近代語コーパスのための形態論情報付与規程の整備

著者	須永 哲矢, 近藤 明日子
雑誌名	近代語コーパス設計のための文献言語研究 成果報告書
ページ	93-117
発行年	2012-10-31
シリーズ	国立国語研究所共同研究報告 ; 12-03
URL	http://doi.org/10.15084/00002767

近代語コーパスのための形態論情報付与規程の整備

須永　哲矢　（国立国語研究所コーパス開発センター）[1]
近藤　明日子　（国立国語研究所コーパス開発センター）[2]

１．近代語コーパスでの言語単位

１．１ 近代語コーパスでの言語単位

近代語コーパスでの言語単位には、「短単位」という言語単位を採用した。「短単位」は『現代日本語書き言葉均衡コーパス』でも採用されている言語単位で、詳細な規程のもと、揺れの少ない単語認定を実現している。実際の単位認定規程は、各ケースごとに細かく定義されているが、ここではその概要として、単位認定の大原則を紹介しておく。現代語についての規程の詳細は、小椋・小磯ほか(2011)を参照してほしい。以下、小椋・小磯ほか(2011)を『規程集』と呼ぶ。近代語については、コーパスの構築に向けて現在規程を検討中であるが、本稿では、規程を整備していくにあたっての問題点を整理する。なお、本稿で述べる問題点は網羅的なものではなく、コーパス作成準備作業において気付いた範囲でまとめたものにとどまることを、はじめにお断りしておく。

１．２「短単位」の概要

《和語》

単純語２語の結合まで、又は単純語１語と接辞１語の結合までを１短単位とする。
（“／”は短単位の切れ目を示す。以下も同様）

【例】／母／　　／母親／　　／母親／代わり／　　／真っ白／

《漢語》

2 字漢語までを１短単位とする。

【例】／大臣／　　／財務／大臣／　　／大臣／級／会合

《外来語》

原語で１語となるものを１短単位とする。

【例】　／オレンジ／　　／カラー／コピー／　　／ビタミン／剤／

[1] tsunaga@ninjal.ac.jp
[2] kondo@ninjal.ac.jp

《付属語》
付属語１語を１短単位とする。

【例】　／が／　　／だ／　　／の／で／

ここで紹介した短単位認定規程は、あくまで大原則であり、実際には、例外を含めて極めて詳細な規程が設けられている。

例えば、連体詞とされる「この」は、もとをたどれば代名詞「こ」と格助詞「の」に分割でき、そう見るならば「付属語１語を１短単位とする」という原則から、付属語である「の」を切り離し、「こ／の」と分割されることになる。しかし現代においては「こ」と「の」に分かれるという意識はもはや薄れており、代名詞「こ」が「の」以外と結合する用法も存在しない。このような事情から、現代語においては「この」に関しては「代名詞＋助詞」のように分割せず、「連体詞」と認定し１短単位とする、と規程として定めてある。
このように、単位認定の仕方が複数想定できてしまう場合や、原則通りに処理しない方が現代語の実態にとっては有用であると判断される場合などに関しても、個々に処理方針を詳細に設定したものが『規程集』となっている。

１．３　時代に合わせた規程整備の必要性

上述のとおり、言語実態に即した詳細な規程といっても、それはあくまで現代語を対象として設定されたものであり、時代が異なれば当時の言語実態と齟齬をきたす場合も少なくない。

例えば、上で例とした「この」に関しては、現代語を扱う限りには連体詞として１短単位として認定したが、平安時代までさかのぼってしまえば、当時においては、「の」以外の助詞も自由に結合でき（「こを」「こは」など）、実例上からも「こ」を単独で代名詞と認めた方が当時の実態に適合すること、などから「こ／の」と２単位に分割する方がふさわしい。

そのため、歴史的言語資料に対し「短単位」を引き続き採用する際には、その時代に合わせて規程にも拡張や変更を加えるなど、整備が必要となる。

歴史的資料を対象とした規程として整備されているのは、平安期の和文資料を対象とした規程（小椋・須永 2012）のみであり、近代語用の規程はまだ設定されていない。そこで、近代語コーパス構築のためには、近代語の言語実態に即した短単位規程を設定しなければならない。そのために、まずは『明六雑誌コーパス』の試作を通じ、作業中に生じた処理上の問題点（従来の規程どおりの処理では近代語の言語実態と齟齬をきたす場合／品詞認定等、処理の仕方が複数想定される事例に関して、従来の規程の判別基準では処理しきれない場合／処理方針を新規に定めなければならない近代語特有の問題、など）を収集し、近代語に合わせた処理方針を検討した。将来的には近代語用の詳細な規程集の作成を最終

目標としているが、本報告書での第一目標は、近代語を処理するにあたっての問題点の整理と、暫定的な処理の方向性を定めることである。

なお、一口に「近代語」と言っても、口語文と文語文で大きく様相が異なる場合があり、近代語用の規程整備という作業においても、実際には文語の場合、口語の場合と分けて設定しなければならない事例もある。そこで将来的には近代語共通の規程のほか、文語のみ、あるいは口語のみに適用される規程を細分していく必要がある。ただ、本稿の範囲内では、試作コーパスとなった『明六雑誌』の文体の多くが文語体であったこともあり、まずは近代文語を処理することを想定した際の、既存の規程の見直しから出発することとした。本稿での「近代語」は、とくに断りがない限りは近代文語を想定している。

２．近代語での単位認定の問題点と、その処理方針

２．１　辞書登録に関わる事項

近代語コーパスの形態論情報の付与にあたっては、形態素解析辞書 UniDic を用いている。近代語コーパス構築作業では、UniDic にそれまで登録されていない、近代語特有の語彙が多数出現する。そこでそれらの語彙を処理するためには、まずは辞書側にその語を登録しなければならないが、現代語から出発した UniDic に近代語特有の語を追加登録しようとすると、問題が生じる場合がある。

２．１．１　表記上、短単位分割が不可能な場合

UniDic に登録する辞書情報も、短単位ごとに登録していく、ということになるが、近代語のテキストを扱っていると、表記の都合上、短単位に分割できないという場合が多数生じる。

【例】

「非る」（あらざる）　　所謂政府なる者亦人に非るなし
「加之」（しかのみならず）　加之官職の設概するに古に簡にして後世に繁く
「不然」（しからざれば）　　不然則又豪奪の賊なり

これらはそれぞれ短単位分割の原則としては「あら／ざる」「しか／のみ／なら／ず」「しから／ざれ／ば」と分割されるべきだが、表記上、そのような分割ができないため、このような出現形に対しての処理方針を定めねばならない。
現時点では大筋において以下のような方針としている。

（１）使用頻度が高く、ある程度慣習的に認められていると見られる出現形に関しては、例外として全体を１短単位として辞書登録する。「非る」「加之」などがこれにあたる。近代語特有の表記に対し、辞書上１短単位と認めたものには、他に以下のようなものがある。

於是　ここにおいて何者　なんとなれば
而　　しこうして
遮莫　さもあらばあれ
乍併　しかしながら

（2）それほど使用頻度が高くないものに関しては、短単位に分割できるよう、テキスト側を書き換える

【例】
「不然」（しからざれば）　原文：不然則又豪奪の賊なり→コーパス：然らざれば・・・

２．１．２　語自体は既登録だが、時代によって品詞認定が異なる場合

時代とともに品詞認定が変わる語もあるので、それらについても処理方針を定めておかねばならない。中古和文では以下のように処理することとしている。

（中古和文）
例えば、現代語として副詞「夜な夜な」が UniDic には登録されているが、中古までさかのぼると、「夜な夜な」に名詞としか認められない用法が出現する。

【例】おはしましし夜な夜なのありさま

このような場合、「夜な夜な」の品詞情報を「副詞」から「名詞－副詞可能」に書き換えるという方法も考えられるが、原則として、既登録の語に対して、別の時代への対応へのために既登録語の品詞情報を書き換えるということは行わない。よって、残る方法としては（1）品詞認定の実際上のずれは多少無視して、既登録の語をそのまま使う、（2）中古和文特有の用法を、別品詞として新規登録する、という 2 つとなる。どちらの方法を採るかの目安は、以下に示すとおりである。

（1）既登録語の品詞が名詞であり、中古和文において副詞用法や形状詞用法が出現した場合→既登録の名詞をそのまま使用

（2）既登録の品詞が名詞以外であり、中古和文において名詞用法が出現した場合→別語として名詞を新規登録

※上記（1）、品詞認定の実際上のずれは多少無視して、既登録の語をそそのまま使うという処理を行うのは、あくまで意味の面でも共通性が見られる場合に限る。意味の面で大

きな違いが認められる場合は、別語として新規登録する。

【例】いかさま
現代語：名詞
中古和文：形状詞を新規登録

既登録の語が名詞であり、時代をさかのぼった結果、形状詞用法も認められた場合、基本方針としては（１）が適用され、既登録の名詞として処理される。しかし、現代語で名詞として登録された「いかさま」は「ごまかし」「いんちき」の意であるのに対し、かつての形状詞用法「いかさま」は疑問、「どのよう」「どんなふう」の意である。このように意味が異なるものに関しては、別語として形状詞「いかさま」を新規登録した。

（近代語）
近代語においても、現代語とは品詞認定が異なる場合の処理は、以下のとおりとする。

原則：既登録の語の品詞情報の書き換えは行わない。

品詞認定が異なる語の処理方針：

（１）既登録語の品詞が名詞であり、近代語において副詞用法や形状詞用法が出現した場合→新語彙素を登録することなく、既登録の名詞を使用する。（中古和文での方針と同様）

（２）既登録の品詞が名詞以外で、近代語ににおいて名詞用法が出現した場合
（２‐A）既登録語が副詞または形状詞で、近代では名詞としての用法もあるが、副詞・形状詞のとしての意味もそのまま有するものは新語彙素を登録することなく、既登録の副詞または形状詞を使用する。

【例】現 UniDic「副詞」―明六「名詞」
習ふに漸次を以てし行ふに歳月を以てし
故に我國方今の景況に於て國人の智識一層を進めば輸出入の差從て一層を加ふべし
縦令兵力今より數層を加ふも野を轉じて文と爲さず
倍氏の理學の大凡後篇に見ゆ

【例】現 UniDic「形状詞」―明六「名詞」
嘗て國家の禍を推すに歸する所は官吏人民の姑息にあり
植民地愈々廣大に至れり

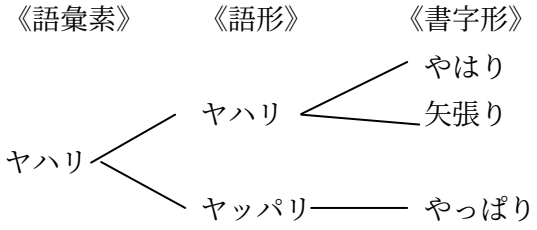
倫理の當然に従て必ず是等の惡風俗を禁止するの憲法を設定する㇀固より緊要の㇀と云ふ可きなり
凡そ簡易明白を歓び
普通の中に特別を附せざる㇀を得ざる者あり

（２‐B）上記（２‐A）にあたらないものは、別語彙素として名詞を新規登録する。

２．１．３　外来語の同語異語判別

形態素解析辞書 UniDic の見出し語は、以下に示すように階層化された形で格納されている（小木曽・中村 2011 参照）。

【例】



新たに辞書情報を登録する際、「語彙素」のレベルから登録すべきか、「語形」レベルで登録すべきか迷う場合がある

【例】

歩く（アリク）・・・・語彙素「歩く（アルク）」とは別の「語彙素」か。
語彙素「歩く（アルク）」の一「語形」か。

このような場合の判断の揺れを抑えるために、『規程集』ではどのようなものを一つの語彙素、または語形にまとめ、どのようなものは別にするかに関しても、「同語異語判別規程」を設けている。ただし、『規程集』での同語異語判別規程はあくまで現代語を対象としたものであり、現代語では想定されていなかった差異の扱いに関しては、新規に方針を定めねばならない。近代語コーパス作成に当たっては、特に外来語において、想定されていなかった同語異語判別の問題が生じる。

【例】

レホルメルス　（reformers）

上例「レホルメルス」は reformer（今日的には「リフォーマー」が一般的なカタカナ表記か）複数形 reformers を読んだものと思われるが、語彙素「リフォーマー」の語形に「レ

ホルメルス」を登録すべきか、語彙素から「レホルメルス」を立てるべきかは、既存の規程からは判断できない。

また、出現形「レホルメルス」に関しては、注釈書その他から原語が「reformers」であることがたまたま推定できたが、そもそも、原語が何なのかわからない外来語も多数出現する。その際、出現した外来語に対し逐一原語を推定し、その上で同語異語判別をしていくというのは､近代語コーパスの試作段階においては煩雑なだけであり､非効率である。そこで、近代語コーパスの試作段階においては、外来語の同語異語判別は以下の通りに行うこととした。

外来語の同語異語判別：

『規程集』所収の規程を適用できる範囲内のものは、その規程に従って判別する。
既存の規程では判断できない語に関しては、原則として出現形をそのまま語彙素として登録する。

補則・外来語の語彙素・語形の表記に関して：

『規程集』では、外来語の語彙素・語形に用いる仮名・符号の種類が定められており、その範囲は近代語での語彙素・語形表記でも遵守するものとする。そのため、出現形の仮名表記が『規程集』に定められた使用可能範囲から外れている場合には、使用可能範囲内の表記に改める。出現形の仮名表記を『規程集』での使用可能範囲内の仮名表記に改める基準に関しては、本稿末尾の資料１を参照のこと。

２．２　単位境界認定に関わる事項

時代が変われば言語の実態も変わる。そのため、「この」の扱いを例に挙げたように、単位境界の認定が、現代語と中古和文で異なる場合がある。これから構築していく近代語の規程は、時代の上では、既存の２つの規程（現代・中古和文）の間に位置することになる。そこで、現代語と中古語とで明らかに扱いが異なるものに関しては、近代語ではどちらの時代と同じ扱いにするか、あるいはどちらとも異なる独自の扱いをするか、を定めていかねばならない。

[１]連体詞

「この」「その」などの「連体詞」の扱いは、以下の通り、現代語と中古和文で単位の切り方が異なる。

現代語：／この／　　／その／　（連体詞）
中古和文：こ／の　　そ／の　（代名詞＋格助詞）
※中古では「連体詞」という品詞を原則として認めない。

近代語では、いわゆる連体詞の扱いに関しては、中古和文に準ずるものとする。
つまり、品詞としての「連体詞」は認めず、現代語での連体詞に当たるものは「代名詞＋格助詞」に分割する。

近代語：こ／の　　そ／の

[2]「異なる」

現代語：／異なる／（動詞）
中古和文：異／なる（名詞＋断定の助動詞「なり」）

近代語では、中古和文側の処理に合わせ、名詞＋助動詞とする。

近代語：異／なる

[3]「－の～」「－つ～」「－が～」
（現代語）
短単位認定規程においては、助詞「の」「つ」「が」は、それだけで１短単位として分割するのが原則だが、現代語においては「の」「つ」「が」を含んだ全体を１短単位と認定した方が適当であるとの判断から、例外的に分割しないものがある。

【例】
「－の～」：　／日の丸／　　／床の間／　　／竹の子／
「－が～」：　／天が下／　　／雁が音／　　／剣が峰／
「－つ～」：　／国つ神／

これらについては規程集内に資料「要注意語」を設け、一覧が整備されている（小椋・小磯ほか 2011 参照）。

※「－の～」で１短単位とするものを選定するにあたっては、以下の事項をおおよその目安とする。

①「の」が読み添えとなっているもの

【例】斎宮（いつきのみや）　　対屋（たいのや）

②「の」の直前の要素が被複形のもの

【例】木の葉　　目の当たり

③「－の～」全体の品詞が名詞以外となるもの

【例】　案の定　　気の毒

④「－の～」が動植物名等を表すもの

【例】　卯の花　　竹の子　　泥の木

※「－つ～」に関しては、現代語で格助詞「つ」はもはや使用されてないため、原則として全ての「－つ～」に関して「つ」を格助詞として分割することはせず、全体を 1 短単位とする。

（中古和文）

中古和文では、当時の実態を踏まえて、「－の～」「－つ～」「－が～」全体を 1 短単位と認定する範囲を限定する。

【例】

現代語：　　／身の程／　　／天が下／

中古和文：　身／の／程　　天／が／下

中古和文において 1 短単位とするものについても、中古和文規程集に一覧が収録されている（小椋・須永 2012 参照）。

※中古和文において「－の～」「－つ～」「－が～」で 1 短単位とするものを選定するにあたっての目安は、現代語の「－の～」に関する目安①～④のうち、①③④を、中古和文の「－の～」「－つ～」「－が～」に適用する。

（近代語）

以上のように、「－の～」「－つ～」「－が～」に関しては、全体で 1 短単位とする範囲が、現代語と中古和文で異なる。近代語では、当面は現代語で定めた範囲に従って 1 短単位と認定する。

２．３　コーパスへの形態論情報付与に関わる事項

実際にコーパス上に形態論情報を付与していく際には、品詞認定等の面で判断に迷う場面がさまざまに生じる。そのような場合のためにも、『規程集』では品詞の判別基準なども設定されている。品詞等の判別に関しても、現代語および中古和文では想定されていなかった近代語特有の問題があるため、それらに対応するために近代語用の判別基準を設定しておく必要がある。以下にはその概略を示す。

以下に示すものはあくまで概略であり、実際の形態論情報付与作業においては、近代語特有の問題が個別的に生ずる場合があり、それらに対しても個別対応の詳細な規程を設けていかなければならない。しかし、現時点で扱った資料は『明六雑誌』のみであり、ここでの個別事例を一般化して規程とするにはまだ資料不足の段階である。今後の作業も含め、実例と、実際の処理方法を蓄積し、なるべく一般化した形でそのような問題に対応できる規程を細部にわたって整備していくことが、将来的な『近代語コーパス用規程集』での課題となる。

２．３．１　活用形・活用型認定

[１]活用型が上二段か四段か判じ難い場合

【例】　恨む　忍ぶ

活用型が上二段か四段か判じ難い場合があるが、その場合、『明六雑誌』の範囲内では、全体的な傾向から見て上二段としておく（近代語全体に関しては未定）。

[２]活用型が文語なのか口語なのか判じ難い場合

出現活用形が文語なのか口語なのか区別がつかない場合は文語としての処理を優先する。

【例】　斬棄てて
（文語下二段／口語下一段　　→　　文語下二段を優先）

活用形上、口語としてしか認められないもののみを口語として処理する。

【例】斬棄てる
（口語下一段）

[３]仮名遣いと活用の行

出現形の仮名遣いからすると活用の行が変わってしまうものは、別語形とする。

【例】

据へて・・・規範的な古典では「据ゑ」。

「文語下二段－ワ行」として既登録。　「文語下二段－ハ行」を新規登録

２．３．２　品詞認定

形態論情報の付与に当たって、最も判断の揺れが生じやすいのは品詞認定である。同一の出現形に対し、選択肢として辞書上に複数の品詞が用意されている場合、そのうちのどれを使うべきかを詳細に規定しておかないと、コーパスの品詞情報が揺れてしまう。例えば疑問文末の「や」に対しては、「係助詞」「終助詞」「副助詞」という 3 種類の処理がありうる。このように、揺れてしまいかねない個所を洗い出し、処理方針を確定させていく、という作業が、最終的な規程集の作成においては必要になってくる。ここでは品詞認定に悩むという事例自体の紹介として、代表的な数例を紹介するにとどめる。近代語用の、個別の判別基準に関しては次節でいくつか詳細に取り上げる。

[１]疑問・反語の助詞「や」「か」

疑問・反語の「や」「か」に関しては係助詞とする。（副助詞、終助詞は使わない）

[2]「間投助詞」と認定したい助詞の扱い

「間投助詞」と認定したい助詞に関しては、「終助詞」として処理する。（UniDic の助詞の分類には「間投助詞」が存在しないため）

「古池や蛙飛び込む水の音」の「や」のような、いわゆる切れ字についても終助詞とする。

[3]出現形「に」の判別基準

出現形「に」に関しては、助詞と認定すべきか、断定の助動詞（「だ」「なり」）の連用形と認定すべきか、判断が揺れやすい。そのため、現代語の『規程集』でも判別基準を設けたが、対象とする時代が異なると、判別基準も修正が必要になる。

現代語での判別基準の概略は、以下のとおりである。

現代語での出現形「に」の判別基準：

①形状詞　　→助動詞「だ」連用形（中古では「なり」）
②文語形容詞連体形（なきにしもあらず）→助動詞「なり」連用形
③他　　→格助詞

しかし時代をさかのぼり、中古和文を対象とする際には、この基準をそのまま適用することはできなくなる。例えば中古和文においては、上記①②以外にも助動詞「なり」の可

能性も生じるため、そのまま適用はできない。そこで中古和文では中古和文用に「に」の判別基準を修正したが、近代語ではまた別個に、現代語とも中古語とも異なる判別基準の設定が必要となったため、近代語用の「に」の判別基準を設定しなおした。近代語での「に」の判別基準は本稿末尾の資料２参照のこと。

２．３．３　読み

コーパスに形態論情報を付与していく際、漢字などに対しても読みを与えていかねばならないが、現実的には読みが不明な場合や、ひとつに確定できない場合も多い。そこで、読みを与える際に迷いそうな場合、どのように処理するかに関しても、方針を定めておかねばならない。

慣例として最も一般的、常識的な読みを採っておく、ということを原則とするが、「一般的、常識的な読み」が確定しない場合は以下の目安に従って読みを与える。

原則１　音読み・特に漢音を優先する。

【例】
一人　イチニン＞ヒトリ
給水場　キュウスイジョウ＞キュウスイバ
重複　チョウフク＞ジュウフク

原則２　音便形であるかわからないものは、元の形を優先する。

【例】
撃て　ウチテ＞ウッテ

原則３　読み添えの「の」を補って読むのは地名、姓のみとする。

【例】
藤原道長　フジワラノ　ミチナガ
中関白　　ナカ＿カンパク（「ナカノカンパク」とは読まない）
後出師表　ゴスイシ＿ヒョウ（「ゴスイシノヒョウ」とは読まない）

原則４　原文にルビがあっても、その読みに従わない場合がある。

[１]漢語に対し、外来語のルビがついている場合

（A）通常の漢語として認められるものは、原文ルビを無視して漢語として読む。

【例】「玩弄品」　原文ルビ：トイス
→コーパス上の読み：ガンロウヒン
「造鉄術」　原文ルビ：アーツオブメイキング
→コーパス上の読み：ゾウテツジュツ

（B）漢語としての用法が見当たらないものは、原文ルビに従い、原文ルビの指す外来語の書字形とする。

【例】「聖礫」　原文ルビ：クルス
「聖礫」という漢語用例見当たらず
→コーパス上の読み：クルス（語彙素「クルス」の書字形）

[2]原文ルビの読みが、熟字訓として定着しているとは言い難いものに関しては、原文ルビを無視して通常の漢語としての読みを与える。

【例】「痴漢」　原文ルビ：バカモノ
→コーパス上の読み：チカン
「吾人」　原文ルビ：ワレラ
→コーパス上の読み：ゴジン

３．今後の課題

以上、『明六雑誌コーパス』での形態論情報付与を事例として、近代語コーパスのための規程整備の必要性を確認し、現時点での近代語用規程の主要なものを概観した。規程整備にあたっては、単位境界認定から品詞認定、表記、読みの問題等、多岐にわたる事項の処理方針を詳細に設定していかねばならない。現代語を対象とした『規程集』は本稿で紹介した言語単位、「短単位」に関する部分のみで２００ページほど、これを前提とした中古和文用の追加規程集（須永・小椋 2012）で１００ページほどの分量となっており、本稿で紹介した近代語用規程の分量と比較しても明らかな通り、本稿は将来作成すべき近代語用規程集の中心原則、しかもその一部という位置づけになる。今後『明六雑誌』以外の近代語資料も見渡す過程で、今回設定した中心原則に詳細な補則を設定したり、修正を施したり、さらには原則から新設するなどして、より詳細な近代語用規程集を作成するのがこれからの課題となる。本稿で紹介した規程はおおよその原則案という位置づけであることは先に述べたが、今後この原則案を基に、近代語用規程として完成させた形のサンプルとして、細部まで整備した規程を本稿末尾に参考資料として掲載する。作業上特に問題になりやすい表記の問題、品詞認定の問題からそれぞれ「仮名表記される外来語の語形の定め方」「出

現形「に」の判別基準」を「資料１」「資料２」として示す。本稿で紹介した、他の規程ひとつひとつに関しても参考資料に掲げるような形で詳細に設定していくというのがこれからの作業となる。

『明六雑誌コーパス』構築という今回の作業に限定しても、現実の言語事実は多種多様であり、原則としての規程を設定してなお、様々な面で個別に判断を求められる場面が多くみられた。そのような「揺れ」になりやすい箇所にも対応し、統一的な処理をするためにこそ、規程は必要なものであり、今後さらなる事例の蓄積を通し、一般化した形で個々の問題に対応できる規程を細部にわたって整備していく予定である。

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』形態論情報規程集　第 4 版（上）（下）（国立研究所内部報告書 LR-CCG-10-05-01,02）』（国立国語研究所）

小椋秀樹、須永哲矢(2012)『中古和文 UniDic 短単位規程集』（『2009-2011 年度科学研究費補助金　基礎研究 (C)「和文系資料を対象とした形態素解析辞書の開発」成果報告書 2（課題番号 21520492））

URL

UniDic	http://download.unidic.org
中古和文 UniDic	http://www2.ninjal.ac.jp/lrc/index.php?UniDic
近代文語 UniDic	http://www2.ninjal.ac.jp/lrc/index.php?UniDic

資料１

仮名表記される外来語の語形の定め方

小椋・小磯ほか（2011）では、現代語においてカタカナ表記される外来語の出現形から語形を定める規程を定める（pp.85-87）。おおよそそれは、「外来語の表記」（平成３年内閣告示）の仮名・符号の表（表1）に基づき、外来語の語形の表記に用いることができる仮名・符号の範囲（表１中の本表・付表A）を定め、その範囲を超える出現形について、範囲内に収める方法を記したものとなっている。

表 1

本表

ア　イ　ウ　エ　オ
カ　キ　ク　ケ　コ
サ　シ　ス　セ　ソ
タ　チ　ツ　テ　ト
ナ　ニ　ヌ　ネ　ノ
ハ　ヒ　フ　ヘ　ホ
マ　ミ　ム　メ　モ
ヤ　　ユ　　ヨ
ラ　リ　ル　レ　ロ
ワ
ガ　ギ　グ　ゲ　ゴ
ザ　ジ　ズ　ゼ　ゾ
ダ　　　デ　ド
バ　ビ　ブ　ベ　ボ
パ　ピ　プ　ペ　ポ
キャ　　キュ　　キョ
シャ　　シュ　　ショ
チャ　　チュ　　チョ
ニャ　　ニュ　　ニョ
ヒャ　　ヒュ　　ヒョ
ミャ　　ミュ　　ミョ
リャ　　リュ　　リョ
ギャ　　ギュ　　ギョ
ジャ　　ジュ　　ジョ
ビャ　　ビュ　　ビョ
ピャ　　ピュ　　ピョ

ン（撥音）
ッ（促音）
ー（長音符号）

シェ　　チェ　　ツァ　ツィ　ツェ　ツォ　　ティ　トゥ　　ファ　フィ　フェ　フォ　　ジェ　　ディ　ドゥ　デュ　フュ

付表A

ウィ　ウェ　ウォ　　ツィ

付表B

イェ　　クァ　クィ　クェ　クォ　　グァ　　ヴァ　ヴィ　ヴ　ヴェ　ヴォ　　デュ　　ヴュ

近代語の短単位においても、外来語の語形の表記に用いることができる仮名・符号は現代語の規程に定める範囲に倣うこととする。ただし、近代語では出現形で用いられる仮名・符号の種類が現代語より多様であるといった、現代語にはない実態があるため、現代語の規程をそのまま近代語に当てはめることはできない。そこで、近代語用に出現形から語形を定める規程を改めてここにまとめるものである。

なお、以下の記述の中でいう「辞書」とは、『大辞林』第 2 版と『日本国語大辞典』第 2 版を指す。両辞書の記述が異なる場合、原則として『日本国語大辞典』第 2 版に従う。

（1） 長音について

ア・イ・ウ・エ・オ列の仮名の後にそれぞれ「ア・イ・ウ・エ・オ」を用いた出現形で、「ア・イ・ウ・エ・オ」が長音を表すと考えられるものについては、原則として長音符号「ー」を用いた形を語形とする。

【例】アパアトメント→アパートメント、アンコオル→アンコール

ただし、辞書の見出し（空見出しを除く。以下同）で「ア・イ・ウ・エ・オ」が用いられている場合は、「ア・イ・ウ・エ・オ」を語形とする。

ア・イ・ウ・エ・オ列の仮名の後にそれぞれ小書き「ァ・ィ・ゥ・ェ・ォ」を用いた出現形は、長音符号「ー」を用いた形を語形とする。

【例】オスカァ→オスカー、ロセツチィ→ロセッチー

イ・エ・オ列の仮名の後にそれぞれ「ヰ・ヱ・ヲ」を用いた出現形で、その「ヰ・ヱ・ヲ」が長音を表すと考えられるものについては、長音符号「ー」を用いた形を語形とする。

【例】ルビヰ→ルビー、ペヱトル→ペートル、レボヲリユーシヨン→レボーリューション

上記の規程を適用した結果、語形で長音符号が複数連続する場合は、一つに統合する。

【例】ウンゾォール（→ウンゾーール）→ウンゾール

（2） 本表の仮名「ツァ」「ツェ」「デュ」「フュ」を用いた出現形について

出現形と同じ仮名を用いた形を語形とする。

（3） 本表・付表 A に見られない仮名・符号について

仮名・符号ごとに以下のように語形を定める。

○「ヂ」

書字形「ヂヱ」は「ジェ」を語形とする。

【例】ブールヂヱー→ブールジェー

拗音「ジョ」の代わりに用いられていると見なせる書字形「ヂヲ」は、「ジョ」を語形とする。

【例】レリヂヲス→レリジョス

上記以外の書字形「ヂ」は「ジ」を語形とする。

【例】ラヂオ→ラジオ

○「ヅ」

書字形「ヅュ」は「ジュ」を語形とする。

【例】（未確認）

上記以外の書字形「ヅ」は「ズ」を語形とする。

【例】ヅボン→ズボン

○「ヴ」

書字形「ヴァ」「ヴヮ」「ヴハ」は「バ」を語形とする。

【例】ヴァイオレット→バイオレット、リヴヮー→リバー、ヴハンクヴアー→バンクバー

書字形「ヴィ」「ヴヰ」は「ビ」を語形とする。

【例】ヴィクトリア→ビクトリア、シヴヰリゼーション→シビリゼーション

書字形「ヴェ」「ヴヱ」「ヴヱ゛」は「ベ」を語形とする。

【例】ヴェクトル→ベクトル、ヴヱネチア→ベネチア、ヴヱ゛イ→ベイ

書字形「ヴォ」「ヴヲ」は「ボ」を語形とする。

【例】ヴォルト→ボルト、レヴヲリユーシヨン→レボリューション

書字形「ヴャ」「ヴュ」「ヴョ」は「ビャ」「ビュ」「ビョ」を語形とする。

【例】ヴュルテンベルヒ→ビュルテンベルヒ

上記以外の書字形「ヴ」は「ブ」を語形とする。

【例】ヘヴン→ヘブン

○「ワ゛」「ヷ」
書字形「ワ゛」「ヷ」は「バ」を語形とする。

【例】アワ゛ンチユール→アバンチュール、ペンシルワ゛ニア→ペンシルバニア

○小書き「ヮ」
「ウヮ」は「ワ」を語形とする。

【例】ドウヮー→ドワー

「ヴヮ」は「バ」を語形とする。

【例】リヴヮー→リバー、ドヴヮー→ドバー

上記以外の書字形「ヮ」は「ワ」を語形とする。

【例】インクヮイアラ→インクワイアラ

○「ヰ」
書字形「ウヰ」は「ウィ」または「イ」を語形とする。「ウィ」「イ」のどちらを語形とするかは、辞書の見出しや原音を参照して語ごとに定めるものとする。

【例】ルードウヰヒ→ルードウィヒ、ダーウヰン→ダーウィン
ウヰスキー→ウイスキー

書字形「テヰ」「デヰ」「フヰ」「ツヰ」はそれぞれ「ティ」「ディ」「フィ」「ツィ」を語形とする。

【例】テヰール→ティール、グランデヰー→グランディー、フヰリツピン→フィリッピン、ツヰング→ツィング

書字形「クヰ」「グヰ」はそれぞれ「クイ」「グイ」を語形とする

【例】クヰーン→クイーン、グヰツチヨク→グイッチョク

書字形「ヴヰ」「ヰ゛」は「ビ」を語形とする。

【例】シヴヰリゼーション→シビリゼーション、スカンディナヰ゛ア→スカンディナビア

イ段の仮名に続く書字形「ヰ」で長音を表すと考えられるものは、長音符号「ー」を語形とする。

【例】ルビヰ→ルビー

上記以外の書字形「ヰ」は「ウィ」または「イ」を語形とする。「ウィ」「イ」のどちらの語形とするかは､辞書の見出しや原音を参照して語ごとに定めるものとする。

【例】ヰーン→ウィーン、サンドヰツチ→サンドウィッチ

○「ヰ゛」「ヸ」
書字形「ヰ゛」「ヸ」は「ビ」を語形とする。

【例】スカンディナヰ゛ア→スカンディナビア

○「ヱ」
書字形「シヱ」「チヱ」「ツヱ」「フヱ」「ウヱ」はそれぞれ「シェ」「チェ」「ツェ」「フェ」「ウェ」を語形とする。

【例】シヱークスピーア→シェークスピーア、マンチヱスター→マンチェスター、カフヱ→カフェ、ツヱペリン→ツェペリン、ノルウヱー→ノルウェー

書字形「ジヱ」「ヂヱ」は「ジェ」を語形とする。

【例】サージヱン→サージェン、ブールヂヱー→ブールジェー

書字形「ヴヱ」は「ベ」を語形とする。

【例】ヴヱネチア→ベネチア、アドヴヱンテージ→アドベンテージ

エ段の仮名に続く書字形「ヱ」で長音を表すと考えられるものは、長音符号「ー」を語形とする。

【例】ペヱトル→ペートル、メッテヱル→メッテール

上記以外の「ヱ」は「エ」または「ウェ」を語形とする。「エ」「ウェ」のどちらを語形とするかは、辞書の見出しや原音を参照して語ごとに定めるものとする。

【例】ヱリザベス→エリザベス、サイヱンス→サイエンス
スヱーデン→スウェーデン

○「ヱ゛」「ゑ」
書字形「ヴヱ゛」「ヱ゛」「ゑ」は「ベ」を語形とする。

【例】ヴヱ゛イ→ベイ、レヱ゛ル→レベル

○「ヲ」
書字形「ツヲ」「フヲ」「ウヲ」はそれぞれ「ツォ」「フォ」「ウォ」を語形とする。

【例】ホーヘンツヲルレルン→ホーヘンツォルレルン、カリフヲルニヤ→カリフォルニヤ、ウヲートルロー→ウォートルロー

書字形「ヴヲ」は「ボ」を語形とする。

【例】レヴヲリユーシヨン→レボリューション

拗音「ョ」の代わりに臨時的に「ヲ」が用いられたと見なせるものは、「ョ」を用いた形を語形とする。

【例】ジヲルジ→ジョルジ、レリヂヲス→レリジョス

オ段の仮名に続く書字形「ヲ」で長音を表すと考えられるものは、長音符号「ー」を語形とする。

【例】レボヲリユーシヨン→レボーリューション

上記以外の書字形「ヲ」は「オ」または「ウォ」を語形とする。「オ」「ウォ」のどちらを語形とするかは､辞書の見出しや原音を参照して語ごとに定めるものとする。

【例】ナポレヲン→ナポレオン、ガリレヲ→ガリレオ、ヲデッサ→オデッサ
コーンヲール→コーンウォール、ヲートルロー→ウォートルロー、ヲルヅヲルス→ウォルズウォルス

○「ヲ゛」「ヴ」

書字形「ヲ゛」「ヴ」は「ボ」を語形とする。

【例】ヲ゛ルト→ボルト、ヲ゛ルテール→ボルテール

○「ヽ」

前の１文字を書字形とする。

【例】クロヽホルム→クロロホルム、ダヽイズム→ダダイズム、ヒッポヽタマス→ヒッポポタマス

ただし、前の文字が濁音・半濁音の場合、清音化した文字を書字形とする場合がある。

【例】（未確認）

上記の結果、他の規程を適用する必要のある形となる場合は、適用後の形を語形とする。

【例】オヽシス（→オオシス）→オーシス

○「ゞ」

前の文字が清音の場合、濁音化した文字を書字形とする。

【例】ハヾローフスク→ハバローフスク

前の文字が濁音の場合、同じ文字を書字形とする。

【例】バヾリヤ→ババリヤ

（4） 本表・付表Aにない、小書き「ャ・ュ・ョ・ァ・ィ・ゥ・ェ・ォ」を含む出現形について

「クュ・スュ・ツュ・ヌュ・ムュ・ルュ・グュ・ズュ・ブュ・プュ」は、それぞれ「キュ・シュ・チュ・ニュ・ミュ・リュ・ギュ・ジュ・ビュ・ピュ」を語形とする。また、「ヅュ」は、「ジュ」を語形とする。

【例】アヴァンツュール→アバンチュール、クルューゲル→クリューゲル、トリブューン→トリビューン、ヂユプュイトラン→ジュピュイトラン

「ケョ・セョ・テョ・ネョ・ヘョ・メョ・レョ・ゲョ・ゼョ・ベョ・ペョ」は、それぞれ「キョ・ショ・チョ・ニョ・ヒョ・ミョ・リョ・ギョ・ジョ・ビョ・ピョ」を語形とする。

【例】テョディー→チョディー、ゲョーテ→ギョーテ、ゼョン→ジョン

「ウァ」は「ワ」を語形とする。

【例】ハーウァード→ハーワード、ショツペンハウァー→ショッペンハワー

ア・イ・ウ・エ・オ列の仮名の後にそれぞれ小書き「ァ・ィ・ゥ・ェ・ォ」を用いた出現形は、長音符号を用いた形を語形とする。

【例】オスカァ→オスカー、ロセツチィ→ロセッチー

拗音「ャ・ュ・ョ」の代わりに臨時的に小書き「ァ・ゥ・ォ」が用いられたと見なせる出現形は、「ャ・ュ・ョ」を用いた形を語形とする。

【例】ギリシァ→ギリシャ、カリゥーム→カリューム、ジォン→ジョン

上記以外のものについては、大書きの仮名に直したものを語形とする。

【例】ワィズマン→ワイズマン、ニコラゥス→ニコラウス、スクェーア→スクエーア、ロェスレル→ロエスレル

ただし、辞書の見出しや原音を参照して異なる形を語形とする場合がある。

（５）　小書き仮名が大書きされた出現形について

本表・付表Ａにある小書きの仮名を用いた形や、上記規程であげた出現形に用いられる小書きの仮名を用いた形について、小書きの仮名ではなく大書きの仮名を用いた形が出現形となる場合がある。その場合、大書きの仮名を小書きに直した上で、必要な規程を適用して語形を定めることになる。

【例】ナチユラル→ナチュラル、インテリゲンツイア→インテリゲンツィア、ヴアイオリン（→ヴァイオリン）→バイオリン、ヴワルカン（→ヴゎルカン）→バルカン、ダブルユー（→ダブルュー）→ダブリュー

大書きの仮名が小書きに直すべきものなのか大書きのままとすべきものなのかは、辞書の見出しや原音を参照して語ごとに定めるものとする。

【例】「シヤ」シヤツ→シャツ、アカシヤ→アカシヤ
「ニヤ」コニヤツク→コニャック、アンモニヤ→アンモニヤ
「テイ」ソサイテイ→ソサイティ、ステイション→ステイション
「デイ」コメデイ→コメディ、デイリー→デイリー

（６）　その他

「ファ・フィ・フェ・フォ」の代わりに「フハ・フヒ・フヘ・フホ」が用いられたと見なせる出現形は、「ファ・フィ・フェ・フォ」を語形とする。

【例】アスフハルト→アスファルト、ソフヒア→ソフィア、フヘルヂナンド→フェルジナンド、カリフホーニア→カリフォーニア

資料２
出現形「に」の品詞判別基準

出現形「に」について、断定の助動詞「なり」の連用形かそれ以外の品詞（格助詞または接続助詞）かの判断基準について、以下に記述する。

原則、次の①～⑤に該当するものは断定の助動詞「なり」とし、それ以外を格助詞または接続助詞とする。

①先行語が形状詞のもの

【例】容易に其舊習を改めし者なるべし（助動詞）
僕竊かに疑なき能はず（助動詞）

ただし、形状詞が名詞的に用いられ、かつ「に」が意味上「～で（～であって）」と解せないものは、格助詞と判断する。これは本稿２．１．２に述べたように、現状の UniDic で形状詞として登録されている語が近代語で名詞として用いられても、既登録の形状詞をそのまま用いるために起こる事象である。

【例】多少の人力財用を無用有害に費す（格助詞）

②「あり」などの存在詞が後続し、意味上「～で（～であって）」と解せるもの（「～に（は）あらず」のように「あり」が打消の助動詞「ず」を伴うことが多い）

【例】是悦ぶべくして惡むべきにあらず（助動詞）
未曾有の事を新に始るには非ず（助動詞）

意味上「～で（～であって）」と解せないものは、格助詞と判断する。

【例】耶蘇教を奉ずる一國ここにあり（格助詞）
政府の強きを致すは天下人民の同心を致すに在り（格助詞）

③「～にして」の形で用いられ、意味上「～で（～であって）」と解せるもの

【例】夫日曜は七曜の一にして、毎週の首なり、（助動詞）
倍根より以前の心靈の理學は空理のみにして實證なし（助動詞）

意味上「～で（～であって）」と解せないものは、格助詞と判断する。

【例】楊朱墨翟の相反せる兩極を合一にして之を社交の一大例規とし（格助詞）
スチプレーション箇條書を頼みにして弊害を防ぎ止めんとするの一法あるのみ（格助詞）

④係助詞「や」が後続し、「〜にやあらむ」と「あらむ」が補えそうなもの

【例】ただ艸木の幹一つにして枝八十に別るるを見て本は一つなりと誤れるにや（助動詞）

⑤やう（様）に、ごとくに

【例】我邦の制とは大に相違する所ある様に覺ゆるなり（助動詞）
正金は年々常例の如くに外出するなり（助動詞）