From vocabulary statistics to corpus-based studies

語彙調査からコーパスへ

宮島　達夫
(国立国語研究所　名誉所員)

キーワード
国立国語研究所，用例カード，基本語彙，生成文法，太陽コーパス

要　旨
国立国語研究所は創立当初から統計的な語彙調査をめざし，新聞・雑誌・教科書・テレビ放送など各種の資料について大規模な調査を行ってきた。それは統計的処理の面で先進的なものだったが，最近の英語圏の調査にくらべると代表性・規模などで問題がある。一方，大量の現代語用例にもとづく記述も国立国語研究所が開拓したものであり，現在開発中の１億語コーパスは，語彙調査と実証的記述の伝統を発展させるものとして期待できる。

1. 国立国語研究所と現代語研究

国立国語研究所が創設されたのは，1948年12月20日である。戦後まもなくで経済の復興におわれていた時期に，なぜこのような研究機関がつくられたのか。それは，日本の民主化にとって言語問題が重要な意味をもつという認識が，日本人のがわにも占領軍がわにもあったからである。明治以来の懸案だった漢字制限（当用漢字）とかなづかい改正（現代かなづかい）は，すでにこの前年に実施されていた。おくればせながら，そのような政策の裏づけとなるべき日本語の実態の科学的な調査が必要だったのである。国立国語研究所設置法第１条には「国語及び国民の言語生活に関する科学的調査研究を行い，あわせて国語の合理化の確実な基礎を築くために，国立国語研究所を設置する。」とある。

その誕生以来，政策と無縁でなかったにもかかわらず，創立からの60年をふりかえってみると，その業績としては，むしろ純粋に学問的なものが目立つ。とくに，現代語の研究を確立したことは，最大の功績といってよい。今からみると奇妙にみえるかもしれないが，現代語の研究は国語研究所の成立によってはじまったのである。それ以前の「国語学」の対象としていたのは国語史であって，それも奈良・平安からせいぜい鎌倉・室町どまりだった。東京大学で最初に東京語をテーマにした卒業論文が提出されたのが1935年だった。そのとき，教授から「近代語をやってもいいけれど，卒業してからこまるぞ」といわれたそうである[1]。言語問題に役立つ研究といえば，当然現代語を対象とすることになる。だが，当時，現代語をおもな対象とする研究者，方言を例外として，厳密にいえば現代標準語を対象とする研究者は，日本中にいなかったのである。初代の所長・西尾実の専門は国語教育と中世文学である。その下で研究部長をして，実質的に研究所を指導し，２代目の所長になった岩淵悦太郎の専門は国語史，とくに音韻史である。語

彙調査の中心になった3代目所長・林大は万葉学者であり，水谷静夫は卒業論文で古事記をとりあげた。方言の柴田武はトルコ語を，野元菊雄はハンガリー語を対象にしていた。かれらは，みな，現代語研究者として国語研究所にはいったのではなく，国語研究所で現代語研究者にそだったのである。全国的にみても，現代語の研究者は国語研究所から各地の大学にうつって，そこでまた新しい研究者をそだてるというケースが少なくない。研究所自身が現代語を研究しただけでなく，その内外で現代語研究者をそだてたことも，国語研究所のおおきな業績である。

2. 統計調査と記述

創立当初の研究所をふりかえると，やることすべてが新しい，という熱気が感じられる。新しかったのは現代語という対象だけではない。若い研究者たちは，つぎつぎに新しい研究方法を身につけていった。人文系の研究では個人研究が中心だが，国語研究所では創立のときから個人研究ではなく共同研究を建て前としてきた。このことが，（当時としては大規模な）語彙調査や全国にわたる方言調査を可能にしたのである。まだめずらしかった録音機で，いちはやくナマの音声を録音して研究したり，理系にしかなかった電子計算機を文系でまっさきに導入したりなど，機器の使用にも積極的だった。社会言語学的な実態調査は世界的にみても早いものに属する。そのような新しい方法のひとつとして，統計の活用がある。数をかぞえるだけの記述統計なら戦前からあったが，検定・推定にいたる新しい統計を武器とした研究は，やはり国語研究所がリードしたものである。「計量国語学会」という学会がある。形式的には国語研究所と関係のない学会だが，創立の中心になったのは，当時の研究所にいた若手研究者たちだった。小さい学会だが，創立1956年で，機関誌『計量国語学』が50年つづいているという，計量言語学では世界に類をみない学会である。

研究所創立後まもなく，国立国語研究所資料集2『語彙調査 —現代新聞用語の一例—』(1952)が出た。これは朝日新聞1ヶ月の統計をとったもので，いわば語彙調査の習作，といった感じのものである。新聞1月分といっても今よりずっと小さく，延べ語数20万語ほどである。また，その前年，国立国語研究所報告3『現代語の助詞・助動詞 —用法と実例—』(1951)が刊行された。話しことばの研究は録音機の導入をまたなければならないが，現代語の書きことばなら言文一致の成立以後いつでも調査の対象にできたはずである。しかし，そうはならなかった。口語文法の本はいくつも書かれたが，それらの多くは実態の調査にもとづいたものではなかった。古代語は研究者の直観によっては記述できない。研究者が古代語の研究から出発したのだから，古代（奈良・平安時代）の言語事実を調査したうえで文法を書くように，まずは現代語の調査をすべきだったのだが，その本格的な作業は『現代語の助詞・助動詞』(1951)にはじまる。

こうして，一方では統計的手法をつかって言語の全体像を巨視的にながめる行き方，もう一方では微視的に言語事実を記述する行き方が，国語研究所の研究のなかに確立した。語彙・文法を対象とした研究について，おもなものを年代順にあげると，表1のようになる。実態調査を中心にしたので，同音語や類義語についての研究のように試験的な手法のものははぶいた。漢字の調査は重要な項目だが，ほぼ語彙調査に付随しているので省略した。調査の名まえは略称をつかっ

た。くわしい名称は付録をみていただきたい。

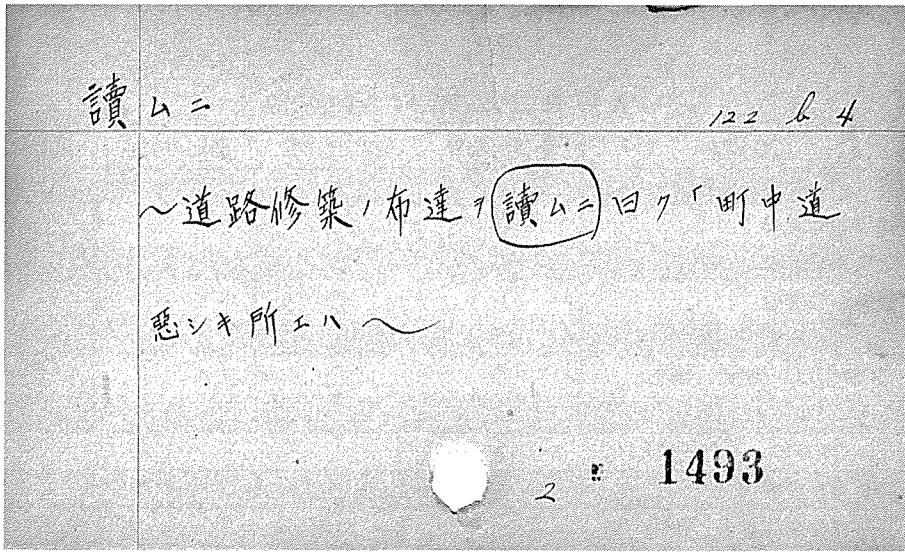
表1　国立国語研究所の語彙調査

年	語彙調査	全体的統計	統計的分析	個別の記述	実例の提示
1951	現代語の助詞・助動詞			◎	◎
1952	朝日新聞調査	◎			
1953	婦人雑誌調査	◎	○	○	
1955	談話語の実態		◎		
1957	総合雑誌調査	◎	○		
1960	『郵便報知』（明治）調査	◎	○		
1960	話しことばの文型			◎	○
	（コンピューター導入）				
1962	雑誌九十種調査	◎	○	○	
1964	（『分類語彙表』初版）				
1967	〈Brown コーパス〉				
1970	新聞調査（電算機）	◎			
1972	動詞・形容詞・アスペクト			◎	○
1983	教科書調査（電算機）	◎	○		
1987	『中央公論』経年調査	◎	○		
1995	テレビ調査（電算機）	◎	○		
1997	国定読本索引	○			◎
2004	（『分類語彙表』増補改訂版）				
2004	日本語話し言葉コーパス			○	◎
2005	雑誌70誌調査	◎			
2005	太陽コーパス			○	◎

3. カードによる用例採集

初期の語彙調査においては，単語をカードに書きとって50音順にならべ，集計する，という方法がとられた。ある程度の文脈は最初の朝日新聞調査のときからつけられていたようである。しかし，この方法は手間がかかる。それで，婦人雑誌調査の途中から，あたらしい方法がとりいれられた。それは，前もって調査すべき箇所をカードに印刷しておき，採集すべき単語なり漢字なりに○をつける，というやり方である。その複製カードも，最初は手書きのガリ版だったが，総合雑誌・雑誌九十種の調査では，邦文タイプライターによる謄写印刷になった。それでもカード作成にあたっては厳密に校正する必要があったが，動詞・形容詞の記述や『中央公論』の経年調査では，原文をコピーしてカードをつくるようになり，校正のわずらわしさもなくなった。

カードの例を図1から図4に示す。

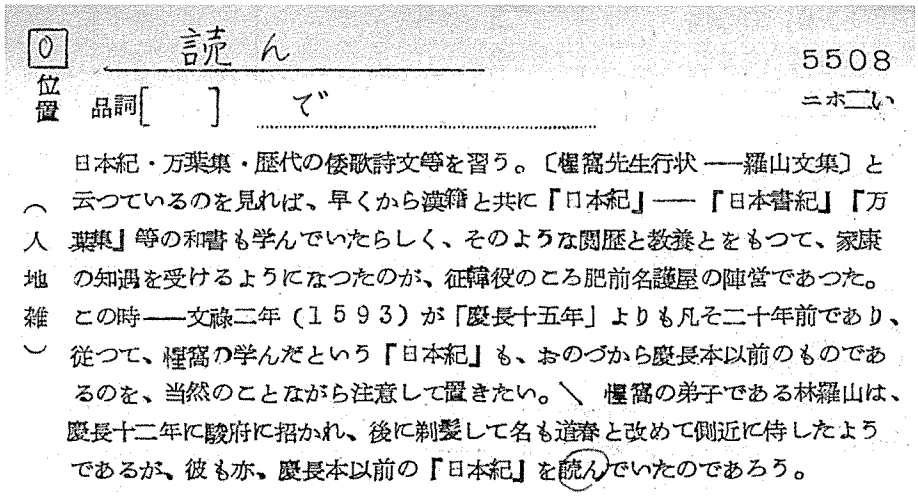

讀ムニ　122 6 4

～道路修築ノ布達ヲ讀ムニ曰ク「町中道

惡シキ所ヱハ～

2 : 1493

図１　手書きカードの例（『郵便報知』）


0 位置　読ん　5508

品詞[　]で　ニホ二い

〔人地雑〕

日本紀・万葉集・歴代の倭歌詩文等を習う。〔惺窩先生行状 ―羅山文集〕と云つているのを見れば、早くから漢籍と共に『日本紀』―『日本書紀』『万葉集』等の和書も学んでいたらしく、そのような閲歴と教養とをもつて、家康の知遇を受けるようになつたのが、征韓役のころ肥前名護屋の陣営であつた。この時―文祿二年（１５９３）が「慶長十五年」よりも凡そ二十年前であり、従つて、惺窩の学んだという『日本紀』も、おのづから慶長本以前のものであるのを、当然のことながら注意して置きたい。＼惺窩の弟子である林羅山は、慶長十二年に駿府に招かれ、後に剃髪して名も道春と改めて側近に侍したようであるが、彼も亦、慶長本以前の『日本紀』を読んでいたのであろう。

No 96　(54年S)

図２　邦文タイプ印刷カードの例１（総合雑誌）

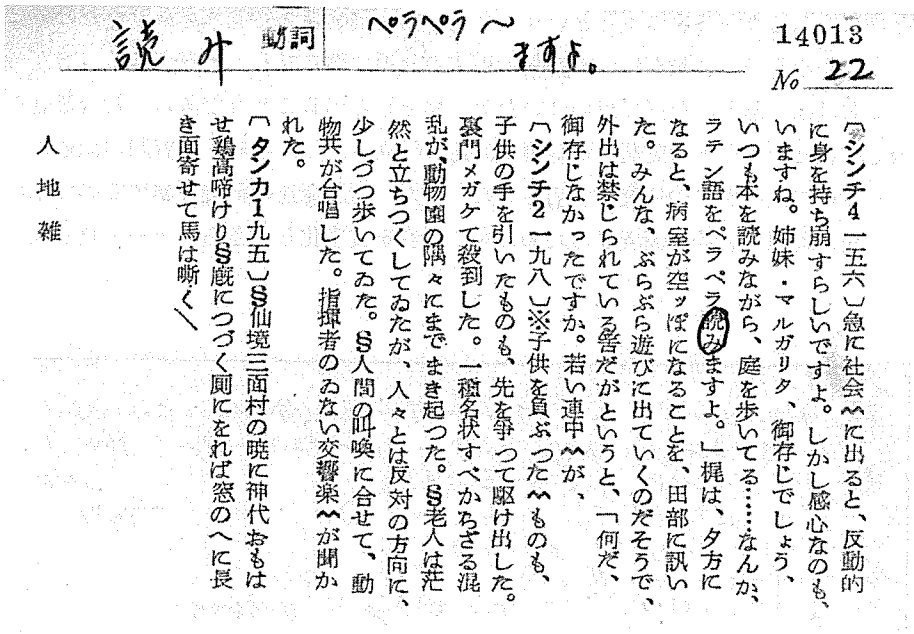

読み 動詞 ペラペラ～ます。 14013 No 22
(シンチ4 一五六)急に社会へに出ると、反動的に身を持ち崩すらしいですよ。しかし感心なのも、いますね。姉妹・マルガリタ、御存じでしょう、いつも本を読みながら、庭を歩いてる……なんか、ラテン語をペラペラ読みますよ。」梶は、夕方になると、病室が空っぽになることを、田部に訊いた。みんな、ぶらぶら遊びに出ていくのだそうで、外出は禁じられている者だがというと、「何だ、御存じなかったですか。若い連中へが、(シンチ2 一九八)※子供を負ぶったへものも、子供の手を引いたものも、先を争って駆け出した。裏門メガケて殺到した。一種名状すべからざる混乱が、動物園の隅々にまでまき起った。S老人は茫然と立ちつくしてゐたが、人々とは反対の方向に、少しづつ歩いてゐた。S人間の叫喚に合せて、動物共が合唱した。指揮者のゐない交響楽へが聞かれた。
(タンカ1 九五)S仙境三面村の暁に神代おもはせ鶏高啼けりS厩につづく厠にをれば窓のへに長き面寄せて馬は嘶く〳〵
人
地
雑

図3　邦文タイプ印刷カードの例2（雑誌九十種）

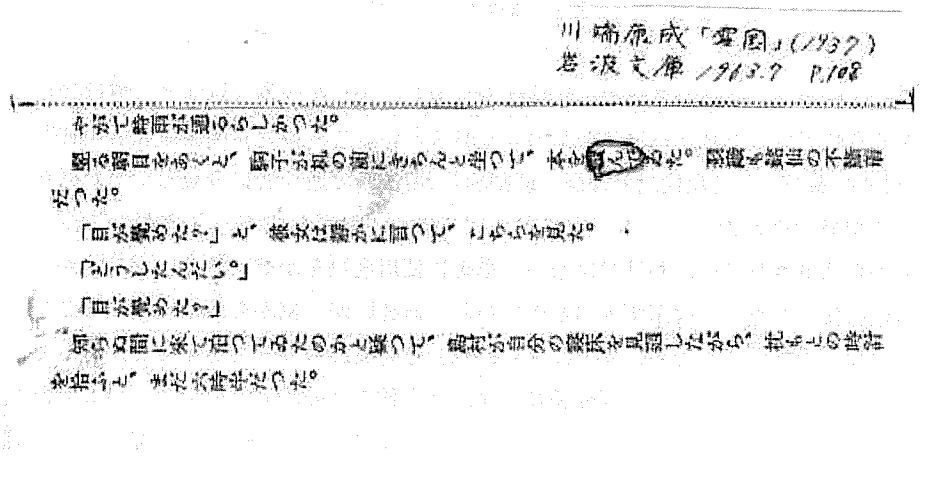

川端康成「雪国」(1937)
岩波文庫 913.7 P.108

図4　コピーカードの例（動詞記述）

原文をコピーしてカードをつくるという〈発明〉は，語彙調査の能率をあげただけではない。カードには，かなり長い文脈がはいるから，用法の分析に役立てることができる。また，たくさん複製しておけば，あとになって思いついたテーマをしらべることもできる。総合雑誌や雑誌九十種の調査に使ったカードは動詞・形容詞の記述に役立ったし，動詞・形容詞の記述のために印刷したカードはアスペクトの研究に利用された。また，言語発達の研究に属するので上の表からははぶいたが，幼児の言語を調査するためにも，録音を文字化してさらにカード化したものが使われた[2]。

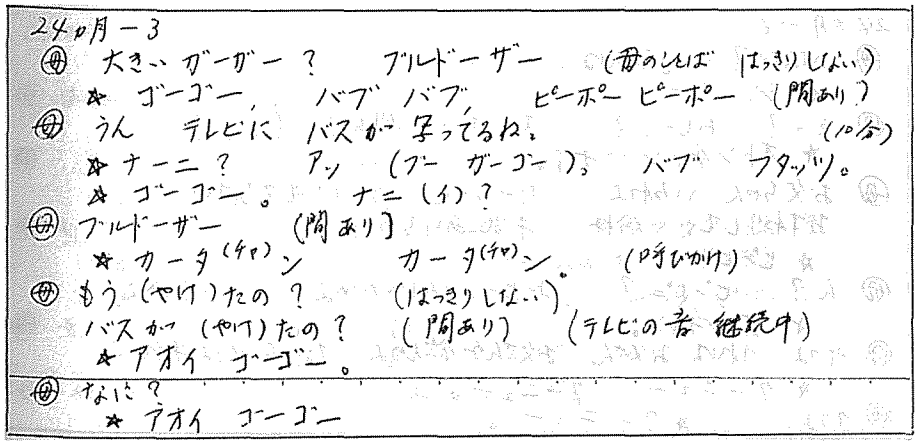


図5　幼児のことば資料

これらは，いわば，コーパスへの一歩手前の，コンピューター導入以前としては最後の段階である。ある意味ではコンピューターなしのコーパスといってもいい。国語研究所の初期に『語彙調査』（頻度の調査）と『現代語の助詞・助動詞』（用法の記述）とに分裂していた研究が，雑誌九十種調査で統合されたのは，カードのおかげである。ただし，カードのもつこのような意味が，最初から認識されていたわけではない。単純に使用度数をかぞえるという目的からすれば，かぞえるのに使ったカードは用ずみのものである。研究所が一橋から西が丘へひっこしたのは雑誌九十種の調査の途中だったが，すでにすんだ総合雑誌調査のカードをすてるか持っていくかを相談した記憶がある（総合雑誌・郵便報知・雑誌九十種などの調査カードは，いまも研究所に保管されている。朝日新聞・婦人雑誌のカードは廃棄されたようである）。手書きのカードからコンピューター利用のあいだに，謄写印刷のカードという段階があったのは日本だけだろうか。

corpus を原義にちかい「言語資料の総体」ととらえるなら，青空文庫でも日本古典文学大系でも，すべての言語作品がはいってしまうが，現在問題になっているコーパスの範囲としては，言語研究のために集められたもので電子化されているもの，といったところだろう。「でも，コーパス言語学は，原則的には，ほとんどのことはコンピュータをつかわない作業で十分に処理することが可能なのです。このことは，しっかりと肝に命じておく必要があります。」（アシュト

ン・バーナード 2004：5）電子化以前のコーパスについてはFrancis(1992)を参照。電子化という条件をはずせば，国語研究所の語彙調査にともなう分析・記述は，「コーパス言語学」とよんでいいだろう（『国語研究所報告25 現代雑誌九十種の用語用字　第3分冊分析』(1964)を参照)。

コーパスの条件としては，さらに，公開性という条件もつけたすべきかもしれない。雑誌九十種の調査データは最近カードからコンピューターに入力された。また，これを受けついだ雑誌70誌の調査は，最初からカードを使わずにコンピューター利用によってなされた。これらは著作権の関係で公開しておらず，国語研究所内部の利用にかぎられているので，その意味ではコーパスとして不完全である。おなじような例は外国にもある。SEU(Survey of English Usage)Corpusは電子化以前のコーパスとして最後の，過渡期的なものだったが，のちに電子化された。しかし，版権の関係でロンドン大学 ユニバーシティー・カレッジの外では使えない，という(Kennedy 1998：19)。

われわれは，調査にあたって，

「正確なる意義解釈の上には用例を集める程大切な事は無い。山成す用例の前には百の疑問も自から氷解し千の議論も立所に鳴を鎮めるものである」（市川三喜）

ということばをモットーとした（『国語研究所報告43』：5）。ただし，実例を重視し，調査にカードを利用したのは，国語研究所だけではない。研究所の外でも，古典語の研究には古くから使われていただろうし，現代語の研究が一般化してからは，当然そこでも利用された。とくに意識的に実例主義をとったのは，奥田靖雄氏を中心とする言語学研究会である。もっとも，そこで作られた膨大なカードのおおくは，文庫本の文学作品を切り抜いてはりつけたもので，カードを印刷・複製して利用する方向には，あまりむかわなかった。当時国語研究所の所員だった高橋太郎・鈴木重幸・宮島達夫らは言語学研究会のメンバーでもあり，国語研究所の研究成果の一部は同時に言語学研究会の成果である，といえる面もあった。生成文法の全盛時代に，日本の学界が全体として主観主義に流されなかった理由の1つは，国語研究所および言語学研究会の用例中心主義がかなりの成果をあげていたことがあると考えられる。

4. 雑誌の語彙調査と基本語彙

日本語の語彙調査は，基本語彙・基本漢字をあきらかにする，という目的をもっていた。雑誌九十種の調査結果を記述した『国語研究所報告21』は「調査結果は実態の記述をまずもって目標とするが，それにとどまらず，基本語彙の選定その他の国語国字問題を考える際の参考資料としても役立つ事を念願した。」（p.1）とのべている。おなじ趣旨のことは，それ以前の婦人雑誌・総合雑誌の調査報告書にもみえる。対象としてまず婦人雑誌をとりあげたのは，女性用語の研究のためではなく，婦人雑誌が生活に密着した記事をのせていて，生活基本語彙をしるのにふさわしい，と考えられたからだった。しかし，最大の範囲をカバーした雑誌九十種の調査でも，日本語の基本語彙をあきらかにする，という目標にはへだたりがある。

140位「身頃」　　343位「増資」
513位「ぬいしろ」　　356位「当社」
718位「えりぐり」　　479位「配当」
785位「ダーツ」　　753位「投資」

など，裁縫や経済関係の用語が上位にならんでいる一方，5,158位に

うたがう，応援，おもちゃ，金持ち，勘，看板，煙，交替，故障，さっさと，姿勢，市民，乗客，新年，スキー，スピード，センター，竹，ためる，知恵，散る，テープ，眠い，のんびり，話しかける，ハンカチ，昼間，文明，ボーナス，翻訳，満員，見本

など，もっと基本的ではないかと思われるものが，ずっと下にある（同じ順位の単語がたくさんならんでいるのは，頻度の同じ単語を同じ順位にするという方針をとったためである）。1994年の雑誌70誌の調査結果も同様である。ここでは，1,000位以内に

417位「素材」　　434位「ポイント」
561位「設定」　　648位「CD」
671位「性能」　　828位「ソフト」
747位「本体」　　858位「アルバム」

など，特定の専門分野に多い単語があるのに対し，5,000位に

梅，ガソリン，固まる，規則，給料，許可，故障，知らせ，杉，すし，ソース，そっと，近ごろ，爪，つらぬく，どっち，番地，ひげ，孫，間近，まね，都，夢中，役所

などがある。

では，英語の調査ではどうか。BNC(British National Corpus)でみることにする。Leech et al. (2001)は，順位よりも度数の表が利用しやすいので，ここでも度数をつかう。ただし，ここでは延べ100万あたりの度数をだしている。総数が1億だから，たとえば20とあるのは100万あたり20，つまり総数では2,000ということになる。

council　348
provision　129
investment　124
sector　110

などの抽象的な単語にくらべて，

bread　38
wise　24
absent　15
boil　12

などの日常語の度数がひくいのは，書きことばを主体にしたためだろうか。つまり，使用度数だけから基本語彙をきめるのは，英語でもムリなのである。それにしても，雑誌九十種ほど特殊な語彙が上位にくることは，少ないようにおもわれる。BNCが英語を代表している度合いは雑誌九十種が日本語を代表している度合いよりも，うえにあるようである。その原因としては，

BNCが単行本を多くしらべていて，これが雑誌単独よりも偏りがすくなかった，ということが考えられる。また，国語研究所の調査では，およそ日本語であるかぎり，ふつうの文章でないものも調査対象にしている。たとえば，将棋の棋譜・相撲の星取り表・洋裁の型紙など，特殊な名詞・数詞がならんでいるだけで，文章になっていないものである。おそらく，BNCの調査には，このような対象はなかっただろう。

国語研究所が日本語教育のための基本語彙を選定したとき，直接参考にしたのは語彙調査の結果ではなく，いわば語彙調査にともなう副産物のようなかっこうで作られてきた『分類語彙表』だった。基本語彙の選定には，50音順の辞典などを台帳にするより，意味分類された語彙表を台帳にする方がいいのは，あきらかである（ただし，旧版『分類語彙表』には雑誌九十種の上位語に印がつけてあるから，語彙調査の結果も，間接的には参考になったはずである)。なお，これからのコーパス利用にあたっても，意味分類が利用される機会はあるだろうが，言語研究に『分類語彙表』を使うことは，すでに多くの実績がある（宮島・小沼 1994)。

国語研究所の語彙調査は，雑誌九十種からあとは，日本語全体の基本語彙をあきらかにし，国語問題に寄与する，という大きな目標をたてていない。つぎの新聞の調査は，語彙調査の結果をだすよりもコンピューター利用の成果をだすためのものだった。中学・高校教科書やテレビ用語の語彙調査は，たしかに全国民の言語生活に大きな影響をもつが，むしろ日本語の重要な1つの側面，位相をあきらかにする，という位置づけがふさわしいようにおもう。最近の雑誌の調査は，以前の九十種調査と同様の使命をもつものではなく，「現代日本の語彙の実態の一面を把握することを目的」としている。教科書・テレビなど各種の分野の位相を前提にして雑誌の位相をあきらかにすること，および40年前の雑誌調査との比較という点に意味がある。こうして，いろいろ問題はあるものの，雑誌九十種調査は，依然として日本語を代表する統計的調査としての位置をしめている。

5. 雑誌語彙調査の評価

それにしても，（今や「大規模」とはよべないかもしれないが）この程度の各種語彙調査が引きつづいてなされている点では，日本語は英語とならぶ特別な言語であるようだ。Leech et al.(2001：ix)はスペイン語・フランス語の統計としてJuilland et al.(1964，1970)を，ドイツ語についてはKaeding(1898)をあげているにとどまる。欧米で語彙調査がへったのは，単なる度数だけでは利用価値がすくないからではないだろうか。見出し語立てをしない調査ならすぐにできるが，利用できない。見出し語立てをするには，たいへんな労力を必要とし，それに見合うだけの価値は，やはりない。日本語と同様に，単なる度数だけで基本語彙をきめるのは，むずかしいはずである。

雑誌九十種の語彙調査は，1962年の発表当時，世界最高の水準にあっただけでなく，ある意味では現在もそうだといっていい。調査した語数だけからいえば，外国には1億語のものもあり，日本でも44万語という九十種の規模をこすものがすくなくない[3]。しかし，それは依然として模範的な語彙調査だともいえる。3つの観点からこの調査を評価することにする。最初の本格的な

コーパスである，アメリカ英語のBrownコーパス(1967)，ついでこれにならったイギリス英語のLOB[Lancaster-Oslo-Bergen]コーパス(1982)とくらべて考える。

5.1. 見出し語立て

最初発表された形ではBrownコーパスもLOBコーパスも，見出し語立てをしていないgraphic wordsの統計であって，take，takes，took，taken，takingが別語とされる一方，助動詞のcanと名詞のcanとは区別がなかった。その後，機械的に処理できる範囲では見出し語のもとに語形が合併されたが，土手のbankと銀行のbankのように人間の目で判別しないと区別がつかないものは，そのままになっている。単語ごとに分かち書きされる表音文字の世界では，graphic wordsの統計は，コンピューターを使えば，なんの苦労もなくできる。英語では「The National Institute for Japanese Language」が6つの単語からできていることは，だれでも分かるし，コンピューターの自動認識にも問題がない。パソコンで百万語の調査をするのは数分ですむだろう。だが見出し語立てはできない。それには数年かかるかもしれない。幸か不幸か，漢字かなまじり文では，最初の段階は簡単だが，あとの処理が大変，というわけにはいかない。まず，入力の手間がアルファベットの比ではない。自動読み取りの精度も格段におちる。しかも，どっちみち，単語ごとに切るのにたいへんな労力がかかる。「国立国語研究所」をどう切るか，という問題には，何とおりかの答えがある。専門家なら単語の認定が確実にできるはずだ，とおもうかもしれないが，じつは逆で，日本語研究の専門家ほど，いろいろな条件を考えて，多様な切り方を作りだす。「国立国語研究所」についても，

国立国語研究所　　(1語)
国立／国語研究所　　(2語)
国立／国語／研究所　　(3語)
国立／国語／研究／所　　(4語)

と，4とおりの単語認定が可能である（雑誌の調査にあたって，実際にとられた方針は「国立／国語／研究／所」と4語にわけることだった。その結果，「研究所」だけでなく「小学校」も「自動車」も「具体的」も，語彙表にはでてこない)。おなじだけの労力を，その先同語別語の判別をして見出し語を立てることにつぎこむとしても，同語別語の判別をしないのにくらべて，いわば五十歩百歩である（5対10では大きいようだが，英語の調査なら1対100以上の開きがでるだろう)。だから，コンピューターを利用するようになっても，日本の語彙調査は人間の目と手で同語別語の判別をして見出し語を立てるのを原則とした。例外は，コンピューターを使った最初の大規模語彙調査である新聞用語の調査で，ここでは人手をできるだけ使わずにコンピューターにやらせるという方針のため，「一月」には和語の「ひとつき」と漢語の「いちがつ」がふくまれ，「いった」「言った」「行った」「言う」「行く」は，それぞれが別語とされた。こんな語彙表を作っても，あまり意味はないが，コンピューターを導入した以上，早く結果がでることを見せなければならない，という理由があったのかもしれない。それ以後の国語研究所の語彙調査は，教科書もテレビ用語も，機械と人間の共同作業で見出し語立てまですました結果をだすのを

原則とし，したがって膨大な量を短時間に処理することはできなかった。

5.2. 標本抽出の方法

雑誌九十種の調査のレベルがたかいというのは，見出し語が立ててあるだけでなく，標本抽出に厳密な無作為抽出がとられ，統計的な管理がしっかりしているからである。母集団は1956年度の雑誌九十種合計226,358ページ。そこから8分の1ページを単位に，227分の1にあたる7,983箇所を抽出する。1箇所あたり，ほぼ55語になる。Brown コーパス・LOB コーパスは，やはり無作為抽出をしているが，2,000語の500箇所だから，九十種にくらべてずっとあらい。そもそも，ある段階で主観をまじえているようで，厳密には，母集団がはっきりしない。ただし，雑誌九十種の調査でも，抽出が厳密に客観的におこなわれたのは，母集団から標本をとりだす段階で，それ以前の，母集団の決定にあたっては主観のはいる余地があった。雑誌の範囲をきめるについては，専門誌・青少年向け雑誌などをのぞいた。売れ行きも考慮しており，評論・芸文については部数1万以上，娯楽・スポーツでは7万以上，とされているが，その根拠となるデータは確実ではないし，線引きは主観的なものである。

5.3. 代表性

しかし，統計的に厳密でないことを，研究のレベルの問題としてとらえるのは，ただしくない。Hofland & Johansson は，LOB コーパスが厳密な無作為抽出をあえて守らなかった理由として，つぎのようにいう。「LOB コーパスの真の代表性は，文章の重要なカテゴリー・下位カテゴリーをふくめるよう計画的に心がけ，盲目的な統計的選択にまかせなかったことから生ずる。」(Hofland & Johansson 1982： 3)。LOB はアメリカ英語に対応するものとして Brown コーパスを忠実になぞったものである。それらのコーパスは，基本的に，A. 図書　B. 新聞・雑誌　C. 政府刊行物　という3種類の資料をもとにしており，これをきめたときに，すでに厳密な無作為抽出をあきらめていたはずである。統計を犠牲にしても，かれらがまもろうとしたのは，英語をよりよく〈代表する〉コーパスをつくることだった。それが成功したことは，アメリカ英語の Brown コーパスとイギリス英語の LOB コーパスとが結果的によく似た上位語をもっていることにしめされている。上位50語はほぼ完全に一致しており，ちがいは Brown の so と LOB の more が相手側の50語にはいっていないことだけである(Hofland & Johansson 1982：18)。また，それらがアメリカとイギリスとの文化的な差を示唆していることも，その代表性を保証する。Brown コーパスでアメリカ英語を，LOB コーパスでイギリス英語を代表させ，それらを比較するという研究がされており，tea がイギリスに，coffee がアメリカに多い，などというのは調べるまでもないが，女性をあらわす she，girl，woman がイギリスに，男性をあらわす he，boy，man がアメリカに多いというのは，調べてみてはじめて分かった結果である。標本抽出のゆれでないとすれば，これがなにを意味するのかは，興味がもたれる点である。

雑誌九十種ではどうか。それが代表するものは，あくまで母集団としての1956年度の雑誌九十種各号の総体，226,358ページであって，それ以上ではない。ここでは標本と母集団との関係が

はっきりしているから，たとえば「～に」と「～へ」の量的比較が標本について可能なだけでなく，母集団についても推定することができる。しかし，われわれがほんとうに知りたいのは，ある年度の雑誌九十種についてではなく，日本語についてである。学問的な態度をたもつかぎり，われわれは，一歩母集団をはなれれば「日本語の書きことば」全体についてはもちろん，次年度の雑誌九十種についてもなにもいえない，ということを，みとめなければならない。これは雑誌について全数調査をしても同じことである。

6. 雑誌語彙調査とコーパスにみる語種・語彙の変遷

しかし，日本語の書きことば全体についての統計がないので，便宜上語彙調査の結果を利用するよりしかたがない。たとえば，日本語の語種分布については，今でも雑誌九十種調査の結果がよくひかれる。以下に，あたらしい雑誌70誌の結果とあわせて，延べ語数の比率を図6にあげる。ただし，九十種調査は雑誌の本文だけを対象にし，広告は調査しなかった。70誌調査では本文・広告をあわせた結果と本文だけの結果とをだしているが，ここにあげるのは本文だけの結果である。

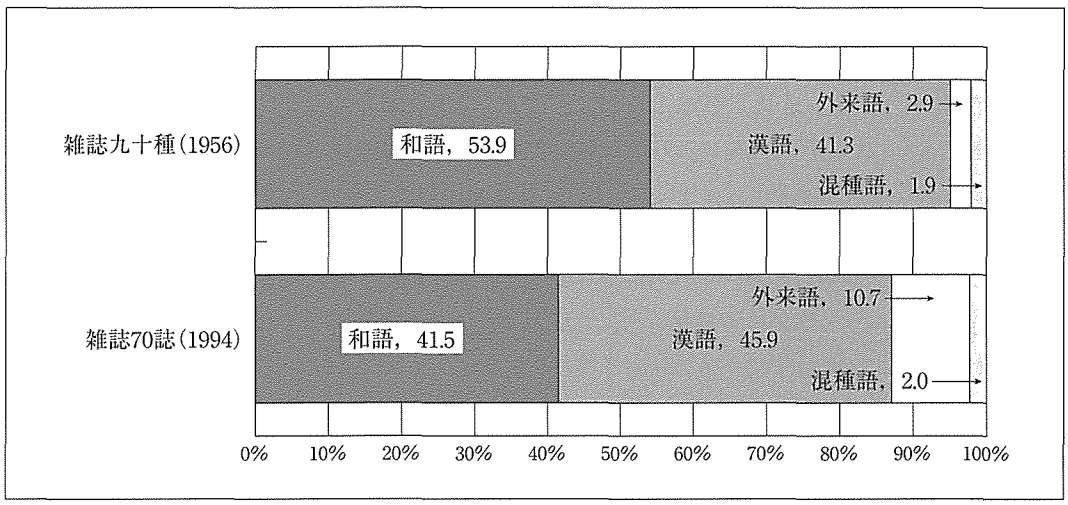


図6　雑誌の延べ語数の語種比率

これでみると，最近は漢語の量が和語をおいこしたこと，外来語が激増していることがわかる。しかし，これは，あくまで雑誌のものだから，新聞や単行本までふくめるとどうなるか，というのが知りたいところである。

国語研究所では，最近『話し言葉コーパス』『太陽コーパス』という2つのコーパスをつくった。これらは，ひとつひとつの音や単語をしらべるのには，ひじょうに有効だが，巨視的に日本語がどうなっていたかをみるわけにはいかない。『太陽コーパス』と雑誌の調査結果をみることにする（表2）。

表2 『太陽コーパス』と雑誌の比較

	寝台	ベッド	食卓	テーブル	汽車	電車	列車	幹線
1895（太陽）	4	−	7	2	120	50	26	5
1901（太陽）	17	−	19	7	116	15	143	3
1909（太陽）	16	1	4	24	135	173	38	4
1917（太陽）	16	−	20	5	86	66	27	1
1925（太陽）	23	21	21	22	93	51	47	6
・・・								
1956（九十種）	9	25	9	24	26	33	33	2
1994（70誌）	1	18	18	42	4	29	17	12

年によって，かなり変動があるが，「寝台→ベッド」「食卓→テーブル」という大勢にあることや，「汽車」が激減したことは分かる。「幹線」がふえたのは，もちろん「新幹線」のせいで，古い例は一般用語としての「幹線」である。『太陽コーパス』では，必要があれば，文脈も出せる。「右鐵道の幹線に於ても亦電氣發動機の採用を見るに至るや否やの一事あるのみ」（1895年9号「全國鐵道概覽」）。このように，いちいちの単語について近代100年の動きは分かるが，語種全体の動きは分からない。その点では，上にあげたような語彙調査の結果には，かなわないのである。

今めざしているコーパスで，国語研究所の2つの伝統，語彙調査があきらかにした巨視的な観点と，『現代語の助詞・助動詞』から用例カード・『太陽コーパス』にうけつがれた微視的な記述とが総合されることを希望したい。

7. コーパス利用とコーパス言語学

コーパスは有用だし必要だが，コーパスと言語研究との関係を問題にする言語学の1分野としての「コーパス言語学」は，言語学全体からみれば周辺的なものである。また，単にコーパスを利用して言語現象をしらべた，という研究を「コーパス言語学」とよぶ必要はない。コーパスの第1の価値は，膨大な用例の量にある。これからの研究には，当然それを利用すべきだが，それは「コーパス言語学」でも「用例言語学」でもない。大量の例文をしらべることは，まさに言語学の王道，限定なしのザ・言語学である。

といっても，むやみに量だけふやせばいいわけではない。コーパスは何がいえるかをおしえるが，何がいえないかはおしえない。コーパスにたまたま非文がふくまれていても，それが非文であることの判定は利用者にまかされる。とくに，インターネットを使うと，かんたんに膨大な用例がえられる反面，そのただしさについては，注意が必要である（田野村 2000）。動詞「ある」の否定は「あらない」ではなくて「ない」である。江戸時代初期の聞き書き『おあむ物語』には「くびもこはいものではあらない」という表現がでてくるが，この1例を当時否定形「あらない」が一般に使われていた証拠にすることはむずかしい。ところが，Google で「あらない」を引いてみたら，なんと47,100件あった。そのなかには，なぜ「あらない」がないか，というようなメ

タ言語や，「消していない部分も修正しなきゃあらないのでがんばります」という誤記，「荒内(あらない)」のような人名もあったが，

「私は決してHCLがとても普及している!!　ことをわかったあらないことは!」

「非常に印象的!私は決してその質のオーロラの長さを見たあらないことは。」

「他の人々上の権限があったあらないことは考える誰かにであり。」

など，英語からの機械翻訳かと思われる，なんとも訳の分からないものが多数あった[4]。ウエブの情報をチェックするためにも，やはり，きちんとしたコーパスを作っておく意味がある。

現実のコーパスは一定量のものだが，一定の範囲で調査する，という必要はない。1億語のコーパスを利用したら，そこでやめずに，例文をふやしたらいい。そのさい，分野・文体のバランスがくずれないように注意する必要はあるが，それはコーパスをつかわない例文採集でもおなじことである。また，コーパス利用を計量的なものにかぎる必要もない。「～らしい」と「～ようだ」のちがいをしらべるのに，計量的にしらべるのは，均衡コーパスの長所をいかすことだが，使用例数をかぞえなくても，大量の用例をしらべることでコーパスは活用されている。

コーパスは道具である。それは言語観・研究法のちがいにかかわらず，あらゆる言語研究者にとって役にたつはずのものである。ただし，話し手の直観を基準にすればいい，とする生成文法の立場とは両立しにくいかもしれない。チョムスキーは，コーパスにたいして，はっきり否定的な意見をのべた。コーパスは現にある（あった）用例だけをとりあげるが，文法はありうべき用例まで問題にしなければならないからである。ただし，これは，まだ大規模なコーパスが現実的なものになっていなかった時期の発言である。その後，コーパスおよびそれにもとづく研究が飛躍的に発展した現在にあっては，その効用を頭から否定するのはむずかしいだろう。生成文法家にとっても，コーパスによって得た例文を話し手の直観で吟味して使えばいいわけだから，利用価値がないことはないはずだ。げんに，「生成文法を学ぶ人のために」という副題のついた『言語研究入門』という本のなかにも，「コーパス言語学」という章がある（園田 2002)。

コーパスを無視するのも，絶対視するのもまちがいである。Fillmore(1992：35)がいうように，思弁的な言語学者とコーパス言語学者とは協力しなければならず，ひとりが両者をかねるのがのぞましい。

将来，今からは想像もつかないような大規模なコーパスが使われると，言語学の理論にとっても根本的な問題をなげかけるかもしれない。その1つは，文法的な文と非文法的な文との境目があいまいになることである。「あらない」がまちがいだということと，それがほとんど出てこない，ということとのあいだには，実質的な差があるのだろうか。ゼロと無限小との差のようなものである。langueとparole，competenceとperformanceとは程度の差にすぎないのではないか，ということである。もう1つは，実際の文には同じ形の句がくりかえし現れるという事実の発見である。言語使用者は今までに使われたことのある表現を使い回ししているのであって，言語はチョムスキーのいうほど生成的ではない，という（赤野 2004：11)。しかし，だからといってこの発見が語彙と文法との差を否定するというのは言いすぎだろう。差があることと連続的であることとは，むじゅんしない（Schönefeld 1999：151-152)。ある日の昼と夜の境目を何時何分何

秒まで厳密にきめたとしても，昼と夜とが瞬間的にかわるわけではないし，その境目が１本の線ではきめにくいからといって昼と夜の差を否定するのはナンセンスである。

注

1　中村通夫「明治初年の東京語研究」。中村氏は卒業後大学や旧制高校の教師にならずに文部省にはいり，国語研究所の初期に話しことば研究室長をつとめた。わたしは，この話を漠然と先輩からきいたような気がしていたが，じつは鈴木重幸氏とふたりで中村氏を訪問して学生時代の話をきいたことがあり，そのときに本人から直接おそわったことかもしれない。なお，文学との境界領域での卒論には，山本正秀「明治小説文章発達史―言文一致を焦点として―」(1933)がある。

2　ちなみに，幼児語は古代語とならんで内省による調査のできない研究分野である。古代人の言語感覚を直接知る方法がないのと同様に，幼児にむかって「ウマウマ」は名詞か動詞かをきくわけにはいかない。McEnery & Wilson(1996：11)参照。

3　雑誌九十種の語数については，53万語とされることがある。これは，不正確で誤解を生じさせる。国研報告25の「調査のデータ概略」によれば，助詞・助動詞以外438,135，助詞・助動詞94,642だから，これらを合計すると532,777になり，概算53万語になる（この数字は，その後修正されたが，大きな変化はないので，「概略」の数字による。なお，以下，「助詞・助動詞以外」を自立語，「助詞・助動詞」を付属語とよぶことにする)。ところが，おなじ「概略」にある母集団の推定値は，自立語１億語，付属語5,600万語である。これによれば，母集団での比率は自立語１に対して付属語0.560であるのに，標本では自立語１に対して付属語0.216で，母集団の半分以下になる。どうしてこのようなくいちがいがおこるのか。おもな理由は，調査の３分の１の段階で付属語の調査をうちきったことによる（国研報告21，p.296参照)。付属語は異なり語数がすくなく，途中までしらべれば，それで概略がわかるからである。したがって標本数が合計53万だったことは正しいが，そこには異質なものが混在しているのである。外国や日本のほかの語彙調査の規模と比較するためにも，付属語の範囲をふくめずに，概算44万語というのがいいと思う。

4　田野村忠温氏の教示によれば，「Google を含む多数のサイトで翻訳に利用されているというSYSTRAN 社の翻訳エンジンがどうやら犯人のようです」とのことである。

参考文献

赤野一郎(2004)「語彙研究とコーパス」『英語青年』149(11)，研究社

G. アシュトン・L. バーナード(2004)北村裕監訳『The BNC Handbook ／コーパス言語学への誘い』松柏社

園田勝英(2002)「コーパス言語学」大津由紀雄ほか編『言語研究入門』，264-275，研究社

田野村忠温(2000)「電子メディアで用例を探す―インターネットの場合」『日本語学』19(6)，25-34，明治書院

宮島達夫・小沼悦(1994)「言語研究におけるシソーラスの利用」宮島達夫『語彙論研究』，539-568，むぎ書房

Fillmore, C. J. (1992) “Corpus linguistics” or “Computer-aided armchair linguistics”, In J. Svartvik (ed.) *Directions in corpus linguistics, Proceedings of Nobel Symposium 82*, 35-60, Berlin/New York:

Mouton de Gruyter.
Francis, W. N.(1992)Language corpora B. C., In J. Svartvik(ed.) *Directions in corpus linguistics, Proceedings of Nobel Symposium 82*, 17-32, Berlin/New York: Mouton de Gruyter.
Hofland, K. & S. Johansson(1982) *Word frequencies in British and American English*, Bergen: The Norwegian Computing Centre for Humanities.
Juilland, A. & E. Chang-Rodríguez(1964) *Frequency dictionary of Spanish words*, The Hague: Mouton.
Juilland, A., D. Brodin, & C. Davidovich(1970) *Frequency dictionary of French words*, The Hague: Mouton.
Kaeding, F. W. ed.(1898) *Häufigkeitswörterbuch der deutschen Sprache*, Berlin: Steglitz.
Kennedy, G.(1998) *An introduction to corpus linguistics*, Essex: Longman.
Leech, G., P. Rayson, & A. Wilson(2001) *Word frequencies in written and spoken English*, Harlow: Pearson Education.
McEnery, T. & A.Wilson(1996) *Corpus linguistics*, Edinburgh: Edinburgh University Press.
Schönefeld, D.(1999) Corpus linguistics and cognitivism, *International Journal of Corpus Linguistics* 4(1), 131-171.

付　録

国立国語研究所報告書類
No.3　現代語の助詞・助動詞 —用法と実例—(1951)
資料集 2　語彙調査 —現代新聞用語の一例—(1952)
No.4　婦人雑誌の用語 —現代語の語彙調査—(1953)
No.8　談話語の実態(1955)
No.12　総合雑誌の用語 —現代語の語彙調査—前編(1957)
No.13　総合雑誌の用語 —現代語の語彙調査—後編(1958)
No.15　明治初期の新聞の用語(1959)
No.18　話しことばの文型 —対話資料による研究—(1960)
No.21　現代雑誌九十種の用語用字 第 1 分冊総記・語彙表(1962)
No.22　現代雑誌九十種の用語用字 第 2 分冊漢字表(1963)
No.23　話しことばの文型 —独話資料による研究—(1963)
No.25　現代雑誌九十種の用語用字 第 3 分冊分析(1964)
資料集 6　分類語彙表(1964)
No.37　電子計算機による新聞の語彙調査 I (1970)
No.38　電子計算機による新聞の語彙調査 II (1971)
No.42　電子計算機による新聞の語彙調査 III (1972)
No.43　動詞の意味・用法の記述的研究(1972)
No.44　形容詞の意味・用法の記述的研究(1972)
No.48　電子計算機による新聞の語彙調査 IV (1973)
No.56　現代新聞の漢字(1973)
研究部資料幼児のことば資料(1)(1981)
No.76　高校教科書の語彙調査(1983)

No.78　日本語教育のための基本語彙調査(1984)
No.81　高校教科書の語彙調査Ⅱ(1984)
No.82　現代日本語動詞のアスペクトとテンス(1985)
No.87　中学校教科書の語彙調査(1986)
No.89　雑誌用語の変遷(1987)
No.91　中学校教科書の語彙調査Ⅱ(1987)
No.99　高校・中学校教科書の語彙調査 分析編(1989)
No.112　テレビ放送の語彙調査Ⅰ(1995)
No.114　テレビ放送の語彙調査Ⅱ(1997)
国定読本用語総覧12総集編(1997)
No.115　テレビ放送の語彙調査Ⅲ(1999)
資料集14 分類語彙表 —増補改訂版—(2003)
日本語話し言葉コーパス(2004)
No.121　現代雑誌の語彙調査(2005)
No.122　雑誌『太陽』による確立期現代語の研究 (太陽コーパス)(2005)
No.125　現代雑誌の表記(2006)

(投稿受理日：2007年5月31日)

宮島　達夫 (みやじま　たつお)
国立国語研究所名誉所員
190-8561　東京都立川市緑町10-2
miya-tt@nifty.ne.jp

From vocabulary statistics to corpus-based studies

MIYAJIMA Tatsuo
Emeritus Staff, The National Institute for Japanese Language

Keywords

The National Institute for Japanese Language, citation slips, basic vocabulary, generative grammar, *Taiyo Corpus*

Abstract

Since its establishment, the National Institute for Japanese Language (NIJLA) has conducted statistical studies of Japanese vocabulary with a number of large-scale surveys on a variety of data, such as newspapers, magazines, textbooks, TV programs and so on. Although advanced in statistical processing of data, the surveys did not achieve the representativeness or the scale manifested in recent surveys conducted in English-speaking countries. NIJLA, however, pioneered descriptive studies based on large-scale data of modern Japanese and is now compiling a hundred-million-word corpus which is expected to further enhance the studies based on vocabulary surveys and empirical description.