An Analysis of the Differences Between Classical and Contemporary Poetic Vocabulary of the Kokinshū

Hilofumi Yamamoto
Tokyo Institute of Technology

Bor Hodošček
Osaka University

1 Introduction

The purpose of the current project is to clarify the relationship between literal (or written) elements and non-literal elements (or connotation) of an ancient language. We will first clarify the differences between the original ancient language and modern language translations of poems in the same literary work, the Kokinshū. In particular, we will examine whether the translations of the Kokinshū use the same words as in a poem (or words corresponding to the modern language) or whether they use words not corresponding to words in a poem. To specify elements written only in the translations, we subtract the elements of original poems (OP: the Kokinshū) from the elements in their contemporary translations (CT), and analyze the residual elements. The differences, therefore, may include two kinds of elements: 1) elements resulting from chronological differences in language; 2) elements added for interpretation. We will subtract the elements of OP from those of CT to account for these differences. While similar attempts have done the subtraction processing manually for the analysis of modern language (Miyazima 1979, 1980, Suzuki 1988, Hasumi 1991), this is the first attempt to subtract a set of linguistic elements from another set by the computer.

2 Methods

We will use the corpus of the Kokinshū by Nakamura et al. (1999). As shown in Figure 1, the poems and the translations are stored as corpora databases and both of them are separated into tokens using the classical poem tokenizer, `kh` (Yamamoto 2007). We convert the tokens into meta-codes, then using the meta-codes, subtract the elements of the original from the elements of their translations. We examine the length of the portion of meta-codes between the two elements.(Figure 2) As an algorithm for matching the elements of CT and OP, we use Longest Common Subsequence (Traum and Habash 2000). An example of subtraction processing with `code2match.c` (Yamamoto 2005, 2009) is shown in Figure 3.

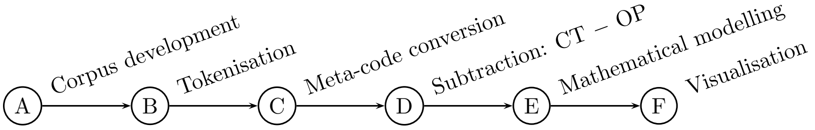

Fig. 1: Flowchart of data processing

Table 1: Summary of the contemporary Japanese translations

translation work (year)	pages	manuscript	method
Kaneko (1933)	1105	Teika	word-for-word
Kubota (1960)	1449	Teika	word-for-word
Matsuda (1968)	1998	Teika	not mentioned
Ozawa (1971)	544	Teika	wording changed
Takeoka (1976)	2278	Teika	word-for-word
Okumura (1978)	434	Teika	intention oriented
Kyūsojin (1979)	1260	Teika	words added
Komachiya (1982)	407	Teika	not mentioned
Kojima and Arai (1989)	483	Teika	not mentioned
Katagiri (1998)	3022	Teika	word-for-word

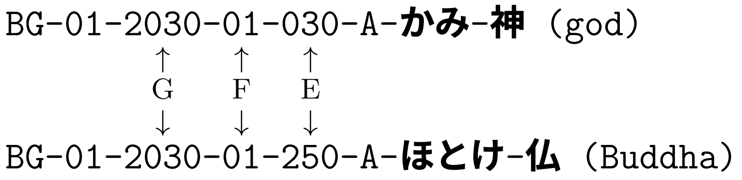
BG-01-2030-01-030-A-かみ-神 (god)
G
F
E
BG-01-2030-01-250-A-ほとけ-仏 (Buddha)

Fig. 2: Level of matching elements: group matching (G); field matching (F); exact matching (E); each level is evaluated by the length of corresponding characters of meta-codes from the first letter.

3 Results

Table 2 indicates a calculation of the components of OP$_{(298)}$. OP$_{(298)}$ refers to a poem by Prince Kanemi. CT$_{(298,\ \text{koma})}$, in turn, refers to the translation of the 298 poem by Teruhiko Komachiya in 1982. 12 elements out of 16 (75 percent) are matched in CT$_{(298,\ \text{koma})}$. One element out of 16 is matched at the field level, and two elements out of 16 are matched at the group level in CT$_{(298,\ \text{koma})}$. One element of OP$_{(298)}$ does not match the elements of CT$_{(298,\ \text{koma})}$. If we assume that matched elements at all the three levels express in CT$_{(298,\ \text{koma})}$, then 15 elements (94 percent) of OP$_{(298)}$ express as the elements in CT$_{(298,\ \text{koma})}$. If we assume that matched elements at all the three levels are expressed in CT(298, koma), then 15 elements (94 percent) of OP(298) are expressed as the elements in CT(298, koma). The remaining 6 percent of elements of

```
 +-------- pair No.
 |  +----- value of matching level, exact=17, field=13, group=10
 |  |  +-- POS No.
 |  |  |
 |  |  | OP element No.+     +- CT element No.
 |  |  | OP element +  |     |  CT element
 |  |  |            |  |     |  |
 1 17 11 *tatsutahime 00 <-> 12 *Tatsutahime (pn.Tatsutahime)
 2 17 47           te 04 <-> 25 te           (hand)
 3 17 47       mukeru 05 <-> 26 mukeru       (toward)
 4 17  2         kami 06 <-> 32 kami         (god)
 5 10 61           no 07 <-> 33 ga           (SUB)
 6 17 47          ari 08 <-> 34 aru          (be)
 7 10 64           ba 09 <-> 35 kara         (because)
 8 17 65         koso 11 <-> 36 koso         (EM)
 9 17  2          aki 12 <-> 38 aki          (autumn)
10 17 71           no 13 <-> 39 no           (CON)
11 17  2       konoha 14 <-> 40 konoha       (leaf of tree)
12 17  2         nusa 19 <-> 45 nusa         (present)
13 17 61           to 20 <-> 46 to           (CRD)
14 17 47        chiru 21 <-> 49 chiru        (fall)
15 13 74         ramu 22 <-> 54 u            (CJR)
```

Fig. 3: An example of the alignment of the matched elements between OP(298) and CT(298, koma). Each line consists of the matched pair ID number (1), the matching level indicated by the value (17), ID number of POS (11) which indicates a place name, OP element (`*tatsutahime`), ID number of OP element, ID number of CT element, CT element (`*Tatsutahime`), and the glossary; * written in different kanji.

Table 2: Result of subtracting the elements of OP(298) from those of CT(298, koma): it indicates the ratio of the ingredients of OP(298).

OP (valid number of element)	= 16
E (ratio of exact match)	12/16 = 0.750
F (ratio of field match)	1/16 = 0.062
G (ratio of group match)	2/16 = 0.125
T (ratio of total match)	15/16 = 0.938
U (ratio of unmatched OP)	1 - T = 0.062

OP(298) do not match against any elements in CT(298, koma). None of the ten modern language translations could be fully expressed with the ancient language. The amount of added information was 80 percent higher than the original.(Table 4) The differences between the theoretical and experimental values were at most 8 percent. Those were rare cases, and in general accounted for around 4 percent.

4 Discussion

Based on the analysis of the differences between the two, we assume that translators attempted to express some cultural elements unfamiliar to modern people. Table 3 is an example of a calculation which indicates the components of CT(298, koma). CT(298, koma) uses the same 12 elements as OP(298). The total number of elements of CT(298, koma) is 41; thus 29 percent of CT(298, koma) is calculated as the component of OP(298). The rest of CT(298, koma), 71 percent, is considered as added annotated text. Ratio `A`, however, does not consist only of newly added components: it should be deconstructed into three kinds of components: 1) the first level of the paraphrased component, `P1`, which can be estimated from the ratio of the elements of the field match `F` and the group match `G`; 2) the second level of the paraphrased component, `P2`, which can be estimated from the ratio of the unmatched elements, since even unmatched elements are assumed to be somehow translated into CT; and 3) the purely added component, `D`, which can be estimated from the ratio of the annotation minus `P1` and `P2`.(Figure 4)

If the estimation from the subtraction of elements of OP from those of CT is correct, the practical value, `D`, can be close to the theoretical value, `H`, and the validity of the

Table 3: Component of CT in case of KKS *298* by Komachiya (1982): `fabs(D-H)` stands for the function of the absolute value of the practical value, `D`, minus the theoretical value, `H`.

CT	(valid number of element)		=	41	
W	(ratio of original word use)	12/41	=	0.293	(E/CT)
A	(ratio of annotation)	1-0.293	=	0.707	(1-W)
	---breakdown of the annotation---				
	P1(ratio of FG paraphrased)	(0.62+0.12)/0.707	=	0.073	(F+G)/A
	P2(ratio of U paraphrased)	(0.707-0.073)*0.062	=	0.040	(A-P1)*U
	D (ratio of purely added)	0.707-(0.073+0.040)	=	0.595	A-(P1+P2)
H	(theoretical value of D)	1-16/41	=	0.610	1-OP/CT
Gap		fabs(0.595-0.610)	=	0.015	fabs(D-H)

Table 4: Amount of added information (N=1000)

		alignment			subtraction		
	translator	min.	mean (SD)	max.	min.	mean (SD)	max.
1	Kaneko	0.16	0.53 (0.09)	0.80	0.18	0.49 (0.09)	0.73
2	Katagiri	0.21	0.49 (0.08)	0.71	0.16	0.44 (0.08)	0.68
3	Kojima Arai	0.15	0.46 (0.09)	0.74	0.10	0.41 (0.10)	0.69
4	Komachiya	0.12	0.44 (0.08)	0.72	0.11	0.39 (0.08)	0.67
5	Kubota	0.15	0.45 (0.09)	0.77	0.13	0.40 (0.09)	0.72
6	Kyusojin	0.10	0.47 (0.08)	0.73	0.11	0.42 (0.08)	0.69
7	Matsuda	0.00	0.44 (0.09)	0.77	0.07	0.39 (0.09)	0.69
8	Okumura	0.06	0.44 (0.08)	0.75	0.11	0.41 (0.08)	0.72
9	Ozawa	0.10	0.46 (0.08)	0.72	0.20	0.44 (0.07)	0.70
10	Takeoka	0.11	0.42 (0.10)	0.74	0.06	0.38 (0.10)	0.69
	mean	0.12	0.46 (0.03)	0.74	0.12	0.42 (0.03)	0.70

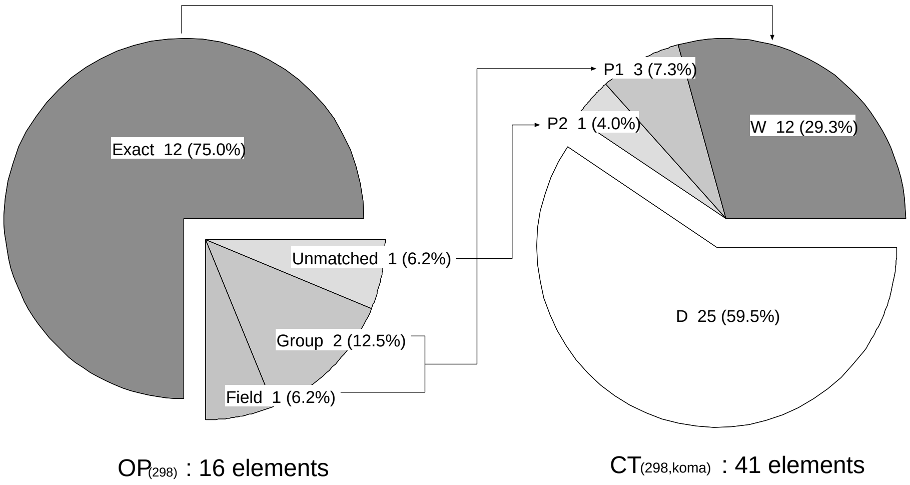

FIG. 4: Pie-charts illustrating the components of OP(298) and CT(298, koma): the ratio of purely added components is estimated based on the number of elements in common in OP and CT.

operation will be supported. In the case of the values between OP(298) and CT(298, koma), the theoretical value is 0.610, the practical value is 0.595, and their discrepancy is 0.015, which means the two values are very close.

5 Conclusion

The current paper discussed the differences between the original poems of the Kokinshū and its translations. We attempted to classify the components of both OP and CT to examine whether or not CT includes added elements, which are the non-literal elements of OP. After subtracting the matched elements between OP and CT from CT, the presence of a residual indicated that CT includes newly added elements. It shows that it is impossible to convert the contents in the ancient language into only their equivalents in the modern language.

References

Hasumi, Yoko (1991) "Dōitsu jōhō ni motozuku bunshōhyōgen ni tsuite no bunseki / Difference of expressions on the same information", *Mathematical Linguistics*, Vol. 18, No. 3, pp. 136–144.

Kaneko, Motoomi (1933) *Kokinwakashū Hyōshaku: Shōwa Shimban*, Tokyo: Meijishoin.

Katagiri, Yoichi (1998) *Kokinwakashū Hyōshaku Jō, Chū, Ge*, Tokyo: Kodansha.

Kojima, Noriyuki and Eizō Arai (1989) *Kokinwakashū*, Vol. 5 of Shin-Nihon bungaku taikei (A new collection of Japanese literature), Tokyo: Iwanami shoten.

Komachiya, Teruhiko (1982) *Gendaigo yaku taishō Kokinwakashū (Kokinwakashū with modern Japanese translations)*, Obunsha Bunko Taiyaku Koten Series, Tokyo: Ōbunsha.

Kubota, Utsubo (1960) *Kokinwakashū Hyōshaku (Vol. 1, 2, 3)*, Tokyo: Tokyodo shuppan.

Kyūsojin, Hitaku (1979) *Kokinwakashū Zen'yaku Chū (Comprehensive annotations of the Kokinwakashū)*, Vol. 1–5 of Kodansha Gakujutsu Bunko: Kodansha.

Matsuda, Takeo (1968) *Shinshaku Kokinwakashū Vols.1 and 2*, Tokyo: Kazama Shobo.

Miyazima, Tatuo (1979) ""Kyōsantō Sengen" no yakugo (Translated terms in the "Communist Manifesto")", in Gengogaku Kenkyūkai ed. *Gengo no Kenkyū (Study of language)*, Tokyo: Mugi Shobo, pp. 425–517.

——— (1980) "'Jodōshi' to 'Hojodōshi' (Auxiliary verbs and subsidiary verbs)", in Society of Modern Language ed. *Kindaigo kenkyū (Study of contemporary vocabulary)*, Vol. 6, Tokyo: Musashino shoin, pp. 455–468.

Nakamura, Yasuo, Yoshihiko Tachikawa, and Mayuko Sugita (1999) *Kokubungaku kenkyū shiryōkan dētabēsu koten korekushon "Nijūichidaishū" Shōhobanbon CD-ROM (Database Collection by National Institute of Japanese Literature "Nijūichidaishū" the Shōho edition CD-ROM)*: Iwanami Shoten.

Okumura, Tsuneya (1978) *Kokinwakashū*, Shinchō Nihon Koten Shūsei, Tokyo: Shinchō sha.

Ozawa, Masao (1971) *Kokinwakashū*, Vol. 7 of Nihon Koten Bungaku Zenshū, Tokyo: Shōgakkan.

Suzuki, Tai (1988) "Weirando "Shūshinron" no Kanji (Kanji in the "Elements of moral science" by Francis Wayland)", in *Gengo no Kenkyū (Study of language)*, Vol. 8 of Kindai Nihongo to Kanji (Contemporary Japanese and Kanji), Tokyo: Meijishoin, pp. 128–164.

Takeoka, Masao (1976) *Kokinwakashū Zen Hyōshaku Jō Ge (the complete annotated edition of Kokinwakashū, Vols. 1 and 2)*, Tokyo: Yubun Shoin.

Traum, David and Nizar Habash (2000) "Generation from lexical conceptual structures", in *NAACL-ANLP 2000 Workshop on Applied interlinguas*, pp. 52–59, Morristown, NJ, USA: Association for Computational Linguistics.

Yamamoto, Hirofumi Hilo (2005) "A Mathematical Analysis of the Connotations of Classical Japanese Poetic Vocabulary", Ph.D. dissertation, Australian National University.

Yamamoto, Hilofumi (2007) "Waka no tame no Hinshi tagu zuke shisutemu / POS tagger for Classical Japanese Poems", *Nihongo no Kenkyu / Studies in the Japanese Language*, Vol. 3, No. 3, pp. 33–39.

——— (2009) "Thesaurus for the Hachidaishu (ca. 905–1205) with the classification codes based on semantic principles", *Nihongo no Kenkyu / Studies in the Japanese Language*, Vol. 5, No. 1, pp. 46–52.