語種を観点とした近代語と現代語の語彙の比較
―形態素解析辞書「近代文語 UniDic」「UniDic」を用いて―

近藤明日子　　小木曽智信
{ kondo, togiso } @kokken.go.jp

独立行政法人　国立国語研究所

１．はじめに

近年、日本語学・国語学の分野でもコーパスを利用した研究の気運が高まりつつある。さらに、従来この分野で大きな位置を占める、古い時代の資料を扱う歴史的研究においても、コーパスの利用が活発になることが期待されている。しかし、歴史的研究においてコーパスの利用が広がるためには解決しなければならない問題がいくつかあり、その最大のものとして、歴史的資料の形態素解析ができないことがあげられる。

発表者等はこの問題を解決する端緒として、近代の文語論説文（主として明治普通文）を対象とした形態素解析辞書「近代文語 UniDic」[i]を開発している。これは現代語用の解析辞書「UniDic」[ii]をベースにすることで、近代文語でも現代語でも同一の語認定基準で形態素解析を行えることを意図した辞書であり、両 UniDic がそろうことによって語レベルでの近代語と現代語の比較研究が可能になる。

本発表では、この両 UniDic を用いた近代文語と現代語の比較研究の一例として、語種を観点とした語彙の考察を行う。語種とは、日本語の語彙をその出自によって分類した種類であるが、日本語の語彙を把握する上で基本的な観点として、これまでの語彙研究でもしばしば取り上げられてきた（林（監修）、1982、pp60-72 など）。解析結果に語種属性を付与できることは両 UniDic の特長の一つであり、本発表は両 UniDic により初めてなし得る研究分野の一つである。

２．語彙調査の方法

最初に、本発表で語彙調査の対象とした資料について説明する。近代語の資料は国立国語研究所（編）（2005）『太陽コーパス』から抽出した。『太陽コーパス』は 1895 年から 1928 年にかけて刊行された総合雑誌『太陽』に基づくコーパスで、1895・1901・1909・1917・1925 年の 5 ヶ年分、計 3409 記事の全文が収録されており、近代語のコーパスとしては類を見ない大規模なものである。この『太陽コーパス』から「近代文語 UniDic」が主な対象とする文語論説文に近い性質を持つ文章として、地の文の文体が文語で、かつジャンルが文学以外（日本十進分類法（NDC）の 1 次区分が 0～8）の記事を抽出し調査対象とした。対象記事数は計 1160 記事である。現代語の資料は、国立国語研究所で構築が進められている「現代日本語書き言葉均衡コーパス」（BCCWJ）[iii]に収録予定の書籍サンプルから抽出した。書籍サンプルの中で近代語の調査対象資料に近い性質を持つ文章として、ジャンルが文学以外（NDC の 1 次区分が 0～8）の可変長サンプルを抽出し調査対象とした。対象サンプル数は計 9496 サンプルである。

資料の形態素解析は、近代語の資料は「近代文語 UniDic」（ver.0.91、非公開バージョン）、現代語の資料は「UniDic」（ver.1.3.10i、非公開バージョン）によって行った。形態素解析器は MeCab0.97[iv]を用いた。

解析された語の同語異語判別は、両 UniDic によって付与される属性[v]のうち、「語彙素読み」「語彙素表記」「語義」「品詞」の 4 属性を用い、これらの値がすべて一致するものを同語と見なし一つの見出し語のもとにまとめた。

そして、両 UniDic の付与する品詞属性の大分類が名詞・代名詞・形状詞・連体詞・副詞・接続詞・感動詞・動詞・形容詞・接頭辞・接尾辞の語を調査対象とし、助詞・助動詞・記号類は対象外とした。

３．語数と語種比率

まず、近代語と現代語それぞれの語数と語種比率を概観・比較してゆく。表 1 は、近代語と現代語の延べ語数・異なり語数を両 UniDic の付与する語種属性別に示したものである。表中の「未確定」の語種とは、本発表で用いた両 UniDic のバージョンでは語種属性が未整備となっている語である。

表 1　語種別の延べ語数・異なり語数

語種	近代語		現代語	
	延べ語数	異なり語数	延べ語数	異なり語数
和語	993,301	9,503	6,838,361	19,538
漢語	1,146,622	26,766	6,644,350	33,218
外来語	10,508	1,291	577,308	10,592
混種語	62,737	1,212	209,191	2,620
固有名	110,194	9,154	637,280	25,445
記号	243	19	26,390	880
不明	1,043	75	37,627	253
未確定	2,168	27	1,900	166
合計	2,326,816	48,047	14,972,407	92,712

さらに表 1 から、和語・漢語・外来語・混種語の 4 語種について、延べ語数の合計に対する語種比率をグラフで示したものが図 1、異なり語数の合計に対する語種比率をグラフで示したものが図 2 である。

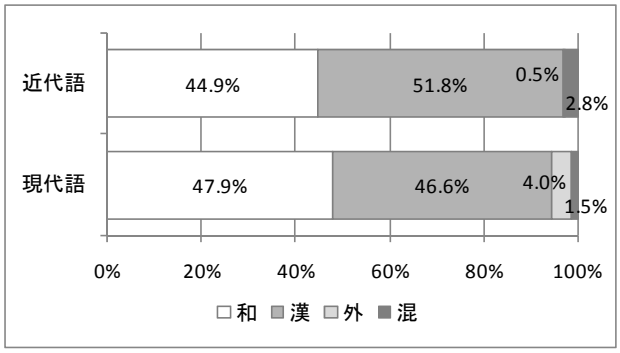

図 1　語種比率（延べ語数）

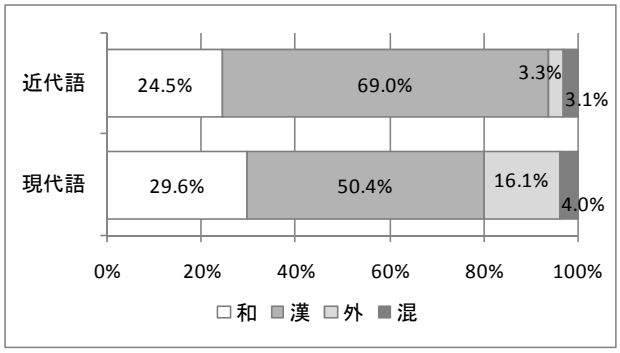

図 2　語種比率（異なり語数）

図 1・図 2 で近代語と現代語の語種比率を比較すると、和語比率と外来語比率は近代語＜現代語、漢語比率は近代語＞現代語という大小関係になっていることが分かる。この傾向は、異なり語数（図 2）のほうで顕著であり、近代語では現代語より漢語が活発に使われ、現代語では近代語より和語・外来語が活発に使われていると言える。

また、延べ語数と異なり語数との間で語種比率を比較すると、近代語・現代語ともに、和語比率は延べ＞異なり、漢語・外来語・混種語比率は異なり＜延べという大小関係になっていることが分かる。これは、近代語でも現代語でも和語は漢語・外来語・混種語よりも高頻度な語が多いことを示唆している。また、漢語比率は、近代語では延べ語数と異なり語数との間で差が大きいのに対し、現代語では差が小さい。これは、近代語より現代語のほうが漢語に高頻度な語が多いことを示唆している。

4．度数段階別の語種比率

3 で得られた示唆について確認するために、近代語・現代語それぞれの語彙を度数の高いものから A～F の 6 段階に分け、各段階での語種比率を考察する。段階分けは、度数の高い語から順に度数を累積したもの（累積度数）の延べ語数に占める割合（カバー率）によって以下のように設定した[vi]。

A　0～60%
B　60～80%
C　80～90%
D　90～95%
E　95～99%
F　99～100%

この方法により近代語・現代語それぞれの語彙の段階分けを行い、各段階に所属する語の異なり語数と度数およびカバー率を示したものが表 2・表 3 である。カバー率が上述の設定と一致していない段階があるが、これは度数が同じ複数の語すべてを累積しないうちに上述の各段階のカバー率の上限に達してしまう場合は、その同度数の語すべての累積度数のカバー率をその段階の上限として再設定することにしたためである。

表 2　近代語の度数段階の設定

段階	異なり語数	度数	カバー率
A	910	126331～375	0.0～60.1%
B	2758	374～83	60.1～80.1%
C	5043	82～27	80.1～90.2%
D	7194	26～11	90.2～95.4%
E	15624	10～3	95.4～99.0%
F	16518	2～1	99.9～100.0%

表 3　現代語の度数段階の設定

段階	異なり語数	度数	カバー率
A	908	683152～2316	0.0～60.0%
B	2947	2315～481	60.0～80.0%
C	5969	480～137	80.0～90.0%
D	9102	136～51	90.0～95.1%
E	27135	50～9	95.1～99.0%
F	46651	8～1	99.0～100.0%

このようにして設定した度数段階別に、和語・漢語・外来語・混種語の 4 語種について、異なり語数ベースで語種比率を求めた。近代語について示したものが図 3、現代語について示したものが図 4 である。

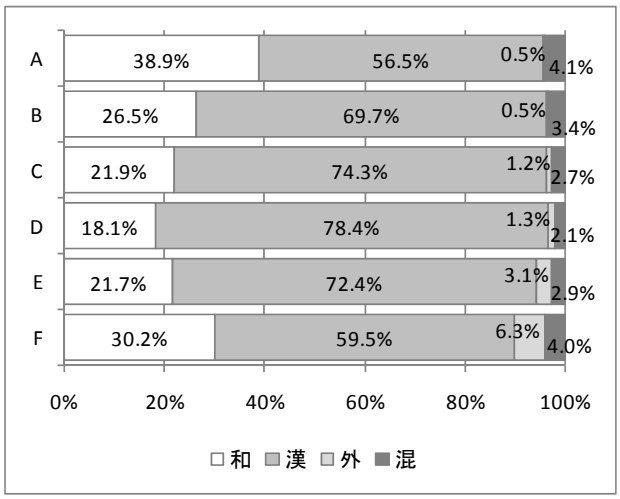

図 3　近代語の度数段階別語種比率（異なり）

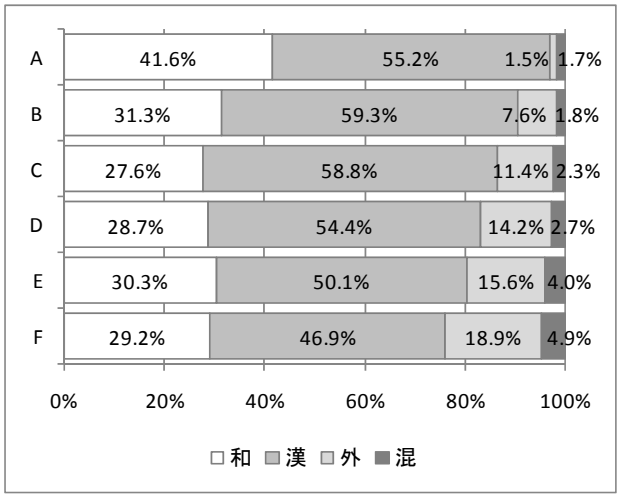

図 4　現代語の度数段階別語種比率（異なり）

まず、和語比率について見ると、近代語・現代語ともに A 段階が最も比率が高い。これは、近代語でも現代語でも和語には高頻度語が多いという 3 で得られた示唆を裏づけるものである。段階ごとの和語比率の変化を見ると、現代語では A～B 段階にかけて大きく減少し、B～F 段階はほぼ一定であるのに対し、近代語では A～D 段階で漸次減少し、D～F 段階にかけて再び増加する。これは、近代語の和語には低頻度な語も多いことを意味する。この理由の一つとして、調査対象資料とした『太陽コーパス』の文語記事に、引用等の形で口語文が混在しており、そこに使われる口語特有の語彙が低頻度語として現れたことが考えられる。近代文語の語彙調査という観点からは、口語文は除外して調査を行うべきであり、今後の課題としたい。

次に、漢語比率について見ると、近代語では D 段階が最も高いのに対し現代語では B 段階が最も高い。これは、近代語より現代語のほうが漢語に高頻度な語が多いという 3 での示唆を裏付けるものである。この背景として、「当初は低頻度語彙を含む広い範囲の語を用いていたところから、次第に限られた範囲の語を繰り返し用いるように、語彙が変化した」（田中、2005）という近代語における漢語の変化があったことが考えられる。つまり、現代語であれば一語で表される内容を、近代語では低頻度の類義語で使い分ける傾向にあったため、近代語では高頻度な漢語が少なくなっていると考えられるのである。

最後に、外来語比率について見ると、近代語・現代語ともに A～F 段階にかけて漸次増加し、F 段階が最も高い。近代語でも現代語でも外来語は非常に低頻度な語が多いことが分かる。また、すべての段階で近代語よりも現代語のほうが外来語比率の高いことも明らかになった。

5．語種別に見る近代語と現代語の語彙交替

以上の考察を踏まえ、具体的な語のレベルで近代語と現代語の語彙を比較するために、和語・漢語・外来語の 3 語種について、度数の多い順に上位 10 語を取り出して、高頻度語の交替について考察する。

まず、和語の上位 10 語を表 4 に示す。網掛けした語は、近代語・現代語の両方で上位 10 語に入ったものを表す（表 5・表 6 も同様）。

表 4　和語の度数上位 10 語

順位	近代語	現代語
1	為る(する)	為る(する)
2	有る	有る
3	其の	居る
4	此れ	事(こと)
5	事(こと)	言う
6	物	成る
7	於く	其の
8	持つ	無い
9	言う	此の
10	無い	物

和語は上位 10 語中 7 語が近代語と現代語で共通している。これは、高頻度な和語はその多くが、文体や年代の差を越えて、高頻度のまま使用されていることを示唆するものである。

次に、漢語の上位 10 語を表 4 に示す（ただし、数詞は除く）。

表 5　漢語の度数上位 10 語

順位	近代語	現代語
1	年（ねん）	的（てき）
2	第（だい）	様（よう）
3	者（しゃ）	年（ねん）
4	氏（し）	者（しゃ）
5	月（げつ）	自分
6	日（にち）	第（だい）
7	的（てき）	月（げつ）
8	会（かい）	人（じん）
9	上（じょう）	性（せい）
10	大（だい）	問題

漢語は上位 10 語中 5 語が共通しており、和語の 7 語より語数が少ない。これは、文体や年代の差を越えて高頻度な漢語が高頻度のまま使用される割合が、和語よりは低いことを示唆するものである。

最後に、外来語の上位 10 語を表 5 に示す。

表 6　外来語の度数上位 10 語

順位	近代語	現代語
1	露（ろ）※	パーセント
2	トン	ページ
3	ドル	システム
4	フィート	サービス
5	マイル	メートル
6	ポンド	テレビ
7	タバコ	データ
8	ガス	グループ
9	インチ	イメージ
10	クラブ	レベル

※「ロシア」のこと

外来語では、近代語と現代語で共通する語は全くない。高頻度な外来語は和語・漢語よりも語彙の交替が甚だしいことを示唆するものである。3 や 4 で見てきたように、外来語比率が近代語より現代語のほうが高いことを併せて考えれば、現代語で高頻度な外来語の多くは、近代語では用いられていない語であることが予想され、事実、現代語の上位 10 語中「サービス」「テレビ」「データ」「グループ」「イメージ」「レベル」の 6 語は近代語では度数 0 である。

6．おわりに

以上、「近代文語 UniDic」「UniDic」によって初めて可能となる研究の一例として、語種を観点とした近代語と現代語の語彙の比較を行った。両 UniDic により、コーパス言語学的な手法による歴史的資料の研究および通時的研究の端緒が開けたと言えよう。ただし、日本語の通時的変化を本格的に研究するためには、例えば近代語から現代語にかけての研究では、近代文語と現代語に対応した両 UniDic だけでなく、近代口語に対応した解析辞書も必要となる。また、より古い時代の資料の研究にはそれに対応した解析辞書が欠かせない。各時代・各文体に対応した解析辞書の開発もまた今後の課題である。

---

[i] http://www.kokken.go.jp/lrc/index.php?UniDic にて ver.0.9 を一般公開中。ver.0.91（未公開バージョン）の解析精度は、境界認定：99.43%、品詞認定：98.42%、語彙素認定 97.90%（いずれも F 値）となっている。

[ii] http://download.unidic.org/

[iii] BCCWJ の基本設計については山崎（2006）を参照のこと。

[iv] http://mecab.sourceforge.net/

[v] 両 UniDic の付与する属性の詳細については、同梱のマニュアルを参照のこと。

[vi] カバー率によって語彙を段階分けする方法は、田中（2008）を参照した。

参考文献

国立国語研究所編（編）（2005）　『国立国語研究所資料集 15　太陽コーパス　雑誌『太陽』日本語データベース』　博文館新社

田中牧郎（2005）　「言文一致と語彙の変化―『太陽コーパス』の二字漢語サ変動詞の分析による―」　『日本語学会 2005 年度秋季大会予稿集』　pp.197-204

田中牧郎（2008）　「語彙レベルの設定」　『特定領域研究「日本語コーパス」言語政策班中間報告書　言語政策に役立つ，コーパスを用いた語彙表・漢字表等の作成と活用』　pp.7-12

林大（監修）（1982）　『図説日本語　グラフで見ることばの姿』　角川書店

山崎誠（2006）　「『現代日本語書き言葉均衡コーパス』の基本設計について」　『特定領域研究「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』　pp.127-136

付記

本発表は、文部科学省科研費・特定領域研究「日本語コーパス」および日本学術振興会科研費・若手 B（課題番号 19720110）による成果の一部を含むものである。