自然言語処理 51－2
（1985. 9. 13）

新聞記事データを対象とする自動単位切り処理とそれに基づく語彙調査

藤崎博也・亀田弘之（東京大学　工学部）

1．はじめに

言語は時代の状況を反映するが故に、時代の変遷とともに変化していく。従って言語の主要な構成要素のひとつである語彙に着目し、その使用状況を調査することにより、一国の言語や社会の現状を把握しその特徴を論ずることが可能となる。またその結果は言語の標準化や国語教育、更にはワードプロセッサ等の自然言語処理技術のための重要な基礎的資料となる。

日本語に関する大規模な語彙調査としては、昭和41年に国立国語研究所が新聞（朝日・毎日・読売の三紙合計約18日分）を対象に行ったものがあるが、その際の言語資料を単語単位に切り分ける作業（以下「単位切り」と称する）はもっぱら人力にたよったものであり、その結果の集計・作表等に計算機を用いたにすぎなかった[1]。一方その後約20年間の計算機技術の長足の進歩により、最近では新聞・雑誌等の計算機による自動製版が可能となり、また国語辞典も計算機可読な形で提供されるようになり、計算機を利用して大規模な語彙調査を行うための環境が整ってきた。そこで筆者らは上記国立国語研究所の調査を参照しながら、新聞（朝夕刊合計84日分）を対象としてさらに大規模な語彙調査を行った[2~3]。

本報では、まずこの研究の概要を略述し、次にここで採用した自動単位切りの方法について述べ、更に得られた調査結果につき報告する。

2．調査の概要

2-1.語彙調査対象の選定

語彙調査となる対象の言語資料としては、新聞・雑誌・小説・歌詞など種々のものが考えられるが、本研究では日本語の標準化の基礎的資料作成を目的としているため、政治・経済から文化・スポーツ等にいたる多種多様な分野の語彙を偏りなく含んでおり、かつ、日常的に国民大多数の目に触れるという見地から、新聞を対象資料として選定した。

対象とした新聞記事は、昭和57年の朝日新聞朝夕刊合計84日分であり、その全数調査を行った。取り扱った文字量は全体で約1220万文字である。

2-2.自動単位切りのための見出し語の収集と単語単位の統一

本研究では、上述したように大量の言語資料を対象としているため自動単位切りを行った。しかし、既存の単一の辞書ではこの目的には不十分であるので計算機可読な新明解国語辞典（三省堂、第３版）（以下「新明解」と略称）や日本語単語機械辞書（九州大学工学部）（以下「九大辞書」と略称）をはじめ、地名・人名・団体名など種々の語彙資料を併合して単位切り辞書の完備を図った。

しかしながら、これらの辞書は編纂方針や使用用途が異なっているため、辞書相互の見出し語に統一性が欠ける問題点が生ずる。更に、同一の辞書内に於いても必ずしも見出し語に統一性があるわけでわない。例えば「新明解」では、国語辞典としての使い易さを考慮しているため、いわゆる“単語”のみならず、連語、慣用句、諺等も見出し語として記載されている場合がある。また、例えば「具体性」は記載されているが、「抽象性」は記載されていない。従って、これらの辞書・語彙資料から自動単位切りのための見出し語を

選定するにあたっては、語彙調査に適した単語単位に統一する必要がある。

本研究では、上記の辞書・語彙資料の中で辞書としての体裁が最も整っている「新明解」の見出し語を便宜上の出発点としたが、「新明解」にはなく他の「九大辞書」・語彙資料に含まれている見出し語に対しても以下に記す規則によって単語単位の統一を図った。

①記号は１文字で１単語とする。ここに記号とは、日立標準漢字コード（16進）でA1A1～A1A5,A1A7A～A1B2,A1BD～A2AE,AFA1 ～AFC4のものをいう。

②サ行変格活用型の動詞の取り扱いは以下の通り。

「新明解」において、

1)サ行変格活用型の動詞として見出し語となっているもの（例：冠する）は、そのまま１単語として扱う。

2)他の品詞の単語として見出しに記載されている単語のうち、文字列「する」をともないサ行変格活用型の動詞となるもの（例：勉強）は、サ行変格活用型の動詞として登録しない。

③複合動詞は２語に分割することを原則とする（例：話し／合う）。但し、「新明解」に予め登録されているものは特に分割しない。

④下記の接尾語はその直前の語と切りはなさない。

的（例：連続的）、化（例：国有化）、性（例：中立性）、者（例：生存者）、力（例：経済力）、感（例：違和感）、論（例：肯定論）、庁（例：経企庁）、主義（例：自由主義）、大学（例：東京大学）、島（例：グアム島）、半島（例：伊豆半島）、諸島（例：カナリア諸島）、湖（例：十和田湖）、川（例：利根川）、山（例：富士山）、さ（例：ゆううつさ）、み（例：面白み）など約33個

⑤下記の接尾語・省略語は１つの独立した単語として扱う。

都（例：東京／都）、府（例：大阪／府）、県（例：広島／県）、区（例：文京／区）、市（例：武蔵野／市）、町（例：山手／町）、村（例：大滝／村）、郡（例：塩谷／郡）、氏（例：山田／氏）、年（例：1980／年）、月（例：五／月）、日（例：26／日）、率（例：普及／率）、法（例：日銀／法）、高・高校・高等学校（例：帝京／高）など大多数の接尾語

⑥下記の接頭語はその直前の語と切りはなさない。

お（例：おみやげ）、ご（例：ご機嫌）の計２個

⑦その他の接頭語は１つの独立した単語として扱う。

同（例：同／会長）、本（例：本／審査）、元（例：元／議員）、前（例：前／首相）、新（例：新／協定）、再（例：再／選挙）、両（例：両／国）、大（例：大／回転）、超（例：超／大型）、各（例：各／機関）など

⑧姓名は姓と名とに分ける（例：山田／太郎）。

⑨上記③～⑧に抵触しない限り、固有名詞はひと纏まりとして扱う（例：北大西洋条約機構）。

2-3.作業の流れ

作業の流れは、図１に示す通りである。＊印を付したブロックは人的処理を、＋印を付したブロックは計算機処理をそれぞれ表わしている。

これらの処理の詳細を、処理①～④は３～６で、処理⑤、⑥は７で、処理⑦は８でそれぞれ述べる。

３．調査対象の選定

昭和57年の朝日新聞の中から、各月内で曜日に重複のないようにして７日分ずつを１年分合計朝夕刊84日分を選定した。その際、新聞の縮刷版を参考にして、表や写真が多いために文章の部分が特に少なくなっている日のもの

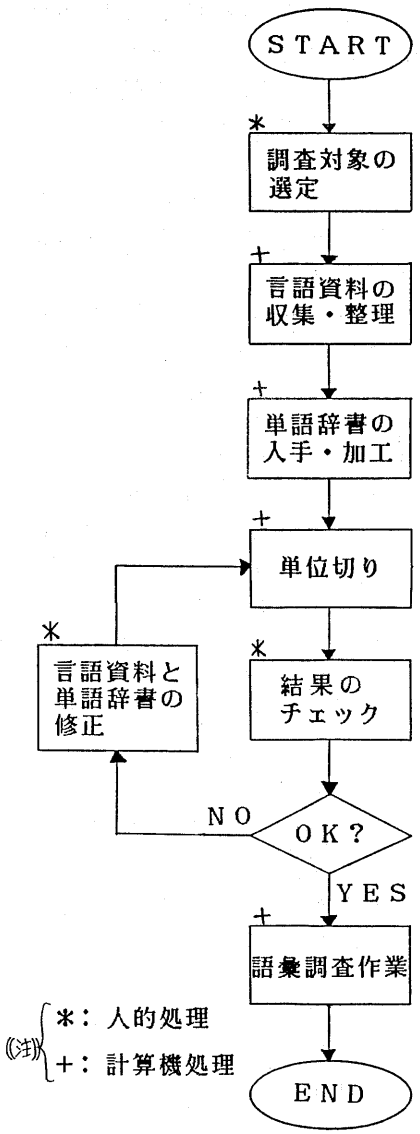

図１．作業の流れ

は避けた。

４．言語資料の収集と整理

調査資料の新聞記事データは、以下の手順で収集・整理した。

①新聞記事データが含まれている原データを新聞社から磁気テープの形で供与を受ける

②計算機を利用して、原データの中から調査に必要な新聞記事データを抽出・作成し、これをプリントアウトする

③プリントアウトした新聞記事データと縮刷版とを見比べ、抽出・作成ミスをチェックする

④このチェックに基づき新聞記事データを修正・整理する

新聞社から入手した原データは、新聞紙面作成・編集のための、絵や写真等のグラフィック情報と記事文等の文字情報とが混在した形となっている。従って、この原データの中から新聞記事データのみを抽出する必要がある。処理②はこのために行った。抽出した記事の種類は、第１面、第２面、第３面、外報面、経済面、スポーツ面、社会面等の記事と家庭欄、科学欄等の記事及び小説であり、商況欄、ラジオテレビ欄、地価欄、囲碁・将棋欄等のものや、グラフィック情報として格納されている広告や大見出し等は抽出しなかった。その後、縮刷版に掲載されていない不要な文字データを処理③により検出し、処理④により新聞記事データの修正・整理を行った。

５．単語辞書の作成

この研究で最終的に利用した単位切り用単語辞書（以下「単語辞書」と略称）は、計算機可読な「新明解」・「九大辞書」と地名・人名などの語彙資料を主素材として、以下の手順で作成した（なお下記の処理②～⑤と⑦は「新明解」、処理⑥は「九大辞書」、処理⑧はその他の語彙資料を用いての処理である）。

①「新明解」と「九大辞書」を磁気テープの形で入手

②副見出しの主見出し化

③動詞・形容詞・形容動詞・接続詞・副詞・連体詞・代名詞の読み（ひらがな表記）の主見出し化

④辞書の語義説明文中に埋もれている単語の抽出とその主見出し化

⑤以上の処理①～④により得られた単語のうち、用言は語尾を変化させ、その活用形もそれぞれ１つの単語として登録する。
⑥固有名詞（人名・地名等）・省略語等の追加登録
⑦諺と慣用句の削除
⑧単位切り結果をもとに、「新明解」、「九大辞書」、地名・人名等の語彙資料に含まれていない単語の追加・削除を行う（詳細は後述の6-1 を参照のこと）

「新明解」の見出しは、主見出しと副見出しの２種からなっており、例えば「アーク灯」は主見出し「アーク」の副見出しとして記載されている。これは辞書編纂上の理由による区別であり、本研究ではこの区別は不要である。従って、これに対処するために処理②を行った。

また同一の単語でありながら、「新明解」と新聞とで表記を異にするものがある。これらのうち、「新明解」の語義説明文中に「... とも書く」という形で載っているものは処理④により対処し、そうでないものは処理③、⑧により対処した。

以上の処理をしても、なお登録もれの単語が存在する。その代表的なものが固有名詞と省略語である。これは処理⑥で対処した。なお、処理⑤は単位切りプログラムを複雑化させないことを意図して行った。処理⑦は単語単位統一のために行った。

６．単位切りの方法とシステムの実現

6-1.単位切りの方法

単位切りの方法としては従来から種々のものが提案されているが、本研究では左端固定型の最長一致法を基本として、これに下記の改良を加えたものを用いた。
①単位切り処理中に単語が抽出されない場合、その直前の単位切り処理に誤りがあるものとしてバックトラッキングを起こし、直前部分の再処理を行う。但し、このバックトラッキングは１回だけである。
②漢数字列・アラビア数字列・アルファベット文字列・かたかな文字列は、それが単語辞書に無くても１つの単語とみなす。ここに漢数字列とは、零、〇、一、壱、二、弐、三、参、四、五、六、七、八、九、十、百、千、万、億、兆と・（中点）のみから構成されている文字列のことをいう。アラビア数字列とは、０、１、２、３、４、５、６、７、８、９と・（中点）のみから構成されている文字列のことをいう。また、かたかな文字列とは、ア、イ、ウ、・・・、ンとー（長音記号）のみから構成されている文字列のことをいう。
③おどり字（例：々）は、その直前の単語の一部とみなす。
④最長一致法では正しく単位切りできない単語列のうち、使用頻度の高いものは連語として辞書に登録しておき、最長一致法で単位切りする際には連語として切り出し、その後別に連語辞書を参照して正しい単語単位に分割する。
⑤単語辞書に登録されていない文字が連続して出現した場合は、これらを纏めて１つの単語とする。
⑥単語の語長は最大20文字とする。

①のバックトラッキングは、例えば「～はきっと～」の場合有効である。この例ではまず、「～／はき／っと～」のように単位切りした後、文字列「っと」の単位切りに失敗するので、バックトラッキングをおこして「～／は／きっと～」と切り直す。

②は例えば、文字列「ガス管」の単位切りでこの文字列が辞書にない場合に有効である。すなわちこの場合、最長一致法に基づき文字列「ガス」を次の単語候補とし辞書の参照を行うが、

この文字列が辞書になくても規則②によりこれを１単語とするのである。この規則により「／ガ／ス／管」と細かく切れることを回避することができる。なお規則②は、文字列「ガス」が辞書にあるかないかによらず「ガス／管」と切るという意味ではなく、もし辞書に「ガス管」があれば最長一致法によりこれを１単位とすることを妨げない。

③は、単語辞書作成に用いた「新明解」の表記におどり字が用いられていなかったので特に設けた。

④は、最長一致法の欠点を補うために導入した。この方法では例えば「なければならない」は、まず１個の連語の形で切り出した後、「なけれ／ば／なら／ない」と分割する。これはひらがな列部分の単位切りに特に有効である。

また⑤は、最長一致法や上記①～～④によっても処理しつくせなかったものが連続して出現した場合、これをひと纏めにするという意味で設けた。

⑥は、新明解国語辞典での最長単語が、「インフェリオリティーコンプレックス」の17文字長となっているのを参考にして決めた。

6-2.単位切りシステムの実現

単位切りシステムは、東京大学大型計算機センターのM280H 上に実現した。プログラム言語は、実行速度・入出力機能・デバッグ機能の３点を考慮し、最適化FORTRAN77 を用いた。新聞記事データは、ディスク上に蓄積し利用する。単語辞書は、アクセス回数が極めて多いので、主記憶に読み込んで利用する。この際、個々の単語が主記憶領域上で連続的に配置されるようにし、読み込み時間を約30% 減少させた。更に、単語辞書へのアクセスは処理時間削減のためハッシング法を用いている。ハッシング関数は図２の線形ハッシング法のものを用いた[4]。この方法では、関数HASH（図２参照）で求めたアドレスで衝突が生じた場合、衝突の生じないアドレスが見つかるまで関数NEXTAD（図２参照）を用いて次のアドレスを求める。なおハッシュテーブルのサイズは登録単語総数の約４倍に相当する素数値とした。

```
**************** FUNCTION DIVISION ****************
      FUNCTION HASH(ITEM,WL,TSIZE)
      INTEGER*4 TSIZE,QUOT,REMAIN,WL,HASH,PARA
      INTEGER*2 ITEM
      DIMENSION ITEM(WL)
      COMMON QUOT,REMAIN
      PARA=1
      DO 10 I=1,WL
         PARA=PARA*ITEM(I)
 10   CONTINUE
      PARA=ABS(PARA)
      HASH=MOD(PARA,TSIZE)
      REMAIN=HASH
      QUOT=PARA/TSIZE
      IF(MOD(QUOT,TSIZE).EQ.0) THEN
        QUOT=1
      END IF
      RETURN
      END
**********************************************
      FUNCTION NEXTAD(ADDR,TSIZE)
      INTEGER*4 QUOT,ADDR,TEMP,NEXTAD,REMAIN,TSIZE
      COMMON QUOT,REMAIN
      TEMP=MOD(ADDR + QUOT,TSIZE)
      IF(TEMP.EQ.REMAIN) THEN
        NEXTAD=-1
      ELSE
        NEXTAD=TEMP
      END IF
      RETURN
      END
```

図２．ハッシング関数のプログラム

以上のシステム構成を採用した結果、一日分（朝夕刊）の新聞記事データに対する単位切り処理時間は、CPU 時間で約70～80秒程までに短縮化された。今後更に、最も実行回数の多いハッシング関数部分をベクトル化するとともに、上記のハッシング関数では区別できない「端末」と「末端」のような単語をも区別するように改良すれば、なお高速化することが期待される。

７．単位切り作業と結果の評価

7-1.実際の単位切り作業

単位切り作業は、①単位切り、②単位切り結果のチェック、③辞書・アルゴリズム等の修正、④再単位切り、の各作業の繰り返しからなる。この繰り返しは、処理精度が充分良くなった時点で終了させる。処理③では、主として単位切り方略の改良と辞書への単語の追加・削除を行った。図３に作業の

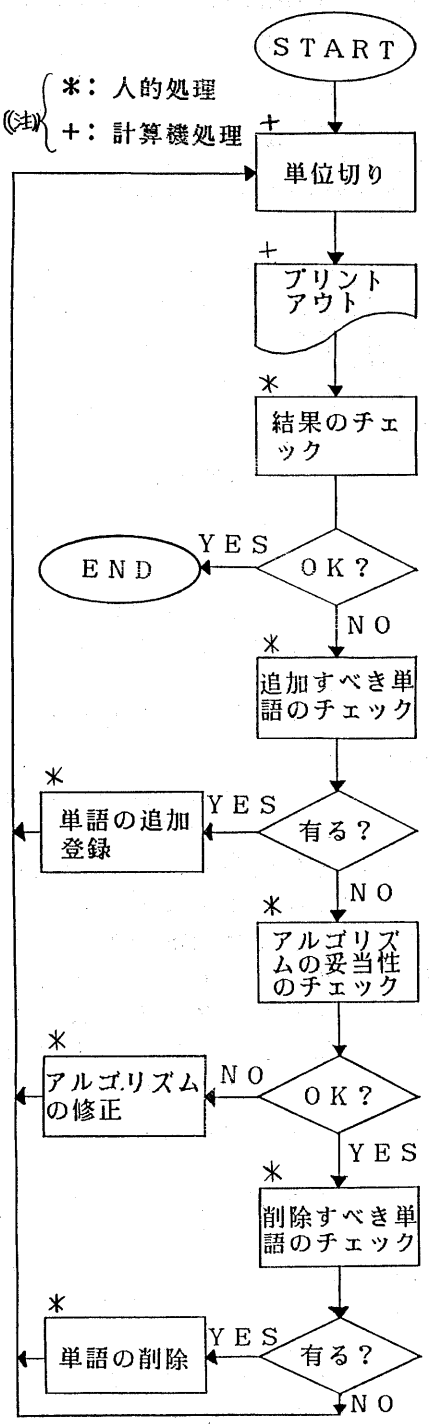

図３．単位切り作業の流れ

流れの詳細を示す。

単語辞書に追加した単語は、固有名詞（例：東京大学）・略記名称（例：通産省）・混ぜ書きや異字等の辞書表記と異なる表記語（例：「近付く」に対する「近づく」、「反撥」に対する「反発」）等の約１万単語である。

削除した単語は、諺（例：旅は道づれ世は情）や慣用句（例：泣いても笑っても）等もともと単語ではないものと、誤った処理結果の誘発原因となる語（例：古語「がな」や方言「のう」等）等約４百単語である。これらのリストの一部を図４に示す。

お手洗い	音楽的	向ケ丘遊園
そりゃあ	海援隊	佐田の山
みそ汁	寒げいこ	子グマ
スモール化	関口	大リーグ
安政	義博	大和なでしこ
磯田	金田一	東叡山寛永寺
永茂	慶応義塾	特ダネ

1）追加した単語の例

あるか	たわけ	になっ
いたも	てくる	によ
がい	であるこ	のこ
がな	となる	のころ
たとい	とみ	のし

2）削除した単語の例

図４．追加・削除した単語の例

また、新聞記事データの誤り（例：ひらがなの「へ」がかたかなの「へ」となっている。もしくはこの逆）等は、見つかり次第対処した。実際の単位切り結果を図５に示す。

なお図５の結果は、利用目的によっては必ずしも十分なものではないが、本研究の目的：①自動単位切りアルゴリズムの研究、②日本語の基礎的資料の作成、を鑑みるならば許容範囲内であると言えよう。

7-2.単位切りの問題点とその対処法

前述の方法による単位切りでは、単語の文法的接続関係、構文規則、単語

政府／は／春闘／終盤／の／焦点／と／なっ／て／いる／三公社五現業／の／賃上げ／問題／に／ついて／、／十三／日／夕／に／給与／関係／閣僚／会議／を／開き／、／公企体／当局／が／有額回答／を／出す／こと／を／認める／方針／を／決めた／。／これ／は／宮沢／官房長官／が／十／日／午前／の／記者／会見／で／明らかに／し／た／もの／。／

1）第一面の記事

音学／の／時間／、／クラッシック／に／は／なじま／ない／。／生徒／の／気／を／引く／ため／の／「／投げ入れ／教材／」／が／「／若者／たち／」／「／ばら／が／咲い／た／」／海援隊／の／「／贈る／言葉／」／など／。／校内／で／わから／ない／裏側／の／交わり／を／「／地下／組織／」／と／いう／。／テレビ／を／見／ながら／マンガ／を／読む／子／が／多い／。／「／ながら／読み／」／　／全国／の／中学／、／高校／の／卒業式／で／警官／が／校内／外／を／警戒／し／た／「／警官／ガード／卒業式／」／は／なんと／十三／％／の／約／二千二百／校／。／

2）特集記事

「／赤坂／の／津田／」／　／丸山／は／福井／に／命じ／た／。／　／「／かしこまり／まし／た／」／　／「／津田／」／は／赤坂／の／一流／料亭／で／ある／。／福井／は／新宿／から／甲州／街道／を／横切り／、／代々木／の／参宮橋／へ／行く／道路／に／入り／、／それから／神宮／域内／の／森林／の／西側／に／沿っ／て／千駄谷／に／出る／道／を／えらん／だ／。

3）小説

図５．単位切りの実例

の意味情報等は利用していないという欠点があるため、場合によっては誤処理が生ずる。以下では、この方法では回避し得なかった問題点に対し、どんな解決方法があるかを簡単に論じる。

（例１）正：／可決／さ／れ／た／
　　誤：／可決／され／た／

（例１）は、本研究で「名詞＋する」というサ行変格活用型の複合動詞を認めなかったことが主因であるが、文字列「され」の直前の単語が名詞の場合「さ／れ」と分離することにすれば対処可能である。同様の例としては、

（例２）正：／成功／さ／せ／た／
　　誤：／成功／させ／た

がある。

（例３）正：／起きる／と／がっかり／する／
　　誤：／起きる／とがっ／かり／する／

この例では、品詞情報に着目することにより対処することが可能である。まず誤処理例を品詞列に置き換えると①動詞１＋動詞２＋動詞３＋動詞４または、動詞１＋動詞２＋名詞＋動詞４となる。動詞３が音便の形となっていることより、その直後に動詞や、名詞は出現しない筈である。従って、「とがっ」は誤った単語単位といえる。次に「とが」を単語候補とすると、これは動詞「とぐ」の未然形と照合する。しかしながら、この場合次の単位切り処理に失敗するので、正しい単位切りではない可能性がある。従ってここでバックトラッキングをおこし「と」を単語として切り出す。これにより以後単語「がっかり」と「する」が切り出され正しい処理がなされる。

（例３）正：／汚れ／が／とれ／なく／なる／
　　誤：／汚れ／が／とれ／なくなる／

例３は最も処理が難しく、文法的接続関係や構文規則の両方を満足しているにもかかわらず誤処理となっている。従ってこれの場合、意味処理を必要とする。同様の例としては、次のものがある。

（例４）正：／それ／は／赤い／の／で／よい／
誤：／それ／は／赤い／ので／よい／

7-3.単位切りの評価

単位切りを評価する場合、①無意味な文字列を誤って抽出していないか、②単語間の文法的な接続関係が正しいか、③意味的にも正しく区切られているか、という３つのレベルが考えられる。本研究では、③のレベルでの評価を行った。

さて単位切りとは、所与の文字列中に単語境界を見い出しその位置に区切り記号を付与することである。従って、誤りの原因は、①単語境界のシフト、②単語境界の欠落、の２つに分類される。これらの例を以下に示す。

1)予断／しに／くい
2)食べる／とがっ／かり／する
3)しから／れ／たそ／う／だ
①の例

4)行く／ことに／なっ／て／いる
5)当て／に／でき／なくなっ／て
6)正しい／とはいえ／ない
②の例

((注))下線は誤処理部分を示している

図６．誤処理の例

以上のことをもとに処理精度を((誤り原因の総個数)/(区切りを付与すべき箇所の個数))･100で定義した場合、本報の処理結果は約99%である。

８．語彙調査結果

語彙調査として、文字頻度分布、漢字KLIC(Key Letter In Context)、単語頻度順位一覧表、KWIC(Key Word In Context)等を作成した。なお、KWICは約670 万行、KLICは約430 万行であり磁気テープに格納されている。

９．おわりに

本報では日本語の現状を語彙の側面から分析し、日本語に関する基礎的資料を作成するために行った新聞語彙調査に関し、調査の概要、自動単位切りの方法・結果及び調査結果を報告した。

謝辞

本研究は、昭和59年度科学研究費特定研究“言語の標準化”総括班新聞委員会の事業の一環として行ったものであり、御助言を戴いた同委員会の各位、データを提供された朝日新聞社、計算機可読な辞書を提供された三省堂及び九州大学吉田教授、データ整理を援助された姫路短期大学田中助教授と東横学園女子短期大学倍賞助教授をはじめとする方々、単位切りに関して助言を下さった埼玉大学荻野講師及び作業に協力して下さった東京大学藤崎・広瀬研究室の諸氏に深く感謝する。

<<参考文献>>

1)国立国語研究所：“電子計算機による新聞の語彙調査（Ⅳ）”，国立国語研究所報告書48，1973.
2)藤崎・亀田・荻野：“新聞記事文の分かち書き処理とそれに基づく語彙調査”，情報処理学会第30回全国大会講演論文集，1985.
3)亀田・藤崎：“大量の新聞記事データを対象とした語彙調査”,情報処理学会第31回全国大会講演論文集，1985.
4)Harold S. Stone：“Introduction to Computer Organization and Data Structures ：PDP-11 Edition,” McGraw-Hill,1975.