

国立国語研究所学術情報リポジトリ

『明六雑誌コーパス』の仕様

メタデータ	言語: Japanese 出版者: 国立国語研究所 公開日: 2020-03-18 キーワード: 作成者: 近藤, 明日子, 田中, 牧郎, KONDO, Asuko, TANAKA, Makiro メールアドレス: 所属:
URL	https://doi.org/10.15084/00002768

『明六雑誌コーパス』の仕様

近藤 明日子 (国立国語研究所コーパス開発センター)¹

田中 牧郎 (国立国語研究所言語資源研究系)²

1. はじめに

本稿では、本プロジェクトで設計している「近代語コーパス」のモデルとして構築した『明六雑誌コーパス』の仕様について説明する。

「近代語コーパス」は、明治時代から昭和時代ごろまでを対象に、近代日本語を代表でき、近世までの日本語から現代日本語への変化の過程を歴史的にたどることができるものにするのが望まれる。そのためには、多種多様な資料をコーパス化の対象にしていく必要があるが、最初に取り組む資料として『明六雑誌』を選んだ。

2. 『明六雑誌』を選ぶ理由

『明六雑誌』は、明治6(1873)年に学術啓蒙を目的に結成された明六社の機関誌で、明治7(1874)～8(1875)年に、1号から43号まで発行された。森有礼、津田真道、西周、西村茂樹、中村正直、加藤弘之、福沢諭吉、箕作麟祥ら16名が執筆している。西洋の近代思想を普及するために書かれた広範な論説が155編おさめられている。大半は漢文訓読風の文語体であるが、中には演説的な口語体も含まれている。思想史上の重要資料とされてきたが、日本語学においても、特に西洋からの新概念を取り入れるために必要とされた新漢語の資料として注目されてきたものである。明治前期の日本語研究資料として重要なものであり、複製本、注釈書、総索引なども整備され³、研究の蓄積もある⁴。このような特徴から、「近代語コーパス」が対象とする時代の初期の資料として、最初に取り組むのに適切なものだと判断した⁵。

本プロジェクトにおける「近代語コーパス」は、2005年に公開した最初の近代語コーパスである『太陽コーパス』を踏まえ、これを発展させる形で設計を行っている。『太陽コーパス』は、雑誌の本文を、記事、引用、擬似的な文の単位で構造化し、マークアップ言語XMLでタグ付けをしたものであり、記事や引用については著者、話者、文体などの情報をタグの中に入れて書き込み、校訂注記や異体字などの情報も、本文の当該箇所にタグ付けをして埋め込んだものである。日本語史研究資料におけるはじめての構造化テキストタグ付きコーパスである(田中2005)。ところで、コーパスとして言語研究に本格的に利用して

¹ kondo@ninjal.ac.jp

² mtanaka@ninjal.ac.jp

³ 『明六雑誌』の注釈書と校訂本文に山室・中野目(1999-2009)があり、複製本・総索引に高野・日向(1998)がある。

⁴ 『明六雑誌』の言語研究に、神奈川大学人文学研究所(2004)や高野(2004)がある。

⁵ 「近代語コーパス」の資料選定の考え方については、本報告書に収録した田中牧郎「近代語コーパスにおける資料選定の考え方」を参照。

いくためには、文章を構造化するだけにとどまらず、単語のレベルまで構造化を行うことが望まれる。『太陽コーパス』の開発段階からそれは意識されていたが、当時の技術では実現が困難であった。しかし、近年の研究の進展により、近代語テキストに対しても、単語に区切り読みや品詞を与えていく形態素解析技術が適用できるようになってきた⁶。そこで、本プロジェクトにおける「近代語コーパス」においても、形態素解析データを含めたコーパスを設計することとした。一般に形態素解析は、文法が整った書き言葉には比較的適用しやすいが、そうでない話し言葉に適用するのはやや困難が伴う。まずは、書き言葉に適用し、その後話し言葉に適用させていく手順を取ることが現実的である。特に、近代語の口語を反映した資料は、地域や階層などによる言語の多様性が大きく、形態素解析技術の実現にはいくつかの研究段階が必要とされると見通されている。また、近代語の文語体書き言葉も、言文一致以前の文体は多様であり、やはり段階を踏んだ研究が求められる。そのような背景から、論説文という等質の文章でありながら、著者やジャンルは多様である『明六雑誌』は、形態素解析技術の適用事例として最初に取り組みのみに好適であると考えられる。

なお、コーパス化の対象とするのは『明六雑誌』全 43 号の全文とする。ただし、(1)表紙、(2)目次、(3)識語・奥付、(4)図表中の文字列、は対象外とする。

3．文字入力の基本仕様

3．1．基本方針

本文テキストの入力はすべて全角文字で行う。

原文の書記体が漢字片仮名交じりの場合、外来語といった一部の語を除き、片仮名を平仮名に変換して入力する。原文の書記体の種類の情報は article タグの属性として表し、片仮名のままとした文字列には span タグを付与する。

3．2．文字集合

使用する文字集合は、JIS X 0213 のうち、(1)康熙字典、(2)UCS 互換字、(3)CJK 統合漢字拡張 B に符号位置が割り当てられる文字、を除外した範囲とする。この範囲にない文字は外字として「≡」で入力し、g タグを付与し、タグの属性として文字の情報を表す。なお、(1)(2)にあたる文字は通用字形に包摂し、(3)にあたる文字は外字とする。

3．3．包摂規準

包摂規準については、JIS X 0213 のものに準拠する。

ただし、『明六雑誌』では JIS の包摂規準の適用できない字形差を持つ文字が多数出現し、それらをすべて外字として「≡」で入力するとコーパスの実用性を損ねかねない。そ

⁶ 近代語テキストへの形態素解析の現状と今後の見通しについては、本報告書に収録した小木曾智信「近代語テキストの形態素解析」を参照。

ここで、本コーパスでは独自に JIS の包摂規準を拡張したものを定義し、それによって字体包摂を行い、JIS 内字を用いて入力し、g タグを付与してタグの属性として拡張包摂規準の適用を表す。

また、この拡張包摂規準を適用してもなお外字となる文字についても、類似の意味・用法を持つ文字が JIS 内にある場合は、なるべくその文字で入力し、g タグを付与してタグの属性として原文の文字の情報を表す。

拡張包摂規準や別字代用の詳細については、本報告書に収録した須永哲矢「近代語文献を電子化するための異体字処理」を参照のこと。

3.4. 特殊な表記

ルビは、ルビの振られた文字列に ruby タグおよび lRuby タグを付与し、その rubyText 属性値によって表す。

割書された文字列は、warigaki タグを付与しその範囲を示す。

濁点の期待される文字に濁点が付いていない場合、濁点の付いた文字で入力し、vMark タグを付与する。

踊り字は、踊り字で繰り返される文字を入力し、odoriji タグを付与する。

漢字のよみを明らかにするために漢字の前後に小さく書かれた仮名や踊り字は、通常の入力とし、特にタグは付与しない。

JIS X 0213 外字の合字は、よみに対応する複数の仮名で入力し、特にタグは付与しない。

漢文に付与された振り仮名や返り点は入力対象外とする。

3.5. 空白

紙面に現れる空白は、常に空白 1 文字で入力する。ただし、レイアウト上複数行に渡って行われる字下げについては、論理行冒頭のみ空白 1 文字を入力する。

天皇等の高貴な人に敬意を表すために、その人に関連する語の直前に表記された空白(敬意欠字)は、空白を入力し、g タグを付与して敬意欠字であることを表す。

3.6. 誤植

誤植と思われる文字は、適切な文字に修正して入力し、corr タグを付与してタグの属性として原文の文字の情報を表す。

3.7. 判読困難な箇所

印刷のかすれや破損・抹消によって、文字の形がまったく残っておらず判読ができない場合、入力を行わず、gap タグを付与して文字の存在を表す。

文字の形が一部残り、元の文字が推測可能な場合、その文字を入力し、unclear タグを付与する。

4 . XML タグセット

本コーパスは、本文テキストに XML によって文書構造・単語・文字・表記に関する情報を付与する。そのための XML タグの一覧は、次の表 1 のとおりである。各タグで表される要素について続く各節で詳説する。なお、要素詳説であげる XML 例では、説明に不要なタグを省略して示す場合がある。

表 1 XML タグセット

タグ名	説明	詳説する節番号
magazine	雑誌 1 号分を表す。	4 . 1 .
front	雑誌中で前付けに相当する文書要素を表す。	4 . 2 .
body	雑誌中で中心本文に相当する文書要素を表す。	4 . 3 .
article	1 記事を表す。	4 . 4 .
titleBlock	前付け・中心本文の中にあり、記事とは認められない文書要素を表す。	4 . 5 .
p	記事中の段落に相当する文書要素を表す。	4 . 6 .
block	記事中にあり、段落とは見なせない文書要素(記事タイトル・記事著者・小見出し等)を表す。	4 . 7 .
figureBlock	図表を表す。	4 . 8 .
warigaki	割書された文字列を表す。	4 . 9 .
quotation	発話部分や他の文献からの引用を表す。	4 . 1 0 .
superS	引用や割書を含むため、複数の s 要素に分割された文を表す。	4 . 1 1 .
s	文を表す。	4 . 1 2 .
odoriji	踊字で表記されていたことを表す。	4 . 1 3 .
span	漢字片仮名交じり文の片仮名を平仮名に変換したテキストを作成する際、片仮名のまま残した文字列を表す。	0
gap	抹消・破損等で判読できない文字列の存在を表す。	4 . 1 5 .
pb	原本での改ページ位置を表す。	4 . 1 6 .
lb	原本での改行位置を表す。	4 . 1 7 .
SUW	語(短単位)を表す。	4 . 1 8 .
ruby	本行の右側に振られたルビを表す。	4 . 1 9 .
lRuby	本行の左側に振られたルビを表す。	4 . 2 0 .
corr	誤植を校訂したことを表す。	4 . 2 1 .
unclear	不鮮明ではあるが字体が推定できる文字を表す。	4 . 2 2 .
vMark	濁点無表記の文字を表す。	4 . 2 3 .
g	外字・敬意欠字等の特殊な文字を表す。	4 . 2 4 .
kanbun	漢文によって書かれた文章に返読・補読を行ったことを表す。	4 . 2 5 .

4 . 1 . magazine 要素

説明

雑誌 1 号分を表す。

属性

title (必須) : 雑誌名

year (必須) : 発行年

issue (必須) : 号番号

version (必須) : XML ファイルのバージョン

XML 例

```
<magazine title="明六雑誌" year="1874" issue="01" version="1.0">
<front>
( ... 中略... )
</front>
<body>
( ... 中略... )
</body>
</magazine>
```

4 . 2 . front 要素

説明

雑誌中で前付けに相当する文書要素を表す。本コーパスでは雑誌タイトルがこれに該当する。

属性

なし

XML 例

```
<magazine title="明六雑誌" year="1874" issue="01" version="1.0">
<front>
<titleBlock>
<block>
<s>明六社雑誌第一號</s>
</block>
</titleBlock>
</front>
<body>
( ... 中略... )
</body>
</magazine>
```

4 . 3 . body 要素

説明

雑誌中で中心本文となる文書要素を表す。本コーパスでは複数の記事からなる部分がこれに該当する。

属性

なし

XML 例

```
<magazine title="明六雑誌" year="1874" issue="01" version="1.0">
<front>
( ...中略... )
</front>
<body>
<article title="洋字を以て国語を書するの論" author="西周" style="文語" script="漢字カタカナ">
( ...中略... )
</article>
<article title="開化の度に因て改文字を発すべきの論" author="西村茂樹" style="文語" script="漢字カタカナ">
</body>
( ...中略... )
</article>
</magazine>
```

4 . 4 . article 要素

説明

中心本文中の各記事を表す。

属性

title (必須) : 記事の題名

author (必須) : 記事の著者名・訳者名

originalAuthor (任意) : 翻訳記事の原著者名

style (必須) : 記事の文体を表す。

- 文語...文語体。文末辞が「なり」「たり」「つ」「ぬ」「き」「けり」のもの。
- 口語...口語体。文末辞が「だ」「ぢや」「である」「です」「ます」のもの。
- 混在...文語体と口語体が混在するもの。

script (任意) : 記事の書記体

- 漢字カタカナ...漢字片仮名交じり
- 漢字ひらがな...漢字平仮名交じり

XML 例

```
<article title="人民の自由と土地の気候と互に相関するの論(一)" author="箕作麟祥" originalAuthor="モンテスキ  
ユウ" style="文語" script="漢字カタカナ">
<block>
( ...中略... )
</block>
<p>
( ...中略... )
</p>
<p>
( ...中略... )
</p>
</article>
```

4 . 5 . titleBlock 要素

説明

前付けまたは中心本文の中の、記事と同位の文書要素で、記事とは認められないものを表す。本コーパスでは前付け中の雑誌タイトルがこれに該当する。

属性

なし

XML 例

```
<front>
<titleBlock>
<block>
<s>明六社雑誌第一號</s>
</block>
</titleBlock>
</front>
```

4 . 6 . p 要素

説明

記事中の 1 段落を表す。

原則として、論理改行を段落末として段落を認定する。ただし、箇条書きのように論理改行による認定がふさわしくないと考えられる部分については、人手により段落の認定を行う。

属性

なし

XML 例

```
<p>
<s>西先生の改文字論を再三熟讀するに其論說痛快精到少しも遺憾なし</s><s>果して此言の如くなる ㊦を得ば實に文運の大進歩にして吾儕操觚者の最も愉快とする處なり</s>( ...中略... )<s>願くは諸先生の高論を以て左の件々を議定あらん ㊦を</s>
</p>
<p>
<s> 第一 會社の名</s><s> 第二 社中人員の定數</s><s> 第三 新に入社する人員を撰ぶの法</s>( ...中略... )<s> 第七 書記並びに掌計者を撰ぶ事</s><s> 第八 日誌出版の法</s>
</p>
<p>
<s> 本朝にて學術文藝の會社を結びしは今日を始めとす</s><s>而して社中の諸賢は皆天下の名士なり</s>( ...中略... )<s>何とぞ諸先生の卓識高論を以て愚蒙の眠を覺し天下の摸範を立て識者の望を曠ふせざらん ㊦を是折る</s>
</p>
```

4 . 7 . block 要素

説明

記事中、段落と同位の要素で、段落とは認められないものを表す。本コーパスでは記事

タイトル・記事著者表示・記事小見出しがそれに該当する。

属性

なし

XML 例

```
<article title="洋字を以て国語を書するの論" author="西周" style="文語" script="漢字カタカナ">
<block>
  洋字を以て國語を書するの論
</block>
<block>
  西周
</block>
<p>
  (...中略...)
</p>
</article>
```

4 . 8 . figureBlock 要素

説明

図表のある部分を表す。空要素。

属性

なし

XML 例

```
<p>
  (...中略...)
</p>
<figureBlock/>
<p>
  (...中略...)
</p>
```

4 . 9 . warigaki 要素

説明

割書となっている文字列を表す。

属性

なし

XML 例

```
その教と同派のものを信ずる某宗徒の爲に此徒を管轄する他國
<warigaki>
  此國も亦耶蘇教を奉ず但し別派なり
</warigaki>
の事に與聞せんと要するは則その理あり
```

4.10. quotation 要素

説明

記事・引用中で、その記事・引用とは発話者や発話場面・文体の異なる文書要素（他文献からの引用や会話・心話など）を表す。

属性

type（必須）：引用の種類

- 会話...会話
- 心話...心話
- 手紙...手紙
- 典拠...手紙を除く他文献からの引用
- 記事説明...記事に対し編集者等が説明を加えるための文書
- 韻文...漢詩を除く韻文
- 漢文...漢文によって書かれた文書。他の引用の種類に該当するものであっても、漢文の形式で書かれていれば type 属性値は「漢文」とする。

source（必須）：引用部分の話し手や書き手、引用元の書名等

style（任意）：引用の文体が上位 article 要素または quotation 要素の文体とは異なる場合の、文体の種類。ただし、type 属性値が「韻文」「漢文」の場合は不要。

- 文語...文語体。文末辞が「なり」「たり」「つ」「ぬ」「き」「けり」のもの。
- 口語...口語体。文末辞が「だ」「ぢや」「である」「です」「ます」のもの。
- 混在...文語体と口語体が混在するもの。

XML 例

例1 会話

```
或人曾て英公使の書記官サトウ氏に語て曰く
<quotation type="会話" source="或人">
英學頗る日本に行はる
</quotation>
とサトウ氏頭を掉て
<quotation type="会話" source="英公使書記官サトウ">
否米學なり
</quotation>
と言へりとぞ
```

例2 心話

```
蓋拷問の苦堪ふべからず常人は乃思へらく
<quotation type="心話" source="常人">
其拷問に苦しまんよりは寧冤罪に死せん
</quotation>
と
```

例 3 典拠

昨年十月の布告に新聞紙發行の條目中
<quotation type="典拠" source="新聞紙發行條目">
「國體を誹り國律を議し及び外法を主張宣議して國の妨害を生ぜしむるを禁ず」「政事法律等を記載するに付妄に批評を加ふるを禁ず」「猥りに教法を記入し政法の妨害を生ぜしむるを禁ずる」
</quotation>
等の箇條あり

例 4 記事説明

<quotation type="記事説明" source="箕作麟祥">
右佛國大學士「モンテスキュー」所著のスピリット、ヲフ、ロウスより抄譯す尚續譯して次号に出すべし
</quotation>

例 5 韻文

本居宣長の
<quotation type="韻文" source="本居宣長">
「式島の日本心を人間ば朝日に香ふ山櫻花」
</quotation>
と詠ぜしは即ち此易直の質を以て我が國民の氣風に烙記を居ゑたる者にて流石に夫れ者だけ能名状したる者と謂ふべし

例 6 漢文

邵康節の詩
<quotation type="漢文" source="邵康節">
尋常巷陌連羅綺、到處樓臺奏管絃、天下泰平無事日、鶯花無限日高眠、
</quotation>
といへるにても想見べし

4.11. superS 要素

説明

割書や引用を含むために、複数の文からなると見なされる 1 文を表す。

本コーパスでは、形態素解析での必要性から、warigaki 要素はその前後とは必ず別の s 要素と認定する。また、複数の s 要素からなる quotatin 要素も同様にその前後とは別の s 要素と認定する。よって、該当要素を含む 1 文は、1 文であるにもかかわらず複数の s 要素に分割される。これらの複数の s 要素をまとめ上げるのが superS 要素である。

属性

なし

XML 例

例 1

```
<superS>  
<s type="fragment">先生の御論にては内養</s>  
<warigaki><s>先生論ずる所即政府官吏の理治</s></warigaki>  
<s type="fragment">外刺</s>  
<warigaki><s>即人民の政府を刺衝するに</s></warigaki>  
<s type="fragment">相平均せざる可らざる内にも外刺を以て殊に緊要と被致候様に相見へ候</s>  
</superS>
```

例 2

```
<superS>
<s type="fragment">昨年十月の布告に新聞紙発行の條目中</s>
<quotation type="典拠" source="新聞紙発行条目"><s>「國體を誹り國律を議し及び外法を主張宣議して國の妨害
をせしむるを禁ず」</s><s>「政事法律等を記載するに付妄に批評を加ふるを禁ず」</s><s>「猥りに教法を
記入し政法の妨害をせしむるを禁ずる」</s></quotation>
<s type="fragment">等の箇條あり</s>
</superS>
```

4.12.s 要素

説明

1 文を表す。文の認定は人手により行う。

属性

type (任意) :

- fragment... 割書や引用を含むために、1 文であるにもかかわらず複数の s 要素に分割された結果生じた、文の一部を内容とする s 要素であることを表す

XML 例

例 1 通常の s 要素

```
<s> 其他語格の若きは後日の成功を待つべし</s>
<s>右聊か愚考を陳じ諸先生の可否を請ふ</s>
<s>敢て採用を望むにあらざると雖ども諸先生幸に電覽を賜はゞ幸甚</s>
```

例 2 type 属性値が「fragment」の s 要素

```
<superS>
<s type="fragment">その教と同派のものを信ずる某宗徒の爲に此徒を管轄する他國</s>
<warigaki>
<s>此國も亦耶穌教を奉ず</s><s>但し別派なり</s>
</warigaki>
<s type="fragment">の事に與聞せんと要するは則その理あり</s>
</superS>
```

4.13.odoriji 要素

説明

踊り字で表記されている箇所を表す。

踊り字が繰り返す文字列を odoriji 要素の内容とし、原文の踊り字は originalText 属性として入力する。ただし、短単位中で直前の 1 字を繰り返す「々」「と」は odoriji 要素とはせず、テキストを「々」「と」のままとする。

属性

originalText : 原文で使われている踊り字

XML 例

例 1 一字点

```
是僕尤恐る<odoriji originalText="ゝ">る</odoriji>所なり
```

例 2 二字点

```
天下ます<odoriji originalText="と">ます</odoriji>昌明の運に進み
```

例 3 同字点

```
果て<odoriji originalText="々">果て</odoriji>は
```

例 3 くの字点

```
顯國代る<odoriji originalText=" / \ ">代る</odoriji>興り
```

例 4 odoriji 要素としない同字点

```
奇々怪々の一案
```

4.14. span 要素

説明

本コーパスでは本文の書記体を漢字平仮名交じりに統一するため、原文では漢字片仮名交じりの文章は片仮名を平仮名に変換してテキストを作成する。しかし外来語といった片仮名表記のままのほうがよいと判断した文字列については、span 要素としてマークアップする。

属性

type (必須) :

- カタカナ...片仮名のまま残す文字列を表す

XML 例

例 1

```
<span type="カタカナ">アベセ</span>二十六字を知り
```

例 2

```
近日<span type="カタカナ">ヘボン</span>の字書又佛人<span type="カタカナ">ロニ</span>の日本語會あり
```

4.15. gap 要素

説明

抹消・損傷等により判読不可能な文字列の存在を表す。空要素。

属性

quantity (任意) : 該当文字列の文字数がわかる場合、その文字数

XML 例

```
然る時は又天地に彌り古今<gap quantity="2"/>き人の主として方向を定むべき者唯善のみ
```

4.16.pb 要素

説明

原文の紙面上での改ページ位置を表す。空要素。

属性

n (必須) : 該当位置から始まるページの番号。原本の丁付けで一丁表となるページから順に、開始の値を「1」とする連番を振る。

originalN (必須) : 該当位置から始まるページの、原本での丁番号と表裏の区別。例えば一丁表ならば属性値は「1オ」とする。

XML 例

```
<pb n="1" originalN="1オ">明六社雑誌第一號 (...中略...) 故に上旨は下達せず下情は上伸せずして全身不遂  
<pb n="2" original="1ウ">の人の如し (...中略...) 然ども此弊に因て斯世の民幸福を蒙るゝを得ず衰弊の極救薬  
すべからざるに至  
(...中略...)  
<pb n="24" original="12ウ"> 第二 社中人員の定數 (...中略...) 何とぞ諸先生の卓識高論を以て愚蒙の眼を覺  
し天下の模範を立て識者の望を曠ふせざらんゝを是祈る
```

4.17.lb 要素

説明

原文の紙面上での改行位置を表す。空要素。

属性

なし

XML 例

```
<pb n="1" originalN="1オ"> <lb/>明六社雑誌第一號  
<lb/> 洋字を以て國語を書するの論 西周  
<lb/>吾輩日常二三朋友の盃簞に於て偶當時治亂盛衰の故政治得失の跡な  
<lb/>ど凡て世故に就て談論爰に及ぶ時は動もすればかの歐洲諸國と比較  
<lb/>するゝの多かる中に終には彼の文明を羨み我が不開化を歎じ果て果ては
```

4.18.SUW 要素

説明

単語 (短単位) を表す。

本コーパスの SUW 要素は、近代の文語文を対象とする形態素解析辞書「近代文語 UniDic」

による解析結果を人手で修正したものである。SUW 要素の各属性の詳細については、「近代文語 UniDic」(<http://www2.ninjal.ac.jp/lrc/index.php?UniDic>)のユーザーズマニュアルや、「近代文語 UniDic」のもととなった現代語版「UniDic」(<http://download.unidic.org>)のユーザーズマニュアルを参照のこと。

属性

orthToken (必須) : 書字形出現形

IForm (任意) : 語彙素読み

lemma (任意) : 語彙素

subLemma (任意) : 語彙素細分類。区別がある場合のみ出力。

pos (必須) : 品詞

form (任意) : 語形

cType (任意) : 活用型。活用語のみ出力。

cForm (任意) : 活用形。活用語のみ出力。

pronToken (任意) : 発音形出現形

kanaToken (任意) : 仮名形出現形

orth (任意) : 書字形基本形。活用語のみ出力。

wType (任意) : 語種

start (必須) : 語の始まる文字位置

end (必須) : 語の終わる文字位置

originalText (任意) : 原文文字列。orthToken 属性値と異なる場合のみ出力。

orderID (必須) : 語の通し番号

BOS (任意) :

- True...文頭に現れる語であることを表す

XML 例

```
<SUW orthToken="洋字" lForm="ヨウジ" lemma="洋字" pos="名詞-普通名詞-一般" form="ヨウジ" pronToken="ヨージ" kanaToken="ヨウジ" orth="洋字" wType="漢" start="100" end="120" orderID="80" section="v">洋字</SUW>
<SUW orthToken="を" lForm="ヲ" lemma="を" pos="助詞-格助詞" form="ヲ" pronToken="オ" kanaToken="ヲ" orth="を" wType="和" start="120" end="130" orderID="90" section="v"></SUW>
<SUW orthToken="以" lForm="モツ" lemma="持つ" pos="動詞-一般" form="モツ" cType="文語四段-タ行" cForm="連用形-促音便" pronToken="モツ" kanaToken="モツ" orth="以つ" wType="和" start="130" end="140" orderID="100" section="v">以</SUW>
<SUW orthToken="て" lForm="テ" lemma="て" pos="助詞-接続助詞" form="テ" pronToken="テ" kanaToken="テ" orth="て" wType="和" start="140" end="150" orderID="110" section="v">て</SUW>
<SUW orthToken="國語" lForm="コクゴ" lemma="国語" pos="名詞-普通名詞-一般" form="コクゴ" pronToken="コクゴ" kanaToken="コクゴ" orth="國語" wType="漢" start="150" end="170" orderID="120" section="v">國語</SUW>
<SUW orthToken="を" lForm="ヲ" lemma="を" pos="助詞-格助詞" form="ヲ" pronToken="オ" kanaToken="ヲ" orth="を" wType="和" start="170" end="180" orderID="130" section="v">を</SUW>
<SUW orthToken="書する" lForm="シヨスル" lemma="書する" pos="動詞-一般" form="シヨス" cType="文語サ行変格" cForm="連体形-一般" pronToken="シヨスル" kanaToken="シヨスル" orth="書す" wType="混" start="180" end="210" orderID="140" section="v">書する</SUW>
<SUW orthToken="の" lForm="ノ" lemma="の" pos="助詞-格助詞" form="ノ" pronToken="ノ" kanaToken="ノ" orth="の" wType="和" start="210" end="220" orderID="150" section="v">の</SUW>
<SUW orthToken="論" lForm="ロン" lemma="論" pos="名詞-普通名詞-一般" form="ロン" pronToken="ロン" kanaToken="ロン" orth="論" wType="漢" start="220" end="230" orderID="160" section="v">論</SUW>
```

4.19. ruby 要素

説明

原本本行の文字列の右側に振られているルビを表す。

属性

rubyText (必須) : ルビとして振られた文字列

rubyBase (任意) : 複数の SUW 要素により構成される文字列に 1 つのルビが振られている場合、先頭の短単位を ruby 要素とする処理を行い、rubyBase 属性に実際にルビの振られている文字列を値として入力する。

XML 例

例 1

```
<ruby rubyText="サフロフ">候</ruby>文
```

例 2

```
<r rt="ケミストリ">化學</r>
```

例 3 複数の SUW 要素により構成される文字列に 1 つのルビが振られている場合

```
<ruby rubyText="コントラソシヤール" rubyBase="國民約束">國民</r>約束
```

4.20. IRuby 要素

説明

原本本行の文字列の左側に振られているルビを表す。

属性

rubyText (必須) : ルビとして振られている文字列

rubyBase (任意) : 複数の SUW 要素により構成される文字列に 1 つのルビが振られている場合、先頭の短単位を ruby 要素とする処理を行い、rubyBase 属性に実際にルビの振られている文字列を値として入力する。

XML 例

例 1

```
<IRuby rubyText="ボストン">波士頓</r>
```

例 2 複数の SUW 要素により構成される文字列に 1 つのルビが振られている場合

```
<IRuby rubyText="ヂウアイン" rubyBase="上帝道">上帝</IRuby>道
```

4 . 2 1 . corr 要素

説明

原文で誤植と見られる文字を表す。

該当文字が本行にある場合、適切に修正した文字を corr 要素の内容とし、originalText 属性に原文文字を値として入力する。type 属性値が「excess」の場合は空要素。

該当文字がルビにある場合、該当ルビを表す ruby 要素・IRuby 要素の rubyText 属性値に適切に修正した文字を入力し、corr 要素の originalText 属性値でルビ全体の原文文字列を表す。

属性

type (必須) : 誤植の種類

- erratum...誤字
- excess...衍字
- omission...脱字

originalText (任意) : 該当文字が本行にある場合、本行の原文文字を表す。ただし、type 属性値が「omission」の場合は不要。また、該当文字がルビにある場合、ルビ全体の原文文字列を表す。

subType (任意) :

- ruby...該当文字がルビにあることを表す。

XML 例

例 1 誤字 (本行)

```
<corr originalText="巳" type="erratum">已</corr>に
```

例 2 衍字（本行）

```
盡く其實を<corr originalText="ヲ" type="excess"/>明示し
```

例 3 脱字（本行）

```
改めざる可か<corr type="omission">ら</注>ず
```

例 4 衍字（ルビ）

```
<corr originalText="クゝ" type="excess" subType="ruby"><ruby rubyText="ク">喰</ruby></corr>つて見た上で
```

例 5 脱字（ルビ）

```
<corr originalText="そ" type="omission" subType="ruby"><ruby rubyText="ほそ">細</ruby></corr><ruby rubyText="ね">根</ruby>
```

4.22. unclear 要素

説明

原本の損傷等により不鮮明ではあるが字体の推定は可能な文字を表す。

属性

originalText（任意）：該当文字がルビにある場合は、該当文字を「 」で表記してルビ全体の文字列を表す。該当文字が本行にある場合は不要。

type（任意）：

- ruby...該当文字がルビにあることを表す。

XML 例

例 1 本行にある場合

```
公法大學生ヒリモア<unclear>及</unclear><unclear>び</unclear>ワツテル兩家
```

例 2 ルビにある場合

```
<unclear originalText=" イロソフィカル" type="ruby"><ruby rubyText="フィロソフィカル">哲理</ruby></unclear>
```

4.23. vMark 要素

説明

濁点の表記が期待されるにもかかわらず、原文では濁点無表記の仮名・記号が使われていることを表す。

本行に該当文字がある場合、濁点を表記した文字を vMark 要素の内容とする。

ルビに該当文字がある場合、そのルビを表す ruby 要素・lRuby 要素の rubyText 属性値には濁点を表記した文字を入力し、vMark 要素の originalText 属性値でルビ全体の原文文字列を表す。

属性

originalText (任意) : ルビに該当文字がある場合は、ルビの原文文字列を表す。本行に該当文字がある場合は不要。

type (任意) :

- ruby...ルビに該当文字があることを表す。

XML 例

例 1 本行の場合

```
談論爰に及<vMark>ふ</vMark>時は
```

例 2 ルビの場合

```
儲其可否は<vMark originalText="トウダ" type="ruby"><ruby rubyText="ドウダ">如何</ruby></vMark>と云った時は
```

4.24.g 要素

説明

JIS X 0213 外字や敬意欠字といった特殊な文字・記号を表す。

JIS X 0213 外字の漢字であるが拡張包摂規準により字体包摂を行う場合、包摂後の字体を入力して g 要素の内容とする。

JIS X 0213 外字で、かつ拡張包摂規準の適用外の漢字であるが、意味・用法の類似する他の漢字での代用が可能な場合、その代用字を入力して g 要素の内容とする。

字体包摂も代用字での代用もしない JIS X 0213 外字の場合、「**ニ**」を入力して g 要素の内容とする。

天皇等の高貴な人に敬意を表すために、その人に関連する語の直前に表記された空白(敬意欠字)の場合は、空白を入力して g 要素の内容とする。

属性

type (必須) :

- 外字...JIS X 0213 外字で、かつ拡張包摂規準の適用外の文字であることを表す。
- 包摂...JIS X 0213 外字であるが、拡張包摂規準により JIS X 0213 内字に包摂した文字であることを表す。
- 敬意欠字...天皇等の高貴な人に敬意を表すために、その人に関連する語の直前に表記された空白であることを表す。

ref(任意) : type 属性値が「外字」の場合、Unicode4.0 の 16 進コードがあるものは「U+」を先頭に加えた文字列を値とし、Unicode 外字の場合は字体記述を値とする。type 属性値が「包摂」「敬意欠字」の場合は不要。

XML 例

例 1 外字 (Unicode 内字)

```
<g type="外字" ref="U+9AF2">𠄎</g>
```

例 2 外字 (Unicode 外字)

```
<g type="外字" ref="衣+丸">𠄎</g>
```

例 3 外字 (代用字による入力)

```
<g type="外字" ref="U+7FA1">羨</g>
```

例 4 包摂

```
<g type="包摂">時</g>
```

例 5 敬意欠字

```
我大日本<g type="敬意欠字"> </g>天皇陛下の特詔を垂れて
```

4.25. kanbun 要素

説明

漢文によって日本語を書き表したと見なされる文字列を訓読するために、返読・補読を行ったことを表す。

ただし、漢籍の引用など日本語を書き表したと見なされない漢文には返読・補読は行わず、quotation 要素とする。

type 属性が「返読前」のものは空要素。

属性

type (必須) :

- 返読前...返読の対象となる文字が、本来あった位置。
- 返読後...返読の対象となる文字が、返読後に移動した位置。
- 補読...訓読によって補読された語。

originalText (任意) : 原文の文字。type 属性値が「返読前」のものは必須。type 属性値が「返読後」「補読」のものは不要。

id (任意) : 返読の対象となる文字の、返読前の位置と返読後の位置を対照するために与えられた XML ファイル内固有の ID。type 属性値が「返読前」「返読後」のものは必須。type 属性値が「補読」のものは不要。

XML 例

例 1 返読

```
此段宜敷御評議を<kanbun type="返読前" originalText="可" id="00001"/><kanbun type="返読前" originalText="被" id="00002"/>遂<kanbun type="返読後" id="00002">被</kanbun><kanbun type="返読後" id="00001">可</kanbun>候<也
```

例 2 補読

```
然ども是<kanbun type="返読前" originalText="不" id="00008"/><kanbun type="返読前" originalText="得" id="00009"/>已<kanbun type="補読">を</kanbun><kanbun type="返読後" id="00009">得</kanbun><kanbun type="返読後" id="00008">不</kanbun>の時なり
```

5. コーパスの公開形式

本コーパスの公開形式は以下の 3 種類である。

5.1 XML ファイル

本文テキストに XML タグによって文書構造・形態論・文字・表記に関する情報を付与した形式で、コーパスの根幹となるデータである。

1号1ファイルとし、全43ファイルからなる。XML ファイルの符号化形式は UTF-8 (BOM なし) である。ファイル名は「m」に続く 4 桁の数字が該当号の刊行年を、次の 2 桁の数字が号番号を表す。例えばファイル名が「m187401.xml」ならば、1874 年刊行の 1 号のデータを収めた XML ファイルということになる。

5.2 形態論情報タブ区切りデータ

XML ファイルから特に SUW 要素に関する情報を抽出し、タブ区切りのデータに成形したものである。ファイル名は「merioku_suw.txt」、符号化形式は UTF-8 (BOM なし) である。1 行目はフィールド名を入力した行で、2 行目以降から 1 行が 1SUW 要素に対応している。

データのフィールドリストを表 2 として示す。

表 2 フィールドリスト

フィールド名	備考
コーパス名	
ファイル名	XML ファイル名に対応
記事題名	article 要素 title 属性に対応
記事著者	article 要素 author 属性に対応
記事原著者	article 要素 originalAuthor 属性に対応
記事文体	article 要素 style 属性に対応
記事書記体	article 要素 script 属性に対応
語連番	SUW 要素 orderID 属性に対応
文字開始位置	SUW 要素 start 属性に対応
文字終了位置	SUW 要素 end 属性に対応
文頭ラベル	SUW 要素 BOS 属性に対応 (B: 文頭、I: 文頭以外)

語彙表 ID	書字形出現形レベルで語を識別する ID
語彙素 ID	語彙素レベルで語を識別する ID
語彙素読み	SUW 要素 lForm 属性に対応
語彙素	SUW 要素 lemma 属性に対応
語彙素細分類	SUW 要素 subLemma 属性に対応
語種	SUW 要素 wType 属性に対応
品詞	SUW 要素 pos 属性に対応
活用型	SUW 要素 cType 属性に対応
活用形	SUW 要素 cForm 属性に対応
語形	SUW 要素 form 属性に対応
書字形基本形	SUW 要素 orth 属性に対応
書字形出現形	SUW 要素 orthToken 属性に対応
原文文字列	SUW 要素 originalText 属性に対応
発音形出現形	SUW 要素 pronToken 属性に対応

5.3 「ひまわり」用データ

文字列検索システム「ひまわり」用のデータである。このデータを「ひまわり」にインストールすることで、GUIによる簡便なコーパスの検索が可能となる。

5.3.1 「ひまわり」へのインストール方法

データの「ひまわり」へのインストールは次の ~ の手順で行う。

ダウンロードしたデータを解凍すると、「meiroku_himawari」フォルダが現れる。その中に次のファイルがあることを確認する。


- Corpora フォルダ...『明六雑誌コーパス』データを格納したフォルダ
- config_meiroku.xml...設定ファイル

「ひまわり」ver.1.3 をインストールする。国立国語研究所「言語データベースとソフトウェア」のホームページ (<http://www2.ninjal.ac.jp/lrc/index.php>) の画面左にある「メニュー」「ソフトウェア」「全文検索システム『ひまわり』」をクリックすると、「ひまわり」のページに移動する。そこに書かれた説明に従い「ひまわり」ver.1.3 のインストールを行う。

「ひまわり」ver.1.3 をインストールすると「Himawari_1_3」フォルダが現れる。その中に、 の「Corpora」フォルダと「config_meiroku.xml」を移動する。その際、コンピュータ環境によっては「Corpora」フォルダの上書きの確認のメッセージが表示される場合があるが、そのまま上書きを許可してよい。

5.3.2 「ひまわり」を使ったコーパスの検索方法

「ひまわり」にインストールしたコーパスデータの基本的な検索方法を説明する。

「Himawari_1_3」フォルダ内の「himawari.exe」(アイコン ) をダブルクリックすると「ひまわり」の起動画面が開く(図1)。

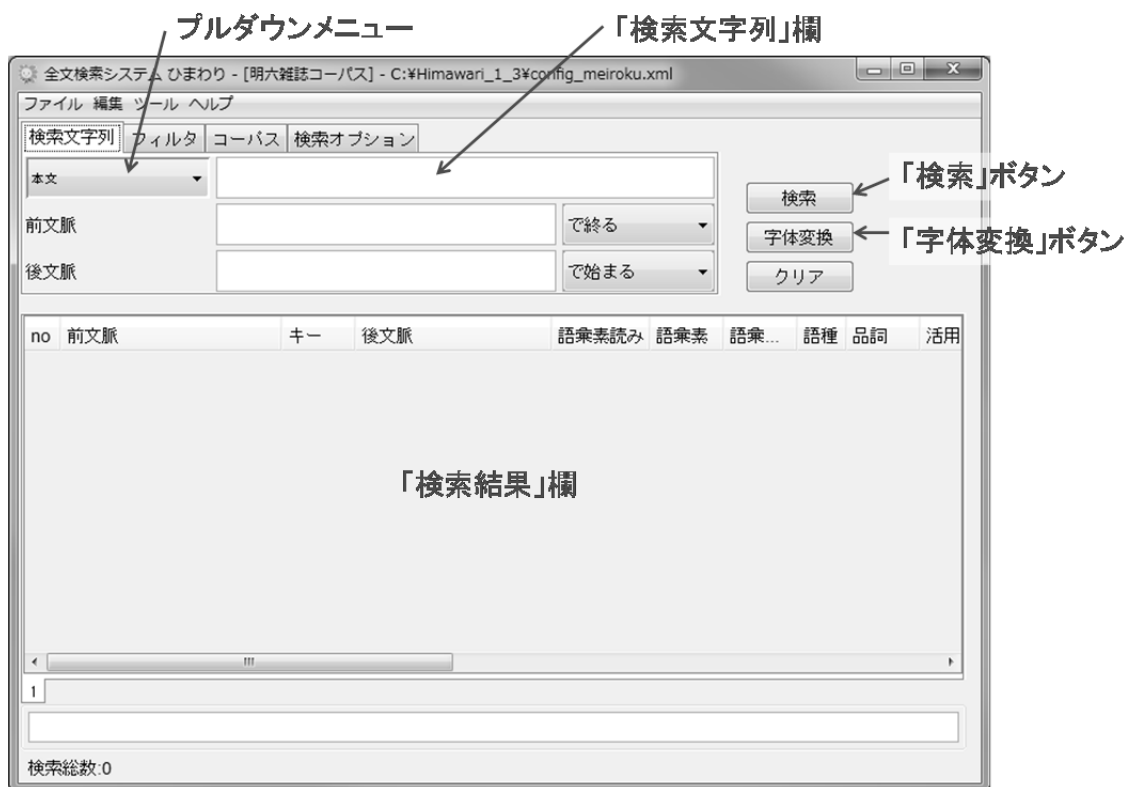


図1 「ひまわり」の起動画面

次に、画面上部の「ファイル」メニュー 「新規」を選択する（図2）。すると、設定ファイルを指定するための画面が現れるので、「config_meiroku.xml」を選択する。

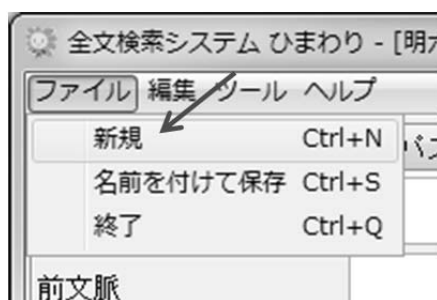


図2 「ファイル」メニュー 「新規」

次に「プルダウンメニュー」（図1参照）で検索対象を指定する。検索対象のリストを表3としてあげる。なお、プルダウンメニューに表示される「完全一致」「部分一致」は検索対象と検索文字列との照合方法を表す。

表3 「ひまわり」検索対象リスト

プルダウンメニュー表示	検索対象
本文	本文テキスト部分

右ルビ/完全一致	ruby 要素 rubyText 属性値
右ルビ/部分一致	
左ルビ/完全一致	lRuby 要素 rubyText 属性値
左ルビ/部分一致	
語彙素/完全一致	SUW 要素 lemma 属性値
語彙素読み/完全一致	SUW 要素 lForm 属性値
語種/完全一致	SUW 要素 wType 属性値
品詞/部分一致	SUW 要素 pos 属性値
活用型/部分一致	SUW 要素 cType 属性値
活用形/部分一致	SUW 要素 cForm 属性値
語形/完全一致	SUW 要素 form 属性値
書字形基本形/部分一致	SUW 要素 orth 属性値

次に「検索文字列」欄（図 1 参照）に検索したい文字列を入力する。「字体変換」ボタン（図 1 参照）をクリックすると、入力文字列に異体字がある場合は異体字を含めた検索ができるように「検索文字列」欄の入力が変換される。そして「検索」ボタン（図 1 参照）をクリックすると「検索結果」欄（図 1 参照）に検索結果が KWIC 形式で表示される（図 3）。



図 3 「ひまわり」での検索結果表示

「検索結果」欄に表示される列のリストを表 4 として示す。

表4 「ひまわり」検索結果列リスト

列名	備考
前文脈	
キー	
後文脈	
語彙素読み	SUW 要素 lForm 属性に対応
語彙素	SUW 要素 lemma 属性に対応
語彙素細分類	SUW 要素 subLemma 属性に対応
語種	SUW 要素 wType 属性に対応
品詞	SUW 要素 pos 属性に対応
活用型	SUW 要素 cType 属性に対応
活用形	SUW 要素 cForm 属性に対応
語形	SUW 要素 form 属性に対応
書字形基本形	SUW 要素 orth 属性に対応
雑誌名	magazine 要素 title 属性に対応
年	magazine 要素 year 属性に対応
号	magazine 要素 issue 属性に対応
ページ	pb 要素 originalN 属性に対応
語連番	SUW 要素 orderID に対応
記事題名	article 要素 title 属性に対応
記事著者	article 要素 author 属性に対応
記事著者	article 要素 originalAuthor 属性に対応
記事文体	article 要素 style 属性に対応
引用種類	quotation 要素 type 属性に対応
引用ソース	quotation 要素 source 属性に対応
引用文体	quotation 要素 style 属性に対応

よりひろい文脈で検索結果を閲覧したい場合は、「検索結果」欄のセルをダブルクリックする。Web ブラウザが起動し、雑誌単位での閲覧ができる。閲覧表示スタイルは次の3種類がある。

- 本文 + 付加情報 (図4)
- 本文 (図5)
- 形態論情報リスト (図6)

閲覧表示スタイルの切り替えは「ひまわり」起動画面の「ツール」メニュー—「オプション」—「閲覧表示スタイル」から行うことができる。



図4 「本文 + 付加情報」スタイルでの文脈表示



図5 「本文」スタイルでの文脈表示

明六雑誌02号(1874年)

語連番	ページ	文頭	書字形出現形	語彙素読み	語彙素	語彙素細分類	語種	品詞	活用型	活用形	語形	発音形出現形	引用種類	引用ソース	引用文体
10	1オ	B	明六	メイロク	明六		固	名詞-固有名詞-一般			メイロク	メイロク			
20	1オ	I	雑誌	ザッシ	雑誌		漢	名詞-普通名詞-一般			ザッシ	ザッシ			
30	1オ	I	第	ダイ	第		漢	接頭辞			ダイ	ダイ			
40	1オ	I	二	ニ	二		漢	名詞-数詞			ニ	ニ			
50	1オ	I	號	ゴウ	号		漢	名詞-普通名詞-助数詞可能			ゴウ	ゴウ			
60	1オ	B	福澤	フクザワ	フクザワ		固	名詞-固有名詞-人名・姓			フクザワ	フクザワ			
70	1オ	I	先生	センセイ	先生		漢	名詞-普通名詞-一般			センセイ	センセイ			
80	1オ	I	の	ノ	の		和	助詞-格助詞			ノ	ノ			
90	1オ	I	學者	ガクシャ	学者		漢	名詞-普通名詞-一般			ガクシャ	ガクシャ			
100	1オ	I	職分	シヨクブン	職分		漢	名詞-普通名詞-一般			シヨクブン	シヨクブン			
110	1オ	I	論	ロン	論		漢	名詞-普通名詞-一般			ロン	ロン			
120	1オ	I	は	ハ	は		和	助詞-係助詞			ハ	ハ			
130	1オ	I	慶應	ケイオウ	慶応		固	名詞-固有名詞-一般			ケイオウ	ケイオウ			
140	1オ	I	藝塾	ギョク	藝塾		漢	名詞-普通名詞-一般			ギョク	ギョク			

図6 「形態論情報リスト」スタイルでの文脈表示

「ひまわり」の利用方法の詳細については、「ひまわり」の利用者マニュアル(「ひまわり」起動画面の「ヘルプ」メニュー「『ひまわり』マニュアル」)を参照のこと。

文 献

- 神奈川大学人文学研究所(2004)『「明六雑誌」とその周辺—西洋文化の受容・思想と言語—』(御茶の水書房)
- 高野繁男・日向敏彦(1998)『明六雑誌語彙索引 付・復刻版「明六雑誌」』(大空社)
- 高野繁男(2004)『近代漢語の研究 日本語の造語法・訳語法』(明治書院)
- 田中牧郎(2005)「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」(国立国語研究所『雑誌「太陽」による確立期現代語の研究 「太陽コーパス」研究論文集』博文館新社)
- 山室信一・中野目徹(1999-2009)『明六雑誌(上)(中)(下)』(岩波文庫)