

# Discovering Characteristic Patterns from Collections of Classical Japanese Poems

Mayumi Yamasaki<sup>1</sup>, Masayuki Takeda<sup>1</sup>,  
Tomoko Fukuda<sup>2</sup>, and Ichirō Nanri<sup>3</sup>

<sup>1</sup> Department of Informatics, Kyushu University 33, Fukuoka 812-8581, Japan

<sup>2</sup> Fukuoka Jo Gakuin College, Ogōri 838-0141, Japan

<sup>3</sup> Junshin Women's Junior College, Fukuoka 815-0036, Japan  
{yamasaki, takeda}@i.kyushu-u.ac.jp

**Abstract.** WAKA is a form of traditional Japanese poetry with a 1300-year history. In this paper, we attempt to discover characteristics common to a collection of WAKA poems. As a formalism for characteristics, we use regular patterns where the constant parts are limited to sequences of auxiliary verbs and postpositional particles. We call such patterns FUSHI. The problem is to find automatically significant FUSHI patterns that characterize the poems.

Solving this problem requires a reliable significance measure for the patterns. Brāzma et al. (1996) proposed such a measure according to the MDL principle. Using this method, we report successful results in finding patterns from five anthologies. Some of the results are quite stimulating, and we hope that they will lead to new discoveries. Based on our experience, we also propose a pattern-based text data mining system. Further research into WAKA poetry is now proceeding using this system.

## 1 Introduction

WAKA is a form of traditional Japanese poetry with a 1300-year history. Most WAKA poems are in the form of TANKA, namely, they have five lines and thirty-one syllables, arranged thus: 5-7-5-7-7. This poetry was usually composed in momentary flashes of inspiration. Most frequently, it was used as a subtle means of communication between lovers and friends, and was therefore an important part of daily life in ancient Japan. WAKA poetry has been central to the history of Japanese literature, and has been studied extensively by many scholars. Recently, since an accumulation of about 450,000 WAKA poems became available in a machine-readable form, it is expected that computers will play an important role in research into WAKA poetry.

In this paper, we focus on the problem of discovering characteristics common to a collection of WAKA poems. As a formalism for characteristics, we use the class of *pattern languages* introduced by D. Angluin [2]. One of the most important subclasses is the class of *regular pattern languages*, in which each variable symbol appears only once. This subclass is sufficiently rich from a practical viewpoint [9]. To characterize WAKA poems, we limit the constant parts of regular

patterns to sequences of *adjuncts*, i.e., auxiliary verbs and postpositional particles. For example, we consider patterns such as  $*BA*ZARAMASHIO*$ , where BA is a postpositional particle, and ZARAMASHIO is a chain of two auxiliary verbs ZARA and MASHI and a postpositional particle O. This pattern corresponds to the subjunctive mood. We call such patterns FUSHI, and call their constant parts *adjunct sequences*. This limitation of the constant parts to adjunct sequences is essential in our characterization. It should be noted that a Japanese sentence is a sequence of *segments*, each of which consists of a word and its subsequent adjuncts. The FUSHI pattern is a reliable model for techniques used in composing WAKA poems, and is closely related to their structure and rhythm.

The goal of this paper is to discover, more or less automatically, significant FUSHI patterns that characterize a given set of WAKA poems, and then report some features that may be due to the times or to the poets' personalities. The difficulties are summarized as follows.

- To identify adjuncts appearing in WAKA poems.
- To give an appropriate definition of the *significance* of FUSHI patterns.

Since we have no delimiters between segments in the Japanese language, the first item requires morphological analysis of WAKA poems. However, many ambiguities, which are not easy to resolve, will arise during this analysis. In this paper, we assume that any substring identical to an adjunct is the adjunct, and therefore we have only to perform pattern matching. This assumption simplifies the discussion: Let  $\Sigma$  be an alphabet, and let  $*$   $\notin \Sigma$  be the gap symbol. A *pattern* is a nonempty string over  $\Sigma \cup \{*\}$ . We say that a pattern  $p$  *matches* a string  $w$  in  $\Sigma^+$  if  $w$  is obtained by substituting strings in  $\Sigma^*$  for occurrences of  $*$  in  $p$ , respectively. A set  $\Pi$  of patterns is said to be a *covering* of a set  $S$  of strings in  $\Sigma^+$  if, for any string in  $S$ , at least one pattern in  $\Pi$  exists that matches it.

Let  $C \subseteq \Sigma^+$  be a set of *adjunct sequences*. A FUSHI pattern is a pattern of the form  $*\alpha_1*\alpha_2*\cdots*\alpha_h*$ , where  $h > 1$  and  $\alpha_1, \alpha_2, \dots, \alpha_h \in C$ . Our problem is then defined as follows.

Given a finite set  $S$  of strings in  $\Sigma^+$ , find the most *significant* covering  $\Pi$  of  $S$  that consists of FUSHI patterns.

Note that, in general,  $S$  has infinitely many coverings. If we have an appropriate definition of the significance of such coverings, then we can determine the most significant covering according to the definition. However, the problem remains of defining the significance appropriately. One such definition was given by Arimura et al. [3]. A *k-minimal multiple generalization* (abbreviated to *k-mmng*) of a set  $S$  of strings is defined as a minimally general set that is a covering of  $S$  containing at most  $k$  patterns. They showed in [3] that *k-mmng* is optimal from the viewpoint of inductive inference from positive data based on identification in the limit [7]. However, we are faced with the following difficulties: (a) we must give an integer  $k$  as an upper bound of the size of coverings in advance; (b) a polynomial-time algorithm exists for finding a *k-mmng*, but it is impractical in the sense that the time complexity will be very large for a relatively large value

of  $k$ ; (c) since a set  $S$  of strings has more than one  $k$ -mmg, we need a criterion to choose an appropriate one.

Another definition was proposed by Br̄azma et al. [5]. The most significant covering of a set  $S$  is defined as the most probable collection of patterns likely to be present in the strings in  $S$ , assuming some simple, probabilistic model. This criterion is equivalent to Rissanen's Minimum Description Length (MDL) principle [8]. According to the MDL principle, the most significant covering is the one that minimizes the sum of the length (in bits) of the patterns and the length (in bits) of the strings when encoded with the help of the patterns. Since finding the optimal solution is NP-hard, a polynomial-time algorithm for approximating the optimal solution within a logarithmic factor is presented.

In this paper, we use the MDL principle to define the significance of FUSHI patterns, and apply the method developed by Br̄azma et al. to the problem of finding significant patterns from a set of WAKA poems. The main contributions of this paper are summarized as follows.

1. A new schema is presented in which the constant parts of regular patterns are restricted to strings in a set  $C$ . Allowing  $C$  to be the set of adjunct sequences yields a reliable model for characterizing WAKA poems.
2. A new grammatical scheme of Japanese language is given as the basis of the above characterization. This scheme is far from standard, but constitutes a simple and effective tool.
3. Successful results of our experiment of finding patterns from five anthologies are reported, some of which are very suggestive. We hope that they will lead to new areas of research. The significance measure for patterns based on the MDL principle is proved to be useful.
4. A text data mining system is proposed, which consists of a *pattern matching part* and a *pattern discovery part*. Using this system, further research into WAKA poetry is now proceeding.

It should be emphasized that the FUSHI pattern of a WAKA poem is not conclusively established even when determined by non-computer-based efforts. The determined pattern may vary according to the particular interests of scholars, and to the other poems used for comparison. Our purpose is to develop a method for finding a set of significant patterns, some of which may give a scholar clues for further investigation. Similar settings can be found in the field of data mining. In other words, our goal is to develop a text data mining system to support research into WAKA poetry.

Unlike data mining in relational databases, data mining in texts that are written in natural language requires preprocessing based on natural language processing techniques with some domain knowledge, and such techniques and knowledge are the key to success [1,6]. In our case, no such techniques are needed; the knowledge used here is merely the set of adjunct sequences, which are allowed to appear in FUSHI patterns as constant parts. It may be relevant to mention that the third author is a WAKA researcher, the fourth author is a linguist in Japanese language, and the first and the second authors are researchers in computer science.

## 2 Method

This section shows why the FUSHI pattern can be a reliable model for characterizing WAKA poems, and presents our grammatical scheme on which the characterization is based.

### 2.1 FUSHI Pattern as a Model of Characteristics

Consider the problem of finding characteristics from a collection of WAKA poems. Most studies on this problem have been undertaken mainly from the viewpoint of preference for KAGO words. By KAGO, we mean the nouns, verbs and adjectives used in WAKA poems<sup>1</sup>. It might be considered that WAKA poems commonly containing some KAGO words are on the same subject matter: in other words, such characterization corresponds to similarity of subject matter. For example, Ki no Tsurayuki (ca. 872–945), one of the greatest of the early court poets, composed many WAKA poems on the cherry blossom. It is, however, simplistic thinking to conclude that the poet had a preference for cherry blossoms. The reason for this is that court poets in those days were frequently given a theme when composing a poem. We must assume that the poets could not freely use favored words.

What then should be considered as characteristics of WAKA poems? The three poems shown in Fig. 1 are very famous poems from the imperial anthology SHINKOKINSHŪ. These poems were arranged by the compilers in one section of the anthology, and are known as *the three autumn evening poems* (SANSEKI NO UTA). All the poems express a scene of autumn evening. However, the reason why the poems are regarded as outstanding poems on autumn evening is that they all used the following techniques:

1. Each poem has two parts: the first three lines and the remaining two lines.
2. The first part ends with the auxiliary verb KERI.
3. The second part ends with a noun.

Such techniques are basically modeled by FUSHI patterns, regular patterns in which the constant parts are limited to adjunct sequences.

WAKA poetry can be compared to IKEBANA, the traditional Japanese flower arrangement. The art of IKEBANA relies on the choice and combination of both *materials* and *containers*. Limitation on the choice of materials, that is, KAGO words, probably forced the poets to concentrate on the choice of containers, FUSHI patterns. The FUSHI pattern is thus a reliable model for characterizing WAKA poems.

---

<sup>1</sup> The term KAGO consists of two morphemes KA and GO, which mean ‘poem’ and ‘word’, respectively; therefore, it means words used in poetry rather than in prose. However, here, we mean the nouns, verbs and adjectives used in WAKA poems.

---

#361 (Priest Jakuren)	
SABISHISA WA	<i>One cannot ask loneliness</i>
SONO IRO TO SHI MO	<i>How or where it starts.</i>
NAKARI KERI	<i>On the cypress-mountain,</i>
MAKI TATSU YAMA NO	<i>autumn evening.</i>
AKI NO YŪGURE.	
#362 (Priest Saigyō)	
KOKORO NAKI	<i>A man without feelings,</i>
MI NI MO AWARE WA	<i>Even, would know sadness</i>
SHIRA RE KERI	<i>When snipe start from the marshes</i>
SHIGI TATSU SAWA NO	<i>On an autumn evening.</i>
AKI NO YŪGURE.	
#363 (Fujiwara no Teika)	
MIWATASE BA	<i>As far as the eye can see,</i>
HANA MO MOMIJI MO	<i>No cherry blossom,</i>
NAKARI KERI	<i>No crimson leaf;</i>
URA NO TOMAYA NO	<i>A thatched hut by a lagoon,</i>
AKI NO YŪGURE.	<i>This autumn evening.</i>

---

**Fig. 1.** The three autumn evening poems from SHINKOKINSHŪ; blank symbols are placed between the words for readability. English translations are from [4].

## 2.2 Our Grammatical Scheme

In the standard framework of Japanese grammar, words are divided into two categories: independent words (or simply, *words*) and dependent words (or *adjuncts*). The former is a category of nouns, verbs, adjectives, adverbs, conjunctions and interjections, while the latter are auxiliary verbs and postpositional particles. A Japanese sentence is a sequence of segments, and each segment consists of a word and its subsequent adjuncts. Verbs, adjectives, and auxiliary verbs can be conjugated. It should be noted that most of the conjugated suffixes of verbs and adjectives are identical to some auxiliary verbs or to their conjugated suffixes. If we regard the conjugated suffixes of words as adjuncts, a segment can be viewed as a word stem and its subsequent adjuncts. This stem-adjunct scheme is far from being standard grammar, but it does constitute a simple and effective tool for our purposes.

We can see that the FUSHI pattern \*REBA\*KOSO\*KERE\* is common in the poems in Figure 2. The occurrences of BA and KOSO in these poems are postpositional particles. However the occurrences of RE and KERE have more than one grammatical category in the standard grammar. In fact, the occurrence of RE in each of the first two poems is a conjugated suffix of a verb, while RE in the last poem is a conjugated suffix of an auxiliary verb. The occurrence of KERE in each of the first two poems is a conjugated suffix of an adjective, while KERE in the last poem is an auxiliary verb. Although the occurrences of the FUSHI pattern in the three poems are thus different, we intend to treat them as if they were the same. Finding FUSHI patterns requires a new grammatical scheme that is

different from the standard; our stem-adjunct scheme is appropriate. As stated previously, the strings appearing as the constant parts of FUSHI patterns are called adjunct sequences. Although an adjunct sequence consists of one or more adjuncts, we treat it as an indivisible unit.

### 3 Optimal Covering Based on MDL Principle

This section presents the definition of optimal covering based on the MDL principle proposed by Brāzma et al. [5], and then shows an algorithm for approximating the optimal covering.

#### 3.1 Definition

Let us denote by  $L(\pi)$  the set of strings a pattern  $\pi$  matches. Consider a pattern

$$\pi = * \beta_1 * \cdots * \beta_n * \quad (\beta_1, \dots, \beta_n \in \Sigma^+)$$

and a set  $B = \{\alpha_1, \dots, \alpha_n\}$  of strings such that  $B \subseteq L(\pi)$ . The set  $B$  can be described by the pattern  $\pi$  and the strings

$$\begin{array}{c} \gamma_{1,0} \gamma_{1,1} \cdots \gamma_{1,h} \\ \gamma_{2,0} \gamma_{2,1} \cdots \gamma_{2,h} \\ \vdots \\ \gamma_{n,0} \gamma_{n,1} \cdots \gamma_{n,h} \end{array}$$

such that  $\alpha_i = \gamma_{i,0} \beta_1 \gamma_{i,1} \cdots \gamma_{i,h-1} \beta_h \gamma_{i,h}$  for  $i = 1, \dots, n$ . Such description of  $B$  is called *the encoding by pattern  $\pi$* . We denote by  $\|\alpha\|$  the description length of a string  $\alpha$  in some encoding. For simplicity, we ignore the delimiters between strings. The description length of  $B$  is

$$\|\pi\| + \sum_{i=1}^n \sum_{j=0}^h \|\gamma_{i,j}\|.$$

---

KOKINSHŪ #193	(Ōe no Chisato)
	TSUKI <u>MIRE BA</u> / CHIJI NI MONO <u>KOSO</u> / KANASHI <u>KERE</u> /
	WAGA-MI HITOTSU NO / AKI NI WA ARA NE DO.
GOSENSHŪ #739	(Daughter of Kanemochi no asom)
	YŪSARE <u>BA</u> / WAGA-MI NOMI <u>KOSO</u> / KANASHI <u>KERE</u> /
	IZURE NO KATA NI / MAKURA SADAME M.
SHŪISHŪ #271	(Minamoto no Shitagō)
	OI <u>NURE BA</u> / ONAJI KOTO <u>KOSO</u> / SE RARE <u>KERE</u> /
	KIMI WA CHIYO MASE / KIMI WA CHIYO MASE.

---

**Fig. 2.** WAKA poems containing pattern \*REBA\*KOSO\*KERE\*.

Let us denote by  $c(\pi)$  the string obtained from  $\pi$  by removing all  $*$ 's. Assuming some symbolwise encoding, we have

$$\|\alpha_i\| = \sum_{j=1}^h \|\gamma_{i,j}\| + \|c(\pi)\|.$$

The description length of  $B$  is then

$$\|\pi\| + \sum_{i=1}^n (\|\alpha_i\| - \|c(\pi)\|) = \sum_{i=1}^n \|\alpha_i\| - \left( \|c(\pi)\| \cdot |B| - \|\pi\| \right).$$

Let  $A$  be a finite set of strings. A finite set  $\Omega = \{(\pi_1, B_1), \dots, (\pi_k, B_k)\}$  of pairs of a pattern  $\pi_i$  and a subset  $B_i$  of  $A$  is also said to be a *covering* of  $A$  if:

- $B_i \subseteq L(\pi_i)$  ( $i = 1, \dots, k$ ).
- $A = B_1 \cup \dots \cup B_k$ .
- $B_1, \dots, B_k$  are disjoint.

When the set  $B_i$  is encoded by  $\pi_i$  for each  $i = 1, \dots, k$ , the description length of  $A$  is

$$M(\Omega) = \sum_{i=1}^n \|\alpha_i\| - C(\Omega),$$

where  $C(\Omega)$  is given by

$$C(\Omega) = \sum_{j=1}^k \left( \|c(\pi_j)\| \cdot |B_j| - \|\pi_j\| \right).$$

Now, the *optimal covering* of the set  $A$  is defined to be the covering  $\Omega$  minimizing  $M(\Omega)$ , or to be the set of patterns in it. Minimizing  $M(\Omega)$  is equivalent to maximizing  $C(\Omega)$ .

### 3.2 Detail of Encoding

In the above definition, the optimal covering varies depending on the encoding method. In the coding scheme in [5], the patterns and the strings for substitutions are coded together with delimiter symbols in some optimal symbolwise coding with respect to a probability distribution. Therefore, the formula of  $C(\Omega)$  contains parameters that are the occurring probabilities of the delimiter symbols and the gap symbol  $*$ .

However, in our case the strings to be coded are of length less than  $m = 32$  because we deal with only the poems consisting of thirty-one syllables. So, we can choose a simple way. We shall describe a string  $w \in \Sigma^*$  as the pair of the length of  $w$  and the bit-string representing  $w$  in some optimal coding with respect to a probability distribution  $P$  over  $\Sigma$ . In practice, we can take  $P(a)$  proportional to the relative frequency of a symbol  $a \in \Sigma$  in a database. We denote by  $\ell_P(w)$

the length of the bit-string representing  $w$ . We also denote by  $n_*(\pi)$  the number of occurrences of  $*$  in a pattern  $\pi$ . Assuming a positive number  $m$  such that  $|w| < m$ , we have

$$C(\Omega) = \sum_{j=1}^k \left( u(\pi_j) \cdot |B_j| - v(\pi_j) \right),$$

where

$$\begin{aligned} u(\pi) &= \ell_P(c(\pi)) - n_*(\pi) \log_2 m + \log_2 m, \\ v(\pi) &= \ell_P(c(\pi)) + n_*(\pi) \log_2 m. \end{aligned}$$

### 3.3 Approximation Algorithm

Since the problem of finding the optimal covering of a set of strings contains as a special case the set covering problem, it is NP-hard. Brāzma et al. [5] modified the problem as below:

Given a finite set  $A$  of strings and a finite set  $\Delta$  of patterns, find a covering  $\Omega$  of  $A$  in which patterns are chosen from  $\Delta$  that minimizes  $M(\Omega)$ .

They presented a greedy algorithm that approximates the optimal solution. It computes the values of

$$u(\pi) - \frac{v(\pi)}{|L(\pi) \cap U|}$$

for all possible patterns  $\pi$  at each iteration of a loop, and selects the pattern maximizing it to the covering. Here  $U$  is the set of strings in  $A$  not covered by any pattern that has already been selected. The value of  $M(\Omega)$  for an approximate solution  $\Omega$  obtained by this algorithm is at most  $\log_2 |A|$  times with respect to the optimal one. The time complexity of the algorithm is  $O(|\Delta| \cdot |A| \cdot \log_2 |A|)$  when excluding the computation of  $\{(\pi, L(\pi) \cap A) \mid \pi \in \Delta\}$ , which requires  $O\left(\sum_{\pi \in \Delta} |\pi| + |\Delta| \cdot \sum_{\alpha \in A} |\alpha|\right)$  time to perform the pattern matching between the patterns in  $\Delta$  and the strings in  $A$ .

## 4 Finding FUSHI Patterns from WAKA Poems

This section describes our experiment of seeking FUSHI patterns in a collection of WAKA poems. To apply the algorithm of Brāzma et al. to our problem, we need a way of identifying adjuncts with less misdetections and a formal definition of adjunct sequences, which are presented in Sections 4.1 and 4.2. Successful results of the experiment are then shown in Section 4.3.

### 4.1 How to Avoid Misdetection of Adjuncts

Since we identify a string that matches an adjunct with the adjunct, there can be many misdetections. To avoid such misdetections, we adopted the following techniques.

- We restricted ourselves to the adjunct sequences appearing at the end of lines of WAKA poems. Obviously, an adjunct sequence can appear in the middle of a line when the line contains more than one segment. However, most of the important adjunct sequences related to FUSHI patterns appear at the end of lines.
- A WAKA poem was written in a mixture of Chinese characters and KANA characters. The former are ideograph characters whereas the latter are syllabic characters. An equivalent written in only KANA characters is attached to every WAKA poem in our database. Suppose that a line of one poem is equivalent to a line of another poem. If we use the one having a shorter KANA string at the end of line, then misdetection of an adjunct will be decreased because adjuncts are written in KANA characters. Based on this idea, we replaced each line of the poem by its ‘canonical’ form.

Although we cannot avoid all misdetections of adjuncts, our system is adequate for finding FUSHI patterns.

### 4.2 Definition of Adjunct Sequences

In our setting, the class of FUSHI patterns is defined by giving a set  $C$  of adjunct sequences. We therefore need a formal definition of adjunct sequences. For the definition, we give grammatical rules about concatenation of adjuncts. An adjunct sequence can be divided into three parts: first, a conjugated suffix of verb, adjective, or auxiliary verbs; second, a sequence of auxiliary verbs; third, a sequence of postpositional particles. Let us denote a conjugated suffix, an auxiliary verb and a postpositional particle by  $Suf$ ,  $AX$ , and  $PP$ , respectively. There are syntactic and semantic constraints in concatenation of  $Suf$ ,  $AX$ , and  $PP$ . The syntactic constraint is relatively simple, and is easy to describe. It depends on the combination of a word itself and the conjugated form of the preceding word. On the other hand, the semantic constraint is not so easy to describe completely. Here, we consider only the constraint between  $AX$  and  $AX$ , and between  $PP$  and  $PP$ . We classified the category of  $AX$  into five subcategories, and developed rules according to the classification. We also classified  $PP$  into six subcategories, and applied rules in a similar way. The set  $C$  of adjunct sequences was thus defined.

### 4.3 Experimental Results for Five Anthologies

We applied the algorithm to five anthologies : KOKINSHŪ, SHINKOKINSHŪ, MINISHŪ, SHŪIGUSŌ, and SANKASHŪ. See Table 1. The first two are imperial anthologies, i.e., anthologies compiled by imperial command, the first completed in 922,

**Table 1.** Five collections of WAKA poems.

<i>Anthology</i>	<i>Explanation</i>	<i># poems</i>
KOKINSHŪ	Imperial anthology compiled in 922	1,111
SHINKOKINSHŪ	Imperial anthology compiled in 1205	2,005
MINISHŪ	Private anthology by Fujiwara no Ietaka (1158–1237)	3,201
SHŪIGUSŌ	Private anthology by Fujiwara no Teika (1162–1241)	2,985
SANKASHŪ	Private anthology by Priest Saigyō (1118–1190)	1,552

and the second in 1205. The differences between the two anthologies, if any exist, may be due to the time difference in compilation. On the other hand, the others are private anthologies of poems composed by the three contemporaries: Fujiwara no Ietaka (1158–1237), Fujiwara no Teika (1162–1241), and the priest Saigyō (1118–1190). Their differences probably depend on the poets' personalities.

Table 2 shows the results of the experiments. A great number of patterns occur in each anthology, and therefore it is impossible to examine all of them manually. In the second column, the values in parentheses are the numbers of patterns occurring more than once. In the experiment, we used these sets of patterns as  $\Delta$ , the sets of candidate patterns. The size of coverings is shown in the third column. For example, 191 of 8,265 patterns were extracted from KOKINSHŪ. The coverings are relatively small in order to examine all the patterns within them.

**Table 2.** Coverings of five anthologies.

<i>Anthology</i>	<i># occurring patterns</i>	<i>Size of covering</i>
KOKINSHŪ	164,978 (8,265)	191
SHINKOKINSHŪ	233,187 (12,449)	270
MINISHŪ	187,014 (16,425)	369
SHŪIGUSŌ	214,940 (14,365)	335
SANKASHŪ	279,904 (12,963)	232

Table 3 shows the first five patterns emitted by the algorithm from KOKINSHŪ. The first pattern \*KEREBABERANARI\* contains the auxiliary verb BERANARI, which is known to be used mainly in the period of KOKINSHŪ. The fourth pattern \*RISEBARAMASHI\* corresponds to the subjunctive mood. The last pattern \*WANARIKERI\* corresponds to the expression “I have become aware of the fact that . . .”. The remaining two patterns are different correlative word expressions, called KAKARI-MUSUBI. The obtained patterns are thus closely related to techniques used in composing poems.

Next, we shall compare pattern occurrences in the five anthologies. Table 4 shows the first five patterns for each anthology, where each numeral denotes

**Table 3.** FUSHI patterns from KOKINSHŪ.

<i>Pattern</i>	<i>Annotation</i>
*KEREBA*BERANARI*	use of auxiliary verb BERANARI
*ZO*SHIKARIKERU*	correlative word expression (KAKARI-MUSUBI)
*KOSO*RIKERE*	correlative word expression (KAKARI-MUSUBI)
*RISEBA*RAMASHI*	the subjunctive mood
*WA*NARIKERI*	the expression of awareness

the occurring frequency of the pattern in the anthology. The following facts, for example, can be read from Table 4:

1. Pattern \*BAKARI\*RAM\* does not occur in either KOKINSHŪ or SHINKOKINSHŪ.
2. Pattern \*WA\*NARIKERI\* occurs in each of the anthologies. In particular, it occurs frequently in SANKASHŪ.
3. Pattern \*MASHI\*NARISEBA\* occurs frequently in SANKASHŪ.
4. Pattern \*KOSO\*RIKERE\* does not occur in SHŪIGUSŌ.
5. Pattern \*YA\*RURAM\* occur in each anthology except KOKINSHŪ.

It is possible that the above facts are important characteristics that may be due to the times or the poets' personalities. For example, (2) and (3) may reflect Priest Saigyō's preferences, and (5) may imply that the pattern \*YA\*RURAM\* was not preferred in the period of KOKINSHŪ. Comparisons of the obtained patterns and their frequencies thus provide a WAKA researcher clues for further investigation.

**Table 4.** FUSHI patterns from five anthologies with frequencies, where A, B, C, D and E denote KOKINSHŪ, SHINKOKINSHŪ, MINISHŪ, SHŪIGUSŌ, and SANKASHŪ, respectively.

<i>Patterns</i>	A	B	C	D	E	<i>Patterns</i>	A	B	C	D	E
*KEREBA*BERANARI*	5	0	0	0	0	*BAKARI*RAM* <sup>1</sup>	0	0	11	8	3
*ZO*SHIKARIKERU*	8	1	0	0	3	*NO*NARIKERI*	19	30	39	19	49
A *KOSO*RIKERE* <sup>4</sup>	11	8	8	0	13	D *RAZARIKI*NO*	0	0	1	6	1
*RISEBA*RAMASHI*	5	2	0	0	4	*YA*RURAM* <sup>5</sup>	0	8	40	24	23
*WA*NARIKERI* <sup>2</sup>	20	26	26	11	52	*NI*NARURAM*	0	2	8	8	7
*KARISEBA*MASHI*	3	6	0	0	1	*MASHI*NARISEBA* <sup>3</sup>	0	2	1	0	10
*NO*NIKERUKANA*	4	11	2	1	4	*KOSO*KARIKERE*	4	4	1	0	8
B *WA*NARIKERI* <sup>2</sup>	20	26	26	11	52	E *NARABA*RAMASHI*	1	0	0	0	8
*KOSO*RIKERE* <sup>4</sup>	11	8	8	0	13	*O*UNARIKERI*	1	0	0	0	7
*MO*KARIKERI*	4	11	8	5	7	*NO*RUNARIKERI*	4	3	4	0	10
*BAKARI*RURAM*	0	0	6	0	3						
*KOSO*NARIKERE*	4	0	5	0	5						
C *YA*NARURAM*	0	2	16	4	7						
*WA*NARIKERI* <sup>2</sup>	20	26	26	11	52						
*NO*NARIKERI*	19	30	39	19	49						

## 5 A Pattern-Based Text Data Mining System

Based on the experience of finding patterns from anthologies described in the previous section, we propose a text data mining system to support research into WAKA poetry. The system consists of two parts: the pattern discovery part and the pattern matching part. The pattern discovery part emits a set of patterns, some of which stimulate the user to form hypotheses. To verify the hypotheses, the user retrieves a set of poems containing the patterns by using the pattern matching part, examines the retrieved poems, and then updates the hypotheses. The updated hypotheses are then verified again. Repeating this process will provide results that are worthwhile to the user.

In practice, a slightly modified pattern is often better than the original emitted by the pattern discovery part, that is, a slightly more general/specific pattern may be preferred. Say the user wants to browse the ‘neighbors’ of a pattern. The proposed system has as GUI a pattern browser for traversing the Hasse diagram of the partial-order on the set of patterns. We have implemented a prototype of this system, and a new style of research into WAKA poetry utilizing the prototype system is now proceeding.

### Acknowledgments

The authors would like to thank Ayumi Shinohara and Hiroki Arimura for valuable discussions concerning this work. We also thank Setsuo Arikawa, Yuichiro Imanishi, and Masakatsu Murakami for their valuable comments.

### References

1. H. Ahonen, O. Heinonen, M. Klementtinen, and A.I. Verkamo: Mining in the phrasal frontier. In *Proc. 1st European Symposium on Principles of Data Mining and Knowledge Discovery* (PKDD'97), 343–350, 1997. **131**
2. D. Angluin: Finding patterns common to a set of strings. In *Proc. 11th Annual Symposium on Theory of Computing*, 130–141, 1979. **129**
3. H. Arimura, T. Shinohara, and S. Otsuki: Finding minimal generalizations for unions of pattern languages and its application to inductive inference from positive data. In *Proc. 11th Annual Symposium on Theoretical Aspects of Computer Science* (STACS'94), 649–660, 1994. **130**
4. G. Bownas and A. Thwaite: *The Penguin Book of Japanese verse*. Penguin Books Ltd., 1964. **133**
5. A. Brāzma, E. Ukkonen, and J. Vilo: Discovering unbounded unions of regular pattern languages from positive examples. In *Proc. 7th International Symposium on Algorithms and Computation* (ISAAC'96), 95–104, 1996. **131, 134, 135, 136**
6. R. Feldman and I. Dagan: Knowledge discovery in textual databases (KDT). In *Proc. 1st International Conference on Knowledge Discovery and Data Mining* (KDD'95), 112–117, 1995. **131**
7. E. M. Gold: Language identification in the limit. *Information and Control*, 10: 447–474, 1967. **130**

8. J. Rissanen: Modeling by the shortest data description. *Automatica*, 14: 465–471, 1978. 131
9. T. Shinohara: Polynomial-time inference of pattern languages and its applications. In *Proc. 7th IBM Symposium on Mathematical Foundations of Computer Science*, 191–209, 1982. 129