# **Development of English translations for use in a visualization** system of classical Japanese poetic vocabulary

# Summary



Hodošček Bor **Osaka University, Graduate School of Humanities** hodoscek.bor.hmt@osaka-u.ac.jp

Yamamoto Hilofumi Tokyo Institute of Technology vamagen@ila.titech.ac.ip

Utilized structured generation techniques with multiple open-source and closed-source LLMs to programmatically create English translations and word-level alignments of classical Japanese poetry. Evaluations demonstrate that chunk- and line-based translations achieve best quality while word-level translations remain useful for generating English tokens for the visualization of Japanese poetic vocabulary. Prompt-based word alignment was also employed to recover token mappings from line- and chunk-based translations. URL https://github.com/borh/hachidaishu-translation

# **1** Introduction and Motivation

Working on visualizing lexical relationships in graphs showing Japanese terms and their corresponding English translations. During the manual development of an English token dictionary, we noticed that GitHub Copilot provided useful completion tokens but sometimes introduced innaccuracies. However, a human-in-the-loop approach was found to be highly effective in our pilot experiment.



%cw -dE 11001,11002,11003,11005,11006,30569,30571,30572,30573,30574,50713,50714,50715,50716,50717,71160,71161,71162,71163

URL https://cuckoo.js.ila.titech.ac.jp/~yamagen/waka/poem.cgi

M 7 -W 2.5 -L 0 -F 1-1 -C 0 -K BG-01-1632-01-01 -Z 1

# **2 Pilot Experiment**

- Translations were produced using ChatGPT-4, Copilot, Copilot with human assistance, and human translations. Findings:
- ChatGPT-4 offered natural and easy-to-understand translations but sometimes diverged too far from the source content.
- Copilot was more accurate but produced less natural translations. With human refinement, it balanced accuracy and fluency.
- Human translations provided the best literary nuance but were costly and time-consuming.

# **③** Results: translations evaluation

Translations are assessed comparing to our golden translation dataset (Rodd, L. R., & Henkenius, M. C. (Eds.). (1996). Kokinshū: A Collection of Poems Ancient and Modern. Cheng & Tsui.). We randomly sample 100 translations and use three different translation granularities (word, chunk and line) to perform a total of a maximum 300 translations per model below. (Model with / in their name are using their HuggingFace Models identifier; the rest are APIbased models). Models were chosen from a variety of best-performing as well as specialist models (webbigdata/ALMA-7B-Ja-V2 for Japanese-English translation).

### **Evaluation metrics employed:**

**chrF**: surface-level metric; F-score of character-level n-grams in translations **METEOR:** considers semantic equivalence through synonym matching and stemming SentenceTransformers (cosine similarity): sentence embeddings capture semantic similarity

chrF			ມີ
			Ť
(WhereIsAI/UAE-Large-V1)		Distribution of evaluations per model, generation method and type ordered by geometric mean of metrics	Ð
claude-3-5-sonnet-20240620 / MAGENTIC-pydantic / chunks / t=0.3			1009
claude-3-5-sonnet-20240620 / MAGENTIC-pydantic / lines / t=0.3			1009
gpt-4o-2024-08-06 / MAGENTIC-pydanti	ic / chunks / t=0.3		1009
gpt-4o-2024-08-06 / MAGENTIC-pyda	ntic / lines / t=0.3		100%
open-mixtral-8x22b / MAGENTIC-cha	at / chunks / t=0.3		989
mistral-large-latest / MAGENTIC-c	chat / lines / t=0.3		1009
open-mixtral-8x22b / MAGENTIC-chat / lines / t=0.3			989
meta-llama/Meta-Llama-3.1-8B-Instruct / HF-regex / lines / t=0.3			1009
Qwen/Qwen2-7B-Instruct / HF-re	egex / lines / t=0.3		1009
weblab-GENIAC/Tanuki-8B-dpo-v1.0 / HF-pydantic / chunks / t=0.3			989
plamo-beta / MAGENTIC-chat / chunks / t=0.3			489
google/gemma-2-9b-it / HF-re	egex / lines / t=0.3		1009
claude-3-5-sonnet-20240620 / MAGENTIC-pydantic / words / t=0.3			1009
tokyotech-llm/Swallow-7b-instruct-v0.1 / HF-regex / lines / t=0.3			1009
meta-llama/Meta-Llama-3.1-8B-Instruct / HF-regex / chunks / t=0.3			1009
webbigdata/ALMA-7B-Ja-V2 / HF-pydantic / words / t=0.3			19
google/gemma-2-9b-it / HF-regex / chunks / t=0.3			1009
microsoft/Phi-3.5-mini-instruct / HF-regex / lines / t=0.3			1009
meta-llama/Meta-Llama-3.1-8B-Instruct / HF-pydantic / lines / t=0.3			1009
plamo-beta / MAGENTIC-chat / lines / t=0.3			719
open-mistral-nemo / MAGENTIC-chat / chunks / t=0.3			609
weblab-GENIAC/Tanuki-8B-dpo-v1.0 / HF-pydantic / lines / t=0.3			1009
gpt-4o-2024-08-06 / MAGENTIC-pydantic / words / t=0.3			1009
google/gemma-2-9b-it / HF-pydantic / chunks / t=0.3			1009
tokyotech-llm/Swallow-7b-instruct-v0.1 / HF-pydantic / lines / t=0.3			1009
meta-llama/Meta-Llama-3.1-8B-Instruct / HF-pydantic / chunks / t=0.3			1009

**Problem:** Manual translation of Japanese tokens into English is highly context-sensitive and time-consuming.

**Proposed solution:** Leverage structured generation with LLMs to programmatically assign English tokens.



**Extract and align poems** 

with golden translations

**Build run** 

configuration

Iterate over models

**Iterate over generation** 

methods (PYDANTIC/

**REGEX/CHAT**)

Iterate over run types

(LINES/CHUNKS/

WORDS)

Translate

No

Yes

Save alignments

0.67

1.00

0.56

Retry

Valid?

Yes

run?

System overview

We use a previously open-sourced dataset of the Hachidaishu, a collection of eight Japanese imperial anthologies (ca. 905-1205). The Hachidaishu database consists of *waka* poems of 5-7-5-7-7 syllabus length.

# Structured generation overview

SYSTEM\_PROMPT = """You are an expert Japanese-English poetry translator. Your task is to faithfully translate given Japanese waka poems into English. Ensure that each translation is between 16 and 40 words long. Provide only the translated poem without any additional notes or formatting unless ordered to do so."

#### Word-level translation with JSON schema via github:jackmpcollins/magentic

class WordSchema(BaseModel) original: str = Field(description="Original Japanese word") translated: str = Field(description="Translated English word")

class WordsSchema(BaseModel) words: list[WordSchema] = Field( description="List of original-translated word pairs."

@magentic.prompt("""{instructions}

Franslate the following waka poem into English, word-by-word. Dutput the translation for each word using at most 3 English words.

‡# Japanes

## Japanese-English word pairs

def magentic\_words\_prompt\_fn(instructions: str, waka: str) → WordsSchema:

Reasons for retries: invalid JSON schema returned (Pydantic), regex generation error (Outlines/Transformers model problem), presence of Japanese text in translation, too short (< 12 words) translation, API instability, etc.

Pydantic models are converted to JSON schemas and can be used for structured generation in the OpenAI and Anthropic APIs through magentic. They can also be used in Outlines' JSON generation.

#### Word- and chunk-level translation with regex via github:dottxt-ai/Outlines

Word-alignment of existing translations class TokenAlignment(BaseModel): original\_token: str = Field(description="A single Japanese word") translated\_token: str = Field(description="Translation of Japanese word")

alignments list[TokenAlignment] = Field( description="""A list of (original\_token, translated\_token) TokenAlignment objects representing the alignment between tokens in the original and translated texts. The provided tokens are space-delimited string and should not be further split. A token can be aligned to multiple tokens; i and the string the alignment between tokens the string and should not be further split. A token can be aligned to multiple tokens; i and the string the alignment between tokens the string and should not be further split.

such cases, include multiple tuples with the same original\_token paired with different translated\_tokens. Unaligned tokens (typically those with

predominantly grammatical function) can be omitted from the alignment list. For disambiguation, if a token appears multiple times, a suffix is appended it; reuse this suffix to ensure correct alignment."""

class TokenAlignmentSchema(BaseModel)

See the

translations:

# Automatically generate regex for each poem. Example: # 逢坂 の ゆふつけ鳥 も 我 ことく 人 や 戀しき ね のみ 鳴 らん

words\_regex = r""" 逢坂: ([a-zA-Z]{1,20}){1,2}([a-zA-Z]{1,20})? の: ([a-zA-Z]{1,20}){1,2}([a-zA-Z]{1,20})? ゆふつけ鳥: ([a-zA-Z]{1,20}){1,2}([a-zA-Z]{1,20})? らん: ([a-zA-Z]{1,20} ){1,2}([a-zA-Z]{1,20})?

chunks\_regex = r" 逢坂 の: [a-zA-Z ]{1,40} ゆふつけ鳥 も: [a-zA-Z ]{1,40} 我 ことく: [a-zA-Z ]{1,40} 人 や 戀しき: [a-zA-Z ]{1,40} ねのみ 鳴 らん: [a-zA-Z ]{1,40}

> Fallback for chat-based generation using 1-shot prompting in magentic

@magentic.chatprompt( magentic.SystemMessage(SYSTEM\_PROMPT) magentic.UserMessage(ONE\_SHOT\_LINE[0]) magentic.AssistantMessage(ONE\_SHOT\_LINE[1]) magentic.UserMessage("Translate the following [...]"),



See the

aligned tokens:

, def magentic\_lines\_chat\_prompt\_fn(waka: str) → str:



Results of structured generation using JSON structured output mode (via Pydantic schema) for word alignment between original Japanese waka poem and its translation using gpt-4o-2024-08-06 and



C



# **5** Discussion

Multiple temperature settings (0.0, 0.3, 0.5, 0.7) were attempted. Full circles are only attested to by one model, circles having number in center indicate agreement of number models.

In general, frontier models (claude-3.5-sonnet and gpt-40) outperformed smaller models, but preclude the use of regexbased generation methods that offer greater control over the exact tokens generated. Regex-based generations were generally superior to JSON schema structured output on open models usable through the Outlines structured text generation library, but their use in not supported in most API-based models.

MT In the spring wind / flowers / gently / from the heart / to be distracted from my thoughts and be

GT oh spring breeze please do / not draw near these petals as / you waft for I would / like to know whether it is / their own wish that they should fall

### mistral-large-latest / MAGENTIC-chat / chunks chrF = 0.30 (0.82) METEOR = 0.04 (0.06) ST = 0.80 (0.96) 契り けん / 心 そつらき / 七夕 の / 年 に 一たひ / 逢 は あふ か は

MT Bound by a promise / Heart yearning, so cruel / On Tanabata night / Once a year we meet

GT our one promised night / of the year will it come to / pass Tanabata / oh too cruel is the young / heart that would deny me this

## Highest scoring examples by chrF, METEOR and ST

D claude-3-5-sonnet-20240620 / MAGENTIC-pydantic / chunks chrF = 0.53

色 より も / か こそ あはれ と / おもほゆれ / た か 袖 ふれ し / やと の 梅 そ も

MT More than its color, / it's the fragrance that moves me: / the plum blossoms / at the house where once / your sleeve brushed against mine. GT more than the color / of the flower the fragrance / delights my senses-- / whose scented sleeve brushed against / the plum blossoms near my house

### Qwen/Qwen2-7B-Instruct / HF-regex / lines

METEOR = 0.56

### 新しき / 年の始に / かくしこそ / 千年をかねて / たのしきをつめ

At the start of the new year, we anticipate a thousand years of joy and fulfillment, nestled within the freshness of the moment, yet deeply rooted in tradition and hope for the future, embodying the eternal cycle of life and,

at the beginning / of this auspicious new year / we anticipate / the joys of a thousand years / by piling stacks of firewood

### Claude-3-5-sonnet-20240620 / MAGENTIC-pydantic / chunks

ST = 0.93

### かは風の / すいしくもある / かうちよする / 浪といもにや / 秋は立らん

MT The cool river breeze / How refreshing it feels / As waves lap the shore / Together they herald / The arrival of autumn GT how cool the breeze / from the river I feel the / refreshing chill / of near-approaching autumn / roll in with each rising wave



Automated evaluations show moderate correlation but also disagreement and need further analysis to determine the most useful metric (or combinations of metrics) to use for guiding the creation of English word tokens.

# **⑦** Limitations and Future Work

Expanding models' context through the integration of detailed linguistic information such as part-of-speech (POS) data and predeveloped semantic labels from previous work (github:yamagen/ hachidaishu-pos) is likely to help. Also, expanding the use of few-shot prompting (only used for chat-based API use currently) will likely improve format conformance for JSON schema based models.

Assessing the optimal alignment with the Leipzig glossing rules when integrating tokens remains an open question. Future work will explore this issue as well as perform contamination checks on the original Japanese waka and their English translations, and expand the golden translation dataset to other establised translations.

JADH2024 2024/9/18-20@University of Tokyo