通時コーパスをどう使うか

近藤泰弘（青山学院大学/国立国語研究所）†

How to Use the Historical Corpus

Yasuhiro Kondo (Aoyama Gakuin University / NINJAL)

1　はじめに

我々のプロジェクトにおいては、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。

- コーパスの対象に含める文献資料をどのようにして選定するか
- 選定した資料をどのように電子化しどのような情報を付与するか
- 古典テキストに対応した形態素解析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。こうした研究は、日本語史上のいくつかの時点の主要資料についてコーパスを試作し、これを活用した日本語史研究を実践することを通して行う。また、コーパスの構築作業における他機関との連携の可能性を探り、コーパス公開のために不可欠な著作権処理の問題についての検討も行い、通時コーパスの構築・公開に向けた諸課題に見通しを付けることを目的としている。

言語資源研究系の現代語コーパスにかかわる研究と連携を取り、コーパス開発センターで実施中の現代語コーパスの構築作業、著作権処理業務などとも関連付けて研究を進めている。プロジェクトの Web サイトは以下の URL である。

http://historicalcorpus.jp

2　通時コーパスの構造

基本的には、XML によるマークアップを施した電子化テキストである。従来の同様な試みとしては、国文学研究資料館の KOKIN ルールおよび SGML によるマークアップを施したコーパスがあるが、今回研究をしているコーパスにおいては、記述的マークアップをさらに言語的な要素にまで及ぼすことにしている。

- XML (eXtensive Markup Language) によるマークアップ

† yhkondo@cl.aoyama.ac.jp

- NINJAL BCCWJ (Balanced Corpus of Contemporary Written Japanese) と極力互換性がある
- 全文コーパス
- UTF-8 Encoding
- 形態素単位のマークアップ (自動的な形態論的解析)
- SUW (Short Unit Word) （短単位）

記述的マークアップが必要となる。具体的には次のようなものとなる。少なくとも、形態論的構造まではすべての収録テキストについてマークアップをする予定であるが、統語論的構造や意味論的構造についてはどの程度可能になるかまだ未定である。

1. 論理構造 cf. title, paragraph, citation, kokka-taikan-number
2. 表記構造 cf. ruby, page
3. 形態論的構造 cf. word, part of speech, inflection
4. 統語論的構造 cf. sentence, clause, phrase,
5. 意味論的構造 cf. Agent, Object,

2.1 XML タグセット

XML のタグセットおよびアトリビュートについてもまだ未定の点が多い。現在は、BCCWJ に習っている点も多いが、今後、古典語としての特質を調査しつつ、適切なタグセットを検討していきたい。

1. sample 文書
2. div　　内部構造
3. p 同上
4. pb Page Break
5. note 頭注
6. ruby ルビ
7. sentence
8. SUW　短単位

2.2 XML アトリビュート

1. (sample) ID, no, title, filename, etc.
2. (SUW) orthToken（出現書字形）、lForm（仮名形）、lemma（語彙素）、pos（品詞）、Form（原形）、PronToken (出現発音形)、w Type(語種)、start（開始文字位置）,end（終了文字

位置)、cType（活用型)、cType(活用形)、orderID（単語出現順番号）

3　マークアップの例（竹取物語）

これは、『竹取物語』の冒頭のマークアップ例である。前節で示した仮のタグセットによってマークアップしたものであるが、だいたいのイメージとしてとらえていただきたい。

¡?xml version=”1.0” encoding=”UTF-8”?¿
¡sample sampleID=”1201_竹取物語” no=”1201” title=”竹取物語”
fileName=”1201 竹取物語_100728”¿
¡div id=”00000001”¿¡div type=”古典本文”¿¡p org=”空 1”¿¡sentence¿
¡SUW orthToken=”　” lForm=”” lemma=”　” lemmaID=”23”
kana=”” pos=”空白” Form=”” pronToken=”” wType=”記号”
start=”10” end=”20” orderID=”10” BOS=”True” /¿
¡note org=”1” text=”1”¿¡/note¿
¡SUW orthToken=”いま” lForm=”イマ” lemma=”今” lemmaID=”2460”
kana=”イマ” pos=”名詞-普通名詞-副詞可能” Form=”イマ” pronToken=”イマ” wType=”和”
start=”20” end=”40” orderID=”20” /¿いま
¡SUW orthToken=”は” lForm=”ハ” lemma=”は” lemmaID=”29321”
kana=”ハ” pos=”助 詞-係 助 詞” Form=”ハ” pronToken=”ワ” wType=”和” start=”40” end=”50”
orderID=”30” /¿は
¡SUW orthToken=”むかし” lForm=”ムカシ” lemma=”昔” lemmaID=”37012”
kana=”ム カ シ” pos=”名 詞-普 通 名 詞-副 詞 可 能” Form=”ム カ シ” pronToken=”ム カ シ” wType=”和”
start=”50” end=”80” orderID=”40” /¿むかし

4　利用方法

BCCWJ のコーパスブラウザである「中納言」による利用を考えている。「中納言」では次のような検索が可能である。古典語の場合は、文法的内省が働かないため、特に「形態論情報を利用した検索」がひじょうに有効である。

1. 形態論情報を利用した検索が可能
2. 短単位検索、長単位検索、文字列検索の機能がある
3. 多数の用例が見つかった場合でも、その全体をダウンロードできる
4. 検索語の前後合わせて最大 10 単位まで条件が指定できる

5. 文脈を長めに（最大前後各 500 語まで）表示することができる

5　研究の方法

5.1　平安時代の「て」節の様相

平安時代語の複文は、従属節と主節との区分をすることが厳密には難しく、節が次々と連なっていく、いわゆる「節連鎖」的な特徴を持っていることが言える。その中でも「て」接続助詞による接続は、連鎖的な機能が強く、そのつながり方について詳しく調査することが必要である。

一般に、係助詞は、連用修飾から接続助詞まで広くその後に用いることができるのに対し、副助詞にはその分布がより狭い。また、発表者が論証したように、副助詞は次の二種類に分類できる（1 種・２種の命名は、小柳智一 (1998) による）。

- 第 1 種　ばかり・まで　格助詞に前接・形容詞連用形に後接しない・副助詞に前接
- 第 2 種　のみ・さへ・だに　格助詞に後接・形容詞連用形に後接・副助詞に後接

このうち第 1 種は、連用修飾や格助詞の後に接続しないため、今回の研究からは当面除外して考えることができる。第 2 種については、いわゆる連用修飾や格助詞、そして、接続助詞の一部に至るまで接続することができるのであるが、それが具体的に「て」助詞との関係においてどのような分布になるかは従来わかっていなかった。発表者は、中古語、特に『源氏物語』を資料として、次のようなことを論証した。

接続助詞「て」と言われるものの中には、現代語と同じように、二種類のものがあり、ひとつは、従属節の従属度としては、いわゆる A 類に属するものである。意味的には「付帯状況」を示す。もうひとつはいわゆる B 類に属すると思われるものである。意味的には「継起」「原因／理由」「並列」を示す。そして、A 類に属するものは、第 2 種副助詞を後接することができるのに対し、B 類に属するものは、第 2 種副助詞を後接することが不可能であった。したがって、第 2 種副助詞を後接している文型を調査することで、A 類の「て」の類型を知ることが可能になる。

- （A 類）
  - –（付帯状況）おぼつかなくてのみ年月の過ぐるなむあはれなりける（源氏物語・若菜下）
  - –（付帯状況）おぼししづみてのみおはするを（夜の寝覚）
- （B 類）
  - –（継起）まことに明け方になりてぞ、宮帰り給ふ（源氏物語・梅枝）
  - –（原因理由）道にてやまひしてなむ、死にける。（大和物語）
  - –（並列）知らぬ国に吹き寄せられて、鬼のやうなるもの出で来て殺さんとしき。（竹取物語）

5.2　平安時代の「て」節と疑似分裂文

今回、コーパスによってさらに精密に「て」節を観察することができた。

現代語では、A 類の「て」節は疑似分裂文とすることができるが、B 類の「て」節はそれが不可能なことが知られている（内丸 2006)。

- （A 類）
  - –（付帯状況）太郎はよそ見をして車を運転していた。
  - – 太郎が運転していたのは、よそ見をしてだ。（疑似分裂文）
- （B 類）
  - –（並列）おじいさんは山に芝刈りに行って、おばあさんは川に洗濯に行った。
  - – *おばあさんが川に洗濯に行ったのは、おじいさんが山に芝刈りに行ってだ。（疑似分裂文）
  - –（継起）電車を降りて、改札を抜けた。
  - – ？改札を抜けたのは、電車を降りてだ。（疑似分裂文）
  - –（原因）台風が近づいて、学校が休みになった。
  - – ？学校が休みになったのは、台風が近づいてだ。（疑似分裂文）

平安時代語の場合の「て」節を「なり」が受ける疑似分裂文を調査してみると、先に A 類として分類したものがほとんどをしめる。

- あえかに見えたまひしも、かく長かるまじくてなり。（源氏物語・夕顔）

ただし、原因理由を表す B 類かと思われる「寄りて」だけは疑似分裂文となることができる。

- （よりてなり）
  - 夏虫の身をいたづらになす事もひとつおもひによりてなり（古今集）
  - おぼろげの願によりてにやあらむ（土佐日記）
  - 苦しきによりてにや。（枕草子）

この「よりて」は、表面的には原因理由を示すように読めるが、平安時代語の段階では、「寄る」という動詞の意味がまだかなり強く、付帯状況を示しているという可能性が高いのではないか。

- さらに許されぬによりてなむ、かく思ひ嘆き侍る。（竹取物語）

6　従属節の従属度と副助詞・係助詞

さて、以上調査してきたように、A 類には、副助詞・係助詞が後接し、B 類には係助詞のみが後接する。また、ここには記述しなかったが C 類には副助詞はもちろん、係助詞も後接しない。この

ようにA類からC類にむかって階層的な秩序があるが、個々の接続助詞にはそれぞれ別な制約もあるようである。たとえば、B類のうち、順接の「已然形＋ば」や「未然形＋ば」には係助詞が後接する。それに対し逆接の「も」「とも」には副助詞はもちろん係助詞も接続しない。この制約は統語的なものではなく、意味的なものだろう。

全体としては、おおよそ次のようになる。

類	形態	接続
A類	て（A類）・ながら・ず・で・用言連用形	副助詞・係助詞後接
B1類 （順接）	て（類）・ば・（く）は・つつ・ず・で・用言連用形	係助詞後接
B2類（逆接)	とも・ども・ものの	どちらも後接せず
C類	を・に・が	どちらも後接せず

このように「て」のA類・B類を正確に分離して記述することで、平安時代語の従属節全体についての見通しも明らかになってくるのであり、コーパスを駆使することで記述の精度を格段に上げることが可能になるのである。

文献

[1] 内丸裕佳子 (2006)「動詞のテ形を伴う節の統語構造について一付加構造と等位構造との対立を中心に一」(『日本語の研究』2巻1号)

[2] 小柳智一 (1998)「中古の「ノミ」について一存在単質性の副助詞一」(『國學院雑誌』99巻7号)

[3] 近藤泰弘 (2012)「平安時代語の接続助詞「て」の様相」(『国語と国文学』89巻2号)

[4] 安永尚志 (1989)『日本古典文学作品本文データベース仕様書（第二版)』(国文学研究資料館)

[5] 安永尚志編 (1998)『講座　人文科学研究のための情報処理　[第3巻　テキスト処理編]』(尚学社)

[6] 安永尚志 (1998)『国文学とコンピュータ』(勉誠社)