計算言語学10－2
（1977. 6. 24）

高校教科書語彙調査の概要

中野洋
斎賀秀夫・土屋信一・鷲岡昭夫・野村雅昭
佐竹秀雄・斎藤秀紀・田中卓史（国立国語研究所）

0．目的

高校教科書の用語用字調査のうち，語彙調査システムが一応の完成をみた。このシステムは、語彙調査だけでなく、用字調査や用語（語の用法の研究）調査の基幹となるシステムである。　それら分析システムを作成する前に、このシステムを報告し，言語情報処理，計量言語学的な立場からの御意見を伺いたい。

1．これまでの語彙調査

国立国語研究所（以下、国研と略称する）では設立以来，日本語の大量語彙調査を続けている。　昭和41年には、電子計算機ＨＩＴＡＣ－3010を導入し、新聞三紙の語彙調査をおこなった。　その成果は，国研報告37，38，42，48に報告した。　この調査には、電子計算機を使うことによって生まれた、それ以前の調査になかったいくつかの特徴がある。　その一つは，多種類の語彙表，分析表が報告されたこと。一つは，作成されたデータは語彙調査以外にも使用されていること。一つは，データの複製が容易に可能なことである。　例をあげれば，次のようである。

1．長単位語彙表，短単位語彙表（各50音順と度数順表）-----集計単位の異なった語彙表

　層別語彙表-----分野別語彙表
　連接表-----短単位が長単位の中でどのように用いられたかを示す表。語構成を示す。
　語種別，品詞別語彙表-----たとえば，外来語表，動詞表など。

2．漢字表-----報告56で報告された。　漢字の使用率表，用例表。
　KWIC-----短単位300万語の用例。現在，進行中。
　その他，言語情報処理，言語統計用のデータとして使用されている。

3．いくつかの研究機関に　コピーがわたされ，言語処理用データとして用いられている。　たとえば，電子総合研究所第747号「電子計算機による自動索引の研究（下）」など。

以上のような特徴は，電子計算機を用いたことによって，はじめて　生まれたものであった。（かし，語彙調査用のデータとしては　二三の欠点もあった。　多数の作業者によって　データのエディットが　おこなわれたため，データの質にむらがあった。　よみがな，品詞情報の付加作業において，文脈がないために正しい情報が　つけられなかった。　異形同語・同形異語の判別が　完全でなかった。

調査対象 資料	調査対象 期間	調査方法	母集団 延べ語数	標本 抽出比（約）	標本 語数 延べ	標本 語数 異なり	語の単位	助詞助動詞の調査	漢字の調査
(1) 朝日新聞1紙	24. 6. 1～6. 30（1か月分）	全数	24万	—	24万	1.5万	β′	×	×
(2) 婦人雑誌2誌（主婦之友（全体））	25. 1～12（1年分）	標本	90万（推定）	1/6	15万	2.7万	α	×	○
（婦人生活（一部））		〃	33万（推定）	1/6.5	5万	1.0万		○（一部）	○
(3) 総合雑誌13誌（改造・世界ほか）	28. 7～29. 6（1年分）	〃	900万（推定）	1/40	23万	2.3万	β	×	○
(4) 現代雑誌90誌（五部門90誌）	31. 1～12（1年分）	〃	1.4億（推定）	1/230	53万	4.0万	β	○（一部）	○
(5) 新聞3紙（朝日・毎日・読売）	41. 1. 1～12. 31（1年分）	〃	〈長単位〉1.2億 〈短単位〉1.8億	1/60	200万	21.3万	α′	○	○
					300万（未集計）		β′		

国立国語研究所の語彙調査一覧（言語生活48.10）

昭和49年度から、高校教科書の用語用字調査がはじまった。　この調査にも電子計算機が使用された。

2、この調査の目的

「現代日本語の用語用字の実態を明らかにするために、これまで、新聞、雑誌九十種、総合雑誌、婦人雑誌を対称として調査を重ねてきた。　以上の諸調査のあとを受けて、国民が一般教養として、各分野の専門知識を身につける時に必要となる用語用字の実態を明らかにすることを目的として、高校教科書について用語用字の実態を調査する。」（国研年報から）

したがって、この調査は高校での教育活動全体における用語の位置を考えるためでは　必ずしもない。また、調査対象として　理科・社会の教科書9冊を選び教材的な性格を持つ国語の教科書を選ばなかったのは　この目的のためである。

3、この調査の特徴

前回の「電子計算機による新聞の語彙調査」とくらべて、この調査の特徴をあげると次のようになる。

(1) 調査対象である教科書の理科・社会に描かれている世界は、新聞や雑誌とくらべると、比較的、閉じた体系をもつ世界であるといえる。　このような世界における語彙の体系を調べることは、他の調査結果との比較に　意味がある。*1

(2) 電子計算機による語彙調査としては質が向上した。

- KWICによるエラーデータの検査を　とりいれたため、単位切り、よみがなつけ、情報つけエラーの発見が容易になった。　作業者による作業のゆれも、ここで統一することができた。
- 異形同語、同形異語の判別をシステムの中に入れた。
  異形同語（「書く・書か」などの活用による変化、「タバコ、煙草、TABACCO」などの表記による変化、「あめがふる　あまだれ、あま合羽」など音の変化）を一語にまとめるため、すべての語に　代表形に判別情報（「ようがある」に、「用」、「そのようだ」に　「様」、「勉強（よう」に　「む」、「へがある」に　「㊎」、「ある日」に「或」）を　KWICを調べることによって　つけることができた。
  この代表形と判別情報をキーにすることによって同語異語判別が可能になる。

(3) これまでのほとんどが　サンプリング調査であったのに対して、この調査は全数調査である。

- 文章における語彙の体系を調べることができる。
- 全数調査と抽出調査の比較をし　言語統計におけるサンプリングについての研究が可能。

---

*1　この種の、他の調査結果との比較によって、そこに描かれている世界、あるいは描かれ方について論じたものに　石綿敏雄「新聞の用語と雑誌の用語」（国研報告54「電子計算機による国語研究Ⅲ」，1974）がある。

4. この調査の概要

4-1. 調査の対象

昭和49年度に使用の高校の社会科・理科の教科書9冊を対象とする。（ただし 延べ語数は最終結果ではない）

	教科	出版社	延べ語数
社会	政治経済	自由書房	83,106
	倫理社会	東京書籍	71,565
	地理B	帝国書院	64,509
	世界史	三省堂	85,860
	日本史	山川出版	96,315
理科	生物I	清水書院	56,757
	化学I	大日本図書	52,918
	物理I	講談社	59,160
	地学I	実教出版	45,618

ただし

- 表紙・目次を除き、奥付の前までを対象とする
- 図表・写真・およびその周辺部分にある説明のことば、巻末の索引年表などは除く。
- 脚注は除く。
- 化学・物理などにある数式・化学式は式であることだけを入力する。
- 人名・地名の上下に付いているアルファベット表記・生没年は除く。

4-2. 調査単位

これまで国研の語彙調査で用いられた言語単位はいくつかあった。それらは下のように長い単位と短い単位に分けることができる。

- 長い単位……文節（橋本文法による）中の、自立語と付属語＝J単位・長単位・L単位
- 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＝B単位・短単位・S単位

これらの単位を今回の調査にも採用すれば、それらの調査結果との比較が容易になる。しかし、すでにもう多種類の単位が用いられているわけだから、これまでの単位を用いる理由は強くない。また、今回の調査の目的や方法の面からも、必ずしもこれまでの単位にはとらわれないということで、調査単位が考えられた。

調査目的にある、「各分野の専門知識を身につける時に必要となる用語」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、基本語彙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」のように分けて調べる必要がある。

また、分析のためにKWICを作ることになると、その単位は短い方がよい。このことから、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。

(1)長い単位（word単位の意でW単位と称する）構文上、かかりうけの機能を中心に考えた。

これまでの長い単位は、文節中の自立語、付属語の概念で定められていたので、「雨の降る日」と「雨の降った日」の「降る」と「降っ」は同一の語とされていた。かかりうけという理論の上から、「降る」と同格なのは、「降った」であると考えられる。また、「降るから」と「降ってから」、「後から」と「降ってから」を比べても、「降って」と「降る」「後」とは同じレベルにあると

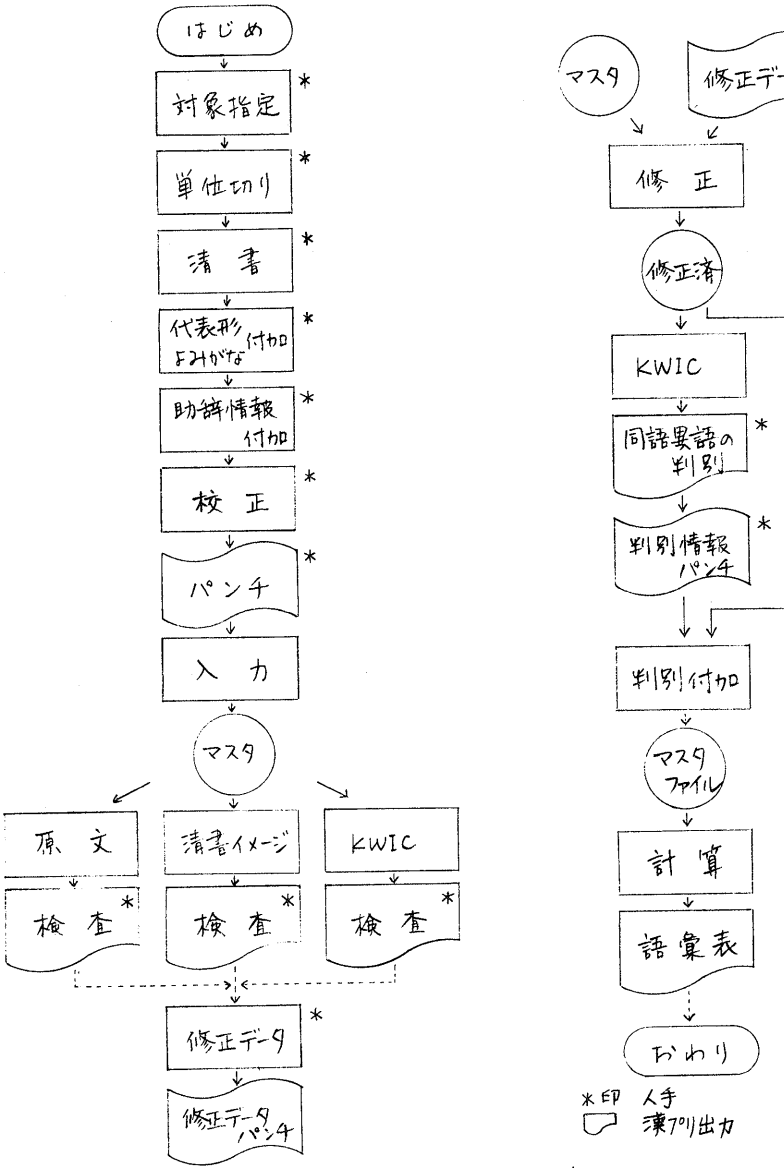
思われる。　これらは、単位を「実質概念を表わすもの」と「関係概念を表わすもの」とに分けるという考え方に従うものである。　このような単位をW単位とした。　関係概念を表わす語を助辞と呼び、これには　助辞情報〈丁〉を付け、いわゆる文節単位にもどれるようにした。

(2)短い単位　原則的に最小単位を採用した（これは形態素 Morpheme に近いのでM単位と名付けた）。

ただし、漢語は一回結合までを一単位とし、混種語の「デモる」「愛す」「愛する」などは一単位とした。　また、「けだもの」「四角い」など、一部分に最小単位を含むものでも、現代語で意味を荷えないものが残る場合は切り離さない。

4-3. 調査の方法　　電子計算機を用いて調査を進める。　作業工程の概要は以下の通りである。

教科書調査フローチャート



使用機器　HITAC-8250，高速漢字プリンタ（C5210），漢字タイプライタ

すでに本研究会で「大量言語処理におけるエラーと対策」（斉藤・霧岡・中野・米田，CL4-1，1975.12.15）と題して報告したように，語彙調査システムにおいては，作業者が多数はいり，調査が長期にわたり，かつファイルが多くなるので，プログラム・システムの作成よりも作業システムの管理，オペレートシステムの充実，ファイルの管理などの方が重要であり，神経を使わなければならない。全体的な管理システムの確立，及び機関と人との効率のよい運用方法が今後の課題である。

4-4 データ例

◦原文イメージ出力　（／はM単位の切れ目を，Ⓢはw単位の切れ目）

●1 生命 の 基礎 と し／て の 細胞● @ 細胞 が 生命 の 基本／単位 て
あ る と いう 細胞／説 は シュライデン （Ⓢ1／8／3／8 ） と シュワン （ 1／8
／3／9 ） に よっ／て となえ／られ／た が ， これ は 生物 の からだ が 細胞 か
ら でき／て いる と いう 事実 だけ を 根拠 に し／て いる の で は ない ． #
細胞 が 自分 と 同じ もの を つくり／だす 細胞／分裂 に よっ／て ， 生物／体 の

◦清書イメージ出力　（清書原稿のイメージの形式で印字）

通し番号	単位	出現形	情報	ルビ	読み	代表形	頁	段落	文	語	修正文種	検査
000001	00	〈 生物1P003〉	.	.	.	.	.	003	0000	000	K	DR
000002	00	●	.	.	.	.	.	003	0000	000		
000003	00	W 1	.	.	.	.	.	003	0101	001	見	N
000004	00	W 生命	〈	.	せい・めい	.	>	003	0101	002	見	K
000005	00	W の	〈J.	.		.	>	003	0101	003	見	
000006	00	W 基礎	〈	.	き・そ	.	>	003	0101	004	見	K
000007	00	W と	〈J.	.		.	>	003	0101	005	見	
000008	00	W し	〈	.		～する	>	003	0101	006	見	DO
000009	00	M て	.	.	.	.	.	003	0101	007	見	

◦KWIC出力　（教科書名，単位，代表形よみ，助辞情報 出典情報（ページ番号，段落番号，文番号，語番号） 文脈を印字）

〖 生物1 〗　M単位

代表形	情報	頁	段落	文	語		出現形
くらべる		181	02	03	020	るふつうの染色体に	くらべ て，長さ・太さとも70～
くる 来		149	01	06	022	いに明らかになって	き ている．1植物の◇題長発生種
くる 来		160	00	01	041	長く解決されないて	き た．ルーは，実験によって発生
くる 繰		125	03	05	027	固有の刺激とともに	くり かえして与えたときに，条件
くる		125	01	02	005	○学習動物は経験こ	くり かえすことによって，それぞ
ぐん		181	00	00	001	ためと考えられる．	群 ）は4つあることがわかった．

◦語彙表例

全体			見出し語	情報	度数	語其内		
順位	使用率	累積使用率				使用率	累積使用率	順位
23	7.3	493.4	なる		23	12.4	237.7	8
23	7.3	500.8	0	〔数〕	23	—	—	—
23	7.0	507.9	イオン		22	11.9	249.6	9
23	6.4	514.3	から	〔助〕	20	—	—	—
23	6.4	520.7	すいそ		20	10.8	260.4	10

5．20分の1サンプリングデータの分析

教科書調査のシステムを設計するために、各教科から20分の1を抽出（抽出単位：ページ）し、テストデータとした。そのうちの4教科（倫理社会、政治経済、生物Ⅰ、化学Ⅰ）について、以下に、分析結果を示す。

5－1．語彙量

	倫理社会		政治経済		生物Ⅰ		化学Ⅰ	
	延べ	異なり	延べ	異なり	延べ	異なり	延べ	異なり
語基*	2259	637	2697	807	1956	513	1847	465
		(3.5)		(3.3)		(3.8)		(4.0)
その他	1297	40	1643	47	1196	38	1268	44
		(32.4)		(35.0)		(31.5)		(28.8)
計	3556	677	4340	854	3152	551	3115	509
		(5.3)		(5.1)		(5.7)		(6.1)

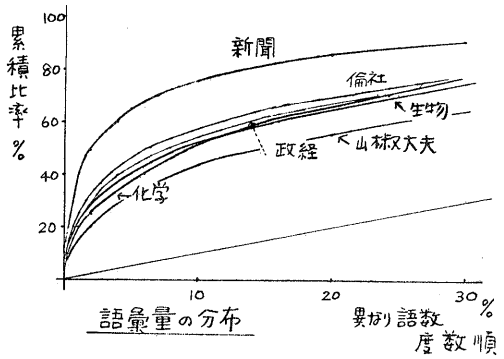
（ ）内は〔延べ÷異なり〕の値

＊語基には接辞（接頭語・接尾語、「あった」の「た」など）を含む。

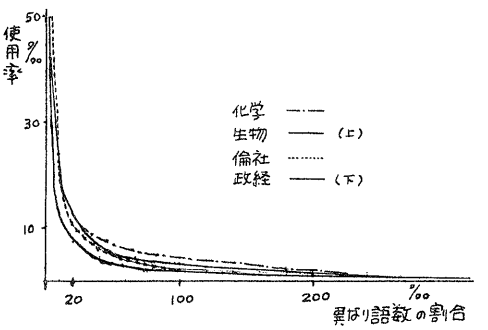
（延べ÷異なり）の値は、一語当りの平均使用頻度数を示す。これは、理科の方が社会より多い。すなわち、理科の方が語彙の豊富さからみれば貧弱だといえる。

その他には、助詞、数字、記号が含まれる。異なりでみれば、理科と社会に差はないが、使用度数からみると理科の方が多い。

5－2．語彙量の分布



語彙量の分布



縦軸に累積比率をとり、横軸に総異なり語数に対する度数順の順位の割合をとると、図のようになる。この図の計算には、記号・数字・助詞（あるいは助詞・助動詞：新聞・山椒大夫）を含まない。

調査の結果、すべて度数1なら図中で下の直線になる。曲線が上にいくほど、上位語の使用率が高くなる。

各科のグラフが直線と同じ角度になるのは、横軸でそれぞれ、政経－45.5、倫社－46.8、生物－54.8、化学－53.8、山椒大夫－42.6％のところである。

累積比率が50％になるのは、それぞれ総語数の6％（倫社）、8％（政経）、9％（生物）、10％（化学）のところである（新聞は2％、山椒大夫は16％）。

したがって、これは、社会の方が、度数1の語の割合が大きく、少数の語で延べ語数の多くをしめることを示している。いいかえれば、理科の方が、ある程度の使用度数をもった語が多いことになる。右のL字型分布の図で、理科の曲線の方が上に分布しているのはこのことを示している。これは、理科の方があるテー

マについて述べる部分（文章）では、同じ語がよく用いられるが、その語は他の部分ではあまり使われない... いいかえれば、いくつかの文章において、用語の共通度は理科の方が低いことによると推察される。*1 次のような文が理科には多いのだと思われる。

「たとえば、子葉が黄色で種子が丸い系統と、子葉が緑色で種子にしわがよっている系統を交配すると、Fはすべて子葉が黄色で種子は丸くなる。」（生物）

この文において、「子葉」「種子」は3回、「黄色」「丸い」「系統」は2回用いられている。しかも、多分、「子葉」「種子」は生物の教科書のあらゆる分野で数多く用いられる語とはいえないだろう。しかし、このことの確認は文章における語彙の体系の調査をまたねばならないだろう。

5－3．特徴語彙・共通語彙の抽出

4教科が比較できる下のような表を作った。

		使用率 ‰			
見出し語	度数	倫社	政経	生物I	化学I
ん	8	1,77	0,74	0,51	0,54
れる	103	15,93	8,89	17,89	4,33
られる	43	4,42	4,82	8,69	1,62
ら	14	1,77	1,48	1,53	1,62
よる	45	6,19	2,22	10,73	2,16
よう	82	10,18	8,52	11,24	7,52
ゆく	16	3,09	0,74	2,55	1,08
もちいる	11	0,44	0,37	1,02	3,78
みる	29	4,42	3,33	1,53	3,78
また	34	5,31	5,19	2,55	1,62

さて、共通・特徴語彙を抽出するために、χ^2値を用いた。

$$\chi^2 = \sum \frac{(f-r)^2}{r}$$

f：実測値　r：理論値

ここで、r_i＝（i教科の総のべ語数）÷（総のべ語数）×（語rの全体度数）

ただし、$f=0$の語は計算しない。

4教科に共通に出現した語は、「ん、れる、られる、もちいる、へんか、ば...」など90語であった。各総異なり語数872語のうちの1割強であり、意外に少なかった（対象は助詞、数字、記号を含まない。χ^2値の小さい語…理論値とのへだたりが少ない語の上位10語は表のとおりである。

---

*1. この傾向が他の文章、新聞や雑誌よりつよいとすると、高校教科書の調査において、サンプリング単位は非常に小さくとらなければならなくなり、サンプリング調査自体がむつかしいものになろう。

学校文法でいう自立語では、「よる・ゆく・もちいる・みる・また・へんか・ぶつ・ば(場)・はじめる・に(ニ)・なる・なに・なか・どうよう・とる・とき・できる・てん・つよい・つける・つくる・つく…」などであった。下線をつけた語は、教科書だけの共通語彙といえるかもしれない。

次に、他にはなくて理科だけ(または社会だけ)に現われた語、1つの教科だけに現われた語を度数の多い順にあげる。これらは、特徴語彙といえよう。

〔理科だけに現われた語〕 53語 ( )内は総度数を示す。

◎式(38)、水素(25)、反応(24)、分子(19)、実験(16)、C(16)、水(16)、H(14)、物質(14)、化合(11)、◆○○(11)、わかる(9)…

〔社会だけに現われた語〕 132語

社会(64)、日本(34)、産業(33)、物(29)、生活(26)、国(24)、おく〔「おいて」の形〕(22)、政治(21)、我(20)、経済(20)、憲法(19)、発展(18)、国民(15)、生産(15)、傾向(14)…

〔生物だけに現われた語〕 254語

発生(25)、細胞(24)、神経(17)、再生(14)、体〔たい〕(13)、子〔し〕(9)、刺激(9)、染色(9)、遺伝(9)、学習(8)、卵〔らん〕(8)、受精(8)、分化(8)…

〔化学だけに現われた語〕 240語

溶液(24)、イオン(22)、操作(17)、管(17)、塩素(16)、酸化(16)、量(15)、m(14)、水〔ず〕(14)、グラム(649)、試験(12)、濃度(12)…

〔倫理社会だけに現われた語〕 325語

大衆(22)、人間〔22〕、現代(16)、思想(12)、情報(12)、課題(11)、活動(11)、機能(8)、一〔ひと〕(5)…

〔政治経済だけに現われた語〕 460語

自治(18)、地方(18)、貿易(14)、輸出(12)、所得(9)、県(9)、議会(9)、外国(8)、明治(8)、万(8)、税(8)、失業(7)…

上の特徴語彙は、調査量を増やせば少なくなるかもしれない。

リストには、我々の生活における教養語彙が浮かびあがっているといえよう。またそれは単に特殊な専門語彙に限らず(倫理社会、政治経済の特徴語彙はもちろんのこと)、「この問題の発生をたどると」「単細胞の人間」「神経をすりへらす」「ヘドロの海を再生させよう」「青酸カリ溶液」などというように、一般語彙に入っている語を多く見出すことができる。

このような小さな分析においてさえ、この調査の重要性をあらためて認識することができる。

付記　4－2調査単位の説明は、霧岡昭夫「高校教科書用語調査における言語単位」(昭和50年度国立国語研究所研究発表会要旨)に従った。また、本研究会での発表(CL4－1、1975、12)にも、この単位について述べている。

国研報告37、38、42、48は国立国語研究所「電子計算機による新聞の語彙調査」および同(Ⅱ)、(Ⅲ)、(Ⅳ)であり、国研報告56は「現代新聞の漢字」である。