日本語学とコンピュータ

丸山直子
（東京女子大学現代教養学部）

日本語学がコンピュータとどうかかわってきたかという歴史を，創立 50 年あまりの学会，計量国語学会の歴史と重ね合わせて振り返る．と同時に，最近ようやく日本語研究においてもさかんになってきたコーパス言語学について，現状と問題点を述べる．広く，データやツールを共有して研究を進めて行くことができるのは大変有益であるが，データもツールも使えるものを使う（使えるものだけ使って満足する）という傾向が見られるのは，反省すべき点である．

Japanese Linguistics and Computer

NAOKO MARUYAMA
（Tokyo Woman’s Christian University）

This talk reflects the history of Japanese linguistics and its relation to information technology. The use of computers in studying Japanese linguistics started around 50 years ago. Mathematical Linguistics Society of Japan, which was established in 1956, has been the central forum for quantitative and mathematical study of Japanese. Today computer-readable linguistic databases and analysis tools are widely available and many researches are conducted using them. Based on these experiences, we discuss tips and best practices of using computers in Japanese linguistic studies.

1. はじめに

日本語学という学問分野がどのようにコンピュータと関わってきたか，また，関わっているかについて述べる．日本語学の学会として最も中心的な学会は日本語学会（旧国語学会）であるが，コンピュータを利用した日本語研究は，情報処理学会の自然言語処理研究会や言語処理学会のような，理系中心の学会が主に担ってきた．そんな中で，計量国語学会は，国語学（日本語学）の研究者が中心となっている学会であり，日本語を数理的・計量的に研究しようとする，50 年以上の歴史を誇る学会である．ここでは，日本語学とコンピュータのかかわりを，計量国語学会の歴史と重ね合わせて振り返る．とともに，最近ようやく日本語研究においてもさかんになってきたコーパス言語学について，現状と問題点を述べる．

2. 日本語学と日本語処理（と東京女子大学）

日本語学というのは，要するに日本語を研究する学問分野であるが，わが国では江戸時代の国学に端を発するとされている．中世においても歌学書等に日本語に関する記述が存在するが，本格的な記述が始まったのは，国学以降である．国学の伝統を受け継いで発達してきたのが国語学（日本語学）であるが，数理的研究，計量的研究は，いつ始まったか．国立国語研究所が 1948 年に設立され，日本語についての本格的な調査を始めてから，と言えるのではないか．当初は，コンピュータを使用していなかった．コンピュータの使用は，1965 年から始まったようである．国語学（日本語学）は，長らくデータ採取にカードを使用してきた（いまだにカードでデータを管理している研究者もいる．日本語の歴史を研究している学者に多い）．コンピュータ使用は，当然，はじめは大型機，のちにＰＣを用いるようになる．国語学者自らプログラミングをすることも，現在より多かったようだ．機械語を学び，Cobol や Basic でプログラミングを行った．自らプログラミング言語を開発した国語学者もいる．東京女子大学名誉教授の水谷静夫氏は，黒川利明氏（当時東芝所属）の協力のもとで，1980 年代に朱唇というプログラミング言語を開発した（黒川（2004））．一時，東京女子大学日本文学科の「言語情報処理」の授業で教えており，現在もＰＣで稼働させることができる．東京女子大学日本文学科の「言語情報処理」は大型機の頃から設けられている科目で，使用する言語は，SNOBOL→LISP→朱唇→Perl と変化してきている．

一方で理系の研究者による日本語情報処理も，1970 年代の仮名漢字変換，1980 年代の機械翻訳等，大変盛んになり，その時期，日本語学・言語学を専攻した学生を採用して，より高度な日本語情報処理を行おうという努力が見られた．東京女子大学の水谷ゼミ・丸山ゼミの学生もかなり，日本語情報処理に携わる企業・研究機関に就職した．情報処理振興事業協会，計量計画研究所他，民間企業多種．今でも，時代の流れに従って従事している内容を変えつつも，それぞれの場で研究に勤しんでいる．

3. 計量国語学会の歩み

計量国語学会は，1956 年 12 月に発足した．国立国語研究所が設立されてから 8 年後のことである．2007 年に創立 50 周年を迎え，2009 年秋『計量国語学事典』を刊行した．日本語を計量的・数理的に研究する分野は，国立国語研究所の研究員を中心に確立され，かつ計量国語学会によって推し進められてきた分野であると言えよう．大型コンピュータが導入される前から，各種語彙調査が行われた.『計量国語学事典』の「計量国語学概説」の項から，「計量国語学」の歴史について少し引用する．

> 推測統計学の標本抽出による用語調査は国研の『現代語の語彙調査　婦人雑誌の用語』(1953) が世界的レベルでも嚆矢となった．(略) 計量国語学にコンピュータが使用されたのは，国研に大型計算機・HITAC3010 が導入された 1965 年からである．これにより，用字用語調査の調査量は飛躍的に伸びることとなった．また計量語彙論，計量文体論，社会言語学などでは，多変量解析のような高度に複雑な統計手法も導入できるようになった．1980 年代半ば以降は，計量国語学で利用されるコンピュータも，その主流は大型計算機からパソコンへとしだいに移っていくようになる．1990 年代に入るとパソコンの性能は 1970 年代の大型計算機よりも向上した．(『計量国語学事典』pp.18-20 より)

計量国語学の分野は，語彙・方言・文法等，多岐にわたる．当初は，文字・表記，語彙が主流であったが，後に日本語処理が盛んになると文法に関わるものが増える．現在は，社会言語学的な分野 (社会や場面との関わりから言語を捉える運用論的分野) が多いように思われる (1998 年に社会言語科学会が設立されたため，計量国語学会での研究発表は減るが，日本語学界全体では，社会言語学的研究が飛躍的に増えていると感じる.)

計量国語学の分野における初期の研究をいくつか紹介しよう．(詳しくは，『図説日本語』『計量国語学事典』参照.)

1）国研の各種調査

婦人雑誌の語彙調査／総合雑誌の語彙調査／現代雑誌九十種の用語・用字／
電子計算機による新聞の語彙調査／高校教科書の語彙調査／
中学教科書の語彙調査／テレビ放送の語彙調査／現代雑誌の語彙調査

2）大野の法則 (1956)・樺島の法則 (1954,1957)

・大野晋 (1956)「基本語彙に関する二三の研究－日本の古典文学に於ける」『国語学』24 号

古典作品における異なり語数の品詞別割合

名詞は，万葉集，随筆グループ，日記グループ，物語グループの順に減少する．
動詞と形容詞は，名詞と逆の順に増大する.

・樺島忠夫 (1954,1957)「類別した品詞に見ら得る規則性」『国語国文』24[6],55-57.

品詞比率 (談話語，戯曲，小説地の文，新聞社説，新聞記事，新聞見出し，『日本文学大辞典』，和歌，俳句)

$M=a-bN$,　$\log I=c-d\log N$,　$V=100-(N+M+I)$

N (名詞)，V (動詞)，M (形容詞・形容動詞・副詞・連体詞)，
I (接続詞，感動詞)

3）安本美典 (1963)・森岡健二 (1969)

・安本美典 (1963)「漢字の将来」『言語生活』137 号，46-54.

1900 年から 1955 年の間に発表された, 100 人の作家による 100 編の小説を対象に，それぞれ 1000 字ずつ抽出して, 漢字含有率が年代とともに減少傾向にあることを指摘.

・森岡健二 (1969)『近代語の成立－明治語彙編』明治書院.

1879 年から 1968 年までの新聞記事 (社会面) の文章が調査対象で，1 年おきに約 1000 字ずつとって，語種と漢字・仮名表記の関連について調査．仮名表記和語の増加を指摘.

4）宮島の類似度C (1970)，水谷の類似度D (1980)

・宮島達夫 (1970)「語いの類似度」『国語学』82 号,42-64.

$C_{AB}=\Sigma_i \min[P_i(A),P_i(B)]$

作品 A・B に共通する 1 組の単語のうち，使用率の小さい方の単語を取り出し，その使用率の総和をとったもの. $P_i(A)$, $P_i(B)$は二つの作品の使用率. この手法で，古典 14 作品の類似度を示した.

・水谷静夫 (1980)「用語類似度による歌謡曲仕訳」『計量国語学』12[4], 145-161.

$D_{A|B}=\Sigma_i P_i(A)=\Sigma F(A)$

語彙 A・B に共通する単語の，語彙 A における使用率の総和をとったもの.
$P_i(A)$は語の使用率, N は延べ語数, F は使用度数である. これを用いて 1935～1959 年の 20 篇の歌謡曲の分類を行っている.

4. 最近の研究

4.1 コーパス言語学（コーパスとツール）

国立国語研究所は，戦後ずっと日本語の研究に勤しんできたが，2009（平成 21）年 10 月 1 日に大学共同利用機関法人 人間文化研究機構の機関となった．外部との連携を重視する方向に転換している．

筆者は，2006 年度から 2010 年度まで国研が主導するコーパスのプロジェクト1に属していた．さらに 2010 年度からは，並行して，共同研究「コーパス日本語学の創成」にも参加している．ここでは，その中で使用しているコーパス，ツールについて述べる．

コーパス言語学は，既に欧米・アジアにおいて盛んに行われている分野であるが，日本は，著作権の問題もあり，遅れをとってきた．国立国語研究所が５年間のプロジェクトで１億語のバランスドコーパスを作るという計画を立て，ほぼ完成した．まだ，１億語という小規模なものなので，今後さらに，時代的にも数量的にも拡張していくことが望まれている．このコーパスは，BCCWJ（現代日本語書き言葉均衡コーパス Balanced Corpus of Contemporary Written Japanese）と名付けられているが，KOTONOHA という名前で現在（2011 年５月現在）デモバージョンが一般公開されている2.

これまで日本語研究によく用いられてきたコーパスとしては，新聞記事ＤＢ，新潮文庫１００冊を代表とするＣＤ化された小説類，国会会議録等が存在する3．新聞記事ＤＢは値段が高いが信頼できるコーパスである．新潮文庫１００冊も，小説のデータの中では，信頼できるデータであるとされている．青空文庫4のようなものもあるが、これはボランティアによって作成されているもので，誤りが多く，日本語研究に使用するには，注意を要する．国会会議録5は，誰にでもアクセスできる．戦後の日本語の経年変化を見ることのできる貴重な資料である（松田編（2008））．その他，現在，WEB からも大量の日本語データを採取できるわけであるが，情報は極めて不安定で，扱いがむずかしい（田野村（2011））．話し言葉のコーパスとしては，国立国語研究所・情報通信研究機構・東京工業大学が共同開発した話し言葉コーパス CSJ の他，東京外国語大学で作成された「BTS による多言語話し言葉コーパス」，日本女子大学で作成された DB「アジアの文化・インターアクション・言語の相互関係に関する実証的・理論的研究」，東京女子大学で作成した DB（非公開）等を使用している．

コーパスツールとしては，データ抽出系とデータ解析系に分けることができる．データ抽出系のツールとしては，国立国語研究所が「ひまわり」6や「中納言」7を提供しており，これは大変使いやすい．文系の言語研究者は，使いやすいものしか使わない傾向がある．また，なるべく，加工されたものでなく，生のデータを自分で見たいという気持ちがある．品詞の分け方，語の認定の仕方も個々の研究者で異なるからである．したがって，アノテーションされていない生のデータを文字列検索するというのが，まずは，好まれる．しかし，文字列検索では様々な限界があるため，「中納言」では短単位検索を実現した．短単位は，国立国語研究所が考案し実践してきた語の単位であるが，現在の語彙調査ではかなり標準化されてきているので，この単位で検索できるのは大変便利である．表記のゆれを超えた検索ができる．立命館大学で開発された「KH-Coder」8も大変使いやすく，学生にも好評である．奈良先端科学技術大学院大学で開発したタグ付きコーパスを管理・検索するためのツール「茶器」9は，多少スキルを要求されるが，きめ細かい検索ができ，優れたツールである．「茶漉」10も，検索できるＤＢが限られているが，重宝されている．

データ解析系としては，形態素解析ツールとして，茶筌11，Juman12，MeCab（和布蕪）13，構文解析（係り受け解析）ツールとして，KNP14，Cabocha15などが使われている．構文解析は，日本語学研究者の間ではそれほど使われていないのが実情である．精度の問題もあるが，それぞれの研究者のニーズに合わせたものにはなし得ないのが大きな理由であると思われる．文の構造をどう捉えるか，係り受け関係をどう考えるかは，まさに研究者の間で考え方の分かれるところなのである．それに比べて，形態素解析の方は，かなり重宝されている．但し，単独で使用するよりは，やはりデータ抽出系のツールに組み込んだ形の方が使いやすい．

かつては，日本語学研究者自らプログラミングすることも多かった（現在でも一部の研究者は，Perl などで，独自にカスタマイズしたソフトを作っている）が，最近は，できあいのツールを使用することが多くなった．ツールが充実してきたことは大変ありがたいが，同時に，いくつか問題点も生じている．

1 文部科学省科学研究費特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：２１世紀の日本語研究の基盤整備」略称「日本語コーパス」
2 http://www.kotonoha.gr.jp/demo/
3 コーパス日本語学のための情報館（http://www30.atwiki.jp/corpus-ling/）参照．
4 http://www.aozora.gr.jp/
5 http://kokkai.ndl.go.jp/
6 http://www2.ninjal.ac.jp/lrc/
7 http://morph.kotonoha.gr.jp/manual/manualTop.aspx
8 http://khc.sourceforge.net/
9 http://chasen.naist.jp/hiki/ChaKi/
10 http://tell.fll.purdue.edu/chakoshi-wiki/
11 http://chasen.naist.jp/hiki/ChaSen/
12 http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
13 http://mecab.sourceforge.net/
14 http://nlp.ist.i.kyoto-u.ac.jp/index.php
15 http://chasen.org/~taku/software/cabocha/

4.2. コーパス言語学の問題点

問題点の一つとして感じているのは，使えるものを使うことで満足する傾向にあるということである．日本語研究者は，データを収集するのが一つの仕事であった．それぞれ独自のデータベースを作成してきた（多くはカードの山であった）のだが，昨今は使えるコーパスを使って研究を行うことが増えてきた．また，生のデータを見なくなる，数値でものを言う，読んだことのない作品の例を挙げる，等の現象が見られる．出てきた数値だけで判断すると，実際にはあり得ない数値が出ていても，そのおかしさに気付かないことがある．また，全体像を把握することなく，局所的にデータを見ていると，例えば，単なる引用であっても，それを著者の言葉遣いである，と解釈してしまうような過ちも犯しやすい。生のデータを眺めることで獲得されてきた，かつての国語学者が持っていた勘のようなものが失われて来ているように感じる．

4.3. コーパスと辞書記述（丸山の研究）

最後に，筆者の研究について，若干述べる．筆者は，特定領域研究「日本語コーパス」の中で，辞書編集班に属し，国語辞書の記述にコーパスを利用する方法について探ってきた．実際に，『岩波国語辞典』第七版改訂作業に，BCCWJ をはじめとするコーパスを使用し，特に，例文の充実をはかる仕事をした．基本的な動詞 1454 語を対象に見直し，結局 537 語について，例文や他の情報の追加を行った．例えば，「カバーする」には，最近の用法として「古い曲をカバーする」のようなものがあり，この場合の語義は「オリジナルの曲に対して，他の歌手や演奏家が，自分の持ち曲として歌ったり演奏したりすること」である。これが『岩波国語辞典』第六版にはなかった．コーパスの種類によって現れる数が異なるものであるが，もとの語義からの派生の仕方が大きいこと，コーパスの種類によってはかなり用例が多いことから，この語義を追加することとした．一方「はまる」には，「ゲームにはまる」のような用法があり，この用法についても，第六版では記述がなかった．これも，くだけた文体のコーパスには大変多く見られる用法であるが，語義の派生の仕方から，この語については新たな語義を立てることなく，「②落ちこむ。陥る。」の用例として「ゲームにはまる」を載せるにとどめた．コーパスは，新たな用法を見つけるのに役立つ．別の語義を立てた方がよいのか，同じ語義の中の用法として用例を示すにとどめた方がよいのかの判断は，コーパスの情報だけではむずかしいが，ヒントを与えてくれるものとしては，活用できると考える．

5. おわりに

日本語学という学問分野がどのようにコンピュータと関わってきたか，また，関わっているかについて，計量国語学会の歩みを中心に述べ，また，最近のコーパス言語学で使用されているコーパスやツールの実態をご紹介した．計量国語学の分野は，むしろかつての方が勢いを持っていた．『計量国語学事典』に次のような文章がある．「計量国語学派の第一世代である渡辺修，水谷静夫，樺島忠夫，安本美典らは，超人的な努力によって，日本の計量国語学を世界的なレベルの研究から始めることに成功した．ところが，その成功が大きければ大きいほど，後に続く世代に高度な数学的な知識を要求することとなり，かえって計量国語学という分野が近寄りがたい存在になったことが考えられる．」しかし，近年充実してきた様々なツールを活用することで，庶民的研究者にも道が開ける可能性を感じる．問題点を認識しつつ，活用していきたいと考えている．

参考文献

黒川利明（2004）『ソフトウェア入門』岩波新書.
計量国語学会編（2009）『計量国語学事典』朝倉書店.
田野村忠温（2011）「日本語研究とインターネット」『特定領域研究「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集』419-426.
林大監修，宮島達夫・野村雅昭編（1982）『図説日本語』角川書店.
松田謙次郎編（2008）『国会会議録を使った日本語研究』ひつじ書房.
丸山直子（2010）「助詞「に」を伴う＜役割＞成分－コーパスに基づく分析－」『日本語文法』10[1], 71-87.
丸山直子（2011）「動詞の格情報－国語辞書の記述とコーパス－」『東京女子大学日本文学』107, 227-245.

コーパス

書き言葉
　青空文庫（http://www.aozora.gr.jp）／新潮文庫の 100 冊 CDROM／新聞記事 DB／
　現代日本語書き言葉均衡コーパス（BCCWJ　http://www.kotonoha.gr.jp/demo/）
話し言葉
　日本語話し言葉コーパス（CSJ）／BTS による多言語話し言葉コーパス／
　アジアの文化・インターアクション・言語の相互関係に関する実証的・理論的研究 DB

ツール

データ抽出系
　KH Coder／ひまわり／中納言／茶器／茶漉
データ解析系
　茶筌／Juman／KNP／MeCab(和布蕪)／Cabocha