『国民之友コーパス』解説書

第 1.1 版

近藤 明日子　（国立国語研究所）

2014

目次

第1章　『国民之友コーパス』の開発経緯と特徴

『国民之友コーパス』は、近代の雑誌『国民之友』の 1887（明治 20）～1888（明治 21）年刊行分である 1～36 号の全文をコーパス化したものである。国立国語研究所共同研究プロジェクト「通時コーパスの設計」[1]（2009-、プロジェクトリーダー：田中牧郎）が中心となって開発を進めた。

これまで国立国語研究所では、『太陽コーパス』[2]（2005）、『近代女性雑誌コーパス』[3]（2006）、『明六雑誌コーパス』[4]（2012-）の 3 種の近代雑誌コーパスを開発・公開しており、『国民之友コーパス』はそのコーパス群の一角をなすものとして設計されている。特に、1895（明治 28）・1901（明治 34）・1909（明治 42）・1917（大正 6）・1925（大正 14）年刊行分をコーパス化した『太陽コーパス』と併せて用いることで、明治中期から大正期にかけての書き言葉の変化を 6～8 年間隔で観察することが可能となる。

原資料である雑誌『国民之友』は、徳富蘇峰の設立した民友社により 1887（明治 20）年から 1898（明治 31）年にかけて刊行された。主に、徳富蘇峰ら民友社社員および当時の著名知識人による政治・社会・経済・文学等の評論や文学作品を掲載する。その執筆者は幅広く、コーパス化の対象である 1887・1888 年刊の 36 号分だけ見ても、高橋五郎・森田思軒・朝比奈知泉・久松義典・依田学海・宇川盛三郎など 80 名以上に上る。また、発行部数は当時発行された雑誌のなかでは第 1 位であり、全国の知識層に広く普及し、彼らの言論活動に大きな影響を与えた雑誌であった（有山、1986）。近代の書き言葉の形成過程を知る上で重要度が高く、近代雑誌コーパス群には欠かせない資料としてコーパス化の対象として選択された。

コーパスの仕様は、『明六雑誌コーパス』の仕様（近藤・田中、2012）を引き継ぎつつ、それを整備・拡張させたものとなっている。『明六雑誌コーパス』は将来的に開発が想定される大規模な近代語コーパスのモデルとして設計され、形態論情報をはじめとする高密度な情報付与や原本画像の参照機能の実装など、開発当時の最新の研究成果を反映した仕様を持つ。ただし、コーパス自体の規模は延べ語数 18 万語と比較的小規模であった。『国民之友コーパス』は、『明六雑誌コーパス』で示された大規模コーパスのモデルを、実際に規模の大きいコーパスに適用した最初の例と言える。

こうして開発された『国民之友コーパス』は、延べ語数 101 万語（記号類除く）と、形態論情報の付与された近代語コーパスとしては公開時点において最大規模のものとなった。これによりコーパスを活用した近代語の研究がさらに進展することを期待したい。

『国民之友コーパス』の開発に携わったスタッフは次のとおりである（括弧内は開発当

[1] http://www.ninjal.ac.jp/research/project/a/corpus/
[2] 国立国語研究所（編）（2005）
[3] http://www.ninjal.ac.jp/corpus_center/cmj/woman-mag/
[4] http://www.ninjal.ac.jp/corpus_center/cmj/meiroku/

時の国立国語研究所職名。途中で転任したものも含む)。

- 開発担当者
  近藤明日子　（プロジェクト非常勤研究員）

- 開発協力者
  小木曽智信　高田智和　田中牧郎　（専任研究者）
  鴻野知暁　須永哲矢　間淵洋子　（プロジェクト非常勤研究員）
  木川あづさ　田口久美子　服部紀子　（技術補佐員）

第2章　『国民之友コーパス』の仕様

1　コーパス化の対象

コーパス化の対象とするのは、国立国語研究所蔵本『国民之友』1～36 号（1887・1888 年刊）[5]の全文である。ただし、次のものはコーパス化の対象外とする。

(1)　表紙
(2)　目次
(3)　識語・奥付
(4)　誤植の訂正記事
(5)　広告

2　コアデータの設定

コーパス全体の延べ語数の約 3%に相当する 24 記事、3 万 5 千語を「コア」データとして、「コア」以外のデータよりも高精度な情報を付与する。特に、①濁点無表記文字の濁点付き文字への校訂、②文境界位置、③形態論情報、に関して、コアデータではツール等による自動処理の後にすべて人手修正を行っており、精度が高い。一方、コアデータ以外は、自動処理の後、時間の許す範囲で人手修正を行ったものとなっている。

コアデータのサンプリングは記事単位とし、全記事を文体により「文語」「口語」の 2 層、ジャンルにより「小説・戯曲・詩歌」「それ以外」の 2 層、計 4 層に層別化し、各層から同量のデータが得られるようランダムサンプリングを行った。

コアデータであることは XML（「4　XML タグセット」で詳説）の article タグの core 属性で表す。

3　文字入力

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。

3.2　文字集合

使用する文字集合は、JIS X 0213 のうち、(1)康熙別掲字、(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。

この文字集合にない外字であっても、コーパスの利便性を考え、できるだけ集合内の文字で入力する方針とする（「3.3　外字」で詳説）。

[5] 原本画像は http://dglb01.ninjal.ac.jp/ninjaldl/bunken.php?title=kokuminnotomo で公開されている。

3.3　外字

3.3.1　非漢字の外字

合字の外字は、合字の表す複数の仮名で入力し、特にタグは付与しない。

類似の意味・用法を持つ文字が集合内にある場合は、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。

以上の処理を行ってもなお外字となる文字は、「**=**」（面区点番号：1-02-14、Unicode コード：U+3013）で入力し、g タグを付与してタグの属性として文字の情報を表す。

3.3.2　漢字の外字

JIS X 0213 の包摂規準を適用できる場合は、それにより字体包摂を行い、集合内の漢字で入力する（康熙別掲字と UCS 互換字はこれにより集合内の漢字で入力する）。特にタグは付与しない。

JIS X 0213 の包摂規準は適用できないが、集合内の漢字とは微細な字体差しかない場合は、コーパス用に独自に定義した追加包摂規準[6]を適用して字体包摂を行い、集合内の漢字で入力する。特にタグは付与しない。

JIS X 0213 の包摂規準および追加包摂規準は適用できないが、類似の意味・用法を持つ漢字が集合内にある場合は、その漢字で入力し、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。

以上の処理を行ってもなお外字となる漢字は、「**=**」（面区点番号：1-02-14、Unicode コード：U+3013）で入力し、g タグを付与してタグの属性として文字の情報を表す。

3.4　特殊な表記

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各種タグの script 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。

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。

濁点の期待される仮名に濁点が付いていない場合、濁点の付いた仮名で入力し、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。

---

[6] 『国民之友コーパス』の追加包摂規準については、須永・堤・近藤ほか（2013）を参照のこと。

3.5　空白

紙面に現れる空白は、その大きさにかかわらず常に全角スペース（面区点番号：1-01-01、Unicode コード：U+3000）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、論理行冒頭のみに全角スペース 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、g タグを付与する。

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 タグを付与しタグの属性として原文の文字の情報を表す。

3.7　漢文体・候文体

漢文体・候文体は、必要な返読・補読・仮名開きを行ったうえで入力し、返読・補読・仮名開きを行った部分に corr タグを付与し、タグの属性として原文の文字の情報を表す。

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は訓読せず、そのまま入力する。

3.8　判読困難な文字列

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は、「␣」（面区点番号：1-07-93、Unicode コード：U+2423）で入力する。

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。

3.9　入力対象外の文字列

図表中の文字列は入力対象外とし、figureBlock タグにより図表の存在を表す。

非日本語（外国語・漢文）からなる段落中の文字列は入力対象外とし、rejectedBlock タグにより段落の存在を表す。

文字の傍らに付けられた傍点・傍線や返り点は入力対象外とし、特にタグは付与しない。

4　XML タグセット

4.1　概要

『国民之友コーパス』は、本文テキストに XML によって文書構造・形態論・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は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。	4.2（p.6～）
front	雑誌中で前付けに相当する文書要素を表す。	4.3（p.7～）
body	雑誌中で中心本文に相当する文書要素を表す。	4.4（p.7～）
back	雑誌中で後付けに相当する文書要素を表す。	4.5（p.8～）
article	記事を表す。	4.6（p.9～）
titleBlock	記事と同位の文書要素で、記事とは見なせないものを表す。	4.7（p.10～）
p	段落に相当する文書要素を表す。	4.8（p.10～）
block	段落と同位の文書要素で、段落とは見なせないものを表す。	4.9（p.11～）
figureBlock	図表の存在を表す。	4.10（p.12～）
rejectedBlock	非日本語（外国語・漢文）からなる段落の存在を表す。	4.11（p.12～）
warigaki	割書されている文字列を表す。	4.12（p.13～）
quotation	上位要素とは発話者や発話場面の異なる文書要素が引用されている部分を表す。	4.13（p.13～）
superS	引用や割書を含むため、複数の s 要素からなると見なされる文を表す。	4.14（p.14～）
s	文を表す。	4.15（p.15～）
odoriji	踊り字で表記されている箇所を表す。	4.16（p.16～）
span	特にマークアップする必要のある文中の文字列を表す。	4.17（p.17～）
pb	原本での改ページ位置を表す。	4.18（p.18～）
cb	原本での改段位置を表す。	4.19（p.18～）
lb	原本での改行位置を表す。	4.20（p.19～）
br	論理改行を表す。	4.21（p.19～）
SUW	語（短単位）を表す。	4.22（p.19～）
ruby	本行の右側に振られているルビを表す。	4.23（p.21～）
lRuby	本行の左側に振られているルビを表す。	4.24（p.22～）
corr	原文の文字に修正を施し、異なる文字としたものを表す	4.25（p.23～）
unclear	不鮮明ではあるが字体が推定できる文字を表す。	4.26（p.25～）
vMark	濁音を表記するにもかかわらず、原文では濁点のない仮名が使われていることを表す。	4.27（p.25～）
g	外字・敬意欠字を表す。	4.28（p.26～）

4.2　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="国民之友" year="1888" issue="21" version="1.0">
<front>
 （…中略…）
</front>
<body>
 （…中略…）
</body>
<back>
 （…中略…）
</back>
</magazine>
```

該当箇所原本画像（21 号 1 ページ～）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=21&pb=1

4.3 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="国民之友" year="1888" issue="21" version="1.0">
<front>
  國民之友第二十一號
  （明治廿一年五月四日第一金曜日發兌）
</front>
<body>
 （…中略…）
</body>
<back>
 （…中略…）
</back>
</magazine>
```

該当箇所原本画像（21 号 1 ページ～）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=21&pb=1

4.4 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="国民之友" year="1888" issue="21" version="1.0">
<front>
 （…中略…）
</front>
<body>
<titleBlock>
  國民之友
</titleBlock>
<article title="政治上の分業" author="＊" ranmei="國民之友" style="文語" script="ひらがな">
  政治上の分業
太古曚昧野蠻の世に於ては、一人若くは一部の人にして、帝王ともなり、醫者ともなり、大將ともなり、僧侶ともなり、占卜者どもなり、裁判官ともなりたるが如きの例甚だ少なしとせず、
 （…中略…）
</article>
<article title="地方官の淘汰" author="＊" ranmei="國民之友" style="文語" script="ひらがな">
 （…中略…）
</article>
 （…中略…）
</body>
<back>
 （…中略…）
</back>
</magazine>
```

該当箇所原本画像（21 号 1 ページ～）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=21&pb=1

4.5　back 要素

説明

雑誌中で後付けに相当する文書要素を表す。本コーパスでは社告内の記事的文章がこれに該当する。

属性

なし

XML 例

```
<magazine title="国民之友" year="1888" issue="21" version="1.0">
<front>
 （…中略…）
</front>
<body>
 （…中略…）
</body>
<back>
<article title="社告" author="＊" ranmei="" style="文語" script="ひらがな">
 （…中略…）
</article>
</back>
</magazine>
```

該当箇所原本画像（21 号 47 ページ～）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=21&pb=47

4.6　article 要素

説明

記事を表す。

属性

title（必須）：記事の題名

author（必須）：記事の著者名。翻訳記事の場合は訳者名。原本に記載される呼称のままではなく、現代の辞典類で一般的な呼称に統一する。著者が複数の場合は、全角スペースにより区切り列挙する。無署名の場合は、「＊」（面区点番号：1-01-86、Unicode コード：U+002A）を入力する。

originalAuthor（任意）：翻訳記事の原著者名

style（必須）：記事の地の文の文体を表す。取り得る属性値は次のとおり。

- 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。
- 項目…文末辞がなく、文語体か口語体か定められないもの。
- 漢文…漢文。
- 外国語…外国語（漢文を除く）。
- 韻文…日本語による韻文。
- 万葉…万葉仮名を使用するもの。

ranmei（任意）：記事が所属する欄名

script（必須）：記事の書記体。取り得る属性値は次のとおり。

- カタカナ…漢字片仮名交じり
- ひらがな…漢字平仮名交じり

core（任意）：

- true…コアデータであることを表す。

XML 例

```
<article title="あいびき（一）" author="二葉亭四迷" originalAuthor="ツルゲーネフ" ranmei="藻鹽草" style="口語"
 script="ひらがな" core="true">
<block>
  あいびき
</block>
 （…中略…）
<p>
秋九月中旬といふころ、一日自分がさる樺の林の中に座してゐたことが有ツた。（…中略…）
</p>
 （…中略…）
</article>
```

該当箇所原本画像（25 号 54 ページ～）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=25&pb=54

4.7　titleBlock 要素

説明

記事と同位の文書要素で、記事とは見なせないものを表す。本コーパスでは雑誌タイトル・欄タイトル、欄や複数の記事に対する説明部分がこれに該当する。

属性

script（任意）：書記体。取り得る属性値は「4.6　article 要素」の script 属性を参照のこと。

XML 例

例 1　欄名

```
<titleBlock>
  時事
</titleBlock>
<article title="市制及び町村制" author="＊" ranmei="時事" style="文語" script="ひらがな">
（…中略…）
</article>
```

該当箇所原本画像（21 号 40 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=21&pb=40

例 2　欄に対する説明

```
<article title="歐洲事件" author="＊" ranmei="時事" style="文語" script="ひらがな">
（…中略…）
</article>
<titleBlock>
（以上時事六月三十日脱稿）
</titleBlock>
```

該当箇所原本画像（25 号 40 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=25&pb=40

4.8　p 要素

説明

段落に相当する文書要素を表す。

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

style（任意）：文体の種類。上位要素と文体が異なる場合に必要。取り得る属性値は「4.6　article 要素」の style 属性を参照のこと。

script（任意）：書記体。上位要素と書記体が異なる場合に必要。取り得る属性値は「4.6

article 要素」の script 属性を参照のこと。

XML 例

```
<p>
<s>我が明治政府は、今ま明治廿一年四月廿五日を以て、市制及町村制を發布せられたり、</s>（…中略…）<s>此法律の趣旨を了解せんには、先づ第一に自治と稱ふる者を了解するを要するものとすべし、<br/></s>
</p>
<p>
<s>自治とは自ら獨立して自分の事を治むるの謂なり、</s>（…中略…）<s>故に丁年以上の者にして自治の權なきものは、不完全のもの、即ちカタワ者なりと謂ふ可し、<br/></s>
</p>
<p>
<s>人相集りて家を成し、又は町村をなし、又は府縣をなし又は國をなすときは、此等の集合躰に於て、共同の利益を有すること少からざるなり、</s>（…中略…）<s>此無形人中には法律を以て創造せられ、又は認定せらるるものあるが故に、之を法律上の人即ち法人とも稱するなり、<br/></s>
</p>
```

該当箇所原本画像（21 号 11 ページ～）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=21&pb=11

4.9 block 要素

説明

段落と同位の要素で、段落とは認められないものを表す。本コーパスでは雑誌タイトル・欄タイトル・記事タイトル・記事著者表示・記事小見出し等がそれに該当する。

属性

style（任意）：文体の種類。上位要素と文体が異なる場合に必要。取り得る属性値は「4.6 article 要素」の style 属性を参照のこと。

script（任意）：書記体。上位要素と書記体が異なる場合に必要。取り得る属性値は「4.6 article 要素」の script 属性を参照のこと。

XML 例

```
<titleBlock>
<block>
  特別寄書
</block>
</titleBlock>
<article title="地方自治を論じ併て市制町村制を論ず（一）" author="宇川盛三郎" ranmei="特別寄書" style="文語" script="ひらがな">
<block>
  地方自治を論じ併て市制町村制を論ず（其一）
</block>
<block>
  宇川盛三郎
</block>
<p>
（…中略…）
</p>
</article>
```

該当箇所原本画像（21 号 11 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=21&pb=11

4.10　figureBlock 要素

説明

図表の存在を表す。空要素。

属性

なし

XML 例

```
<p>
（…中略…）今試に近日出版したる日本支局萬國福音同盟會の調査したる統計中にて重なる教會の統計表を摘記すれば則ち左の如し
</p>
<figureBlock/>
<p>
勿論或る確なる人より聞けば組合教會の如きは教會の數三十一とあれども其の實は三十九若しくは四十なる可しとのことなれば或は間まかかる誤謬なきにもあらざる可れども其の大体に於ては必らず信ずるに足らん（…中略…）
</p>
```

該当箇所原本画像（2 号 16 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=02&pb=16

4.11　rejectedBlock 要素

説明

漢文や外国語のような非日本語からなる段落の存在を表す。空要素。

属性

type（必須）：段落の種別

- kanbun…漢文
- foreign…外国語

lineN（必須）：段落が占める行数

XML 例

```
<p>
（…中略…）所謂「敷島の大和心を人問はば、朝日に香ふ山櫻かな」
</p>
<rejectedBlock  type="foreign"  lineN="5"/>
<p>
とは偶然自然に發揮したる感情を直寫したるものに非らずや、（…中略…）
</p>
```

該当箇所原本画像（10 号 15 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=10&pb=15

4.12　warigaki 要素

説明

割書されている文字列を表す。

属性

なし

XML 例

```
ラザレフ村より五露里の間は雑木繁茂する山地に因て通し此山地を下りてダウビホザ河
<warigaki>
廣さ五十「サーゼニ」
</warigaki>
を過ぎ是よりアヌチノに至るまでは盡くダウビヘ河の谷地に沿ふ
```

該当箇所原本画像（11 号 22 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=11&pb=22

4.13　quotation 要素

説明

上位要素とは発話者や発話場面の異なる文書要素が引用されている部分を表す。

属性

type（必須）：引用の種類

- speech…会話・心話・演説等の引用
- citation…他文献からの引用や記事に対する説明等

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）：文体の種類。上位要素と文体が異なる場合に必要。取り得る属性値は「4.6 article 要素」の style 属性を参照のこと。

script（任意）：書記体。上位要素と書記体が異なる場合に必要。取り得る属性値は「4.6 article 要素」の script 属性を参照のこと。

XML 例

例 1　type 属性が「speech」の quotation 要素

```
此に於てコンスタンチノプル駐劄の佛國公使モンテベロ氏は土帝に見へ、此の問題に附て竊かに佛國政府の意を陳べて曰く、
<quotation type="speech" source="佛國公使モンテベロ">
佛國は斯る譯の分らぬ問題を賛成する能はず、土帝の主權を分割して英國に與るを賛成する能はざるなり
</quotation>
と、
```

該当箇所原本画像（11 号 4 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=11&pb=4

例 2　type 属性が「citation」、style 属性が「漢文」の quotation 要素

```
<quotation type="citation" source="唐朝の詩人" style="漢文">
長安少年無遠圖、畢生唯羨執金吾、
</quotation>
とは是れ唐朝の詩人が當時の少年を諷刺したる句なり、
```

該当箇所原本画像（2 号 15 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=02&pb=15

例 3　type 属性が「speech」、style 属性が「口語」の quotation 要素

```
吾人は後藤氏の演説したるに敬服す、（…中略…）今試に其の大意を左に掲げん、
<quotation type="speech" source="後藤" style="口語">
　諸君よ諸君は固より有名の人物であられます又愛國の志士であられます（…中略…）此の同胞をして進んで獨立不羈の國旗を海外に飜へすことは我々平生自ら信じて疑はざる所であります又敢て他人に讓らざるの覺悟であります
</quotation>
此の演説は、極めて極めて單簡なれども、以て當夜宴會の目的と、及び後藤氏近日の政治上に關する意見の大概を知るに足らん、
```

該当箇所原本画像（10 号 35 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=10&pb=35

4.14　superS 要素

説明

割書や引用を含むため、複数の s 要素からなると見なされる文を表す。

本コーパスでは、形態素解析での必要性から、warigaki 要素と quotatin 要素はその前後とは必ず別の 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　warigaki 要素を含む superS 要素

```
<superS>
<s type="fragment">ラザレフ村より五露里の間は雜木繁茂する山地に因て通し此山地を下りてダウビホザ河</s>
<warigaki>
<s>廣さ五十「サーゼニ」</s>
</warigaki>
<s type="fragment">を過ぎ是よりアヌチノに至るまでは盡くダウビヘ河の谷地に沿ふ</s>
</superS>
```

該当箇所原本画像（11 号 22 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=11&pb=22

例 2　quotation 要素を含む superS 要素

```
<superS>
<s type="fragment">此に於てコンスタンチノプル駐劄の佛國公使モンテベロ氏は土帝に見へ、此の問題に附て竊かに佛國政府の意を陳べて曰く、</s>
<quotation type="speech" source="佛國公使モンテベロ">
<s>佛國は斯る譯の分らぬ問題を賛成する能はず、</s><s>土帝の主權を分割して英國に與るを賛成する能はざるなり</s>
</quotation>
<s type="fragment">と、</s>
</superS>
```

該当箇所原本画像（11 号 4 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=11&pb=4

4.15　s 要素

説明

文を表す。

文境界は、コアデータはすべて人手により認定した。コア以外のデータは、最初に「、」「。」と改行文字を手がかりとして仮の文境界を自動認定し、次にその仮の文境界直前の短単位（SUW 要素）の品詞や活用形を手がかりとした自動修正（一部、人手による修正）により認定した。

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。

style（任意）：文体の種類。上位要素と文体が異なる場合に必要。取り得る属性値は「4.6 article 要素」の style 属性を参照のこと。

script（任意）：書記体。上位要素と書記体が異なる場合に必要。取り得る属性値は「4.6 article 要素」の script 属性を参照のこと。

XML 例

例 1　通常の s 要素

```
<s>皇帝陛下及び皇后宮には先帝二十年の親祭を行はしめ給んが爲めに去る一月廿五日東京を御發輦西京へ行幸在らせられ給へり</s>
<s>それ西京は一千年來の帝都にして殊に維新中興の大舞臺にてありき</s>
<s>而して先帝の英武剛明に在し給ふは殆んど後三條帝の上に出づ</s>
<s>蓋し維新の大業は今上の功徳に出ると雖も亦た先帝の遺烈に馮らずんばあらず、</s>
```

該当箇所原本画像（1 号 1 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=01&pb=1

例 2　type 属性値が「fragment」の s 要素

```
<superS>
<s type="fragment">ラザレフ村より五露里の間は雑木繁茂する山地に因て通し此山地を下りてダウビホザ河</s>
<warigaki>
<s>廣さ五十「サーゼニ」</s>
</warigaki>
<s type="fragment">を過ぎ是よりアヌチノに至るまでは盡くダウビヘ河の谷地に沿ふ</s>
</superS>
```

該当箇所原本画像（11 号 22 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=11&pb=22

4.16　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、同一 SUW 要素内の直前の漢字 1 字を繰り返す「々」「ゝ」は odoriji 要素とはせず、テキストを「々」「ゝ」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

```
擅ま<odoriji originalText="ゝ">ま</odoriji>に選擧法を變改して
```

該当箇所原本画像（15 号 16 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=15&pb=16

例 2　二字点

```
國會開設の準備の責に任ぜらる<odoriji originalText="ゝ">る</odoriji>ことを信ずればなり
```

該当箇所原本画像（1 号 6 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=01&pb=6

例 3　同字点

```
頃日速記<odoriji originalText="々">記</odoriji>者に命じて其の大意を筆記せしめ
```

該当箇所原本画像（9 号 16 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=09&pb=16

例 3　くの字点

```
嗟呼開け<odoriji originalText="/ \ ">開け</odoriji>博覧會も開け
```

該当箇所原本画像（1 号 4 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=01&pb=4

例 4　odoriji 要素としない同字点

```
茫々又た茫々何の邊に漂ふ乎
```

該当箇所原本画像（1 号 3 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=01&pb=3

4.17　span 要素

説明

特にマークアップする必要のある文中の文字列を表す。原文の漢字片仮名交じりの片仮名を平仮名に変換して電子テキスト化する際、片仮名表記のままとした文字列や、上位要素とは文体・書記体が異なる文字列をマークアップするために用いる。

属性

type（任意）：

- カタカナ…原文の漢字片仮名交じりの片仮名を平仮名に変換して電子テキスト化する際、外来語など片仮名のままとした文字列を表す。

style（任意）：文体の種類。上位要素と文体が異なる場合に必要。取り得る属性値は「4.6 article 要素」の style 属性を参照のこと。

script（任意）：書記体。上位要素と書記体が異なる場合に必要。取り得る属性値は「4.6 article 要素」の script 属性を参照のこと。

XML 例

例 1　type 属性が「カタカナ」の span 要素

```
<s><span type="カタカナ">ヘンリー</span>、<span type="カタカナ">ジョージ</span>氏は米國有名なる社會主義の學士にして著す所進歩及貧乏、土地疑問等の書あり</s>
```

該当箇所原本画像（1 号 23 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=01&pb=23

例 2　style 属性が「外国語」の span 要素

```
<s>而して此の如き辨護は英語の所謂<span style="外国語">Ｔｏ　ｄｏ　Ｊｕｓｔｉｃｅ</span>なるものにして十九世紀の學士識者が何れも是認する所なれども舊學派安政年間の思想若くは其復古に志ある大發明家は非なりとせらるるかも知れず甚だ大膽ながら敢て御質問に及び候</s>
```

該当箇所原本画像（26 号 36 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=26&pb=36

例 3　script 属性が「カタカナ」の span 要素

```
<s>是より先き、アメリカン、パアテーと申す<span script="カタカナ">「米國は米國人の支配たるべし」</span>の主義を奉ぜる一政黨、當カリフオルニヤに起り、首として外國より入來する、無資産、不道徳、の者共を嫌惡し、今に尚之れを嫌惡する最中なり、</s>
```

該当箇所原本画像（28 号 37 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=28&pb=37

4.18　pb 要素

説明

原本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。原則、原本のページ番号と一致する。ただし、原本では同一号内で新たにページ番号が振り直されていたり、ページ番号がなかったりする場合は、コーパス独自に連続するページ番号を付ける。

XML 例

```
<pb n="1"/>　國民之友第九號
（…中略…）吾人は
<pb n="2"/>其の罪に伏せんことを甘ずればこそ、天下具眼の人士に向て、豫じめ今日に於て、戒嚴せられんことを願ふ、（…中略…）諸氏の擧動を信ぜ
<pb n="3"/>ざるなり、
```

該当箇所原本画像（9 号 1 ページ～）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=09&pb=1

4.19　cb 要素

説明

原本の紙面上での改段位置を表す。空要素。

属性

n（必須）：該当位置から始まる段の番号。

XML 例

```
<pb n="2"/><cb n="1"/>吾人が待ちに待たる長崎事件の談判も既に穩便に決着せり（…中略…）既に彼國活眼政治家の李鴻章曾紀澤諸氏の看破す
<cb n="2"/>る所にして果して此の如く穩便に決着したるは吾人が尤も兩國の爲めに祝し（…中略…）地方官と
<pb n="3"/><cb n="1"/>議會とは其間圓滑を尚び相敵視することなくして利害を謀る事
```

該当箇所原本画像（1 号 2 ページ～）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=01&pb=2

4.20　lb 要素

説明

原本の紙面上での改行位置を表す。空要素。

属性

なし

XML 例

```
<pb  n="2"/><cb  n="1"/><lb/>吾人が待ちに待たる長崎事件の談判も既に穏便に決着
<lb/>せり蓋し長崎事件は一小事件なり清國水夫と日本巡査
<lb/>との一時の喧嘩に過ぎざる事件なり固より立談の間に
<lb/>其の局を結ぶ可き也而して彼の清國なるもの動もすれ
<lb/>ば毛を吹て疵を求め、藪を探りて蛇を出だすが如きの
<lb/>傾向ありしは何ぞや
```

該当箇所原本画像（1 号 2 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=01&pb=2

4.21　br 要素

説明

論理改行を表す。空要素。

属性

なし

XML 例

```
<p>
<s>我が明治政府は、今ま明治廿一年四月廿五日を以て、市制及町村制を發布せられたり、</s>（…中略…）<s>此法律の趣旨を了解せんには、先づ第一に自治と稱ふる者を了解するを要するものとすべし、<br/></s>
</p>
<p>
<s>自治とは自ら獨立して自分の事を治むるの謂なり、</s>（…中略…）<s>故に丁年以上の者にして自治の權なきものは、不完全のもの、即ちカタワ者なりと謂ふ可し、<br/></s>
</p>
<p>
<s>人相集りて家を成し、又は町村をなし、又は府縣をなし又は國をなすときは、此等の集合躰に於て、共同の利益を有すること少からざるなり、</s>（…中略…）<s>此無形人中には法律を以て創造せられ、又は認定せらるるものあるが故に、之を法律上の人即ち法人とも稱するなり、<br/></s>
</p>
```

該当箇所原本画像（21 号 11 ページ～）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=21&pb=11

4.22　SUW 要素

説明

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」[7]による解析結果を人手で修正したものである。修正範囲は、コアデータはすべての SUW 要素、コア以外のデータは一部の SUW 要素とした。情報の精度（適合率）はコアデータで 99%以上、コア以外のデータで 95%以上を確保している。

語の単位として採用した短単位（Short-Unit Word）は、国立国語研究所により設計された言語単位で、基準がわかりやすくゆれが少ないという特長を持つ。短単位の規程は、適応する資料の時代や文体ごとに定められ、現代語用の規程集（小椋・小磯・冨士池ほか、2011、第 3 章）と中古和文用の規程集（小椋・須永、2012）が公開されている。しかし、近代語に関しては定まった規程はまだない。そこで『国民之友コーパス』では暫定的に、文語用の規程として『明六雑誌コーパス』開発時に検討された近代語の規程案（須永・近藤、2012）を中古和文用の規程で補ったものを用い、口語用の規程として現代語の規程を用いることとした。

SUW 要素の各属性の詳細については、上述の各種短単位規程集や「近代文語 UniDic」のユーザーズマニュアル、「近代文語 UniDic」のもととなった現代語用「UniDic」[8]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

lForm（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。本コーパスで特に使用する属性値は次のとおり。

- 仮…pos 属性値が「名詞-固有名詞-人名-一般」「名詞-固有名詞-地名-一般」の要素のうち、語彙素・語形の認定を行っていないもの。lForm・lemma・form・pronToken・kanaToken 属性には仮に orthToken 属性と同値を入力している。

pos（必須）：品詞。本コーパスで特に使用する属性値は次のとおり。

- メタ（誤り）…語形や表記等が誤っている語の引用
- 外国語…外国語
- 漢文…漢文
- 言いよどみ…言いよどみ
- 読取不可…判読できない文字を含むもの

form（任意）：語形

cType（任意）：活用型。活用語のみ必要。

cForm（任意）：活用形。活用語のみ必要。

pronToken（任意）：発音形出現形

[7] http://www2.ninjal.ac.jp/lrc/index.php?UniDic
[8] http://sourceforge.jp/projects/unidic/

kanaToken（任意）：仮名形出現形

orth（任意）：書字形基本形。活用語のみ必要。

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

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="吾人" lForm="ゴジン" lemma="吾人" pos="代名詞" form="ゴジン" pronToken="ゴジン" kanaToken="ゴジン" orth="吾人" wType="漢" start="308630" end="308650" orderID="204230" BOS="True">吾人</SUW>
<SUW orthToken="は" lForm="ハ" lemma="は" pos="助詞-係助詞" form="ハ" pronToken="ワ" kanaToken="ハ" orth="は" wType="和" start="308650" end="308660" orderID="204240">は</SUW>
<SUW orthToken="後藤" lForm="ゴトウ" lemma="ゴトウ" pos="名詞-固有名詞-人名-姓" form="ゴトウ" pronToken="ゴトー" kanaToken="ゴトウ" orth="後藤" wType="固" start="308660" end="308680" orderID="204250">後藤</SUW>
<SUW orthToken="氏" lForm="シ" lemma="氏" pos="接尾辞-名詞的-一般" form="シ" pronToken="シ" kanaToken="シ" orth="氏" wType="漢" start="308680" end="308690" orderID="204260">氏</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="308690" end="308700" orderID="204270">の</SUW>
<SUW orthToken="演説" lForm="エンゼツ" lemma="演説" pos="名詞-普通名詞-サ変可能" form="エンゼツ" pronToken="エンゼツ" kanaToken="エンゼツ" orth="演説" wType="漢" start="308700" end="308720" orderID="204280">演説</SUW>
<SUW orthToken="し" lForm="スル" lemma="為る" pos="動詞-非自立可能" form="ス" cType="文語サ行変格" cForm="連用形-一般" pronToken="シ" kanaToken="シ" orth="す" wType="和" start="308720" end="308730" orderID="204290">し</SUW>
<SUW orthToken="たる" lForm="タリ" lemma="たり" subLemma="完了" pos="助動詞" form="タリ" cType="文語助動詞-タリ-完了" cForm="連体形-一般" pronToken="タル" kanaToken="タル" orth="たり" wType="和" start="308730" end="308750" orderID="204300">たる</SUW>
<SUW orthToken="に" lForm="ニ" lemma="に" pos="助詞-格助詞" form="ニ" pronToken="ニ" kanaToken="ニ" orth="に" wType="和" start="308750" end="308760" orderID="204310">に</SUW>
<SUW orthToken="敬服" lForm="ケイフク" lemma="敬服" pos="名詞-普通名詞-サ変可能" form="ケイフク" pronToken="ケーフク" kanaToken="ケイフク" orth="敬服" wType="漢" start="308760" end="308780" orderID="204320">敬服</SUW>
<SUW orthToken="す" lForm="スル" lemma="為る" pos="動詞-非自立可能" form="ス" cType="文語サ行変格" cForm="終止形-一般" pronToken="ス" kanaToken="ス" orth="す" wType="和" start="308780" end="308790" orderID="204330">す</SUW>
<SUW orthToken="、" lForm="" lemma="、" pos="補助記号-読点" orth="、" wType="記号" start="308790" end="308800" orderID="204340">、</SUW>
```

該当箇所原本画像（10 号 35 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=10&pb=35

4.23　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

```
人を動すは<ruby rubyText="エノルヂー">力</ruby>なり、
```

該当箇所原本画像（1 号 19 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=01&pb=19

例 2

```
<ruby rubyText="や">八</ruby><ruby rubyText="つか">束</ruby><ruby rubyText="ひげ">鬚</ruby>生たる蝦夷も<ruby rubyText="いも">紅裙</ruby>と酌む<ruby rubyText="なさけ">情</ruby>の道は<ruby rubyText="かはら">異</ruby>ざりけり
```

該当箇所原本画像（2 号 33 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=02&pb=33

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="ケヤレス" rubyBase="不注意">不</ruby>注意
```

該当箇所原本画像（36 号 23 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=36&pb=23

4.24　lRuby 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

```
<lRuby rubyText="コングレツセス、">會盟</lRuby><lRuby rubyText="コンフエレンセス">會議</lRuby>
```

該当箇所原本画像（19 号 64 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=19&pb=64

例 2

```
<lRuby rubyText="たち">立</lRuby><lRuby rubyText="をう">徃</lRuby><lRuby rubyText="じやう">生</lRuby>
```

該当箇所原本画像（10 号 4 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=10&pb=4

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

```
<lRuby rubyText="ライト、オフ、レゲーション" rubyBase="駐剳權">駐剳</lRuby>權
```

該当箇所原本画像（19 号 63 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=19&pb=63

4.25　corr 要素

説明

原文の文字に修正を施し、異なる文字としたものを表す。

文字の修正は次の場合に行う。

(1)　誤植の校訂

(2)　漢文体・候文体を訓読するための返読・補読・仮名開き[9]

該当文字が本行にある場合、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」「返読前」「仮名開き前」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・lRuby 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）：修正の種別

- erratum…誤植の種類が誤字であることを表す。
- excess…誤植の種類が衍字であることを表す。
- omission…誤植の種類が脱字であることを表す。
- 返読前…返読の対象となる文字の、訓読前の文字と位置を表す。
- 返読後…返読の対象となる文字の、訓読後の文字と位置を表す。助動詞などは仮名に開いてテキスト化する。
- 補読…補読された文字を表す。
- 仮名開き前…返読は伴わないが仮名に開く対象となる文字の、開く前の文字と位置を表す
- 仮名開き後…返読は伴わないが仮名に開く対象となる文字の、開いた後の

[9] 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は訓読せず、そのまま入力する。

文字と位置を表す

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。type 属性値が「erratum」「excess」「返読前」「仮名開き前」の場合に必要。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。

id（任意）：返読・仮名開きの対象となる文字の、訓読前の位置と訓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」「仮名開き前」「仮名開き後」の場合に必要。

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

```
敵となり、<corr originalText="昧" type="erratum">味</corr>方となり、
```

該当箇所原本画像（1 号 12 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=01&pb=12

例 2　衍字

```
ヘンリ<corr originalText="リ" type="excess"/>ージョージ著
```

該当箇所原本画像（3 号 41 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=03&pb=41

例 3　脱字

```
社會の理にあ<corr type="omission">ら</corr>ざる乎
```

該当箇所原本画像（1 号 35 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=01&pb=35

例 4　ルビ中の誤字

```
<ruby rubyText="なぐさ"><corr originalText="かぐさ" type="erratum" subType="ruby">慰</corr></ruby>むる日
```

該当箇所原本画像（24 号 32 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=24&pb=32

例 5　返読・仮名開き

```
焼板にてのちのちは彌得難き物は一編貳百疋づつにてゆづり<corr type="返読前" originalText="被" id="3108"/>成<corr type="返読後" id="3108">れ</corr>候<corr type="仮名開き前" originalText="半" id="3109"/><corr type="仮名開き後" id="3109">は</corr><corr type="仮名開き後" id="3109">ん</corr>や
```

該当箇所原本画像（28 号 31 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=28&pb=31

例 5　返読・補読

```
之れを裁判し<corr type="返読前" originalText="不" id="1901"/><corr type="返読前" originalText="得" id="1902"/>止<corr type="補読">を</corr><corr type="返読後" id="1902">得</corr><corr type="返読後" id="1901">ず</corr>腕力を以て脅迫の處置を行はしめたるものなり
```

該当箇所原本画像（11 号 19 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=11&pb=19

4.26　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。

本行に該当文字がある場合、推定した文字を unclear 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 属性値には推定した文字を入力し、unclear 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）：該当文字がルビにある場合は、該当文字を「␣」（面区点番号：1-07-93、Unicode コード：U+2423）で入力した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

```
併<unclear>せ</unclear>て東洋平和の爲めに祝して止まざる所なり
```

該当箇所原本画像（1 号 2 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=01&pb=2

例 2　ルビにある場合

```
<ruby rubyText="エステルライヒ"><unclear originalText="エ␣テルライヒ" type="ruby">澳斯</unclear></ruby>
```

該当箇所原本画像（25 号 60 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=25&pb=60

4.27　vMark 要素

説明

濁音を表記するにもかかわらず、原文では濁点のない仮名が使われていることを表す。

vMark 要素は、コアデータではすべて人手により認定し、コア以外のデータでは濁点自

動付与システム「AYTC（文鳳）にごり ONLY バージョン」[10]による処理結果を一部人手により修正して認定した。

本行に該当文字がある場合、濁点付きの仮名を 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 属性値には濁点付きの仮名を入力し、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

```
先帝の遺烈に憑ら<vMark>ず</vMark>ん<vMark>ば</vMark>あら<vMark>ず</vMark>、
```

該当箇所原本画像（1 号 1 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=01&pb=1

例 2　ルビの場合

```
<ruby rubyText="あじ"><vMark originalText="あし" type="ruby">味</vMark></ruby><ruby rubyText="き">氣</ruby>なき境遇に立つも、
```

該当箇所原本画像（5 号 23 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=05&pb=23

4.28　g 要素

説明

コーパスの使用する文字集合（「3.2　文字集合」参照）の範囲にない外字で、代用字や「=」で入力されたものを表す。また、敬意欠字を表す。

外字で、かつ JIS X 0213 の包摂規準および追加包摂規準（須永・堤・近藤ほか、2013）の適用外の文字であるが、意味・用法の類似する他の文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。

代用字での入力も適用できない外字は「=」（面区点番号：1-02-14、Unicode コード：U+3013）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、全角スペース（面区点番号：1-01-01、Unicode コード：U+3000）を

[10] http://cl.naist.jp/~teruaki-o/AYTC/

入力して g 要素の内容とする。

属性

type（必須）：

- 外字…外字で、かつ JIS X 0213 の包摂規準と追加包摂規準の適用外の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「敬意欠字」の場合は不要。

XML 例

例 1　外字（代用字による入力）

```
<g type="外字" ref="U+51CF">減</g>少
```

該当箇所原本画像（1 号 36 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=01&pb=36

例 2　外字（代用字による入力）

```
ウ<g type="外字" ref="小書き「ヰ」">ヰ</g>ルヘルム
```

該当箇所原本画像（2 号 23 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=02&pb=23

例 3　外字（=による入力）

```
<g type="外字" ref="U+89BC">=</g>次
```

該当箇所原本画像（2 号 35 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=02&pb=35

例 4　外字（=による入力）

```
合<g type="外字" ref="「丞」の最終画の代わりに「巴」">=</g>
```

該当箇所原本画像（30 号 40 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=30&pb=40

例 5　敬意欠字

```
但だ我が<g type="敬意欠字">　</g>皇上の至仁至徳に在すの故を以て、
```

該当箇所原本画像（1 号 20 ページ）

http://dglb01.ninjal.ac.jp/ninjaldl/convert.php?title=kokuminnotomo&issue=1&pb=20

5　データの種類と形式

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のファイルで、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 36 ファイルからなる。XML ファイルの符号化形式は UTF-16LE（BOM あり）、改行コードは LF である。ファイル名は「k」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「k188701.xml」ならば、1887 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　「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、わかりやすいユーザーインターフェイスによるコーパスの検索・閲覧が可能となる。原本画像の参照機能も持つ。

5.2.1　「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手順で行う。

① データ kokumin_himawari.zip をダウンロードする。Windows 機の場合は、kokumin_himawari.zip を右クリックし、［プロパティ］>［全般］でセキュリティのブロックが解除されていることを必ず確認する。

② kokumin_himawari.zip を解凍すると「kokumin_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…『国民之友コーパス』データを格納したフォルダ
- config_kokumin.xml…「ひまわり」用設定ファイル
- .himawari_package_info…パッケージインストール設定ファイル

③ データの対応するバージョンの「ひまわり」をインストールする。国立国語研究所コーパス開発センターWeb サイトの「ツール」ページ（http://www.ninjal.ac.jp/corpus_center/tool.html）から、「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」のインストールを行う。

④ 「ひまわり」をインストールすると「Himawari_X」（X には「ひまわり」のバージョン番号に対応した数字が入る）フォルダが現れる。その中の「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（図 1）が開く。画面上部の「ファイル」メニュー―「インストール」（図 2）を選択し、解凍した「kokumin_himawari」フォルダを指定して『国民之友コーパス』データをインストールする。

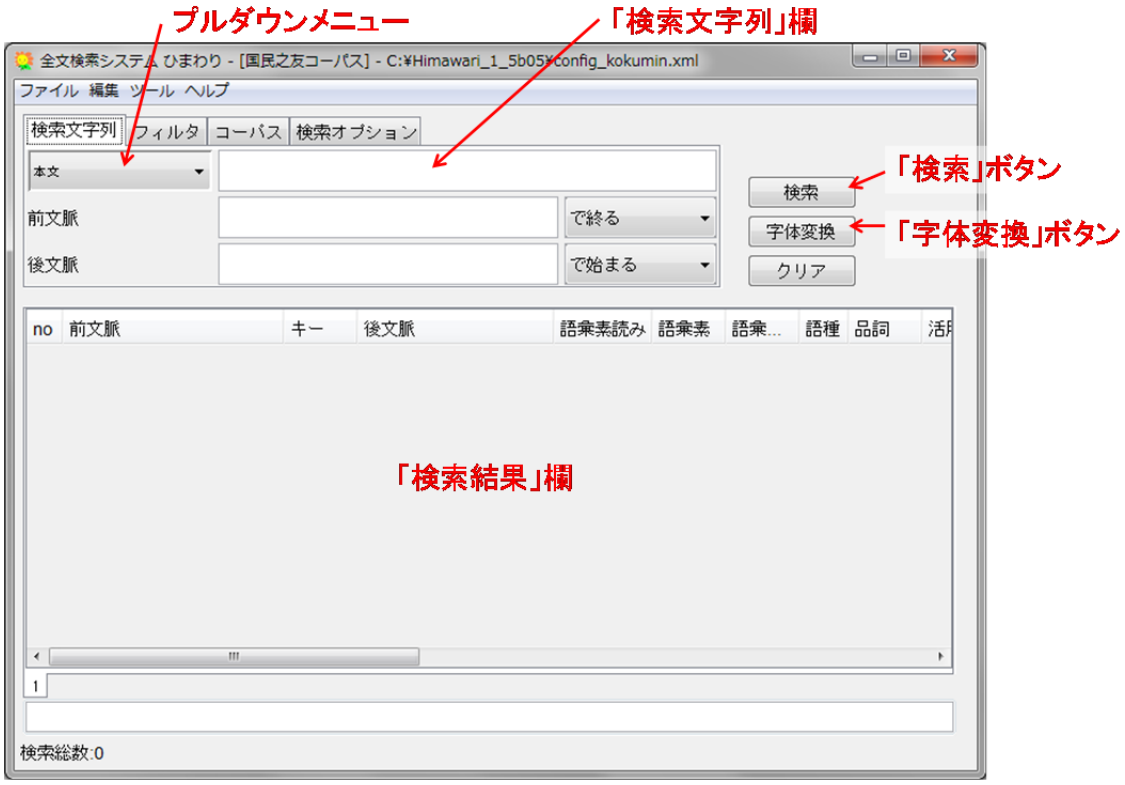

図 1　「ひまわり」の起動画面

図 2　「ファイル」メニュー―「インストール」

5.2.2　「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・閲覧方法を説明する。

まず、「ひまわり」の起動画面（図 1）上部の「ファイル」メニュー―「新規」を選択する（図 3）。設定ファイルを指定するための画面が現れるので、「config_kokumin.xml」を選択する（『国民之友コーパス』データのインストール直後や前回起動時の設定が保存されている場合は、この手順は省略できる）。

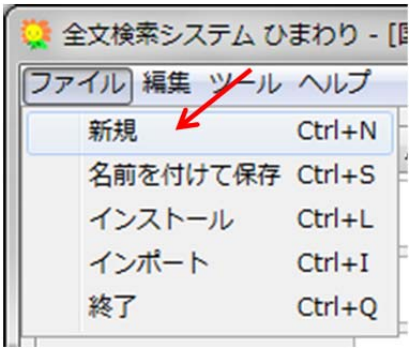

図 3　「ファイル」メニュー―「新規」

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 2 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 2　「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
語彙素/完全一致	SUW 要素 lemma 属性値
語彙素/部分一致	
語彙素読み/完全一致	SUW 要素 lForm 属性値
語彙素読み/部分一致	
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/完全一致	SUW 要素 orth 属性値
書字形基本形/部分一致	
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	

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 4）。

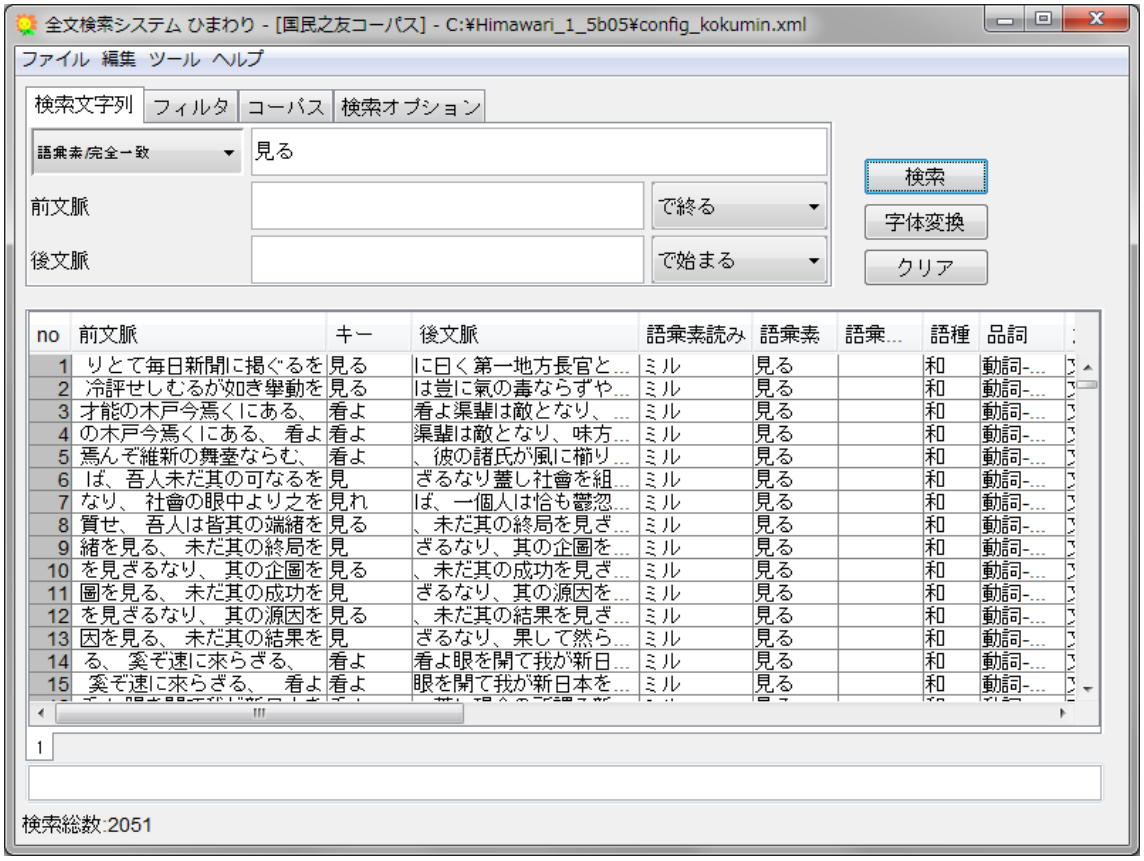

図 4　「ひまわり」での検索結果表示

「検索結果」欄に表示される列のリストを表 3 として示す。

表 3　「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 n 属性に対応
段	cb 要素 n 属性に対応
語連番	SUW 要素 orderID 属性に対応
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
コア	article 要素 core 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

「検索結果」欄の「ページ」列のセルをダブルクリックすると、Web ブラウザが起動し、該当ページの原本画像が閲覧できる（図 5）。原本画像は国立国語研究所 Web サイト（http://dglb01.ninjal.ac.jp/ninjaldl/bunken.php?title=kokuminnotomo）で公開されているものを参照している。

図 5　原本画像の閲覧

「検索結果」欄の「ページ」列以外のセルをダブルクリックすると、Web ブラウザが起動し、雑誌単位あるいは記事単位での文脈閲覧ができる。

閲覧表示スタイルは次の 4 種類がある。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。

- 本文（図 6）
- 本文＋画像（図 7）
- 本文＋付加情報（図 8）
- 形態論情報リスト（図 9）

図 6　「本文」スタイルでの文脈表示

図 7　「本文＋画像」スタイルでの文脈表示

図 8　「本文＋付加情報」スタイルでの文脈表示

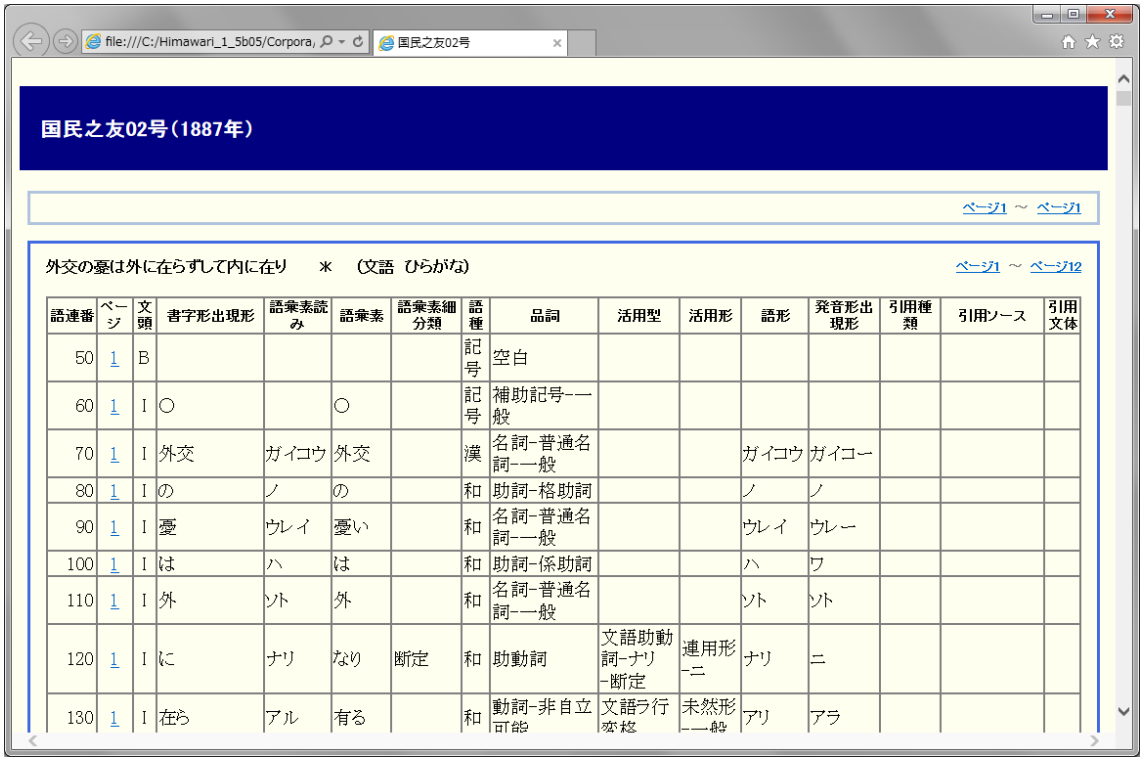

図 9　「形態論情報リスト」スタイルでの文脈表示

「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。

5.3　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「kokumin_suw.txt」、符号化形式は UTF-16LE（BOM あり）、改行コードは LF 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 1 つの SUW 要素に対応している。

データのフィールドリストを表 4 として示す。

表 4　形態論情報タブ区切りデータのフィールドリスト

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
年	magazine 要素 year 属性に対応
号	magazine 要素 issue 属性に対応
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
コア	article 要素 core 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
発音形出現形	SUW 要素 pronToken 属性に対応

5.4　著者情報タブ区切りデータ

コーパス中の記事の著者・原著者に関する情報のリストを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「kokumin_author.txt」、符号化形式は UTF-16LE（BOM あり）、改行コードは LF 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 1 人の著者・原著者に対応している。

データのフィールドリストを表 5 として示す。

表 5　著者情報タブ区切りデータのフィールドリスト

フィールド名	備考
記事著者	article 要素 author 属性に対応。author 属性値に複数人が列挙されている場合は 1 人ずつに分割した値とする。本文テキストのない article 要素 author 属性は含まない。
記事原著者	article 要素 originalAuthor 属性に対応。本文テキストのない article 要素 originalAuthor 属性は含まない。
生年	辞典類の記述に拠る。不明なものは空欄。
没年	辞典類の記述に拠る。不明なものは空欄。
所属	辞典類や原文での記述に拠る。不明なものは空欄。
分野	辞典類の記述に拠る。不明なものは空欄。

5.5　記事情報タブ区切りデータ

XML ファイルから特に article 要素・titleBlock 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「kokumin_article_titleBlock.txt」、符号化形式は UTF-16LE（BOM あり）、改行コードは LF 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 1 つの article 要素・titleBlock 要素に対応している。

データのフィールドリストを表 6 として示す。

表 6　記事情報タブ区切りデータのフィールドリスト

フィールド名	備考
年	magazine 要素 year 属性に対応
号	magazine 要素 issue 属性に対応

記事題名	article 要素 title 属性に対応。figureBlock 要素の場合は「figureBlock」を入力
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 type 属性に対応
記事書記体	article 要素 script 属性に対応
コア	article 要素 core 属性に対応
冒頭ページ	article 要素・titleBlock 要素の始まるページ番号
冒頭段	article 要素・titleBlock 要素の始まる段番号
末尾ページ	article 要素・titleBlock 要素の終わるページ番号
末尾段	article 要素・titleBlock 要素の終わる段番号
冒頭語連番	article 要素・titleBlock 要素の最初の SUW 要素 orderID 属性に対応
末尾語連番	article 要素・titleBlock 要素の最後の SUW 要素 orderID 属性に対応
冒頭文字開始位置	article 要素・titleBlock 要素の最初の SUW 要素 start 属性に対応
末尾文字終了位置	article 要素・titleBlock 要素の最後の SUW 要素 end 属性に対応

参考文献

有山輝雄（1986）「言論の商業化―明治 20 年代の民友社―」『コミュニケーション紀要』4、pp.1-23（http://www.seijo.ac.jp/graduate/gslit/orig/journal/communication/pdf/scom-04-01.pdf よりダウンロード可）

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011）『『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 4 版（上）（下）』国立国語研究所（http://www.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-01.pdf と http://www.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf よりダウンロード可）

小椋秀樹・須永哲矢（2012）『中古和文 UniDic 短単位規程集　平成 21（2009）－平成 23（2011）年度科学研究費補助金基盤研究（C）「和文系資料を対象とした形態素解析辞書の開発」研究成果報告書 2』（https://dl.dropboxusercontent.com/u/73297026/report/unidic-EMJ_rulebook2012.pdf よりダウンロード可）

国立国語研究所（編）（2005）『太陽コーパス―雑誌『太陽』日本語データベース―』博文館新社

近藤明日子・田中牧郎（2012）「『明六雑誌コーパス』の仕様」『近代語コーパス設計のための文献言語研究 成果報告書』pp.118-143（http://www.ninjal.ac.jp/corpus_center/cmj/doc/07kondo.pdf よりダウンロード可）

須永哲矢・近藤明日子（2012）「近代語コーパスのための形態論情報付与規程の整備」『近代語コーパス設計のための文献言語研究　成果報告書』pp.93-117（http://www.ninjal.ac.jp/corpus_center/cmj/doc/06sunaga.pdf よりダウンロード可）

須永哲矢・堤智昭・近藤明日子・木川あづさ・服部紀子（2013）「明治中期雑誌の異体漢字と JIS 漢字―『国民之友』を事例として―」『じんもんこん 2013 論文集』pp.201-208