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,
WAKA

- WA → Japanese / Japanese style
- KA → song

Early established *waka*

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

Example from the Kokinshu

Example 2

- 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	きゐるうぐひす	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

Waka: Stylistic and rhetorics perspective

- 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 derived from the practice of poetic expression
- Not too short, not too long, ‘just right’ for expressing emotions
- One theory suggests the involvement of
  - pleasantness of phonetics and rhythm (the 5-7 pattern),
  - length of breath,
  - ease of recitation, and
  - transmission.

Poetic ideas pack into the 31-syllable form

- The 31-syllables 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.

Methods

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 patterns, especially predicative elements. i.e. tense, aspect, ...
  ← elements making a poem longer.
- Lexical constructions such as proper nouns.
- Rhetorical techniques such as implications.

Materials

- A) *Kokinshu*: a collection of 1,000 *waka* poems → Hachidaishu Classical Japanese Poetic Vocabulary Dataset on Zenodo contains the original poems of the Hachidaishu (including the Kokinshu) and their semantic codes.
- B) Modern Japanese translations: 10 sets of translations → Parallel corpus of original poems and their translations Kokinwakashu Hyoshaku by Motoomi Kaneko translation sentence vocabulary dataset
  - only Kaneko Motoomi's translation is available on Zenodo

B: Ten sets of contemporary translations

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

<!--

Methods

Using a parallel corpus of pre-tokenized *waka* (OP) and modern Japanese
translations (CT),

- align *waka* (OP) with contemporary translations (CT)
- use the BG-code (WLSP: word list semantic principle) semantic principle codes
  to match words at 3 levels of categorical similarity
- subtract and model poetic construction

Subtraction

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.

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: hierarchical semantic categories

```
 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 matching BG-

Examples of 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
```

Pair token table: `code2match -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
 |   |   |                         |  |     |   |
 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
```

Extract residual of Kaneko no. 5: `code2match -r`

Residual tokens reveal what the translation needs to say that the original poem
leaves unsaid. (Example output:)

```
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
...
```

Element breakdown between OP and CT: `code2match -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:                                               fabs(D-H)= 0.029
```

Figure: Ingredients of the translation of Kokinshu No. 298 by Komachiya.

Predicate alignments between OP and CT: `code2match -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
```

The compression of poetic thought into 31-syllable form: Questions

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

> Even a statistician would hesitate to give a definitive answer here.

Considerations in approach

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

By asking AI? But how are we going to explain the results...

→ ***we need accountability of the results.***

Why is it important that researchers go "hands-on"?

- **No "black box"** — Manually validate data-hypothesis links using explainable
models (critical for linguistics).
- **Small examples = deeper insight** — E.g., tracing "春霞 → 春 + ... + 霞" or
"ふりつつ → 降り降りして" reveals transformation logic.
- **Hands-on exploration** — Prioritizes understanding processes over just
results, essential for nuanced linguistic analysis of complex data.

We believe that John Tukey's Exploratory Data Analysis (EDA) is a good start.

Results of the hands-on process

Nouns avoided in *waka* (top 20 residuals)

- **Abstract & deictic nouns**
  - 花 (flower), こと (thing), それ (that), もの (thing), これ (this)
  - ⇒ Preference for concrete, symbolic imagery
- **Time & season terms**
  - 時 (time), 春 (spring), 昔 (past), 今 (now)
  - ⇒ General time words give way to specific seasonal words
- **Terms of self & agency**
  - 人 (person), 自分 (self), 内 (inside), 外 (outside), はず (should)
  - ⇒ Avoidance of explicit self-reference

Key insights

- **Concrete / symbolic keywords**
  - *Waka* retains vivid imagery; abstract/utility nouns are cut
- **Poetic temporal expression**
  - General time nouns are replaced by evocative seasonal or momentary phrases
- **Anthology bias**
  - 梅 (plum), 桜 (cherry), 雪 (snow) less frequent in some collections

Comment on nature themes as residuals

- **Unexpectedly low direct frequencies**
  - i.e., *ume* (plum), *sakura* (cherry), and *yuki* (snow)
  - Often subsumed under the generic term "flower" or conveyed metaphorically:
    yuki/snow as 白き/white 花/flower
- **A promising focal point for comparative studies on thematic selection**
  - Why do these specific nature terms appear less frequently in *waka*?
  - Why did not poets choose simple, direct expressions for these themes? (Such
    as *ume* or *sakura*?)

<!--

Command executed

```
./c2m.sh kokin kaneko 1-100 -d \
  | awk '{print length($0), $0}' \
  | sort -nr \
  | nl \
  | head -10
```

- Extracted the top 10 longest predicate mappings between the Kokin and Kaneko
  corpora
- Each line shows:
  - Length in characters

```
$ ./c2m.sh kokin kaneko 1-100 -d| awk '{print length($0), $0}' | sort -nr | nl | head
-10
     1 148 PRED: kaneko  86 [21|ふく|らむ|22] => [07|吹か|ぬ|時|に|も|、|雪|の|よう|に|ひた
すら|散る|が|、|それ|さえ|以て|惜しく|ある|もの|を|、|また|この上|どのように|烈しく|散れ|
と|いっ|て|、|こう|も|風|が|吹く|の|で|あろ|う|57]
     2 111 PRED: kaneko  78 [21|まちみ|て|ちら|ば|ちら|なむ|29] => [35|待っ|て|み|て|、|い
よいよ|来|ぬ|時|に|こそ|、|散る|なら|ば|お前|の|勝手|に|散っ|て|貰お|う|わ|64]
     3  98 PRED: kaneko  36 [11|をり|て|かざさ|む|15]  => [27|折り取っ|て|、|我が|容貌|の|
老|も|隠れる|か|どう|か|と|、|試し|に|挿頭し|て|みよ|う|58]
     4  96 PRED: kaneko  11 [01|き|ぬ|04] =>  => [02|た|と|世間|の|人|は|いう|が|、|まだ|
鴬|は|鳴い|て|い|ない|、|自分|は|何でも|鴬|の|鳴か|ぬ|31]
     5  94 PRED: kaneko  76 [12|をしへよ|いき|て|うらみ|む|18] => [32|教え|て|くれ|よ|、|然
らば|、|そこ|に|行っ|て|思う存分|恨み|を|いお|う|52]
     6  91 PRED: kaneko  61 [04|くははれ|る|05]  => [11|加わっ|て|長く|なっ|た|今年|なり|と
も|、|人|の|心|に|は|なぜ|厭か|れ|は|せ|ぬ|39]
     7  85 PRED: kaneko  77 [06|ちり|な|む|11]  => [08|散る|なら|ば|、|自分|も|一緒|に|何
処|へ|なり|と|退散|し|て|しまお|う|32]
     8  85 PRED: kaneko  74 [03|ちら|ば|ちら|なむ|ちら|ず|11] => [07|散る|なら|ば|勝手|に|
散っ|て|貰お|う|、|たとえ|散ら|ず|22]
     9  83 PRED: kaneko  45 [00|くる|と|あく|と|めかれ|ぬ|ものを|12] => [12|と|いっ|て|は|
見|、|夜|が|明ける|と|いっ|て|31]
    10  80 PRED: kaneko  63 [01|こ|ず|03] => [07|来れ|ば|こそ|、|この|桜|を|花|と|は|見|ま
すれ|、|若し|今日|来|ぬ|28]
```

-->

Key observations from predicate correspondence analysis

- Substantial expansions

  Eg. “ふく|らむ” ⇒ “吹かぬ時に…雪のようにひたすら散るが…” (148 chars)

- Frequent verb classes

  Eg. scattering (散る), blooming (咲く), seeing (見る), falling (降る)

- *Waka* elaboration patterns
  - Addition of temporal/conditional clauses
  - Shift from simple verb forms to more poetic constructions

Word Types

- Chinese word construction techniques applied to *waka*:
  - person + action (e.g., 人言 person speaks, 人来 person comes) ...
    not: 人の言葉 someone speaks words, 人の来る someone comes somewhere
  - noun + noun (e.g., 山川 mountain and river, 山野 mountain and field) ...
    not: 山の川 mountain's river, 山の野 mountain's field
  - noun modifier + modified noun (e.g., 朝露 morning dew, 白露 white dew) ...
    not: 朝に降りている露 morning-falling dew, 白く光った露 white-shining dew

→ These are one of the compression methods in *waka*.

Summary of poetic compression techniques encountered

- Poetic compression is a key feature of *waka*.
- It involves condensing complex ideas into concise phrases.

Technique description	Example
Compressing sentences into words	梅の花 (ume no hana)
Using Chinese characters to condense meaning	朝露 (asa tsuyu)
Avoiding Repetition for emphasis	降りつつ (furi tsutsu)
Abstracting emotions	鳴く→cry(birds)..cry(human)
Omitting unnecessary words	白露 (shira tsuyu)
Leaving interpretation to the reader	白...雪/snow, 花/flower

Questions for discussion

- What kinds of patterns do you observe when comparing poetic originals with
their translations in your own corpus?
- Have you found similar cases of word expansion (e.g., a single poetic word
becoming a phrase in translation)?
- Do you ever annotate or align poetic lines manually before analysis? How do
you balance structure and meaning?
- Can we think of ways to represent translation divergence not just as loss/gain, but as stylistic transformation?
- How might this Japanese example (e.g., "朝露" → "露が朝に降りている") resonate
with condensation in your own poetic tradition?

Conclusion

Methods for compressing poetic thought into 31-syllable form
→How to pack ideas into the closet?

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

> This is how we approached translation analysis in Japanese *waka*.
> We wonder
> how similar or different your poetic traditions behave in such transformations. Let's explore this together.

References

- Kamitani, Kaoru, (1999). Kokinwakashu yogo no goiteki kenkyu (Lexical Study of Kokinwakashu vocabulary), Izumi Shoten, Osaka.
- Sachi Kato, Masayuki Asahara, Nanami Moriyama, Asami Ogiwara, and Makoto Yamazaki (2021). Opposite Information Annotation on Word List by Semantic Principles, Journal of Natural Language Processing, Vol.28, No.1, 60-81, DOI https://doi.org/10.5715/jnlp.28.60.
- John W. Tukey, (1977). Exploratory Data Analysis, Addison-Wesley, Reading, MA.

References (cont.)

- Yamamoto, H. and Hodošček, B. (2019). An Analysis of the Differences Between
Classical and Contemporary Poetic Vocabulary of the Kokinshu, The 9th
Conference of Japanese Association for Digital Humanities, JADH2019)
Localization in Global DH, 68--71.
- 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]