歴史コーパスとは何か

近藤泰弘

１　コーパスの分類と歴史コーパス

　コーパス（corpus・複数形はコーポラcorpora）とは、語源的にはラテン語の「身体」という意味であるが（例えば、Corpus Christiとはキリストの体、またそれを記念する祝日の意味）、古くから文学や法律などの図書・典籍の「全集」「集成」の意味で用いられてきた。例えば、古代ギリシャの医学書の集成であるHippocratic Corpus（ヒポクラテス・コーパス）などが著名である。しかし、1950年代後半から言語学のための言語資料の集成の意味で用いられるようになる（OEDによる）。その時代の代表的な業績はRandolph Quirk等によるSurvey Corpusであり、このコーパスは紙のカードに書かれていた（１）。そして1964年に米国のブラウン大学で完成した世界初の電子化コーパスであるThe Standard Corpus of Present-Day Edited American English（通称Brown Corpus）以来、コーパスは、もっぱら電子化コーパスあるいは機械可読コーパスの意味で用いられるようになった。テキストコーパスと呼ばれることもあるが同じ意味である。前述のSurvey Corpusも後に電子化されLondon -Lund Corpusの一部となって現在に至っている。
　似たような意味でテキストデータベース(Text Database)あるいはフルテキストデータベース（Full-Text Database）という用語も用いられるが、これは、論文や図書の書誌情報や商品の売り上げ数値などの典型的なデータベースの対義語として用いられるものであり、たとえば特許文書の全文とか、法律の全文や新聞の紙面など、実用的にその全文をサーチして利用できるようなデータベースを指す用語である。したがって、単一あるいは複数の文学作品のテキストの集成の場合も、それをテキストデータベースと呼ぶこともあり得るのであり、その点でコーパスとの区別がつけにくい場合もある。例えば、シェークスピアの全作品の全文が電子化されて掲載されているウェブサイトが複数あるが、それをシェークスピアコーパスと呼んでいるサイトもあれば、シェークスピアデータベースと呼んでいるサイトもある。本稿では混乱を招かないように、単一の作品の電子化されたものから複数の作品の集成までを含めて、すべての電子化テキストやアーカイブを単純にコーパスと呼ぶこととしたい。
　ところで、コーパスにはいくつかの分類基準がある。単純に列挙していくと、次のようになる。

○ 位相（話し言葉・書き言葉）
○ アノテーションの種類（プレーンテクスト・品詞タグ・統語タグ）
○ 代表性（サンプリング・資料そのまま・レジスタの多様性）

○ 地域性（標準語・方言）
○ 時代（現代語・古代語・共時・通時）
○ 配布方法（オンライン・ダウンロード・DVD・CD）
○ ライセンス（オープンアクセス・無料登録・有料登録・商用）

これらについて若干の解説を施しておく。話し言葉と書き言葉との別や、地域性については特に問題はないものと思う。まずアノテーションについて一言しておこう。まず、出典情報などをコーパスに埋め込むことが第一に必要なことである。コーパスは、平文（plain text）そのものでも文字列検索によって必要な単語を検索できるが、同音異義語や動詞・名詞同形語などについては、検索結果が不十分なものになるため、品詞などの形態論情報を付け加えることが必須である。このような作業をアノテーションと称する。

出典情報は、初期のコーパスでは、プログラミング上の都合から、固定長形式（テキストの各行の開始部分に固定された桁数のテキストの番号や行数を記載する方法）で置かれたが、＜＞の形のタグを用いて、出典を示したCOCOA形式も広く用いられた。後で述べる『源氏物語』テキストデータベースもCOCOA形式によっている。形態論情報は、Brownコーパスの段階から、品詞を独自の規格でタグ付けすることで表記する方法が行われていた。また、その品詞タグ（part-of-speech (POS) tag）付けはコンピュータで自動化されていた。また、近年のものでは、係り受け関係を示した統語タグを付けたコーパス(Parsed Corpus)もかなり一般化している。

なお、出典情報も、形態論情報も、そのアノテーションの実際には、初期の独自タグやCOCOA形式のタグから、SGMLそしてXMLといったマークアップ言語によるものが一般化してきている。これはコーパスのアノテーションを統一化しようとするTEI(Text Encoding Initiative)の運動が、最初はSGML、現在はXMLをその記述言語として採用していることが大きい。従って、現在開発されている新しいコーパスはいずれもXMLによるマークアップを施されていることになっている。国立国語研究所の「日本語歴史コーパス」、オクスフォード大学の「上代日本語コーパス」（２）いずれもXMLによるマークアップである。なお、前者は現時点ではTEIには従っていないが、後者はTEIにも準拠している。

歴史的な面で述べれば、厳密には通時コーパスと歴史コーパスは異なるものであり、現代語の通時コーパスというものもあり得るものである。しかし、現実には、現代語（共時）コーパスと、歴史（通時）コーパスという分類があると考えて大過はない。英語コーパスにおける最初の体系的な歴史コーパスであり通時コーパスでもあるのはHelsinki Corpus of English Texts (Diachronic Part)であり、1991年に完成したものである。これはOld English、Middle English、Early Modern Englishの3つの時代区分から160万語を集成したものであって、基本的英語史の研究はこのコーパスでかなりの部分が実現できる。このHelsinkiコーパスは、一部が統語タグを加えられてThe Penn-Helsinki Parsed Corpus of Middle English, second edition、およびThe Penn-Helsinki Parsed Corpus of Early Modern Englishとなっている

（３）。
　以上のような過程を経て、現在に至っているわけであるが、ここからわかるように、ここまでの研究過程を経て、現在の歴史・通時コーパスに要求されている水準は、出典情報・形態論情報のアノテーションを、XMLによるマークアップで示すこと、さらには、統語情報までも加えられているとなおよいということである。このような経緯を知った上で、次には日本語における歴史コーパスの現状について述べてみたい。

２　日本における最初の歴史コーパス

　日本語における史的言語（古典語）を扱ったコーパスの最初となるのは、長瀬真理（東京女子大）等による『源氏物語テキスト・データベース』である（４）。1990年に公開されたものであるが、底本は、小学館の『日本古典文学全集』（旧版）である。このコーパスは『源氏物語』だけであったが、
１　当初からタグづけされておりCOCOA方式で出典情報などを入れてあったため、ただちにOCP(Oxford　Concordance Program)などで扱うことができる
2　英語（サイデンステッカー訳）とのパラレルコーパスである。
1　東京大学大型計算機センターで公開された他、オクスフォードテキストアーカイブにも登録され広く全世界に広がったなどの先進的な仕様を持っている。筆者も、後に千葉大学の土屋俊氏の研究室でこのデータベースを作る時に用いたOCR装置を見せていただいたことがあるが、この時は、千葉大学の加藤尚武・坂井昭宏氏が中心となって、実際の作成作業は千葉大学で行われたようだ。
　COCOA形式はこの当時に広く使われたタグによるマークアップの形式であるが、次のようなタグ形式である。
<W 紫式部> 著者名
<T 源氏物語> 書名
<C きりつぼ> 章名
<P 93> ページ数
COCOAは英国のAtlas研究所で1967年頃に開発された人文科学のためのテキスト処理プログラムの名称であり、word Count and Concordance Generator on Atlasの略称とされる。このCOCOAが扱うタグ形式なのでCOCOA形式（フォーマット）と呼ぶ。COCOAの後継としてオクスフォード大学で作られたOCPもCOCOAフォーマットを処理することができたため、オクスフォードテキストアーカイブにおいても標準的なタグ形式として使われた。
　OCPは大型計算機用のソフトウェアであったが、パソコンにも移植された。この『源氏物語』データベースが公開された当時に、日本においても、OCPの日本語化の作業が長瀬氏を中心に行われ、MS-DOS用（後にWindows用）のソフト

ウェアのMicro-OCP（日本語版）が沖田電子技研から発売された。OCPの使い方については、長瀬氏の『コンピューターによる文章解析入門―OCPへの招待』（５）に詳しく記載されている。
　いろいろな意味で、この『源氏物語』データベースは、日本における最初の本格的な古典語コーパスであったというに留まらず、現時点から見ても、一般的な底本を用い、標準化されたタグを備え、そのための日本語化された処理系を用意し、全世界からアクセスできる場所に無料で公開されたということでコーパスとしての完成度がきわめて高いものだったと言える。形態論情報などのアノテーションが付いていなかったのだけが、最新のコーパスには劣る部分であるが、それは時代的制約からやむをえなかったと言える。

３　「日本語歴史コーパス」の構想と仕様

　国立国語研究所の「日本語歴史コーパス」（６）については、国立国語研究所の共同研究プロジェクトの「通時コーパスの設計」で底本の選択や、当初に取りかかる作品名など、おおよその構想がなされ、データ形式の決定、形態素解析、データの補正、形態素解析辞書とコーパスの保守などの実装のすべてを、国立国語研究所のコーパス開発センターが行った（７）。
　日本語歴史コーパスをどのようなものとすべきかにはいろいろな考え方があった。写本や版本から本文を作成し、それによってコーパスを作成するという方法もあれば、広く流布している本文に依拠するという立場も考えられた。しかし、時間的な制約と、実際にできあがるコーパスの利便性を考えると、広く流布しているものによるべきという結論に達し、小学館の『新編日本古典文学全集』を主たる底本とすることになった。もっとも広く使われている古典文学全集であること、本文がネットで有料公開されており、本文や注釈をどこでも入手可能であること、また、小学館から本文データをXML形式（写植印刷用）で入手することが可能であったことなど、多くの好条件があった。
　一般的に日本語の古典コーパスの特色として考えられるのは、ひじょうに多くのジャンルにわたって、古い時代から多くの文献が存在していることである。Helsinkiコーパスと比較しても、Helsinkiコーパスの分類名で、Fiction(小説)　、Romance(ロマンス)、History(歴史)、Travelogue(旅行記)、Drama　Mystery(神秘劇)など、両者に共通するジャンルも多くある。また、共通するものの中でも、詩（poem）、日記（diary）などが日本古典の場合、特に重要なものとなるだろう。逆に、Helsinki　　コーパスに存在するが、日本語歴史コーパスに存在しないものには、Law(法)、Philosophy(哲学)、Science(科学)、Sermon(説教)などの分野がある。これらは、日本の古典では漢文で書かれるものに相当するため、日本古典文学全集の類には収載されていない。ここには、日本語の歴史コーパスに、日本で作成された漢文文献を入れるかどうかという大きな問題がある。岩波書店の『国書総目録』などは、純漢文文献や、貴族日記などの和化漢文も「国書」という枠で収載してあるわけだが、今回のコーパスでは、漢文文献は除外することとした。これは、日本漢文の場合

の形態論情報の付け方など技術的な問題が多いのであり、次の課題としたい。また、文学全集を底本にした時点で、仮名消息、法語、訓点資料などの非文学作品の日本語文献も対象外となってしまった。これについては、今後十分に計画していくべきものだろう。

日本語の歴史コーパスの一般論としては、Helsinkiコーパスなどと比較して、次のような特色が存在する。

１　古墳時代の金石文から近代に至るまで2000年近くの間に各種の大量のテキストが存在する
２　作者が明確なテキストがひじょうに古い時代から存在する
３　和歌のような文学作品の形態が古代から近代まで続いており、同一のジャンルの通時的なコーパスが可能
４　同一言語が同じ地域で現代まで使われている
５　多様な表記の様式があり、いずれの場合も分かち書きがされていない

これらの特色を生かすためにも、日本語歴史コーパスでは、もともとある印刷物の叢書（一種のコーパス）をもとに電子化することがもっとも簡単な方法であったと言える。

また、古典語のコーパスを現代語のコーパスと比較した場合には、その分量が圧倒的に少ないことが特徴だ。また、底本として古典文学全集を採用することとすると、その対象はすべて文学作品である。これによってもうひとつの方針が立てられる。それは、サンプリングの問題である。現代語コーパスはそのテキストの総量が多いため、何らかの形でサンプリングする必要がある。特に、ひとつの文学作品（たとえば小説など）であっても、BCCWJ（現代日本語書き言葉均衡コーパス）ではその一部ページだけを固定長あるいは可変長でサンプリングして用いている。したがってBCCWJからもとの作品を復元することは不可能である。しかし、古典語コーパスの場合は、例えば『源氏物語』の中で「桐壺」の一部だけが収載されているというようなことは許されないだろう。近現代語においても『青空文庫』のように全文が収載されたコーパスに特有の有用性は存在するが、古典語の場合は、必ず全文が収められている必要がある。もちろん、知られているすべての作品をすべて入れるわけにはいかないので、ひとつのジャンル（例えば、洒落本）の中の一部の作品を収めるという意味でのサンプリングは必要であるが、作品の一部を抜き出すという意味でのサンプリングは不必要である。今回の歴史コーパスでもそのような方針にしたがった。

最後に「日本語歴史コーパス」のタグ仕様について触れておきたい。当然のことながら形態論情報を含むために、そのアノテーションのためのマークアップを施さなくてはならない。これまでの大型のコーパスでは、形態論情報がないものとしては、前述の「源氏物語データベース」、『新編国歌大観』（角川書店）「大系本文データベース」（国文学研究資料館）「ジャパンナレッジ新日本古典文学全集」（小学館）「バージニア大学日本語イニシャティブ」「CDROM版源氏物語本文研究データベース」（勉誠出版）などがある。形態論情報があるものとしては『角川国文大観源氏物語』（角川書店）「オクスフォード大学上代日本語コーパスThe Oxford Corpus of Old Japanese」（オクスフォード大学）がある。日本語歴史コーパスでは、小木曽智信の開発した中古和文UniDicを用いて形態素解析を行い、短単位・長単位の形態

論情報をXMLで付してある。XMLのタグセットは独自であり、BCCWJとの互換性が高い。オクスフォード大学のコーパスもXMLでマークアップしてあるが、こちらはTEI準拠であり、現在のところ日本語歴史コーパスとの互換性はない。将来的に何らかの調整をはかる予定である。この他に今回の日本語歴史コーパスの仕様としては、文字コードとしてはUTF-8(ユニコード)を用いており、ただし漢字集合としてはJISの第1水準から第4水準まで（すなわちJISX0213）にとどめてある。また、付加的なアノテーションとしては、会話・和歌などの位相的な部分の明示、資料の成立年代、などが加えてある。

４　「日本語歴史コーパス」の展開

「新編日本古典文学全集」からの作品もまだ公開されていない部分も多い。現在のところは平安時代編ということで平安時代の作品の一部が公開されているわけだが、現在作業が進行中のものとしては『今昔物語』がある。また、全集以外でも重要と思われるものは収載される予定であり、当面『大蔵流狂言』（清文堂書店版）「洒落本」（洒落本大成）などの入力を進めている。いずれにせよ、形態素解析は自動化されているが、現代語とは異なり、自動化されたものをそのまま公開することは不可能である。それは絶対的なデータ量が少ないため、全部をなるべく正確な形で公表しなくてはならないからである。したがって、データ修正の手間・人員を考えるとおのずから一定時間にできる量には限りがある。徐々に進めていく他はないだろう。

筆者は、プロジェクト「通時コーパスの設計」の一員であり、今後もコーパスのあり方について考察していきたいと思っているが、個人的に計画しているのは、用例集および単語N-gramの作成である。公開されているコーパスは、中納言のインターフェースを通したKWICコンコーダンスであるが、コーパスの元データからは多様なデータが作成可能だ。

考えられるうちもっとも有用なものは、用例集である。KWICコンコーダンスは特定の単語をターゲットにしてその用例を集めることができるが、異なった種類の多くの単語の用例を通覧することはできない。しかし、辞書と同じように単語の用例を通覧することができれば、それは有用なデータとなるはずだ。具体的には、『古典対照語彙表』（笠間書院）に用例数だけでなく、用例そのものが付載されたようなものである。あるいは、「日本語歴史コーパス」の単語を用例とした「古語辞典」（ただし意味の記述などはない）と言い換えてもいい。コーパスから自動的に作る場合は、時代・用言の活用・その他を勘案しつつ平均的に用例を収集して入れることができる。それらをもとに古語辞典などの編纂に役立てることもできるだろう。

もうひとつは、単語N-gram集である。Googleなどが収集したサイトのテキストデータを形態素解析して単語に分解してからそのN-gramを収集し、研究用に公開している。それと同じように、古文を単語に分解したコーパスから、単語N-gramを作れば、いろいろな作品の対照研究に便利である。従来古典語のN-gramとしては文字N-gramの研究が多いが

（８）、単語N-gramも別のメリットはある（９）。文字N-gramでは掛詞など要素に分解しにくいものまで一致度を調査できるが、無意味な文字列も多いので無駄が多い。単語N-gramはその点すべてが存在する、意味のある単語列なので無駄が少ない。ただ、すべての資料について統一的な形態論情報がついた資料を基にしないと意味のないこととなるため、自前で形態素解析を統一的に行うか、歴史コーパスのようなものをもとにするしかない。

現代日本語書き言葉均衡コーパスでも、多くは「中納言」を通してコンコーダンスとして用いているが、もっと他の用途が開拓されてもよいし、そのような使い方によるデータも公開されてもいいと思う。

５　「日本語歴史コーパス」の活用

現状では、「古典文学全集」を用いているため、本文は、歴史的仮名遣い・漢字仮名交じり文に校訂されている。したがって、表記や音韻の研究には不適当であり、語彙あるいは文法の研究に適している。稿者も文法研究として、動詞のアスペクトの研究（１０）などを、コーパスを使って行って発表してきている。その際にもっとも有効なのは、「中納言」で作成したKWICコンコーダンスデータをエクセルでクロス集計表とする方法である。つまり、前接語や後接語との接続の数をクロス集計することで、統語的な秩序を観察することになる。これはコーパス以前から行われていることであるが、「日本語歴史コーパス」を用いると、すべての助動詞とすべての接続助詞との相互承接というような、手作業では絶対にできないレベルのクロス集計が簡単にできてしまうので、まったく次元の違う研究が可能になる。

ただ、その解釈はコンピュータではできず、文法理論との兼ね合いで、研究者自身がなすべきものであることは言うまでもない。一例をあげれば、中古語のモダリティ研究で「まほし」を扱った（１１）。「まほし」はモダリティの中ではもっとも外側にあるもので、三人称者の自らの意志を示すことができる。「まほしがる」という動詞形、「まほしさ」という名詞形、「まほしげ」という状態名詞形を持つことなども、他のモダリティの助動詞とは異なっている。ただ、話し手の誂え望む気持ちを示す用法については、モダリティに準ずるといってもいい部分もある。さらに調査が必要であろう。上代語の「まくほし」との差など、コーパスでは従来わからなかった詳細な接続の既述が可能であり、さらに分析を続けてみたい。

いずれにせよ、コーパスを用いた古典語の語彙・文法研究は、緒についたばかりであり、今後、飛躍的な発展が望める分野であることは間違いのないことである。

注

1　Crystal, David, and Quirk, Randolph (1964). Systems of Prosodic and Paralinguistic Features in English. The Hague: Mouton.）に、この紙のカードのコーパスの話題が資料としてとりあげられている。
2　本特集の別稿で、フレレスビッグ教授による解説論文がある。
3　ペン・ヘルシンキコーパスについては、プロジェクトのホームページ参照。
http://www.ling.upenn.edu/hist-corpora/
4　長瀬真理「日本語―英語対照「源氏物語」のテキストデータベースの作成に関する基礎的研究」（『情報知識学会誌』一巻一号・一九九〇）
現在でもオクスフォードテキストアーカイブから自由にダウンロード可能である。
（日本語版）http://www.ota.ox.ac.uk/desc/2246
（英語版）http://www.ota.ox.ac.uk/desc/2245
5　長瀬真理・西村弘之『コンピューターによる文章解析入門―OCPへの招待』（オーム社・一九八六）
6　WebサイトURLは http://maro.ninjal.ac.jp/
7　小木曽智信他「「日本語歴史コーパス平安時代編」先行公開版について」
http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no3_papers/JCLWorkshop_No3_33.pdf
8　文字N-gramを用いた古典語の解析手法については次を参照。
近藤みゆき「ｎグラム統計処理を用いた文字列分析による日本古典文学の研究『古今和歌集』の「ことば」の型と性差」（『千葉大学人文研究』二九号・二〇〇〇）
9　単語N-gramを用いた古典文学作品の解析については次の論文がある。
太刀岡勇気「中古日記文学の計量国語学的分析と異本間の関係性の客観分析―『和泉式部日記』と『更級日記』を題材に一」（『計量国語学』二九巻六号・二〇一四）
10　近藤泰弘「電子化コーパスを用いた古典語のテンス・アスペクト研究」（『日本語学』三二巻一二号・二〇一三）
11　近藤泰弘「日本語モダリティの史的変遷」（『モダリティⅠ・理論と方法　ひつじ意味論講座３』ひつじ書房・二〇一四）

　なお、このモダリティ論文中で、「まほし」の人称制限に触れた部分があるが、尊敬語との相互承接の解釈でやや足りなかった部分がある。これについては別途再論したい。この場所を借りて付言しておく。

(記)　本稿は、科学研究費基盤研究（B）「平安時代の言語リソースの構築」の研究成果の一部である。また、国立国語研究所共同プロジェクト「通時コーパスの設計」（田中牧郎プロジェクトリーダー）の研究成果である。

（こんどう・やすひろ　青山学院大学教授）