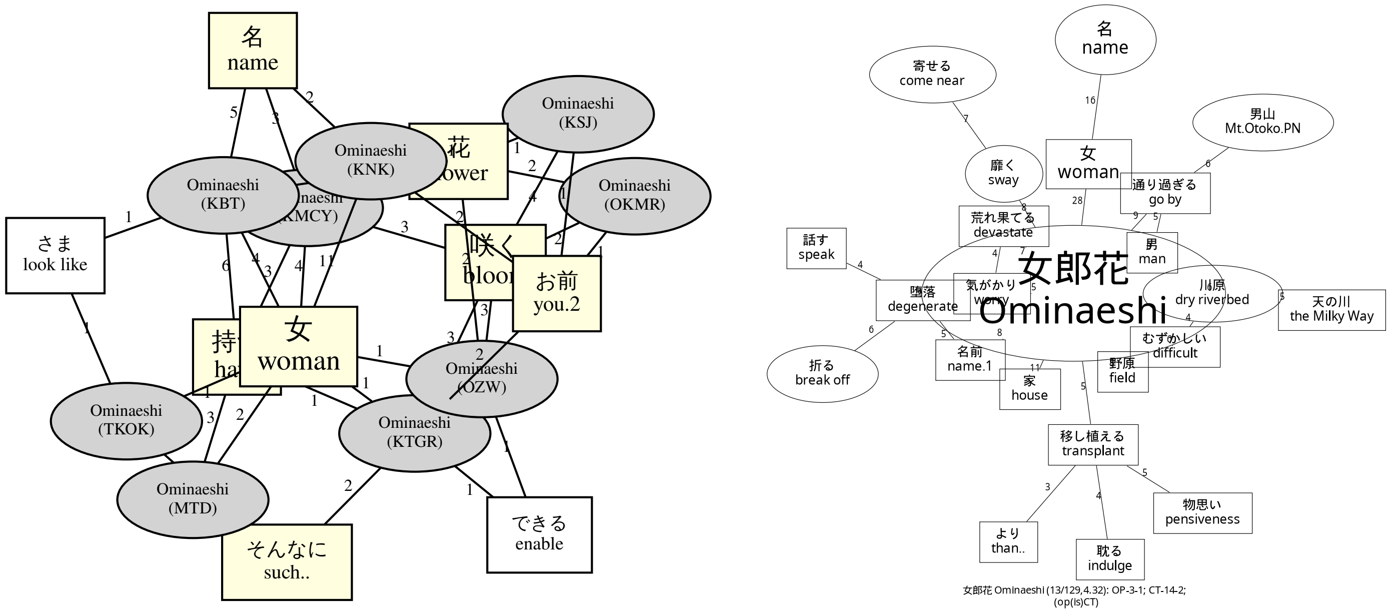
Figure 9: Network visualizations of "*ume*" (梅, en. plum): Square nodes indicate additions in translations; elliptical nodes indicate poetic words.

4.1 Case-Specific Observations Compared with the Dictionary

***"Ume"* (en. plum; Figure 9).** The translation word "fragrance" (ja. 香/移り香/薫る) was near the hub of the "plum" network generated by Implementation A as well as Implementation B. According to Katagiri (1983, 82–83), praising the fragrance of "plum" became common in classical Japanese poetry after the *Kokinshū*. The results agreed with the core description of the poetic word "plum". Moreover, Implementation B presented a potentially significant node, "woven hat", which is not directly mentioned in the "plum" entry of the dictionary. The dictionary contains an entry *"ume no hanagasa"* (梅の花笠, en. woven hat of plum blossoms), which often suggests a woven hat stitched by a warbler (Katagiri 1983, 83). Given the connection between "plum" and "woven hat"; "plum" can also retain a link with "warbler", a frequently utilized pair in classical Japanese poetry and Japanese visual arts. Besides, an explanatory addition "distinction" ties "plum" with "snow" indirectly, from which we might deduce that the cluster portrays a situation where it is difficult to distinguish between "snow" and "plum". Similarly, in Figure 9b, we can deduce from which classical Japanese poem each cluster of the network originates. For instance, the cluster "snow-break off-distinction-plum-which" could stem from No. 337[15] in the *Kokinshū*, and the cluster "plum-woven hat-warbler-hide-sew" could stem from the No. 36.[16] Conversely, fig. 9a fails to replicate the context of the original poem, whereas it can display the core explanatory addition ("fragrance") as a hub for various translation versions more clearly.

[15] "[W]hen snow has fallen/flowers appear on all the/trees clusters of white blooms/from which of them/can I pluck the fragrant plum blossoms" (translated by Rodd et al. 1996, 143).

[16] "[T]hey say the thrush weaves/a rain hat of flowering/plum perhaps I too/may pluck a spray and make a/garland to conceal my age" (translated by Rodd et al. 1996, 59).



(a) Visualization A: grey = queried word, yellow = common misalignment.

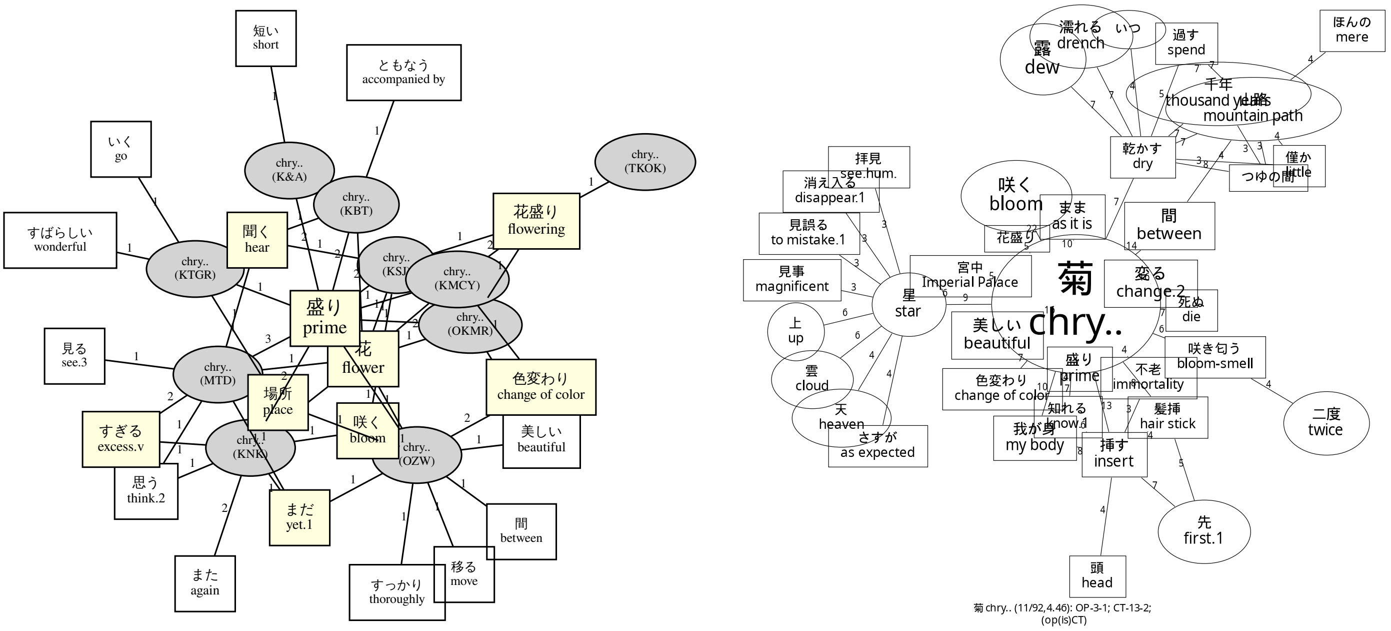
(b) Visualization B: *cw* threshold = 3, 14 for poem/translation.

Figure 10: Network visualizations of *"ominaeshi"* (女郎花, en. golden valerian): Square nodes indicate additions in translations; elliptical nodes indicate poetic words.

"*Ominaeshi*" (en. golden valerian; Figure 10). The explanatory "woman" was the hub in the "golden valerian" network formed by Implementation A (Figure 10a). The visualization showed that the flower holds a feminine image due to its "name" (ja. 名) of "woman" (ja. 女/をみな). This outcome aligned with Katagiri (1983, 481–482): During the Heian period (794–1185), the majority of poems concerning golden valerian depicted the flower as "woman". Figure 10b also included the connotative term "woman" although "woman" was not in a central position in the network. Besides, the hypernym "flower" is frequently added in translations (Figure 10a) and Figure 10b visualized the antonym terms "male" and the "Mountain of Man" (Mt. Otoko, ja. 男山), which are also aspects of connotation in some theories (Eco 1976). Moreover, each cluster in Figure 10b can reflect part of the narratives of corresponding poems, poems No. 226, No. 227, and No. 230, respectively.

"*Kiku*" (chrysanthemum; Figure 11). Both visualizations visualize the connotative term "color change" (ja. 色変わり/移る/変わる), which was described in Katagiri (1983, 127–129): During the Heian period, people admired the gradual change in color of white chrysanthemums to red due to the cold weather. According to Katagiri (1983, 128), after the revival of the Chongyang Festival[17] in the fifth year of Emperor Saga's reign (814), "chrysanthemum" came to be used in many classical Japanese poems. Chrysanthemum was an essential part of the Japanese Chongyang Festival because it was believed to have the effect of prolonging one's life. Therefore, "chrysanthemum" was often used in classical Japanese poems to convey the sense of longevity. As for the connotative words regarding longevity, Figure 11a only presented "peak (of blossom season)" (ja. 盛り), which is a common addition at the hub; Figure 11b presented "one thousand years" (ja. 千年), "never get old" (ja. 不老), and "peak (of blossom season)" (ja. 盛り/花盛り)

[17] In Japan, the festival is also known as the Chrysanthemum Festival. According to Katagiri (1983, 128), the chrysanthemum was introduced to Japan along with the Chinese Chongyang Festival.



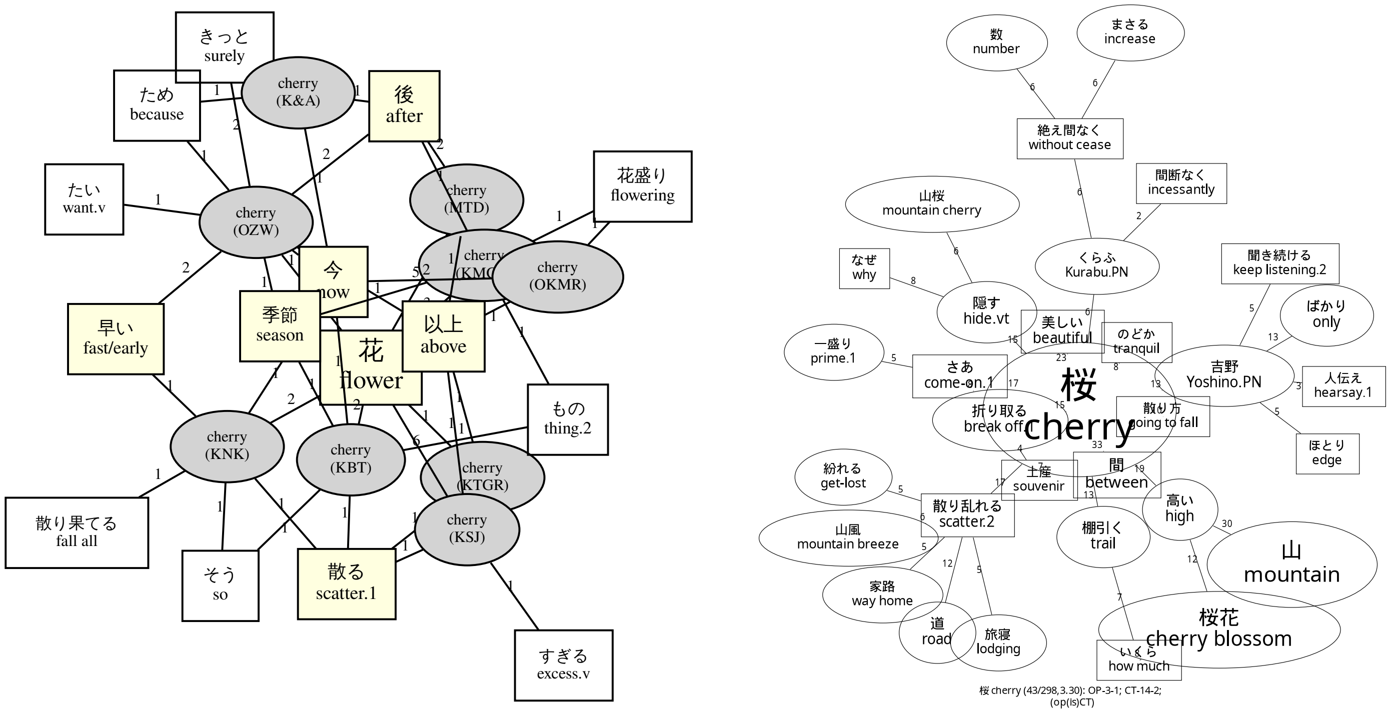
(a) Visualization A: grey = queried word, yellow = common misalignment.

(b) Visualization B: *cw* threshold = 3, 13 for poem/translation.

Figure 11: Network visualizations of “*kiku*” (菊, en. chrysanthemum): Square nodes indicate additions in translations; elliptical nodes indicate poetic words.

while these words are not at an important position in the network (Figure 11b). Besides, Figure 11b presented “hear” (parallel image for “chrysanthemum” as *kakekotoba*) as a common addition. Nevertheless, the two methods failed to visualize the relationship between the poetic word “chrysanthemum” and the Chongyang Festival, which should be essential background knowledge unfamiliar to contemporary Japanese novice readers. This was because the Chongyang Festival not appearing in the translation text.

“*Sakura*” (en. cherry; Figure 12). The dictionary (Katagiri 1983, 172–173) includes *“sakura”* as a compound item *“sakura-bana”* (桜花, en. cherry blossom). According to Katagiri (1983), most cherry blossom poems in the *Kokinshū* are about falling/scattering (ja. 散る) cherry blossoms which are often related to the transience of human life. We could observe connotative words regarding the sense of falling, such as “to be or to lay scattered about” (ja. 散り乱れる), in Figure 12a. On the other hand, Figure 12a indicated that four translators incorporated “fall” (four added general “fall”; one added 散り果てる, en. all falling) into their translations of “cherry” poems. Moreover, the terms “season”, “now”, and “later” (ja. 以上 and 後) were common additions in Figure 12a, which might suggest that “cherry” as a flower could have a strong relationship with time, evoking a mood of lamenting life’s impermanence and the passing of spring (Katagiri 1983, 173). Furthermore, Figure 12b illustrates the robust connection between “cherry” and “mountain” (inclusive of “mountain breeze”, “Yoshino”, “mountain cherry”). According to Katagiri (1983, 456), “cherry” and “Mt. Yoshino” have maintained a significant relationship since the *Shin Kokinshū*, the eighth anthology of the *Hachidaishū*, while instances of the duo are scarce in the *Kokinshū*. From the non-literal associations visualized in Figure 12b, we discern that “mountain” is a vital context for cherry blossoms. This could provide a foundation for the forthcoming shifts in the usage of “cherry”.



(a) Visualization A: grey = queried word, yellow = common misalignment.

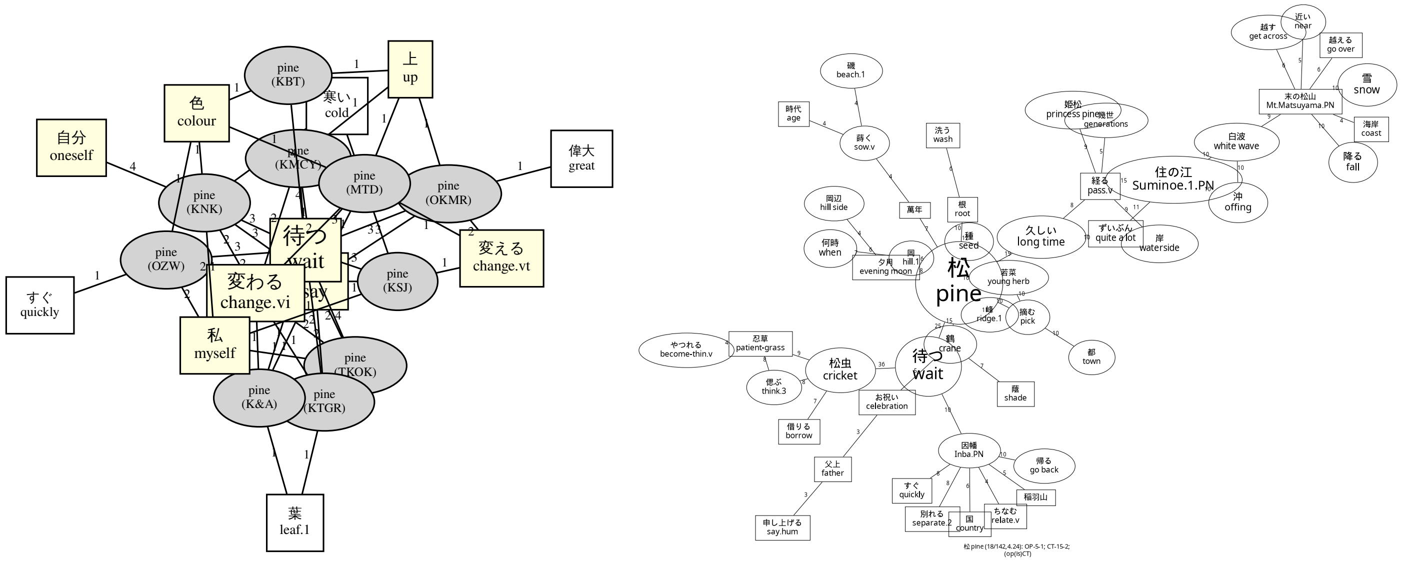
(b) Visualization B: *cw* threshold = 3, 14 for poem/translation.

Figure 12: Network visualizations of "*sakura*" (桜, en. cherry): Square nodes indicate additions in translations; elliptical nodes indicate poetic words.

***"Matsu"* (en. pine; Figure 13).** The most salient node in the two networks is "to wait". The results were consistent with Katagiri (1983, 383–384): The word "pine" is a *kakekotoba* that holds another meaning of "to wait" (ja. 待つ, the same kana character with "pine" in Japanese), which implies "long-awaited", to celebrate the eternal nature of the pine tree. Poets associated it with the word "one thousand years old" (ja. 千歳) or with "crane" (ja. 鶴) and "wisteria" (ja. 藤), which also symbolize the one-thousand-year longevity (Katagiri 1983, 383–384). We identified these terms in Figure 13b. On the other hand, Figure 13a displayed "change" and "color", as frequently added non-literal elements, from which we can deduce that translators aim to highlight the longevity of pine because its "color" never "changes". Nevertheless, both visualizations failed to visualize many descriptions in the dictionary. For example, poets use the poetic word pine as "evergreen pine" (ja. 常盤の松) or "pine green" (ja. 松の緑). "Pine" is also known for "*kadomatsu* (gate pine)" (ja. 門松),[18] where the gods resided, and "pine wind" (ja. 松風), the wind that blows through the treetops of pine trees (Katagiri 1983, 384). They are rarely explained in translations.

***"Yamabuki"* (en. kerria; Figure 14).** Figure 14b showed that "kerria" was "reflected" (ja. 映る) by the "river" or "underwater" (ja. 水底/河底). The results agreed with the comment in Katagiri (1983, 439–440): Since "kerria" often blooms near water, poets usually use "kerria" with the verb "to be reflected/to fade" (ja. 移ろう/映ろう). Conversely, Katagiri (1983, 439–440) notes that since the Heian period, "kerria" had become an important symbol of the end of "spring", and "kerria" often occurs with the nouns "well (noun)" and "frog". From these perspectives, neither Implementation A nor Implemen-

[18] *"Kadomatsu"* are traditional Japanese decorations made for welcoming ancestral spirits of the harvest in the New Year.



(a) Visualization A: grey = queried word, yellow = common misalignment.

(b) Visualization B: cw threshold = 5, 15 for poem/translation.

Figure 13: Network visualizations of "*matsu*" (松, en. pine): Square nodes indicate additions in translations; elliptical nodes indicate poetic words.

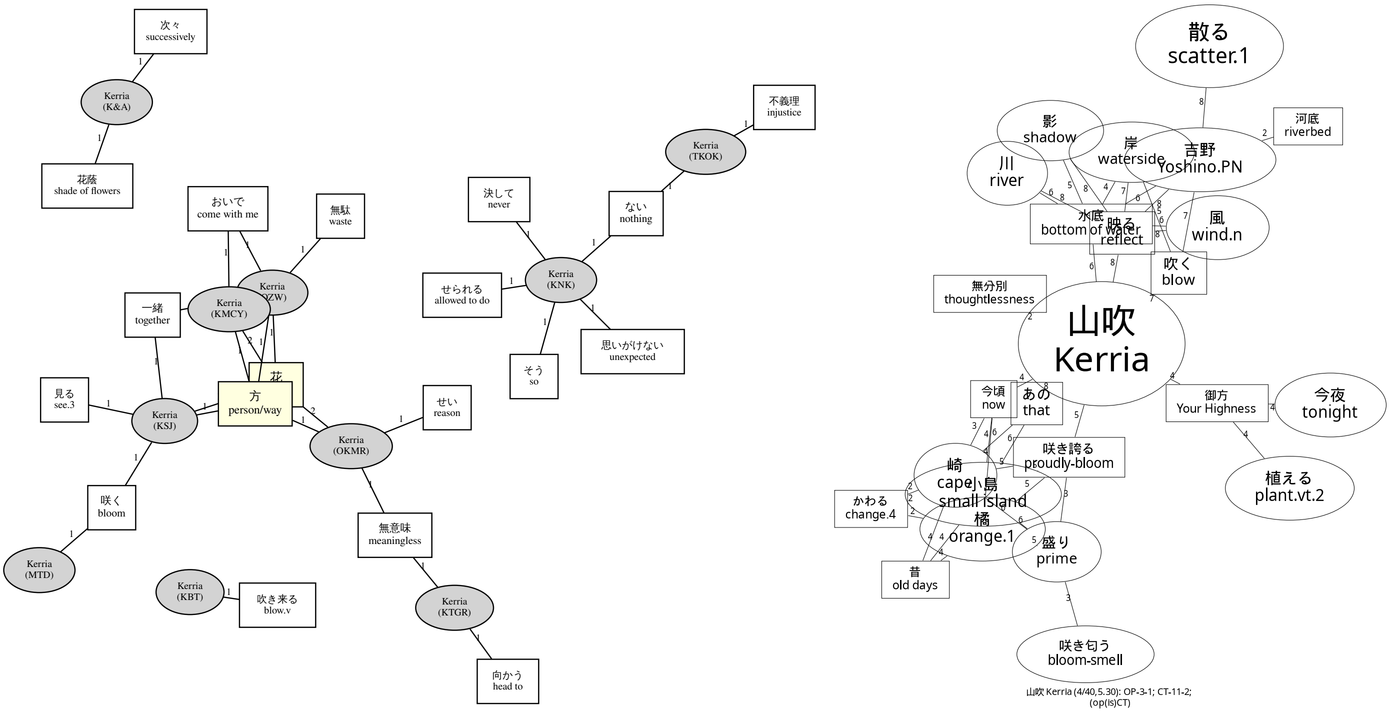
tation B could visualize such connotative information. Such connotative information is typically represented by phraseological patterns that appear in the original texts and are translated as is, without any additional explanation and context in translations.

4.2 Summary

Based on the case study of poetic flora words, we can summarize that for each flora word, non-literal elements visualized by the two set difference-based Implementations could serve as hints to the connotation of classical poetic Japanese vocabulary.

Implementation A could divide the additions into common ones (hubs) and individual ones (peripheries in networks), and hence visualize essential explanatory additions to reflect the translators' common understanding of the poets' field of experience; Implementation B could visualize the main contexts in original poems linked by explanatory additions in translations for the flora words. From the visualization by Implementation B, we could infer part of the narratives stemming from the original poems. The two methods, especially Implementation B, can cover most of the descriptions regarding connotation in the classical poetic Japanese dictionary (Katagiri 1983), while the non-literal visualization strategy could not visualize two types of connotation: (a) explicit phraseological patterns in the original poem, for which minimal explanation was not added in translations; (b) encyclopedic knowledge that was beyond the texts in translations.

Implementation A occasionally visualized domain-general knowledge as a hub, which was often the hypernym of queried words, such as "flower" for "cherry". Since such words were not function words, we could not exclude them. Although the hypernym and hyponym for the queried words are viewed as connotations in some definitions, for our objectives, hypernyms and hyponyms are not the ideal connotation candidates. This is because these semantic relations are domain-general rather than domain-specific knowledge only for poetic words.



(a) Visualization A: grey = queried word, yellow = common misalignment.

(b) Visualization B: cw threshold = 3, 11 for poem/translation.

Figure 14: Network visualizations of *"yamabuki"* (山吹, en. kerria): Square nodes indicate additions in translations; elliptical nodes indicate poetic words.

5. Discussion

This section discusses the implication of the project of connotation visualization for classical poetic Japanese vocabulary, including a summary of the differences between the results of Implementations A and B, as well as the contributions and limitations.

5.1 Characteristics of the Visualizations

Both Implementations A and B are based on Schramm's communication model and visualize non-literal elements in complementary sets of original poems and translations. However, they take different strategies to perform set difference calculations and hence have different presentations of connotations.

Implementation A visualizes non-literal elements reflecting rhetorical techniques, hypernyms, and core associations; Implementation B visualizes non-literal elements reflecting rhetorical techniques, hyponyms, antonyms, and a wide array of associations (e.g., phraseological pattern, contextual information). The pragmatic aspects (e.g., translators' value judgment) of connotation were included in the visualizations; whereas, both Implementation A and B cannot visualize sociolinguistic aspects (connotative terms reflecting gender, style, social class, and region of the poets) and emotional aspects (positive/negative/neutral) of connotation. These aspects are also less mentioned in Katagiri's dictionary, which might be left for further studies using approaches from corpus-based variationist linguistics and stylistics to explore.

For the objective of supplementing the descriptions for poetic language dictionaries, Implementation B could help incorporate minimal contextual information from original poems through clusters in the co-occurrence networks. On the other hand, Implementation A could demonstrate the consistent use of explanatory additions among experts to reflect core connotations. In contrast, a dictionary with selective consciousness may occasionally overlook some dynamic connotations visualized by the methods.

5.2 Contributions

The two visualizations demonstrated how to utilize explanatory additions in the translations of literary texts to visualize the inaccessible part of the connotation of historical, literary languages. We showed practically that not only dictionaries but also translations are feasible resources as a medium to access connotation in a historical, literary language. Employing traditional statistical natural language processing algorithms and information theory-based methods enhanced our interpretation when the historical literary text data is a low resource.

Although our operationalization of connotation is only an imperfect simplification, it demonstrates that removing all the semantically equivalent elements between the parallel texts and visualizing the leftover additional information is a transparent way of approaching the connotation.

The misalignment-based implementation, which does not use language-specific semantic category annotation, can be applied not only to contemporary Japanese translations of classical Japanese poetry but also to translations in other languages. On the other hand, the co-occurrence-based implementation can cover a large part of the connotation described in the dictionary and also provide essential contexts stemming from original poems.

The two implementations may also provide some theoretical considerations as follows.

Firstly, connotation could be considered a relative concept in practice. The misalignment-based Implementation A yields only a minimal number of non-literal elements related to connotation, suggesting that many connotations may still be accessible to contemporary Japanese people, such as domain-general knowledge. Consequently, translators may choose not to add words for this connotation to translations. In other words, the connotation found at the intersection of contemporary and ancient Japanese remains unseen in the non-literal set defined by Implementation A. However, we could regard this imperfect visualization as a dynamic feature of Implementation A, where connotation is seen as a relative concept. The connotative information in one culture is relative to other cultures. Absolute connotation is considered an open set so far; while we do not have to visualize all the elements of the connotation even when the elements are general among different cultures. What kind of information is the essential connotation should be answered relatively through a specific comparison between the source culture and the target culture.

Secondly, most of the lexical connotations lie in the interrelation of words. When there is no direct explanatory addition for a queried poetic word, Implementation A might not provide a visualization for the word (which does not mean the word has no connotation). In contrast, the co-occurrence patterns in Implementation B can help visualize the unseen elements in Implementation A. The explanatory addition may be not directly for the queried word itself, but for the co-occurrence relation that the queried word holds. This reflects that connotation exists not only in the poetic word itself but is also implied in the interrelation between poetic words.

5.3 Limitations

The operationalization and the two implementations leave the following problems:

Firstly, encyclopedic knowledge and sociocultural context could not be extracted. Our methods could never visualize complete narratives and sociocultural contexts beyond texts. For example, both methods could not visualize the connection between the poetic word "chrysanthemum" and the Chongyang Festival, which could be common tacit knowledge among poets during the Heian period. If such knowledge and sociocultural context do not appear in the translation and original texts, we are unable to visualize them.

Secondly, the presumption that novice readers share no field of experience with poets is extreme. In reality, domain-general connotations are shared between ancient Japanese poets and contemporary novice Japanese readers. Therefore, based on our operationalization, we could not visualize non-literal elements regarding these parts of connotation.

Thirdly, the study did not provide a quantitative evaluation of the implementations. We assumed a stance where we had no knowledge about the subject of study as everything was observed only through visualizations. We did not presume any specific hypotheses to be answered from the methods. This stance makes the methods difficult to evaluate. In future works, we should apply the methods to tackle particular issues. For instance, we can scrutinize philological annotations for specific hypotheses of poetic words and subsequently assess them. Although we do not intend to artificially impose arbitrary limits on the elements that do or do not fit into connotation, we may explore various ways of extracting connotations from other sources, namely, automatic extraction from dictionaries, and compare them with the results of the current methods.

Fourthly, the algorithms used in the two implementations are open to improvement. We used IBM model 2 in the first implementation and *cw* value in the second implementation. However, the two algorithms were not systematically evaluated and were not the only options. We need to provide additional algorithm options in a suite of the workflow and compare among different algorithms in the future.

Despite the limitations, the methods is an attempt to reflect on the fundamental challenges encountered in the studies of meaning in classical Japanese literary languages and how to deal with them in a principled way. The methods themselves are not the ultimate solution to the challenges, but they can be a starting point for further studies.

6. Conclusion

To offer an independent lexical connotation visualization tool as a supplement to dictionaries of classical poetic Japanese, this paper aims to visualize connotation in the classical poetic Japanese vocabulary in the *Kokinshū* by presenting non-literal elements of each word. The non-literal elements were extracted by calculating the set difference between the *Kokinshū* and its ten translation versions, inspired by Schramm's (1954) communication model.

This paper outlines the motivation for using non-literal elements revealed by translations as a projection of connotation and discusses the relationship between connotation and non-literal elements, as well as the relationship between the dictionary and translations. We argued that although non-literal elements are not equivalent to connotation, when holding an explanatory addition in translations, non-literal elements could reflect part of connotation. Moreover, an explanatory addition of connotation in translation could illuminate the unconscious part of connotation, which is rarely covered by a dictionary.

From a technical standpoint, the paper presents two different implementations for visualizing non-literal elements based on Schramm's model: One is a word-misalignment-based visualization; the other is a co-occurrence pattern-based visualization. We applied these two visualizations to the six most frequent floral words in the *Kokinshū*. As a result, word misalignment-based visualization could visualize the common explanatory addition by different translators, reflecting the most robust and essential perception of poetic floral words among the translators; co-occurrence pattern-based visualization covered a wide range of connotation descriptions in the poetic Japanese dictionary (Katagiri 1983). For our purpose of supplementing a dictionary, co-occurrence pattern-based implementation provided a more comprehensive visualization. On the other hand, misalignment-based visualization could show the common part and the variation among different translation versions, which was limited while more detailed.

In the future, we will incorporate optional algorithms into the workflow and apply the methodology to more specific theoretical assumptions to contribute to the studies of Japanese poetry.

7. Data Availability

The *Kokinshū* Data can be found on Zenodo: `10.5281/zenodo.4735848`; the translation data is protected by copyright.

8. Software Availability

Scripts for replicating Implementation A can be found on Github `https://github.com/nehcx/kokinMisalign`; a demonstration of Implementation B can be found on the authors' website `https://cuckoo.js.ila.titech.ac.jp/~yamagen/waka/poem.cgi`.

9. Acknowledgements

This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2112, and JSPS KAKENHI Grant Number JP18K00528, JP23KJ0910, JP23K00545.

10. Author Contributions

Xudong Chen: Conceptualization, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing

Yamamoto Hilofumi: Conceptualization, Project administration, Investigation, Supervision, Resources, Methodology, Software, Writing – original draft, Writing – review & editing

Hodošček Bor: Conceptualization, Project administration, Resources, Data curation, Software, Writing – original draft, Writing – review & editing

References

Allaway, Emily and Kathleen McKeown (2021). "A Unified Feature Representation for Lexical Connotations". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2145–2163. 10.18653/v1/2021.eacl-main.184.

Asahara, Masayuki and Yuji Matsumoto (2000). "Extended Models and Tools for High-Performance Part-of-Speech Tagger". In: *Proceedings of the 18th Conference on Computational Linguistics*. Vol. 1. Association for Computational Linguistics, 21–27. 10.3115/990820.990824.

Bloomfield, Leonard (1933). *Language*. H. Holt and Company.

Brown, Peter E., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer (1993). "The Mathematics of Statistical Machine Translation: Parameter Estimation". In: *Computational Linguistics* 19 (2), 263–311.

Burrows, John (2007). "All the Way Through: Testing for Authorship in Different Frequency Strata". In: *Literary and Linguistic Computing* 22 (1), 27–47. 10.1093/llc/fqi067.

Chandler, Daniel (2002). *Semiotics: The Basics*. Routledge. 10.4324/9780203166277.

Chen, Xudong, Bor Hodošček, and Hilofumi Yamamoto (2022). "Tango Araimento no Ayamari Taioo wo motiita Utakotoba no Konoteeshon Kenshutsu/Connotation Detection for Classical Poetic Japanese Vocabulary Using Word Alignment Mismatch[単語アライメントの誤り対応を用いた歌ことばのコノテーション検出]". In: *Proceedings of Symposium for Humanities and Computer 2022* [**人文科学とコンピュータシンポジウム** 2022 **論文集**] 2022 (1), 111–118.

Du, Keli, Julia Dudar, and Christof Schöch (2022). "Evaluation of Measures of Distinctiveness: Classification of Literary Texts on the Basis of Distinctive Words". In: *Journal of Computational Literary Studies* 1 (1). 10.48694/JCLS.102.

Dunning, Ted (1993). "Accurate Methods for the Statistics of Surprise and Coincidence". In: *Computational Linguistics* 19 (1), 61–74. https://aclanthology.org/J93-1003 (visited on 11/29/2023).

Eco, Umberto (1976). *A Theory of Semiotics*. Indiana University Press.

Hall, Stuart (2018). "8. Encoding and Decoding in the Television Discourse [originally 1973; republished 2007]". In: *Essential Essays, Volume 1. Foundations of Cultural Studies*. Ed. by David Morley, 257–276. 10.1515/9781478002413-014.

Hjelmslev, Louis (1969). *Prolegomena to a Theory of Language*. Rev. Engl. ed., reprinted. Univ. of Wisconsin Pr.

Hodošček, Bor and Hilofumi Yamamoto (2022). "Development of Datasets of the Hachidaishū and Tools for the Understanding of the Characteristics and Historical Evolution of Classical Japanese Poetic Vocabulary". In: *Digital Humanities 2022 Conference Abstracts*. The University of Tokyo, 647–648.

Kalouli, Aikaterini-Lida, Rebecca Kehlbeck, Rita Sevastjanova, Katharina Kaiser, Georg A. Kaiser, and Miriam Butt (2019). "ParHistVis: Visualization of Parallel Multilingual Historical Data". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Association for Computational Linguistics, 109–114. 10/gh547n.

Kaneko, Motoomi (1933). *Kokinwakashu Hyoshaku: Showa Shimban/An Annotated Kokinwakashu: The New Showa Edition* [古今和歌集評釈: 昭和新版]. Meijishoin.

Katagiri, Yoichi (1983). *Utamakura utakotoba jiten zoutei ban/Dictionary of Poetic vVocabulary Additional Version*[歌枕歌ことば辞典増訂版]. Kadokawa Shoten.

— (1998a). *Kokinwakashu zen hyoshaku/A Complete Annotated Edition of Kokinwakashu* [古今和歌集全評釈]. Vol. 1. Kodansha.

— (1998b). *Kokinwakashu zen hyoshaku/A Complete Annotated Edition of Kokinwakashu* [古今和歌集全評釈]. Vol. 2. Kodansha.

— (1998c). *Kokinwakashu zen hyoshaku/A Complete Annotated Edition of Kokinwakashu* [古今和歌集全評釈]. Vol. 3. Kodansha.

Koehn, Philipp (2010). *Statistical Machine Translation*. 1st ed. Cambridge University Press.

Kojima, Noriyuki and Eizo Arai (1989). *Kokinwakashu* [古今和歌集]. Iwanamishoten.

Komachiya, Teruhiko (1982). *Kokinwakashu:Gendaigo Yaku Taisho/Kokinwakashu: With Modern Japanese Translations* [古今和歌集：現代語訳対照]. Obunsha Bunko/Obunsha Series [旺文社文庫]. Obunsha.

Kubota, Utsubo (1960a). *Kokinwakashu Hyoshaku/An Annotated Kokinwakashu* [古今和歌集評釈]. Vol. 1. Tokyodo.

— (1960b). *Kokinwakashu Hyoshaku/An Annotated Kokinwakashu* [古今和歌集評釈]. Vol. 2. Tokyodo.

— (1960c). *Kokinwakashu Hyoshaku/An Annotated Kokinwakashu* [古今和歌集評釈]. Vol. 3. Tokyodo.

— (1994). "Kago no hensen/Transition in Poetic Vocabulary [歌語の変遷]". In: *Gekkan Gengo/Language monthly*[月刊言語] 267, 58–65.

Kudo, Taku, Kaoru Yamamoto, and Yuji Matsumoto (2004). "Applying Conditional Random Fields to Japanese Morphological Analysis". In: *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 230–237. https://aclanthology.org/W04-3230 (visited on 12/12/2023).

Kyusojin, Hitaku (1979). *Kokinwakashu zen chushaku/Comprehensive Annotations of Kokinwashu* [古今和歌集 全注釈]. Vol. 1. Kodansha gakujutsu bunko/Kodansha Academic Collection of Japanese Literature [講談社学術文庫]. Kodansha.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schutze (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Matsuda, Takeo (1968a). *Shinshaku Kokinwakashu Hyoshaku: /A New Annotated Edition of Kokinwakashu* [新釈古今和歌集]. Vol. 2. Kazamashobo.

— (1968b). *Shinshaku Kokinwakashu Hyoshaku: /A New Annotated Kokinwakashu* [新釈古今和歌集]. Vol. 1. Kazamashobo.

Matsumoto, Yuji, Akira Kitauchi, Tatsuo Yamashita, Osamu Imaichi, and Tomoaki Imamura (2002). *Morphological Analysis System ChaSen Version 2.2.9 Manual*. https://users.monash.edu/~jwb/chasen-2.2.9.pdf (visited on 12/12/2023).

Mounin, Georges (1976). *Les problèmes théoriques de la traduction*. Gallimard.

Nakano, Hiroshi, Ooki Hayashi, Hisao Isii, Makoto Yamazaki, Masahiko Ishii, Yasuhiko Kato, Tatuo Miyazima, and Akio Tsuruoka (1994). *Bunrui goi hyo furoppi ban/Word List by Semantic Principles, floppy disk version* [**分類語彙表 フロッピー版**]. Vol. 5. okuritsu Kokugo Kenkyujo Gengoshori datashu/National Language Research Institute Language Resource [国立国語研究所言語処理データ集]. Dainippon Toten.

Okumura, Tsuneya (1978). *Kokinwakashu* [**古今和歌集**]. Shincho nippon koten shusei/Shincho Collection of Classical Japanese Literature [新潮日本古典集成]. Shinchosha.

Ozawa, Masao (1971). *Kokinwakashu* [**古今和歌集**]. Thirteenth. Nihon koten bungaku zenshu/The Complete Series of Classical Japanese Literature [日本古典文学全集]. Shogakukan.

Paquot, Magali and Yves Bestgen (2009). "Distinctive Words in Academic Writing: A Comparison of Three Statistical Tests for Keyword Extraction". In: *Corpora: Pragmatics and Discourse*. Ed. by Andreas H. Jucker, Daniel Schreier, and Marianne Hundt. Rodopi, 247–269.

Rashkin, Hannah, Sameer Singh, and Yejin Choi (2016). "Connotation Frames: A Data-Driven Investigation". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 311–321. 10.18653/v1/P16-1030.

Rodd, Laurel Rasplica, Mary Catherine Henkenius, and Tsurayuki Ki, eds. (1996). *Kokinshū: A Collection of Poems Ancient and Modern*. 1st pbk. ed. C&T Asian Literature Series. Cheng & Tsui Co.

Rössler, Gerda (1979). *Konnotationen: Untersuchungen. Zum Problem der Mit- und Nebenbedeutung*. Steiner.

Schöch, Christof, Daniel Schlör, Albin Zehe, Henning Gebhard, Martin Becker, and Andreas Hotho (2018). "Burrows' Zeta: Exploring and Evaluating Variants and Parameters". In: *Book of Abstracts of the Digital Humanities Conference*. the Digital Humanities Conference. ADHO. https://dh2018.adho.org/burrows-zeta-exploring-and-evaluating-variants-and-parameters/ (visited on 11/29/2023).

Schramm, Wilbur Lang (1954). *The Process and Effects of Mass Communication*. University of Illinois Press.

Spärck Jones, Karen (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". In: *Journal of Documentation* 28 (1), 11–21. 10.1108/eb026526.

Stede, Manfred (1999). *Lexical Semantics and Knowledge Representation in Multilingual Text Generation*. Springer US. 10.1007/978-1-4615-5179-9.

Stubbs, Michael (2002). *Words and Phrases: Corpus Studies of Lexical Semantics*. Blackwell Publishers.

Takeoka, Masao (1976a). *Kokinwakashu zen hyoshaku: Kochu nanashu shusei/A Complete Annotated Edition of Kokinwakashu: with Seven Version of Old Annotations* [**古今和歌集全評釈: 古注七種集成**]. Vol. 1. Yubunshoin.

— (1976b). *Kokinwakashu zen hyoshaku: Kochu nanashu shusei/A Complete Annotated Edition of Kokinwakashu: with Seven Version of Old Annotations* [**古今和歌集全評釈: 古注七種集成**]. Vol. 2. Yubunshoin.

Voloshinov, V. N. (1986). *Marxism and the Philosophy of Language*. Trans. by Ladislav Matejka and I. R. Titunik. Harvard University Press.

Yamamoto, Hilofumi (2005). “A Mathematical Analysis of the Connotations of Classical Japanese Poetic Vocabulary”. PhD thesis. The Australian National University.

— (2007). “Waka no Tame no Hinshi Taguzuke Shisutemu/POS Tagger for Classical Japanese Poems [和歌のための品詞タグづけシステム]”. In: *Nihongo no Kenkyu/Studies in the Japanese Language* [**日本語の研究**] 3 (3), 33–39. 10.20666/nihongonokenkyu.3.3_33.

— (2009). “Bunrui Kodo Tsuki Hachidaishu Yogo No Sisoraasu/Thesaurus for the Hachidaishu (ca. 905-1205) with the Classification Codes Based on Semantic Principles [分類コードつき八代集用語のシソーラス]”. In: *Nihongo no Kenkyu/Studies in the Japanese Language* [**日本語の研究**] 5 (1), 46–52. 10.20666/nihongonokenkyu.5.1_46.

Yamamoto, Hilofumi and Bor Hodošček (2018). “A Study on the Distribution of Cooccurrence Weight Patterns of Classical Japanese Poetic Vocabulary”. In: *Proceedings of the 8th Conference of Japanese Association for Digital Humanities (JADH2018) “Leveraging Open Data”* (2018), 179–182.

— (2019). “An Analysis of the Differences Between Classical and Contemporary Poetic Vocabulary of the Kokinshu”. In: *The 9th Conference of Japanese Association for Digital Humanities (JADH2019) “Localization in Global DH”* , 68–71.

— (2021). *Hachidaishu Vocabulary Dataset*. 10.5281/zenodo.4744170.