形態素解析用辞書UniDicへの語種情報の実装と政府刊行白書の語種比率の分析

小椋秀樹　　小木曽智信　　原裕　　小磯花絵　　冨士池優美

独立行政法人国立国語研究所

1. はじめに

国立国語研究所では，1976年から2005年までの30年間に出版された日本語の書き言葉を対象とした『現代日本語書き言葉均衡コーパス』（*Balanced Corpus of Contemporary Written Japanese*，以下BCCWJと略す。）の構築を進めている[1]。

BCCWJは，国語学・情報工学をはじめとする幅広い分野での活用を目指したコーパスであり，そのために様々な研究用の付加情報を付与する。これらの付加情報のうち形態論情報については，言語単位として，用例検索に適した短単位とBCCWJに格納したサンプルの言語的特徴の解明に適した長単位の2種類を採用した。この2種類の言語単位に基づいて，更に見出しや品詞等の情報を付与する[2]。

BCCWJは，1億語から成る大規模なコーパスであるため，形態論情報の付与には自動形態素解析システムが必須のものとなる。現在，国立国語研究所では，短単位の自動解析に使う形態素解析用辞書UniDic[3]への見出し語の追加など整備拡充作業を行い，解析精度の向上を図っているところである。また，2007年度からは，新たにUniDicの見出し語への語種情報の付与も開始し，既にすべての見出し語（約11万語）に語種情報を一通り付与した。語種情報は，現時点では他の形態素解析用辞書にはないものであり，UniDicの大きな特徴の一つと言うことができる。

UniDicに語種情報を実装したことによって，短単位解析の結果に，見出し・品詞などとともに語種も出力できるようになった。今後はUniDicを利用すれば，様々なテキストを対象として，同じ長さの言語単位を用いて語種構造の分析を容易に行うことができるようになる。UniDicへの語種情報の実装は，コーパスを使った日本語研究の可能性を広げるものと言うことができる。

本稿では，UniDicの見出し語に付与した語種情報の概要を紹介する。また併せて語種情報を活用した研究の一例として，BCCWJに格納する1976年から2005年刊行の政府刊行白書のデータを取り上げ，その語種比率の変動について報告する。

2. 語種情報の概要

2.1 語種の分類

日本語の語種は一般に，和語，漢語，外来語と，これら3種類の語種のうち異なる2種類以上の語が結合した混種語の4種類に分けられる。UniDicでは，この4種類のほかに固有名，記号，語種不明の3種類を加えた7種類に分類した。以下，各分類について簡単に説明する。

和語：日本固有の語。

漢語：近代以前に中国から入った語のほか，和製漢語（例:大根，返事）も漢語とした。

外来語：欧米系の諸言語から入った語のほか，和製英語も外来語とした（例:アフレコ，ナイター）。また，梵語等を中国で音訳した語に由来する語（例:盂蘭盆）やアイヌ語から入った語（例：昆布），中国以外のアジア諸国語から入った語（例：キムチ），近代以降に入った中国語（例：シュウマイ）も外来語とした。

混種語：和語・漢語・外来語のうち異なる2種類以上の語種の語が二つ以上結合した語（例：本箱）。漢語・外来語であったものの末尾が活用するようになった語（例：トラブる）。

固有名：人名・地名・組織名・商品名等。

記号：句読点・括弧などの補助記号や，箇条書きの項目名として使われた「ア」「イ」などの記号。

語種不明：語源が未詳であるため語種の判定ができない語。

2.2 語種の判定基準

UniDicの見出し語の語種を判定するに当たって，利用した資料は，『新潮現代国語辞典』第2版（新潮社，以下『新潮』と略す。）である。

『新潮』を用いたのは，この辞書が見出し語の語種を示しているからである。『新潮』は，見出し語が漢語・外来語の場合は片仮名で，和語及び不明の場合は平仮名で表記しているため，その表記を手掛かりにして語種を知ることができる。

UniDicの見出し語で『新潮』の見出しにない語は，『精選版日本国語大辞典』（小学館，以下『精選』と

略す。)を主たる資料として語種判定を行った。また,『新潮』の語種判定に従い難いと判断した場合は,『精選』等を参照し，独自に語種を判定した。例えば次のような語がある。

あまのじゃく:『新潮』の見出しは「あまのジャク」となっており，和語と漢語との混種語としている。しかし『精選』には「天探女（あまのさくめ）の転訛とする説が有力」とある。この説を採用して和語とした。

ごまかし・ごまかす:『新潮』では「ごまかし」を「ふくらんで中空の「胡麻（ごま）菓子」から」とし,「ごまかす」はその活用したものとして混種語とする。ただし「かす」を接尾語とする説もあり，語源が断定できないため,「ごまかし」「ごまかす」とも語種不明とした。

ちょうちょう（丁々）:『新潮』の見出しは片仮名表記で漢語としている。しかし「ちょうちょう」は擬音語で,「丁々」は当て字とする説によって,和語とした。

ところで,『新潮』を語種判定の資料に使う場合に注意すべき点がある。それは，見出し語が和語の場合のほか，語種が不明である場合も平仮名で表記している点である。見出し語が平仮名表記のものを一律に和語とすると，語種が不明であるため平仮名表記されていた語まで和語と判定してしまうことになる。そのため，見出し語が平仮名で表記されている場合,『新潮』の注記や他の辞書等を参照して和語とすべきか不明とすべきか適宜判断した。その結果,次のようにUniDicでは語種不明としたものもある。

だいだいいろ（橙色）:『新潮』の見出しは平仮名表記。「橙」は「代々」からの転とする説が有力だが，語種不明と判断した。語種不明の「橙」と和語「色」との結合であるため，語種不明とした。

わく（枠）:『新潮』の見出しは平仮名表記。注記に「前項（筆者注:矍）の字音か」とある。語源未詳と考え，語種不明とした。

以上のように，一つの辞書によって語種の判定を行うとしても，実際には種々の資料を参照して個別に判断する必要のある場合が少なくない。そのため,UniDicの語種判定は，国語史を専門とする研究者が担当した。

2.3 語種情報の実装

UniDicでは，表記が異なっても同じ語であれば,一つの見出し語にまとめ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と呼んでおり，この語彙素の下に語形，更に語形の下に書字形という階層が設けられている。語種情報は，これら三つの階層のうち最も上位の語彙素に付与した。その際，次の略称等を用いた。

和　… 和語　　漢　… 漢語
外　… 外来語　　混　… 混種語
固　… 固有名　　記号 … 記号
不明 … 語種不明

2.4 UniDic見出し語の語種比率

以上のようにしてUniDicの見出し語（語彙素）の語種を判定し，語種情報を付与した。ここでUniDicの見出し語の語種比率を見ておくことにする。

UniDicの見出し語（109,864語）を語種ごとに分類し，その語数と比率とを**表1**に示した。なお，表1には固有名・記号を除いた，いわゆる一般語（74,775語）の語種比率も示した。

UniDicの見出し語全体の中で最も多いのは漢語で,33.4%を占めている。次いで，固有名（29.6%），和語(22.2%)の順となっている。漢語は，一般語に限るとほぼ半数近くを占めている。

表1 UniDic 見出し語の語種比率

語 種	語 数	比率（全体）	比率（一般語）
和　語	24,356	22.2%	32.6%
漢　語	36,670	33.4%	49.0%
外来語	10,139	9.2%	13.6%
混種語	3,348	3.0%	4.5%
固有名	32,477	29.6%	―
記　号	2,612	2.4%	―
不　明	262	0.2%	0.4%

UniDicの見出し語全体の中で漢語・和語が55.6%,一般語に限ると81.6%を占めるのは，UniDicの整備拡充作業の中で国語辞典等を参照して漢語・和語を中心に未登録語を追加したことが関係していると思われる。また，先行して構築作業が進んでいるBCCWJの白書データ(約500万語)から未登録語の登録作業を行ったことも漢語の比率を高くしている要因と考えられる（白書の漢語の比率については後述する。)。

混種語の比率が低いのは，短単位の認定規定が関係している。例えば，サ変動詞「する」は1字漢語と結合する場合を除き単独で1短単位すると規定しているため,「運動する」「アドバイスする」は「運動/する」「アドバイス/する」と2短単位に分割される。また，外来語は原則として原語で1語に当たる要素（最小単位）を単独で1短単位すると規定している。そのため「オレンジ色」「テレビ局」が「オレンジ/色」「テレビ/局」と2短単位に分割される。

「運動する」「アドバイスする」「オレンジ色」「テレビ局」を1単位とすれば混種語となるが，以上のよ

うにそれぞれ2単位に分割されるため，各単位は漢語や外来語・和語に分類されることになる。その結果，混種語の比率が低くなっていると考えられる。

3. BCCWJ・白書データの語種比率

3.1 調査の概要

本節では，語種情報を活用した研究の一例としてBCCWJに格納する政府刊行白書のデータを取り上げ，その語種比率について分析を行う。

白書は，各省庁が作成する年次報告書である。白書は，国会に提出されるとともに，国民に向けて発表されることから，国が作成する文書の中でも極めて重要なものであり，それゆえ書き言葉の中でも特に公共性の高いものと位置付けられる。このような公共性の高い書き言葉における語の使用実態を語種の面から検討しようというのが，この語種比率調査のねらいである。

BCCWJに取られた白書の種類（タイトル数）は40種類で，内容は国土交通，外交，安全，教育，環境，福祉，科学技術，経済，農林水産の9分野にわたっている。

調査対象は，白書データのうち語彙や文字の統計的な調査に適した固定長サンプル（記号等を除く1,000字で構成されるサンプル）1,500である。このサンプルは，1976年から2005年の30年間に刊行された白書からランダムにサンプリングされたものである。この白書データをMeCab-0.96，UniDic-1.3.7で解析した。延べ語数は1,201,290語である。

語種比率の調査では，この中から助詞・助動詞・固有名詞・記号を除いた，いわゆる一般語を対象とした。また一般語のうち語種が不明のものも，今回は対象外とした。最終的に調査対象としたのは延べ語数で721,496語である。

なお，ここで助詞・助動詞を対象外としたのは，これらがすべて和語で，頻度も非常に高いことから，助詞・助動詞を含めて調査を行うと，延べ語数における和語の比率が著しく高くなってしまうためである。従来の国立国語研究所の語彙調査でも，語の頻度や語種比率を調査する際には，助詞・助動詞を別に集計している。今回もそれと同様の扱いをしたということである。

3.2 調査結果

対象とした白書の刊行年である1976年から2005年までの30年間を5年ごとに，第1期から第6期までの6期に分けて，各期における語種比率を調査した。

異なり語数における語種比率を**表2**に，延べ語数における語種比率を**表3**にまとめた。表2・表3ともに，各語種の語数と比率とを示した。

表2 白書の語種比率（期別・異なり語数）

	第1期	第2期	第3期	第4期	第5期	第6期
和	1,326	1,277	1,290	1,347	1,225	1,269
	19.9%	19.3%	19.6%	20.2%	19.2%	19.1%
漢	4,661	4,621	4,595	4,518	4,399	4,523
	70.1%	70.0%	69.7%	67.8%	68.9%	68.1%
外	509	563	558	648	636	712
	7.7%	8.5%	8.5%	9.7%	10.0%	10.7%
混	151	145	154	146	121	137
	2.3%	2.2%	2.3%	2.2%	1.9%	2.1%
合計	6,647	6,606	6,597	6,659	6,381	6,641

表3 白書の語種比率（期別・延べ語数）

	第1期	第2期	第3期	第4期	第5期	第6期
和	33,689	32,084	31,792	31,561	30,888	30,639
	27.0%	26.6%	26.2%	26.4%	25.9%	26.5%
漢	84,086	82,713	83,509	81,759	82,426	79,318
	67.4%	68.5%	68.7%	68.4%	69.1%	68.5%
外	3,733	3,437	3,776	3,908	3,670	3,681
	3.0%	2.8%	3.1%	3.3%	3.1%	3.2%
混	3,263	2,471	2,421	2,276	2,276	2,120
	2.6%	2.0%	2.0%	1.9%	1.9%	1.8%
合計	124,771	120,705	121,498	119,504	119,260	115,758

(1)異なり語数の語種比率

表2から分かるように,最も比率が高いのは漢語で，すべての期で約70%を占めている。それに次ぐのが和語で，各期とも約20%である。更に詳しく見ると，漢語の比率は第1期が最も高く70.1%で，その後緩やかに減少し，最も低い第4期では67.8%となっている。第1期と第6期とを比べると，2.0%の減少である。和語の比率は第4期で20.2%と最も高いが，他の期はいずれも約19%で，全期を通じて余り変動がない。

外来語の比率は第1期の7.7%から第2期・第3期では8%台，第4期では9%台，第5期以降は10%台となっており，徐々にではあるが増加している点が注目される。

以上のように，各期とも漢語の比率が非常に高いが，年代的な変化を見ると，第1期から第6期にかけてやや減少している。その一方で，外来語は徐々に増加していることが確認された。

(2)延べ語数の語種比率

表3を見ると，最も比率が高いのは漢語で，異なり語数における語種比率と同様，各期とも7割近い比率となっている。次に比率が高いのが和語で，すべての期で26%前後である。外来語の比率は低く，各期とも3%前後である。

異なり語数における語種比率では，漢語の比率が減少している一方，外来語の比率が増加するという年代的な変化が見られた。しかし延べ語数の語種比率は，期ごとに多少増減はあるものの，第1期から第6期にかけて増加又は減少するというような傾向は見られない。

(3)異なり語数と延べ語数の語種比率の比較

次に異なり語数における語種比率と延べ語数における語種比率とを比較する。

漢語は異なり語数における比率と延べ語数における比率の差が，他の語種に比べて小さい。第1期では異なり語数における比率の方が2.7%高いが，それ以降差は更に小さくなっている。

和語は，すべての期で延べ語数における比率が異なり語数における比率よりも7%程度高い。これは，漢語1語当たりの平均使用度数が各期とも約18であるのに対し，和語の平均度数が各期とも約24から25と漢語より高いことに起因している。和語は，語種比率では異なり・延べともに漢語には及ばないが，漢語よりも高頻度の語が多いと言える。

外来語は，各期とも異なり語数における比率が延べ語数における比率より高くなっている。またその差も第1期では4.7%であったが，次第に広がって第6期では7.5%となる。外来語1語当たりの平均度数を見ると，第1期は7.3であったものが，第6期では5.2となっている。このことから，外来語については，新しい語が使われるようになり種類は増えるが，そのほとんどが余り定着していない低頻度の語，あるいは特定の分野で使われる専門性の高い語であり，結果として延べ語数での比率が増えないと考えられる。

4.終わりに

以上，本稿では，形態素解析用辞書UniDicに実装した語種情報の概要を報告するとともに，語種情報を利用した日本語研究の一例として，BCCWJに格納する白書データの語種比率の調査結果を報告した。

BCCWJは2010年度の完成を目指して，現在構築を進めているところであり，その作業の中でUniDicについてもBCCWJのサンプルを基に辞書未登録語の登録作業を行っている。今後も新しく登録した見出し語に語種情報の付与を進めていく。

語種の計量的な研究を行うためには，テキストに出現したすべての語に対して語種を付与しなければならない。これはかなりの労力を要する作業であり，この種の作業を個人が行うことは容易ではない。そのため，日本語の研究において語種の計量的な研究は，活発に行われているとは言い難い。

しかし，今後はUniDicを使えば，同じ長さの言語単位を使って，様々なテキストの語種比率について容易に分析を行うことができるようになる。これにより，語種研究の進展が期待されるところである。

また，本稿では語種情報を用いた研究の一例として，BCCWJの白書データを対象とした語種比率調査の結果を報告した。その結果は，以下のとおりである。

(1) 異なり語数・延べ語数ともに漢語の比率が約7割で，最も高い。異なり語数では第1期から第6期にかけて比率がやや減少するが，延べ語数の比率にはほとんど変動がない。
(2) 和語は，異なり・延べともに漢語に次ぐ比率となっている。異なり語数の比率よりも延べ語数の比率の方が高いが，これは和語に高頻度の語が多いことによる。
(3) 外来語は，異なり語数では第1期から第6期にかけて比率が約3%増加しているが，延べ語数の比率にはほとんど変動がない。

上記の結果から，白書が漢語を多用する固い文体である可能性が示唆される。このように漢語比率が高い要因としては，白書が各省庁の年次報告書という性格から，法律用語・行政に関する専門用語を多用していることがあると考えられる。これらの点については，白書以外の資料を対象に語種や使用されている漢語について調査を行い，その結果と比較する必要がある。今後の課題としたい。

また，外来語の増加ということがよく言われるが，今回の調査から，白書に関しては，外来語の使用度数の増加ではなく，種類の増加と見ることができる。外来語の種類が増加しているが，その多くは定着度の低い段階，限られた分野での使用という段階にとどまっており，その結果，1語当たりの平均度数が低く，延べ語数での比率が増加しないと考えられる。このことが現代一般の外来語の使用実態として言えることなのかどうか，更に他の資料を調査していく必要がある。この点についても今後の課題としたい。

注

1 BCCWJの設計については，山崎誠（2007）を参照。
2 BCCWJの形態論情報については，小椋秀樹ほか(2007)，小椋秀樹（2007），冨士池優美ほか（2008）を参照。
3 UniDicについては，伝康晴ほか（2007）を参照。

参 考 文 献

小椋秀樹ほか（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（2007）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』短単位規程集 Version 1.2』(LR-CCG-06-01)．

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.

冨士池優美ほか（2008）「「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」本論文集所収.

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.

付記 本研究は，文部科学省科研費特定領域研究「日本語コーパス」による補助を得た。