形態素解析辞書「中古和文 UniDic」を用いた古文単語帳作成

大津　千尋，　三日市　綾花，　須永　哲矢（昭和女子大学）†

Compilation of Classical Literature Wordbooks Using an Electrical Dictionary for Morphological Analysis "Chuko-Wabun UniDic"

Chihiro Ohtsu,　Ayaka Mikkaichi,　Tetsuya Sunaga　(Showa Women's University)

要旨

形態素解析辞書「中古和文 UniDic」の教育転用の一例として、古文単語帳の作成を試み、作成方法の紹介と、作成結果から読み取れる言語事実の報告を行う。作成方法の概要は以下の通り。１）高校の古典教科書をテキストデータ化し、「中古和文 UniDic」により形態素解析、解析結果を Excel に出力する。２）解析結果をもとに高校の教科書に使用されている語の語彙頻度表を作成する。３）頻度表をもとに、単語帳に収録すべき古文単語を選定し、実例に基づいた訳語を充てる。今回の研究では、まずは特定の教科書 1 冊を元に単語帳の作成を目指し、「教科書に載るテキストの高頻度語」を明らかにした。教科書に出現する自立語延べ約 6500 語、異なり約 1500 語を対象に調査したところ、異なり語数にして全体の 2 割程度、300 語強でテキスト全体の約 7 割をカバーできることが明らかになった。ここで作成した単語リストを別の教科書テキストに対しても適用したところ、ほぼ同等のカバー率を得ることができ、有効性が確認できた。

１．はじめに

国立国語研究所「中古和文 UniDic」の公開により、特に機械処理の知識を持たない一般ユーザーであっても、歴史的資料に対して機械処理を行った研究が可能になっている。「中古和文 UniDic」は、現代語を対象とした従来の解析辞書では無力であった古典資料に対し、高精度で解析することを可能にした画期的な形態素解析辞書であり、実際これを利用したデータとして国立国語研究所「日本語歴史コーパス」の公開も始まっている。古典語のみならず、近年さまざまなコーパスが公開され、研究環境は充実しているが、コーパスを利用するという場合には、調査対象は自動的にコーパス化されているもののみに限られる。しかし研究目的によっては、コーパス化されている範囲と調査したい範囲が異なるという場合も十分ありうることで、そのような場合には自分でデータを作るということになる。その際には、特別な知識がない一般ユーザーにとっても使用しやすい UniDic は非常に有用である。形態素解析辞書「中古和文 UniDic」の利用の可能性は研究利用にとどまらず、須永(2014)のように教育面においても、主に高等学校での古典学習教材等、さまざまな活用法がありうる。本稿では、形態素解析辞書「中古和文 UniDic」の教育転用の一つとして、古典教科書本文をもとに形態素解析を行ったデータをもとに古文単語帳の作成を試み、その手順の紹介、および有効性の検証を行う。

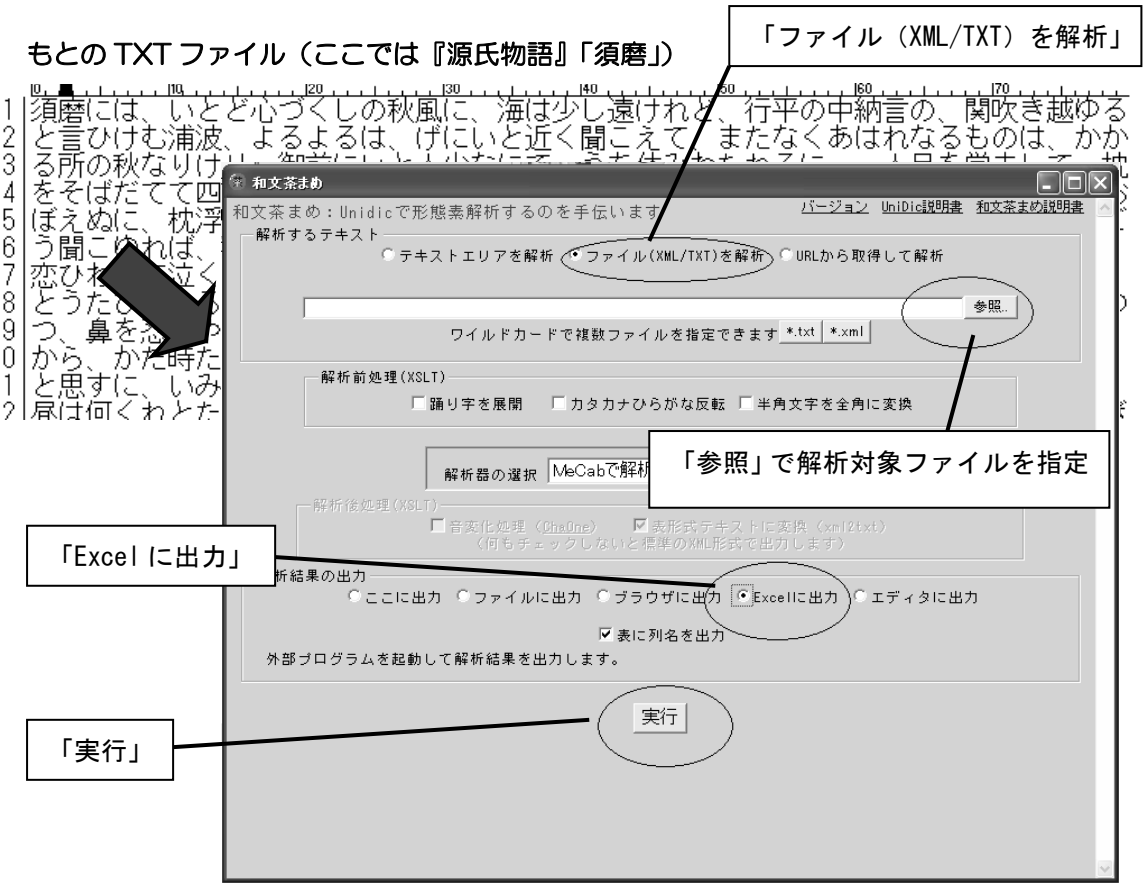
２．形態素解析辞書「中古和文 UniDic」とその利用

形態素解析とは、簡単に言えば「機械が自動で品詞分解して、活用の種類や活用形を書き出してくれる」というものである。公開されている「中古和文 UniDic」は中古和文 UniDic ホームページより無償でダウンロードできる。利用するには「MeCab0.96」以降（こちらも

†21112069@st.swu.ac.jp　21112318@st.swu.ac.jp　tsunaga@swu.ac.jp

無償）がインストールされていることが前提となるが、それも含め、ホームページでの指示に従ってダウンロード・インストールを行えば、特に機械処理に関する詳しい知識がなくとも、誰でも手軽に形態素解析を行う環境を手に入れることができる。

実際の操作にあたっては操作用ツール「和文茶まめ」が用意されており、ユーザはマウス操作で簡単に解析が行えるようになっている。古典本文を txt 形式で用意しておけば、あとはこの操作画面でファイルを指定してやれば、自動で品詞分解が完了する。（おおよそのイメージは図 1 参照)。



「和文茶まめ」（中古和文 UniDic の操作画面）

品詞分解が自動で行われた Excel ファイル

	1	2	3	4	5	6	7	8	9	10	11
1	出典	文境界	書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
2	源氏須磨	B	須磨	スマ	スマ	スマ	名詞-固有名詞-地名-一般			スマ	固
3	源氏須磨	I	に	ニ	ニ	に	助詞-格助詞			ニ	和
4	源氏須磨	I	は	ワ	ハ	は	助詞-係助詞			ハ	和
5	源氏須磨	I	、			、	補助記号-読点				記号
6	源氏須磨	I	いとど	イトド	イトド	いとど	副詞			イトド	和
7	源氏須磨	I	心づくし	ココロズクシ	ココロヅクシ	心尽くし	名詞-普通名詞-一般			ココロヅクシ	和
8	源氏須磨	I	の	ノ	ノ	の	助詞-格助詞			ノ	和
9	源氏須磨	I	秋風	アキカゼ	アキカゼ	秋風	名詞-普通名詞-一般			アキカゼ	和
10	源氏須磨	I	に	ニ	ニ	に	助詞-格助詞			ニ	和
11	源氏須磨	I	、			、	補助記号-読点				記号

図 1　操作画面「和文茶まめ」での操作と、出力される Excel ファイル

形態素解析を通し、機械が品詞分解をした結果、さまざまな情報が付与されるが、その中に「語彙素」という情報がある。「語彙素」とはいわば辞書見出し形であり、実際の表記・

活用形がどうであれ、辞書形・代表表記に戻したうえで語を表示する列であり、たとえば本文内の出現形が「はしる」であろうと「走ら」であろうと、語彙素レベルでは「走る」に統一される（図 2)。そこで、この「語彙素」列を利用することで、日本語で語を数える際の難関である、表記や活用形などの語形のゆれを乗り越えて、単語の数を自動で、正確に数え上げることが可能になる。

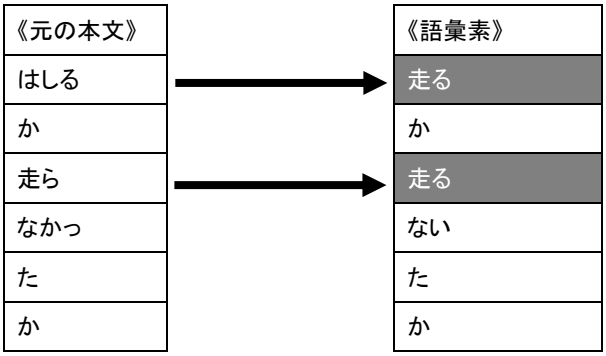


図 2　「語彙素」列のイメージ

３．古文単語帳の作成

上述の「語彙素」列を利用することにより、頻出語を抽出することが可能となる。「中古和文 UniDic」以前は、表記や活用の問題があり、古文テキストから単語を自動的に取り出すことは困難であった。表記や活用の問題が深刻でない英単語においては、機械処理をもとにした学習参考書・英単語帳が数多く見られるのに対し、古文単語帳の方ではそのような客観的根拠をもとにしたものがさほど見られなかったのは、このような事情によると思われる。そこで今回は、「中古和文 UniDic」での形態素解析を利用し、出現頻度という客観的根拠をもとにした単語帳の作成を試みたいと考えた。収録語数・レベルなどによって目標設定は変わりうるが、今回は第一回めの試作ということもあり、教科書に出現する単語を対象とし、必要最低限の入門的な単語帳、というレベルを想定している。

３．１　作成元となるテキスト

今回の単語帳作成元となる古文テキストは、高校の教科書 1 冊分とした。対象とした教科書は第一学習社『古典Ｂ』(2015 年度版)。『古典Ｂ』の中には中世以降、近世までのテキストも収録されており、中古のいわゆる「古典」とは毛色の違う作品も多い。「中古和文 UniDic」は中古語を対象としていること、また、学校教育において中世以降の作品に触れることはあっても、文法教育や単語教育の面においては実際のところ中古語に照準が合わせられていることを考え併せ、調査対象は中古のものに限定した。今回の試作で元とした古典作品は表 1 に示す 8 作品 26 話、総語数は 1 万 2860 語である。

表 1　単語抽出元とした作品（第一研究社『古典Ｂ』収録部分）

作品名	収録タイトル	語数
枕草子	「宮に初めて参りたるころ」「古今の草子を」「二月つごもりごろに」「ふと心劣りとかするものは」「この草子、目に見え心に思ふこと」	1764
源氏物語	「須磨の秋」「住吉参詣」「明石の姫君の入内」「紫の上の死」「薫と宇治の姫君」	3360
紫式部日記	「若宮誕生」「日本紀の御局」	585
更級日記	「門出」「源氏の五十余巻」「大納言殿の姫君」	1227
大鏡	「雲林院の菩提講」「花山院の出家」「道長と伊周―弓争ひ―」「時平と道真」	3692

	「兼通と兼家の不和」「道隆と福足君」「三舟の才」「道長と隆家」	
堤中納言物語	「このついで」	814
とりかへばや物語	「父大納言の苦悩」	659
しのびね物語	「偽りの別れ」	759
	計	12860

３．２　UniDic の単位認定と、単語帳作成面での精度

「中古和文 UniDic」はあくまで機械プログラムによって自動で品詞分解しているのであり、自動解析結果にはエラーも生じる。中古和文 UniDic は、平安仮名文学作品に対しては高い解析精度を実現しており、中古和文 UniDic　Ver0.5 の段階で、単位境界（品詞の切れ目が正しいか）で 99.3%、品詞認定で 97.8%という解析精度が報告されている（中古和文 UniDic ホームページほか）が、教科書のテキストに対してはどの程度の精度をもって解析が可能なのかは検証しなければならない。実際の作業においては、データの正確さのためには自動解析結果を人の目で確認、エラーを修正する必要がある。今回は自動解析に加え、人手による確認・修正作業も行った。今回解析に使用した「中古和文 UniDic」は Ver1.4(2014 年 3 月公開)である。

また、「中古和文 UniDic」が自動で「単語に分ける」という際の言語単位についても補足しておかねばならない。「中古和文 UniDic」は、国立国語研究所のデータ共通の言語単位として「短単位」という単位を採用しており、表 1 の語数もこの「短単位」の数による。「短単位」認定の詳細については規程集が公開されているためそちらを参照されたいが、一般的な高校教育での単語認定と、形態素解析結果の「短単位」としての語認定での相違点として注意せねばならないのは、以下の 2 点である。

①解析結果の 1 語は、一般的な高校教育での 1 語より小さい場合がある。

例えば高校教育では「吹き越ゆ」「大納言」などで 1 語とする方が一般的であるが、「中古和文 UniDic」では「吹く」＋「越ゆ」、「大」＋「納言」の 2 単位として解析される。

②解析結果の品詞・活用形認定は、一般的な高校教科書と異なる場合がある。

大きく異なるのは以下の 2 点である。（1）形容動詞の認定：UniDic の品詞体系では「形容動詞」はなく、いわゆる形容動詞語幹を「形状詞」、続く「なり」は断定の助動詞と認定する。例えば「きよらなり」は学校教育では形容動詞 1 語という認定だが、UniDic では形状詞「きよら」＋助動詞「なり」となる。（2）完了の助動詞「り」が接続する活用形は、学校教育では已然形が一般的であるのに対し、UniDic では命令形と認定する。高校の古典教材作成の用途・目的によっては、以上 2 点に注意し、修正が必要となる。

しかし、今回の目的は単語帳の作成であり、単語帳のための頻出語洗い出しという目的からは、上記①②はさほど問題にならない。まず①についてであるが、学校教育に倣って「吹く」「越ゆ」とは別個に動詞「吹き越ゆ」を認定し、別動詞として新たに指導するよりも、「吹き越ゆ」も分割して「吹く」と「越ゆ」の中に解消して処理する方が、一般性が高く、効率的である。このような複合語については、複合によって、元の語の足し算からは導けないような意味が生じる場合のみ、注意せねばならないが、大部分の、意味の足し算で複合語の意味も導けるような場合に関しては、むしろ UniDic のように分割して元の語だけを意識させる方が効率的である。①および②（1）に関しては、品詞認定と品詞分解の切れ目を示す教材を作成する、というような用途にとっては致命的だが、古文が読めるように、よく出る語を洗い出す、という用途にとっては問題は生じない。①に関しては可能な限り基本的な語に分解しておいた方が複合語として項目を立てるよりも一般性が高く有用であるし、②（1）の「形容動詞」／「形状詞＋なり」という認定の差についても、UniDic での「形状詞」を形容動詞として数え上げればよいだけの話であり、問題はない。②（2）に関しても、単語帳作成という範囲では、代表形としての「語彙素」が取り出せればよい

のであって、活用形の認定の違いは問題にならない。

以上のような観点から、単語帳作成のために単語抽出を行うという目的において、「中古和文 UniDic」が高校古典教科書に対してどの程度の精度を実現しているのか、エラーチェック作業を通して検証したところ、1 万 2860 語のうち、「語彙素」「品詞」レベルで語認定が誤っていたのはわずか 1 か所であった。高校の古典教科書に収録されるテキストは、高校生に読みやすいよう、表記、仮名遣いが統一された整ったテキストになっており、このようなテキストに対しては、「中古和文 UniDic」は通常以上の精度を達成できることが実証された。活用形などの認定込みで、別の学習教材を作成する場合、活用形レベルでのエラーを拾うとなるとエラーはもう少し増えるが、それとてたいした量ではなく、作業面において十分実用に足る精度と言える。極端な話、単語帳のための語彙頻度表を作成するだけなら、自動解析のままエラーチェックをしなくてもさして問題がないほどであると見てよかろう。

表 2　単語抽出目的における誤解析状況

作品名	語数	エラー
枕草子	1764	なし
源氏物語	3360	なし
紫式部日記	585	なし
更級日記	1227	なし
大鏡	3692	「さいつごろ」→接頭辞「さ」+「いつ頃」　（本来は先／つ／頃）
堤中納言物語	814	なし
とりかへばや物語	659	なし
しのびね物語	759	なし
計	12860	1か所

３．３　解析結果をもとにした語彙頻度表の作成

「中古和文 UniDic」では、解析結果を Excel に出力することができるので、解析結果をそのまま Excel データとして利用し、簡単に語彙頻度表を作成することができる。方法は人によってさまざまであるが、ここでは作業の中心となる手順の一例を紹介する。

（1）「語彙素」列をコピーする。

	A	B	C	D	E	F	G	H	I	J	K
1	出典	文境界	書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
2	紫式部日記	B	十月	カミナズキ	カミナヅキ	神無月	名詞-普通名詞-一般			カミナヅキ	和
3	紫式部日記	I	十	ジュー	ジュウ	十	名詞-数詞			ジュウ	漢
4	紫式部日記	I	余	ヨ	ヨ	余	名詞-数詞			ヨ	漢
5	紫式部日記	I	日	ニチ	ニチ	日	名詞-普通名詞-助数詞可能			ニチ	漢
6	紫式部日記	I	まで	マデ	マデ	まで	助詞-副助詞			マデ	和
7	紫式部日記	I	も	モ	モ	も	助詞-係助詞			モ	和

図 3　「語彙素」列を利用

（2）新しいシートにコピーした「語彙素」列を 1 列あけて 2 列コピーする。一方の列（図 4 ではC列）に対し、「データ」＞「重複の削除」で重複の削除を行う。A列がテキスト出現順に単語が並んでいるのに対し、C列は重複を削除したことにより、そのテキストの異なり語のリストとなる。この時点で、A列に並んでいる語の総数が延べ語数、C列の語の総数が異なり語数ということになる。

図 4　「重複の削除」を利用し、延べ語・異なり語リストを作成

（3）異なり語リストをもとに、延べ語の列における各語の出現数を計算する。ここでは COUNTIF 関数を使用する。COUNTIF 関数とは、指定した条件に一致するセルの個数を計測する関数で、図 5 のとおり、結果を表示させたいセルに直接「=countif」と入力する。（範囲,検索条件）の「範囲」は計測する範囲、「検索条件」は、ここでは計測対象とする語となる。図 5 では、「=countif(A:A,C2)」と指定しているが、これは「A:A」（A列全て、つまりテキスト上に出現した延べ語リスト）から、「C2」のセルにある文字列「昔」と一致するセルの数をカウントするよう指定していることになる。範囲や検索条件の指定は、直接入力せずとも、マウスのカーソル移動・指定でも可能である。

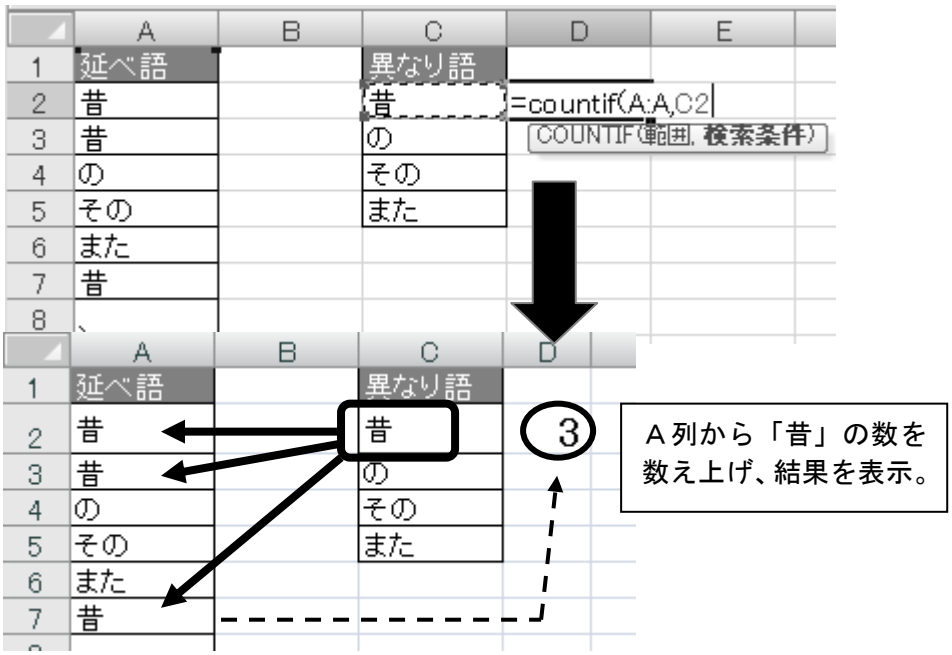


図 5　COUNTIF 関数の利用

（4）以上の操作で、単語の出現頻度を算出することが可能となる。この後は「並べ換え」などを利用し、高頻度順に並べ直したりすればよい。

４．古文単語帳の試作

以上の手順を利用して作成した語彙頻度表をもとに、高頻度語を抽出し、古文単語帳の作成を試みる。まず古文単語帳に収録する品詞の範囲であるが、助詞・助動詞といった付属語はむしろ「文法」の要点であり、数の上でも有限で、文法教育の側でカバーされる。このため、「単語帳」の収録対象は自立語に限定し、さらに固有名を排除することとした（頻出の固有名も将来的には収録すべきかと考えられるが、今回の試作では除外）。この時点で、元になるテキストの総語数は延べ 6488 語となる。

表 3　調査対象となる自立語（固有名除く）の延べ語数・異なり語数

延べ	異なり
6488	1485

４．１　「よく出る単語」の抽出方法

さて、各テキストの語のリストから「よく出る単語」を抽出するわけだが、何をもって「よく出る」とするかについては幾つか別の考え方がありうる。一つは素直に、教科書の対象テキスト全体から、出現数の高い順に語を取りだしていく、という方式であるが、この場合、ある作品のある箇所にのみ多数登場するが、他の作品ではほとんど登場しない、という語があった場合、たまたま教科書に載った箇所の特殊性ゆえに、高頻度語に位置づけられてしまう可能性もある。そこで別の方法として、「その語が何作品にまたがって出現するか」という尺度も導入することとする。今回対象となる古典作品は表 1 に示した 8 作品であり、総数は問わず、複数作品に出現した語を「よく出る」とする見方である。作品は問わず、全体の総数順で「よく出る」と認定した「総語数方式」と、総数は問わず、出現した作品数で「よく出る」と認定した「作品数方式」の２種を試し、有効性に差があるのかを以下で検証する。

４．２　総語数方式・作品数方式による単語抽出とカバー率

まず総語数方式で 4 回以上出現する語を抽出したところ、345 語であった。調査対象テキスト全体の異なり語数が 1485 であるため上位 23%を切り出したことになる。この 345 語で、実際のテキスト全体の自立語のうちどの程度がカバーできるかを算出したところ、72%がカバーできることが明らかになった。

続いて作品数方式であるが、作品数方式では出現作品数を 4 回以上とすると 325 語がこれに該当し、総語数方式で 4 回以上出現した語の語数とほぼ同じ規模になる。この場合のカバー率も 70%と、総語数方式とさほど差は出なかった。実際、両方式で抽出した 345 語・325 語のうち、278 語が共通であった。参考までに表 5，6 に各方式の上位 10 語を挙げるが、その大部分がどちらの方式で抽出しても取りだせるものであることがわかる。

総語数方式であれ、作品数方式であれ、よく出る単語の上位 2 割、300 語程度で、実際のテキストの 7 割ほどがカバーできるのである。

表 4　総語数方式・作品数方式のカバー率

	語数	作品全体の異なり語数に対するカバー率	作品全体の延べ語数に対するカバー率
総語数方式、4 回以上	345	23.2%	72.4%
作品数方式、4 作品以上	325	21.9%	70.6%

表 5　総語数方式による上位語 10 位（数字は出現語数）

形容詞・形容動詞		動詞		副詞		名詞	
なし	51	給ふ	371	いと	91	事	135
いみじ	43	す	129	かく	24	人	95
あはれなり	30	あり	114	然	20	程	59
をかし	21	思ふ	97	ただ	17	物	56
めでたし	18	見る	75	少し	16	様	49
怪し	15	言ふ	74	げに	16	心	45
あさまし	14	出づ	72	いかに	16	方	37
近し	12	侍り	67	なほ	16	世	36
とし	11	成る	51	え	14	一	27
悲し	11	申す	51	しばし	13	前	26

表 6　作品数方式による上位語 10 位（数字は作品数）

形容詞・形容動詞		動詞		副詞		名詞	
なし	8	給ふ	8	いと	8	一	8
いみじ	8	す	8	ただ	8	物	8
あはれなり	8	あり	8	いかで	8	方	8
近し	7	思ふ	8	かく	8	内	8
をかし	7	見る	8	え	7	世	8
口惜し	7	言ふ	8	しばし	6	程	8
怪し	6	出づ	8	少し	6	様	8
あさまし	6	侍り	8	なほ	6	人	7
心苦しい	5	成る	8	げに	5	事	7
悲し	5	覚ゆ	8	しばし	5	心	7
（他にも 5 作品出現語多数）		（他にも 8 作品出現語多数）		（他にも 5 作品出現語多数）		（他にも 7 作品出現語多数）	

※白抜きは総語数方式・作品数方式ともに出現

４．３　人による単語選定と、意味記述

以上、実数にして 300 語ほどでテキストの 7 割をカバーできる単語リストを得ることができるが、ここから人手の作業が残されており、この人手作業を経てこそ、単語帳の実用性は高まると考える。第一に意味記述の問題がある。形態素解析から作れるのは単語リストまでであり、教科書に合わせて必要十分な意味を記述していくのは人間の仕事ということになる。また、単語リストから覚える必要のない語を、人間の目で排除していくことで、単語数はさらに減らすことができる。たとえば表 5 の頻出名詞を見ると 1 位は「事」2 位は「人」3 位は「程」…となっており、これらは現代語にも共通する基本語彙であって、「古文単語」としてとりたてて覚える必要はない。「300 語ほど」とした語数の中にはこのような語も多数含まれるため、人間の目で選定していけば「カバー率 7 割の入門用の単語帳」は、より少ない語数で実現することが可能となる。現代では使わない古文特有の単語、および現代でも使う語ではあるが古文特有の意味・用法をもつ語を重点的に洗い出して記述

していくことで、より効率的な単語帳が作成できるはずである。

以上の手順で作成した語彙表をもとに、単語を予備的に選定したところ、この約 300 語から、実際覚える必要のある語は 120 語ほどという見通しを得た。「いみじ」や「具す」などに代表される、現代で使わない古文特有の単語としては 56 単語、「めでたし（→古典語では「すばらしい」）」や「驚く（→古典語では「目が覚める、気付く」）」のように、現代でも形式自体は使うが、古文特有の意味・用法をもつ単語として、64 単語というのがその内訳である。選定基準や、選定語そのものについては今後とも検討を要すると考えているため、今回のここでの報告はあくまで予備調査としての見通しにとどまるが、実用面を考慮し、人間の目で単語選定をすることによって、今回の語彙リストにおいては「古文単語」として覚えるべき基本語彙は半数以下になることが確認された。

５．実用性の検証

今回の語彙リスト作成の段階で、頻出語上位 300 語ほどで教科書の 7 割がカバーできることが明らかになった。ただしこれはあくまで１つの教科書をもとにした結果である。データを取る元となったテキストに対し、カバー率を測定したのであるから、この時点でカバー率が高くなるのはある意味当然といえる。ここで作成した単語リストが、他の同レベルのテキストでも有効なのか、あるいはあくまで今回対象とした教科書限定の単語帳なのかを明らかにせねば、このような単語帳の作成法が本当に有効なのかは判断ができない。そこで今回は検証実験として、作成した単語リストを、別の教科書の、今回採られていない話に対して適用し、その場合のカバー率を測定することとした。対象としたのは『大和物語』より「旅寝の夢」、今回データ採取対象の教科書には収録されていないが、教科書一般の定番である『源氏物語』より「葵の上と物の怪」「藤壺の里下がり」、および、後の時代の作品として『徒然草』より「あだし野の露消ゆるときなく」である。教科書に収録されている分量ということもあり、各話の総語数はさほど大きくない規模での検証実験である。

表 7　効果の検証に用いた別教科書のテキストと、その自立語総語数

	大和物語	源氏「物の怪」	源氏「藤壺」	徒然草	計
自立語総語数	102	435	409	97	1043

カバー率の検証結果は表 8 のとおりで、別教科書に適用しても、同時代の作品であればデータ採集元となった教科書とほぼ変わらない効力を発揮することが明らかになった。また、時代の異なる『徒然草』に対しては、やはりカバー率がやや下がることも確認された。

以上の検証から、教科書 1 冊をもとにした入門用の単語リストが、別教科書に対しても適用できる、一般性の高いものであると判断してよかろう。

表 8　別教科書に適用した際のカバー率の検証

	（元データ教科書）	大和物語	源氏「物の怪」	源氏「藤壺」	徒然草
総語数方式	72.4%	71.6%	69.0%	70.0%	64.0%
作品数方式	70.6%	70.6%	67.9%	68.0%	57.8%

また、今回試作した単語リストに収録された語が、これら別教科書において異なり語としてどの程度出現するのかという、稼働率の算出も試みた。表 7 のとおり、テキスト量がさほど大きくないため、検証に用いた 4 話を統合した上で、総語数方式・作品数方式の双方のリストと突き合わせ、稼働率を測定したところ、1000 語ほどのテキストを相手に 56%ほどの稼働率を見せ、汎用性の高さが証明された。なお、参考までに作品別にも稼働率を

算出したが、検証対象となる自立語総数が 100 語ほどの『大和物語』や『徒然草』は、当然稼働率は低く 1 割程度であり、「葵の上と物の怪」、「藤壺の里下がり」といった自立語総数 400 語程度のテキストになると、3 割台の稼働率を見せるようになる。これが 1000 語ほどのテキストに対しては稼働率 5 割半ばとなる。

表 9　別教科書を対象にした際の稼働率の検証

	徒然草	大和物語	源氏「藤壺」	源氏「物の怪」	4 話統合
	（97 語）	（102 語）	（409 語）	（435 語）	（1043 語）
総語数方式	11.6%	11.6%	33.6%	35.4%	56.5%
作品数方式	10.5%	11.4%	34.9%	36.7%	56.2%

以上の検証により、これらの単語リストは、カバー率の面でも、稼働率の面でも高成績と評価してよく、この単語リストは利用に際して、効率の良いものであると言えよう。

６．おわりに

以上、「中古和文 UniDic」を利用した学習教材開発の一環として、本稿では解析結果をもとにした単語帳作成の流れと、実効性の検証を行った。今回の研究で頻出語上位 300 語ほどで、古典教科書の 7 割ほどがカバーできること、また、語彙採集元とは別の教科書に対しても同様の有効性が見込めることが明らかになった。今後の作業としては、今回の単語リストをもとに実際に覚えるべき語の選定と、意味記述が待っているが、予備調査を通して得た見通しとしては、上位 300 語のうち、覚えるべき語は 120 語に減らせる見込みである。120 語覚えれば 7 割カバーできる、というのは非常に効率的であると考えられる上に、実際の学習上コストとしては、覚える語は 120 語より増やして、200 語、300 語程度にしてもまだまだ現実的な語数といえる。よって今後は、意味記述の精密化など、これに続く作業を継続するのはもちろんであるが、並行して、語彙リストをさらに拡充し、8 割程度をカバーできる単語帳作成なども目指していきたい。

文　献

小木曽智信・小椋秀樹・田中牧郎・近藤明日子・伝康晴（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 人文科学とコンピュータ』Vol.2010-CH-85(No.4) pp.1-8

小木曽智信・小椋秀樹・近藤明日子・須永哲矢(2010)「形態素解析辞書「中古和文 UniDic」とその活用例」『日本語学会 2010 年度秋季大会予稿集』 pp.243-248

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』短単位規程集第 4 版』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

小椋秀樹・須永哲矢(2012)『中古和文 UniDic 短単位規程集』平成 21（2009）- 平成 23（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」研究成果報告書 2（課題番号 21520492、代表者　小木曽智信）

須永哲矢(2014)「形態素解析辞書「中古和文 UniDic」を利用した古典学習教材の作成」『第 6 回コーパス日本語学ワークショップ予稿集』pp.11-20

関連 URL

日本語歴史コーパス「中納言」 http://maro.ninjal.ac.jp/

中古和文 UniDic http://www2.ninjal.ac.jp/lrc/index.php?UniDic

MeCab　http://taku910.github.io/mecab/