『日本語歴史コーパス　明治・大正編 I 雑誌』（短単位データ 1.2）テキストの凡例と「中納言」表示項目について

2019 年 3 月 29 日
間淵洋子　近藤明日子　服部紀子　南雲千香子

１．はじめに

『日本語歴史コーパス　明治・大正編 I 雑誌』（短単位データ 1.2）は，これまで別々に公開されてきた短単位形態論情報付きコーパス『明六雑誌コーパス』『国民之友コーパス』，全文テキストコーパス『太陽コーパス』『近代女性雑誌コーパス』（いずれも，XML 形式および全文検索システム『ひまわり』形式による）を統合し，検索アプリケーション「中納言」での利用に即した短単位形態論情報付きコーパスとして再構築したものに，『東洋学芸雑誌』のデータを新たに追加したものである。

本コーパスの本文は，『日本語歴史コーパス(CHJ)』の他のサブコーパスと整合的な形態論情報を付与するために，雑誌本文に対して校訂を加えたものとなっている。また，本コーパスの本文は，収録対象の雑誌巻号について全文を含んだものになっているが，検索対象サンプルは雑誌の 1 記事を単位として分割されており，各サンプル（＝各記事）は個別の ID と書誌的情報（作品・作者・原資料等に関する情報）を有する。

この文書では，コーパス本文の成り立ちと，検索アプリケーション「中納言」における検索結果の表示項目について，例示しながらその概要を示す。

なお，本コーパスでは，研究上必要と思われる情報を，できるだけ原文の状況に即して記述するよう努めたが，不十分・不適切な箇所が残存する可能性もある。適宜，原資料の情報を基に，原文を確認されることを推奨する。

２．テキストの凡例

２．１　テキストに使用する文字

本コーパスの本文テキストに使用した文字の範囲は，JIS X 0213（JIS の文字コード規格）の文字集合（JIS 漢字の第 4 水準までを含む）に準拠している[1]。変体仮名はこの文字集合に含まれないため用いず，JIS 内の平仮名によって表す。また，この文字集合に含まれない漢字については，以下の順で，異体の文字に包摂・代用した[2]。

(1) JIS X 0213 の「6.6.3　漢字の字体の包摂規準」に若干の拡張を施した本コーパス用の包摂規準に基づいて，JIS 内の文字に包摂する。

(2) (1)の包摂規準を適用できない字形差をもつ漢字は，類似の意味・用法を持つ同音・同訓の JIS 内の文字で代用する。

---

[1] ただし，①JIS X 0213 附属書 7　2.1　b)に掲載される，戸籍法施行規則付則別表“人名用漢字許容字体表”（昭和 56 年法務省令 51）の漢字，及び常用漢字表（昭和 56 年内閣告示第 1 号）のかっこ書き内の漢字（“いわゆる康熙字典体”）のうち，JIS X 0208 で包摂していた漢字，②JIS X 0213:2004 において UCS との互換のために追加された 10 字，③Unicode における CJK 統合漢字拡張 B については，これを用いない。

[2] 異体字の拡張包摂および代用については，須永・堤・高田(2011)，須永・堤・近藤ほか(2013)を参照のこと。

(3) (2)による代用が不可能な文字は，外字として「〓」（げた記号，JIS 面区点 1-02-14，U+3013）で表す。

なお，原資料の状態（印刷のかすれや破損・抹消）によって判読が困難な文字は，「␣」（空白記号，JIS 面区点 1-07-93，U+2423）によって表す。

２．２　本文校訂

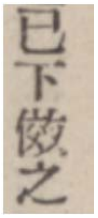
本コーパスでは，本文に対して，『日本語歴史コーパス（CHJ）』の他のサブコーパスや『現代日本語書き言葉均衡コーパス（BCCWJ）』等の他のコーパスと整合的な形態論情報を提供することを目的として，形態素解析辞書 UniDic に基づいた短単位情報を付与している。そのため，UniDic による形態素解析に適した解析用本文を用意する必要があり，原資料の雑誌本文に対して，いくつかの改変を施している。

例えば，UniDic は，通常の日本語文と異なる漢文式の語順や，片仮名表記による活用語尾には対応していないが，本コーパスが収録対象とする明治・大正時代の日本語の書記体には，漢文体の混入や，漢字片仮名混じり文が少なからず見られる。そのため，これらを現代の一般的な語順や表記法に変換して解析本文として用いる必要がある。また，当時の表記法としては一般的であった，踊り字による文字や語の繰り返し，濁点を用いない濁音表記などについても，形態素解析にはそぐわないため，対応が必要である。このような，近代に特有の表記に対して原資料の本文を改変することを，ここでは「校訂」と呼ぶ。

以下に，具体的な校訂の方法について，事例に基づき説明する（実例の出典をサンプル ID により示す）。なお，解析用の本文はこれらの校訂後の本文であるが，「中納言」では，校訂前の文字列を原文の状態に近い形で電子化したものを「原文文字列」「原文 KWIC」として表示させることができるほか，一部で原資料の画像を参照することもできる。利用に際しては，必要に応じて原文の情報を確認されたい。

［漢文式語順］

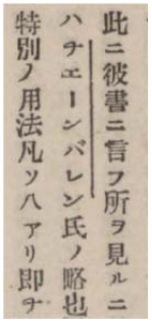
漢文式語順（漢文引用を除く）については，訓読した日本語文の語順に入れ替える。その際，付属語は仮名表記に置き換えるほか（例①：60M 明六 1874_03003），必要に応じて付属語を補足する（例②：60M 国民 1887_02018）。

①	②
	

① **原文**　可被遂候　／　**校訂**　遂らるべく候
② **原文**　已下倣之　／　**校訂**　已下之に倣

［漢字片仮名混じり文］

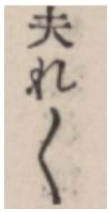
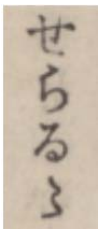
漢字片仮名混じり文は，片仮名部分を平仮名に置き換える。その際，外来語等，今日も片仮名表記が一般的なものは，置き換えの対象としない（例：60M 国民 1887_02018）。



原文　此ニ彼書ニ言フ所ヲ見ルニ(中略)ハチエーンバレン氏ノ略也
校訂　此に彼書に言ふ所を見るに(中略)はチエーンバレン氏の略也

［踊り字］

踊り字は，繰り返される文字・語に置き換える（例①：60M 国民 1887_05021，例②：60M 国民 1887_01020）。ただし，「人々(ひとびと)」「徐ゝ(ジョジョ)」等，1 短単位内部で直前の 1 字を繰り返す「々」「ゝ」は置き換えの対象としない。

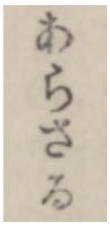
①	②
	

① **原文**　夫れ／＼　／　**校訂**　夫れ夫れ

② **原文**　せらるゝ　／　**校訂**　せらるる

［濁点無表記］

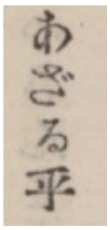
濁音が期待される箇所に濁点付き仮名が用いられていないものは，濁点の無表記と判断し，該当の濁音を表す濁点付き仮名に置き換える（例：60M 国民 1887_02006）。ただし，清濁両形がある語[3]については，同一サンプル内での統一等を除き，置き換えの対象としない。



原文　あらさる　／　**校訂**　あらざる

［誤植］

原文の誤植（脱字・衍字・前後文字列の転倒・誤字）と思われる表記は，訂正する（例：60M 国民 1887_01018）。ただし，仮名遣いの誤りや，語形のバリエーション，当時通用していたと考えられる同音漢字による異表記[4]などは，訂正の対象としない。



原文　あざる乎　／　**校訂**　あらざる乎

３．テキストの範囲とサンプル

３．１　収録対象テキスト

本コーパスは，原則として当該の雑誌各冊の全テキストを収録対象としているが，次の要素は収録対象外とする。

・表紙
・目次・目録
・刊記・識語
・口絵や図表を中心とする記事
・漢文や欧文からなる記事
・前号の誤植等を訂正する記事

---

[3] 当該の語に清濁両形があるかどうかの判定は，原則として『日本国語大辞典　第二版』（小学館）によるものとする。清濁両形が辞書の「見出し」にある場合のほか，「語義説明」内に“(「○○」とも)”の形で異語形を示す場合は，清濁両形があるものと判断する。

[4] 語形のバリエーションかどうかの判定は，注 3 清濁両形の有無の判定に準ずる。また，通用の異表記かどうかの判定は，①『日本語国語大辞典　第二版』及び②近代語のコーパスによる出現状況によるものとする。①は，見出しの「漢字表記」のほか，「用例文」中の表記，「表記」欄の表記などを通用の異表記とみなす。①が適用できない場合，②近代語コーパス（公開済みのもののほか，内部資料を含む）において，複数のサンプルに出現し，出現数が少なくない表記を異表記とみなす。

・記事不掲載を謝罪する記事
・雑誌の販売価格・販売方法，広告料等を告知する記事
・広告

また，コーパスの収録対象とした文書要素のテキストのうち，次の部分は対象からは除外する。

・図表・挿絵・写真中のテキストとそのキャプション
・漢文や欧文からなる段落

３．２　雑誌の構造とサンプル

雑誌は，複数の記事を収録したものであり，雑誌を構成するそれぞれの記事は，書き手や主題が異なる個別の文章・作品である。そのため，本コーパスにおいては，それぞれの記事を個別の作品として区別し，それぞれに書誌的・言語的な情報を付与することとした。そして，この方針に基づき，記事を単位とした「サンプル」というテキストの範囲を定めた。

個々の「サンプル」は，サンプル ID という個別に認識される ID を持つ。サンプル ID は 15 桁からなり，構成は表 1 の通りである（網掛けは記号や数値の意味を表す）。

表 1　本コーパスにおけるサンプル ID の構成

1-2 桁目		3 桁目		4-5 桁目		6-9 桁目	10 桁	11-15 桁目
時代通し番号		ジャンル		作品 ID		成立時期	区切り記号	作品内での出現順通し番号
60	明治・大正	M	雑誌	明治	明六雑誌	1874, 1875	-	号 2 桁[6]+記事連番 3 桁
				東洋	東洋学芸雑誌	1881, 1882		
				国民	国民之友	1887, 1888		
				太陽	太陽	1895, 1901, 1909 1917, 1925		
				女雑	女学雑誌	1894, 1895		
				女世	女学世界	1917		
				婦倶	婦人倶楽部	1925		

このうち，ID 末尾 3 桁目の記事連番について，以下に雑誌の構造との関係を示す。

本コーパスに収録した雑誌本文には，「誌名」や「欄名」など，個々の記事ではなく，雑誌そのものについての記載が含まれている（図 1 の白地部分）。これらは，雑誌全体を構成する本文ではあるが，個々の記事（図 1 の網掛け部分）とは性質を異にしているため，本コーパスでは，記事とは別に，“雑誌本体の構造要素”としてひとまとまりに扱っている。この“雑誌本体の構造要素”を記事連番「000」とし，内部の個々の記事について，出現順に「001」「002」…と ID を与える。

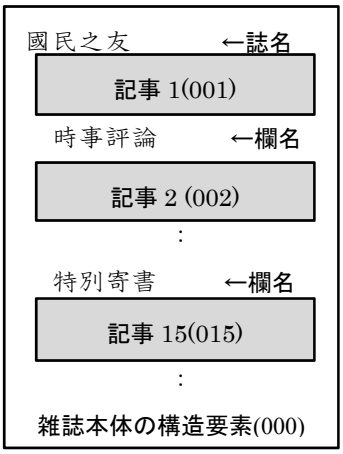


図 1　雑誌の構造

6 『女学雑誌』の場合は，雑誌表示の号番号ではなく，刊行年内での刊行順連番（1 始まりの整数）に直した値を使用する。例えば,『女学雑誌』第 387 号は刊行年の 1894 年において 27 番目に刊行された号であるから，サンプル ID における号番号は「27」とする。

上記の通り，記事連番「000」は，雑誌全体の本文から，記事連番「001」以下の個々の記事の本文を取り除いた，非常に特殊なテキストになっている。利用にあたっては，この点に留意されたい。

３．３　サンプルの種別

本コーパスには，「コア」と「非コア」と呼ばれる二つの種類のサンプルが存在する。限られた開発期間において全てのデータに高精度の形態論情報を付与するのは困難であったため，『明六雑誌』以外は人手処理を加え高精度の形態論情報付与を保証する「コア」と，機械解析の結果に一部修正を加えた「非コア」とに分けた。

使用に際しては，「非コア」は形態論情報の整備が不十分であり，「コア」に比して形態論情報の精度が低いことを留意されたい。このサンプルの種別は，「中納言」の検索画面での検索対象絞込みに用いることもできる（図 2）。

図 2　「中納言」の検索対象選択画面

なお，全てを「コア」とした『明六雑誌』を除き，他の雑誌における「コア」の設計に際しては，①言語の経年変化，②文体の差による言語の差異，③テキストの性質の差による言語の差異などを捉えることができるデータセットとすることを方針とし，以下の通り 3 段階層別ランダムサンプリングによって「コア」を定めた。

(1)　**「雑誌」**と**「発行年」**による計 10 層（『東洋学芸雑誌』1 層：1881-2／『国民之友』1 層：1887-8／『太陽』5 層：1895，1901，1909，1917，1925／『女学雑誌』1 層：1894-5／『女学世界』1 層：1909／『婦人倶楽部』1 層：1925）を第一の層別基準とし，各層 3-4 万語程度を取得する。

(2)　**「ジャンル」**（2 層：文芸<小説・戯曲・詩歌>，非文芸），**「地の文の文体」**（2 層：口語，文語）を掛合わせた 4 層を第二の層別基準とし，各層 7,500-10,000 語程度を取得する。

ただし，一部の層については次のように調整を行った。

- 『東洋学芸雑誌』1881-2 年は，文芸のサンプルが少なくかつ詩歌に偏ることや，地の文の文体が文語のサンプルのみであることから，「ジャンル」と「地の文の文体」を掛け合わせた層別を行わなかった。
- 『国民之友』1887-8 年は，殆どが文語のサンプルであるため，口語のサンプルは全てをコアとして取得し，残りの文語のサンプルの中から文芸・非文芸がほぼ等分になるようランダムサンプ

リングにより抽出した。

・『太陽』1909，1917 年の文芸層は，文語のサンプルの殆どが 1,000 語以下の韻文でバリエーションに乏しいため「地の文の文体」による層別を行わず，『太陽』『婦人倶楽部』1925 年は殆どが口語のサンプルであるため「地の文の文体」による層別を行わなかった。

(3) **「サンプル長」**による言語的特徴の差異を考慮し，サンプルの長さに偏りのないよう取得する。サンプル長分布調査に基づき，2000 語未満，4000 語未満，6500 語未満，それ以上の 4 層に層別し，偏りのないよう調整しサンプルを取得する。ただし，『東洋学芸雑誌』はサンプル長に偏りが少ないため，調整は行わなかった。

4．「中納言」の表示項目

本コーパスは，言語研究を目的とした利用に資する，様々な情報を有している。これらの情報は，コーパス検索アプリケーション「中納言」の検索結果表示画面において参照可能な形で提供される（図 3）。

図 3　「中納言」の検索結果表示画面

表 2 に，「中納言」の検索結果表示画面で参照できる付加情報のうち，初期設定で表示される項目と，本コーパスで特に注意が必要な項目として「文体」「部」「性別」「出版社」について，内容を示す。

表 2　「中納言」の検索結果表示項目（「*」を付けたものはオプション項目）

情報種別	項目名	内容
コーパス情報	サンプル ID	検索対象が含まれるサンプルの ID（→3.2 節）。
	開始位置	検索対象の先頭の文字の，サンプル内における位置を表す 10 刻みの連番。
	連番	検索対象の，サンプル内における短単位の 10 刻みの連番。
	コア	検索対象の含まれるサンプルのコア・非コアの別を表す。1 がコア，0 が非コアを示す。
形態論情報[7]	前文脈	検索対象の前方文脈。

7 形態論情報の個々の項目の意味は，「原文 KWIC」「原文文字列」「振り仮名」を除き，UniDic の見出しに対応している。BCCWJ や CHJ の仕様と同様であるため，詳細は小椋・小磯・冨士池ほか(2011)，国立国語研究所コーパス開発センター（近藤明日子）編(2016)を参照されたい。

ただし，漢文・外国語・未詳の方言・非コアデータの一部の外国地名・外国人名等，一部の文字列で UniDic ベースの形態論情報を付与しないものがある。これらは，形態論情報に関する項目のうち，「前文脈」「キー」

情報種別	項目名	内容
	キー	検索対象の書字形出現形（表記形）。
	後文脈	検索対象の後方文脈。
	原文 KWIC	上記項目「前文脈」「キー」「後文脈」に対する，校訂前の底本に近い形のテキスト（2.2 節参照）。上記各項目の下段に示す。
	語彙素読み	検索対象の語彙素（下記項目「語彙素」参照）の読み。片仮名表記である。
	語彙素	検索対象の語彙素の表記。語彙素は，単語の様々なバリエーション（語形，活用形，表記形など）を統合した辞書の見出しに相当するもので，一般の和語・漢語は漢字平仮名表記，外来語・人名・地名は片仮名表記である。
	語形	検索対象の語形。語形は，語彙素では統合される，語形の別（例：語彙素「矢張り」に対する「ヤハリ」「ヤッパリ」など）や活用型の別（例：語彙素「読む」に対する「ヨム（五段-マ行）」「ヨム（文語四段-マ行）」「ヨメル（下一段-マ行；可能動詞形）」など）等を区別した語の個々の形に相当するもので，片仮名表記である。
	語形代表表記	検索対象の語形に対する代表的な表記形で，語彙素に準じた表記である。
	品詞	検索対象の品詞で，UniDic の体系に基づく。学校文法における「形容動詞」は，語幹が「**形状詞**」，活用語尾が「**助動詞**」に分割される点に注意が必要である。
	活用型	検索対象活用語の活用の型。口語活用は活用の型と行で「**五段-サ行**」のように，文語活用は「文語」が加わり「**文語四段-サ行**」のように示される。検索対象の本文情報「文体」項目（下記項目「文体」参照）の値が「文語」である活用語には文語活用型を，「口語」である活用語には口語活用型を割り当てる。ただし，口語文体内にあっても口語活用型が存在しない語や活用形（文語形容詞「ごとし」，文語助動詞「き」，文語二段活用の連体形語形など）については文語活用型を用い，同様に，文語文体内にあっても文語活用型が存在しない語や活用形（口語助動詞「ない」，口語動詞「ある」等の終止形など）については，口語活用型を用いた(**例**：60M **明六** 1875_42001)。 **例**）其身分に因て夫々の權理**ある**と云ふ㆓は（**五段-ラ行/終止形-一般**）
	活用形	検索対象活用語の活用形。文法的に特定の活用形が期待される箇所（単語同士の接続関係や文末等）で，それとは異なる形態が用いられている場合は，語の形態に即して活用形を割り当てる(**例**：60M **女世** 1909_08036)。 **例**）御堂の邊に漫歩き**いたせ**し時、（**文語四段-サ行/已然形-一般**）

「後文脈」「原文 KWIC」「品詞」「原文文字列」「振り仮名」以外の項目が空欄となっている。

情報種別	項目名	内容
	原文文字列	検索対象の校訂前本文（原文）の文字列（→2.2 節）。
	振り仮名	検索対象に付された振り仮名の文字列[8]。原資料における振り仮名の誤植は訂正したものを示す。訂正の基準は本文校訂における誤植の判定（→2.2 節［誤植］）に準ずる。
本文情報	本文種別	検索対象が「地の文」以外の場合の，その種別。文献等からの引用，記事に対する雑誌記者・編集者の説明・解説・注釈等は「**引用**」，会話・独話・心内発話等の引用は「**会話**」と示す。
	話者	検索対象の本文種別が「引用」である場合の典拠文献名や著者名，「会話」である場合の話者名。
	文体*	検索対象が含まれる文の文体。日本語文については「**文語**」「**口語**」の別を示す。両者が分かちがたく混然とした文は「**混在**」とする。 文体の別を示すに適さない「**漢文**」「**外国語**（漢文以外の外国語文）」「**韻文**（日本語の韻文）」は，その種別を示す。 なお，一つのサンプル内で，引用文に異なる文体が用いられているなど，複数の文体が混在しているものは，引用範囲について個々に文体情報を付与する。
作品情報	ジャンル	検索対象が含まれるサンプルの，文章内容に基づく分類。小説・戯曲・詩歌の類を「**文芸**」，それ以外を「**非文芸**」と示す。また，『太陽』『女学雑誌』『女学世界』『婦人倶楽部』は，上記の分類に加えて「/」で区切り 3 桁の NDC（日本十進分類法）番号を示す。
	作品名	検索対象の含まれるサンプルが収録された雑誌の名称。
	成立年	検索対象の含まれるサンプルが収録された雑誌の発行年。
	巻名等	検索対象の含まれるサンプルのタイトル。本文中にタイトルが含まれる場合は，その文字列を示す。本文に含まれない場合は目次のタイトル表記等を，目次からも判明しない場合は代替として，欄名やサンプル本文の冒頭文字列などを〔 〕を付けて示す。サンプル ID 下 3 桁（記事連番）が「000」のサンプルは空欄とする。
	部*	検索対象の含まれる雑誌の部分け。『太陽』のみ同一雑誌内で部分けを行い，成立年により「1895・1901・1909・1917・1925」の 5 部に分ける。他の資料は部分けを行わず，空欄とする。

[8] ただし，表示対象は原本にある右ルビ（横書きの場合は上ルビ）のみ。また，雑誌『太陽』については，原本にある右ルビのうち，(a)文芸ジャンルのサンプルのすべて，(b)1895 年 1 号・4 号・7 号，1909 年 2 号・8 号・13 号，1917 年 2 号，1925 年 2 号・7 号・12 号のすべて，(c)熟字訓や読みの確定しづらい漢字列に振られたものの一部が表示対象。

情報種別	項目名	内容
作者情報	作者	検索対象が含まれるサンプルの著者名。著者の認定は，本文・目次の記載，その他原資料に関する目録や解説書・研究書等に基づく。ペンネーム等を用いて複数の名前でサンプルを書いている著者は，一般的に知られている呼称に統一した場合がある。また，著者が不明なものは「＊」で示す。 日本の作品の著者・翻訳作品の原作者は氏名に「(作)」を，翻訳作品の訳者は氏名に「(訳)」を付して示す。複数の著者（翻訳作品の原作者・訳者を含む）は「/」で区切って併記する。 著者が「国立国会図書館典拠データ検索・提供サービス（Web NDL Authorities）」（http://id.ndl.go.jp/auth/ndla/）に収録されている場合は，そのウェブページへのリンクを付与する。
	性別*	検索対象が含まれるサンプルの著者の性別を示す。不明な場合は「＊」で示す。
	生年	検索対象が含まれるサンプルの著者生年。西暦 4 桁で示す。不明な場合は「＊」で示す。生年の認定は，各種人名事典・Web NDL Authorities，雑誌本文での記述等に基づく。
底本情報	底本	検索対象が含まれる原資料。『明六雑誌』『国民之友』は，雑誌名に「〈 〉」で号数を加えて示す。『太陽』『女学雑誌』『女学世界』『婦人倶楽部』は，雑誌名に「〈 〉」で発行年と号数を加えて示す。『東洋学芸雑誌』は雑誌名に「〈 〉」で号数を加え，その後ろに版数を示す。
	ページ番号	検索対象の原資料における出現ページ番号。『明六雑誌』（和装本）は丁数に加えて「オ」(表)「ウ」(裏)により表裏の別を示す。他の雑誌は，原則として原資料のページ番号を示すが，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，コーパス独自に付けたページ番号とする。
	出版社*	検索対象が含まれるサンプルが収録された雑誌の出版社を示す。
その他	底本リンク	検索対象の原資料画像へのリンク。『明六雑誌』『国民之友』は，国立国語研究所蔵本画像へのリンクを「Ninjal」ボタンで，『太陽』は小学館「ジャパンナレッジ JK Books」提供画像（閲覧には，「ジャパンナレッジ JK Books」の契約が必要）へのリンクを「JK」ボタンで示す。『明六雑誌』『国民之友』はページ単位，『太陽』は見開き単位での対応となっている。なお，『東洋学芸雑誌』『女学雑誌』『女学世界』『婦人倶楽部』については原資料画像へのリンクを提供していない。

情報種別	項目名	内容
	参考リンク	検索対象の参考資料画像へのリンク。こちらはコーパスの原資料ではないことに注意が必要である。『東洋学芸雑誌』は，国立国語研究所蔵『東洋学芸雑誌』第二版へのリンクを「Ninjal」ボタンで示す。なお，『東洋学芸雑誌』以外の雑誌については，参考資料画像へのリンクを提供していない。

参考文献

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』形態論情報規定集 第 4 版（上）（下)』(特定領域研究「日本語コーパス」平成 22 年度研究成果報告書）国立国語研究所

国立国語研究所コーパス開発センター(近藤明日子)編(2016)『近代文語 UniDic 短単位規程集 Ver.1.1』

須永哲矢・堤智昭・高田智和(2011)「明治前期雑誌の異体漢字と文字コード―『明六雑誌』を事例として―」『じんもんこん 2011 論文集』2011(8), pp.381-388

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』2013(4), pp.201-208