ICOT Technical Report: TR-124

TR-124

計算機による日本語の用語・固有名詞の校正

石井　暁（東芝）

July, 1985

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03) 456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

計算機による日本語の用語・固有名詞の校正

石　井　　暁

（株）東芝　情報通信システム技術研究所

概　　要

日本語文章の校正を行う日本語校正支援システムの構築を試みた。校正は用語についてと固有名詞について行った。前者は主に用語集に基づく校正である。後者は企業役員の異動記事と、データベースとの対照を例にとった。処理系の制約から約10分の１のデータを用いて実験を行い、有効性を確認する事ができた。またPrologによりシステムを記述したので、その評価も併せて行い、望まれる性能を明らかにし、得られた知見を述べた。

本研究は、著者が（財）新世代コンピータ技術開発機構（ＩＣＯＴ）に在籍中に、第５世代コンピータの研究開発の一環として行った物である。本研究の機会を与えられ、また、この発表を許されたＩＣＯＴの関係各位に深く感謝するものである。

目　次

1．はじめに

文章の校正作業は計算機化が望まれ、更に計算機化に向いた作業と考えられる。特に新聞記事は守るべき表記の規準がはっきりしており、その校正は計算機化に向いていると考えられる。そこで、新聞社で現在手作業で行われている校正作業を調査した結果、現状の技術で、辞書や用語集を用いた用語の校正と人名録等を用いた固有名詞の校正の２種が実現の可能性が高い事が解っている。［１］

そこで用語の校正と固有名詞の校正について、必要な知識の調査、収集を行い、［２］［３］日本語校正支援システムの構築を試みた。本文では使用した知識の質・量と共に校正の実験結果について述べる。

また、システムの記述はＤＥＣ2060上のＤＥＣ10 Prolog を用いて行ない、Prologの評価も併せ行ったので、その結果についても述べる。

2．用 語 の 校 正

校正システムの構成を一口に述べると第１図に示す様になる。即ちシステムは蓄えられた知識を用いて入力文中の修正すべき箇所の指摘、可能なら更にどう修正すべきかを表示する。実際の動作においてはシステム全体をまとめて一体として実行させる事が処理系の制約のため不可能であったため、用語の校正と固有名詞の校正に分けて実行させた。そこで、この章では用語の校正について、つぎの章では固有名詞の校正について述べる。尚、図の左側の知識ベースの入力、編集については異質の手続き型の処理であり、別に述べる。

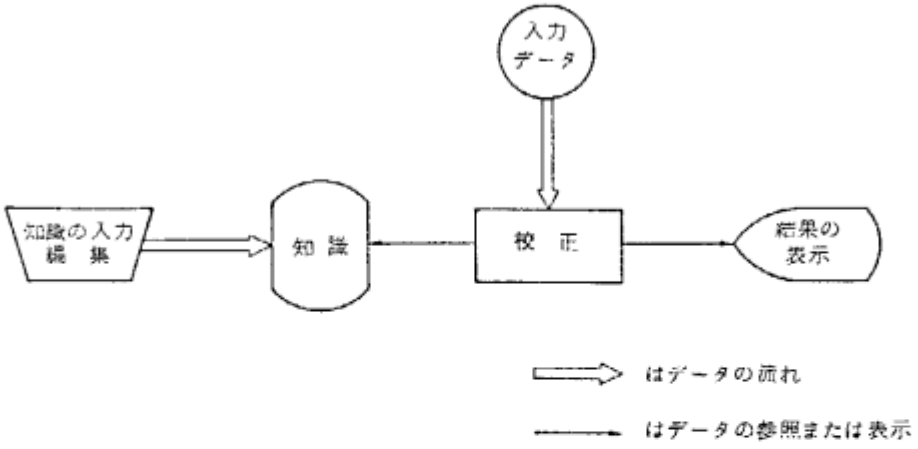
2.1　構成及び使用した知識

用語の校正については更に３段のステップに分けて実行させた。処理系の制約が主な理由であり、また、開発の容易さのためでもある。第２図に従って説明する。

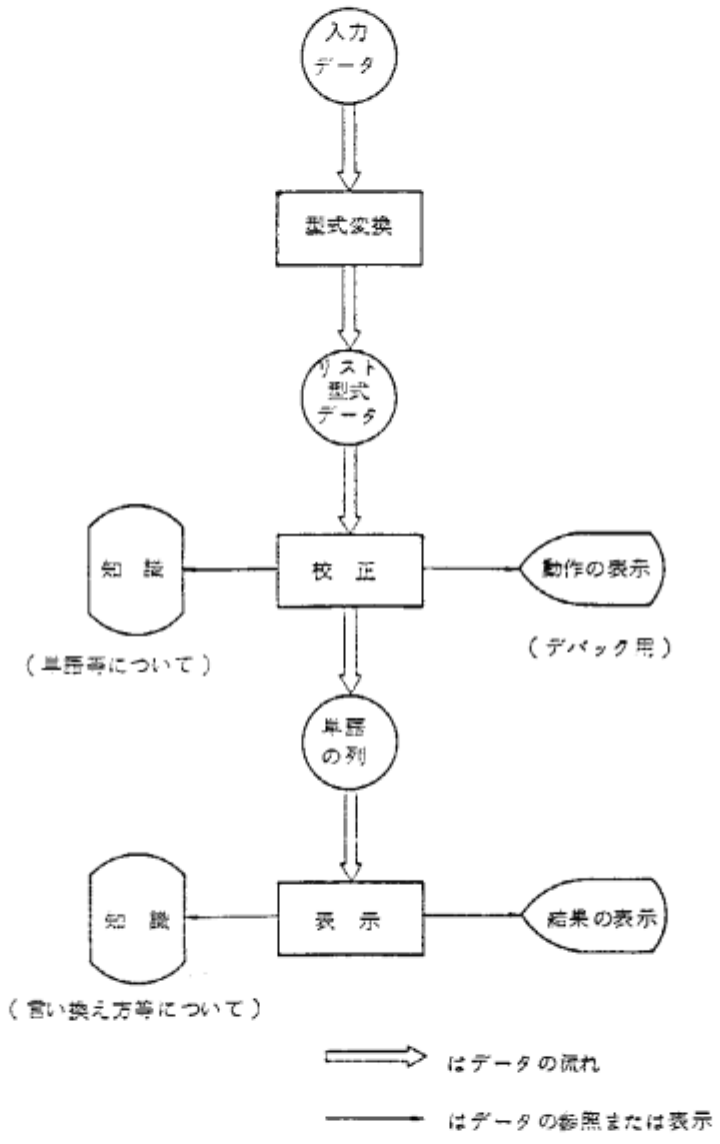
2.1.1　入力データの型式変換

入力データをリスト型式に変換するステップである。リストの各要素がＪＩＳ漢字コードの１文字となる。ただし使用したＤＥＣ 10 Prologには漢字コードがないため、実際は１バイトのascii 文字を２文字使用している。

この部分は手続き的な処理であり、詳細は省略する。



第１図　システムの全体構成


入力
データ
型式変換
リスト
型式
データ
知　識
校　正
動作の表示
（単語等について）
（デバック用）
単語
の列
知　識
表　示
結果の表示
（言い換え方等について）
はデータの流れ
はデータの参照または表示

第２図　用語の校正

2.1.2 校　　正

この３つのステップの中心となる処理であり、リスト型式の入力文を単語分けしつつ校正し、結果をリスト型式で出力する。出力リストの各要素が１単語の情報であり、それも更にリストになっている。

各単語についての出力情報は以下の物である。

単語（入力文での形）
品詞
その単語の終止形
活用形
その語が誤りであるか否か
その語が言い換えるべき語であるか否か
見出し番号

最後の見出し番号は単語を簡単に識別するための数字である。以下、このステップでの処理について、使用する知識と処理アルゴリズムに分けて述べる。

(1) 使用する知識

辞書［４］、常用漢字表［６］、朝日新聞社の用語集［５］を基に次の様な知識を使用した。詳細は既に報告した［３］ので、項目と出典、更にPrologによって表した知識の量を挙げるに止める。

・活用語尾の知識［４］

動詞、形容詞の活用語尾の知識である。約 140行である。

・助詞、助動詞の知識［４］

助動詞の活用の知識、接続条件の知識を含む。約 210行である。

・単語辞書［５］［６］

約27,000語の辞書を用意した。これには誤りとされている語、他に言い換えるべきとされている語も含んでいる。それらについては各々その旨の情報を含ませてある。

(2) 処理アルゴリズム

処理アルゴリズムを第３図に従って簡単に説明する。

校正は、単語切り出しの述語が再帰的に呼び出される事により行われる。単語の切り出しは、まず字種（漢字、ひらがな、かたかな、アルファベット、記号、数字、句読点）の判定を行い、それに基づき、最長一致法により必要な辞書の検索を行う。辞書にない場合は、適当な所まで未知語とする。一応この様に語が切り出されると活用語尾の処理をする。最後に結果を出力して終る。

2.1.3 校正結果の表示

最後のステップが校正結果の表示のステップである。これも使用する知識と処理アルゴリズムに分けて述べる。

(1) 使用する知識

常用漢字表、朝日新聞社の用語集を基に、次の様な知識を使用した。これについても詳細は既に報告した［３］ので、項目と知識の量を挙げるに止める。

・出典の知識

上記の単語辞書の各語についての出典の情報である。約35,000行である。

・言い換え方の知識

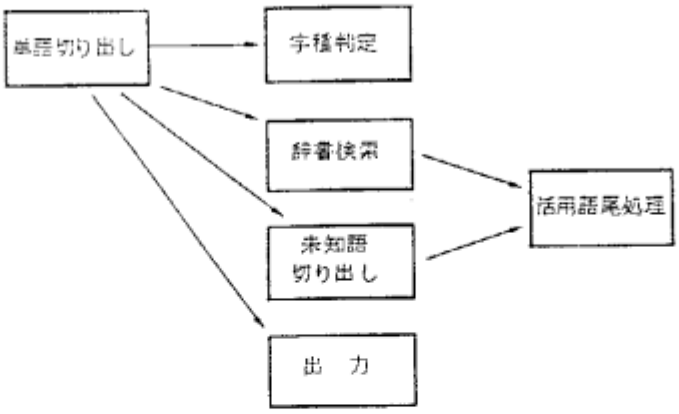
誤りとされている語、他に言い換えるべきとされている語について、代りに用いるべき語の知識である。行数は約 6,900行である。

(2) 処理アルゴリズム

このステップの処理アルゴリズムは前のステップのそれと比較すれば単純であり、基本的には前のステップの出力をそのまま印刷する物である。この際、用語の出典情報を付加し、また用語が誤りであり、または言い換えるべきとされていればそれを表示し、更に代りに用いる用語を表示する。

2.2 校正結果

次にプログラムの実行結果を述べる。ここで、用いた例文は朝日新聞社が1984年 7月 1日から10日までに作成した紙面データより選ぶ様に努めた。ただし、文章を短く



第３図　校正アルゴリズム

したり、誤りを加えたりしたため、紙面と全く同じ文章を扱った例はない。

また、この評価実験において、前に述べた単語辞書、出典の知識、言い換え方の知識について、全てを用いる事が処理系の制約で不可能であり、次の様に一部のみを使った。

- 単語辞書

  活用のない語の知識を除き、記号で始まる語の知識も除いた。その結果、行数は約 3,300行になった。また、語の読みも除いた。

- 出典の知識

  誤り、又は言い換えるべきとされている語についてのみ用いた。その結果、行数は約 6,200行となった。

- 言い換え方の知識

  500行のみを用いた。

プログラムの実行において、システムから何らかの指摘をするのは次の様な場合である。いずれも何らかの問題が存在し、校正が必要である可能性があり、当システムの有効性をうかがわせる物である。

2.2.1 未 知 語

第４図に“足元から雨が降る。”という文章の解析結果を示す。ここで“雨”が未知語となっている。（活用しない語は単語辞書から除いたためである。）未知語を減らすためには、第一に単語辞書の充実が必要がある。

別の問題として、固有名詞の扱いがある。本研究でも固有名詞の知識を扱っては

入力文　　足元から雨が降る。

校正結果 単語	品詞	語幹	活用形	誤り？	言い換え？	見出し番
足元	名詞	足元		ーー	ーー	7320
から	助詞	から		ーー	ーー	0
雨	名詞（推定）	雨		ーー	ーー	ーー
が	助詞	が		ーー	0	
降る	動詞	降	終止	ーー	ーー	40977
。	句点	。		ーー	0	

第４図　用語の校正結果（その１）

いるが、量、質共、改良の余地がある。即ちより多くの人名の収集と共に地名、機関名等、異なった種類の固有名詞の収集が、この種のシステムにおいて重要な課題となる。

2.2.2 誤り又は言い換えるべき用語

次に第５図に“足下から雨が降る。”という文の解析結果を示す。これは前の例と比較して“あしもと”の表記を変えた物である。“足下”という表記は使われず、“足元”という表記を使う様にとの朝日新聞の用語集 245ページの知識が示されている。尚、正しい用語についても出典は表示できる様になっているが、今回の実験では処理系の制約から、正しい用語の出典の知識を格納していないために表示されていない。

2.2.3 ひらがな、かたかなの混用

次に第６図に“再検討すべきだと思う。”という文の解析結果を示す。ここで、“べ”はかたかなが入っている。そのため単語分けがうまく行っていない。ただし、この例は“べ”をひらがなにしても“べき”という文語的な表現がシステムに入っていず、やはり単語分けがうまく行かない。文語的表現の扱いも将来の課題である。

入力文　　足下から雨が降る。

校正結果 単語	品詞	語幹	活用形	誤り？	言い換え？	見出し#
足下	名詞	足下		ーー	言い換え	7 3 0
	言い換えの根拠					
	出典					
		朝日、用字用語集		2 4 5	ページ	
		常用漢字表		8 ページ		
	代わり言葉					
	足元					
から	助詞	から		ーー	ーー 0	
雨	名詞（推定）	雨		ーー	ーー	ーー
が	助詞	が		ーー	0	
降る	動詞	降	終止	ーー	ーー	4 0 9 7 7
。	句点	。		ーー	0	

第５図　用語の校正結果（その２）

2.2.4 かな書きされた語

最後に第７図に“ゆるやかになった。”という文の解析結果を示す。これは常用漢字内の漢字の使い方であってもかな書きされる可能性がある例として取りあげた。処理アルゴリズムでの対応も可能だが、漢字、かなの使い分けにつき、常用漢字表以外の知識の必要性を感じさせる例である。

校正結果

単語	品詞	語幹	活用形	誤り？	言い換え？	見出し#	
再検討	名詞（推定）			再検討	ーー	ーー	ーー
す	名詞（推定）		す	ーー	ーー	ーー	
べ	名詞（推定）		べ	ーー	ーー	ーー	
きだと	名詞（推定）			きだと	ーー	ーー	ーー
思う	動詞	思	終止	ーー	ーー	6 4 3 8	
。	句点	。		ーー	ーー	0	

第６図　用語の校正結果（その３）

入力文　　ゆるやかになった。

校正結果

単語	品詞	語幹	活用形	誤り？	言い換え？	見出し#	
ゆるやかになった	名詞（推定）			ゆるやかになった		ーー	ーー
ーー							
。	句点	。		ーー	ーー	0	

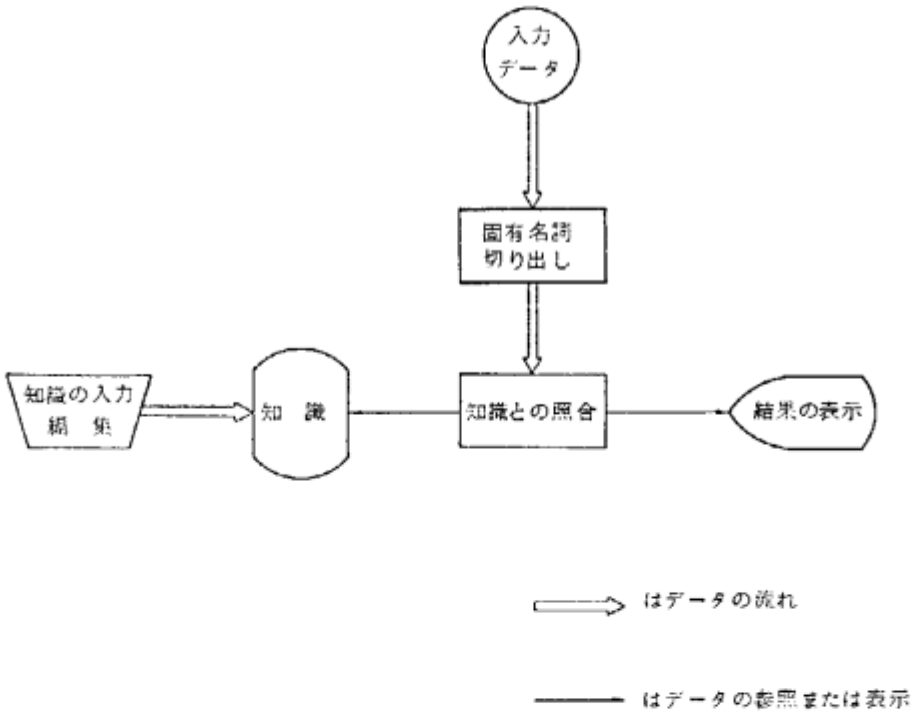
第７図　用語の校正結果（その４）

3．固有名詞の校正

次に固有名詞の校正について述べる。固有名詞、特に、人名、その人の所属する機関及びその人のその中での地位についての校正に当っては適当な人名録の選択が必要となる。即ち入手し易い資料で、かつ、その中の人名がよく文章に表れる物でなければならない。［２］

ここでは上場企業の役員の異動記事についての校正の例を述べる。この種の知識は必ずしも記事に多く表れるわけでないが、資料が既にデータベースとして整備されているため、有効と考えられる。［２］また常用漢字以外の漢字の扱いの一例を併せて述べる。これは固有名詞には常用漢字以外の漢字が比較的多く表れるので、ここで併せて実験した物である。

3.1　構成及び使用した知識



第８図　固有名詞の校正

固有名詞の校正は第８図に示した様に行われる。ここで固有名詞の切り出しに際しては固有名詞が決まったフォーマットの記事で使われる事が多い［２］ため、フォーマットの約束を利用できる事が多く、本実験でもその様にしている。しかし、基本的には前に述べた用語の校正における単語分けと同種の処理であるので、特に述べない。

用いた知識は次の様な物である。

3.1.1 上場企業についての知識

全上場企業について企業名の知識である。東洋経済新報社のデータベースから次の様な述語を作成した。尚、データベースの内容は役員四季報［７］の内容とほぼ同じである。

```
kaisya(1301 , [´極´, ´洋´]).
```

引数は各々証券コード、企業名である。企業名として、正式名の他、省略した名前についても作成した。また利用したデータベースには上場企業以外にも約 100の機関（非上場企業、経済団体、政府機関等）も含まれており、約 2,200の機関の知識である。

行数は約 4,400行である。ここで、漢字等には、ＪＩＳ漢字コードを使用する。ただし、使用したPrologには漢字というデータタイプがないため、実際は ascii文字を２文字使用する。次の 3.1.2で述べる知識についても同様である。

3.1.2 上場企業の役員についての知識

全上場企業の全役員についての知識である。（ 3.1.1で述べた他機関の役員を一部含む）これも東洋経済新報社のデータベースより次の様な述語として作成した。

```
yakuin(1301 , ´Ｄ´, ´Ｙ´, [´志´, ´水´], [´廣´, ´典´],
       [´し´, ´み´, ´ず´], [´ひ´, ´ろ´, ´の´, ´り´]).
```

引数は各々証券コード、役職コード、代表権の有無、姓、名、姓の読み、名の読みである。

行数は約32,000行である。

3.1.3 漢字の知識

常用漢字とその読み方の知識である。常用漢字表に基づく物と朝日新聞社の用語集の漢字表に基づく物を用意した。詳細は既に別に述べた［３］ので省略する。

行数は約 4,100行である。

3.2 校正結果

次に実行結果を述べる。この評価において、処理系の制約のため、以下に示すデータ類のみを用いた。

- 企業名の知識

  証券コード1950から3011までおよび3863の 139社のみの知識を用いた。行数は約 280行である。

- 企業の役員名の知識

  上と同じ証券コードの企業 139社の役員のみの知識を用いた。行数は 3,100行である。また、姓の読み及び名の読みを除いた。

- 漢字の知識

  読み方を除いた。それに伴って１つの漢字については１行のみとしたので、行数は約 1,940行となった。

- プログラム

  人事異動の特定のフォーマットの記事のみを解析できる様な短い物を用いた。

次にいくつかの例を用いて校正結果を述べる。用いた文章は２．で述べた新聞記事から採った。ただし前に述べた様に新聞記事そのものではない。

3.2.1 記事の確認

第９図に示す例がデータの確認を行った例である。まず入力文から会社名等の候補を切り出す。ここで候補とは、切り出しを字面のみから機械的に行っているとい

```
入力文      ◇片倉工業（６月１８日）退任（取締役）守田健二

会社候補      片倉工業
H候補       ６月１８日
新役職候補      退任
旧役職候補      取締役
名前候補      守田健二

会社名あり      コード番号  3001
新役職あり
旧役職あり
名前あり
役職ＯＫ
```

第９図　固有名詞の校正結果（その１）

う意味である。次にこれらの語と格納されている知識を比較する。その結果、記事に問題となる記述が見つからなかったという事を示している。最後の役職ＯＫとは、知識の中では、この人の役職が旧役職（取締役）となっているという意味である。

3.2.2 更新ずれ

次に第10図の例では、前と同様だが、知識の中でこの人の役職が新役職（専務）となっている事を示している。データベースの更新が当システムの目的に合わせて行われていないための更新のずれのためであろう。

3.2.3 常用漢字以外の漢字の扱い

最後の例を説明する。普通に実行すると第11図に示す様に会社名が見つからない。これは、この会社の名前が知識の中に入っているのだが、“十條製紙”と“十条紙”の２つの表記でのみ入っているので、単に探しても見つからないためである。

そこで、常用漢字以外の漢字については特別な扱いをすると、第12図に示す様に会社名が検索され、資料では異なった漢字が使われている事が示される。これは漢字でなくとも（例えばゲタ記号）であっても同様に動作する。

入力文　◇三和大栄電気興業（６月２９日）専務（取締役）鳥山好三

会社候補　三和大栄電気興業
日候補　６月２９日
新役職候補　専務
旧役職候補　取締役
名前候補　鳥山好三

会社名あり　コード番号　1958
新役職あり
旧役職あり
名前あり
役職　新役職とおなじ

第１０図　固有名詞の校正結果（その２）

```
入力文　　◇十条製紙（６月３０日）専務（常務）遠藤健一郎

会社候補　　十条製紙
日候補　　６月３０日
新役職候補　　専務
旧役職候補　　常務
名前候補　　遠藤健一郎

会社名見つからない　　以下無視する
```

第１１図　固有名詞の校正結果（その３）

```
入力文　　◇十条製紙（６月３０日）専務（常務）遠藤健一郎

会社候補　　十条製紙
Ｈ候補　　６月３０日
新役職候補　　専務
旧役職候補　　常務
名前候補　　遠藤健一郎

会社名あり　　資料では　十條製紙　　コード番号　3863
新役職ああり
旧役職あり
名前あり
役職ＯＫ
```

第１２図　固有名詞の校正結果（その４）

4．Prologの評価

既に述べた様に本校正システムの処理系はＤＥＣ2060上のＤＥＣ－10 Prologである。この言語の評価を定量的及び定性的の両面から行った。処理時間は１人の利用者が使っている場合の数字である。

4.1 定量的な評価

定量的な評価としては、扱えるプログラムの大きさ及び処理速度を考える。勿論これらは言語仕様の評価というより処理系の評価という面が強い。しかし純理論的な言語の研究は別として、言語とその処理系は表裏一体であり、この様な評価を行った。

既にシステム構成や機能の評価に伴って当システムで扱えるプログラムの大きさや速度について触れてきた。ここでそれらをまとめると以下の様になる。ここでプログラムのうち大半はいわゆるデータであるので、プログラムの大きさはデータの件数で表す。

(1) 校正に関して

単語辞書は量的に12%のみ用いた。質的には読みが入っていない。また、例にとりあげた程度の短文でないと記憶容量不足で解析できない。速度的には第４図から第６図程度の例で10～15秒程かかる。最長一致のアルゴリズムを“簡単に”コーディングしてある物が、かなり時間がかかっている様である。

(2) 出典、言い換え方について

出典の知識が18%、言い換えの知識が７%のみ、用いた。速度については(1) 程は問題にならない。

(3) 上場企業名、その役員名について

企業名は６%のみ用いた。役員名は量的には10%のみ、質的には読み方を除いて用いた。漢字の知識は質的には読みを除き、それに伴って量的には47%のみを用いた。速度については、第12図に示した漢字の知識を用いる場合で約20秒かかる。それ以外ではかなり高速で、特に問題はない。その差の８～９割前後は常用漢字表の検索に、他が完全マッチングではないリストのマッチングにかかっている様である。

以上を総合すると、今回の実験程度に処理のステップ分けをするとして10～20倍程度のデータの扱いと処理ステップの間の容易な連絡方法が必要となる。もしステップ分けが不可能なら、少なくとも 100倍のデータの扱いが必要となる。また速度的には知識が現在の10～20倍に増えた時で現在の1000倍程度が実用的には必要となろう。勿論これらにはコンパイルの効果やコーディング上の改良の効果も含めて考える。更に、言語仕様の変更や手続き型言語との結合も考える。

ここで使用した処理系で扱える記憶容量は 800Kbyte、処理速度は２Klips程度と

考えられる。

4.2 定性的な評価

定性的な評価としては書き易さ、読み易さ、デバックし易さ、変更し易さ等が考えられている。しかし、これらを客観的に評価する方法は確立されてはいない。そこで、ここでは従来からの手続き型言語を用いて同程度の機能のプログラムを作成し、全体のステップ数、モジュールの数、モジュール当りのステップ数を比較する事により、客観的な評価を試みた。また、最後に得られた知見を列挙する。

(1) 手続き型言語との比較

用語の校正の基礎となる簡単な単語分けプログラムをPrologとFortran（大体水準7000のレベル）で作成し、比較した。このプログラムは極く小規模であり、次の様な物である。

- 単語は名詞、動詞、助動詞、助詞各１語
- 活用についての知識も上記の４単語に必要な物のみ
- 未知語の扱いなし

（ｉ）ステップ数

Prolog	643ステップ
Fortran	834ステップ（common文をinclude）
	1097ステップ（include 文を使わず）

比をとると 0.77 又は 0.59 となる。

（ii）モジュール数

Prolog	80（述語の種類）
Fortran	17（サブルーチン）

平均ステップ数は

Prolog	8
Fortran	50（common文をinclude）
	65（include 文を使わず）

となる。

(2) その他の知見

その他の得られた知見を列挙する。

（ｉ）Prologが優れていると思われる点

- モジュールが理解し易い。モジュール内のステップ数が少なく、またラベルがなく、流れが単調だからである。
- 任意表の文字列の扱い、また任意項数のデータ（リスト）の扱いが容易である。
- common文がなく、副作用が少い。

- 変数名の長さが長くできる。

(ii) Prologで改良の余地があると思われる点

- 人間の通常の考え方との差異がある。例えば、繰り返しを再帰呼び出しで表す場合である。時には、手続き型言語で記述した方が解り易いと思える場合もあり、それとリンクがとれる事が望ましい。
- エラー時に思いがけない動きをする。例えば変数名の綴りをタイプミスした際など、処理系のチェックの充実が可能であろう。

5. おわりに

用語の校正及び固有名詞の校正について、用いた知識や校正結果について述べ、更にシステムを記述した言語であるPrologの評価についても述べた。

用語の校正には辞書から作った活用語尾、助詞、助動詞の知識、朝日新聞社の用語集を中心に常用漢字表も用いて作成した単語辞書、出典の知識、言い換え方の知識を用いた。処理系の制約から約10分の1にあたる約10,000行の知識を用いて動作させ、評価した。その結果、未知語、誤り又は言い換えるべき用語、ひらがな、かたかなの混用、かな書きされた語について校正システムの指摘が得られた。

固有名詞の校正の例として上場企業の役員の異動記事の校正をとりあげた。東洋経済新報社のデータベースから知識を作成し、その10分の1たらずを用い、記事とデータベースの対照を行った。固有名詞には常用漢字外の漢字が比較的多いことから朝日新聞社の用語集から作成した常用漢字の知識を組み合わせた例も示した。

Prologの評価については、今回の実験程度に処理のステップ分けをするとして現在の制約の10～20倍の知識を扱える記憶容量と、処理ステップ間の容易な連絡が必要となる事を述べた。また、その程度の知識量で、現在のインタプリティブな実行の1000倍程度の速度が必要な事を述べた。最後にFortran との比較からプログラムが理解し易い等の知見を述べた。

本研究に当り、既に本文中でも触れた様に朝日新聞社提供の新聞記事のデータ及び東洋経済新報社提供の上場企業の役員データを使用した。また常用漢字表、更に朝日新聞社の用語の手びきの一部を計算機に入力して使用した。データの提供及び資料の使用の承認を頂いた朝日新聞社及び東洋経済新報社の関係各位に深く感謝するものである。また知識の入力、編集に関してはシャープ㈱の援助を頂いた。併せて謝意を表す。

文　　献

［１］石井　暁：“新聞における校正・校閲の実データによる調査”、ＴＲ－０３９、新世代コンピュータ技術開発機構（1983）
これの概要が次の物になっている。
Ishii,S.：“Study of Proofreading Techniques Used at a Japanese Newspaper”、情報処理学会第２８回（昭和５９年前期）全国大会講演論文集（Ⅱ）2M-7、pp.1205～1206、情報処理学会（1984）

［２］石井　暁：“新聞における人名の使い方の調査”、ＴＭ－００６２、新世代コンピュータ技術開発機構（1984）
これの概要が次の物になっている。
Ishii,S.：“Study of a Newspaper's Treatment of Proper Names”、情報処理学会第２９回（昭和５９年後期）全国大会講演論文集（Ⅱ）4J-5、pp.1449 ～1450（1984）

［３］石井　暁：“日本語の漢字・用語の校正のための知識”、ＴＭ－００９２、新世代コンピュータ技術開発機構（1985）
これの概要が次の物になっている。
Ishii,S.：“Knowledge for Proofreading Japanese Word Usage”、情報処理学会第３０回（昭和６０年前期）全国大会講演論文集（Ⅱ）3G-3、pp.1635 ～1636（1985）

［４］久松他：“角川国語辞典”６３版、角川書店（1982）

［５］朝日新聞社用語幹事編：“朝日新聞の用語の手びき”第１８版、朝日新聞社（1984）

［６］大蔵省印刷局編：“常用漢字表”、大蔵省印刷局（1982）

［７］東洋経済新報社編：“役員四季報１９８５年版”、東洋経済新報社（1984）