Development of an Asymptotic Word Correspondence System between Classical Japanese Poems and their Modern Translations

Hilofumi Yamamoto∗† **Hajime Murai†** **Bor Hodošček‡**

∗ University of California, San Diego **†Tokyo Institute of Technology** **‡Meiji University**

Introduction

- This project will develop an automatic word concordance system for parallel texts comprising of Classical Japanese poem texts and their associated modern translations.
- By using these parallel texts, we will clarify the details of language change within Japanese in an objective procedural manner that is not influenced by human observations.
- Our aim is to develop the thesaurus of classical Japanese poetic vocabulary using the system.

Problem

What is Waka?

Tatsuta-Hime.. (5 syllables)
tamukuru KAMI no (7)
arebakoso (5)
aki no konoha no (7)
nusa to chirurame (7)

because Princess Tatsuta
has a god to whom she offers brocades,
the leaves of trees
in autumn will scatter as an offering.

1. Orthography Problem

龍田, 立田, 竜田, たつた indicate all same: a place 'Tatsuta' in Nara prefecture!

2. Unit size Problem

Does 卯の花 consist of one word or 卯/の/花 three words?

3. Attribution Problem

Is 卯の花 a name of flower or bean curd refuse?

4. Polysemy/PUN Problem

海松藻 'mirume' a kind of sea weed means 見る目 (human eyes) as well.

Methods

Material: ***Kokinshū*** a.k.a. *Kokinwakashū* is: the first anthology compiled by the order of Emperor Daigo (ca. 905), which contains about 1,100 poems. And 10 sets of **their Comtemporary Japanese Translations (CT)**

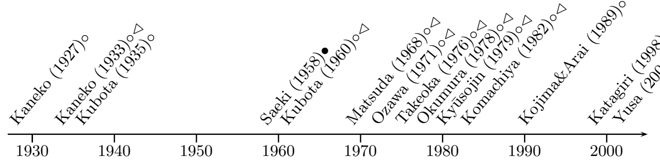

Figure 1: Dates of publication of annotations of the *Kokinshū*: ○ indicates that it has CT; ● indicates that it does not include CT; ▷ indicates that it is used in this project.

Mutual Co-occurrence Rate: Murai (2010)

$$mcr(o,t) = p(o\,|\,t)\ p(t\,|\,o)$$

where, o indicates a token in original texts; t, a token in translation texts; $mcr(o,t)$, the mutual co-occurrence rate; $p(o|t)$, the rate when a token o and t occur at the same time in corresponding texts which are original texts and translation texts.
→ when mcr is large enough, it will be estimated that token o and t are contextually equivalent.

Result

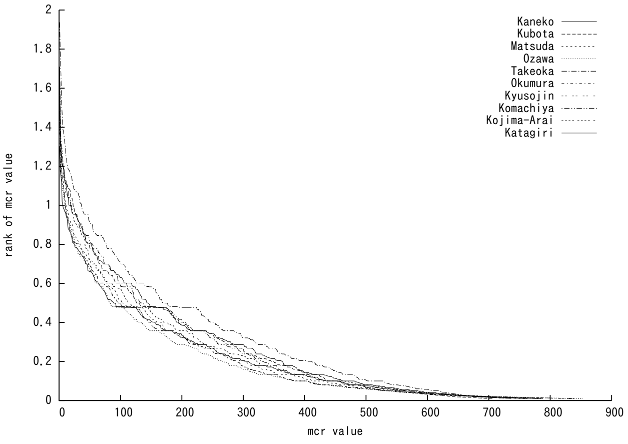

Figure 2: Distribution of Mutual Co-occurrence Rate: original text *Kokinshū* and ten sets of its translation texts.

Good or poor estimated pairs

Table 1: Good estimated pairs and poor estimated pairs; the values of good pairs are the first ten items (over 1.3); and the values of poor pair items are the last ten items (lower 0.01).

no.	good	pairs	poor	pairs
1	鳴く	鳴く cry	異なり	あの
2	風	風 wind	雫	どうして
3	世の中	世の中	此の	この
4	人	人 human	随に	まま
5	春	春 spring	匂ふ	美しい
6	秋	秋 autumn	見る	せい
7	時鳥	時鳥 cuckoo	連れ	つく
8	時鳥	ほととぎす	立ち返る	言う
9	散る	散る fall	有り	つく
10	見る	見る see	有り	まさしく

Conclusion

1. This project has already begun: the parallel corpus of the Kokinshū has been constructed.
2. We are now working on the development of computer software and the optimization of the calculation methods.

Reference

• Murai, Hajime. 2010 Extracting the interpretive characteristics of translations based on the asymptotic correspondence vocabulary presumption method: Quantitative comparisons of Japanese translations of the Bible. Journal of Japan Society of Information and Knowledge Vol. 20, No. 3, 293–310.