異なる文体の混在するテキストに対する複数辞書切り替えによる解析手法の提案

間淵 洋子・小木曽 智信（人間文化研究機構　国立国語研究所）

国立国語研究所では現在，形態論情報を付与した『太陽コーパス』を構築している．文語から口語への文体移行期に刊行された総合雑誌『太陽』には，文語と口語という性質の大きく異なる複数の文体が混在する文章が多く含まれるため，文語文用解析辞書と旧仮名遣いの口語文用解析辞書のいずれかを指定して用いる従来の形態素解析手法では，精度を保つことが困難である．

そこで，本コーパスの構築にあたっては，テキストが有する文体情報を利用し，複数の辞書を切り替えて，部分ごとに適応する辞書によって解析する手法を試みた．

この手法の有用性を確認するため，評価用のデータを作成し，従来手法との解析精度を比較した結果，提案する複数辞書切り替え手法によって，解析精度が向上することを確認できた．

Proposal for a Morphological Analysis of Texts Mixing Different Styles by Selecting between Multiple Dictionaries

MABUCHI Yoko / OGISO Toshinobu
(National Institute for Japanese Language and Linguistics)

The National Institute for Japanese Language and Linguistics is presently constructing the morphologically annotated ‘Taiyo Corpus’. In contrast with many other historical Japanese corpora, this corpus includes many texts mixing two different writing styles: classical and colloquial. For the construction of this corpus we propose a new morphological analysis method using multiple dictionaries. By the proposed method, we select and apply the dictionary appropriate to the writing style in every text block. Comparison shows that the precision of the proposed analysis exceeds that of previous analyses.

１．まえがき

国立国語研究所（以下，国語研）では，現在，『日本語歴史コーパス』[1]の一環として，形態論情報を付与した『太陽コーパス』の構築を進めている．『太陽コーパス』は，明治期の言文一致運動を経て，書き言葉が文語体から口語体へと入れ替わる過渡期のテキストを収録したものである．そのため，文語体・口語体，それぞれで書かれた文章が含まれるほか，文語体の地の文に口語体の会話文が現れるような，一文章中に複数の文体が混在する文章を含む点が特徴的である．

このような異なる文体の混在するテキストに対して，今回，これまで国語研で開発してきた文語用，口語用それぞれの形態素解析辞書を切り替えて解析する新しい手法を試みた．本発表では，この解析手法導入の背景となる『太陽コーパス』の既存の情報・データ仕様を示すと共に，実際に用いた手法について，従来手法との比較を通して，その有用性を報告する．

２．『太陽コーパス』の形態論情報付与における問題点

『太陽コーパス』とは

現在，形態論情報を整備している『太陽コーパス』は，2005 年に公開されたテキストコーパスである『太陽コーパス』 [2]に対して，新たに形態論情報を付与して 2016 年春以降公開予定のコーパス（約 960 万語）である．

元となる 2005 年度版『太陽コーパス』は，明治後期～大正期の総合雑誌『太陽』（博文館刊）の 1895 年，1901 年，1909 年，1917 年，1925 年の全文を格納した約 1450 万字のテキストコーパスである。形態論情報は付与されていないものの，記事を単位として，雑誌巻号・発行年等の書誌情報，著者情報，文体・書記体情報が付されているほか，文章中の引用情報（範囲，発話か典拠引用かの種別，発話者や典拠，文体等），文字・校訂情報などが詳細に XML 形式でマークアップされている[3]. 言文一致を経て文語文から口語文へと移り変わる，現代語の確立期の文章を大規模に収集したコーパスとして資料価値が高く，日本語学分野の研究に多く利用されているが，近年の形態論情報付きの現代語コーパス・歴史コーパスの充実を受けて，形態論情報の付与が望まれてきた．

このような要求を背景に，国語研ではこれまで，雑誌『太陽』と同時代の近代語テキストに対する

形態論情報付与に適した解析辞書として，近代文語文用の解析辞書「近代文語 UniDic」[4]，および，旧仮名遣いで書かれた口語文用の解析辞書「旧仮名口語 UniDic」[5]の開発を行い，これらを用いて，近代語の形態論情報付きコーパスのパイロットとして『明六雑誌コーパス』を構築・公開[6]，続いて『国民之友コーパス』を公開した[7]．これらの先行する形態論情報付き近代語コーパスを礎に，『太陽コーパス』の形態論情報整備は進められている．

『太陽コーパス』構築の問題点

しかし『太陽コーパス』は，規模や含まれる文章の文体等において，先行のコーパスと大きな性質の差を有している．表 1 に形態論情報付き近代語コーパスの語数，表 2 に『太陽コーパス』の文体別記事数を示す．

表 1 から分かるように，『太陽コーパス』は『明六雑誌コーパス』の約 53 倍，『国民之友コーパス』の約 9.5 倍にも上る圧倒的な分量を含む．また，言文一致の進んでいない年代の先行コーパスが，文語体を中心とした文章から成るのに対して，『太陽コーパス』は収録年を反映し，文語・口語両文体の文章を有するほか，両文体が一つの文章に混在している文章をも含む点で，性質が極めて異なっている．表 2 に見るように，『太陽コーパス』においては，文語記事中の 3-10%の記事が口語体の混在する記事となっており，また，口語記事中の 20～50%近い記事に文語体が混在している．

表 1 形態論情報付き近代語コーパスの概要

雑誌名	収録年	語数(万語)	文体
明六雑誌	1874-1875	18	文語
国民之友	1887-1888	101	文語中心
太陽	1895, 1901, 1909, 1917, 1925	960	文語，口語，混在

表 2 『太陽コーパス』の文体別記事数

年	文語	左のうち口語混在	口語	左のうち文語混在
1895	690	43(6.2%)	52	25(48.1%)
1901	484	26(5.4%)	183	41(22.4%)
1909	297	18(6.1%)	420	119(28.3%)
1917	139	5(3.6%)	398	118(29.6%)
1925	60	6(10%)	1024	422(41.2%)

このような多様な文章を含むコーパスに対して，従来の手法を用いて一定精度を保った形態論情報付きコーパスを構築するのは，非常に困難であるという問題がある．特に，文語と口語のように差異の大きい文体が一つの文章に混在する場合の形態素解析については，文語用・口語用それぞれの解析辞書の守備範囲に適合しない文章範囲の解析において精度を落とすことが予想される．例として，文語辞書（近代文語 UniDic）で口語文「これぢや耐らん」を解析した結果を表 3 に，口語辞書（旧仮名口語 UniDic）で文語文「立去らむとしたりしが」を解析した結果を表 4 に示す．

表 3 文語辞書による口語文の解析結果例

書字形	語彙素	読み	品詞	解析活用型	正解
これ	此れ	コレ	代名詞		
ぢや	で	デ	助詞-格助詞		
耐ら	堪る	タマル	動詞-一般	***文語四段-ラ行***	五段
ん	***む***	***ム***	助動詞	***文語助動詞-ム***	助動詞「ず」

表 4 口語辞書による文語文の解析結果例

書字形	語彙素	読み	品詞	解析活用型	正解
立去ら	立ち去る	タチサル	動詞-一般	***五段-ラ行***	文語四段
む	***ず***	***ズ***	助動詞	***助動詞-ヌ***	文語助動詞-ム
と	と	ト	助詞-格助詞		
し	為る	スル	動詞-非自立可能	***サ行変格***	文語サ変
たり	たり	タリ	***助詞-副助詞***		文語助動詞「たり」
し	***為る***	***スル***	***動詞-非自立可能***	***サ行変格***	文語助動詞「き」
が	が	ガ	助詞-接続助詞		

表内の太斜字体は，誤解析箇所である．いずれの場合も，助動詞の認定や解析活用型の選択などに解析の誤りが生じることが分かる．これらは，文体に即した適切な辞書による解析においては，生じない誤りである．

3．複数辞書切り替えによる解析手法の提案

そこで，形態論情報を付与した『太陽コーパス』の構築にあたっては，既にコーパスに付与されている文章ブロックの文体情報を利用し，異なる文体の混在するテキストに対して，複数の解析辞書を切り替えて解析する手法を新たに用い，精度向上を図ることを試みた．

既公開の形態論情報付き近代語コーパス構築においては，それぞれの記事（解析対象となるファイルの単位に相当）に付与された文体の情報に基づき，文語体の記事では「近代文語 UniDic」を，口語体の記事では「旧仮名口語 UniDic」を選択し解析を行ってきたが，今回の提案手法では，べー

スとなる記事の文体に対応する辞書を辞書 1，他方を辞書 2 と設定し，下位のブロック要素に上位ブロックと異なる文体情報が現れた場合，その文章ブロックに対して辞書 2 を用いて解析を行う．これにより，文章ブロックの文体に即して辞書を選択し解析を行うことができるようになる．

例えば，図 1 に示した尾崎紅葉「取舵」という小説では，地の文が近代文語文体で書かれている一方で，登場人物の会話文は砕けた口語文体であり，両者は使用語彙や語法の面で大きな差のあることが分かる．この場合，従来手法では，記事（article 要素）の文体情報「文語」（style 属性値）に即して文章全体を「近代文語 UniDic」で解析することになる．

しかし，提案手法では，ベース文体「文語」に即して辞書 1 に「近代文語 UniDic」を，辞書 2 に「旧仮名口語 UniDic」を設定し，ベース文体の文章ブロック（地の文）を辞書 1「近代文語 UniDic」で，内部に現れる引用範囲（quotation 要素）の文体情報（style 属性値）に基づき，口語の文章ブロック（会話文）を辞書 2「旧仮名口語 UniDic」で解析する．

<article title="**取舵**" author="**尾崎紅葉**" style="**文語**">
:
渠は此惨澹と滓熱とに面を皺めつつ、手荷物の鞄の中より何やらむ取出して、忙々立去らむとしたりしが、忽ち左右を顧て …**辞書 1 解析対象**

<quotation type="speech" source="**学生**" style="**口語**">
「皆様、これぢや耐らん。ちと甲板へお出なさい。涼くツて奈何なに心地が快か知れん。」
…**辞書 2 解析対象**
</quotation>

是空谷の跫音なり。 …**辞書 1 解析対象**
:
</article>

『太陽』1895 年 1 号「取舵」尾崎紅葉

図 1 コーパスの文体情報と解析辞書切り替え

この提案手法では，それぞれの解析辞書が，本来得意とする文体の文章を，分担して解析することができるため，より的確な解析結果を得られることが期待できる．

4．評価

評価方法

本提案手法（複数辞書切り替え）を評価するために，従来手法（単独辞書）との比較により精度評価を行った．

評価に際しては，ベースとなる文体（地の文の文体）が文語である記事と口語である記事とに分けた上で，①提案手法，②従来手法（ベース文体に対応する単独辞書使用），③参考用準従来手法（ベース文体に対応しない単独辞書使用）の 3 手法による解析精度を求めて比較した．

使用した解析辞書は，文語用に「近代文語 UniDic」（人手修正済み文語体近代語コーパス約 67 万語を学習データとして生成したもの），口語用に「旧仮名口語 UniDic」（人手修正済み口語体近代語・現代語コーパス約 248 万語を学習データとして生成したもの），見出し語数約 142 万語，解析器は「MeCab(ver.0.993)」を用いた．

評価データ

評価データは，データ構築中の『太陽コーパス』に含まれる，短単位情報人手修正済みの文体混在テキスト 24 記事，約 12.7 万語である（上記学習データを含まない）．表 5 にその内訳を示す．また，評価データの文章例を図 2，3 に示す．

表 5 評価用コーパスの内訳

ベース文体	記事数	語数（万語）	記事例
文語	9	4.9	樋口一葉「ゆく雲」，幸田露伴「新學士」，田山花袋「淺間横斷記」
口語	15	7.8	南方熊楠「蛇に關する民俗と傳説」，佐々木信綱「賀茂真淵雑話」

上がりて見れば既長火鉢に火もありて鐵瓶も沸り居れり、
臺所には水瓶あり手桶あり小桶あり、
塵取り、俎板、さるぼう、柄杓、よろづさし、庖丁さし、杓子の數々、竈に釜茶釜、棚に鐵鍋青銅鍋の大中小、庖丁鐵灸餅網鐵串、灰篩、火箸、小刀、悉皆揃へり、瀬戸物類の一つも見えぬを不審する時、二人乗の車の音ごろごろとして、
これは皆水口へ廻はして
といふ聲は浪奴に疑ひ無し、
浪奴戻つたか
と云へば、
旦那様御歸りになりましたの
と云ひながら青繪の手水鉢の品よきを手にして坐敷を通り抜け、縁側より竿石の上にそつと置き、
手水鉢には惜しいなれど何と好いではござりませぬか、
源公、これに水をなみなみ顧ひたい、
お勝殿臺所づかひの瀬戸物の受取り方が濟んだらば一寸手水鉢の柄杓を持つて來て下さい、
（1895 年 5 号「新學士」幸田露伴）

図 2 評価データの文章例
文語ベースの記事（斜字体は口語部分）

古今要覽稿巻五三一に
「凡そ十二辰に生物を配當せしは王充の論衡に初て見たれども、淮南子に山中未の日主人と稱ふるは羊也、
莊子に未嘗爲牧、而群生於奥と云るを釋文に西南隅の未地と云しは羊を以て未に配當せしも其由來古し」
と論じた。
果して其通りなら十二支に十二の動物を配る事戰國時既に支那に存したらしく、淮南子に巳日山中稱寡人者社中蛇也と有る、
蛇を以て巳に當たのも前漢以前から行れた事だらう歟。
凡て蛇類は好んで水に近づき又之に入る。
沙漠無水の地に長じた蛇すら能く水を泳ぎ、印度で崇拜さるる帽蛇は井にも入れば遠く船を追て海に出る事も有り。
去れば諸國で所謂水怪の多くは水中又水邊に棲む蛇で有る。
（1917 年 1 号「蛇に關する民俗と傳説（一）」南方熊楠）

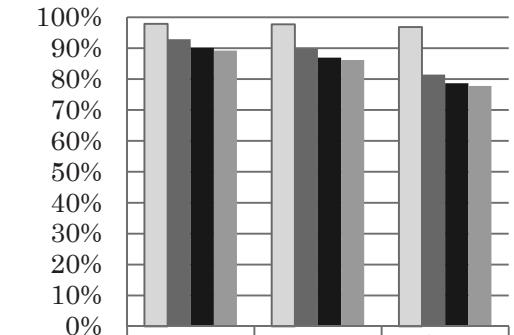
図 3 評価データの文章例
口語ベースの記事（斜字体は文語部分）

評価結果

精度評価の結果を，図 4 および図 5 に示す．図 4 はベースとなる文体が文語である評価データを，図 5 はベースとなる文体が口語である評価データを，それぞれ上記①②③の手法により解析した結果の精度をグラフにしたもので，数字は F 値（Precision と Recall の調和平均）をパーセント表示にした値である．

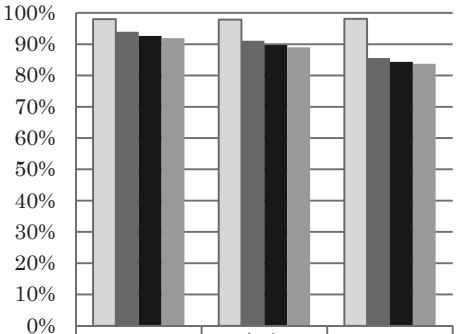
各手法の評価レベル「L1 境界」は単語境界の認定の正しさを，「L2 品詞」は，L1 に加え品詞認定の正しさを示す．「L3 語彙素」は L1・L2 に加えて語彙素（辞書見出し）認定が正しいことを意味し，例えば，「金」が「きん」か「かね」かを正しく解析できたかどうかといった点を評価したものである．「L4 発音形」は L1〜L3 に加えて語形認定が正しいことを意味し，同一の語彙素ながら，読みのバリエーションがある場合，例えば「言語」が「ゲンゴ」か「ゴンゴ」か，「行く」の連用形「行(て)」が「ユキ(テ)」か「イッ(テ)」か等を正しく解析できたかどうかといった点を評価したものである．

主に，「L3 語彙素」の認定について解析精度を見てみると，図 4 より，文語ベースの記事の場合，①提案手法は②従来手法（近代文語用辞書単用）より 3.2 ポイント，③参考用準従来手法（旧仮名口語用辞書単用）より 11.3 ポイント程度上回っていることが分かる．また，図 5 より，口語ベースの記事の場合，①提案手法は，②従来手法（旧仮名口語用辞書単用）より 2.9 ポイント程度，③参考用準従来手法（近代文語用辞書単用）より 8.3 ポイント程度上回っていることが分かる．



	提案手法	近代文語 Unidic	旧仮名口語 Unidic
L1 境　界	97.855%	97.698%	96.824%
L2 品　詞	92.883%	89.750%	81.459%
L3 語彙素	90.092%	86.939%	78.669%
L4 発音形	89.220%	86.138%	77.761%

図 4 文語記事の解析精度評価



	提案手法	旧仮名口語 Unidic	近代文語 Unidic
L1 境　界	98.003%	97.853%	98.143%
L2 品　詞	93.974%	91.114%	85.610%
L3 語彙素	92.633%	89.717%	84.365%
L4 発音形	91.948%	89.024%	83.754%

図 5 口語記事の解析精度評価

以上の評価結果より，いずれのベース文体においても，提案手法は従来手法より解析精度が向上することが確認できた．

UniDic の解析精度と提案手法

ここで，解析に用いた「近代文語 UniDic」と「旧仮名口語 UniDic」の標準的な解析精度と，今回の評価における解析精度を比較しておく．「L3 語彙

素」認定の F 値における「近代文語 UniDic」の近代文語文解析精度は 96.48%[4],「旧仮名口語 UniDic」の口語文解析精度は 93.99%[3]であるが，今回の精度評価では，提案手法を含めたいずれの手法においても，特に「近代文語 UniDic」においてその精度をかなり下回る結果となった．

これは，評価データとした文体の混在するテキスト（表 5 参照）に，文学的性質の強いものが多く，独特・難解な語用・語法が多いことや，それに相関して現段階での評価コーパスの人手修正が不十分であることに起因することが推測される．

4．エラー分析

前節に示した提案手法の解析精度を踏まえ，提案手法において，どのような場合に解析の誤りが生じ，どのような誤りが多いのかについて調査を行った．評価レベルごとに，エラーの類型を示す．

「L1 境界」のエラー

単位境界のエラーについては，以下の 2 種のエラーが目立った（「／」は単位境界を示す）．

(1) 助詞を含む接続詞等の連語的要素と，分割されるべき要素間のエラー
　例）世の無常を嘆息し遂に死を決して毒の盛つてある杯を飲まんとする時に當つて、
　誤：決して（接続詞）　正：決し(動詞連用形)／て(接続助詞)
(2) 三字以上の漢字連続における語構成上の曖昧性に基づくエラー
　例）青年といふものは人生の好時節である、
　誤：好時／節　　正：好／時節

「L2 品詞」のエラー

品詞のエラーについては，以下のパターンが目立った．

(1) 同形異品詞語，助詞下位分類間のエラー
　例）第三十八議會に於て既に一種の奇功を奏し又た第三十九議會に於ても或は然らんかと思はるる人は
　誤：副詞　　正：接続詞
(2) 同一出現形となる助動詞・助詞，動詞連用形・連用形転生名詞等のエラー
　例）些とお遊びに入らして、
　誤：名詞　　正：動詞連用形（「お＋連用形＋になる」型敬語表現）
　その他語例）助動詞「だ」≒格助詞「で」「に」
　助動詞「たり」≒格助詞「と」

「L3 語彙素」のエラー

語彙素（辞書の見出し語）の認定エラーについては，以下のパターンが目立った．特に，同一の漢字表記に対して，異なる複数の別語が割り当てられる場合のエラーが多く見られた．

(1) 異語同表記形の曖昧性による選択エラー
　例）又宇内造艦界の大勢が何故に滔々として大艦主義に傾きつつあるのであらうか
　誤：ナゼ　　正：ナニユエ
　その他語例）居る(イル≒オル)，入る(ハイル≒イル)，出る(デル≒イデル)等
(2) 平仮名表記形の曖昧性による選択エラー
　例）無師無錢を口實にして學問成就せずといふものあらば、
　誤：物（モノ）　　正：者（モノ）

「L4 発音形」のエラー

語形・発音形のエラーについては，以下のパターンが目立った．特に，文語文における音便形を，口語活用型と判定する誤りが多く見られた．

(1) 解析活用型や活用形の選択エラー
　例）先切幕をきつて出たまふ時どうやらもつたいがあるやうでなし。
　誤：五段-ラ行　　正：文語四段-ラ行
(2) 複数の読みを有する語の読み選択エラー
　語例）日本(ニホン≒ニッポン)，何(ナニ≒ナン)，人(ニン≒ジン)等
(3) 数詞の読みの選択エラー
　例）此日は一週の終の日にはあらざりしかど、
　誤：イチ　　正：イッ

上記に示したエラーの多くは，辞書の選択や辞書項目の有無によるものではない．出現形（表記形）の読みの曖昧性による語彙素や語形の誤認定については，テキストに既存の読み情報（ルビ等）の利用などによって解消される可能性があり，また，品詞選択や語構成の判断に関するエラーについては，学習用コーパスの充実等により解決が図れる可能性がある．いずれも今後検討する．

その他のエラー

その他のエラーとして，以下のものが少なからず見受けられた．これらは，今後，辞書の拡充や評価データの整備によって解消されるものであると思われる．

(1) 外国語等の未知語，活用語尾の特殊表記
　例）其大さを記して四足有りと言ぬを見ば
　誤：連用形「ミ」　　正：仮定形「ミレ」
(2) 文学的な特殊な熟字訓に相当する表現
　例）有ふれた本邦の蛇の中で一番大いから之を支那の巨蟒に充た者か
　誤：巨(キョ)／蟒(ウワバミ)　正：巨蟒(オロチ)

(3) 評価データのエラー
例) 此邊土人の云るには
誤：邊／土人　　正：邊土／人

更に，これらの他に，評価データの文境界情報が正しくないこと（句読点を一律に文区切りとしているため，本来文末でない箇所が文末になっている）に起因する，活用形（終止形／連体形間等）の判定誤りが多くあったが，これらは，解析対象データの整備によって解消される可能性が高い．

また，評価データの文体情報に誤りがあったこと（タグ付けの誤りや，タグ付け対象外の箇所における文体の混交）によって，文体情報を手掛かりとした辞書切り替え手法が，有効性を発揮できないものもあった．これについては，文体情報が付加されていないデータへの適用も視野に入れ，今後検討すべき問題である．

5．まとめと今後の課題・展開

本研究で提案した，異なる複数の文体が混在するテキストに対して，文体に適した複数解析辞書を切り替えて解析する手法により，精度向上を図ることが可能であることが分かった．

一方で，本手法を用いてもなお，解決されない解析の誤りがあることが判明した．今後は，エラー分析結果を元に，更なる解析手法や解析辞書の開発に取り組む必要がある．

『太陽コーパス』の文章のように，文体が混在するテキストは，近代語のテキストにおいては，特に明治期の小説（地の文が文語，会話文が口語）や，明治後期以降の古典籍に関する学術的論説（地の文が口語，引用文が文語）などで多く見られるため，今後これらの資料をコーパス化する際には本手法を適用することが妥当である．

また，近代語テキストに限らず，現在国語研において開発中の近世口語資料のコーパスなど，大きく異なる複数の文体を含むテキストの解析に利用できるほか，以下のような場面への応用も可能であると考える．今後も検討を続けたい．

- 「地の文」（書き言葉）と「会話文」（話し言葉）のように，スタイルの異なる文章ブロックを，最適な辞書で解析する．
- これまで文体別にテキストの範囲を切り分けた上で解析を行っていたものを，多様な文体の文章を含むより大きな文章ブロック（雑誌，新聞のような多様な記事を収録した媒体全体など）に対して，記事の切れ目を意識することなく，解析する．
- 自動文体判別等の手法を取り入れることで，文体を意識することなく解析する．

6．あとがき

本研究の提案手法により解析し，整備を進めている形態論情報付きの『太陽コーパス』は，2016年春以降に，国語研の既公開形態論情報付きコーパスと同様，コーパス検索システム『中納言』によって公開することを予定している．今回の評価およびエラー分析に基づくコーパス情報の修正・整備も，喫緊の必須課題である．

解析手法についての研究や解析辞書の整備を進めることと同時に，精度の確保され充分な規模を有するコーパスを開発することで，両者の相乗的な発展が実現されるものと考える．今後も並行的に研究およびコーパス構築を続ける予定である．

付記

本研究は，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」（プロジェクトリーダー：田中牧郎），及び，「通時コーパスによる日本語史研究の新展開」（第 3 期準備プロジェクト，プロジェクトリーダー：小木曽智信）の成果を発展させたものである．

参考文献

1）田中牧郎：『日本語歴史コーパス』の構築，日本語学，Vol.33，No.14，pp.56-67(2014).
2）国立国語研究所：太陽コーパス，博文館新社(2005).
3）田中牧郎：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，国立国語研究所編，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―，pp.1-48，博文館新社(2005).
4）小木曽智信：旧仮名遣いの口語文を対象とした形態素解析辞書，じんもんこん 2012 論文集，7，pp.25-32 (2012).
5）小木曽智信，小町守，松本裕治：歴史的日本語資料を対象とした形態素解析，自然言語処理，Vol.20，No.5，pp.727-748(2013).
6）近藤明日子，小木曽智信，須永哲矢，田中牧郎：『明六雑誌コーパス』の開発―近代語コーパスのモデルとして―，第 2 回コーパス日本語学ワークショップ予稿集，pp.329-334(2012).
7）近藤明日子：『国民之友コーパス』解説書　第 1.1 版，国立国語研究所（オンライン），入手先〈 http://www.ninjal.ac.jp/corpus_center/cmj/doc/kokumin_manual_v1_1.pdf〉（2014）．