〈共同研究プロジェクト紹介〉通時コーパスの設計 日本語通時コーパスの設計について

著者	近藤 泰弘
雑誌名	国語研プロジェクトレビュー
巻	3
号	2
ページ	84-92
発行年	2012-10
URL	http://doi.org/10.15084/00000709

国語研プロジェクトレビュー
NINJAL Project Review　Vol.3 No.2 pp.84–92（October 2012）

日本語通時コーパスの設計について

Developing the Diachronic Corpus of Japanese

近藤 泰弘（KONDO Yasuhiro）

1. はじめに

日本語の歴史的研究は，本来，各種の言語データを用いて，それらから各種言語事実を導き出す研究手法が中心であった。なかでも，日本語の歴史的研究が本格的にそのような科学的な手法を取り入れるようになったのは江戸時代の本居宣長の研究に始まると言ってよい。宣長の研究は，語彙・文法・表記・音韻・日本漢字音など多岐に渡るが，いずれも，上代から平安時代までの各種文献を収集し，それらの用例をもとに，記述的な成果を積み上げたものである。

明治時代になり，欧米の近代言語学，特に歴史言語学の様々な理論が入ってきてからも，日本語史研究の主流は，宣長に発する国学系の研究態度を大きく変えることはなかった。山田孝雄の記述的古典語文法（『奈良朝文法史』『平安朝文法史』『平家物語の語法』），橋本進吉の上代特殊仮名遣いの研究など，いずれも長い日本語研究の流れの正統的な発展である。

どのような言語でもそうであるが，古典語研究は，すでにその言語の話者がいないため，いわゆる内省によって研究することがまったくできない。したがって，日本語史研究が資料の記述を中心とした研究になることには，本来，必然性があったわけである。日本語史研究のための資料を電子化して，簡単に検索や集計ができることは，この流れの中で必然的な要請である。その意味で，日本語の様々な文献の中で，言語研究のための各種のアノテーション（付加情報）を施し，世界に向けて最初に公開されたテキストが，現代語ではなく代表的古典語文学作品の『源氏物語』であったことは，偶然ではなかったと言える（近藤（2003）で詳述した）。

さて，このようなわけで，古典語研究のために，主要な古典語資料を集成し，ひとつのまとまった電子化資料（電子コーパス，以下「コーパス」と略す）とすることはきわめて重要な課題である。特に，以前とは異なり，生成文法の考え方による言語研究の方法が普遍化していることは，コーパスの必要性をより大きくしている。現在の日本語研究，特に文法研究では，ある文が「言える」「言えない」あるいは「文法的」「非文法的」であるという判断を行い，その判断によって，言語事象の理論化を行うということが多い。これは言うまでもなく，チョムスキーによって始められた生成文法的な手法であるが，日本語史研究の分野でもおおいに使われている。しかしながら，最初に書いたように「内省」が働かない古典語においての「文法的」「非文法的」の判断は，用例の有無によるしかない。多くの古典語作品の

中において特定の単語の連接や文型の有無を調査することはひじょうに手間がかかることであり，ここにコーパスの意義がある。

本プロジェクトでは，このような見地にたち，平安時代から江戸時代までの代表的な文学作品のうち，言語資料として適したものをとりあげコーパスとすることを試行している。本稿では，そのコーパスの概要を述べ，また，同時に古典語の通時コーパスの設計における必要な事項について記していきたいと思う。なお，本プロジェクト発足以前の日本語古典語コーパスの様子については，先にも述べた近藤（2003）や安永（1998）を参照されたい。なお，当面，今回のプロジェクトの目標としては，平安時代和文に限定した「試行版コーパス」というべきものを作成することとしている。

また，本稿では，プロジェクトの研究内容のうち

1. コーパスにどのような情報を付加するか，つまりコーパスアノテーションの問題
2. 形態素解析に関わる諸問題

の二つの問題を中心に述べていきたい。

次の節ではその前提として「通時コーパス」の概念について略述する。

2. 古典語コーパスか通時コーパスか

我々のプロジェクトは「通時コーパスの設計」ということで，「通時」ということをうたっている。また，その内容は主に古典の文学作品である。これについてまず説明しておきたい。

そもそも，通時的に見るということと，古典語を扱うことは必ずしも同じことではない。明治・大正・昭和・平成という時代区分にしたがって，現代語の形成を「通時的」に観察することも可能であるし，また，平安時代のうち，西暦 1000 年代に限定して資料を収集し，「共時的」なコーパスを作成することもまた可能である。しかしながら，現時点においては，「通時コーパス」というものが日本語においてまったく存在しないのであるから，まずその最初のモデルを提示するものとしては，特定の時代に限定されたコーパスよりも，古い時代から新しい時代までの古典語を広く概観することができるものであることが望ましいだろう。なお，現代語のコーパスと，古典語のコーパスと直接に接続した通時コーパスを作ることはいろいろな点で難しいし，すぐに必要でもないので，これは次の課題とすべきだろう。

ということで，現在作るべきものは，「古典語コーパス」であり，「通時コーパス」であり，日本語の「歴史的コーパス」でもあるということになる。古典語の共時的コーパスとしては，共同研究を行っているオクスフォード大学の上代日本語コーパス（http://vsarpj.orinst.ox.ac.uk/corpus/）がそうである。また，通時コーパスとしては，国文学研究資料館の大系本文データベース（http://base3.nijl.ac.jp/）がそれに近いものと言える。しかし，前者は古典語コーパスではあるが通時コーパスではない。また，後者は，形態素解析がまったくなされていないため，言語学的な研究対象としてはやや不十分である。その意味で，本プロジェクトが目指すコーパスを作成する意義がある。

3.　アノテーションの必要性

従来作られた日本語の古典語のコーパス的なものはいくつかあるが，それらは，多くの作品や資料を含んでいるが，翻字された原文がそのままテキスト化されているものか，単語に分割されて品詞情報が付けてはあるが，単一あるいは，少数のテキストのみに止まるかのいずれかであった。このいずれも，言語学的な目的でのコーパスとしては不十分である。

コーパスに付加される様々な情報を「アノテーション」と呼ぶが，このアノテーションにはいろいろな種類がある。原文の写本の丁数，翻字された活字本のページ数，物語の巻数，和歌の歌番号，作者，作品の成立年代など，いわゆる出典の情報を付けるのはまずは当然である。しかし，言語学的な目的を持ったコーパスとしては，形態論的な分析を施し，それに関するアノテーションを付加することは絶対条件である。単語に分割し，品詞情報，読みの情報，語彙素（lemma）の情報などを付けることであるが，これを情報処理の分野では，昔から「形態素解析」と称するので，本稿でも同様に呼ぶ。この形態素解析は，言語学で言う「形態素」に分けるのではなく，後述する「統語解析（係り受け解析）」の前段階で行う，形態論的な分析を総称するものである。この形態素解析によるアノテーションを付加することは，特に日本語のコーパスではひじょうに重要であるので，若干これについて述べておく。なお，語彙素については特に重要であるので，後の 5 節で詳述する。

そもそも欧米の言語のように，単語分かち書きの正書法を持つ言語では，もともとのテキストをそのまま入力するだけで，スペースで区切られることで形態素解析の第一段階は終了していると言える。もちろん，品詞情報などを付ける必要はあるが，漢字の読みなどの問題もないため，とりあえず，もともとのテキストそのままでもいろいろな研究目的に耐えうる。いずれにせよ，形態論的なアノテーション，あるいは最低，単語への分割がされているかどうかは，コーパスとして本質的な問題である。

そのことは，以下のように説明するとわかりやすい。単語分割されていないテキストデータでも電子的なものであれば，検索することは容易だ。特定の単語（単語に分割されていない場合には，正確には「文字列」）の全用例を容易に列挙できる。しかしながら，「テキスト全体の総単語数を求める」「結合しやすい二つの単語を頻度数で表示する」などの目的のためには，単語への分割がなされていないと不可能である。また，「動詞の一覧を出す」などのためには，品詞情報のアノテーションが必要となる。また，「「歩く」という動詞の活用形も含めてすべての例をあげる」には，「語彙素」のアノテーションが必要になる。このように，形態論的なアノテーションは，言語学的研究のためのコーパスとして必須なものであることがわかる。

理想的には統語論的なアノテーションによって，述語にかかる補語や副詞の状況，連体修飾などの状況もわかるとよいが，これは，文末の動詞や形容詞に近い位置の格助詞を指標にして調査することで，擬似的に統語論的な分析をすることはできる。

以上のように，形態論的なアノテーションは，コーパスにとって付加的な情報ではなく，本質的なものであると言えよう。形態論的なアノテーションは，日本語のコーパスとしては，すでに「現代日本語書き言葉均衡コーパス」（BCCWJ）など，国立国語研究所のこれまでに

作成したコーパスで行われているため（前川・山崎（2009）等参照），通時コーパスでもできるだけ，それらと互換性のある方法でのアノテーションを行うことが望ましいと考えている。

4. 形態論的アノテーションの自動化

従来から存在する，形態論的なアノテーションが施された古典語の代表的な電子テキストとしては，角川古典大観『源氏物語』（伊井春樹（1999））がある。これは，単語単位に分割し，品詞情報などを付加したものであり，言語学的な目的に耐える。これはおそらくすべて手作業で単語分割などをしたものと思われるが，『源氏物語』単一のものであるので手作業で済んだが，さらに収録範囲を大きくして，コーパスとしていく場合には，アノテーションを手作業だけで行うことはひじょうに困難である。

困難であるというのは，作業を遂行できないという意味だけではなく，統一することが困難であるというのが一番大きな要因である。人間による作業はどれだけ詳しい作業マニュアルを決めても，作業にあたる人間が異なると，その結果は異なってくる。大きな作業になればなるほど，分担作業となり，その結果，作業者ごとに異なった単位の分割や品詞付けを行ってしまう危険性が増大する。

しかし，形態論的アノテーションを自動化するソフトウェア（いわゆる形態素解析器）を用いるならば，かなりの高精度で，均一なアノテーションを施すことができ，それはどれだけコーパスが大きくなっても精度に変化がない。したがって，コーパスが小さいうちから，形態素解析器を用いて，アノテーションを行うことがかならず必要である。

現在のところ，現代日本語を対象として，自動形態素解析を行うことのできるソフトウェアで，研究用に自由に用いることができるのは次の四つである。

1. JUMAN（ジュマン）・京都大学黒橋・河原研究室
2. ChaSen（チャセン）・奈良先端科学技術大学院大学松本研究室
3. MeCab（メカブ）・京都大学情報学研究科−日本電信電話株式会社コミュニケーション科学基礎研究所 共同研究ユニットプロジェクト
4. KyTea（キューティー）・Graham Neubig，笹田鉄郎，森信介

1 と 2 と 3 は，形態素解析用の辞書を備えており，その辞書を形態論的な面で考慮したものに取り替えることで古典語に対応することができる。4 は，形態素解析済みのテキストをトレーニングデータとして与えることで解析モデルを形成する，機械学習を全面に出したタイプのものである。これらはいずれも古典語においても対応可能な設計になっているが，現在のところは，このうち MeCab を主に用い，そのための形態素解析専用辞書としては，プロジェクトメンバーの小木曽智信氏等の開発（小木曽他（2011）参照）になる「中古和文 UniDic」「近代文語 UniDic」を用いている。UniDic は現代語用にも開発されており，ChaSen にも適用可能であるが，ここでは古文専用の MeCab 用バージョンを使っている。現在，コー

パスの主たる対象としているのは，試行版コーパスの内容である平安時代の和文であるが，平安時代和文に MeCab ＋中古和文 UniDic1.2 を適用した場合，未知語がない環境では単位境界で 99.31 パーセントが正確に切れる。

この古文対応の UniDic は，今回の通時コーパスプロジェクトのキーテクノロジーであると言える。現在のところ，平安時代和文を中古和文 UniDic で，そして平安時代の漢文訓読系資料を近代文語 UniDic で処理することを試みているが，中世語や近世語については十分な成果を得ていない。さらに研究が必要なゆえんである。

5.　通時コーパスの機能について

今回の我々のコーパスに，通時コーパスとして必要な機能がどのように実装されているかについて，ここで述べてみたい。まずもっとも重要なことは，特定の資料＝文学作品の成立年代をアノテーションとして入れておくことである。これがあれば，同一の単語や文型（語連続）を，年代順に配列し，統計をとったりすることが可能になるわけである。年代のアノテーションが加えられていれば，たとえば，格助詞の「に」と「へ」の消長，係助詞の消滅，動詞の連体形と終止形の合一などをコーパス上で詳細に観察できるはずである。

次に，もうひとつ，通時コーパスとして重要なアノテーションについて述べておきたい。それは「語彙素」（lemma）である。共時的なコーパスでは，語彙素は，主に活用形の差異を吸収するために用いられる。具体的には，「走らない」「走ります」「走る。」「走る時」「走れば」等を一括して「走る（語彙素）」で検索したり，取り出したりすることができるため，文法研究・語彙研究に欠かせないアノテーションであると言える。また，語形のバリエーション（「蝿」についての，ハエやハイなど）を吸収することも可能である。通時的に見た場合には，この共時的に見た場合の語彙素のアノテーションの特性とともに，もうひとつのメリットがある。たとえば「落ちる」という一段動詞の場合，平安時代では「落つ」という終止形であり，活用も「落ちず」「落ちたり」「落つ」「落つる時」「落つれば」のように二段活用になっており現代語とは異なる。いわゆる二段動詞の一段化の現象による変化である。しかしながら，これを通常の語形検索で検索すると「落つ」と「落ちる」が異なる動詞となってしまい，これでは通時的研究が求める検索とは言えない。通常の文法史研究・語彙史研究のためには，「落つ」も「落ちる」も同様に検索・収集ができなくてはならない。それで，この場合も，両方の語形を通して，たとえば「落ちる（語彙素）」のアノテーションを付すことによって，通時的な観点を加味した上での検索が可能になるわけである。また，「歌ふ」と「歌う」のような音韻の変化および表記上の変化も，やはり「語彙素」として単一のものに吸収できるため，統一的に検索できる。

以上のようなことによって，どのようなタイプの研究が可能になるかについて，次に述べておこう。大きなエクセルの表のようなものを考える。横軸に日本語のすべての単語の一覧を，縦軸に時代順に言語資料を並べるとする。つまり，横が数万から数十万（単語の異なり語数），縦が数千（対象とする言語資料・文学作品の数）である。それぞれがクロスした表の欄には用例数を記載する。これによって，ある単語がある時代の作品にあるかないかが一

覧できる。また，そのクロスした部分に用例数だけでなく，用例へのリンクを置いておけば，具体的な用例も直ちに通覧することが可能になる。これはいわば「通時的日本語用例集」というものになる。このようなものを作る場合，横軸のすべての単語の一覧は，語彙素であることが必要である。しかも，先に述べた「落ちる」「落つ」のようなタイプの差も吸収できるような語彙素の設定が必要となる。

このような表を作ると意外な言語事実が見えてくる。たとえば「商い」（語彙素）という語の出現を調べると，平安時代では和文系の作品としては『竹取物語』だけに出現するものであり，その他は『大唐西域記長寛点』や『類聚名義抄』といった漢文訓読系のものだけである。『竹取物語』への出現もこの観点からすると，漢文訓読の語彙であろう。『源氏物語』『大鏡』など大部の作品にはまったく出てこない。中世になっても『蒙求抄』など訓読文にのみ出現する。そうではないタイプの資料に出てくるのは『狂言』を待たなくてはならない。もちろん，江戸時代の諸作品にはごく普通の語として用いられる。この傾向は動詞「商う」（語彙素）についても同様である。つまり，古くは漢文訓読（しかもほぼ漢籍に限られる）だけに用いられていた語なのであるが，それが資料的な制約により，一度，潜伏してしまった形となるが，後に漢文訓読語から一般の語へと拡張されたため，再度，表面に現れた語であるというようなことがわかるのである。

このようなタイプの表としては『古典対照語い表』（宮島（1971））が先駆的なものであるが，語数が 23000 語と少なく，また語彙素としては，活用形のみが集約されており，表記の差は入っていない。時代も『万葉集』から『徒然草』までの 14 作品にとどまっている。また『古典対照語い表』はそれまでに作成された語彙索引を改編して作成されたため，形態論的な単位の不統一を修正するために多大な労力がかかっている。先までに述べた形態論的アノテーションを統一的に行い，成立年代のアノテーションを付したコーパスからなら，すべて完全に自動的にこのような「通時的日本語用例集」を作成することが可能である。

6. 通時コーパス作成上における困難な点

古典語の通時コーパス作成の上での困難な点について述べてみたい。問題は，語彙・文法の面と，表記の面とでは異なる点が多いので，それぞれについて分割して記述していく。

6.1 語彙・文法

古典語資料は現代語ではないため内省が働かない。そのため，そもそもある語形がどういう意味であるかがはっきりしないことも多い。そのような場合は，形態素解析も不可能である。むしろコーパスを作成し，研究を行うことで，問題が解決するという順序であろう。また，日本の古典，特に和歌特有の問題として，「掛詞」がある。「掛詞」は同じ単語に二重の意味を持たせるため，同一の単語形態に二つの意味や属性があることになる。この場合，主たる意味だけを採用するか，あるいは従たる意味もなんらかのアノテーションを工夫して採用するかという問題が生じる。ひじょうに難しいところである。もし掛詞のアノテーションを採用することとなれば，単語の切れ目自体が異なることもあるため，単に単語内部に新た

な属性を付すだけでは対応できないなど，問題は複雑である。これらの問題に対してはまだ解決手段がない。今後，十分に研究が必要である。

6.2　表記

表記の上で問題となるのは，二つである。ひとつは，出典となったテキストの情報をどれだけ反映するかということである。こちらは，複数の層にわかれたテキストをどのようなアノテーションに反映させるかということになり，技術的な問題はあるものの方針さえ決定できればそれほどの問題はない。それに対して，複数の表記の層のうち，何をもっとも重要な基盤としてコーパスを作成するかという問題はさらに困難が多い。日本語の古典語の場合，特に平安時代資料の場合，原資料となる写本は，ほとんどが平仮名である。その平仮名文を，漢字仮名交じり文としたものがいわゆる校訂本文であるが，漢字仮名交じり文は，もとの写本の解釈であって，いわば，一種の「現代語訳」であるとも言える。したがって，校訂者によって種々の解釈が可能になり，多様な校訂本文が生じる。そのため，何を基本としてコーパスを作成していいのかがきわめて不安定な状態となる。表記形による検索を行う場合に，校訂本文のゆれは微妙な問題を引き起こす。近代語・現代語のように印刷された底本をもとにコーパスを作成できる場合とは異なっているのであり，この問題もひじょうに難しいが，解決が必要である。

7.　資料のサンプリングの問題

通時コーパスの現在の段階（試行版コーパス）では，特定資料からのサンプリングは一切していない。つまり，特定の文学作品を採用した場合には，その全文を使う，全文コーパスの方針で進めている。現代語日本語コーパスでは BCCWJ において，固定長および可変長のサンプリングが行われており，BCCWJ からはもとの文献を復元することはできないようになっている。これは，著作権処理を容易にし，また多種類の文献を制限ある容量に収めるために必要な処理であったと思われるが，古典語においては，次の理由から，サンプリングの必要性はないと考える。

1. 資料の全体量の少なさ
   小学館の古典全集などの代表的文学作品全集でも語数にして，せいぜい 500 万語程度の延べ語数しかないものであり，現代語に比べると圧倒的に少ない。一億語の BCCWJ と比較すればはるかに小さなものであり，コーパスの容量が大きすぎるという問題は存在しない。
2. 文学作品への配慮
   資料はいわゆる文学作品が多く，その一部だけを取り出した場合，文学研究者にとってまったく価値のないコーパスとなってしまう。これは避けたほうがいい。文学研究においてはある作品のすべてが存在することが絶対に必要である。
3. 文章法研究への対処

古典語の場合，書き言葉として残された資料が唯一のものである。したがって指示詞（特に文脈指示）の研究や，主題の研究，物語のテンス，話法など，文脈を考慮した研究が重要になってくる。文脈における旧情報，新情報の問題などの語用論的な研究も重要である。これらの場合，文章の一部を切り出した部分的な文章では，それらの研究はほとんど不可能となる。古典語研究の重要な分野を占めるこのような文章法あるいは語用論的な研究についての配慮をするとすれば，サンプリングをすることには慎重にならなくてはならない。

以上のような点から，ある作者の全作品のうち，とりあげない作品があるという形でのサンプリングはあるかもしれないが，一作品のうちの部分のみを取り出すという形のサンプリングは，通時コーパスには不適当であると考えられる。

8. 現状のまとめと展望

通時コーパスプロジェクトでは，2013 年度までに試行コーパスを完成させることにしている。具体的には，小学館新編日本古典文学全集のうち，代表的な平安時代和文を中心とした試行版コーパスを作り，公開にむけての研究を行う予定である。そこでは，コンコーダンサ（索引作成ソフトウェア）として，BCCWJ の検索ツールとして公開されている「中納言」を用いる。形態素解析が施され，作成年代・作品名・作者等の情報を加えたアノテーション付きの通時コーパスは，日本語研究にひじょうに大きな武器を加えることになると考えている。

その後，国語研究所の本格的な通時コーパスプロジェクトが実施されることになれば，さらに時代や資料の位相を広げて，同様の方針でコーパスを作成していくことが期待されるのである。

●参照文献●

伊井春樹（編）(1999)『CD−ROM 版角川古典大観「源氏物語」』東京：角川学芸出版.
近藤泰弘(2003)「古典語のコーパス」『日本語学』22(5): 62−81.
前川喜久雄・山崎誠(2009)「現代日本語書き言葉均衡コーパス」『国文学解釈と鑑賞』74(1): 15−25.
宮島達夫（編）(1971)『古典対照語い表』東京：笠間書院.
小木曽智信・小椋秀樹・近藤明日子・須永哲矢(2011)「形態素解析辞書「中古和文 UniDic」とその活用例（日本語学会 2010 年度秋季大会研究発表会発表要旨）」『日本語の研究』7(2): 104−105.
安永尚志(1998)『国文学研究とコンピュータ』東京：勉誠社.

《要旨》 国立国語研究所共同研究プロジェクト（基幹型）「通時コーパスの設計」では，日本語の史的研究に用いることができる本格的な「通時コーパス」を構築する準備段階として，コーパスの設計にかかわる諸問題について研究している。その中で，「選定した資料をどのように電子化しどのような情報（アノテーション）を付与するか」「古典テキストに対応した形態素解析等をどのように行うか」など，通時コーパス設計のための重要問題を中心に，基礎的な研究を展開している。

Abstract: In preparation for the development of the "Diachronic Corpus" to be built at the National Institute for Japanese Language and Linguistics, basic research is being conducted on diachronic corpus design. Based on typical materials from several periods ranging from ancient times to early modern times, an experimental model of the "Diachronic Corpus" will be created, and at the same time, research focused mainly on the following two points will lead to the actual construction of a partial corpus.

1. How classical texts are digitized, and what kinds of information (variant texts, text notations, variant characters, quotations, writing styles, etc.) are added
2. How morphological analysis appropriate for the vocabulary and the grammar of each period and writing style is conducted

近藤 泰弘（こんどう・やすひろ）

青山学院大学文学部教授。博士（文学）（名古屋大学）。東京大学助手，日本女子大学助教授，青山学院大学助教授を経て，1998 年 4 月より現職。

2009 年 10 月より国立国語研究所言語資源研究系客員教授。

主な著書・論文：『日本語記述文法の理論』（ひつじ書房，2000），『新訂日本語の歴史』（共編，日本放送出版協会，2005）.

受賞：第 4 回新村出記念財団研究助成金（新村出記念財団，1986），第 28 回金田一京助博士記念賞（金田一京助博士記念会，2000）.

社会活動：日本語学会理事，日本語文法学会評議員，東京大学国語国文学会評議員.

基幹型共同研究プロジェクト「通時コーパスの設計」

プロジェクトリーダー　近藤泰弘（青山学院大学　文学部　教授／国立国語研究所　言語資源研究系　客員教授）

プロジェクトの概要

国立国語研究所で構築する「通時コーパス」の開発に先立ち，通時コーパスを設計する基礎的な研究を行う。将来に作成される「通時コーパス」の原形となるべきコーパスを試作し，どのような問題点があるかを洗い出すことを目的としている。予定では，最終年度である 2013 年度に平安時代の和文を中心とした試行版コーパスを完成させ，研究者にも提供できることを目指している。また，そのコーパスの評価も行い，日本語史研究のためにどのような利用方法があるかについても研究を行っている。コーパスの概要については，次の web サイトにも記載があるので参照されたい。http://historicalcorpus.jp