計算言語学 21－4
（1980. 3. 14）

大規模漢字データの検証（姓，名ファイルを用いて）

田中康仁　日本ユニバック㈱

はじめに

漢字システムの導入は社会に省力化と効率化をもたらしている。しかしその影で漢字の入力とか，校正作業というあまりよくない作業や，単純作業を産み出している。このような分野の研究が十分行われないと，一方での合理化効果は他方の悪い作業の犠牲の上に成り立つことになる。これではせっかくの効果も大きな立場からながめると意味のないものになる。

漢字入力については色々な方式が検討されている。しかしこれら入力方式はスピードばかりに重点がおかれ，データの品質についてはあまり考えられていない面がある。漢字入力のスピードが上がっても校正作業が相変らず手作業で膨大な時間と経費をかけていたのでは漢字入力のスピード向上もあまり意味がなくなってしまう。そこでここでは漢字データの誤りの自動抽出について考えてみることにする。

1　誤りデータについて

1－1　誤りデータの処理

誤りデータの処理については次のことに注目しなければならない。

① 誤りデータの検出（誤り自動抽出，……）

② 誤りデータの修復（誤りの発見にともない，これを修正する方法）

①の誤りデータの検出は数値データ，カナ文字データについて長いデータ・プロセシングの歴史の中で方法論がみつかっているが，漢字データについてはほとんど研究されていないのが実情であり，誤りは人手によって見つけ出すものという考えが定着している。これでは漢字入力費用を安くするとか入力の処理時間を短縮するとか，品質を向上させるといった問題が困難になってくる。

②の誤りデータの修復については漢字エディターや漢字ディスプレイ等の方法が，バッチ処理による文字修正などの研究，実用化によりかなり見通しがついている。

ここでは特に①の誤りデータの検出について考えてみることにする。

1－2　大規模データの誤りの特徴

誤りを含んだデータは，データ・プロセスの各部門で問題を引き起こすものである。又いくら立派な処理システムが完成しても誤りデータが多ければ全体が無意味なものとなる。そこで誤りデータの特徴について調べてみる。

① 小規模のデータを取り扱っている実験室や研究室では通常考えられないようなことが大規模システムのデータの中には発生する。（誤りデータの量的拡大は質的変化をもたらす。）

② 誤りデータの除去を考えるとき数千件のデータの誤りは簡単に調べることができ，修正することも考えられるが，数１０万件，数１００万件のデータでは誤り除去の方法について１つのサブ・システムとしての方法を考えなければならない。

③ 誤りデータは放置すれば累積する傾向があり，このため早めに除去しなければならない。

④ 毎日発生するデータでは数千件のデータでも誤りの検出，修正の方法について大規模データの取扱いと同じように考えなければならない。

このような特徴から大規模データの誤りの調査は大きな問題であり，又システマテイックに解決方法を考えなければならないということが判る。

2　誤りデータの自動抽出方法と実験結果

ある大規模の漢字姓名データと当社で持っている姓，名ファイルを照合し，アンマッチ・データを出し，その中から正しいデータと誤りデータを見つけ出す。

姓名マスターのカナ文字と漢字に一致するものは正しいとみなし，それ以外は誤りである可能性が強いものとしデータの分離して，その分離されたデータをさらに調べる。

例えば次のような誤りがある。

姓について

カナ	漢字	カナ	漢字
アイズ	今津 → （今→会）	トミヨ	トヨミ →（カナの誤り）
シシド	宍戸 →（宍→宍）	トシヤ	敏夫 →（トシヤ→トシオ）
マスダ	竹田 →（竹→升）	ヒデユキ	英芝 →（芝→之）
クリハラ	粟原 →（粟→栗）	フミコ	矢子 →（矢子→文子）

このような誤りは原票を調べるまでもなく誤りと判断のつくものである。

実　験　(1)

実験結果

①　対象件数　：　1,499,000

②　調査日時　：　昭和５２年１２月 ～ 昭和５３年６月

③　実験に使用した姓名マスターの件数

姓　　１０７,５４９件

名　　１５０,９３０件

④　実験結果

	全データ	アンマッチ・データ	エラー
姓	1,499,000件	38,688件	6,728件
	（１００%）	（２５９%）	（０４５%）
名	1,497,526件	１１３,７２２件	7,621件
	（１００%）	（７５%）	（０５１%）

アンマッチ・データは姓，名マスターに含まれないデータである。この中より誤りデータを抽出したものがエラーである。

この誤りデータをさらに分析すると誤りの種類は次のようになっている。

①　カナ文字の誤り　　６割　　③　原票の書きちがい　　２割

②　漢字入力の誤り　　２割

実　験　(2)

①　対　象　件　数　：　6,717,238件

②　調　査　日　時　：　昭和５３年９月～昭和５４年８月

③　実験に使用した姓名マスターの件数

姓　　１２７,４３１件

名　　２６３,２７１件

④　実　験　結　果　：

	全データ	アンマッチ・データ	エラー
姓	6,717,238件	２０２,１４０件	８１,６８０件
	（１００％）	（３.０％）	（１.２％）
名	6,717,238件	４１１,３５３件	１０３,４０５件
	（１００％）	（６.１２％）	（１.５３％）

アンマッチ・データはエラー・データを含んだ件数である。さらにアンマッチ・データからエラー・データを取り除くと次のようになる。

	正しいデータ	種　類
姓	１２０,４６０件	２８,９８６件
名	３０７,９４８件	１７５,２１３件

姓名マスターに新しく見つかった種類を追加すると次のようになる。

姓　　１５６,４１７件　　名　　４３８,４８４件

このように姓名マスターが充実することにより，次のデータ・チェックではアンマッチ・データの中にはエラー・データがほとんどで正常データは少ないという状態になるであろう

3. 実験結果の考察

- これら2つの実験に使用したデータはいづれも外注会社で校正を行い納品したデータである。
  これからもわかるように，人手の校正がいかに不十分であるかがよく判る。
  この意味からも機械的チェックによる校正の偉力を感じさせられる。

- 実験(1)では150万件のアンマッチ・データ（姓,名 合計件数）が約15万件ありこの校正に約1ケ月の労力を要した。
  又実験(2)では670万件のアンマッチ・データ（姓,名 合計件数）が約61万件あり，この校正に約3ケ月強の労力を要した。
  今後は姓名マスターの充実により労力はもっと減るであろう。

- 実験(1)で得られたアンマッチ・データのうち正しいデータを姓・名マスターに追加したため実験(2)のアンマッチ・データの中で正しいデータの割合は減っている。
  これはマスターを充実させることによりこのエラー・チェック・システムがより充実していくことを示している。

姓について

	アンマッチ・データ	アンマッチ・データ中の正しいデータ	エラー・データ
実験(1)	38,688件 （2.59%）	31,960件 （2.14%）	6,728件 （0.45%）
実験(2)	202,140件 （3.0%）	120,460件 （1.8%）	81,680件 （1.2%）

名について

	アンマッチ・データ	アンマッチ・データ中の正しいデータ	エラー・データ
実験(1)	113,722件 （7.5%）	106,101件 （6.98%）	7,621件 （0.52%）
実験(2)	411,353件 （6.12%）	307,948件 （4.59%）	103,405件 （1.53%）

実験(1)の姓について

$$\frac{\text{アンマッチ・データ中の正しいデータ}}{\text{（全データ）－（エラー・データ）}} = \frac{214}{9955} = 0.0214$$

実験(2)の姓について

$$\frac{\text{アンマッチ・データ中の正しいデータ}}{\text{（全データ）－（エラー・データ）}} = \frac{18}{988} = 0.0182$$

実験(1)と実験(2)を比較して正しいデータの分離が実験(2)で少なくなっていることが判る。

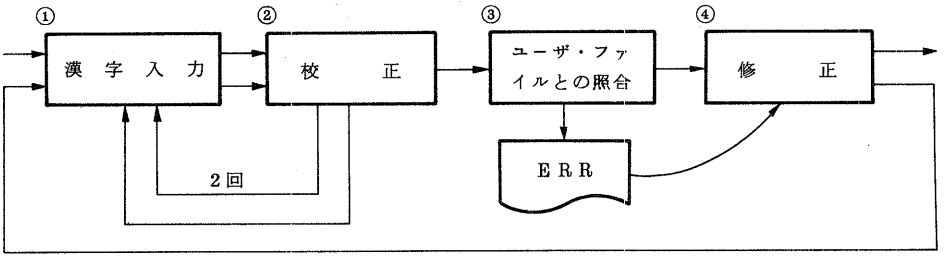
名についても姓と同じように上式で計算すると
実験(1)の場合，実験(2)の場合で 0.0701，0.0466 となり，これを比較して正しいデータの分離が実験(2)で少なくなっていることがわかる。

4　人手による校正と姓名ファイル照合による校正について

4－1　作業プロセス

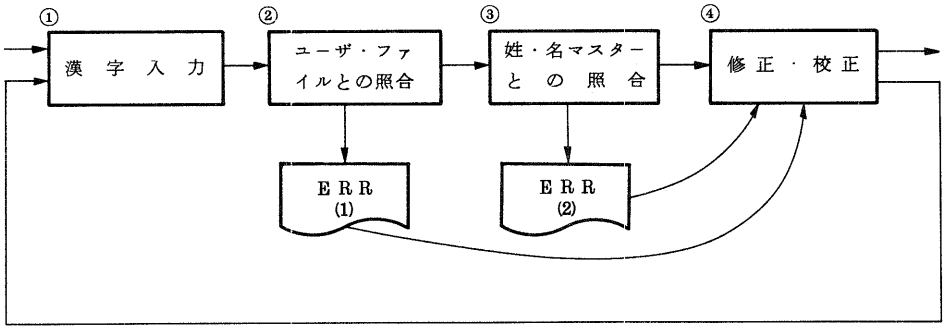
姓名ファイルと照合することにより実際の校正作業のプロセスがどのように変わるか比較してみると次のようになる。

人手による校正方式



① は原票を見ながら漢字の姓名を入力する。
② 校正は人手によって1文字1文字チェックする。
③ ユーザ・ファイルとの照合によりデータの抜け，重複をチェックする。この作業はコンピュータ処理で行う。
④ 修正伝票を起し再度漢字入力を行う。

姓名ファイルと照合する方式



① は原票を見ながら漢字の姓名を入力する。
② ユーザ・ファイルとの照合によりデータの抜け，重複をチェックする。この作業はコンピュータで行う。
③ 姓・名マスターとの照合により誤りらしきデータの抽出を行う。この作業はコンピュータで行う。
④ 修正伝票を起し再度漢字入力を行う。
修正伝票を発行するための誤りデータ，誤りデータらしきものが，すでにコンピュータ処理で見つけ出されているためこの作業は従来のものにくらべはるかに楽である。
校正もアンマッチ・データ特に注目しておけばよい。

4－2　人手による校正方式と姓名ファイルと照合する方式の特徴

次に，これら両方式の特徴について述べてみる。

人手による方式の特徴

① 校正者の能力によって，品質にむらが起る。
② 多勢の人を管理しなければならない。
③ 少量のデータには向いているが多量のデータには向かない。
④ 校正者の作業場所，労働条件の改善などで多くの費用が必要，短期間の場合には割安である。

姓名ファイルと照合する方式の特徴

① 人手の介入が少ない。校正者の能力によって作業品質の高低がない。約1／10程度の人手で済む。
② 品質の保障が得られる。誤りの発見が早い。
③ 処理時間が早い。（大量のデータを処理すると）数万件～10万件程度以上の場合。
④ コンピュータ処理が中心である。
⑤ 大量データの校正には安く処理できる。
⑥ 校正作業にともなう精神的負担が少なくてすむ。

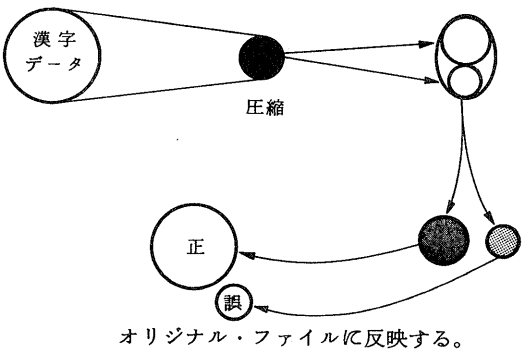
5. 姓名ファイルを持たない場合のエラー・チェックについて

姓名ファイルを持っていない場合はこのようなエラー・チェックが行えないが，特にこのようなファイルが無い場合でも少し工夫をすることにより省力化した方法が考えられる。

次にその方法について述べる。

① オリジナル・データは10～20万件以上のデータとする。
② 姓，名にデータを分け同一の姓，名をおのおの1件としてその件数を加える。このようにするとファイルの圧縮がはかれる。
③ 頻度3件以上のグループと頻度2件以下のグループに分ける。これはデータのエラー・チェックを行いやすくするものである。前者は誤りが一般的に少なく，後者は誤りを比較的多く含んでいる。
④ エラー・チェックを行い誤ったデータは削除し正しいデータ・ファイルを作る。データ・チェックは2回以上行う。姓，名ファイルの小型ファイルが完成する。
⑤ オリジナル・ファイルと④で作られたファイルを照合し誤りデータだけを抽出する。

以上の内容を判りやすく図にすると次のようになる。



データ・チェックを行いやすくするため頻度3件以上のグループと頻度2件以下のグループに分けチェックする。

正しいデータ
誤りデータ
} に分離する。

姓名ファイルを持たない場合のエラーチェックとして漢字1字づつに姓名に使われるカナ文字を付けておき，機械的にこれらカナ文字を組合せユーザ・ファイルにあるカナ文字と一致するか否か調べ，一致すれば正しいとみなし，一致するものがなければ誤りらしいとして再度調べる方法が考えられる。

しかし，この方法は誤りらしいデータが多く抽出されるので 問題が多いであろう．

6. おわりに

ここでは姓名という分野の校正作業について考えたが，この考え方を拡大し，日本語の分かち書き，辞書による照合を行えば一般文の校正にもこの考えはあてはまるであろう。

7. 参考文献

〔1〕 田中康仁 "漢字データのチェック方法（名前のチェックについて）"
SYSTEMS 3，4月号 1975年

〔2〕 田中康仁 "漢字データのチェック方法（名前のエラー・チェックについて）"
SYSTEMS 11，12月号 1975年

添　付　資　料

ｼﾅﾀﾆ	新谷
ｼﾅﾀﾞ	品田
ｼﾉｲ	椎野
ｼﾉｻﾞｷ	篠塚
ｼﾉﾀﾞ	篠崎
ｼﾊﾞｳﾗ	柴原
ｼﾊﾞｶﾞｷ	柴田
ｼﾊﾞｻﾜ	柴沢
ｼﾊﾞﾀ	柴田
ｼﾊﾞﾀ	芝崎
ｼﾊﾞ	芝登
ｼﾌﾔ	渋谷
ｼﾌﾞﾔｼ	渋谷
ｼﾌﾞﾔｼ	澁谷
ｼﾌﾞﾔ	澁江
ｼﾏｲﾁ	島
ｼﾏｻﾞｷ	島■
ｼﾏｽｴ	■末
ｼﾏｾ	島崎
ｼﾏｾ	嶋瀬
ｼﾏﾀ	島
ｼﾏﾀ	舌間
ｼﾏﾀﾞ	■田
ｼﾏﾄﾓ	島本
ｼﾓﾀ	トタ
ｼﾓﾀﾞ	下田尾
ｼﾓﾄﾌﾀﾅ	下栃棚
ｼﾓﾄﾖﾄﾞﾒ	下豊
ｼﾓﾄﾘ	霜島
ｼﾓﾊﾗ	下村
ｼﾓﾑﾗ	下平
ｼﾓﾔﾅ	下■
ｼﾓ	■
ｼﾓ	卞
ｼﾔｱ	車
ｼﾔﾘﾀ	捨田利
ｼｭｳﾄﾞ	清土
ｼｭﾀﾞ	柴田
ｼｮｳﾊﾞﾔｼ	荘林
ｼﾗｲ	白戸
ｼﾗｲ	白水
ｼﾗｲ	白田
ｼﾗｻｶ	反坂

実験(1)のエラー・リスト・サンプル（姓）

ｻﾀﾞｵ	■男
ｻﾀﾞｷ	忠嘉
ｻﾀﾞｺｳ	貞子
ｻﾀﾞｺ	サタ子
ｻﾀﾞｺ	えだ子
ｻﾀﾞﾉﾘ	完訓
ｻﾀﾞﾉ	サダソ
ｻﾀﾞ	サタ
ｻﾁｴ	サチコ
ｻﾁｺ	■智子
ｻﾁｺ	■致子
ｻﾁｺ	幸江
ｻﾁｺ	幸子
ｻﾂ	サワ
ｻﾄｼ	■
ｻﾄﾐ	つとめ
ｻﾄﾐ	覚己
ｻﾄ	さそ
ｻﾉ	さと
ｻﾌﾞﾛｳ	信也
ｻﾌﾞﾛｳ	喜三郎
ｻﾌﾞﾛｳ	已三郎
ｻﾌﾞﾛｳ	正一
ｻﾌﾞﾛｵ	三武郎
ｻﾜ	サク
ｻﾝｳｴﾓﾝ	三右ェ門
ｻﾝｺ	ミ子
ｻﾝｿﾞｳ	■三
ｻﾞｴﾓﾝ	左ェ門
ｻﾞｺ	和子
ｼｲｷﾁ	末吉
ｼｲﾀｶ	栄孝
ｼｶ	さか
ｼｹﾞｵ	宏
ｼｹﾞｵ	哉夫
ｼｹﾞｵ	哉雄
ｼｹﾞｷ	哉樹
ｼｹﾞｺ	シケ子
ｼｹﾞｺ	しげる
ｼｹﾞｺ	小志げ
ｼｹﾞｺ	茂
ｼｹﾞｼｻ	重久
ｼｹﾞｼﾞ	繁一
ｼｹﾞﾀｶ	■孝
ｼｹﾞﾉﾌﾞ	信重

実験(1)のエラー・リスト・サンプル（名）