Word count by use of computer and the lemmatization processing

電子計算機による用語調査と同語異語の処理

石　綿　敏　雄

1. 電子計算機による用語調査と同語異語

ワード・カウントの単位についての二つの重要な事項として，単位の長さということと，単位の幅ということがある。後者は筆者がつくったことばであって，まだあまり広くは使用されていない。単位の幅というのは，長さに対して用いたことばであって，その内容は同語異語といわれるものに等しい。すなわちカウントの基礎として，どの範囲のものをまとめてカウントするかというときの，わくのようなものである。このわくがきまっていなければ，どのようにカウントしてよいかわからず，このわくがすべて異なったものと考えれば，異なり語と延べ語の数が一致し，すべての語は度数1になってしまう。それゆえ，ワード・カウントの単位としての最も重要な基礎の一つである。

いま英語の例をあげると

「しごと」	「はたらく」
my works	I work
his work	he works

というばあい，(my) works の works と (he) works の works は語形が同じであるが，その意味用語は全く別である。(his) work の work と (I) work の work も，語形は全く同じあるが，意味用法は別である。そこでこれは，語の内容 (Inhalt) 本位には

(my) works と (his) work を

(I) work と (he) works を

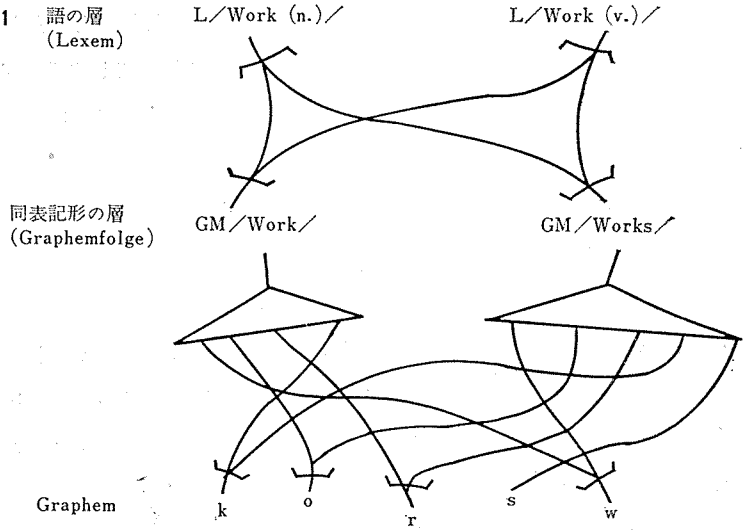
まとめてカウントすべきである。言語表現の側 (Ausdrucksseite) に現われた形をもって，単純にカウントするのは問題である。しかしこれを形の上から

work と works に分けてカウントすることがあるのである。

問題である，といったのは，実はよくない，というべきであるが，また別にそのレベルでのカウントが意味をもつことがあるからである。たとえば言語情報処理など，機械で処理することを考えるばあいには，これが重要な意味をもつ。もちろんそこからさらに Inhaltsseite へとさかのぼる作業をするのではあるが，その段階での多義性（Mehrdeutigkeit）がどのくらいあるかをさぐるばあいの，第一の作業単位は，まずこれであるといってよいからである。これは，あくまで言語情報処理の立ち場であって，いわゆるワード・カウントのばあいには，このままではよくないことは目に見えている。つまり用語調査とすると，やはり語のレベルで整理すべきなのである。

ここでレベルという語を用いたが，これはどちらかといえば生成文法の用語であって，シドニー・M・ラムの *stratificational grammar*「成層文法」の用語でいえば *stratum*「層」である。ラムの *representation* に十分忠実ではないがこの関係を図示すると，たとえば

図1



のようになるであろう。この三つの *statum* を混同することは許されない。そしていわゆる *semantic count* は語の層の上位にある意味素の層でのカウントである。「分類語彙表」はこの *sememic stratum* のテーブルであるというこ

とができよう。ワード・カウントは *lexemic stratum* であると考えられる。だから *morphemic stratum* から *lexemic stratum* へと移すことを考えなくてはならない。

この作業は，機械処理のばあいなかなかやっかいである。機械でこの形態素（それも考え方によってはこうもいえないのであるが）までは比較的らくに処理できる，つまり文字列のカウントまではらくにできるのであるが，その先へは一歩 Inhalt の側にはいる（つまり意味用法に多少ともふれなければならない）ので，むずかしくなるのである。

われわれが行なった「婦人雑誌の用語」「総合雑誌の用語」「現代雑誌九十種の用語用字」では，全体が人間作業であったため，そのような操作が随時行なわれた。しかし機械を使用すると，中間で手を入れることがむずかしいので，前編集か後編集かにまかせなければならなくなってしまう。

アメリカのブラウン大学で行なった現代アメリカ英語の調査（*Corpus* 1014 232 *words*……1972年夏国語研を訪問された J. B. キャロル氏より今は "five million words" の調査が，すんでいることを直接聞いた）では，この問題をまったく処理せずに通りすぎてしまった。つまり文字列調査である。その報告書にも

> *Homographs* (*word identical in spelling but different in pronounciation and meaning*) *and homologs* (*word indentical in spelling and pronounciation but different in meaning*) *are lumped together as the same type. Thus* sow, *'plant seed' and* sow, *'female pig' are not distignuished nor* bear, *'carry' and* bear, *'animal'.*

と書いてある。それればかりか，

> *Variant spellings of phonologicalley and lexically identical words are listed and counted separately. Thus* catalog *and* catalogue *are separated as are* non-conformist *and* nonconformist.

であり，

> *Morphological and syntactically variant gaphic forms of lexically identical words are listed seperately. Thus* cannot, can't *and* c'n *appear*

as separate types, while can not *will simply be counted as one instance each of can and not. In fact it will not be possible to derive from these talbes an accurate count of auxiliary* can, *even if the morphological variants are counted, since all tokens of the noun* can *are lumped with verbal* can.

といっている。したがって全体としてはずいぶん制約の多いものになる。たとえば文体の分析などにとって大変つごうの悪いものである。Lorge がやったように *semantic* な *count* はとてもできない。それは *computer technology* の限界の外にある。この本を利用する人々に，このような *homograph* その他の問題に注意してほしいということを *advice* したいといっている。すなわち

These consequences of basing the list uncompromisingly on the graphic word as unit undeniably restrict the usefulness of the count, especially for stylisticae analysis. But it is hard to see how anyother procedure short of a completely "semantic" count like that of Lorge is possible. In the present state of the art a semantic count, even if desirable, is beyondthe reach of computer technology. We can only advise the user of the word lists in this book to be aware of how homographs, variant spellings, andmophological variation may influence his conclusions.

このような結論について二つのことが感じられる。

その一つは，電子計算機による用語調査の一つの考え方として，この homographs などの未整理な段階で一応の語彙表とし，あとはそれを用心しながら使うというものである。計算機の現段階とするとこれが一つの現実的な考え方でもある。そのような結果について問題を含みながらも，計算機の早い処理能力を生かして，むしろ大量に処理する，という方に力を注ぐことも，一つの有用な考え方とすべきであろう。

もう一つ，感じられることは，上記の英文が，われわれが作成した報告書に書かれたことと，全く軌を一にするものであることである。筆者は国語研報告37「電子計算機による新聞の語彙調査」で次のように書いた。

今回の語彙調査は，電子計算機を用いて行なう第一回のものであるので，

技術的に必ずしもすべての見通しがつけられていたわけではない。解決できなかった問題として，大きなものは漢字の読み（「通った」のカヨッタ，トオッタなど）も含めて，同語異語の判別，異形同語の処理がなされていないのである（3ページ）

また

この調査ではいわゆる同語異語の判別を行なっていない。得られた単位は同表記同形の語について度数をカウントした表である。表記形が同じである「いき」（「粋」の意）と「いき」（「行き」の意）とは区別されず同じ語としてカウントされ，表記形が異なる「いき」（「行き」の意）と「行き」とは別な語として別にカウントされ，整理された度数表である。すなわち異なった語でも表記が同じであれば区別されず，同じ語でも表記や語形が異なれば別の語として処理されている。

と書いている。このように見てくると細部にあっては日本語と英語の，この種の問題に相違が見られても，基本的な部分ではかなり一致するところが多い問題をかかえていると見るべきであろう。特にコンピュータで取り扱う，という点からみて，そうである。これをどのように解決することができるか。とにかくわれわれの処理法（新聞語彙調査）も，クチェラ氏らの英語の処理法も同じレベルでとどまっているのである。

用語調査や用語総索引作成のためにコンピュータを用いることは，ヨーロッパでも広く行なわれているが，この問題もやはりそこに登場しているようである。たとえば ILT の，Monica RöSSING-HAGER, Wortindex zu George Büchner : Dichtungen und Übersetzungen, の書評に次のような文がある。（全体としては alphabetischer Wortformenindex なのであるが，そのうしろにひん度表がついていて）

Häutigkeitsliste

Hier sind die Wortformen in der Rangfolge ihrer Häufigkeit geordnet, beginnend mit dem höchsten Vorkommen.……

1. Hier noch mehr als beim Index fällt die *Nicht-Homogenität* des indizierten Textes auf. Es handelt sich zwar um Texte von Büchner,

aber was besagte es, wenn man Z.B. weiss, dass die Form *frei* elfmal vorkommt, wenn wan nicht weiss, wo das Wort eigentlich vorkommt. Hier sollte die Frequenz doch besser pro Werk spezifiziert werden.

Vergleiche zwischen den unterschiedenen Werken würden auf diese Weise möglich. Auch die *Nicht-Lemmatisierung* ist hier mehr als beim Index ein grosses Handikap.

やはりこの種の語い表作成にあたって，Index のときよりも一層，ひん度のばあいに Nicht-Lemmatisierung すなわち同語異語の処理がなされていないことが欠陥であると指摘されているのである。日本・アメリカのみならずヨーロッパでも，この問題は同じように現われてくる。

1970年の夏ドイツのアーヘンのライン・ウエストファリア工業高等専門学校百年祭の催しの一つとして開催された，"Literatur und Datenverarbeitung" の会合で，さまざまな研究発表が行なわれた。このなかで，たとえば文学者のシラーの作品の総索引，哲学者カントの作品の総索引作成などの報告も出ており，その他の各種の報告や発表が行なわれたが，この会議のはじめに，ザールブリュッケン大学のハンス・エッガース教授が行なった発表は，この稿のテーマに関連して，きわめて興味深いものである。エッガース氏の実質的提案はあとまわしにして，ここでは問題提起の部分を紹介しよう。

エッガース氏はまず，このような索引作成にあたって，完全な語い表がつくられたという例がないということをいい，これまでにつくられたものはなお不十分で，要求を満足させるものでないという。

Dennoch wissen die Bearbeiter ebenso wie die Benutzer, dass die Indices noch viele Wünsche offen lassen.

コンピュータはなるほど非常に正確なしごとをし，索引作成にあたっても作成者が望んだとおりの，語形排列をやってくれる。しかし，ドイツ語の例をとると，語形屈折変化が多く同一語がばらばらになってしまう。分離動詞などの *discontinuity* があること，Homograph が多いことなどで問題が多い。（これはまさに同語異語の問題である)。

これを解決しようとすると，どうすればよいか。あるばあいにはプリエディトがなされたし，またポストエディトをしてもよいわけである。同語異語判別の操作がすんだ結果立てられる見出し語（Lemma）が与えられるためにはこの二つの方法しかない。「コンピュータをどこにどのように使うかについて，レクシコグラファーは現在のところ次の二つの方法を用いている。事前に大いにプリエディトをやっておき，Lemma が与えられている状態にしておくものと，事後に手を加えて言語学的に Lemma を書き添えてゆくものとである。」すなわち

Wo Electronenrechner zu Hilfe genommen worden, verwenden Lexikographen heute in allgemeinen eine der beiden folgenden Methoden: Entweder wird zu jeder Wortform des zu bearbeitenden Textes im Vorwege das Lemma angegeben, unter dem die Form im Ausdruck erscheinen soll, oder es wird im nachhinein jeder einzelne Beleg philologisch bearbeitet und seinem Lemma zugeordnet.

ところが，このどちらの方法も非常に時間がかかる。(Beide Wege erfordern einen sehr erheblichen Zeitaufward)。しかもその作業たるや単調（eintönig）で，同じことのくりかえし（sich ständig wiederholende Arbeit）である。プリエディトのしごとは，「モーレツ」なしごとだ（eine sehr intensive Vorbereitung der Texte)。(エガースはいっていないが，その上，誤の絶無は期しがたいものである)。

加うるに，われわれがとりかかろうとするものは，カントにしろ，ゲーテにしろ，トーマス・マンにしろ，ぼう大な作品である（und da die interessierenden Texte meist sehr umfangreich sind,)。コンピュータを用いたにしろ，全体で大きなしごとなのである。

ドイツ語で問題になる点を少しくわしく説明すると，まず，ドイツ語の（名詞の）曲用および（動詞の）活用について，同じ Lemma にはいるものが，ばらばらになってしまうことである。たとえば ESSEN（食べる）という動詞はそのままの形で文中に現われることもあるが，ASS, GEGESSEN, ISST のように活用しても現われる。これらはばらばらなところにならべられてしま

う。次にいわゆる分離動詞で，AUFESSEN（食べつくす）は不定形で，他の形は文中では ISST……AUF のように，活用することもある。最後に Homograph があるが，たとえば LIEBE という形は（コンピュータの中ではすべて大文字なので）「愛」という名詞，「親愛な（女性名詞など）」という形容詞「（私は）愛する」という動詞でもありうる。FIBEL という語は「入門書」という語と「留め金」という語とあって，別な語である。

ハンス・エッガース氏は，それについて，そのグループで開発した自動構文解析のプログラムを使用して，自動同語異語処理（automatische Lemmatisierung）を行なうことを提案しているが，これについては，3. の将来における解決法のなかで，紹介することにする。

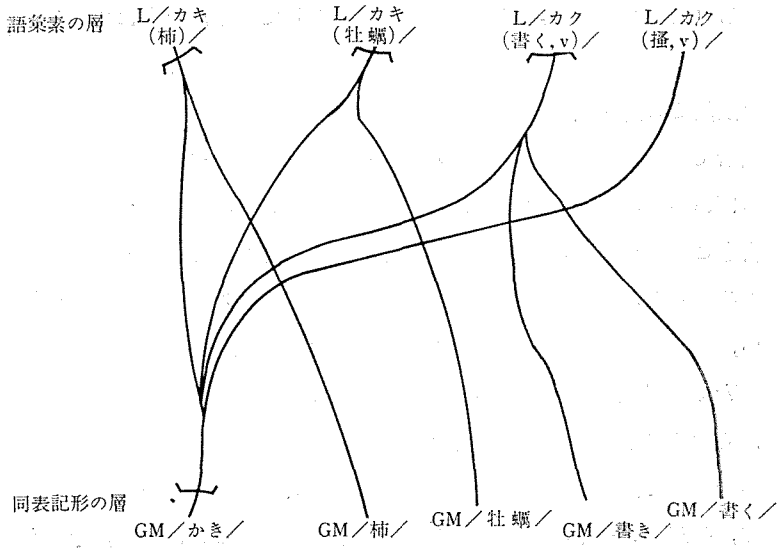
以上米日独の例について，コンピュータによる用語調査ないしは用語総索引の作成において，同語異語の処理に問題があることを指摘したわけである。用語総索引のばあいには Wortfqrmenindex というのがかなりの意味をもち，コンピュータ言語学（Computerlinguistik）のばあいには Wortform による Häufigkeitlisteをつくることも有意味であるが，語彙論のレベルでの要求はやはりそのままでは不便なこと，つまり，*lexemic stratum* でのカウントをすべきこと，しかしそのためには前編集あるいは後編集という手続きをとらねばならず，それが大変な作業になる，というようなことを述べた。

われわれの新聞用語調査その他でも，前編集や後編集が行なわれていたりする。そこで，この際，コンピュータをどのように活用することが，マン・マシン一体化のしごとを要するこのぼう大な作業のしごとについて有効であるか，という点について，二，三の提案をしてみたいと思う。この提案には大きく分けて二つの種類のものがあり，まず特別の手段を講じないで現在の段階でできること（2. で扱う）と，将来の実施を目ざして現在から準備をすすめて行くという方向のもの（3. で扱う）とである。

さて，その前に，用語調査における同語異語判別のオペレーションについて述べよう。例を日本語にもどして考えてみる。

いま「かき」「柿」「牡蠣」「書き」「書く」というような語形がコンピュータでまず取り扱いうる形（同表記語）の *stratum* で得られたとし，これを

図2



語彙調査で整理すべき *lexemic stratum* の形「柿」「牡蠣」「書く」「掻く」などと比較してみるとする。この関係を *stratificational grammar* の *representation* の方式を借りて（「借りて」というのは「忠実にのっとって」ということではない）示すとすると図2のようになる。

ここで，「柿」「牡蠣」「書く」のような漢字表記の語は大体においてそのまま *lexemic stratum* にもっていけるが，「かき」のばあいには大変である。「かき」のばあいの *upward* (*unordered*) *or* はすべてのLの段階のものと関係づけられている。一般にかな書きの語にはこの危険が多いといえよう。しかし漢字で表記されていても「方」のホウ，カタや「間」のアイダ，ケン，マ，カン，「風」のフウ，カゼ，「上手」のウワテ，カミテ，ジョウズなどいくらでもある。

さてこのように同表記語の層からしわけられてきたものを，語彙素の層でまとめるというオペレーションも必要である。「かき」と書かれたもののなかで「柿」を意味しているものと，「柿」と書かれているものをまとめるということである。（もちろん内訳がわかっていた方が，あとで使いやすい表ができよう）。同語異語の処理は，同形異語の判別と異形同語の集合という二つの操作

にあるといえるが，成層文法の表記を借りれば，*upward or* と *downward or* に当たる（ただし一般の調査のばあい，方向としては下から上にである）。

2. 現段階での解決案

現段階では，まず前編集か後編集であることは先に述べたとおりである。前編集のばあいは，やはりコードブックを作って（いわば辞書をつくって），それを書きこむようになるだろう。「かく」を「書く」と「掻く」とに分けるとするとそうせざるを得ない。「分類語彙表」を利用することができるが，このばあい，もともと *sememic stratum* に属するはずであるから，いきなり *semantic count* になっていわゆる *word count* を通りこしてしまう（したがって，そこからえられるものは基本語彙ではなくて基本意味である）。もちろんそのつもりで使えばそれでよいのであるが。このようなコードには，辞書をきめて，それをコードブックに使うという手もある。このばあい，きめないと作業に不都合がおこることは，筆者が国語研報告25「現代雑誌九十種の用語用字」第三分冊で述べたとおりである。すなわち，「あつい」（「暑い」「篤い」「暑い」「熱い」），「つぐ」（「次ぐ」「継ぐ」「注ぐ」）などを，「三省堂国語辞典」「辞海」「岩波国語辞典」について調べてみると，すべて一致しないのである。（もっとも外国でも事情は同じで，たとえばアンリ・ミッテランは

> La distribution des formes homographes à' l'intérieur d'articles uniques ou multiples du dictionnaire ne se fait pas sans quelque arbitraire.

といっている）。

前編集のばあい，あとでさまざまな加工をし，長く使うばあいには有効であろう。目的によってその程度をどのようにきめるかに問題がある。

後編集のばあいには，すべての語についてでなく，利用者が見当をつけて，必要なもののみを判別してゆけばよい，という点に利点がある。このばあいできれば総索引形式のを利用するのでなく，文脈つきの KWIC の形式でアウトプットされれば，作業がずっとらくになる。日本語のばあい，漢字プリンタを使うと便利であるが，漢字の読みを排列の上でどう生かすかに一工夫が要る。

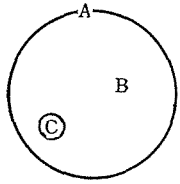
かえって漢字コードつきのカナ KWIC の方が使いやすいことがあるかもしれない。用語総索引のばあい，Worformeindex でよいということがあるし，文脈つきであれば，それで十分かもしれないのである。文脈によるソートが可能なばあい，それで，ずいぶん便利なことができる。(文献15)。

前節で述べたように，コンピュータを使用して大量の調査を行なうばあいは Wortform によるカウントもある程度意味をもち，ばあいによっては上位数千語にのみ同形語の内訳を書く，つまり同形異語の弁別のみを行なう，というのでも利用度は高くなろう。またばあいによっては一部分について詳細に同語異語の弁別を行ない，全体はそこから類推できる形にするという方法も考えられる。(図3)

前編集のばあい，すべての語につけるのでなく，よく使われる語はリストしてオミットし，それについてあとからコンピュータで書きこむという方法も考えられる。ジップの法則によって，ひん度の高い語についてこのような処理法が有効であることが考えられる。アルバイタ (Bearbeiter) は簡単なリストならおぼえられるものである。

以上の程度のことは，だれでも考えつくことで目新しいことではないだろう。いくつかの考え方を列挙してみたにすぎない。ここで一貫したシステムとして，筆者はカード・システムを提案してみたいのである。これは前編集でも後編集でもなく，いわば中編集である。

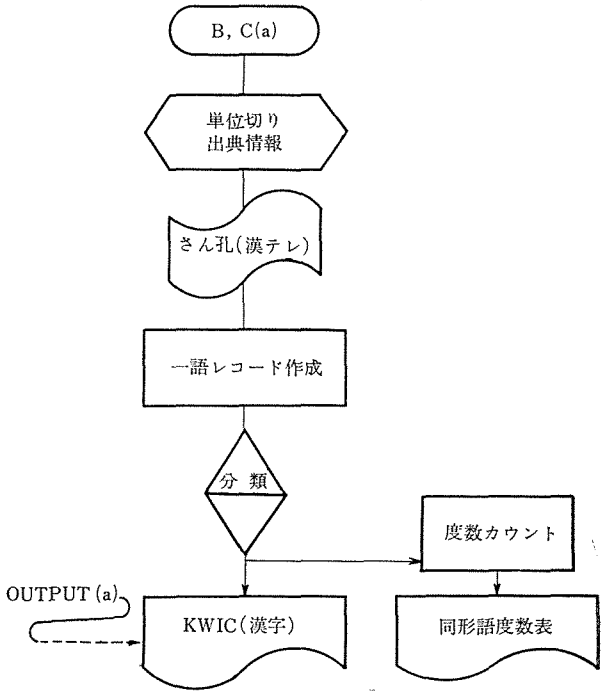
はじめのインプットの際は紙テープの方が，カードより便利なことが多いので，紙テープを使うのがよかろう。(日本語のばあい，少なくとも単位切りはしておく必要がある。漢字に読みを入れておく（書き添えておく）となお便利である。) この入力を計算機で処理して一種の KWIC をつくり（文脈つきレコードの作成と分類)，これを出力するのである。出力に際しては，ラインプリンタで打ち出すといわゆる KWIC になってしまい，そこからあとは後編集になってしまう。それでカードに一語一語文脈つきでアウトプットする。このカードは OCR 文字があるばあいには OCR カードでもよく，また IBM カードでもよい。IBM カードのばあいは事後に印字する。このカードについて，同語異語の並べかえを行なう。つまり同語のグループ化を人間の手で行なうの

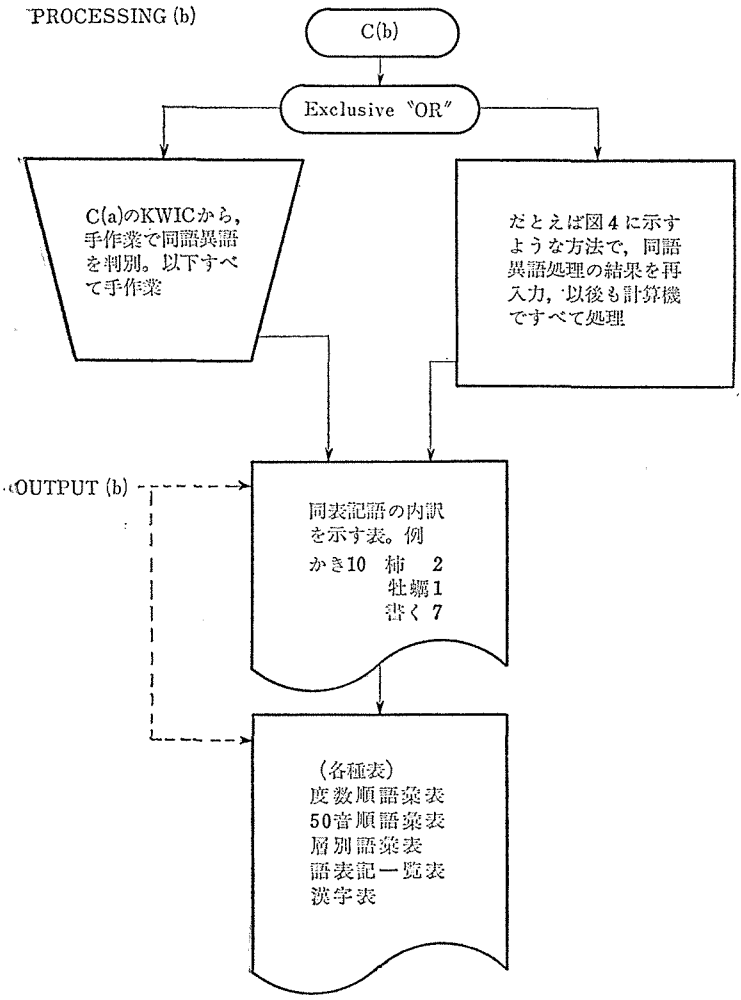


A＝B＋C（標本全体）
B－同形語調査する部分
C－同形語調査(C(a)) と
　　同語異語判別調査(C(b)) をする部分
B－単位切り，出典情報(新聞の層別など)
C－同上および漢字に読みがな

INPUT

PROCESSING (a)




PROCESSING (b)
C(b)
Exclusive "OR"
C(a)のKWICから，手作業で同語異語を判別。以下すべて手作業
だとえば図4に示すような方法で，同語異語処理の結果を再入力，以後も計算機ですべて処理
OUTPUT (b)
同表記語の内訳を示す表。例
かき10 柿 2
牡蠣1
書く 7
(各種表)
度数順語彙表
50音順語彙表
層別語彙表
語表記一覧表
漢字表

である。できれば見出しカードを用意して（初回は人間の手でつくらなければならないが，二度目からは機械でつくるとよい），これをはさみこんでいく。人間の手は，同じ語形の語について，カードに印字された文脈を見ながら分類，グループわけをして，見出しを立てるだけでよいのである。この文脈を見ながらグループ分けをするという作業自体がきわめてむずかしい作業であるので，その部分だけを人間が行なう。あとは，そのカードを，そのままインプットする。一種のターン・アラウンドである。OCR カードでも IBM カードでも，この条件をみたしてくれる。コンピュータのなかにはいったものはこのようにして人間の手で Lemmatisierung されているから，あとはすきなように処理すればよい。さてこの中間の再グループ化作業であるが，あらかじめ文脈でソートされていると同じ性質のものが並んでいることが多いので，（語の意味用法は文脈に具現化されていることが多いのでそうなるのは当然である）作業がかなりらくであり，早くできる。同形語の判別と異形同語の集合は，見出し語を立てるということとカードをまとめて動かすということだけですんでしまう。

このことは文脈ソートのできる KWIC を筆者がもって使っていることから，大体誤りなく想像できるのである。このようにすれば，漢テレ入力のばあい，漢字の読みと単位切り程度の前編集で，かなりらくに Lemmatisierung ができると考えられる。途中整理のためだけにカードを作るのは，特に大量のカードを作るのはよくない，と考えられるかもしれないが，このカードは用語用字の分析のためにあとあとまで使えるのである。決してむだにはならない。しかも，すきなように並べかえ，情報をつけて，なん度でも再入力できる利点がある。

以上のことをフローチャートにかいておく（図4）。これは一応漢テレを使い，漢字の読みを入れ IBM カードを使用したケースを考えてみた。こまかい *modification* も考えられるが，わかりにくくすることをおそれて，大まかな考え方だけを書く。

ただ，このような大量のアウトプットが理論的には可能であるが，実際問題として大変だということはあるかもしれない。カードの保管管理が，データ量

が大きいだけに，大変だろうという心配がある（もっとも他の方法を取っても，何かの意味でそれはある）。したがってあまりぼう大な作業にはむいていないかもしれない。まあ百万語前後の Corpus について精確な調査をしたいときに有効であろう。

このシステムの利点は，中間でアウトプットしたカードそのものが，比較的少量の人間の手で行なわれる処理のと，まったくそのまま再入力の媒体となることである。もし必要であれば修正もでき，しかも入力時よりも検査しやすい（同じ種類のデータがソートされて並んでいるので）ので，検査と修正が，人間の作業の観点からみて有効に行なわれるという利点もある。これによって次の語彙表作成に進むこともでき，さらにさかのぼって原文に情報を加えるということも，プログラムで可能になる。

初めの入力がかなやローマ字であるばあいは，処理が簡単で，カードも見やすいものができるが，漢字であると，できれば漢字の読みがほしい。単位切りがしてあれば，漢字解読ルーチンを使うこともできよう。これは経験上大体エラーは1パーセント以下だから，百枚に1枚のカードについて読み誤りを直すということが必要になってくる。どちらがよいか，設計者の考えにより，微妙なところである。これをたとえば IBM カードに出すときは，たとえば

ゲン（＃C）キン（9A）

のようにして「現金」を現わすことができる。このコードは国語研究所のコードであって，エンコード・デコードのハンドブックがすでに用意されている。いちいちデコードするのはやっかいなようであるが，ソートされて出てくるので，最初の一系列について行なえばあとにつづく同文字列は処理しなくてもすむ。意外にらくである。このことも，すでに筆者は同様の KWIC を作って実験ずみなのである。

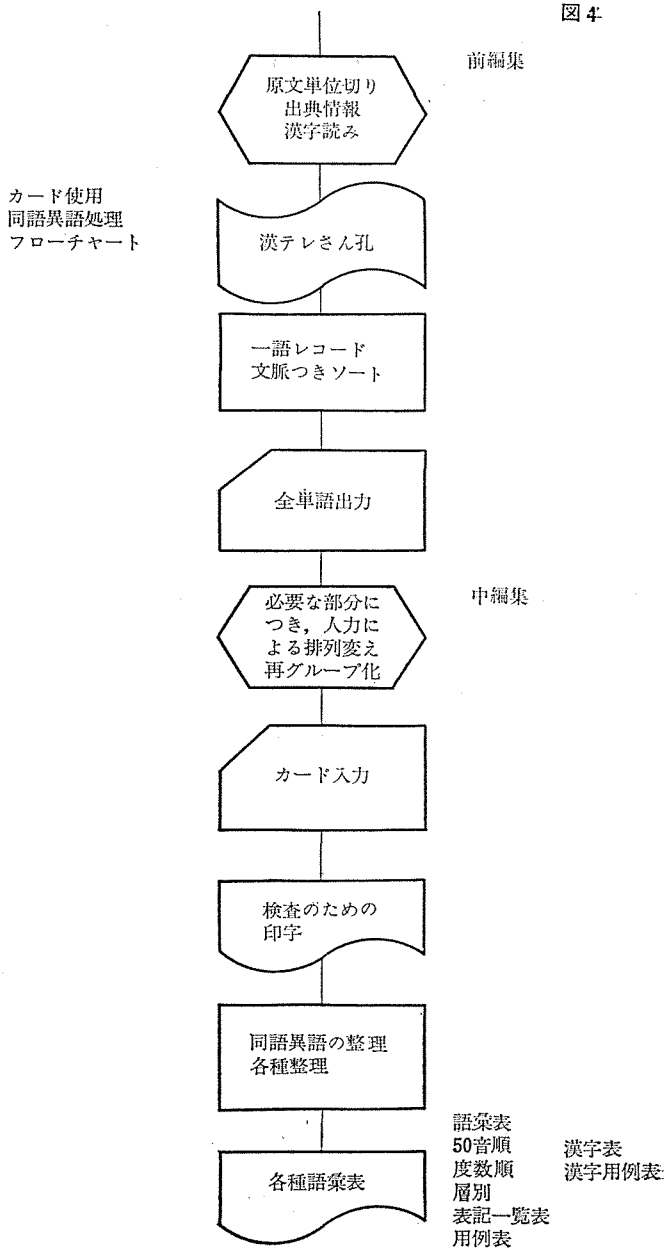
なお漢字の読みの入れははたとえば

漢〔かん〕字〔じ〕

のようにするのが，今の国語研究所では普通になっている。しかしこれにこだわらずに，他の方式を考えてもよい。

途中でカードに出すのでなく，その間の手続きがディスプレーに出されて処

図 4


前編集
原文単位切り
出典情報
漢字読み
カード使用
同語異語処理
フローチャート
漢テレさん孔
一語レコード
文脈つきソート
全単語出力
中編集
必要な部分につき，人力による排列変え
再グループ化
カード入力
検査のための
印字
同語異語の整理
各種整理
語彙表
50音順
度数順
層別
表記一覧表
用例表
各種語彙表
漢字表
漢字用例表

理するように設計することもできよう。このばあい，使える入力があることが前提である。

3. 将来の解決案

ここに述べることは，現段階ですぐにできることではないが，将来にわたって研究開発を続け，いつの日かに，実現されることを期待する，という解決策である。その方向はやはり automatische Lemmatisierung である。

そこで，さきに述べたエッガース氏が，アーヘンでどんなことを述べたか，ということになる。

まずいうべきは，エッガース氏自身，自動化が必ずしも完全にできるといっていないことである。(Ich will nicht behaupten dass die vollautomatische Herstellung von Wörterbuchern möghich ist.)。完全にできるためには，意味の問題が計算機で扱えるように，形式化されたばあいにのみ可能である。そうではなくてエッガース氏は，これまでより更に広い預域で，コンピュータがこのしごとで活用されることを期待しているのである。

エッガース氏はまず語の曲用活用の処理について，コンピュータのなかに辞書をたくわえ，テキストのなかの語と比較する。辞書は語幹をたくわえ，語尾はその語幹につけられたインフォーメーションと，それに応じたアルゴリズムによって作られたサブルーチンによって処理してゆく，というのである。この方法はプログラムによって解決ずみである。

統辞的な Homograph や同形語の処理も，実際上すでに解決されているのである。

die Erkennung mehrteiliger Formen und die Auflösung syntaktischer Homographen, sind praktisch bereits gelöst.

というのである。これは大変なことである。このためにはエッガース氏たちが，ザールブリュッケンでどんなしごとをしてきたかを紹介しなければならない。（これについては「ドイツのコンピュータ言語学」（文献15）という小文でも紹介してあるので，参照していただきたい。ここで大すじだけ述べる。なおここでの原典は主として文献3である)。

エッガース教授らのしごとのそもそもの目的は現代ドイツ語書きことばのシンタクスについての記述的な研究であった。ところで現代語の資料は日日に新たに作られているので，この大量のデータ処理のためにコンピュータを用いることにしたのである。だから研究当初の目的からすると機械による統辞分析は従属的なものであった。研究者たちはより高次の目的について片時も忘れたことがなかったのであるが，問題自体が広範囲であり，そのなかのある部分に長くかかわってくると，次第にそのことに比重がかかってくることになった。自動解析のばあい operativ—な文法の考え方が必要になるが，そのような方向も一つのアプローチだと次第に考えられてくるようになったのである。この企画は1956年にはじまり，エッガース教授を中心にすぐれた協力者たちの強力な協同作業によって推進されたのである。

グループの方向は最初から固定したものとせず，誤りなどの修正がいつでも可能なような，ヒューリスティックであることにまず努力した，生成文法もまだ十分発達していないときであったので，スタートにおいては昔ながらの taxonomisch な方法がとられていたが，漸次修正されていった。

機械による分析とすると，L・テニエールの依存モデル Abhängigkeitsgrammatik を使うこととし，そこから格の多義性をとり除いたり，syntaktisch な Homograph を見分ける必要が生じたのである。

そこで大量のデータを整理して，Homograph をタイプによって分類し，50の型を得た。この50ののグループについてこれを見分けるサブルーチンを作ったのだった。「われわれは長年この問題に没頭した (Dieses Problem hat uns jahrelang beschäftigt.)」とエッガースはいう。そうして「そのうち二，三のプログラムについてなお二，三の改良すべき点があるにしても，原則的には成功したといえよう。」というのである。問題になったのはこれだけでなく，もちろん他の構文解析上のさまざまな部分においてそれがあり，それの解明プログラムの開発とそれに伴う言語の operativ な記述が行なわれ，構文解析が大量のデータのコンピュータ実験において成功していったのである。コンピュータ言語学について手きびしい批判をした言語学者レオ・ワイスゲルバーも，ボン大学の LIMAS の業績とハンス・エッガース氏らの業績は高く評価して

いる。

ER SCHREIBT SEINE TATEN AUF. というような文を扱うとき，ここから Lemma である AUFSCHREIBEN が得られるのは，AUFが前置詞でなく動詞の前部分であると決定されたときようやくきまり，TATEN が TUNという動詞の変化形でなく，TAT という名詞の複数形であるとわからなければ TAT へまとめることはできない。このような語形の整理にあたっては，文法的な情報と統辞的な分析への道程においてのみはじめて可能となるのである。同語異語の操作をすることすなわち「ある見出し語に属する語をまとめようとするばあい」には，「自動構文解析の方法によって得られよう。そのほかの方法では得られないだろう。」という。すなわち次のごとくである。

Aber die Zuordnung alles Zugehörigen zu einem Lemma dürfte man mittels der automatischen syntakischen Analyse (aber nicht anders) erreichen können.

アーヘンの会議では，エッガース氏は automatische Lemmatisierung の解決にあたっては，まず全体を systematisch に考えるべきこと，はじめに文法理論を確立しておき，操作のアルゴリズムはあとからいくらでも修正可能なようにしておくべきこと，このプログラム運転のための基礎データには十分な Corpus すなわち大量のデータをもとにした記述がなければならないことを述べている。

一つの用語総索引を作るために，一つの用語調査を行なうためだけに，このようなルーチンをつくるのは，その犠牲があまりにも大きすぎるかもしれない。しかしいったんできあがれば，次々と用語調査，総索引作成を行なうばあい，有利に利用できる。コンピュータを作うばあいの利点はそこにある。

以上でエッガース氏らのしごとの紹介を終わるが，ここで大事なことは，最近のコンピュータ言語学が理論だけに走って，実際の地道なこのような研究をおろそかにしていることである。いやこれはコンピュータ言語学だけでなくコンピュータ・サイエンス全般を通じていえることかもしれない。「今日必要なのは実際に“もの”（ハードウエアまたはソフトウエア）をつくり，十分多くの実例に適用することによって検証する研究であり，単なる理論的解析ではな

い」と高橋秀俊氏はいわれる。

最後にはじめに紹介したアメリカのブラウン大学での例と，ドイツのザールブリュッケン大学での例を比較して，筆者の感想を述べておきたい。ブラウン大のあげた問題点のすべてを，ザールブリュッケン大の処理が含んでいるとは考えられず，すくなくともその一部ではないかと考えられるが，その重要な一部と考えられるであろう。むしろかなりの部分ということができるかもしれないし，相当の部分が一致しているといってもよいと思う。ことにそのうちの *'sow'* などをあげた部分は語い的にひん度がそう高くない部分に属しているから，用語調査のばあい語彙表の大勢に影響が及ぶものではない。語彙表の大勢に影響が及ぶのはエッガース氏たちが扱ったような用語の部分である。

このように考えてみると，ブラウン大とザールブリュッケン大の間に大きな距離があることが感ぜられる。少なくともその姿勢においてである。このような問題が非常に複雑な内容をかかえ，その分類，整理に，特に多くの頭脳を要することはいうまでもない。問題はその解決に実際に手を下して，とりかかるかどうかである。そして一方は *"beyond the reach of computer technology"* と達観してしまった。手をこまねいていればいつまでも昔のままである。現在でもそのままなのである。一方はそのしごとが重要であると認識した ("So stellte sich uns als erstes Hindernis das problem der Homographen in den Weg")。そうしてそれに「長年没頭」した。そしてついに"sind praktisch bereits gelöst" ということになった。現在ではできあがっているのである。

筆者はもちろん，ザールブリュッケンの業績を高く買うのである。もちろん方法そのものについては意見がないではないが。

日本語において，同形異語を弁別するために，どのような問題があるか，データについて調べてみなければならない。それも大量のデータについて調べてみる必要がある。さいわい，われわれは大量の KWIC のアウトプットをもっている。これはそのための最良のデータである。次回から，これを利用して同形異語の弁別アルゴリズムを考えていきたい。

参考文献

1. Sydney. M. LAMB: Outline of stratificational grammar, 1966
2. H. KUC ERA and W. N. FRANCIS: Computational analysis of present-day American-English, 1967
3. H. EGGERS u. a.: Elektronische Syntaxanalyse der deutschen Gegenwartssprache, 1969
4. Literatur und Datenverarbeitung, Bericht über die Tagung im Rahmen der 100-Jahr-Feier der Rheinisch-Westfälischen Technischen Hochschule Aachen, 1972
5. L. WEISGERBER: Die geistige Seite der Sprache und ihre Erforschung, 1971
6. W. Martin: Monika Rössing-Hager, Wortindex zu Geroge Büchner Dichtungen und Übersetzungen, ITL 13 1971
7. H. MITTERAND: Les mots francais, ≪QUE SAIS-JE?≫ 1963
8. L. TESNIERE: Eléments de syntaxe structurale, 1959
9. 高橋秀俊「電子計算機随想」『情報処理 13-7』1972
10. 「電子計算機による新聞の語彙調査」国語研究報告37, 1970
11. 「現代雑誌九十種の用語用字」国語研究報告25, 1964
12. 中野洋「品詞認定の自動化」国語研究報告39所収, 1971
13. 石綿敏雄「電子計算機による語彙調査の一実験」国語研論集所収, 1964
14. 〃 「ドイツのコンピュータ言語学」『計量国語学62』1972
15. 〃 「KWIC の設計」『計量国語学60』1971