高校教科書用語調査における言語単位

高校教科書用語調査における言語単位

鷲 岡 昭 夫

Ⅰ はじめに

ⅰ 言語単位の条件

(1) 調査目的を達成することができること

(2) 文法理論その他の理論の上で矛盾がないこと

(3) その他

ⅱ 過去の用語調査で用いた単位

(1) 単位の種類の系統

A 長い単位……文節（橋本文法による）中の，自立語と付属語＝α 単位・長単位・L単位

B 短い単位……最小単位（現代語で意味をになう最小の言語単位）の0～1回結合＝β 単位・短単位・S単位

(2) 一回の調査で用いた単位の種類数

A 単数……手作業の処理。主にα 単位またはβ 単位が用いられた。

B 複数……機械処理

a 「電子計算機による新聞の語彙調査」（二単位使用）

原文→長単位処理→長単位集計→短単位処理→短単位集計

b 「漱石・鷗外の用語研究」（三単位使用）

原文→C単位処理→L単位処理→S単位処理→CLS単位別集計

Ⅱ

ⅰ 過去の単位を用いるかどうか

国立国語研究所では，過去数回にわたって用語調査を行っている。今回は標題にあるように高校教科書の用語調査をすることになった。文章を対象とした用語調査では，まず文章から調査単位として語を抽出しなければならない。

一連の語彙調査の結果と比較するためには，言語単位は他と同じであったほうが便利である。しかし，今回，われわれが調査にはいるときには，すでにⅠⅱで述べたように多種類の単位が用いられていたわけであるから，過去の単位と同じにするという理由は弱くなる。また，今回の調査の目的や方法の面から検討をする必要もあるように思われた。そういうことから，今回の調査単位は，過去の単位には必ずしもとらわれない，ということを前提に検討されたものである。

ⅱ 今回の調査の性格と単位

(1)調査目的

今回の調査の主な目的は，「義務教育より高い段階における知識体系の形成にあずかる言語の実態を知る」ということにある。また，雑誌・新聞・小説なで，以前に調査したものの用語と教科書の用語とをくらべてみることももちろん予定されている。

(2)機械処理の進歩

昭和４９年度から国立国語研究所に漢字プリンターが導入された。これは，漢字かな交り文を，１分間1000行前後という高速で出力できるものである。この機械を用いて編集し出力したものを作業台帳として校正や情報付加の作業をすることができるようになった。また，文脈つき総索引（KWIC索引）も作れるので，必要な語の検索も可能になった。前回の新聞調査から用いはじめた電子計算機とあわせて処理の効率が格段にあがったと言える。

(3)単位の性格

「知識体系の形成にあずかる言語の実態を知る」ためには，「百年戦争」とか，「一次関数」など，ある長さで特別の意味をもつ語を調べる必要があるし，一連の用語調査から基本語を定めようとする目的からは，それらを「百・年・戦争」「一・次・関数」のように，基本的な語に分けて調べる必要がある。また，前項で述べたように検索に用いるとなるとなるべく短い単位のほうがどこからでもひくことができる。以上のことから，今回の調査は，長い単位と短い単位を用いることになった。

(4) 今回の単位

A長い単位

長い単位（Word単位，略してW単位とする）は，構文上，かかりうけの機能（展叙，統叙などと言われることもある）を中心に考えた。

結合の上，すなわち語構成の上から考えた単位の方が理論上すぐれているとか，「一次関数」や「第一次世界大戦」を一単位とするなら「ピタゴラスの定理」のように二文節以上にわたるものも一単位としたくなる等の反対意見も出たが，かかり受けの機能から考えた単位でも理論的に矛盾しないようにしうるものであるし，展叙・統叙の単位は文構成上の基本的単位であり，とくに生の文章から調査単位を切り出すのに有効である，などの理由で結局，かかりうけの上からの単位とすることに決まった。

ただし，今までの長い単位は，文節中の自立語，付属語の概念で定められていたので，たとえば，「雨の降る日」の「降る」と「雨の降った日」の「降っ」を同一の語としていた点にかかりうけという理論論の上から疑問を感じて検討をしなおした。すなわち，さきの「雨の降った日」の「降った」は，「雨の降る日」の「降る」と構文の上では同格だと考えられるのである。また「降るから」と「降ってから」「後から」と「降ってから」を比べてみても，「降って」と「降る」「後」と同じレベルにあるように思われる。

最近の構文論で，単語を「実質概念を表すもの」と「関係概念を表すもの」に分けるという学説がかなり行われるようになっているが，今回の調査でもこれをとり入れたわけである。

ついでながら，関係概念を表す語，われわれはこれを助辞と呼び一語－W単位にすることにしたが，これに助辞情報＜J＞を付けることにした。こうすることにより助辞の研究に便利なほか，W単位＋助辞，すなわちいわゆる文節単位（C単位）が自動的に合成できる，という利点があるからである。

B短い単位

短い言語単位は，検索のためにも，基本語を選ぶ手がかりとするためにも，原則的に最小単位（→Ⅰⅱ１B）を採用することにした（これは形態素Morpheme に近いのでM単位と名付けた）。

ただし，和語の「かわ」「やま」「ゆく」と，それに対する漢語の「川（せん）」「山（さん）」「行（こう）」などは，レベルが異なるように感じられるので，漢語（混種語中で結合している漢語要素を含む）は，一回結合までを一単位とすることになった（漢字漢語の研究用には文脈付き漢字KWIC を別途作成することにした）。

また，外来語や一字漢語に活用語尾の付いた「デモる」「愛す」「信じる」等では，語尾の「る」「す」「じる」等を切り離さないことにした。このため，「愛する」「信じる」の「する」「ずる」も切られないことになった。

「けだもの」「四角い」など，一部分に最小単位的なものをふくんでいる語でも，あとに現代語で意味を担えないと考えられるものが残る場合は切り離さないことにした。

Ⅲ

ⅰ 単位切り作業

作業の進め方は，原文（教科書）に，まずW単位の切れ目へ赤鉛筆で／を入れて分割し，次いでW単位の間に黒鉛筆で／を入れてM単位に分割する方式をとった。以下に単位切りの例を示す（ここではW単位の切れ目を／，M単位の切れ目を〃で示す）。

／現代〃社会／は／，／産業〃社会〃化／の／面／に／おい〃て／も／，／また／大衆〃社会〃化／の／面／に／おい〃て／も／，／さま〃ざま／な／問題／を／含ん〃で／おり／われ〃われ／の／不安／や／悩み／も／多く／は／これ〃ら／の／問題／と／深く／結び〃付い〃て／いる／。／

／最近／隕石／の／研究／が／進み／，／地球〃型〃惑星／が／誕生〃し〃た／とき／の／岩石〃質〃粒子／は／隕石／で／ある／と／いう／考え〃方／が／有力／と／なっ〃て／き〃た／。／

単位を切った原文は，原稿用紙へM単位ごとに一行ずつ転写するが，そのとき，単位の頭に，W単位先頭のものにW，その他のものにMという情報を付けることにした。前の例文について一部分を例示すると，つぎのようになる。

W現代
M社会
Wは＜J＞
W，
W産業
M社会
M化
Wの＜J＞
W面
Wに＜J＞

Wおい
Mて
Wも＜J＞
⋮

上の例で，Wから次のWが現れるまでが一W単位である。また，＜J＞の付いたW単位はすべて前のW単位に付けるといわゆる文節になるわけである。

このあと，この原稿用紙にさまざまな情報を付加し，漢字テレタイプでパンチして電子計算機処理にまわすのである。付加する情報については機会を改めて発表するつもりである。

ii W単位，M単位と他の単位との比較

(1) 短い単位系

(M単位)	(S単位)	(短単位)	(β単位)	cf.(W単位)
地球	地球	地球	地球	地球型惑星
型	型	型	型	
惑星	惑星	惑星	惑星	
が	が	が	が	が
誕生	誕生	誕生	誕生	誕生した
し	し	し	し	
た	た	た	た	
とき	とき	とき	とき	とき
の	の	の	の	の
岩石	岩石	岩石	岩石	岩石質粒子
質	質	質	質	
粒子	粒子	粒子	粒子	
は	は	は	は	は
隕石	隕石	隕石	隕石	隕石
で	で	で	で	で
ある	ある	ある	ある	ある
と	と	と	と	と
いう	いう	いう	いう	いう
考え	考え方	考え方	考え方	考え方
方				
が	が	が	が	が
有力	有力	有力	有力	有力
と	と	と	と	と
なっ	なっ	なっ	なっ	なって
て	て	て	て	

き た 。 結び 付い た 。 昭和 四 十 九 年 。	き た 。 結び 付い た 。 昭和 四十 九 年 。	き た 。 結び付い た 。 昭和 四十 九 年 。	き た 。 結び付い た 。 昭和 四十 九 年 。	きた 。 結び付いた 。 昭和四十九年 。

(2) 長い単位系

（W単位）	（L単位）	（長単位）	（α単位）
地球型惑星 が 誕生した とき の 岩石質粒子 は 隕石 で ある と いう 考え方 が 有力 と なって	地球型惑星 が 誕生し た とき の 岩石質粒子 は 隕石 である と いう 考え方 が 有力 と なっ て	地球型惑星 が 誕生し た とき の 岩石質粒子 は 隕石 で ある と いう 考え方 が 有力 と なっ て	地球型 惑星 が 誕生し た とき の 岩石質 粒子 は 隕石 で ある と いう 考え方 が 有力 と なっ て

きた	き た	き た	き た
。	。	。	。

(3) まとめ

A ○ M単位は，漢語の場合は他の短い単位とほぼ一致するが，それ以外は最小単位と同じ。

○ M単位では数量は，文字一字ごとに一単位としたが，他の短い単位では位を表わす十百千…等は切らない。

B ○ W単位は，代名詞，副詞，連体詞，接続詞，感動詞で他の長い単位とほぼ同じである。名詞については，名詞連続をすべて一単位とする点，L単位と等しく，長単位よりやや長く，α単位よりも長い。また，動詞・形容詞は，終止・連体形，命令形，および連用中止（形容詞は連用修飾も）の連用形の場合のみ他の長い単位と等しくそれ以外の場合は他より長くなる。

○ 助詞・助動詞の多くはW単位の中に含まれ，名詞，および活用語の連用修飾形（連用形や連用形に「て」の付いたもの）につくもののみ助辞とされ一W単位となっている。

Ⅳ おわりに

電子計算機を用いて大量な語彙調査を行うとなると，単位切り作業は，多人数でやることになる。したがって，単位もおのずから，誰にでも解りやすくまた，統一がとりやすいものでなければならなくなる。

今回の単位は，冒頭にあげた単位の条件の(1)(2)を優先的に考れることにしたが，結果的に上のこともじゅうぶんに満たすものとなった。

現在，W単位四十五万語，M単位六十万語（いずれも推定延べ語数）の単位切り作業が終わり，入力，処理にかかっているところである。

細部ではまだ修正しなくてはならないところも今後出てくる心配があるが，今一応作業が終わった段階で発表することにした。