情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2011-CH-92 No.5
2011/10/8

日本語通時コーパスの設計について

近藤泰弘[†1]

古典語のコーパスの設計が，現代語のコーパスの場合とどのように異なるかという問題について改めて考え直す．具体的には，どのような観点でコーパス化する資料を選定するか，どのように古典本文を電子化し，どのような情報を付与するか，各時代・各文体の語彙や文法に対応した形態素解析をどのように行うかなどである．

Design of a Japanese Diachronic Corpus

Yasuhiro Kondo[†1]

This paper shows the grounds for selecting materials for the corpus and how classical texts are digitized, and what kinds of information are added and how morphological analysis corresponding to the vocabulary and the grammar of each period and writing style is conducted.

1. はじめに

国立国語研究所の基幹型プロジェクトのひとつとして，発表者がプロジェクト・リーダーとなり「通時コーパスの設計」を行っている．本発表では，その紹介を兼ねて，通時コーパスにどのような問題点があるかについて述べる．

まず最初にプロジェクトの紹介をしておきたい．プロジェクトは，将来に行うべき日本語の通時的な総合コーパスの試験プロジェクトという位置づけであり，2010 年から 2015 年にかけて行われている．メンバーは以下の通りである．

近藤泰弘（青山学院大学／国立国語研究所），高山善行（福井大学），山元啓史（東京工業大学），岡部嘉幸（千葉大学），村上　謙（埼玉大学），山田昌裕（恵泉女学院大学），田中牧郎（国立国語研究所），小木曽智信（国立国語研究所），岡崎友子（就実大学），川村大（東京外国語大学），Bjarke Frellesvig（オクスフォード大学）,Stephan Horn（オクスフォード大学）,Kerri Russel（オクスフォード大学）

現在までに行ったこととしては次のようなものがあげられる．

(1)　『小学館日本古典文学全集』の入力・形態素解析
(2)　江戸語資料の入力
(3)　自動単位切りの手作業による訂正（「大納言」＝データベース操作ツール）
(4)　自動単位切りの精度向上（中古和文 UniDic の改良）

具体的には，メンバーの小木曽智信氏が中心になって開発されている中古和文 UniDic を用いて，形態素解析を行った古文データをもとに作成している．2015 年までに方法論にめどをつける。

それ以後，予算処置などが可能になれば，NINJAL 全体の事業として通時コーパスを拡張していく可能性もある．コーパスだけでなく、利用ツール (中納言・少納言)・辞書などを積極的に作成することも考えている．また，近代語（明治時代語）プロジェクトと協力して，コーパスを作っていくことも必要である．

現在作業を行っているのは次のテキストである．

- 古今和歌集
- 竹取物語
- 伊勢物語
- 大和物語
- 平中物語
- 土佐日記
- 蜻蛉日記
- 落窪物語
- 堤中納言物語
- 枕草子
- 源氏物語
- 和泉式部日記

†1 青山学院大学 / 国立国語研究所
Aoyama Gakuin University / National Institute for Japanese Language and Linguistics

- 紫式部日記
- 更級日記
- 讃岐典侍日記
- 今昔物語集
- 平家物語
- 近松門左衛門集

2. 入力のフォーマット

- XML 形式
- BCCWJ（書き言葉均衡コーパス）に準拠
- 全文コーパス
- UTF-8 コーディング
- 形態論的単位によるマークアップ
- SUW（国語研・短単位）による分割

3. XML タグセット

- sample 文書
- div　　内部構造
- p 同上
- pb Page Break
- note 頭注
- ruby ルビ
- sentence
- SUW　短単位

4. アトリビュート

(sample) ID, no, title, filename, etc. (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）、wType(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）

参 考 文 献

1) 小木曽智信，間淵洋子，前川喜久雄：階層的形態論情報を考慮した『現代日本語書き言葉均衡コーパス』の公開用 XML フォーマット,『現代日本語書き言葉均衡コーパス』完成記念講演会予稿集，pp.35–42 (2011).
2) 小木曽智信，小椋秀樹，田中牧郎，近藤明日子，伝康晴，中古和文を対象とした形態素解析辞書の開発,『情報処理学会研究報告 人文科学とコンピュータ』Vol.2010-CH-85(No.4), pp.1–8(2010).
小木曽智信，小椋秀樹，近藤明日子，須永哲矢，形態素解析辞書「中古和文 UniDic」とその活用例,『日本語学会 2010 年度秋季大会予稿集』，pp.243–248(2010)
3) 高田智和，山口昌也：BCCWJ「書籍コーパス」の JIS 外字,『現代日本語書き言葉均衡コーパス』完成記念講演会予稿集，pp.29–34 (2011).
4) 国立国語研究所他：形態素解析辞書 UniDic，言語データベースとソフトウェア（言語資源公開），入手先⟨http://www2.ninjal.ac.jp/lrc/index.php?UniDic⟩（参照 2010-09-09）.

付　　録

A.1 XML フォーマットサンプル

```
<?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="  " pos="空白" Form="" pronToken="" wType="記号"
start="10" end="20"
morphID="10"
BOS="True" />  <note org="1" text="1"></note><SUW orthToken="いま"
lForm="イマ" lemma="今"
pos="名詞-普通名詞-副詞可能" Form="イマ" pronToken="イマ" wType="和"
start="20" end="40"
morphID="20" />い<SUW orthToken="は" lForm="ハ" lemma="は"
pos="助詞-係助詞" Form="ハ"
pronToken="ワ" wType="和" start="40" end="50" morphID="30" />は<SUW
```

orthToken="むかし" lForm="ムカシ" lemma="昔"
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"
morphID="40" />むかし<SUW orthToken="、" lForm="" lemma="、"
pos="補助記号-読点"
Form="" pronToken="" wType="記号" start="80" end="90"
morphID="50" />、
<SUW orthToken="たけ" lForm="タケ" lemma="タケ"
pos="名詞-固有名詞-地名-一般"
Form="タケ" pronToken="タケ" wType="固" start="90"
end="110" morphID="60" />
たけ<SUW orthToken="とり" lForm="トリ" lemma="鳥"
pos="名詞-普通名詞-一般"
Form="トリ" pronToken="トリ" wType="和" start="110"
end="130" morphID="70" />
とり<SUW orthToken="の" lForm="ノ" lemma="の"
pos="助詞-格助詞" Form="ノ"
pronToken="ノ" wType="和" start="130" end="140"
morphID="80" />の<ruby
rubyText="おきな"><SUW orthToken="翁" lForm="オキナ"
lemma="翁"
pos="名詞-普通名詞-一般" Form="オキナ"
pronToken="オキナ" wType="和"
start="140" end="150" morphID="90" />翁</ruby><SUW
orthToken="と"
lForm="ト" lemma="と" pos="助詞-格助詞" Form="ト"
pronToken="ト"
wType="和" start="150" end="160" morphID="100" />と<SUW
orthToken="いふ"
lForm="イウ" lemma="言う" pos="動詞-一般" Form="イウ"
cType="文語四段-ハ行"
cForm="連体形-一般" pronToken="ユー" wType="和"
start="160" end="180"
morphID="110" />いふ<SUW orthToken="もの" lForm="モノ"
lemma="物"
pos="名詞-普通名詞-サ変可能" Form="モノ" pronToken="モノ"
wType="和"
start="180" end="200" morphID="120" />もの<SUW orthToken="あり"
lForm="アル"
lemma="有る" pos="動詞-非自立可能" Form="アリ"
cType="文語ラ行変格"
cForm="連用形-一般" pronToken="アリ" wType="和"
start="200" end="220"
morphID="130" />あり<note org="2" text="2"></note>
<SUW orthToken="けり"
lForm="ケリ" lemma="けり" pos="助動詞" Form="ケリ"
cType="文語助動詞-ケリ"
cForm="終止形-一般" pronToken="ケリ" wType="和"
start="220" end="240"
morphID="140" />けり<SUW orthToken="。" lForm=""
lemma="。" pos="補助記号-句点"
Form="" pronToken="" wType="記号" start="240" end="250"
morphID="150" />。
</sentence>