「中古和文 UniDic」における言語単位の設計

小椋秀樹　　須永哲矢　　小木曽智信　　近藤明日子　　田中牧郎

国立国語研究所

1. はじめに

国立国語研究所は，2011年度の公開を目指して『現代日本語書き言葉均衡コーパス』(以下，BCCWJ)の構築を進めている[1][2]。BCCWJには長短2種類の言語単位に基づく形態論情報をはじめ[3]，種々のアノテーションが施されており，これらを利用することで，コーパスを活用した現代日本語の研究が，今後大きく進展することが期待される。

一方，日本語の歴史的研究に関しても，国立国語研究所において共同研究プロジェクト「通時コーパスの設計」の下，通時コーパス設計の基礎的な研究や一部のコーパスの構築が計画されている[4]。

BCCWJ の構築計画においては，現代語用の形態素解析用辞書 UniDic が開発され[5]，BCCWJ に対して高い精度での形態論情報付与を実現している。通時コーパスの構築，更にはそれを活用した日本語の歴史的研究を考えた場合，現代語用の形態素解析用辞書だけでなく，歴史的資料の形態素解析に適した辞書の開発が求められる。

小木曽，小椋，近藤は，現代語用の UniDic を基に近代の文語論説文を対象とした「近代文語 UniDic」を開発し，一般に公開した[6]。それに続く歴史的資料を対象とした形態素解析用辞書として，筆者らは，平安時代（中古）の仮名文学作品を対象とした「中古和文 UniDic」の開発を進めている[7]。

本稿では，「中古和文 UniDic」の言語単位の設計等について述べるとともに，中古語をどのように単位認定し，辞書登録したのかについて報告する。

2. 「中古和文UniDic」の概要

2.1 開発の経緯

筆者らは，現代語用の UniDic を基にして日本語の歴史的資料の形態素解析を行うための形態素解析用辞書の開発を進めている。既に，近代の文語論説文を対象とした「近代文語 UniDic」を開発・公開しており，現在は中古の仮名文学作品を対象とした「中古和文 UniDic」の開発を進めている。

「近代文語 UniDic」は近代文語論説文という漢文訓読調の，漢語を多く含む文体を対象としたものであったため，和語が大半を占める中古の仮名文学作品を解析した場合には十分な解析精度を得ることができない。そこで，新たに中古の仮名文学作品を対象とする「中古和文 UniDic」を開発することとした。

2.2 作成方法

「中古和文 UniDic」を作成するためには，現代語用の UniDic，「近代文語 UniDic」に登録されていない語を登録することが必要である。そのための作業方法として，国語辞典の見出し語などを基に中古和文で使われる語を追加することが考えられる。

しかし，「中古和文 UniDic」の基になる 2 種の UniDic には，一般的な語の多くが既に登録されているため，国語辞典を利用した登録作業は効率的ではない。それよりも，実際に中古和文のテキストに出現した仮名遣いや送り仮名等が異なる異表記形を登録していく方が中古和文の解析に必要な語を効率的に登録できる。

そこで，「近代文語 UniDic」で解析した中古の仮名文学作品のテキストを人手修正して学習用コーパスを作成し，その過程で必要になった見出し語を UniDic に登録するという方法を取った。

「中古和文 UniDic」の開発に当たり，整備した学習用コーパスは表 1 のとおりである。

表1:「中古和文UniDic」の学習用コーパス

ジャンル	作　品	延べ語数
歌物語	伊勢物語	14,654
	大和物語	26,519
日　記	土佐日記	7,953
	紫式部日記	20,327
	更級日記	16,658
作り物語	源氏物語	173,020
合　計		259,131

2.3 解析精度

最新版の「中古和文 UniDic」の解析精度を，表 2 に示した。表 2 には現代語用の UniDic，「近代文語 UniDic」による中古和文の解析精度も併せて示した。

語彙素認定で約 97%という解析精度は，現代語用の UniDic で現代語のテキストを解析した場合の約 98%と比べて，必ずしも高い解析精度とは言えない。しかし現代語用の UniDic に比べて学習用コーパスの

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

規模が約 1/10 と非常に小さいことを考えると，高い解析精度を実現できていると言える。

表2:各種UniDicによる中古和文の解析精度

	中古和文 UniDic0.5	近代文語 UniDic1.1	(現代語) UniDic1.3.12
単位境界	99.31%	91.09%	81.86%
品詞認定	97.77%	83.78%	59.25%
語彙素認定	97.10%	78.73%	55.77%
発音形認定	96.64%	77.89%	55.05%

3. 「中古和文UniDic」の言語単位

3.1 採用した言語単位

中古の仮名文学作品については，既に数多くの索引が刊行されている。しかし，索引によって見出し語の認定の仕方に違いがある，一つの索引の中でも見出し語の認定に不統一があるといった問題が指摘されている[8]。そのため，作品間の語彙の比較などを容易に行うことができない。

「中古和文 UniDic」では，言語単位に現代語用の UniDic,「近代文語 UniDic」と同じ短単位を採用した。短単位は，単位境界の認定のほか品詞や見出し等の付与についても詳細な規定を設けており，ゆれの少ない言語単位を実現している。このゆれの少ない短単位を用いることで，形態素解析結果を利用した，中古の仮名文学作品間の語彙比較等が可能になるとともに，現代語用の UniDic,「近代文語 UniDic」の解析結果との比較による日本語の通時的な研究も可能となる。

3.2 短単位認定規定の概要

「中古和文 UniDic」で採用した短単位の認定規定について，その最も基本的な部分（短単位認定の原則）を確認しておく。

短単位は，次に示すように，語種等の別によってどのように短単位を認定するかが定められている。

《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。

【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/ /オレンジ/色/

《付属語》付属語 1 語を 1 短単位とする。

【例】 /が/ /だ/ /の/で/

また，以下のような例外を設けている。

《例外1》造語力の高い接辞・補助用言（これらを付属要素と呼ぶ。）は単独で 1 短単位とする。

【例】 /相/次ぐ/ /汗/ばむ/ /書き/易い/

《例外2》付属語を構成要素に持つもので現代語で 1 語化しているものは，付属語を分割しない。

【例】 /あく=まで/ /例え=ば/

このほか人名・地名などについても規定を設けている。

4. 中古語の短単位認定

4.1 短単位認定規定の中古語への適用

3 節で述べたように「中古和文 UniDic」では短単位を採用した。これにより，我々は 3.2 節に示したものをはじめとする短単位の諸規定に基づいて中古語の単位認定を行うことになる。

しかし短単位の認定規定は，現代語を基に作成したものであるため，中古語に対してそのまま適用できるわけではない。特に 3.2 節に示した規定のうち《例外 1》《例外 2》を何に対して適用するかについては，現代語と中古語とで判断が異なることが十分考えられる。《例外 1》では，現代語で造語力の高いものが中古語でも高いとは限らないし，またその逆も考えられる。《例外 2》では，1 語化しているか否かの判断は当然，現代語と中古語とで異なる。

しかし別の立場から見ると，《例外 1》《例外 2》を何に適用するかについて，中古語の実態を踏まえて考えることにより，中古語の実態に即した短単位認定規定の拡張を図ることができる。またその結果，短単位も中古語の研究に適したものとなる。

上にも述べたとおり，短単位の認定規定は現代語を基に作成しており，中古語には合わない面もある。これについては，規定の拡張を行う必要があるが，その場合でも，3.2 節に示した規定の枠内で拡張を行っていくのである。その結果，個別に見た場合，単位の認定に違いが生じるとしても，その単位を実現している基本的な考え方は同じであり，現代語用の UniDic,「近代文語 UniDic」と「中古和文 UniDic」との互換性は保たれると言えよう。なお，もしこのような方針を取らなければ，何を 1 語とするかという基本的な考え方が異なる，全く別の単位が「中古和文 UniDic」の中に混在することになる。

以下，ここで述べた考え方に基づいて，どのように中古語の短単位認定等を行ったか具体的に見ていくこととする。

4.2 UniDicの階層構造の活用

UniDic では，表記や語形の違いにかかわらず，同じ語であれば，同一の見出しを与えるという方針を取り，語を階層化した形で登録している。この階層の最上位を語彙素（国語辞典の見出しに相当）と呼んでおり，この語彙素の下に語形（形態の違いを区別する層），更に語形の下に書字形（表記の違いを区別する層）という階層を設けている。

語を，このような階層構造で登録した辞書を用いて形態素解析を行うことによって，例えば，ある語について，どのような語形の変異や表記のゆれが，どの程度あるのかという情報を容易に得ることができる。

我々は，この UniDic の階層構造を活用して，中古語と現代語・近代語とを統一的に扱うことを実現し

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

た。

「中古和文 UniDic」の学習用コーパスに出現した語のうち，現代語・近代語の UniDic に登録されていない語は，全くの新規登録語として新しく語彙素を立てた上で，語形・書字形を登録した(表 3)。この際，短単位認定や品詞等の付与は，現代語用の UniDic 等と同じ規定によって行う。

表3:UniDicへの中古語の登録(1)

語彙素	語形	書字形
サワラカ【爽らか】	サワラカ	さはらか
		爽らか
		さわらか

中古語には，現代語・近代語と語形が異なるものがある。例えば，「アキンド【商人】」は，中古では「アキビト」という現代とは異なる語形であった。このような語形が異なるものについては，既に立項されている語彙素の下に新規の語形として追加した。また，動詞，形容詞，助動詞，動詞型・形容詞型接尾辞の文語活用型も語形の層に追加した。

中古の仮名文学作品の表記を見ると，仮名書きされた語が多く，仮名遣いも現代語とは異なり歴史的仮名遣いが用いられている。これら異表記形については，書字形の層に追加した。(以上表 4)

表4:UniDicへの中古語の登録(2)

語彙素	語形	書字形
アキンド【商人】	アキンド	商人
		あきんど
	アキビト	**商人**
		あきびと
		あき人
ワラウ【笑う】	ワラウ（口語・五段）	笑う
		わらう
		嗤う
	ワラウ（文語・四段）	**笑ふ**
		わらふ
		嗤ふ

※ 太字は中古和文UniDicで新規追加したもの。

以上のように，現代語と語形，表記の異なる語を UniDic の階層構造を利用して現代語・近代語と同一の語彙素の下に登録することで，中古語のテキストに出現する様々な語形・表記を現代語・近代語とともに統一的に扱うことを実現している。

4.2 中古語の実態に即した拡張

(1)連体詞

現代語と中古語とでは，1 語化の度合いや文法的な振る舞いに違いのあるものがある。

現代語用の UniDic では，連体詞「この」「その」を 1 短単位としている。一方，中古語では「こ」「そ」が「こは忍ぶなり」(伊勢物語)，「そはいかに」(更級日記）のように単独で代名詞として用いられた例があり，「この」「その」がまだ 1 語化していないと考えられる。

このような例を踏まえ，「中古和文 UniDic」では，「こ」「そ」を代名詞と認め，「の」を付属語の認定規定に基づき 1 短単位とすることとした。つまり，「/こ/の/」「/そ/の/」のように 2 短単位としたのである。

また，現代語の連体詞「同じ」は形容詞から転じたものであるが，中古では形容詞として用いられている。そのため，「中古和文 UniDic」では「同じ」を形容詞とした。これと同様の例としては，「さる（然)」があり，「中古和文 UniDic」では，「さるかた」のような例については動詞「さり（然有)」の連体形としている。

なお，連体詞「さる」を認めず，動詞「さり」を認めることから，接続詞「さりとて」も「/さり/とて/」の 2 短単位に分割する。

(2)補助用言

先にも述べたとおり，付属要素（3.2 節《例外 1》）は，現代語と中古語とで違いが生じ得るものである。実際，コーパスを見ていると，現代語では付属要素となっていないものの中にも，中古語ではかなり造語力の高いものがある。例えば，次に挙げる動詞である。

行く（次第に～になるの意）　例：荒れ行く
詫ぶ（～しあぐむの意）　例：慰め詫ぶ

「行く」「詫びる」共に，現代語では付属要素としていない。この基準を，中古語にそのまま適用すると，「行く」については「更け行く」「打ち解け行く」「衰ひ行く」「増さり行く」「静まり行く」「弱り行く」「重り行く」など，「詫ぶ」については「逃れ詫ぶ」「忘れ詫ぶ」「あり詫ぶ」などを一つ一つ登録する必要が生じ，辞書登録作業の面で効率的とは言えない。また，解析結果を使った研究においても，これら補助用言は単独で 1 短単位として切り出されている方が扱いやすいと考えられる。

3.2 節で示したように，造語力の高い補助用言は単独で 1 短単位として扱おうというのが，短単位における基本的な考え方である。「中古和文 UniDic」では，その考え方に基づいて，「行く」「詫ぶ」を付属要素とすることとした。

なお，UniDic では，補助用言としての用法を持つ動詞に「動詞-非自立可能」という品詞を与える。「行く」は現代語で「～ていく」という形で補助用言としても使われるため，現代語用の UniDic において既に「動詞-非自立可能」として登録されている。一方，「詫ぶ」は現代語では補助用言用法を持たないため，「動詞-一般」となっている。これについては，文語形のみ「動詞-非自立可能」とした。

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

4.3 語の読みの問題

現代語においても「私」を《ワタクシ》と読むか《ワタシ》と読むか,「重複」を《チョウフク》と読むか《ジュウフク》と読むかなど読みを定めにくいものがある。

このように読みが定めにくい語については，現代語用の UniDic では，(1)現代における漢字使用の目安である常用漢字表の音訓による，(2)一般に規範的とされる読みを採用するといった基準を立て，それに基づいて一律に読みを決めている。上の例で言えば,「私」は常用漢字表に基づいて《ワタクシ》を,「重複」は規範的な読みである《チョウフク》を一律に採用している。

中古語でもこれと同様に読みの認定に迷う例がある。例えば，最も読みの定めにくいものとして接頭辞「御」がある。「御」の読みには《オ》《オン》《オオン》《ゴ》《ミ》の五つが考えられるため，基準を立てなければ，不統一が生じやすくなる。

このような例については,「中古和文 UniDic」でも，現代語と同様に基準を立て，その基準に従って読みを与えることとした。

「御」については,《オオン》とするのを原則とした。これは,『日本国語大辞典』(日国オンライン)で《オン》は院政期からで，中古は《オオン》という判断をしており，また古典の注釈書類でも「御」に《オン》という読みを与えたものは見られなかったことによる。

しかし《オオン》という読みも，あくまで原則であり,「御」が結合する語によっては《オ》《ゴ》《ミ》などで読むべきものがある。それらについては，個別に検討を加え，別途一覧表を作って，それに従って作業を行うこととした。以下，その例を挙げる。

《オ》　御前（おまえ）
《ゴ》　御椅子（ごいし）
《ミ》　御局（みつぼね）　御弟子（みでし）…

以上のように，語の読みの問題についても,「中古和文 UniDic」では現代語用の UniDic 等と同様の方法で対応している。

5. 終わりに

以上，本稿では,「中古和文 UniDic」における言語単位の設計等を報告するとともに，現代語を基に作成した短単位の認定規定をどのように中古語に適用し，辞書登録を行ったかについて具体例を挙げて解説した。本稿で述べたのは，中古和文を対象とした短単位認定規定の一部である。「中古和文 UniDic」の学習用コーパス整備に当たって，現代語と同様に，詳細な規定を整備している。

今後に残された課題はまだ多くあるが，一つ例を挙げると，名詞と形状詞の判定に係る問題がある。

「空」は現代語では名詞としてしか使われないが，中古語では「心も空にて」(源氏物語) など，心が空虚であるという意の形状詞としても用いられている。

UniDic では，名詞と形状詞の両用法を持つ語には「名詞-普通名詞-形状詞可能」が，形状詞としてのみ使われる語には「形状詞-一般」又は「形状詞-タリ」が品詞として付与される。

「空」の場合，現代語には形状詞用法がないため，中古語の「空」を別語彙素とすべきか，現代語・近代語で既に登録されている「空」と同じ語彙素にまとめた上で，品詞を「名詞-普通名詞-一般」から「名詞-普通名詞-形状詞可能」に変更すべきか，判断が難しい。現時点では，ひとまず別語彙素の立項も品詞の変更もせず，心が空虚であるという意の「空」も「名詞-普通名詞-一般」としている。今後，同種の他の例も含めて現代語用の UniDic 等との互換性を保持した処理案を検討する。

このような問題に対する規定を作成しつつ，更に学習用コーパスの整備を進め,「中古和文 UniDic」の解析精度の向上を図っていきたい。

参 考 文 献

[1] 山崎誠(2007)「『現代日本語書き言葉均衡コーパス』の基本設計について」『特定領域「日本語コーパス」平成 18 年度公開ワークショップ（研究成果報告会）予稿集』,127-136.

[2] 前川喜久雄(2008)「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発」『日本語の研究』4-1,82-95.

[3] 小椋秀樹ほか(2011)国立国語研究所内部報告書『『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』

[4] 共同研究プロジェクトの紹介「通時コーパスの設計」http://www.ninjal.ac.jp/research/project/a/corpus/

[5] 伝康晴ほか(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,国書刊行会.

[6] 小木曽智信ほか(2008)「近代文語文を対象とした形態素解析辞書・近代文語 UniDic」『日本語学会 2008 年度春季大会予稿集』,211-218.

[7] 小木曽智信ほか(2010)「中古和文を対象とした形態素解析辞書の開発」『情報処理学会研究報告』Vol.2010-CH-85,49-58.

[8] 宮島達夫(1969)「総索引への注文」『国語学』76,110-122.

関連URL

UniDic：http://download.unidic.org/
「近代文語 UniDic」「中古和文 UniDic」：
http://www2.ninjal.ac.jp/lrc/

付記　本研究は，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」(平成 21-23 年度，代表者:小木曽智信)による成果の一部である。

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.