姓名の漢字仮名変換システム

中村祥次郎(日本ユニバック㈱) 田中康仁(日本ユニバック)

•はじめに

漢字を取り扱う時に常に問題になることの1つにふりがなの問題があります。この問題はただ単に漢字と仮名のデータ処理が可能になることによって問題が解決したわけではありません。

漢字の入力と同時にふり仮名も入力しなければならないのか？等色々な問題があります。ふり仮名はデータの分類や索引などで重要な意味があります。そこで漢字と仮名の関係はどのようになっているかどのようにすれば処理できるか調べてみました。

1. 漢字1文字に対する仮名文字の割合.

毎日われわれは日常生活の中に漢字を使用していますが漢字1文字に対する仮名文字の桁数はと聞かれるとまどうものです。日本人の姓6万種名3万5千種を調べてみると次のような結果が得られました。

漢字：仮名＝1：1.686 (姓)
漢字：仮名＝1：1.767 (名)

この結果からみると名前の方が0.1文字だけ仮名が長いことがわかります。この原因については別の機会に分析することにします。漢字に対する仮名の割合はこの結果からわかるように平均1.6～1.8文字です。

これにより仮名文字の桁数から漢字の桁数を割り出すと逆を行うこともできます。

2. 漢字の単語に対する仮名の割合

漢字の単語の読み方は日常文においてほぼ一定しておりそっもろっも単語の読み方がある場合はめずらしいものです。1か1姓や名についてはそっ以上の読み方があるものを簡単に見つけることができます。たとえば角田(ツノダ ツノタ カクタ カクダ スミタ スミダ カドタ)では7種類の読み方があり、村主(ムラヌシ スグリ)では2種類の読み方があります。日本人の姓6万種類、名前3万種類で漢字の読み方を調べると次のようになります。

表1 1つの漢字姓名に対する読み方

読み方件数	姓 件数	名 件数
1	42,934	30,267
2	6,748	3,007
3	1,318	543
4	347	119
5	88	30
6	25	11
7	9	8
8	3	6
9	1	1
10	1	0
11	0	1
12	0	0

この表からわかるように漢字に対する読み方は1～3種類で例外的なものでもせいぜい10種類程度です。これは仮名文字の単語に対する漢字の単語つまり同音異義語の発生に比べると発生頻度が少ないということがわかります。

姓や名の同音異義語の発生頻度を調べると次のようになります。

表2

同音異字件数	姓	名
1	24,779	5,727
2	6,706	1,684
3	2,428	889
4	1,288	511
5	623	367
6	362	216
7	227	183
8	160	153
9	88	123
10	55	91
11	46	71
12	34	80
13	29	59
14	15	44
15	11	33
16	9	40
17	7	25
18	6	22
19	1	23
20	2	24
21	3	21
22	0	24
23	1	16
24	1	12
25	1	12
26	1	17
27	0	8
28	0	7
29	1	15
30	0	9
31以上	0	145
計	36,888	

この表からわかるように1つの発音に対して20件～30件というものも現われます。漢字の姓名の読み方が読み方に対する音化かすく仮名文字の姓名に対する漢字の発生割合は非常に多いことがわかります。

3. コンピュータによるふり仮名付けについて

コンピュータによるふり仮名を考えると全ての処理が自動的に処理されなければならないと考えますが何もすべての処理を自動的に処理する必要はなくある一部分を人間の介入により現実に近づけることが望ましいと考えています。これから2つの処理方法について述べてみます。

3.1 姓と名のファイルによる処理方法

日本人の姓と名前のファイル(姓10万件名前15万件)をディスク・ファイル上に格納漢字の姓と名前にふり仮名を付けなければならないファイルから取りだし漢字ディスプレイ上に表示しこれと同時に漢字の姓漢字の名前により姓のファイル名前のファイルを検索し該当する姓と名前のカナ文字を表示し原票と照合しキーボードまたはライトペンによる入力により正しいふり仮名付けを行ないます。もし該当するカナ文字の姓や名前がない場合はキーボードより入力します。

次の例によって具体的に説明します。ふり仮名を付けたいファイルはディスク上にカタログします。

```
← 01234567
→ 東　猛
→ 1アズマ 2ヒガシ
  3タケシ 4イサム 5ツヨシ

← 1,3
→ アズマ　タケシ
← Y
```

- 顧客NOのキーイン
- ファイルの漢字名の表示
- 姓名ファイルより該当するカナ表示
- 1,3
- 確認の表示
- 正しいという確認

以上のような漢字ディスプレイ上の画面処理でふり仮名付けを行ないます。

次にこの方式の長所欠点について述べてみます。

長所　(1) 正確であり確実である。
　　(2) プログラムが簡単である。

欠点 (1) 単語を集め続けなければならない。
(2) ファイルの容量が大きいのでミニコン程度では処理できない。
(3) 単語のないものに対しては人間の力に頼らなければならない。
(4) 有限な単語から有限な単語への振り仮名付けであるためシステムの発展性がない。

この方法の長所，欠点を十分考えた上でシステム設計されることを望みます。

次に別の方法による漢字仮名変換の方法について述べてみましょう。

3.2 文字単位にふり仮名を付ける。
文字単位にふり仮名を固定して行なう方法であると早のみこみしないでいただきたい。このような誤解を避けるために文字と姓，名について予備知識について述べてみましょう。

(1) 文字と位置
文字には位置関係があります。たとえば2つの名前について考えてみましょう。

康仁　　徳之助

「康」「徳」という文字は先頭の文字であります。「之」は文字と文字の中間にある文字です。「仁」「助」は末尾にある文字です。これらを簡単に表現するため「先頭の文字」は「頭」「中間の文字」は「中」末尾の文字は「尾」とコード付けをします。

(2) 読み方
「康」「仁」「徳」「之」「助」これら文字には文字単位に読み方が付けられています。この読み方は辞書による読み方ばかりでなく人名としての読み方もあります。ここではこれらを含めて考えてみます。

(3) 文字の読み方と音訓
読み方には音読み訓読みなどがあります。

(4) 文字の読み方と頻度
読み方は1つの文字に3～5種類あります。しかし読み方の頻度の多いものとあまり使われない読み方もあります。また(1)の文字と位置の説明のように同一文字でも先頭の文字に現われる読み方と末尾に現われる文字では読み方の頻度が違ってきます。

(5) 読み方の正当性
「康仁」「徳之助」の読み方は「ヤスヒト」「トクノスケ」です。「ヤスヒト」「トクノスケ」のカナ文字を調べると次のことに気づきます。つまりヤの次にはスという文字は接続することがある。その次にはヒとケという文字は接続している等ということです。このことにより姓名のカナ文字の連結関係を表わしたマトリックスを考えます。

後続文字

先行文字	ア ・・・・・・・・・・ ン
ア ・・・・・・・・・・ ン	

この表によりある姓又は名に対してのふり仮名付けをした場合の正当性の検証を行うことができます。つまり妙な発音のふり仮名をした場合この表により削除することができます。しかしこの表だけですべての検証は行なえません。

予備知識として5つの事柄を考えてみました。次にこれらの知識を基にして文字についての実際のデータを集めてみました。電話帳より2万5千件の名前を調査し文字別に読み方位置別に頻度を集計しました。

表3

文字	位置	読み方	頻度
康	△	ヤスシ	35
	△	コウ	5
	頭	ヤス	988
	〃	コウ	111
	〃	ミチ	2
	〃	ノブ	1
	尾	ヤス	410
	〃	コウ	7
仁	△	ヒトシ	104
	△	マサシ	27
	△	ジン	9
	△	シノブ	4
	頭	ニ	146
	〃	ジン	75
	〃	ヒト	74
	〃	ヨシ	7
	〃	キミ	3
	〃	ジ	3
	〃	マサ	3
	〃	サト	1
	〃	シノ	1
	〃	トシ	1
	〃	ミ	1
	〃	ヤス	1
	中	ニ	33
	尾	ヒト	100
	〃	ジ	16
	〃	ジン	12
	〃	トシ	1

文字	位置	読み方	頻度
徳	△	トク	14
	△	ノリ	5
	△	イサオ	1
	頭	トク	1062
	〃	ノリ	101
	〃	ヨシ	4
	〃	ヤス	2
	〃	イツ	1
	尾	ノリ	333
	〃	トク	162
	〃	ヨシ	1
之	△	ヒサシ	1
	頭	ユキ	54
	〃	クニ	1
	〃	シ	1
	〃	ノ	1
	中	ノ	1550
	尾	ユキ	2053
	〃	ジ	50
	〃	ノ	22
	〃	シ	18
	〃	ヤ	1
	〃	ヨシ	1
助	△	タスク	5
	頭	スケ	112
	〃	ジョ	1
	尾	スケ	2007
	〃	ヨシ	9
	〃	ズケ	1

この一次をまとめてみると表3のようになります。

康仁

表4

順位	読み方	ウェイト計算					カナ文字連結性
1	ヤスヒト	988	×	100	=	98,800	○
2	ヤスジ	988	×	16	=	15,808	○
3	ヤスジン	988	×	12	=	11,856	○
4	コウヒト	111	×	100	=	11,100	○
5	コウジ	111	×	16	=	1,776	○
6	コウジン	111	×	12	=	1,332	○
7	ヤストシ	988	×	1	=	988	○
8	ミチトシ	2	×	100	=	200	○
9	ノブトシ	1	×	100	=	100	○
10	ミチジ	2	×	16	=	32	○
11	ミチジン	2	×	12	=	24	○
12	ノブジ	1	×	16	=	16	○
13	ノブジン	1	×	12	=	12	○
14	ミチトシ	2	×	1	=	2	○
15	ノブトシ	1	×	1	=	1	○

徳之助

表5

順位	読み方	ウェイト計算					カナ文字連結性
1	トクノスケ	1062	×	2007	=	2,131,434	○
2	ノリノスケ	101	×	2007	=	202,707	○
3	トクノヨシ	1062	×	9	=	9,558	○
4	ヨシノスケ	4	×	2007	=	8,028	○
5	ヤスノスケ	2	×	2007	=	4,014	○
6	イツノスケ	1	×	2007	=	2,007	○
7	ノリノヨシ	101	×	9	=	909	○
8	ヨシノヨシ	4	×	9	=	36	○
9	ヤスノヨシ	2	×	9	=	18	○
10	イツノヨシ	1	×	9	=	9	○
	トクノズケ	1062	×	1	=	1,062	×
	ノリノズケ	101	×	1	=	101	×
	ヨシノズケ	4	×	1	=	4	×
	ヤスノズケ	2	×	1	=	2	×
	イツノズケ	1	×	1	=	1	×

次にこの表をもとにして「康仁」「徳助」にふり仮名を付けてみます。「康仁」の場合は康は先頭に位置しますのでヤスコウミチノブと読むことが可能です。また「仁」は末尾に位置しますのでヒトジジントシと読めます。これらおのおのの読み方と出現頻度により読み方のウエイト計算を1順序づけを行なうと表4のようになります。「徳之助」についても同じように読み方と出現頻度をもとにしてウエイト計算を行ない順序づけをすると表5のようになります。「之」は読み方がノとだけしか読まないためウエイト計算から省きました。

「徳之助」の場合トクノスケ ノリノスケ ヨシノスケ ヤスノスケ イソノスケのふり仮名が考えられます。

イがノの次にスが接続することがありません。そのためこの5つの読み方の可能性はありません。5つの読み方が発音しても妙なことに気づくことでしょう。

このように「康仁」「徳之助」の読み方は「ヤスヒト」「トクノスケ」が順位1位となり正しくふり仮名ふりが行なえました。

参考までに運転免許証では「康仁」「徳之助」は「ヤスニ」「トクユキ」となっています。

この方式の長所は
1. ファイルの容量が小さくてよくミニコンでも十分処理できる。
2. 単語を集めなくてもよい。
3. 有限のもので無限の組合せ対応汎用性がある。

欠点は
1. いろいろな読み方が発生する可能性がある。
2. 特殊な読み方に対して弱い。

があります。

コンピュータによるふり仮名づけの2つの方式を組み合わせるとさらに良い結果が得られます。

4. コンピュータによる実験の準備。

日本人の姓名（姓10万，名15万）のデータを分析し表3のようなものを作成しなければなりません。このためにはデータ収集のために大変な苦労がかかります。そこでこの作業をなんでも簡単にする方法を考え，機械的処理と単純作業により処理することを考えました。

1. 表6は姓，名ファイルを単純リストしたものです。

2. 表7は姓名の漢字の個数だけデータを増加し1桁づつデータをずらし分類しました。

3. 表7を見ながら表8のデータを作る。

4. 表8のデータをファイルに収めた。この作業を姓，名について行った。

5. コンピュータによる実験。

4.で準備した漢字カナ変換データにもとづき，表4，表5ような計算を行いシュミレーションを行った。

実験のデータ・テープは適当なデータがないため，姓名のテープ（姓10万，名15万）を用いた。

この姓名ファイルには姓，名を単純に集めたもので但しの姓，名にウエイトを加えて（佐藤，鈴木，田中などは多いのでその頻度を考える）処理すべきであるがそれは行なわなかった。

表6

SEQ-NO	よ み	姓	SEQ-NO	よ み	姓
1000193	アイツキ	相槻	1000237	アイハナ	相花
1000194	アイツチ	合土	1000238	アイハマ	合浜
1000195	アイツボ	合坪	1000239	アイハラ	合原
1000196	アイツヤ	會津屋	1000241		相原
1000197	アイヅ	会津	1000244		逢原
1000198		合津	1000245	アイバ	会場
1000199		相津	1000246		合場
1000200	アイヅカ	相塚	1000247		合庭

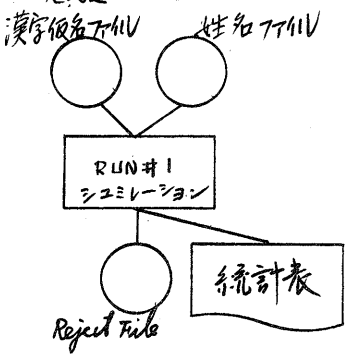
表7

NO.	文字の位置と移動	CODE	読み方	名　前
85139	琢路	017746	タクジ	琢路
85140	琴路	017746	キンジ	琴路
85141	琴路	017746	キンヂ	琴路
85142	琴路	017746	コトジ	琴路
85143	琴路	017746	コトヂ	琴路
85144	環路	017746	カンジ	環路
85145	環路	017746	タマジ	環路
85146	甚路	017746	シゲジ	甚路
85147	甚路	017746	シゲヂ	甚路
85148	甚路	017746	シゲミチ	甚路
85149	甚路	017746	ジンジ	甚路
85150	甚路	017746	ヂンジ	甚路
85151	由路	017746	ユウジ	由路
85152	由路	017746	ヨシジ	由路
85153	由路	017746	ヨシミチ	由路

表8

SEQ.NO	文字	論理コード	位置	頻度	音/訓	仮名文字	旧番号
16441	鼎	22744	尾	2		カネ	16476
16442	鼓	22747	頭	1		ツツミ	16481
16443			尾	1		コ	16480
16444	齢	23006	頭	2		トシ	16484
16445			尾	3		トシ	16482
16446			尾	2		ヨ	16483
16447	龍	23024	頭	45		タツ	16488
16448			頭	29		リユウ	16487
16449			頭	1		ウタ	16491
16450			頭	1		タキ	16490
16451			頭	1		ロウ	16489
16452			中	1		リユウ	16492
16453			尾	5		タツ	16486
16454			尾	4		リユウ	16485

実験プロセス



姓名ファイルには漢字の姓名とふりがなが入っているため次のような処理を行った。漢字の姓名により漢字仮名ファイルの表により表4，表5で示したふりがな付の計算を行い文字の連結性を調べコンピュータで作成したふりがなを五十音順に分類し姓名ファイルのふりがなかテーブルの何個目で一致するか調べてみた。このテーブルと一致しなかったものはリジェクトファイルにアウトプットした。

表9，表10の分析結果は次のように見る。

順位1はコンピュータ内部で発生エセウェイト計算の最大のものと姓名ファイルのふり仮名とが一致した件数である。順位2はウェイト計算で2位のものと姓名ファイルのふり仮名とが一致した件数である。

この分析結果を棒グラフにしたものが図1，図2である。

この結果は予想以上の良い結果であった。

6. 実験結果の分析。

- 姓名の漢字仮名変換をディスプレイを使用して行う場合1つの画面に10～20ケ対応する仮名文字を表示すれば十分であることがわかった。

- この実験の処理時間
  1件の処理に大型計算機で200ms程度必要であった。
  しかしこれは各種統計を取るために色々な計算を行っているので大変な処理時間を必要としているが実際のシステムでは50ms以下で十分処理できると思われる。

- 姓の実験結果より，名の実験結果が良いのは実験の途中でシステムの改良を行ったためである。

- リジェクトファイルが1割程度発生したがこの原因は次の理由による。

  1. 読みがなのつかないもの
     世界 ヒロシ，東海林 ショウジ

  2. 変換テーブルの不備，ミス

  3. 途中にノのあるもの
     井上 イノウエ

  4. くり返し記号のもの
     々ゝこゞ等
     佐々木 ササキ

1.3.以外のものはプログラムの改良により処理可能である。この処理を行えばリジェクトファイルはごくわずかになる予定である。

7. 次期システム。

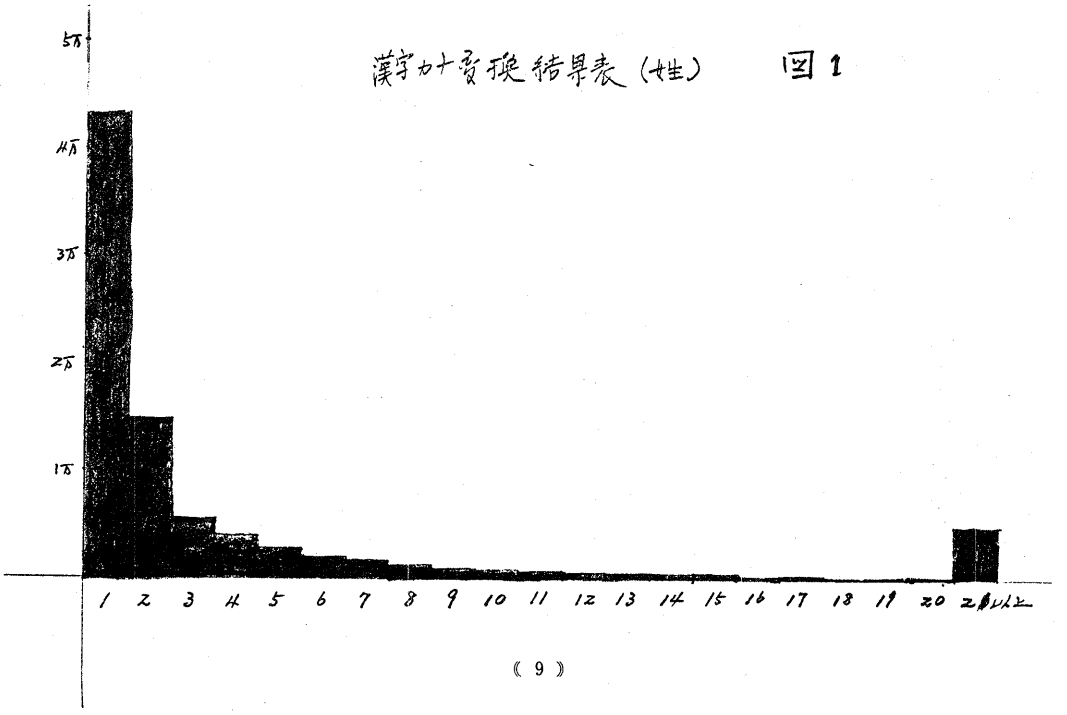
この実験結果にもとづき図3の漢字ディスプレイを使用したふり仮名付システムを日本ユニバック総合研究所と協力し作成する予定である。

表9

姓の分析結果

順位	件数	%	累積%
1	43,317	49.869	49.870
2	15,240	17.545	67.415
3	6,648	7.653	75.069
4	4,020	4.628	79.697
5	2,787	3.208	82.905
6	1,935	2.227	85.133
7	1,553	1.787	86.921
8	1,138	1.310	88.231
9	971	1.117	89.349
10	771	0.887	90.236
11	715	0.823	91.059
12	480	0.552	91.612
13	491	0.565	92.177
14	396	0.455	92.633
15	381	0.438	93.072
16	281	0.323	93.395
17	294	0.338	93.734
18	234	0.269	94.003
19	258	0.291	94.272
20	186	0.214	94.486
21以上	4,764	5.484	100.000
合計	86,860.	100.000	——

リジェクト件数
14,873件

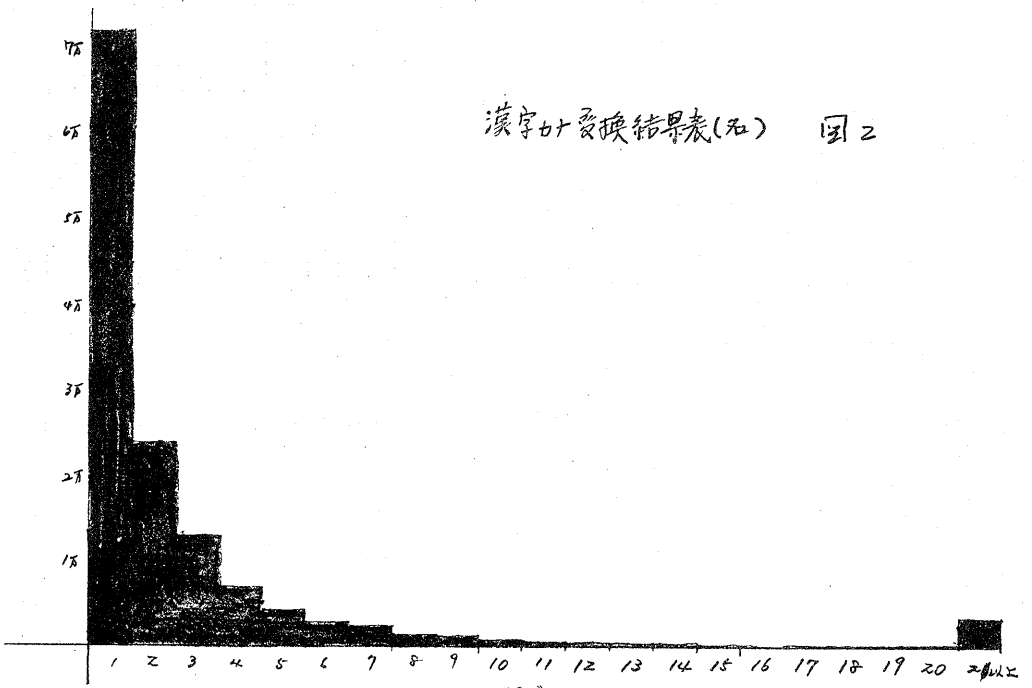


漢字カナ変換結果表（姓）　図1

表10　　名の分析結果

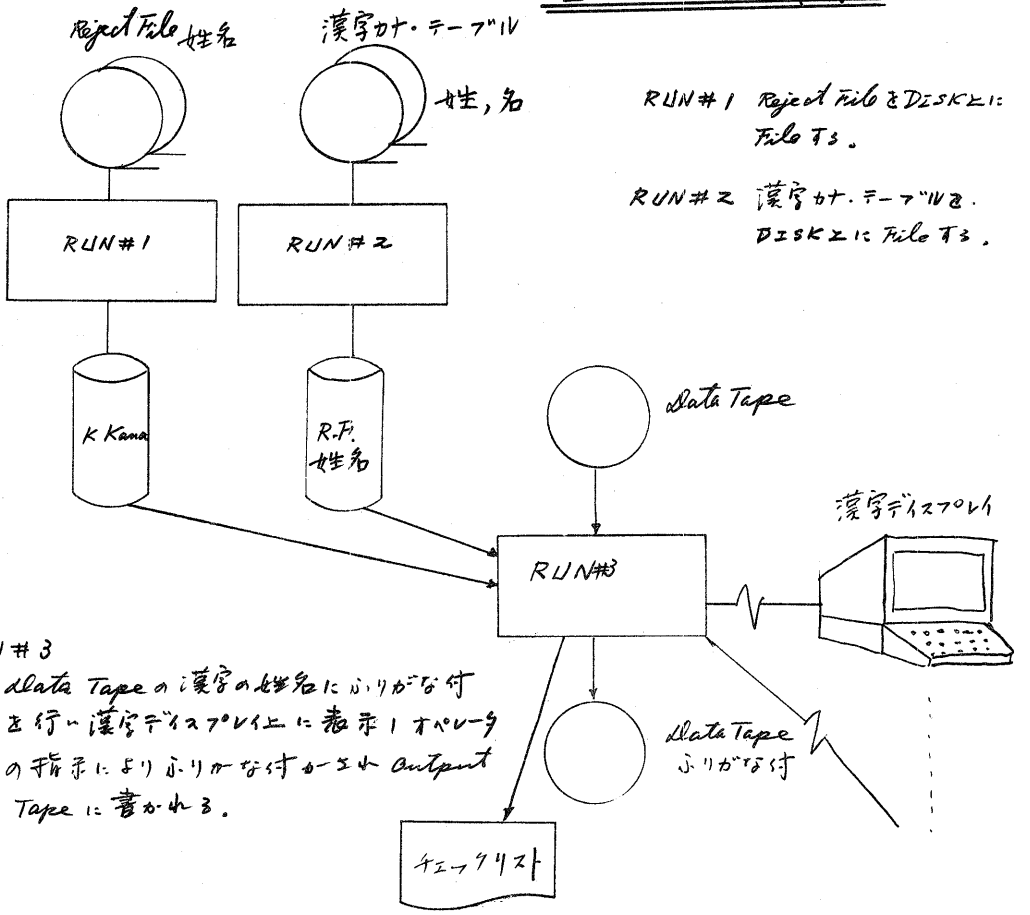
順位	件数	%	累積%
1	72,588	51.968	51.968
2	24,454	17.507	68.475
3	13,324	9.539	79.014
4	7,082	5.070	84.084
5	4,457	3.190	87.275
6	2,902	2.077	89.353
7	2,408	1.723	91.077
8	1,523	1.090	92.167
9	1,414	1.012	93.179
10	1,152	0.824	94.004
11	872	0.624	94.628
12	618	0.442	95.070
13	696	0.498	95.569
14	490	0.350	95.919
15	439	0.314	96.234
16	394	0.282	96.516
17	385	0.275	96.791
18	292	0.209	97.000
19	295	0.211	97.211
20	229	0.163	97.375
21以上	3,664	2.623	100.000
合計	150,928	100.00	——

リジェクト件数
11,250件



漢字カナ変換結果表(私)　図2

図3. システム完成図



RUN#1 Reject File を DISK に File する。

RUN#2 漢字カナ・テーブルを DISK に File する。

RUN#3
Data Tape の漢字の姓名にふりがな付を行い漢字ディスプレイ上に表示しオペレータの指示によりふりがな付きで Output Tape に書かれる。

8. おわりに.

このシステム開発してくださった多くの方たちに感謝致します。特に文字連結表の統計表を作成してくださったシステム統括第一部システム推進部安藤恵美子さん，日本ユニバック総合研究所の方々，データ・チェックをして下さった池谷静香さんに感謝します。
中村祥次部長にはこのシュミレーションプログラムを作成してもらいました。

9. このシステムについての問い合せ先

日本ユニバック株式会社 システム統括第一部システム推進部 田中康仁
(03)585-4111 (EX 3218)
東京都港区赤坂2-17-51

漢字カナ変換システム開発過程とスケジュール

0. ～ 1975年7月　思考時期
1. 1975年8月～12月　データ収集
   ファイルの作成
2. 1975年1月～2月　概要説明書の作成
3. 1976年3月～4月　シュミレーション・プログラムの作成
   ファイルの作成（更新）データ・チェック
4. 1976年5月～6月　シュミレーション実施，
   結果の分析
5. 1976年7月～10月　漢字ディスプレイを用いたシステムの開発予定
6. 1976年11月～12月　評価，次の計画立案

参考文献

1. 人名辞典（番号案内用）　名古屋電話番号案内局　昭和46年6月1日発行
2. 計量言語学　ヨミガナ方式によるカナの漢字変換　田中章夫（国立国語研究所）
3.