万葉集の検索システム

藤沼良一，志村栄一　（山梨大学）

1．まえがき

原始日本語の母音は8通りで，古代日本語の文書，古事記，万葉集などには，まだ，その名残りがあるという．したがって，古代の文書を扱うには，この点を考えておく必要がある．

この研究は，万葉集の検索システムの開発を試みたもので，開発には，山梨大学の乙号機FACOM－ZKO－30をFORTRANによって行う．しかし，これはカナ文字が扱えないので，適当な時期にINTEL　8080　を用いた研究室の計算システムに切換えるつもりである。

システムの開発には，まず，その目的を明らかにすることが必要である。検索システムとしては「問い合せ」と「索引書作り」とが対象になろう。索引書としてはKWIC，KWOCなどが，よく知られている。問い合せでは，自然言語によるもの，記号化した人工言語によるものなどがあろう。そして，答には，どのようなものを作るかという点を決めねばならない。

開発には，目的を明らかにした上で，とりかかるべきであるか，ここでは，これを後日にまわし，文字コードの問題と作者名の問題とを先に検討した。

2．文字の種類

まず，FORTRANを用いて開発するという前提に立ち，万葉集をどのように表わしたらよいかという問題について考える。

万葉集を扱うには，当時の日本語の音を表わす文字と，中国式の音を表わす文字とが必要である．さらに，数字を利用しての表記の簡単化と，いくつかの記号とがいる。すなわち，表記に使う広い意味の字は，いわゆる文字と数字，記号とになる．

<字> → <文字> | <数字> | <記号>

<文字> → <日本音> | <中国音>

ここに，→はBackus記号の::＝と同じ意味に使っている。

日本音は，母音だけか，子音母音とからできていて，子音だけのものはないと考える。母音は，古代には甲乙にわかれていた。これを甲母音，乙母音ということにしよう．

<日本音> → <母音> | <子音><母音>

<母音> → <甲母音> | <乙母音>

甲母音アイウエオはローマ字方式を用いて表わす。

<甲母音> → A | I | U | E | O

乙母音イエオについては，1文字では適当な文字がないので，：をつけた2字で表わした．

<乙母音> → I: | E: | O:

ローマ字表記法にはヘボン式，日本式などがあり，これに応じて，子音の表わし方も，いくつかがある．ここでは訓令式を用いる．

<子音> → K | G | S | Z | T | D | N | H | B | P | M | Y | R | W

訓令式の利点は，活用，連濁などの扱いが簡単になることの他に，古代日本語の音は，訓令式に近い発音であったということもある。

中国式の音には日本語の音の他に，拗音，促音，撥音が加わる。

〈中国音〉→〈日本音〉｜〈拗音〉｜〈促音〉｜〈撥音〉

これらは，訓令式によれば，

〈拗音〉→〈子音〉Y〈母音〉

〈促音〉→〈子音〉〈子音〉

〈撥音〉→N｜N'

母音としては甲母音だけで充分で，乙母音を考える必要はなかろう。

訓令式では，促音を表わすのに，同じ子音を重ねて用いる。これは訓令式の欠点の1つで，これの改良案としてQを用いる方式が提案されている。ここでは，Qの代りにCを用いることにする。

〈促音〉→C〈子音〉〈母音〉

撥音にNを用いると，母音の前ではナ行と混同する。これをさけるため，母音の前ではN'を用いる。したがって2通りの表記が必要になる。この複雑さをさけるため，撥音は常に2字N:で表わすことにする。

〈撥音〉→N:

:を用いても，乙母音と混乱する恐れはない。

数字，記号はつぎの通りとする。

〈数字〉→0｜1｜……｜8｜9

〈記号〉→,｜.｜(｜)｜+｜-｜*｜/｜=｜'｜:｜␣　（␣は空白）

3，文字コード

前節の検討の結果，才葉集を記すのにいる字の種類は，FORTRANの文字で英字20，数字10，記号4で，計34コである。したがって，これらを表わすには6ビット以上が必要になる。乙号機では16ビット，8080では8ビットを単位としているので，8ビットを単位とした文字コードを作る。

(1)　1字1バイト方式

最も簡単なものは，FORTRANの1字を1バイト（8ビット）に表わす方法である。これは簡単であるが，つぎの欠点がある。

(ア) 記憶容量の無駄

扱うべき文字の種類は34で，5ビット強で表わせる。これに8ビットを使うのであるから，約3ビット，40%の無駄となる。

(イ) 記事修正の困難

たとえば，甲母音を乙母音にかえること，あるいは乙母音を甲母音にかえるには:を加え，あるいは除かねばならない。記憶装置のなかみを，このようにずらすことは計算機の苦手とする仕事の1つである。しかも，このような修正は，かなり生ずると思える。

(2)　カナ文字1バイト方式

前案の欠点を直す一方法は，カナ1字を1バイトに表わす方法である。日本音では母音8コで3ビット，子音14コで4ビットあればよい。したがって，8ビットをわけ，5+3，あるいは4+4方式ができる。しかし，5+3方式では，母音のいらない撥音N:が扱えなくなる。そこで，4+4方式について考えを進めよう。

つぎに，中国音を考えると，拗音，促音，撥音が加わるから，4＋4方式で扱えるかどうかをしらべよう。

まず，拗音については，

〈子音〉Y＋〈母音〉

〈子音〉＋Y〈母音〉

とする2つの方法があろう。前の案は，〈子音〉Yを子音として扱う方法である。この案において，拗音記号Yに結びつく子音は，ヤ行Yとワ行W以外のすべてである。したがって，子音の種類は14＋12＝26で4ビットでは表わせない。

これに対し，後の方式は，Y〈母音〉を母音として扱うものである。いわば，拗母音を設定し，ロシヤ語の軟母音のように扱うものである。

〈母音〉→〈甲母音〉｜〈乙母音〉｜〈拗母音〉

拗音記号Yのつく母音はA，U，Oの3つだけであるから計11コで，4ビットで表わせる。

促音についてしらべよう。促音記号Cの扱いは，

C〈子音〉＋〈母音〉

とする方式と，前の母音とむすび

〈母音〉C＋〈子音〉＋〈母音〉

とする方式とがある。

促音記号Cは，すべての子音とむすびつくわけではなく，K，S，T，Pの4コに限られる。しかし，CK，CS，CT，CPという子音を設けると，子音は14＋4＝18となり，4ビットでは表わせない。

前2の〈母音〉Cという方式では，Cはすべての母音と結びつくが8＋5＝13で，4ビットで表わせる。

〈母音〉→〈甲母音〉｜〈乙母音〉｜〈促母音〉

しかし，拗母音3コは扱えるが，拗促母音3コまでは扱えない。この問題については，あとで再び考えよう。

撥音N:について考えよう。

N＋:

N:＋φ

とする2つの方式がある。前のものは，子音Nに母音:がついたとする考え，後のものは子音N:に母音なしとする考えである。

後の方式は，子音を増やすという方式であるが，4ビットを越えることになり4＋4方式では扱えなくなる。したがって，前の母音を増やす方式によらざるをえない。

以上の音の他に，必要と考える子音として

KW，GW，TS，DZ，F

がある。FはHWとするとしても，子音の数は増やせないから，母音として扱う方式を考えねばならない。これは，撥音記号:が子音Nにしかつけないものと考えたように，特別な子音，あるいは子音にはつけない母音とし，不完全母音と名付けよう。

〈不完全母音〉→W｜S｜Z｜:｜C

そうすると，つぎのようなコード案ができる。

I．第1案

〈文字コード〉→〈子音部〉〈母音部〉

<母音部>→<母音>｜<不完全母音>
<母音>→<甲母音>｜<乙母音>｜<拗母音>
この方式によるコードの具体案の1例をあげよう。

コード	子音部	母音部(I案)	母音部(II案)	数字記号
0	母音	:	:	0
1	K	A	A	1
2	G	I	I	2
3	S	I:	I:	3
4	Z	U	U	4
5	T	E	E	5
6	D	E:	E:	6
7	N	O	O	7
8	H	O:	O:	8
9	B	YA	AC	9
10	P	YU	IC	,
11	M	YO	UC	.
12	Y	W	EC	␣
13	R	S	OC	
14	W	Z	W	
15	記号	C	Y	

II 第2案
母音に促母音を加え，拗母音を除くとオ乙の案ができる。
<母音>→<甲母音>｜<乙母音>｜<促母音>
この具体案をあげると，つぎのようになる。子音部には変化がないから，母音部のみをあげる。
この案ではTS，DZの音は表わせないことになる。
2つの案による表記例をあげよう。子音部S，母音部Bからできる文字コードを(S，B)で表わす。

	I	II
花	(H,A)(N,A)	(H,A)(N,A)
神	(K,A)(M,I:)	(K,A)(M,I:)
絵画	(K,W)(−,I)(G,W)(−,A)	(K,W)(−,I)(G,W)(−,A)
京都	(K,YO)(−,O)(T,O)	(K,Y)(−,O)(−,O)(T,O)
国家	(K,O)(−,C)(K,A)	(K,OC)(K,A)
出張	(S,YU)(−,C)(T,YO)(−,O)	(S,Y)(−,UC)(T,Y)(−,O)(−,O)

ここに，−は母音の印を示す。
ここにあげた例だけでは，両者の優劣は決めにくいので，音便の立場からくらべてみよう。口語には，動詞の活用に際し，音便がある。
書きた→書いた（イ音便）
呼みた→呼んだ（撥音便）

勝ちた → 勝った　(促音便)
これを，両方式で表わしてみよう．

	I	II
(K,A)(K,I)(T,A)	(K,A)(-,I)(T,A)	(KA)(-,I)(T,A)
(K,A)(M,I)(T,A)	(K,A)(N,:)(D,A)	(K,A)(N,:)(D,A)
(K,A)(T,I)(T,A)	(K,A)(-,C)(T,A)	(K,AC)(T,A)

この例のように，II案では，促音便に対して語長が変ると，語形の変化が面倒であるという欠点がある。また，TS，DZの扱えないこともあり，I案の方がすぐれていることがわかる。

4、作者名の扱い

万葉集のもつ情報の量をしらべよう。岩波版日本古典文学大系万葉集1～4巻をもとに概算すると

4巻×3000頁×400字×0.5＝24万字

である。この式の係数0.5は，同書が原文と訓み下し文とから構成してあるためのものである。万葉集には，約4500首の歌があるが，これのみを対象としても

31字×4500首＝14万字

となる。これは，すべてが短歌だけとしての計算であるから，実際はもっと多くなる。検索の効率をあげるためには，このような多量の情報を，如何に圧縮するかということになる。ここでは，まず作者名について検討してみよう．

作者名のような長さの変るものを扱うには，(1)可変長方式，(2)固定長方式との2つの方法がある。記憶容量，検索時間を物指しとして，これらの評価比較を行う。

(1)可変長方式

作者名を可変長のまま扱う方法について考えよう。いま，作者名の平均長を10字と仮定すれば，これの総字数Kは

K＝10字×4500首＝4.5万字

となる。これが，必要記憶容量の目安になる。

つぎに，ある作者の全作歌をとり出すに要する検索時間をしらべる。この時間は，作者名をしらべるに要する字数で評価できよう。名前の違うことを知るには平均長の半分，すなわち5字しらべればわかる。したがって，全部の検索には

H＝5字×4500首＝2.3万字

の比較がいる。このKとHとが比較の物指しとなる。

(2)固定長方式

作者名を固定長にするにはいろいろな方法があるが，辞書を利用するのが簡単である。名前の辞書に要する記憶容量は

10字×500人＝0.5万字

となる。ここでは，作者は500人として計算した。

作者は500人であるから，この番号は2バイトで表わせる。したがって歌集中の作者名は

2字×4500首＝0.9万字

となる。よって、総字数Kは、これらの和となる

K＝1.4万字

名前の検索には、辞書の検索が加わる

辞書検索：　5字×250人 ＝ 0.13万字

歌集検索：　1字×45000首 ＝ 0.45万字

計　　H＝0.6万字

(3) 語方式

作者名を見ると、同じ語が何人かの名前に現われる。たとえば、大伴家持、大伴東人、大伴旅人の如くである

これから、語辞書を用い　語番号の列として名前を表わす中間の方式が考えられる。この方式についてしらべよう。

作者名は3語，1語は4字でできると仮定しよう。この仮定で，作者名の延べ語数は

3語×500 ＝ 1500語

となる。異なり語は，この半分とすれば750語となるこれにより字数は

語辞書：　4字×750語 ＝ 0.3万字

歌集：　2字×3語×45000首 ＝ 2.7万字

計　　K＝3万字

となる。

検索時間は，語辞書を3回さがして作者名が決まり、歌集の検索には，語番号が2字であるから、つぎのようになる。

辞書検索：　2字×375語×3回 ＝ 0.22万字

歌集検索：　3字×45000首 ＝ 1.35万字

計　　H＝1.6万字

となる。

以上の計算結果を表Ⅰすると、つぎのようになる

	可変	固定	語
容量	4.5	1.4	3.0
時間	2.3	0.6	1.6

この表Ⅰよれば，固定方式が最もすぐれ，語方式，可変長方式の順になっている。しかし，データの作成　誤の発見などの立場からは，語方式がすぐれている。たとえば，て母音を用いるABE：（阿倍）を、甲母音にしたABEの誤りは，語辞書方式により，容易に発見できる。

5. 作者名表の作成

前節の検討結果にもとずき，語辞書を利用した作者表を作る作業を行った。使用した計算機は2号機FACOM　270－30で，FORTRANを用いた。作業は，次のような順序で行った。

(1) 作者カードの作成

資料2の時代順作者人名録をもとにし，前記資料1を参照して作者カードを作る。作者カードの構成は、作者番号，作者名，必要なときは注釈を加えてある．

〈作者カード〉→〈作者番号〉〈作者名〉〔＊〈注釈〉〕

ここに、〔……〕は、なくてもよい項目を意味する。
資料2の人名録は、ほぼ50音順に配列してあるので、作者番号も、その順になっている。

(1) 101 AOZI NO: OKISIMA
(2) 102 AGATA NO: INUKAHI NO: WOTOME
(3) 103 AGATA NO: INUKAHI NO: HITOKAMI
(4) 104 AGATA NO: INUKAHI NO: MITIYO:
(5) 105 AGATA NO: INUKAHI NO: MOTIWO
(6) 106 AGATA NO: INUKAHI NO: YOSIWO
(7) 107 AKI: NO: OHOKIMI
(8) 108 AKI: NO: WOSA NO: OBITOMARO: * SURUGA NO: SAKIMORI
(9) 109 ASAKURA NO: MASUHITO * KAMITUKE:NO NO: SAKIMORI
(10) 110 ASADA NO: YASU
(11) 111 ASUKA NO: OHOKIMI
(12) 112 ASUKA NO: NADO:MARO:
(13) 113 ATUMI NO: OHOKIMI
(14) 114 ATO NO: TOSITARI
(15) 115 ATO NO: TOBIRA NO: WOTOME
(16) 116 ABE: NO: OKINA
(17) 117 ABE: NO: OKIMITI
(18) 118 ABE: NO: WOMINA
(19) 119 ABE: NO: KOOHODI
(20) 120 ABE: NO: SAMIMARO:
(21) 121 ABE: NO: TUGIMARO:
(22) 122 ABE: NO: TOYOTUGU
(23) 123 ABE: NO: BOU
(24) 124 ABE: NO: BOU * ABE: NO: TUGIMARO: NO: ZINAN:
(25) 125 ABE: NO: HIRONIHA
(26) 126 ABE: NO: MUSIMARO:
(27) 127 AMA NO: BOU
(28) 128 AMA NO: INUKAHI NO: WOKAMARO:
(29) 129 AMU NO: MOROTATI
(30) 130 AHATA NO: BOU
(31) 131 AHATA NO: OHOKIMI
(32) 132 AHATAME NO: WOTOME
(33) 133 ARAUDI INASIKI
(34) 134 ARIMA NO: MIKO

(35) 201 IHOMARO:
(36) 202 IKUSA NO: OHOKIMI
(37) 203 IKUTAMABE NO: TARUKUNI * TOHOTUAHUMI NO: SAKIMORI
(38) 204 IKEDA NO: BOU
(39) 205 IKEBE NO: OHOKIMI
(40) 206 ISIKAHA NO: OKINA
(41) 207 ISIKAHA NO: IRATUME
(42) 208 ISIKAHA NO: HUHITO
(43) 209 ISIKAHA NO: OHOBA
(44) 210 ISIKAHA NO: WOMINA * HUDIHARA NO: SUKUNAMARO: NO: IUMA
(45) 211 ISIKAHA NO: WOMINA * OHOTU NO: MIKO NO: MIYA NO: MAKATATI
(46) 212 ISIKAHA NO: KAKE NO: WOMINA
(47) 213 ISIKAHA NO: KIMIKO
(48) 214 ISIKAHA NO: TARUHITO
(49) 215 ISIKAHA NO: TOSITARI
(50) 216 ISIKAHA NO: BOU * MIYAMARO: KA KIMIKO KA
(51) 217 ISIKAHA NO: BOU
(52) 218 ISIKAHA NO: HIRONARI
(53) 219 ISIKAHA NO: MIMITI
(54) 220 ISOUDI NO: NORIMARO:
(55) 222 ISONO:KAMI NO: KATUWO
(56) 223 ISONO:KAMI NO: BOU * OTOMARO: KA

(2) 語辞書の作成

作者カードから，作者名を構成する異なり語をとり出し，これを並べかえる。そして，語番号をつける。こうしてできたのがつぎの表である。

この時点では，文字コードの編成はしていないのでFORTRANの文字順に並らべてある。また，語番号が100以下のものは，数字記号とした。こうして得た異なり語の総数は，ほぼ600語である。

KOTONARIGO NO MEIZIBIKI
(GOSU= 597 GO)

1	*
2	1
3	2
4	3
5	4
6	5
7	5NEN:
101	ABE:
102	ADUMABITO
103	ADUMAMARO:
104	AGATA
105	AHAMARO
106	AHATA
107	AHATAME
108	AHUMI
109	AKAHITO
110	AKAMARO:
111	AKIMOTI
112	AKINIHA
113	AKI:
114	AMA
115	AMU
116	ANAGA
117	ARAMIMI
118	ARAUDI
119	ARIMA
120	ASADA
121	ASAKURA
122	ASAME
123	ASUKA
124	ATO
125	ATUMI
126	AUZI
127	AYUMARO:
201	BEN:KI
202	BOU
203	BUNIN:
401	DOUYOU
501	ENITATI
502	ENO:WI
503	EUZI
701	GEN:MEI
702	GEN:SYOU
703	GITUU
704	GIYAUMONI
705	GO:NO:TANIWOTI
706	GUWANIGOZI
707	GUWANINI
801	HADA
802	HAHA
803	HAKURI
804	HAKUTUU
805	HANISI
806	HANIUDI
807	HARIMA
808	HASETUKABE
809	HASIHITO
810	HATA
811	HATIMARO:
812	HATORIBE
813	HATUSE
814	HEGURI
815	HEGURIUDI
816	HIBITO
817	HIMATURI
818	HIMEMIKO
819	HINO:KAMI
820	HINO:KUMA
821	HIOKI
822	HIROKAHA
823	HIROKATA
824	HIROME
825	HIROMIMI
826	HIRONAHA
827	HIRONARI
828	HIRONIHA
829	HIROSE
830	HIROSIMA
831	HIROTARI
832	HIROTO
833	HIROTUGU
834	HITATI
835	HITO
836	HITOKAMI
837	HITOMARO:
838	HITONA
839	HITONUSI
840	HITOTARI
841	HITOWOSA
842	HITUGIMIKO
843	HITUZI
844	HIZEN:
845	HOHUSI
846	HOUZI
847	HOZUMI
848	HUBITOUDI
849	HUBUKI
850	HUDIHARA
851	HUDIHARABE
852	HUDIWI
853	HUHITO
854	HUKUSI
855	HUMI
856	HUMIMOTI
857	HUMINARI
858	HUMUYA
859	HUNADO
860	HUNE
861	HURU
862	HURUHITO
863	HUSASAKI
864	HUSIMARO:
865	HUUSI
901	IHANO:HIME
902	IHE

(3)　作者名表

再び作者カードを読み直し，語辞書の番号によって作者名を表わしたのが，次の表である。

この表からわかることは，駿河の防人，上野の防人などという注釈を除けば，殆んどが3語で表わせる。

SAKUSYAMEI NO GOKUBANGO LIST

```
101 = 126,1422,1533,
102 = 104,1422, 921,1422,2329,
103 = 104,1422, 921,1422, 836,
104 = 104,1422, 921,1422,1337,
105 = 104,1422, 921,1422,1354,
106 = 104,1422, 921,1422,2526,
107 = 113,1422,1507,
108 = 113,1422,2323,1422,1501,   1,1958,
* *  1422,1910,
109 = 121,1422,1312,   1,1119,1422,1910,
110 = 120,1422,2515,
111 = 123,1422,1507,
112 = 123,1422,1401,
113 = 125,1422,1507,
114 = 124,1422,2060,
115 = 124,1422,2048,1422,2329,
116 = 101,1422,1532,
117 = 101,1422,1530,
118 = 101,1422,2320,
119 = 101,1422,1162,
120 = 101,1422,1914,
121 = 101,1422,2068,
122 = 101,1422,2065,
123 = 101,1422, 202,
124 = 101,1422, 202,   1, 101,1422,2068,
* *  1422,2603,
125 = 101,1422, 828,
126 = 101,1422,1360,
127 = 114,1422, 202,
128 = 114,1422, 921,1422,2318,
129 = 115,1422,1352,
130 = 106,1422, 202,
131 = 106,1422,1507,
132 = 107,1422,2329,
133 = 118, 919,
134 = 119,1422,1324,
201 = 903,
202 = 908,1422,1507,
203 = 909,1422,2030,   1,2049,1422,1910,
204 = 906,1422, 202,
205 = 905,1422,1507,
206 = 923,1422,1532,
207 = 923,1422, 922,
208 = 923,1422, 853,
209 = 923,1422,1503,
210 = 923,1422,2320,   1, 850,1422,1955,
* *  1422,2073,
211 = 923,1422,2320,   1,1522,1422,1324,
* *  1422,1338,1422,1303,
212 = 923,1422,1112,1422,2320,
213 = 923,1422,1142,
214 = 923,1422,2029,
215 = 923,1422,2060,
216 = 923,1422, 202,   1,1339,1101,1142,
* *  1101,
217 = 923,1422, 202,
218 = 923,1422, 827,
219 = 923,1422,1326,
220 = 927,1422,1419,
222 = 925,1422,1137,
223 = 925,1422, 202,   1,1539,1101,
```

あとがき

語辞書を利用するのは，かなり有利な方法と思うが，駿河の防人などという句辞書の用意も有効かも知れない。これからは更に，統計的な検討が必要であろう。

おわりに，研究に種々助成して下さった本講座伊藤誠助教授，プログラムやデータの作成をして頂いた武藤厚子技官に感謝します。

参考文献

(1) 岩波版　万葉集 1－4　日本古典文学大系
(2) 武田祐吉校註　万葉集上、下　角川文庫