代表性を有する書き言葉コーパスの電子化フォーマットについて

間淵 洋子　　山口 昌也　　柏野 和佳子　　田中 牧郎

独立行政法人 国立国語研究所

1　はじめに

国立国語研究所では，現在，代表性を有する書き言葉コーパスの構築を計画している。この計画は，現代日本語の書き言葉を対象としたバランストコーパスを構築する初めての試みであり，語彙調査，文字調査，文法研究などの言語学的研究，辞書編纂や教育などの産業的な応用など，さまざまな要請に対応する大規模汎用コーパスとしての役割が期待されている。コーパスサイズは１億語強が想定されている[1]。

母集団は，新聞，雑誌，書籍と主とし，生産実態と流通実態に基づき，統計的な代表性の確保を狙う。生産実態を捉えた母集団からは一定文字数のサンプル(「固定長サンプル」と呼ぶ)を，流通実態を捉えた母集団からは記事・節・章などの言語的まとまりを持つサンプル(「可変長サンプル」と呼ぶ)を抽出し，母集団・サンプル長の異なる２種類のコーパスを構築する[2]。

２種類のコーパスは，いずれも XML によって研究用付加情報が付与される。書誌情報や文字情報の他に，国語研究所が既に公開している『日本語話し言葉コーパス』に準拠した形態論情報，そして，文書構造に関する情報が付加されることが，本コーパスの特長である。

本稿では，本コーパスにおける電子化フォーマットの設計・仕様を紹介し，効率的なデータ作成のための方策について述べる。

2　電子化フォーマットの設計

2.1　基本方針

電子化作業にあたっては，どのような内容を，どのような形式で，何を用いて電子化するのかということが問題となる。そこで，本コーパスの電子化フォーマットは，以下の方針に基づいて設計を行なった。

- 言語分析，自然言語処理に適した形式であること
- シンプルで分かりやすく，必要に応じて拡張できる形式であること

この方針は，本コーパスで行なう自動形態論情報付与を考慮すると共に，コーパスが目標とする各研究方面への活用や，今後作成されるコーパスの電子化フォーマットの雛形としての利用を視野に入れたものである。

これに基づき設定した，記述情報，形式化方式，文字集合を以下に示し，次節にて詳細を述べる。

1. **記述情報：A) 書誌情報，B) 文書構造情報，C) 文字情報，D) サンプリング情報**
2. **形式化方式：**XML により，独自の文書型を定義
3. **文字集合：**JISX0213:2004
4. **符号化方式：**UTF-16

2.2　研究用付加情報

2.2.1　記述情報

A) **書誌情報**

サンプリング時の指標である，出版年，メディア(新聞，雑誌，書籍など)，ジャンルの他，リソースを参照するための情報として，書名，著者名，出版社などを管理する。

著者については，氏名だけでなく，言語分析に有用な，生年，性別，出身地などの情報を含めて管理する。コーパス本体とは別にデータベース化し，サンプル本文内部に付与された ID とリンクして参照できるようにする。

B) **文書構造情報**

記事，見出し，段落，引用，文などの枠組みを用意し，テキストを構造化して表現する。これらの枠組みは，以下の目的を満たすものとして策定したものである。

- ある一定のまとまりをもつ文書(記事を想定)が有する，階層性を表現できること
- メディアやジャンルによる文書構造の差異を表現できること
- 文体，語彙，文法に差異の見られるような要素を区別できること

文書構造情報の枠組みについては，3 節に詳細を示す。

C) **文字情報**

文字の読みに関する**ルビ**，誤植等の**校正注**，文字集合に含まない文字や記号**(外字)**などの情報を付与する。

D) **サンプリング情報**

サンプリング時に決定するサンプル抽出基準点(乱数による縦横交叉点から決まる 1 文字)の情報を付与する。

---

[1]コーパスの概要・設計方針については，山崎他を参照。

[2] 実態調査の概要，およびサンプリング手法については，丸山他 2a，丸山他 b を参照。

2.2.2 形式化方式

形式化の方式として，XML によるマークアップを採用する。

XML(eXtensible Markup Language)は，構造化されたデータをテキスト形式で記述することができ，拡張性に優れた記述言語である。タグの検証が可能で，形式の妥当性を確保しやすく，検索や情報抽出，データ形式の変換が容易にできるという特徴もあり，コーパス記述に適した方式として，一般的に用いられる。既に国語研究所で公開している『太陽コーパス』『日本語話し言葉コーパス』等のコーパスも，XML による記述を採用しており，その流れを継承するものでもある。

2.2.3 文字集合

文字集合は，JISX0213:2004 に準拠したものを用いる。

国内規格としては，最も符号化文字数が多く，国内の印刷事情を考慮した異体字についての明確な包摂規準を持ち，コードマッピングにより，Unicode 規格との互換性を確保できる利点がある。

3 電子化フォーマットの仕様[3]

3.1 既存コーパスの電子化フォーマット

近年のコーパス言語学の発展により，大規模コーパスの構築や活用に対する，機運の高まりが見られるが，コーパスと銘打ったものに限らず，言語の電子化データは，これまでも多く作成されてきた。

これらの言語資料は，さまざまなフォーマットで電子化されている。欧米の大規模コーパスの多くが，電子文書やコーパスの電子化共通規格を提供する TEI(Text Encoding Initiative) [4] や，CES(Corpus Encoding Standard)[5]などの標準規格に準拠する電子化方式を採用する一方，国内では，統一的な言語資料の電子化規格が，今のところ確定していない。そこで，本コーパスの電子化フォーマットは，国際規格との連携を保ちつつ，日本語書き言葉の特性を捉えたものを目指す。策定に際しては，コーパスを構成する，新聞・雑誌・書籍を中心に，様々なメディアの文書からサンプルを取り，文書構造の把握を試みた。この試行を元に，上記の目的に合致する枠組みを検討･決定した。要素の一部を表 1 に示す。

3.2 仕様の特徴

ここでは主に，記事単位のサンプル(可変長サンプル)に付与される文書構造に関する仕様について述べる。この仕様の特徴は以下の３点である。

1. 言語の階層構造を，入れ子関係によって示す。
   記事>**クラスタ**>**段落**>**文**等の要素間の親子関係や**クラスタ**の入れ子によって文書の論理構造を明示的に記述

表 1 要素リスト(抜粋)

情報	要素	XML タグ	意味 (* 括弧内は属性による記述)
文書構造	サンプル	sample	サンプルの範囲(サンプルタイプ, 書誌 ID)
	記事	article	同一著者, 同一テーマによる文章の単位(記事 ID)
	タイトル	title	ある範囲の文書要素に対する代表記述
	クラスタ	c	タイトルが包括する文書要素の範囲
	概要	abstract	記事についての概要, 要旨や前文
	著作者情報	authorsData	文書作成に携わった人･組織の表示
	図表付随	fgBlock	図表, 写真と, それに付随するキャプション
	項目	list	箇条書き, 順序付き･定義語付きリストなどの並列要素
	注	noteBody	脚注･後注･傍注など, 本文行に含まれない注要素
	段落	p	一字下げで始まり, 改行で終わる, 文のまとまり
	文	s	句点, "!", "?"などの記号類によって区切られる語のまとまり
	引用	quotation	他文献からの引用や, 発話の書き起こし(引用タイプ)
	韻文	verse	詩や和歌などの韻文
文字	ルビ付き文字	ruby	本文行外に振り仮名等が付されている文字(ルビテキスト)
	外字	missingChar	文字集合外の文字(Unicode 番号, 字体説明)
	校正注	correction	原文における誤植を修正した箇所(誤植タイプ, 原文)

[3] 現仕様では，block 要素約 30，inline 要素約 15 を定義している。形態論情報付与の形式は現在策定中であり，今回の発表では触れない。

[4] http://www.tei-c.org/

[5] http://www.cs.vassar.edu/CES/, http://www.cs.vassar.edu/XCES/

2. 文書中の役割・位置づけを明確に示す。
   タイトル，**概要**，**キャプション**など，文書内での特別な役割を持つ要素を抽出でき，文書内容把握，文書要約，文書タイプ分類などへの活用を可能にする。
3. 言語的特徴・差異が想定される要素を，的確に選択できる。
   文体・語彙・文法的差異の見られることが知られる**タイトル**，**キャプション**，**引用(文献引用・会話)**などを区別でき，要素間の比較や，必要要素の抽出・不要要素の排除を容易にする。

3.3 形式化の例

3.3.1 階層構造の記述とタイトルの抽出

文書内容の把握において重要な役割を果たす要素であるタイトルの抽出と記述形式の例を示す。

新聞や書籍などの定型のテキストとは異なり，雑誌のタイトルの付け方は多種多様である。タイトル周辺に多数の修飾要素が配されるものが多く，それらの位置付けは, タイトルとも記事本文とも異なる。また, トピック, 内容要約，一部抽出など，いずれもタイトルとしての性質を持つ要素が重層的に複数表示されるものがある。これらを，タイトル付属要素と捕らえ形式化する[6]。

```
<titleBlock>
こんなステキな生き方、理想のキャリア・ライフをめざしたい！
<title>憧れの大人の女優６</title>
</titleBlock>
```

このフォーマットは, タイトルの重層性を示すと共に, 新聞タイトルや書籍の章節タイトルとの関連を考慮し, title 要素の等質性を確保する形式にもなっている。

3.3.2 発話書き起こしの記述

発話の形式化の例として，雑誌等に多く見られる対談やインタビューの書き起こし記事の例を示す。

発話書き起こしには，話者表示と共に発話が示されているもの，発話のみが示されているものなど，多様なスタイルがある。これらについて，(1)話者の交代によって区切られる 1 話者の 1 発話範囲を示す, (2)話者表示がある場合には，1 発話範囲とセットで記述する，(3)発話以外のものを排除できる形で記述することを考慮した結果が，以下のフォーマットである[7]。

```
<quotation type=”speech”>
<spMarker>ジャマイカ</spMarker>
<p><s>　「でも、ホントはオレ控えめで大人しいんですよ？</s><stageInline>（一同爆笑）</stageInline><s>人見知りするしね」</s></p>
</quotation>
```

このフォーマットは，シナリオや小説の会話部等の記述にも適用でき，本文から会話部を，会話部から会話以外の要素を抽出・排除できる。入れ子関係により発話の話者を表現する形式にもなっている。

4 電子化作業

4.1 電子化作業における環境整備の目的

電子化作業はコーパス作成の核であり，膨大なデータを，短期間に高い精度で作成することが求められる。

大規模データの作成には，多くの作業者が様々な環境で作業に携わる。各作業者が効率良く作業を行なう必要があり, また, データを的確にチェックする必要がある。

この問題に対処するために，電子化作業を幾つかの段階に分け行なうこととし，各作業段階においては，様々な作業環境の整備を行なうこととした。次節で，作業の流れに沿って詳細を示す。

4.2 電子化作業の各段階の詳細

4.2.1 サンプル紙面の確認と文字入力

抽出されたサンプルについてサンプル抽出規準点，範囲, データ化不要箇所等を確認した後, 文字を入力する。

この際，固定長サンプルの範囲確定に必要となる独自規定による文字計測と，JISX0123 へのサポートが必要となる。そこで, 文字入力環境作業には emacs を導入し, 独自規定による文字計測，JISX0213 規格文字の入力等に対応するマクロを実装して，作業の効率化を図った。

4.2.2 タグの簡易入力

次に，文書構造を表わすためのタグを入力する。

この時，XML タグそのものを入力するのは，非効率的で誤りも多い。より高速かつ簡便にタグ付けをするため, 簡易タグをテキストデータ内に入力する方式を取る。

この簡易タグ入力方式は，例えば title 要素を,「t」と「/t」で挟んで記述するというもので，高速にタグ付けの作業を行なうことができる(図 1･上段)。

簡易タグ入力についても, 上記 emacs マクロで入力の支援と入力結果の検証を行ない，作業の効率化とデータ形式の妥当性を確保する。

エディタ，マクロの整備の他に，文字入力の形式や,

---

[6] 例：金子「憧れの大人の女優６」,『SCREEN』58(14),2003,pp.65-70

[7] 例：「B-Press」,『BACKSTAGE PASS』19(12) ,2003,pp.171-175

類似文字の包摂や使い分け，簡易タグ付けについてのマニュアルを作成した。マニュアルは，実例の分析に基づき作成し，豊富な例と詳細なルールを示すことで，判断の揺れを極力排除するよう心掛けた。

4.2.3 自動 XML 変換

簡易タグ入力の済んだテキストは，perl スクリプトにより，XML ファイルへ変換を行なう。簡易タグの置き換えの他に，テキストを手がかりに，段落，文，改行位置などの要素を自動判別し，タグを付与する(図 1・下段)。

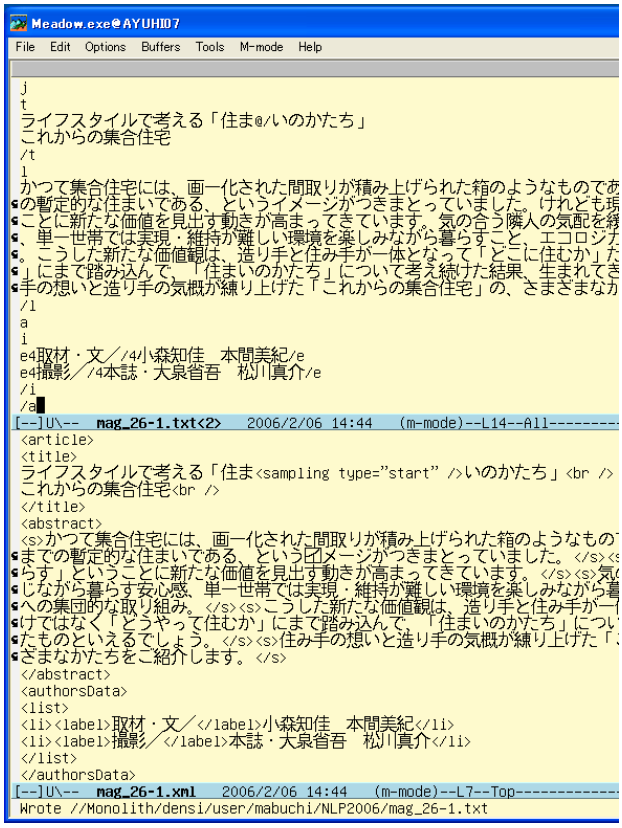

図 1　簡易タグ入力ファイル(上)と変換した XML ファイル(下)[8]

4.2.4 データチェックと修正

タグ付けのデータチェックは，XSLT を介して各要素を明示的に表現する形で web ブラウザに表示して行なう。要素別に表示を変え，階層の深さと背景色の濃さを関連付けるなど，文書の階層構造についても，視覚的に把握できるよう工夫している(図 2)。

4.2.5 DTD チェックと情報追加

構造・文字情報のチェック→簡易タグ付けテキストの修正→XML への変換→チェックの循環によって，データを確定し，DTD による形式チェックを行なう。

簡易タグ入力の対象外となっている一部の属性等の情報を追加し， XML ファイルが完成する。

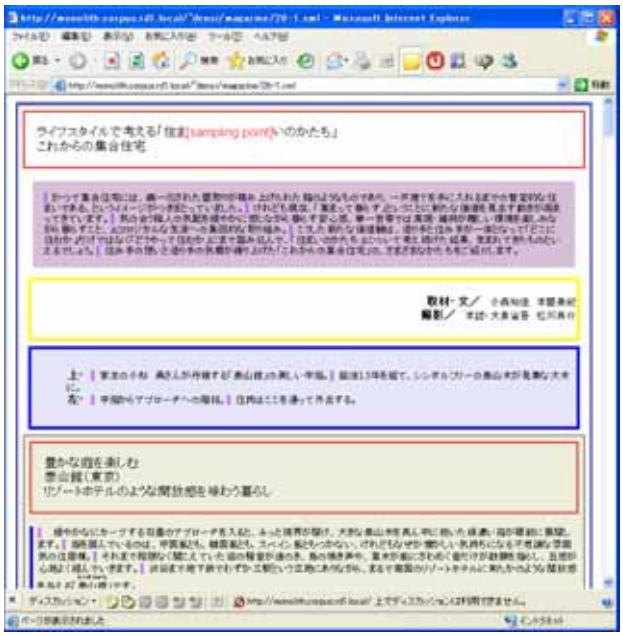

図 2　構造化データチェック画面

5　おわりに

国立国語研究所で構築を計画している，代表性を有する書き言葉コーパスの電子化フォーマットの設計方針と仕様を紹介した。また，データ作成のための作業環境整備の必要性と，実際の環境整備について述べた。

仕様や環境整備には，不十分な点も多い。今後のコーパス作成を通じて，研究利用により適した形式の策定，高精度な大規模データをより効率的に作成できる環境の整備を目指したい。

参考文献

[1] 山崎他「代表性を有する書き言葉コーパスの設計」,本予稿集所収,2006

[2] 丸山他 a「現代日本語の書き言葉に関する生産実態と流通実態」,本予稿集所収,2006

[3] 丸山他 b「代表性を有する書き言葉コーパスのサンプリング手法について」,本予稿集所収,2006

[4] 齊藤・中村・赤野『英語コーパス言語学 基礎と実践（改訂新版）』,研究社出版，2005

[8] 例：小森，本間「ライフスタイルで考える「住まいのかたち」これからの集合住宅」,『家庭画報』46(11),2003,pp.388-405　(※図 2 同様)