

Plotting Poetry 2025

Transforming Poetic Thought into Waka:

How to Pack the Skeleton into a 31-Syllable Closet

- Bor Hodošček, [The University of Osaka](#)
- Hilofumi Yamamoto, [Institute of Science Tokyo](#)

thought2waka

Basics of WAKA

Classical Japanese Poetry, **WA****KA**

- **WA** → Japanese / Japanese style
- **KA** → Song

Early Established Waka

- The Man'yōshū: est. around 7-8th century in Chinese notation. written in Chinese characters, but read in Japanese.
- The Kokinshū: est. ca. 905 in Japanese notation. written in Japanese characters, and read in Japanese.
- Before the Man'yōshū, Kanshi (Chinese poetry) was the dominant form.

Style and Rhetorics

- Include only 31 syllables with 5,7,5,7,7 sounds

	Japanese	Romaji	English Translation
5	うめがえに	ume ga e ni	at the plum branch
7	きめるうぐひす	kiiru uguhisu	warbler came
5	はるかけて	haru kakete	cries over spring
7	なけどもいまだ	nake domo imada	even though it cries
7	ゆきはふりつつ	yuki ha furi tsutsu	snow keeps falling

Theme: **Waiting for the arrival of spring**

Style and Rhetorics

- Express natural views and emotions in a simple sentence:
 - plum branch, warbler, spring, snow
- Use of rhetorics to create a poetic atmosphere:
 - Pun (kakekotoba)
 - Pillow words (makurakotoba)
 - Introductory words (o-kotoba)

Preface of Kokinshū: Kanajo

やまとうたは、人の心を種として、
よろづの言の葉とぞなれりける。
世の中にある人、ことわざ繁きものなれば、
心に思ふことを、見るもの聞くものにつけて、言ひ出せるなり。

Japanese poetry (yamato-uta) takes the human heart as its seed,
and from it grows a myriad of words and leaves.
Since people living in this world are
surrounded by countless events,
they express what they feel in their hearts
by attaching it to the things they see and hear.

Preface of Kokinshū: Kanajo

- Does not mention the 31-syllable form
- The format is driven from the practice of poetic expression
- Not too short, not too long, just right for expressing emotions
- One theory suggests that the pleasantness of phonetics and rhythm (5-7 pattern),
- The length of breath, and ease of recitation and transmission are involved.

Poetic ideas pack into 31-Syllable Form

- The 31-syllable is the final form of the poem, not the initial one.
- The constraint of Waka is the construction of 5,7,5,7,7 syllables.
- Poets create a poem under the 5 segments of 5,7,5,7,7 syllables constraint.
- It is the first step to shorten ideas to fit to 5 or 7 syllables.

Poetic Rules may include:

- Omission of grammatical elements
- Inversion of word order
- Symbolic substitution
- Nominalization
- Manipulation of ambiguity
- Compression of meaning
- Expansion of meaning
- Reinterpretation of context
- ...

Obtain some typical conversion patterns from both

- OP: original poems, and
- CT: contemporary translations

Through the comparison of OP and CT, we can obtain:

- Grammatical pattern, especially predicative elements.
i.e. tense, aspect, ← elements making a poem longer.
- Lexical construction such as proper nouns.
- Rhetorical techniques → such as implications.

Material

- A) Kokinshu: a collection of 1000 waka poems
- B) Modern Japanese translations: 10 sets of translations
 - Parallel corpus: a dataset of original poems and their translations

A: Kokinshu 1000 original dataset (OP)

- **Hachidaishu Classical Japanese Poetic Vocabulary Dataset** on Zenodo contains the original poems of the Hachidaishu (including the Kokinshu) and their semantic codes.
- <https://zenodo.org/records/14001396>
- Creators: Yamamoto, Hilofumi and Hodošček, Bor
- Published: October 28, 2024 / Version v1.0.1
- Hachidaishu classical Japanese poetic vocabulary dataset
- DOI [10.5281/zenodo.14001396](https://doi.org/10.5281/zenodo.14001396)

B: Ten sets of the Translations

No.	Translator	Year	Pages	Manuscript	Translation Style
1.	Kaneko Motoomi*	1933	1,105	Teika	Literal translation
2.	Kubota Utsubo	1960	1,449	Teika	Literal translation
3.	Matsuda Takeo	1968	1,998	Teika	Free translation
4.	Ozawa Masao	1971	544	Teika	Changes word order and grammar
5.	Takeoka Masao	1976	2,278	Teika	Literal translation
6.	Okumura Tsuneya	1978	434	Teika	Respects author's intent
7.	Kusojin Hitaku	1979	1,260	Teika	Supplements words
8.	Komachiya Teruhiko	1982	407	Teika	Unknown
9.	Kojima Noriyuki & Arai Eizo	1989	483	Teika	Unknown
10.	Katagiri Yoichi	1998	3,022	Teika	Literal translation

Kokinwakashu Hyoshaku by Motoomi Kaneko

- only Kaneko Motoomi's translation is available on Zenodo.
- [Kokinwakashu Hyoshaku by Motoomi Kaneko translation sentence vocabulary dataset](#)
- <https://zenodo.org/records/13942707>
- Hilofumi Yamamoto, Bor Hodošček, and Xudong Chen
- Published October 16, 2024 / Version v1.0.1
- DOI [10.5281/zenodo.13942707](https://doi.org/10.5281/zenodo.13942707)

Methods

- Using a parallel corpus of waka (OP) and modern Japanese translations (CT)
- Align waka (OP) with contemporary translations (CT)
- Using BG-code (word list semantic principle)
semantic principle codes to match words by 3 levels of categorical similarity.

Steps

- Step 1: Prepare Kokinshu 1000 original dataset (OP).
- Step 2: Prepare 10 kinds of translation datasets (CT).
- Step 3: Divide both OP and CT sentences into tokens.
- Step 4: Attach Meta codes based on WSLP (semantic principle codes) to each token.
- Step 5: Compare OP with CT by Meta codes.
- Step 6: Describe the predication construction patterns.
- Step 7: Describe the noun phrase construction patterns.
- Step 8: Modeling of poetic construction.

Subtraction

$$\mathbf{CT - OP = Residual}$$

- We will subtract the elements of OP from the elements of CT.
- In other words, we will find out what the CT needs to say that the OP does not say.

Parallel Comparison between OP and CT

Kokinshu No. 3 CT by kaneko

OP : はるがすみ.たてる.や.いづこ.みよしの.の.よしの.の.やまに.ゆき.は.ふりつつ

Gloss: spring haze.arise.Q.where?.Miyoshino.of.Yoshino.of.Mt.snow.falling

Spring haze—where does it rise? On Mount Yoshino in Yoshino, the snow keeps falling and falling.

CT : 春には成ったが、長閑な霞の立っているのは何処の辺か、この吉野の里の吉野山には
雪が降り降りして、一向に春めきもしない。

Gloss: spring-----haze.arize-----where---Q-----Yoshino--MtYoshino-
snow--fallfall-----

Spring has arrived, but where is that gentle haze drifting? Here in the Yoshino village, on Mount Yoshino, snow keeps falling and falling, and it shows no sign of spring at all.

We anotated each poem and each translation as the following:

OP: Kokinshu No.3

1	KW000003	111	1	02	00	00	BG-01-5152-09-040-A	はるがすみ はるがすみ 春霞 spring haze
1	KW000003	111	3	02	00	00	BG-01-1624-02-010-A	-- はる 春 spring
1	KW000003	111	3	02	00	00	BG-01-5152-09-010-A	-- かすみ 霞 haze
1	KW000003	211	0	47	25	04	BG-02-1513-01-010-A	たて たつ 立つ
1	KW000003	212	0	74	68	20	BG-09-0010-03-030-C	る り り
1	KW000003	213	0	65	00	00	BG-08-0065-14-010-C	や や や
1	KW000003	221	0	14	00	00	BG-01-1700-02-100-C	いづこ いづこ 何処
1	KW000003	311	0	11	00	00	CH-29-0000-20-010-A	みよしの みよしの 御吉野
1	KW000003	312	0	71	00	00	BG-08-0071-01-010-A	の の の
1	KW000003	411	0	11	00	00	CH-29-0000-20-010-A	よしの よしの 吉野
1	KW000003	412	0	71	00	00	BG-08-0071-01-010-A	の の の
1	KW000003	421	0	02	00	00	BG-01-5240-05-010-A	やま やま 山
1	KW000003	422	0	61	00	00	BG-08-0061-05-010-A	に に に
1	KW000003	511	0	02	00	00	BG-01-5153-07-010-A	ゆき ゆき 雪
1	KW000003	512	0	65	00	00	BG-08-0065-07-010-A	は は は
1	KW000003	521	0	47	28	03	BG-02-1540-10-010-A	ふり ふる 降る
2	KW000003	521	2	47	28	03	BG-02-5150-03-010-A	ふり ふる 降る
1	KW000003	522	0	64	00	00	BG-08-0064-15-010-A	つつ つつ つつ

CT: Kaneko No.3

1	kaneko	0003	0	02	00	00	BG-01-1624-02-010-A	春 はる 春 spring
1	kaneko	0003	0	61	00	00	BG-08-0061-05-010-A	に に に
1	kaneko	0003	0	65	00	00	BG-08-0065-07-010-A	は は は
1	kaneko	0003	0	47	17	06	BG-02-1220-01-030-A	成っ なる 成る
1	kaneko	0003	0	74	54	01	BG-09-0010-04-010-A	た た た
1	kaneko	0003	0	64	00	00	BG-08-0064-04-010-A	が が が
1	kaneko	0003	0	79	00	00	BG-16-0079-01-010-A	、 、 、
1	kaneko	0003	1	18	00	00	BG-03-3010-02-140-A	長閑 のどか 長閑
1	kaneko	0003	2	18	00	00	BG-03-5150-02-040-A	-- のどか のどか
1	kaneko	0003	0	74	55	06	BG-09-0050-01-030-A	な だ だ
1	kaneko	0003	0	02	00	00	BG-01-5152-09-010-A	霞 かすみ 霞 haze
1	kaneko	0003	0	61	00	00	BG-08-0061-07-010-A	の の の
1	kaneko	0003	0	47	13	05	BG-02-1513-01-010-A	立っ たつ 立つ
2	kaneko	0003	2	47	13	05	BG-02-1521-06-020-A	立っ たつ 立つ
3	kaneko	0003	2	47	13	05	BG-02-3330-11-020-A	立っ たつ 立つ
4	kaneko	0003	2	47	13	05	BG-02-3391-02-110-A	立っ たつ 立つ
1	kaneko	0003	0	64	00	00	BG-08-0064-16-010-A	て て て

... continues

Meta-code system

BG-01-2030-01-030-A-かみ-神 (god)
↑ ↑ ↑
GFE
↓ ↓ ↓
BG-01-2030-01-250-A-ほとけ-仏 (Buddha)

- G: Group match... 10 digits
- F: Field match..... 13 digits
- E: Exact match..... 17 digits

The three matching levels are judged by the length of BG-code digits.

Code Categories with English annotation

BG-01-1000-00-000-X:demonstrative_pronoun
BG-01-1100-00-000-X:class,kinds
BG-02-1000-00-000-X:abstract_relation
BG-02-1110-00-000-X:relation
BG-03-3100-00-000-X:language_and_speech
BG-03-3400-00-000-X:personal_affairs
BG-04-1100-00-000-X:conjunction
BG-05-0000-00-000-X:prefix
BG-06-0000-00-000-X:infix
BG-07-0000-00-000-X:suffix
BG-08-0061-00-000-X:case_particle
BG-09-0000-00-000-X:auxiliary_verb
BG-10-0000-00-000-X:auxiliary_verb_and_auxiliary_adjective
BG-11-0000-00-000-X:relative_pronoun
BG-12-0000-00-000-X:word_endings
BG-13-0000-00-000-X:preposition_and_postposition
BG-14-0000-00-000-X:meaning_unknown
BG-15-0000-00-000-X:proper_noun
BG-16-0000-01-000-X:punctuation
BG-17-0000-00-000-X:wordplay_handling
BG-18-0000-00-000-X:counting

Computer Tools

code2match.c

- Align waka with contemporary translations
- github: <https://github.com/yamagen/code2match>

```
% cat op_file.txt ct_file.txt | code2match -a
```

code2match -h

```
% code2match [-ahv] file....  
-a    print all data  
-b    print between check  
-c    print calculation table  
-d    print predicate part out  
-e    once matched out (bag of words option)  
      use it with other options  
-i    print calculation in line style  
-l    print token list table  
-o    print original poem out  
-p    print pair token table  
-r    print residual  
-s    print valid on  
-t    print title  
-u    print unmatched portion  
-h    print this help  
-v    print code2match version  
(c) 2025 H. Yamamoto yamagen@ila.titech.ac.jp
```

Pair Token Table: -p

+----- number of pair			value of exact=17, field=13, group=10			number of POS			number of OP token			number of CT token		
+-----			+-----			+--			-----+			+-----		
									OP token			CT token		
									OP token			CT token		
1	13	2							春	01	<->	00	春	
2	17	2							霞	02	<->	10	霞	
3	17	47							立つ	03	<->	12	立つ	
4	13	65							や	05	<->	26	か	
5	17	14							何処	06	<->	20	何処	
6	17	71							の	08	<->	21	の	
7	17	11							吉野	09	<->	30	吉野	
8	17	71							の	10	<->	31	の	
9	17	2							山	11	<->	37	山	
10	17	61							に	12	<->	38	に	
11	17	2							雪	13	<->	40	雪	
12	17	65							は	14	<->	02	は	
13	17	47							降る	16	<->	43	降る	
14	10	64							つつ	17	<->	47	て	

Print Residual: -r

Residual tokens reveal what the translation needs to say that the original poem leaves unsaid.

```
CT A--B--C--D--E--F--G--H-----
7 0 1 0 -1 64 0 0 BG-08-0064-16-010-A て て
10 0 1 0 -1 61 0 0 BG-08-0061-02-010-A が が
12 0 1 0 -1 16 0 0 BG-01-1624-05-010-A 冬 冬
13 0 1 0 -1 16 0 0 BG-01-1612-01-060-A 時分 時分
14 0 1 0 -1 61 0 0 BG-08-0061-01-010-A から から
15 0 1 0 -1 57 0 0 BG-03-1000-01-010-A この この
17 0 1 0 -1 61 0 0 BG-08-0061-08-010-A へ へ
21 0 1 0 -1 18 0 0 BG-03-1600-03-020-A 頻り 頻り
22 0 1 0 -1 72 0 0 BG-08-0072-02-010-A に に
33 0 1 0 -1 47 3 7 BG-02-3420-01-010-A し する
36 0 1 0 -1 55 0 0 BG-03-1200-03-060-A 一向 一向
37 1 1 0 -1 47 8 2 BG-02-1624-02-110-A 春めか 春めく
42 1 1 0 -1 74 59 1 BG-03-1200-02-090-A ぬ ぬ
45 0 1 0 -1 21 0 0 BG-01-1010-01-020-A こと こと
46 1 1 0 -1 69 0 0 BG-08-0069-30-010-A よ よ
47 0 1 0 -1 61 0 0 BG-08-0061-03-010-A へ へ
```

Elements breakdown between OP and CT: -c

OP(original poem; valid number of items)	= 16
E (ratio of exact agreement)	$11/16 = 0.688$
F (ratio of field agreement)	$2/16 = 0.125$
G (ratio of group agreement)	$1/16 = 0.062$
T (ratio of total agreement)	$14/16 = 0.875$
U (ratio of unmatched)	$1 - T = 0.125$

CT(contemporary translation; valid number of items)	= 39
W (ratio of original word use)	$11/39 = 0.282$
A (ratio of annotation)	$1 - W = 0.718$
- breakdown of the annotation -	
P1(ratio of FG paraphrased)	$(F+G)/V = 0.077$
P2(ratio of U paraphrased)	$(A-P1)*U = 0.080$

D (ratio of purely added)	$A - (P1+P2) = 0.561$
H (theoretical value)	$1 - 16/39 = 0.590$
Gap:	$\text{fabs}(D-H) = 0.029$

Predicate alignments between OP and CT: -d

```
$ cat data/kokin/0005.db.txt data/kaneko/0005.db.txt | src/code2match -d
PRED: kaneko    5 [09|かけ|て|なけ|ども|13] => [19|かけ|て|頻り|に|鳴く|けれども|24]
PRED: kaneko    5 [18|ふり|つつ|19] => [30|降り降り|し|て|34]
```

```
$ cat data/kokin/0007.db.txt data/kaneko/0007.db.txt | src/code2match -d
PRED: kaneko    7 [12|きえあへ|ぬ|15] => [20|消え|て|果て|ず|25]
PRED: kaneko    7 [22|みゆ|らむ|23] => [41|見える|の|で|あろ|う|46]
```

op predicate

ct predicate

Script to run code2match

```
#!/bin/sh

# This script compares two directories containing Waka poems and their translations.
if [ "$#" -lt 3 ]; then
    echo "Usage: $0 <dir1> <dir2> <id> [option]"
    exit 1
fi

DIR1="$1"
DIR2="$2"
ID=$(printf "%04d" "$3") # ID can be 1-9999, so we format it to 4 digits
OPTION="$4"              # Optional argument for code2match

cat "$DIR1/$ID.db.txt" "$DIR2/$ID.db.txt" | ../src/code2match $OPTION
```

Script: loop 1-1000 to run code2match

```
#!/bin/sh

# args: $1 = kokin directory name (e.g., kokin)
#       $2 = contemporary translation directory name (e.g., kaneko)
#       $3 = poem ID or range (e.g., 1, 100, or 1-100)
#       $4 = optional argument for code2match (e.g., -d, -r)

SRC=../src/code2match

# judge if $3 is a range or a single number
if echo "$3" | grep -qE '^[0-9]+-[0-9]+$'; then
    START=$(echo "$3" | cut -d- -f1)
    END=$(echo "$3" | cut -d- -f2)
else
    START=$3
    END=$3
fi

# Loop through the specified range or single number
for i in $(seq "$START" "$END"); do
    FILE1="$1/$(printf '%04d' "$i").db.txt"
    FILE2="$2/$(printf '%04d' "$i").db.txt"

    if [ -n "$4" ]; then
        cat "$FILE1" "$FILE2" | "$SRC" "$4"
    else
        cat "$FILE1" "$FILE2" | "$SRC"
    fi
done
```


The Compression of Poetic Thought into 31-Syllable Form

- How to detect the compression of poetic thought into 31-syllable form?
- Should we use multivariate analysis of the parallel corpus?
- What variables do we need to consider?

- Even a statistician would hesitate to give a definitive answer here.
- We will observe the patterns of compression one by one.

So far, we've sketched out the problem—but how do we proceed?

Asking AI? Not bad, but...

- John Tukey's Exploratory Data Analysis (EDA) is a good start.
- We will seek the evidence but more than that,
→ ***we need the accountability of the results.***

Results

- Identify and classify poetic strategies
- Analyze how poetic thought is transfigured
- Uncover underlying rules (overt and covert)
- Explore the implications of compression
- Simulate the transformation process:

Discussion

- Explore poetic compression in modern Japanese
- Analyze constraints in poetic expression
- Discuss implications for translation and interpretation
- Consider cultural and linguistic factors

Conclusion

The ways of the compression of Poetic Thought
Into 31-Syllable Form (the Closet of skeleton)

- Word Compression
- Predicate Compression
- Shortening by removing grammatical elements

Word Types

- Chinese word construction techniques applied to Waka
 - Two chinese characters combination methods.
 - person + action (e.g., 人言, 人来, a person speaks, a person comes)
... not: 人の言葉, 人の来る, someone speaks words, someone comes somewhere
 - noun + noun (e.g., 山川, 山野, mountain and river, mountain and field)
... not: 山の川, 山の野, mountain's river, mountain's field
 - noun modifier + modified noun (e.g., 朝露, 白露, morning dew, white dew)
... not: 朝に降りている露, 白く光った露, morning's dew, white dew
- These are one of the **compression methods** in Waka.

Predicate sections

- The simplest verb form can express variously.

Content words

- No modifications.
- Noun and adjective expand images

Remarks

- The 31-syllable form is not a fixed structure but a flexible framework.
- Poets use the 31-syllable form to express their emotions and thoughts in a concise manner.
- Use of hypernyms to indicate the general meaning of the poem.
- Use of generic/shorter nouns (hana = flower) rather than specific nouns (hana tachibana = the flower of orange).

Future research directions

Reference

- Kamitani, Kaoru, (1999). Kokinwakashu yogo no goiteki kenkyu (Lexical Study of Kokinwakashu vocabulary), Izumi Shoten, Osaka.
- Yamamoto, H., Hodošček, B., & Chen, X. (2024). Hachidaishu Part-of-Speech Dataset (1.0.1) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.13940187>
- Yamamoto, H., Hodošček, B., & Chen, X. (2024). Kokinwakashu Hyoshaku by Motoomi Kaneko translation sentence vocabulary dataset (v1.0.1) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.13942707>
- Hilofumi Yamamoto. Thesaurus for the Hachidaishu (ca. 905-1205) with the classification codes based on semantic principles, The Study of Japanese Linguistics, The Society of Japanese Linguistics, Vol. 5, No. 1, pp. 46-52, Jan. 2009.