国立国語研究所学術情報リポジトリ

How to Correctly Morphologically Analyze Text Containing a Mixture of Old- and New-Style Kanji Scripts

メタデータ	言語: jpn 出版者: 公開日: 2021-07-16 キーワード (Ja): キーワード (En): 作成者: 松田, 謙次郎, MATSUDA, Kenjiro メールアドレス: 所属:
URL	https://doi.org/10.15084/00003440

国立国語研究所論集 (NINJAL Research Papers) 21: 123–132 (2021)
ISSN 2186-1358

新漢字と旧漢字が混在したテキストからの短単位形態素の抽出について

松田謙次郎

神戸松蔭女子学院大学／国立国語研究所 共同研究員

要旨

旧字体と新字体の混在するテキストは，形態素解析において誤解析の原因となることが多く，その対策としては形態素解析辞書の記載に異体字を加える方法，そして予め漢字を新字体に置換しておく方法，また複数の辞書を使い分けるといった方法が考えられる。本稿では字体置換 6 通りと，辞書の使い分け 3 通りを掛け合わせた 18 組の組み合わせで國 / 国，會 / 会，關 / 関 3 対の旧 / 新字体の対を含んだテキストの形態素解析を行うことで，目的とする漢字を含む形態素がどれほど正確に切り出せるのかを検討した。データとして第 1 ～ 10 回までの国会会議録を用いた。結果は，漢字置換で隣接する漢字が旧字体の場合に旧字体に置換し，隣接しない場合は新字体とするという置換法（デフォルトを新字体とする日和見置換）と，すべてについて近代文語 UniDic を用いるか，1949 年の当用漢字字体表告示を境として，それ以前では近代文語 UniDic を用い，それ以後では現代語書き言葉 UniDic を用いる方法が，もっとも正確に当該漢字を含む短単位形態素を切り出せるというものであった。形態素解析辞書の記載に異体字を加える方法には，異体字が記載されていない形態素が出現した場合に対応ができないという欠点があるのに対して，漢字置換と辞書の使い分けを活用する方法は，そうした場合にも柔軟に対応が可能であるという利点があることを主張した *。

キーワード：形態素解析，国会会議録，旧字体，新字体，当用漢字字体表

1. はじめに

筆者は現在国立国語研究所機関拠点型基幹研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」の一環として，当用漢字字体表告示前後の日本語書き言葉における旧字体から新字体への変遷を辿るべく，国会会議録の分析を行っている（松田 2019）。その過程で，少なくとも國 / 国のペアでは，それを含む漢語（例：国会，国際，国務）の違いが各字体での出現率に大きな影響を及ぼすという仮説を得るに至った。この仮説を検証するためには国会会議録のテキストから形態素解析により漢語を正しく抽出する必要があるが，そこで戦後直後期の国会会議録のような，漢字旧字体と新字体が混在するテキストから，形態素解析を通して特定漢字を含む短単位[1]形態素を正しく抽出することの困難さに直面した。

* 本稿は，国立国語研究所の共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」（プロジェクトリーダー：小木曽智信）の研究成果であり，JSPS 科研費 19H00531，16H03420，18K00632 の助成も受けている。執筆にあたっては，小木曽智信先生，乙武北斗先生，高田智和先生，田中牧郎先生，前川喜久雄先生より貴重な御教示を賜った。また，論文中で使用したダウンローダは堤智昭先生の作成によるものである。この場をお借りして厚く御礼申し上げる。ただし，本稿の責任はすべて著者にある。

[1] 短単位とは，形態的側面に着目して規定された言語単位である。国語研の「日本語話し言葉コーパス」や「現代日本語書き言葉均衡コーパス」，さらに形態素解析辞書 UniDic の見出し語として採用されており，基準がわかりやすく，ゆれが少ないという特徴がある。たとえば「国立国語研究所が日本語コーパスを構築した」という文は，短単位だと「国立 / 国語 / 研究 / 所 / が / 日本 / 語 / コーパス / を / 構築 / し / た」のように分割される（小椋 2014: 74）。短単位については，小椋ほか（2007），小椋ほか（2011）も参照のこと。

歴史的文書の形態素解析については，国語研のコーパス構築プロジェクトの目覚ましい進展により，現在では時代別，さらに書き言葉と話し言葉といったレジスター差に基づく辞書も整備されている。辞書を適宜使い分けることにより，当然解析の精度は向上する。また，解析の前処理として，正しい解析がなされるように予め対象とする文字を変換しておくことも考えられる。たとえば，特定の漢字について，そのすべてを新字体あるいは旧字体に変換することが考えられる。それでは，戦後間もない時期の書き言葉に見られるような，旧漢字と新漢字が混在したテキストから，できるだけ正しく特定漢字を含む短単位形態素を抽出するには，これらの方法をどのように組み合わせれば良いのであろうか。本稿は，このような問題意識から出発して計画・実施した実験の結果を報告するものである。

2. 問題の所在

戦後直後の国会会議録では，(1) ～ (3) にあるように，漢字の旧字体と新字体との入り混じった使用が認められる。

(1) この規定が設けられることによりまして、一切が國会の御承認を得て支給され、又給與としてこれが表面に現われまして、その人に対する給與額として考慮の中に入れられる。［1948 年 12 月 21 日 第 4 回国会参議院大蔵・人事・労働連合委員会第 5 号］

(2) さて、簡易生命保險及び郵便年金積立金は現在総額百三十億円をこえる巨額を算しているのでありますが、元來この積立金は保險年金事業の責任準備金を本体となすものでありまして、御承知の通り、國営たると民営たるとを問わず、保險のような事業では、事業経営者が資産の運用利まわりと死亡率、事業費等を経営上の要素をよくにらみ合せて保險料率その他を定め、事業全般の運営をいたすのが当然でありまして、資産の運用をみずからいたさないような保險年金事業は、完全なる独立企業とは申しがたいのであります。［1949 年 5 月 12 日 第 5 回国会衆議院本会議第 27 号］

(3) 第一は、一般病害虫の中で広範囲に亘り急激に発生して、農作物に甚大な損害を及ぼす虞れのある病害虫を指定し、これら指定病害虫の異常発生時に備えて、国において農薬の備蓄及び防除機具の備付を行い、必要に応じ農薬の譲與、譲渡及び機具の貸付をなし得ることとなし、更に防除実施者に対して補助を行うことができることにしました。［1951 年 6 月 2 日 第 10 回国会参議院農林委員会第 44 号］

ここで漢語（形態素）に注目すると，下線部にあるように，同一漢語（形態素）内で旧漢字と新漢字で統一されている場合とそうでない場合が存在する。こうした場合，(4) にあるように新字体漢字のみを含む文字列は，漢字を置換することなく形態素解析器 MeCab[2] と「現代語書き言

[2] https://taku910.github.io/mecab/（2021 年 2 月 15 日確認）. MeCab については Kudo et al.（2004）を参照のこと。

葉 UniDic」[3] を用いて正しく形態素解析 [4] され，(5) のような旧字体のみからなる漢語を含む文字列は，同様に「近代文語 UniDic」[5] を用いて正しく解析される。しかし，(6) (7) にみるように，両者が同一形態素内で混在する場合は，どちらの辞書を使用しても本来は同一短単位であるはずの「國営」を 2 つに分割し，誤った解析を行ってしまう。ただし，すべての場合にこうした解析がなされるわけではなく，(8) (9) にあるように，旧字体＋新字体の組み合わせでも，少なくとも「國民」に関しては正しい切り分けがされることもあり，形態素により異体字の許容範囲が異なることがわかる（斜線は形態素の切れ目を，［ ］内は使用された辞書を表す）。

(4) 国営 / 自動 / 車 / 全体［現代語書き言葉 UniDic］

(5) 各 / 國營 / 重要 / 産業 / の［近代文語 UniDic］

(6) 國 / 営 / 競馬 / 賞金［現代語書き言葉 UniDic］

(7) 國 / 営 / 競馬 / 賞金［近代文語 UniDic］

(8) 國民 / 金融 / 公社［現代語書き言葉 UniDic］

(9) 國民 / 金融 / 公 / 社［近代文語 UniDic］

このような異体字の許容範囲の差は，ひとつは辞書の記載への異体字の登録に起因するものと考えられる。そこで，こうした異体字による誤解析への当然の対処法として，当該漢字を含むすべての形態素記載に異体字表記を含めるという解決法が浮上する（高田 2002）。辞書への異体字の登録は，たしかに今回のような誤解析への有力な解決法である。ただし，たとえば國 / 国の場合，第 1 〜 10 回国会会議録中の漢語の異なり数は 357 件に及ぶ。これらの中にはすでに異体字表記が含まれているものが相当数あるとしても，辞書登録に要する手間は無視できないものとなろう。

別な対策法として，たとえばすべての旧字体を対応する新字体に置換して形態素解析を行い，解析後に元の漢字に置換し直す方法が考えられる。さらに，国語研の「通時コーパス」プロジェクトの進展により新規に開発された近代文語 UniDic のような時代別辞書の活用も解決法の候補となり [6]，これらの組み合わせも有力な解決法の候補となるはずである。

[3] https://unidic.ninjal.ac.jp/（2021 年 2 月 15 日確認）. 現代語書き言葉 UniDic については，伝ほか（2007）を参照のこと。
[4] 工藤（2018）によれば，本来「形態素解析」とは単語（もしくは形態素）への分割，品詞の推定，語形変化の処理の 3 つの処理を含むものであり，この立場に立てば本稿のように形態素への分割のみを取り扱う場合に「形態素解析」という名称を使用することはふさわしくない。ここでは便宜的に形態素解析という名称を用いるが，本稿ではあくまで当該漢字を含む単語 / 形態素の切り出しのみに注目しており，品詞推定や語形変化処理にはまったく触れることはないことに注意せられたい。
[5] https://unidic.ninjal.ac.jp/（2021 年 2 月 15 日確認）. 近代文語 UniDic については，小木曽ほか（2013）を参照のこと。
[6] 性質の異なる要素が混在するテキストの形態素解析について辞書の切り替えで対応するという点では，本稿のアプローチは『太陽コーパス』の形態素解析にあたり，文語文用辞書と口語文用辞書を切り替える手法を提案した間淵・小木曽（2015）のそれに近いものと言える。

本稿では，辞書登録という選択肢に一定の有効性を認めつつも，あえて辞書を変更することなく，形態素解析に当たっての辞書選択とテキスト中の文字の置換を柱とした対策の有効性を，実験的手法により追究する。その目的は，旧字体と新字体の混在したテキストからの形態素解析において，辞書選択と文字置換のみによって対応する場合の最適な組み合わせを同定し，またその組み合わせでどれほどまでに正確な形態素の切り出しが可能なのかを見極めることである。

3. データ

データは第 1 回国会（1947 年）から第 10 回国会（1951 年）までの 5 年間の国会会議録のうち，衆議院と参議院の会議それぞれ 4,092 件，5,301 件の会議の会議録を使用した。データは「国会会議録検索システム」[7]（https://kokkai.ndl.go.jp）で公開されているものについて，堤智昭氏作成のダウンローダを用いて 2018 年 7 月 5 日から 12 日にかけてダウンロードし，記号類と空白を削除するなどの前処理を施したものを使用した。表 1 に第 1 〜 5 回（当用漢字字体表告示前），第 6 〜 10 回（当用漢字字体表告示後）までの会議数と文字数を衆参別に集計したものを掲げる。

表 1　使用した国会会議録データの文字数と会議数

	参議院	衆議院	合計
第 1 〜 5 回			
文字数	136,326	187,349	323,675
会議数	2,095	2,909	5,004
第 6 〜 10 回			
文字数	128,335	152,737	281,072
会議数	1,997	2,392	4,389
合計			
文字数	264,661	340,086	604,747
会議数	4,092	5,301	9,393

4. 方法論

4.1 漢字対の選択

分析対象とした漢字は，「國 / 国」，「會 / 会」，「關 / 関」の 3 組の新旧漢字対とした。國 / 国の他にこれら 2 対を選んだ理由は，國 / 国のみに関わるような個別的要因をできるだけ取り除き，漢字選択の効果をできるだけ平均化することにある。國 / 国以外の漢字の選択基準は，その生起頻度数である。會 / 会は分析期間の国会会議録中でもっとも使用頻度の高い漢字対であり，國 / 国は 2 番目に，そして關 / 関は 3 番目に高頻度な漢字対であった。

4.2 漢字の抽出・形態素解析とその判定

第 3 節で挙げたデータの中から，当該の漢字を含む文を抽出し，そこからさらに各組 500 文を

[7] 国会会議録検索サイトについては松田（2008）を参照のこと。ただし，現在のインターフェースは松田（2008）の頃から大きく改善されており，API を用いたデータ入手も可能である（川瀬・清水 2015，岡田 2018）。

ランダムサンプリングにより抽出した。これらの文について，以下に記す 18 の条件に従って漢字置換を行った上で，辞書を選択して形態素解析を施した。形態素解析器は MeCab version 0.996，辞書は現代語書き言葉 UniDic（version 2.3.0），近代文語 UniDic（version 1603）を使用した。

形態素解析出力から，目視により当該漢字を含む漢語が正しく切り出されているのかを判定し，正しく切り出されたものの数を数えた。なお，ここではもっぱら該当漢字を含む漢語の切り出しにのみ着目しており，同一文中に含まれる他の形態素・単語の切り出し，さらに切り出された形態素の品詞情報などは一切考慮していない。一般に形態素解析の性能を評価するには F 値が用いられる（工藤 2018）が，今回はこうした事情から，正確に抽出された件数と総数に対するその割合のみを計算することにした。また，同一文中に当該漢字が複数回使用された場合，そのすべてを計数対象とした。よって 3 組いずれの場合にも分母が 500 にはならないことに注意されたい [8]。

4.3 要因

実験は漢字の置換方式と，辞書の選択の 2 要因の組み合わせでデザインした。漢字置換方式要因は，置換なし（つまり元テキストそのまま）と置換ありに大別される。置換ありは，1. すべてを新漢字か旧漢字に置換する方式，2. 1949 年の当用漢字字体表告示を境として [9]，それ以前のテキスト中の新漢字をすべて旧漢字に置換し，それ以後のテキスト中の旧漢字をすべて新漢字に置換する方式，3. 当該漢字に隣接する漢字の新 / 旧別により置換方向を変化させる方式，の大きく 3 種類の置換方式を採用した。3 の方式は，当該漢字の片側もしくは両側に隣接する漢字が新漢字 / 旧漢字であれば新漢字 / 旧漢字に置換し，どちらにも新漢字 / 旧漢字がなければ旧漢字 / 新漢字に置換する方式である（次頁図 1）。この最後の「どちらにも新漢字 / 旧漢字がなかった場合」に置換される種別を「デフォルト」と呼ぶことにする。また，この方式では周囲の漢字種別にあわせて置換されることから，以後この方式を「日和見置換」と称することにする。それぞれの方式の下部区分を含めると，漢字の置換方式では，全部で 6 つの区別を設けることになる。

辞書の選択では，1. すべてのデータについて近代文語 UniDic を使用，2. すべてのデータについて現代語書き言葉 UniDic を使用，3. 全データのうち 1949 年 4 月 28 日の当用漢字字体表告示以前は近代文語 UniDic を，告示以後は現代語書き言葉 UniDic を使用，の 3 区分を設けた。字体表告示前後で辞書を入れ替えるのは，それを境として旧字体と新字体の生起頻度が大きく変化することが予想されるからである。

漢字の置換方式と辞書の選択によって，各漢字対について 6 × 3 = 18 通りの形態素解析出力

[8] さらに國 / 国と會 / 会については，ランダムサンプリングされた文中のそれぞれの漢字の直前・直後の文字に OCR エラー（松田 2008: 19ff.）がそれぞれ 1 件ずつ発見され，これが形態素解析に影響を及ぼしたと考えられるため，これらを集計から除外した。

[9] 当用漢字字体表の内閣告示は 1949 年 4 月 28 日であった。これに対して第 5 回国会は同年 5 月 31 日に会期末を迎えており，第 6 回国会は同年 10 月 25 日に開会していることから，告示前はほぼ第 1 ～ 5 回国会に，告示後は 6 回以降に対応することになる。ここではデータ量をできるだけ等分にするために，告示後のデータを第 6 回から 10 回に求めることにした。

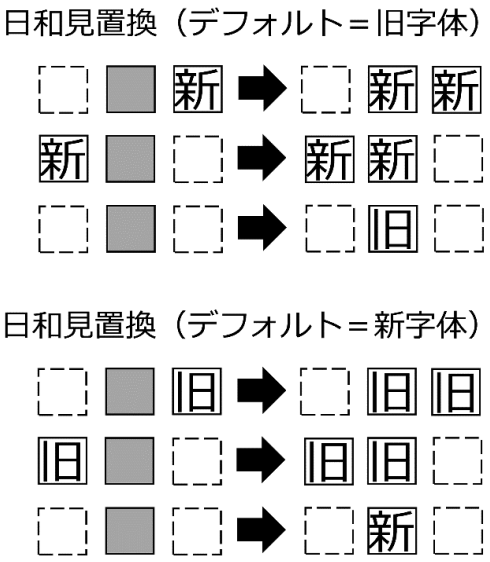


図 1　2 種類の日和見置換

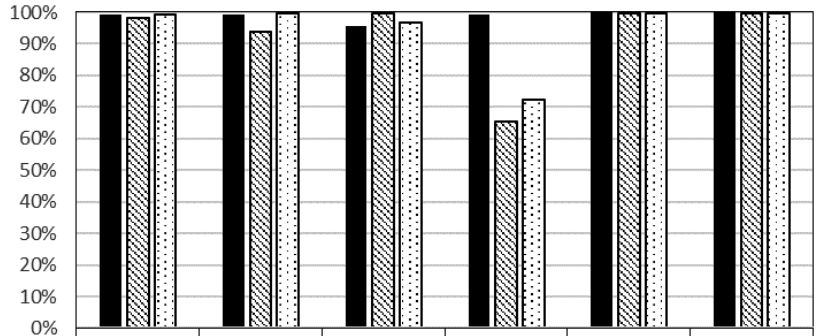
が得られることになる。これらを上記 §4.2 にあるように分析し，当該漢字を含む形態素の切り出しをもっとも安定して正確に行う組み合わせを求めた。

5. 結果と分析

図 2 〜 4 に実験の結果を示す。図中にある「置換なし」は，何らの文字置換もしなかったことを，「新旧」は当用漢字字体表告示前では当該漢字すべてを旧字体に，告示後では新字体に置換したことを，「新字体」は当該漢字すべてを新字体へ置換したことを，「旧字体」は当該漢字すべてを旧字体へ置換したことを，「日和見・旧」は旧字体をデフォルトとする日和見置換を，「日和見・新」は新字体をデフォルトとする日和見置換を施したことを示す。

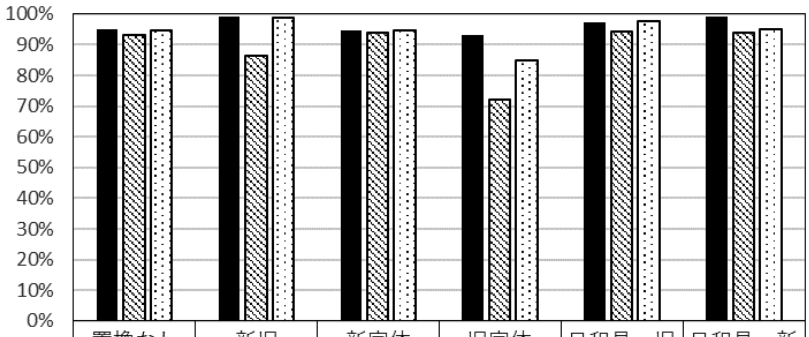
グラフでは全体的に國 / 国と關 / 関が同様な傾向を示しているのに対して，會 / 会がやや異なった分布傾向を呈している。3 対に共通するのは，一概に日和見･旧，日和見･新の成績が高いこと，とりわけ日和見新が一貫して好成績を見せている点である。國 / 国と關 / 関では，二つの日和見方式はほぼ完璧に近いパフォーマンスを見せている。會 / 会ではやや精度が落ちるものの，概して日和見方式の精度は高いと判断して良いであろう。

もう少し細かく見ていくと，3 対を通して一貫して高い正解率を出しているのは，新旧－近代文語 UniDic･辞書入替，日和見新－近代文語 UniDic･辞書入替の組み合わせであることがわかる。中でも日和見新・近代文語 / 辞書入替の 2 つが最高のパフォーマンスを見せている。新旧も辞書入替も当用漢字字体表交付を境界とした入れ替えという点で共通点を持ち，やはり告示以前のテキストと告示後のテキストでは文字種別の分布に大きな差があることが窺われる。日和見置換は，同一漢語内での新旧字体差を均す効果があり，結果的に告示前テキストでは旧字体のみの漢



	置換なし	新旧	新字体	旧字体	日和見・旧	日和見・新
■近代文語UniDic	99.41%	99.41%	95.54%	99.11%	99.85%	100.00%
▨現代語書き言葉UniDic	98.07%	93.91%	99.55%	65.23%	99.55%	99.55%
□辞書入替	99.41%	99.70%	96.88%	72.36%	99.70%	99.70%

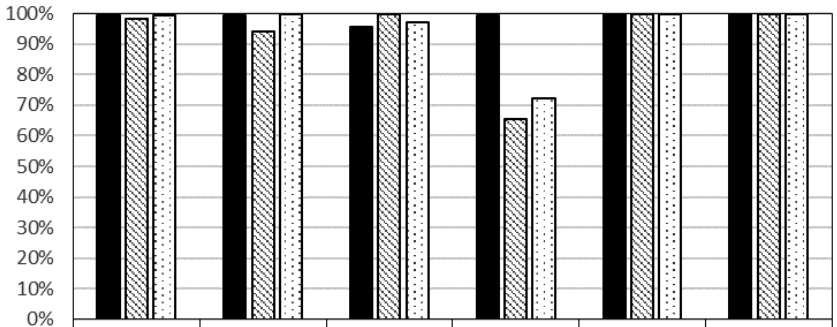
図 2　國 / 国の漢字置換・辞書選択の組み合わせによる分布



	置換なし	新旧	新字体	旧字体	日和見・旧	日和見・新
■近代文語UniDic	95.03%	99.35%	94.51%	93.07%	97.12%	99.22%
▨現代語書き言葉UniDic	93.07%	86.27%	93.73%	72.29%	94.12%	93.73%
□辞書入替	94.51%	98.69%	94.51%	84.97%	97.65%	95.03%

図 3　會 / 会の漢字置換・辞書選択の組み合わせによる分布

語を，告示後では新字体のみの漢語を増加させることにつながり，これが正しい形態素切り出しに貢献しているものと考えられる。



	置換なし	新旧	新字体	旧字体	日和見・旧	日和見・新
■近代文語UniDic	99.41%	99.41%	95.54%	99.11%	99.85%	100.00%
▨現代語書き言葉UniDic	98.07%	93.91%	99.55%	65.23%	99.55%	99.55%
□辞書入替	99.41%	99.70%	96.88%	72.36%	99.70%	99.70%

図 4　關 / 関の漢字置換・辞書選択の組み合わせによる分布

6. 議論

実験結果からは，文字置換と辞書選択を組み合わせることで，特定漢字を含む短単位形態素について，非常に正確な切り出しが可能であることが判明した。ただし，この方法の適用には，文字置換処理＋辞書切り替えを伴う形態素解析＋置換した文字の復元と，相応のコストが掛かることにも留意する必要がある。

これに対して辞書登録による対応は，テキストから辞書への異体字登録が必要な単語を洗い出し，一度登録してしまえばそれ以上のコストは掛からない。第 2 節でも述べたように，辞書への異体字の登録は，たしかに今回のような誤解析への解決法の有力候補である。

しかし，辞書登録による対処法にも弱点がある。それは，未登録単語への対応が困難な点である。国会会議録で使用される単語には当然その会議録というジャンルとともに，当時の時代背景も反映しているはずである。これは今回解析を行ったテキストとまったく異なる同時代のテキスト（たとえば新聞や小説）を分析した場合に，明らかになることであろうと思われる。テキストが変わる毎に新たな異体字登録が必要になるとすると，時代の変動と共に移り変わる漢語に柔軟に対応するためには，異体字の辞書登録よりは，高い精度で切り分けが可能であるのであれば，今回検討したような漢字置換と辞書入替による対応の方が望ましいという議論も成り立つ。漢字置換と辞書入替の組み合わせであれば，新旧漢字の混在した，別ジャンルのまったく性格の異なるテキストであっても対応可能である。特定コーパスを離れて広く柔軟に同様なテキストに対応が可能であるという点で，漢字置換・辞書入替方式には，辞書登録方式に比べて一日の長があるように思われる。今後さらに整備が進むことが予想される戦後期をカバーするコーパスの構築を考慮した場合，こうした視点はとりわけ重要なものとなろう。

7. 結論・今後の課題

3 つの漢字ペアに対する文字置換と辞書選択を組み合わせた実験の結果から，本稿では新字体をデフォルトとする日和見置換法と，近代文語ないし辞書入替の組み合わせによる方法が，もっとも正確に当該漢字を含む短単位形態素を切り出せるということを示した。この方法は，文字置換と辞書切り替えを組み合わせるという点で新しい手法であるが，辞書登録による方法に比べて，新たなテキストへの対応も柔軟に対応できるという長点があることも主張した。

実験結果については，統計解析を含めた，さらに詳細な分析が可能である。分析という点では，特定漢語の誤解析を検討するような，エラー分析も行う必要がある。また，今回の実験では辞書の切り替えはすべて当用漢字告示を境として行うこととしたが，これをもっと細かな単位，たとえば新旧字体の現れ方により文単位で解析用辞書を切り替えるようにすれば，さらなる解析精度の向上が図れる可能性がある。これらすべて今後の課題としたい。

参照文献

岡田祥平（2018）「日本語コーパスとしての『国会会議録検索システム検索用 API』―計量的研究の精緻化・深化の可能性―」『新潟大学教育学部研究紀要 人文・社会科学編』11(1): 31–51.

小木曽智信・小町守・松本裕治（2013）「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』20(5): 727–748.

小椋秀樹（2014）「形態論情報」山崎誠（編）『書き言葉コーパス―設計と構築―』（講座日本語コーパス第 2 巻）68–88. 東京：朝倉書店.

小椋秀樹・小木曽智信・小磯花絵・冨士池優美・相馬さつき（2007）「「現代日本語書き言葉均衡コーパス」の短単位解析について」『言語処理学会第 13 回年次大会発表論文集』，720–723.

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）』（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（JC-D-10-05-01））https://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-01.pdf（2021 年 2 月 15 日確認）

川瀬直人・清水茉有子（2015）「国会会議録フルテキスト・データベース Web API 開発の背景とその利用状況分析」『情報の科学と技術』65(12): 531–536.

工藤拓（2018）『形態素解析の理論と実装』東京：近代科学社.

高田智和（2002）「電子化辞書とねじれの漢字」『計量国語学』23(5): 241–254.

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（2007）「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22: 101–123.

松田謙次郎（2008）「国会会議録検索システム概論」松田謙次郎（編）『国会会議録を使った日本語研究』1–32. 東京：ひつじ書房.

松田謙次郎（2019）「国会会議録における当用漢字の拡散過程」「昭和・平成書き言葉コーパスによる近現代日本語の実証的研究」研究発表会（於国立国語研究所，2019 年 10 月 19 日）.

間淵洋子・小木曽智信（2015）「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」『じんもんこん 2015 論文集』，125–130.

Kudo, Taku, Kaoru Yamamoto, Yuji Matsumoto (2004) Applying Conditional Random Fields to Japanese Morphological Analysis. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. 230–237. https://www.aclweb.org/anthology/W04-3230（2021 年 2 月 15 日確認）

How to Correctly Morphologically Analyze Text Containing a Mixture of Old- and New-Style Kanji Scripts

MATSUDA Kenjiro

Kobe Shoin Women's University / Project Collaborator, NINJAL

Abstract

Japanese texts containing a mixture of old- (*kyūjitai*) and new- (*shinjitai*) style kanji scripts pose a serious problem for an automatic morphological analyzer. However, recent developments in various dictionaries by era, undertaken by the corpora project at NINJAL, have brought about a new opportunity to solve this problem. Another promising solution is to replace the script in the text in some way, so that the analyzer can correctly identify the characters/morphemes. We designed an experiment with three dictionary selection methods and six replacement methods using three pairs of old/new kanji scripts (國/国, 會/会 and 關/関) to determine which combination would result in the most precise analysis. An analysis of the text data from the Minutes of the National Diet between 1947 and 1951 demonstrated that, of the 18 combinations, two dictionaries gave the best results. These were, the Contemporary Written Japanese UniDic dictionary up to the public notification of the Table of Script Styles of Jōyō Kanji on April 29, 1949, and The Modern Literary UniDic. With these, we coupled a replacement of a kanji script with an old counterpart when its immediate neighbor was also an old one, and with a new one when it was not. Although the addition of the different scripts to the dictionary entries would be another viable solution, our method is more desirable in that it is applicable to a wider range of texts without dictionary entry modifications.

Keywords: morphological analysis, Minutes of the National Diet, kyūjitai, shinjitai, the Table of Script Styles of Tōyō Kanji