日本語教育用ＣＡＩシステムの試作―文字と語彙学習―

白井英俊，荒木卓也，内藤佳有（東大　工）

１．はじめに

外国人が日本語を学習する場合、その学習過程は　(1)文字の習得　(2)語彙・文法の学習　(3)日本文化の学習　と進んでいくと思われる。日本語は諸外国語と比較して文字が多様であり、(1)の文字習得に時間がかかる。自習することが可能になれば、学習効果は増大する。東大工学部では「外国人研究留学生のための日本語教育システム」としてＣＡＩ（Computer Assisted Instruction）システムをマイコン上で製作している。現在までに試作したシステムとしては次のようなものがある。

1)　かな学習
2)　手書き文字認識・批評
3)　コンピュータ漢英辞典
4)　漢字テスト

本稿では1)～4)のシステムの概要について報告する。なお、使用しているハードウェア構成は図１のとおりである。

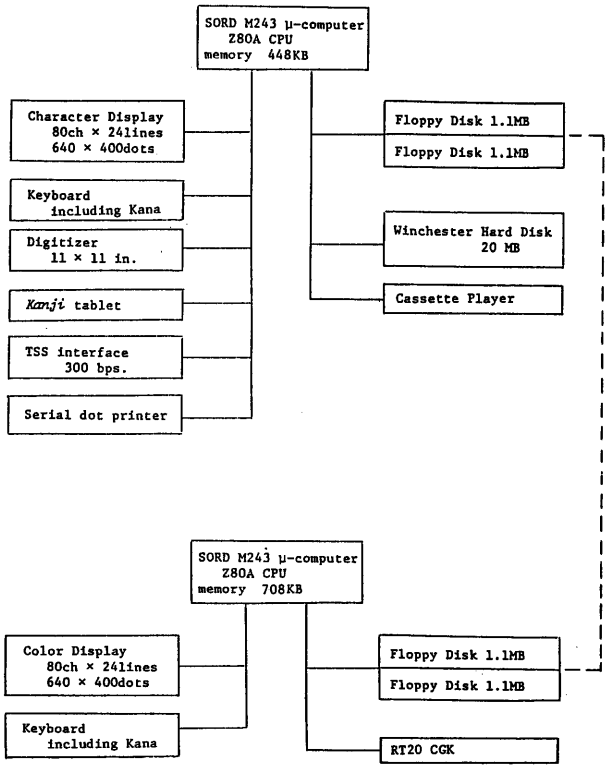

図１．ＣＡＩシステムハードウェア構成

2．かな学習

日本語の学習を始めたばかりの人を対象としたもので、次の２種類がある。いずれも、かなの書き方と発音とを自習することが可能である。

1)　五十音によるかな（ひらがな，カタカナ）の学習

2)　ローマ字入力指定によるかな（ひらがな，カタカナ）の学習

1)は五十音にしたがって順次かなを表示し、テープデッキを連動させてその読みを与える。さらにそれぞれのかなを含んだ単語とその意味（英語）を表示するとともに読み方を与える。2)はかな１文字に相当するローマ字を入力すると、まずそのひらがな，カタカナおよびその文字を含む五十音の行を表示する（図２(a)）。次にこの行に含まれるかなを用いた単語を意味（英語）とともに表示する。なお指定された文字は色をかえて表示する（図２(b)では太字で示してある）。

これらのプログラムで使用しているかなデータはすべて特別に作成したストロークデータで、筆順通りに表示される。またプログラムは、表示する内容を記述するテキストファイル部とこのテキストファイルを解釈・実行するインタプリタ部とからなっており、インタプリタ部さえ移植すればどのような機種のマイコンでも使用可能である。テキストファイルは今回考案した簡単なＣＡＩ言語で記述されている（図３）。

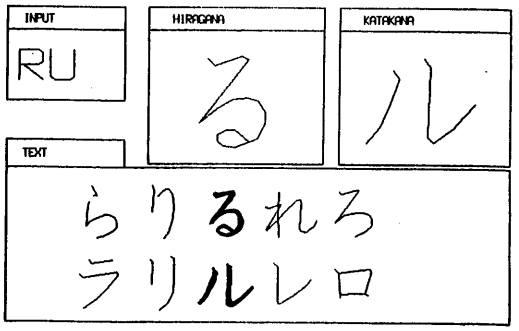

図２(a)　かな表示画面

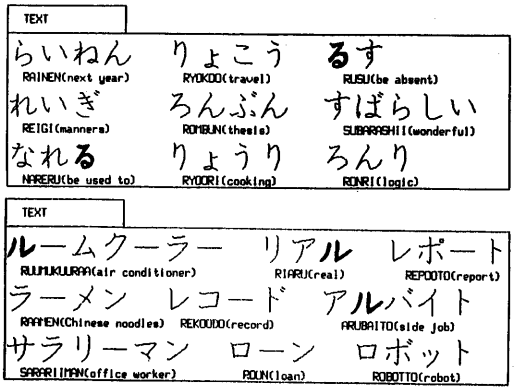

図２(b)　単語表示画面

```
CSIZE 16,32
MOVE 200 365
TEXT          Chapter 1
.SIZ 100
.POS 20 260
.DRW $"$$$&$($*
.WAT 2
MOVE 68 248
CSIZE 8,16
TEXT     A          I          U          E          O
.POS 20 138
.DRW $+$-$/$1$3
.WAT 2
MOVE 68 126
TEXT    KA          KI         KU         KE         KO
.POS 20 16
.DRW $5$7$9$;$=
.WAT 2
MOVE 68 4
TEXT    SA          SHI        SU         SE         SO
.BEL
```

図３．テキストファイルのＣＡＩ言語

3．手書き文字認識・批評[1]

ディジタイザ上に任意の文字（ただし現在、ひらがな，カタカナ，教育漢字のみ）を書くことにより、その文字に対する批評を得ながら日本語の文字を独習するためのシステムである。本システムは文字データベース管理部，認識部，批評部の3部分から構成されている。文字データベース管理部は各文字のストロークや特徴の情報を定義したデータベースとそのデータの更新・修正を行うプログラムから成る。このデータベースをもとにして認識・批評が行われる。図4に「か」のデータ例を示す。

```
Character = か  ($+)
Number of strokes =  3
Stroke number  1  Options :  9
Stroke number  2  Options :  5                4
Stroke number  3  Options :  4                3
 1 Prop. : INTERSECT            Stroke 1  =  1     Stroke 2 =  2
 2 Prop. : NOT INTERSECT        Stroke 1  =  2     Stroke 2 =  3
 3 Prop. : LEFT                 Stroke 1  =  2     Stroke 2 =  3
 4 Prop. : START POINT ABOVE    Stroke 1  =  2     Stroke 2 =  1
 5 Prop. : CLOCK-WISE CURVE     Stroke 1  =  1     Value    =  4
 6 Prop. : STARTING DIRECTION   Stroke 1  =  1     Value    =  2
 7 Prop. : NUMBER OF SEGMENTS   Stroke 1  =  1     Value    =  1
```

図4　データベースの内容例（「か」についてのデータ）

認識部は書かれた文字からストローク情報を抽出し、弁別木をたどり、その文字の認識を行う。多少誤まった文字でも受け入れられるように構成されている。そして、入力文字のストローク情報とデータベースの情報とを比較して、書かれた文字に改善の必要があればその旨を画面に表示するのが批評部である。図5に批評例と書かれた文字から抽出したストローク情報の例を示す。通常画面に表示されるのは批評のみである。

```
CRITIQUE :

Correct !

Stroke        Direction      Length
 1            CW ( 5 )         416          Start =  2  End =  7
 2               5             361
 3               3             181


For LINES  -- First, Last coordinate points
For CURVES -- Top line indicates MAX and MIN coordinates

            Second line indicates First, Last
Stroke        X1   X2   X length          Y1   Y2   Y length
 1           441  304  -137               362   83  -280
 1 Curve -   303  324    21               336   82  -254
 2           391  244  -147               434  220  -214
 3           511  588    77               409  305  -104
```

Character found か

図5　批評および抽出したストローク情報の例
（「か」についてのデータ）

4. コンピュータ漢英辞典

漢字の読みや意味などを簡単に得ることができるなら、やはり学習の能率は向上する。特に外国人にとって漢字の読みを知るためには大変な努力が必要である。

コンピュータ漢英辞典には、ＪＩＳ第１水準の漢字2,965文字に対して次のような情報が登録されており、1), 2), 3), 4), 6) またはこれらの組み合わせにより、漢字を検索することができる。

1) 部首
2) 画数
3) 四角号碼[2] (付録１参照)
4) Simコード (付録２参照)
5) ネルソン漢英辞典の参照番号
6) 読み(音読み, 訓読み)
7) 意味(英語)
8) 熟語およびその読みと意味(英語)

1), 2), 3), 4) の各特徴から特定の漢字は、バイナリサーチによって求める。6)の読みについては、高速に検索するためにハッシュ表を用いている。さらに、ハッシュ表を用いることにより、新規データ追加時、再びソートする必要がなく、データのメンテナンスが楽になる。また、読みから漢字を求めるだけでなく、漢字から読み・意味・熟語を求める必要がある。そこで、読み・意味・熟語は図５に示すようなファイル構成で記憶されており、記憶容量の圧縮をはかっている。特に英単語で語尾だけが異なるもの(—ed, —tion, —ing など)は基本形と語尾を別々のファイルにすることによって重複を避けている。読みは読みハッシュ

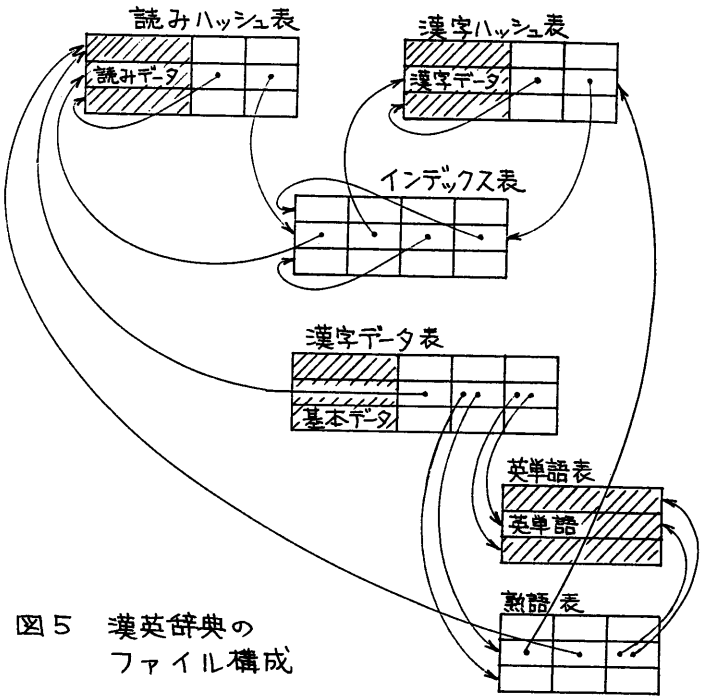

図５ 漢英辞典の
ファイル構成

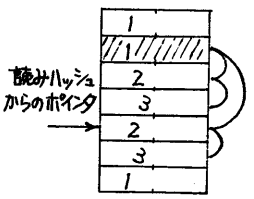

図６ 読みデータ
ファイルの構成

表からさらにポインタを出し、図6に示すような2バイト単位のファイルをさしている。つまりコード体系を工夫して重複部分を共通使用して、データの圧縮をはかっている。

以下「親」という漢字の情報を求めることを例にとって、各動作の説明を行う。画数（16画）を手がかりに漢字を捜そうとすると、この条件を満たす漢字が複数あるため、図7に示すような画面が得られる。ここでNo. 67を指定すると「親」に対する上記1)～5)の情報がまず画面の上部に表示され、さらに「親」の書き方が一画毎に示される（図8(a)）。条件の指定の仕方により漢字が一定に限定される場合はすぐにこの画面が表示される。次に、読み（音読みはカタカナ，訓読みはひらがな）と意味（英語）が表示される（図8(b)）。また、その漢字を用いた熟語およびその読みと意味（英語）を見ることも可能である（図8(c)）。

ところで、画数だけを手がかりに漢字を捜そうとすると図7のように非常にたくさんの漢字があり、音読み順にソートされてはいても、読みを知らない人が、この中から「親」を捜し出すのは大変である。本システムでは、これを容易にするために、複数の特徴を指定する機能がある。例えば画数＝16画と部首＝「見」

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
穐	鮎	緯	謂	窺	毅	穎	頴	衝	燕	鴛	憶	穏	壊	憾	骸	獲	橿	禅	鴨	憾	翰	諌	遠	館
25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
舘	機	橘	錦	選	覆	橋	興	審	凝	錦	薫	憩	頸	澱	憲	賢	諺	醐	衝	鋼	墾	錯	錆	餐
50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74
諮	鴨	篠	編	錫	儘	樹	聾	獣	縦	憔	蕪	鞘	壌	嬢	錠	新	親	錐	錘	整	醒	積	贋	膳
75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99
操	醒	濁	樽	壇	薬	謀	諭	踏	澱	踏	燈	糖	湯	頭	薫	橡	嘲	曇	燃	濃	薄	縛	噺	繁

図7　画数＝16画の漢字（一部）

親　《教育漢字》　部首：見　画数：16　四角号マ：6910
Nelson：4293　SIM code：020907

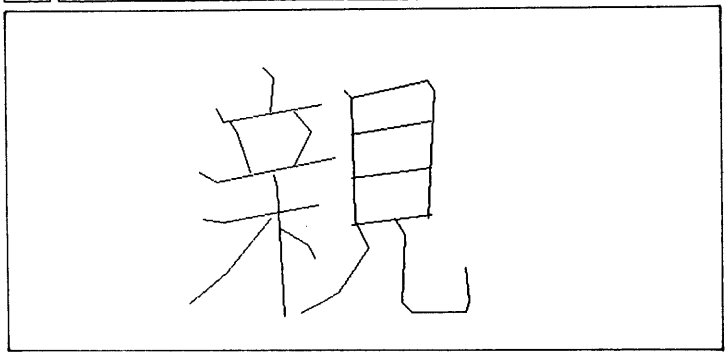

図8(a)　「親」の基本データと書き方

親　《教育漢字》　部首：見　画数：16　四角号マ：6910
Nelson：4293　SIM code：020907

Its readings and meanings are ---

1.シン	(intimacy; parents; relative)
2.おや	parent
3.した・しい	intimate; friendly
4.した・しむ	be friendly

図8(b)　「親」の読みと意味

親　《教育漢字》　部首：見　画数：16　四角号マ：6910
Nelson：4293　SIM code：020907

Some examples are ---

1.懇親会	：　こんしんかい a social meeting;a get-together
2.親子	：　おやこ parent and child
3.親心	：　おやごころ parental affection
4.親族	：　しんぞく a relation,a relative
5.親友	：　しんゆう a bosom friend;one's best friend
6.肉親	：　にくしん a blood relation

図8(c)　「親」の熟語と意味

（コード＝１４７）を指定することにより即座に「親」を求めることができる。「岩」のように、部首を２つ（「山」と「石」）指定して検索することも可能である。このように、複数の特徴（1），2），3），4），6）の組み合わせ）を指定することによって、検索が容易になる。

読み入力では、熟語を検索することも可能で、熟語の各構成語について「親」で示したものと同様な情報を得ることができる。また、部首，Simコードによる検索を用いれば、同一の部首あるいはつくりをもつ漢字を得ることができ、文字学習という点からみれば非常に有益であると考えられる。

なお、図８(a)の書き方を表示する漢字は、教科書体の活字にもとづいて特別に作成したデータで、現在、教育漢字９９６文字のデータが入力されている。

5．漢字テスト

語彙学習のためのプログラムで

1） 表示した文章中の漢字の読みを入力させる

2） 表示した文章中の空白に入る漢字を選択させる（図９）

ものとがある。これらは難易度に応じて９ランクに分類されており、能力に応じたテストが可能である。また、受験者の成績，学習状態を管理するマネジメント機能も有している。

将来、３項で述べた手書き文字認識機能と結合した筆記によるテストシステムも製作する予定である。

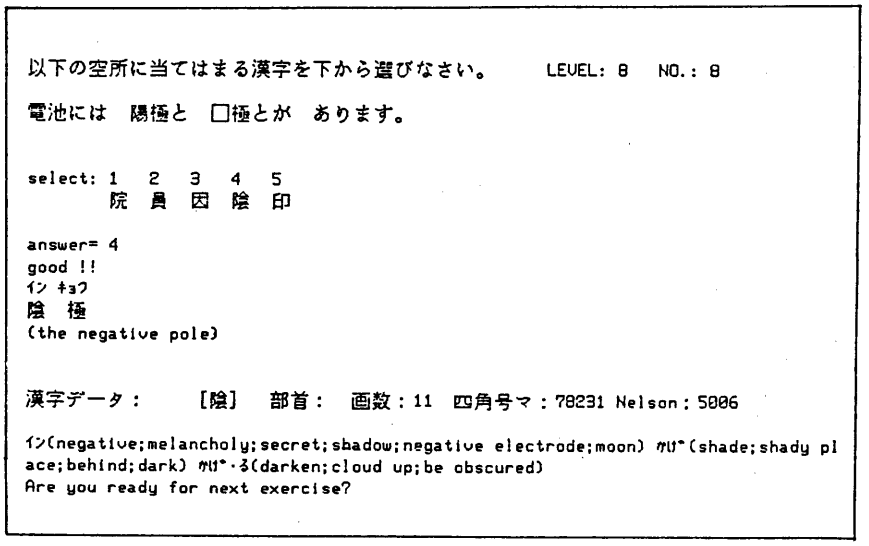

図９ 漢字テストの画面

6．おわりに

外国人留学生の日本語教育支援のために試作したシステムの概要について述べた。今後はこれらのシステムを実際に使用してもらい、問題点の洗い出しを行い、さらに改善していく予定である。また、語彙・文法の学習を支援するシステムも今後開発していく予定である。

参考文献
(1) ディヴィッド，白井："手書き文字の認識と批評を行うシステム"，電子通信学会技術研究報告（パターン認識と学習）ＰＲＬ８３－１１，１９８３年６月
(2) 諸橋ら："新漢和辞典"，大修館書店，１９８１年４月

付録１．四角号碼

四角号碼は中国の王雲五という学者によって考案されたもので、漢字を形でとらえて分類する方法である。分類の方法は次のとおりである。

1) 漢字の四すみの形をあらかじめ分類してある１０種類のパターンにあてはめて、０～９の番号をつける。
2) その四すみの番号を、左上・右上・左下・右下の順に並べて４桁の数字にする。
3) 複数の漢字が同一番号となる場合、これらを区別するために、右下すみでとったすぐ上の形をとり、これにパターンをあてはめ５桁目とする。

例えば「漢」は

左上	丶：３
右上	十：４
左下	ノ：１
右下	乀：３
右下の上	十：４

から「３４１３ｋ」となる。

（以上　大修館「新漢和辞典別冊付録」による）

付録２．Simコード

Simコードはウィーン工科大学のSimoncsics博士によって考案されたもので、漢字を形と画数の情報にもとづいて分類する方法である。このコードは第１～第５コードからなり、それぞれ６桁の数字で構成されている。以下、各コードについて説明する。

1) 第１コード

漢字を形でとらえて、あらかじめ決めてある１３種類のパターンにあてはめ、その番号（００～１２）をつける。例えば、左右に分かれるのであれば「０２」，上下に分かれるのであれば「０８」というぐあいである。そして、パターンのそれぞれの部分に含まれる画数を２桁の数字で表わし、３・４桁目と５・６桁目とする。したがって３・４桁目と５・６桁目の和は総画数となる。

例えば「漢は」

左右に分かれる　　０２

左の部分「シ」の画数　０３
右の部分「莫」の画数　１０
から「０２０３１０」となる。

2) 第２コード

分類したパターンの残りの部分（例えば「０２」では右側，「０８」では下側）に相当する部分の第１画目と第２画目の形と関係をコード化して表わす。

例えば「漠」の残りの部分（右側）は「莫」で、その第１画目と第２画目は「艹」である。

第１画目のパターン　２１
第２画目のパターン　２２
この関係（交わる）　２４

以上より第２コードは「２１２２２４」となる。

3) 第３コード

同様に、残りの部分の最後の一画とその前の一画の形と関係をコード化したものである。

4) 第４コード
5) 第５コード

第２，３コードと同様なことを、分類したパターンに含まれる部分（例えば「０２」では左側，「０８」では上側）に適用したものである。