代表性を有する現代日本語書き言葉コーパスの設計

山崎誠，前川喜久雄，田中牧郎，小椋秀樹，柏野和佳子
小磯花絵，間淵洋子，丸山岳彦，山口昌也
秋元祐哉，稲益佐知子，吉田谷幸宏

独立行政法人 国立国語研究所
E-mail:{yamazaki,kikuo,mtanaka,ogura,waka,koiso,mabuchi,maruyama,masaya,
akimoto,inamasu,daya}@kokken.go.jp

1　なぜ現代語のコーパスが必要か

1.1　国語研究所の書き言葉実態調査の問題点

国立国語研究所では1948年の創立以来，日本語の実態調査を書き言葉・話し言葉の両面から行ってきた。書き言葉については，不特定多数の人が接する媒体を中心に複数の語彙調査や実例に基づく意味用法の記述[1]を行ってきた。これらの調査は，その結果を通して戦後の言語政策への貢献を果たしただけでなく，学術的には現代日本語研究の基盤形成の一端を担ったと言ってもよいだろう。

国語研究所の書き言葉の実態調査は，方法論的にも早くから現在のコーパス言語学的な考え方に近いアプローチをとっており，その意味で先駆的なものであったと言えるが，現代的な観点で評価すると，データに関して次のような問題点が指摘できる。

(1)データの選択について

一回の調査で扱う言語資料が単一の媒体だけである。例えば，雑誌，新聞，教科書などについてそれぞれ別個に調査をしているが，複数の媒体を同時に調査したことはない。

(2)データの総量について

全体の言語量が少ない。これまで最大の調査は，新聞3紙1年分から1/60の抽出比で調査したもので，延べ語数は300万（短単位）である。

(3)個々のデータの長さについて

ランダムサンプリングで抽出されたひとつの抽出箇所はたかだか100文字前後の大きさしかなく，文脈を考慮した分析には不十分である。

(4)データの提供について

著作権処理を行っていないため，原データを公開することができない。学界の共有財産として活用される機会を失っているばかりでなく，分析結果の再現性が保証されない。

1.2　現代語研究におけるデータの役割

過去の言語と違って，現代語のデータは至る所にあり，入手も容易であるせいか，どのようなデータをどれくらい扱えばよいかという資料論が本格的に論じられていない。現代語研究の中核を占める文法研究では，内省によるデータ生成が主流である。少し前のものであるが，図1は，1990年代の現代語文法研究の論文において，データ（用例）をどのように獲得しているかを調べたものである[2]。

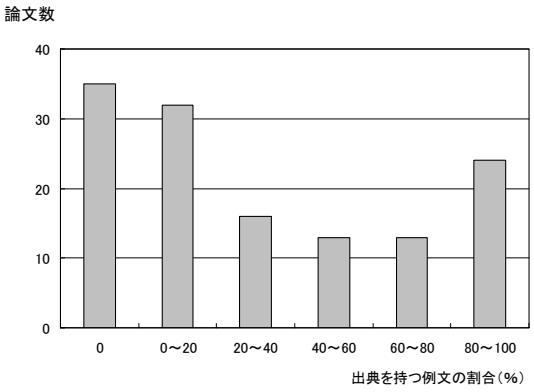

図1 実際のデータから用例を引いている割合

図1によると，実例を全く用いない内省だけのタイプがいちばん多く(0%の級)，その次が0～20%，その次が80～100%である。この分布から，作例をもっぱらとする内省タイプの論文が主流であるが，その一方で実例をもっぱらとする論文も存在していることが分かる（山崎(2000)）。

[1] 『現代語の助詞・助動詞』(1951)，『動詞の意味・用法の記述的研究』(1972)，『形容詞の意味・用法の記述的研究』(1972)などを指す。

[2] 1991年～1998年に発表された現代語文法に関する論文から1/10の無作為抽出による調査。用例に出典が記載されていれば，実例の引用，出典が無ければ作例（内省によるデータ生成）とみなした。

内省タイプの研究は，定性的分析を志向しているものがほとんどであるが，実例を引用するタイプの研究が定量的分析を志向しているかというと必ずしもそうではなく，いわば存在証明としてデータを利用するだけのものが多いようである。すなわち，現代語のデータはまだじゅうぶんに活用されているとは言い難い。とくに定量的な分析に基づく記述やモデルの構築が手薄である。

しかし，このような状況は，研究者をとりまくデータ環境が整備されることによって変わってくる可能性がある。早くからコーパス言語学の可能性に着目し，実践してきた田野村は「日本語コーパスの現状と課題」として，次の４つの点を指摘している（田野村(2000)）。

(A) 日本語研究のために作られたコーパスがない。

(B) 多様なジャンルのテキストをブレンドしたコーパスがない。

(C) 話しことばのコーパスがない。

(D) 文法情報付きのコーパスがない。

このうち，(C) の話し言葉のコーパスについては，国立国語研究所が情報通信研究機構（旧通信総合研究所）・東京工業大学と共同開発した『日本語話し言葉コーパス(CSJ)』が完成したことで解決された。（同時に，(A) (D) も実現したと言ってよい）。残る課題は(B) ということになる。

２　書き言葉コーパスの基本方針

第１節で紹介したような経緯を踏まえ，国立国語研究所では，現代日本語書き言葉コーパスの構築を研究開発の柱に据え，2006 年度から５か年計画で実施することになった。この計画の基本理念は，以下のとおりである。

①現代日本語の縮図となるコーパス

これまで研究所が行ってきた語彙調査の手法を生かし，コーパスがその母集団の統計的な縮図になるよう設計する。それにより，母集団における言語的諸特性の分布が縮図において過不足なく再現でき，母集団における分布を高い精度で推測できるようになる。

②汎用的な目的に供するコーパス

言語研究（語彙・文法・文字）以外にも，応用面として，辞書編集や言語政策，日本語教育などでも使えることを意図し，多様な日本語の姿を捉えることができるよう設計する。また，言語変化に対応するためには，同じ設計のコーパスを繰り返し構築するなど定点観測的な工夫も必要である。

③公開可能なコーパス

収録する著作物の利用許諾を得て，公開を目指す。インターネット上からの簡易検索のほか，共起条件を指定できる検索ツール等もあわせて提供する。

④既存のコーパスとの調和

解析単位の仕様を『CSJ』に合わせ，短単位，長単位の２種類の解析を行う。

３　書き言葉の広がり

母集団からの統計的縮図としてのコーパスを作成するためには，まずその母集団となる書き言葉の全体像を把握しなければならない。このためには，書き言葉の有りようについての整理が必要になる。

3.1　言語生活的観点から見た書き言葉

書き言葉に対して私たちの取りうる行動は「書く」「読む」のどちらかである。一般人を想定すると，新聞・書籍・雑誌などはもっぱら読むだけのものであり，メールやメモなどは読むことも書くこともある。書き手の数は通常一人であるが，読み手は一人とは限らない。これを図２のように整理してみる。

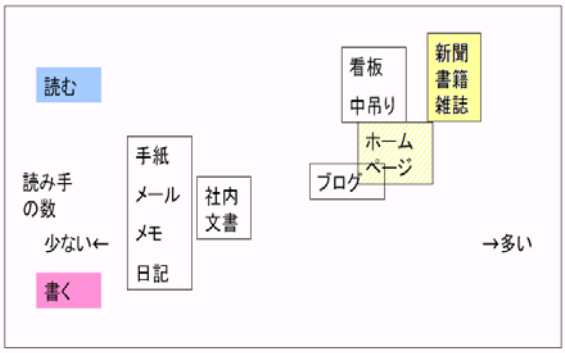

図２　「読み・書き」と読み手の数で捉えた書き言葉の分布

図 2 から左下の特定個人向けの私的な伝達にかかわる文章と右上の不特定多数に対する公共的な伝達にかかわる文章が 2 つのグループを形成することが分かる。今回のコーパスは，母集団を確定した上でその縮図を作る方法をとるため，私的な伝達にかかわる文章については収録の対象としない。私的な書き言葉の母集団を決定することが困難だからである。逆に，公共的な書き言葉の多くは，既存の目録やリストが整っていて，継続的に母集団を把握することができる。今回のコーパスの収録対象としては，新聞・雑誌・書籍に代表される「公共的な書き言葉」が目的に合致している。

3.2　伝達過程から見た書き言葉

公共的な書き言葉は，一方で，市場に出回る商品としての性質を持っている。具体的なモノとして書き手により生産され，必要なだけ複製されて，雑誌や書籍であれ

ば，書店・コンビニ・図書館などに在庫として置かれ，読み手の興味を引けば，読み手の手に渡ることになる。新聞は，宅配と駅売りとでスタイルが異なるが，駅売りを考えた場合，書き手と読み手との間に市場にとどまる段階を考えることができる。

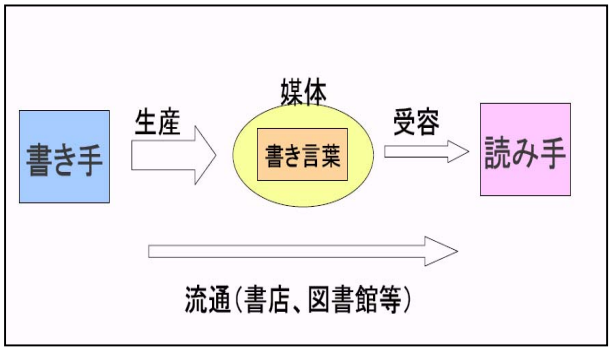

図３ 書き言葉の伝達過程

図３の各段階（生産，流通，受容）のいずれの段階も「書き言葉の実態」としての意味を持つ。したがって，この３種類の実態に即したそれぞれの母集団に基づくコーパスを作ることが望ましいと言える。この３つの段階は，書き手によって生産された書き言葉が，読み手に届くまでに淘汰されるプロセスを表している。

3.3 情報流通センサス

経済産業省が毎年実施している「情報流通センサス」においても情報の発生（提供）と消費との間にいくつかの段階を設けて情報量を測定している。例えば，平成15年度版の報告書では，2003年の１年間における新聞，雑誌，書籍，図書館に関する情報量は表１のようになる。

表１ 媒体ごとの情報量

	原発信情報量[3]	発信情報量[4]	消費情報量[5]
新聞	4.73E+09	2.31E+15	1.57E+14
雑誌	4.74E+09	4.26E+14	8.43E+13
書籍	4.25E+09	8.59E+13	3.27E+13
図書館	2.15E+13	4.30E+13	2.46E+13

(単位：ワード)

※漢字仮名交じりの場合，１ワード＝約3.3文字。

[3]各メディアを通じて流通した情報量のうち，当該メディアとしての複製や繰り返しを除いたオリジナルな部分の情報の総量。
[4]各メディアの情報発信者が，１年間に送り出した情報の総量。複製を行って発信した場合及び同一の情報を繰り返し発信した場合も含む。
[5]各メディアを通じて，１年間に情報の消費者が実際に受け取り，消費した情報の総量。

表１によると，新聞・雑誌・書籍の原発信情報量の比は，ほぼ１：１：１であるが，図書館はそれらの１万倍の情報を発信していることになる。しかし，発信情報量では新聞が群を抜いてトップに立ち，消費情報量でも雑誌との差は縮まるものの新聞が一番読まれている媒体であることが分かる。

４　２種類のサブコーパス

書き言葉の実態の捉え方の違いで３つの母集団が考えられるが，このうち受容に関する部分は人の選好がもっとも関係するところであり，入念な調査を実施しないとその実態把握が難しい[6]。そこで，今回のコーパスでは，生産実態を反映したサブコーパスと流通実態を反映したサブコーパスの２つを構築することにした。

4.1　生産実態を捉えたコーパス

書き言葉を生み出す書き手の立場を重視したもので，売れ行きや知名度に関係なく，どのタイトルも同じ確率で選ばれる。後述の流通実態をとらえたコーパスに比べると，言語的属性の多様性が確保される。また，生産された時点を特定することで，使用実態の時間的断面を浮き彫りにすることができ，経年調査を実施するのに向いている。このコーパスの予想される問題点は，極端に専門的な語彙が入ってしまうことである。例えば，日本書籍総目録2003年版によると，2002年に発行された書籍のうち異なりで約２割が専門書である。すなわち，生産実態を捉えたコーパスの中にもほぼ同様の確率で専門的な文章が入り込むということである。

4.2　流通実態を捉えたコーパス

書き言葉を流通させている市場の立場から捉えたもので，書店の在庫や図書館の所蔵など広い意味で社会の需要を反映している書き言葉とも言える。極端に専門的な文章が排除されることによって，より一般的な用語用字を調べるのに適している。また，資料年代にある程度の時間的な幅が生まれる（長期間流通段階にとどまる資料がある）ため，通時的な観察が可能になる。このコーパスの予想される問題点は，各メディア（書籍，雑誌，新聞）により流通のあり方が異なるため，獲得できる資料年代が不統一になることである。

なお，生産実態及び流通実態を具体的に把握する方法については，丸山(2006a)に詳しい。

4.3　２種類のサブコーパスの仕様

[6] ベストセラーのリストや図書館でよく貸し出された本のリストなど，受容の実態を反映したデータの活用が考えられる。

2つのコーパスの特徴を生かすために,それぞれに次のような仕様を与える。

表2 2種類のサブコーパスの仕様

	生産実態Corpus	流通実態Corpus
サンプル長	記号を除く 1000字	ひとまとまりの文章(=記事)
資料年代	2001年～2005年	1975年～2005年
語数	1000万短単位	1億短単位

サンプル長については，流通実態コーパスでは，段落を超えるひとまとまりの文章（これを記事と呼ぶ）で1万字を超えないもの抽出単位とする。従って，サンプル長は記事により長さが変わる。

資料年代については，生産実態コーパスでは，短期間を対象にして，繰り返し構築できるようにするため，語数を1000万語程度に抑えてある。一方，流通実態コーパスの方は約30年間の変化を観察できるようにするが，年代による資料の分布が一様でないため，生産実態に近い分布になってしまうという問題がある。

いずれのサブコーパスも母集団の量的構成比を推定により求め，それを反映させた縮図を作る（各媒体，各ジャンルごとに構成比を算出する）という手順を経る。

5 サンプリングの設計

5.1 コーパスへの収録条件

既存の目録等はそのままでは母集団にならない。コーパスの基本方針に照らして除外すべきデータがあるからである。実際にサンプリングを行う際にも，抽出箇所が除外対象であれば，そこはスキップすることになる。

除外されるのは，次のような部分である。

(1)文字以外の手段が主体となる表現（写真，絵，楽譜，地図，漫画，図表など）

(2)日本語以外で表現されているもの（外国語）。

(3)文字であっても,文章になっていないもの(語の羅列,名簿，索引など)

(4)現代語ではないもの（資料の刊行年や著者の生年を基準に判定)

(5)本文に含めない部分（広告，目次，奥付，前書き，解説など)

(6)その他

著作権処理が複雑なもの（問題集など)。

5.2 サンプル箇所の指定

母集団をジャンル等によって各層に分け，その中で一定の抽出比に従い，ランダムに抽出箇所を決めていく。具体的なサンプリング手法については，丸山他(2006b)の発表を参照されたい。

6 研究用付加情報

書き言葉コーパスでは,『CSJ』及び『太陽コーパス』(2005)で培われたタグ付けの経験を生かし，文書構造が的確に再現されるようタグセットを用意した。これらはルビ・外字等の文字情報などとともにXMLにより記述される。サンプリング時に得られる書誌情報や著者に関する情報も合わせて参照できるようにしている。

タグ及び電子化フォーマットの詳細については，間淵(2006)を参照されたい。

7 形態論情報

『日本語話し言葉コーパス』での単位設計に合わせて短単位，長単位の2種類の言語単位に基づき，形態論情報（代表形，代表表記，品詞，語種など）を付与する。また，品詞体系については『CSJ』で採用された体系を見直しつつ統一を図る。

8 終わりに

本稿では，向こう5年間で構築する現代日本語書き言葉コーパスの設計方針について説明した。

現代日本語の縮図（あるいは日本語の平均像）という考え方には，懐疑的な意見もあるが，この縮図はその代表性ゆえに日本語の特性を計る基準データとしての役割が期待できる。日本語の諸特性がこの縮図の中でどのような分布を示すのか，様々な事例を観察することによって，新たな記述研究への道が開けるのではないだろうか。

参考文献

国立国語研究所(2005)『太陽コーパス　雑誌『太陽』日本語データベース』博文館新社

田野村忠温(2000)「用例に基づく日本語研究－コーパス言語学－」『日本語学』19-5，明治書院

間淵洋子他(2006)「代表性を有する書き言葉コーパスの電子化フォーマットについて」本予稿集所収

丸山岳彦他(2006a)「現代日本語の書き言葉に関する生産実態と流通実態 ―代表性を有する書き言葉コーパスのための基礎調査―」本予稿集所収

丸山岳彦他(2006b)「代表性を有する書き言葉コーパスのサンプリング手法について」本予稿集所収

山崎誠(2000)「文法研究と用例－実例と作例の割合－」『日本語学』19-6，明治書院