『明六雑誌コーパス』の語彙量

近藤　明日子　（国立国語研究所コーパス開発センター）[1]

１．本稿の目的

本稿は、『明六雑誌コーパス』の語彙量の概要について報告するものである。

２．凡例

２．１．報告の対象

この報告では、『明六雑誌コーパス』の XML ファイルの SUW 要素（詳細は本報告書に収録した近藤明日子・田中牧郎「『明六雑誌コーパス』の仕様」を参照）1 つを 1 語として語彙量を集計する。ただし、SUW 要素のうち、形態論情報の付与の対象外としたものについては報告の対象外とする。対象外としたものは次のものである。

(1) 英語等の外国語の原語表記（SUW 要素 pos 属性値が「英単語」）
(2) 日本語のローマ字表記（SUW 要素 pos 属性値が「ローマ字文」）
(3) 漢文（SUW 要素 pos 属性値が「漢文」）
(4) 前後の文字が判読できないため形態論情報が付けられないもの（SUW 要素 pos 属性値が「読取不可」）

２．２．同語異語判別

異なり語数をカウントする際の同語異語判別には、『明六雑誌コーパス』の形態論情報付与の基盤となった、近代文語文を対象とする形態素解析辞書「近代文語 UniDic」の語彙素レベルを用いる。語彙素レベルとは辞書の見出し語に相当するもので、語形の揺れや書字形の違いを吸収し同語として扱うものである。

[1] kondo@ninjal.ac.jp

３．語彙量の報告

３．１．品詞別語彙量

品詞別に延べ語数・異なり語数を示す（表 1）。

品詞の分類は SUW 要素の pos 属性値の大分類に拠る。

表 1　品詞別語彙量

	延べ語数	異なり語数
名詞	58,428	10,823
代名詞	4,020	42
動詞	28,433	1,224
形容詞	2,298	126
形状詞	1,507	365
副詞	5,790	239
連体詞	4,626	17
接続詞	2,344	28
感動詞	52	8
接頭辞	1,062	45
接尾辞	2,070	198
助詞	52,199	62
助動詞	15,720	29
記号	43	19
補助記号	1,534	13
空白	479	1
合計	180,605	13,239

３．２．著者別語彙量

著者別に延べ語数・異なり語数を示す（表 2)。

延べ語数・異なり語数とも、記号類（品詞が「記号」「補助記号」「空白」の語）を除いて集計する。

著者の分類はコーパスの XML の article 要素 author 属性に拠る。よって、article 要素に含まれない各号の雑誌タイトル部分は集計対象外となる。

表 2　著者別語彙量

	延べ語数	異なり語数
西周	35,424	4,549
阪谷素	31,934	4,428
津田真道	26,187	3,887
西村茂樹	15,402	1,964
中村正直	12,121	2,310
杉亨二	11,502	2,187
森有礼	9,990	1,857
神田孝平	9,306	1,717
加藤弘之	7,236	1,111
箕作麟祥	5,611	1,073
福沢諭吉	4,623	1,008
柏原孝章	3,637	827
清水卯三郎	1,597	498
柴田昌吉	1,339	421
津田仙	1,068	419
箕作秋坪	1,068	341
合計	178,045	

３．３．文体別語彙量

記事の文体別に延べ語数・異なり語数を示す（表 3)。また、延べ語数における文体比率を示す（図 1)。

延べ語数・異なり語数とも、記号類（品詞が「記号」「補助記号」「空白」の語）を除いて集計する。

文体の分類はコーパスの XML の article 要素 style 属性に拠る。よって、article 要素に含まれない各号の雑誌タイトル部分は集計対象外となる。

表 3　文体別語彙量

	延べ語数	異なり語数
文語	167,832	12,642
口語	8,394	1,690
混在	1,819	651
合計	178,045	

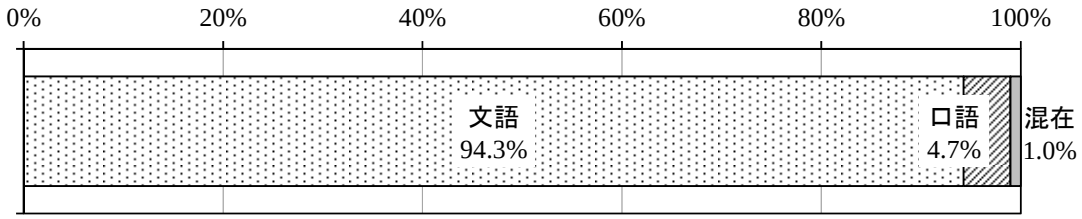

図 1　文体比率（延べ語数）

３．４．語種別語彙量

語種別に延べ語数・異なり語数を示す（表 4)。また、延べ語数および異なり語数での和語・漢語・外来語・混種語の比率を示す（図 2)。

延べ語数・異なり語数とも、記号類（品詞が「記号」「補助記号」「空白」の語）と助詞・助動詞を除いて集計する。

語種の分類は SUW 要素の wType 属性に拠る。wType 属性値の意味は次のとおりである。

和…和語
漢…漢語
外…外来語
混…混種語
固…固有名（品詞が「名詞-固有名詞」のもの）
記号…記号

表 4　語種別語彙量

	延べ語数	異なり語数
和	59,779	2,287
漢	43,965	9,504
外	559	250
混	3,685	378
固	2,638	682
記号	4	2
合計	110,630	13,103

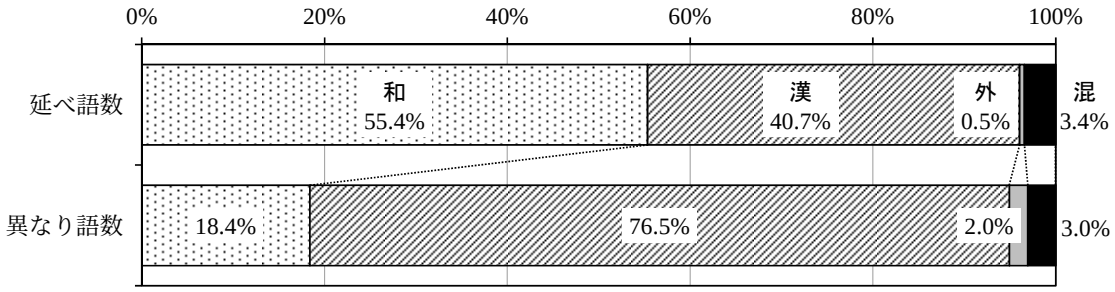

図 2　語種比率

３．５．文体・語種別語彙量

文語記事・口語記事ごとに語種別の延べ語数・異なり語数を示す（表 5・表 6)。また、文語記事・口語記事ごとに異なり語数での和語・漢語・外来語・混種語の比率を示す（図 3)。

延べ語数・異なり語数とも、記号類（品詞が「記号」「補助記号」「空白」の語）と助詞・助動詞を除いて集計する。

文体の分類はコーパスの XML の article 要素 style 属性に拠る。よって、article 要素に含まれない各号の雑誌タイトル部分は集計対象外となる。

語種の分類は SUW 要素の wType 属性に拠る。

表 5　語種別語彙量（文語）

	延べ語数	異なり語数
和	56,513	2,054
漢	41,288	9,269
外	497	227
混	3,514	360
固	2,406	652
記号	4	2
合計	104,222	12,564

表 6　語種別語彙量（口語）

	延べ語数	異なり語数
和	2,644	684
漢	1,884	778
外	60	36
混	138	53
固	109	69
記号	0	0
合計	4,835	1,620

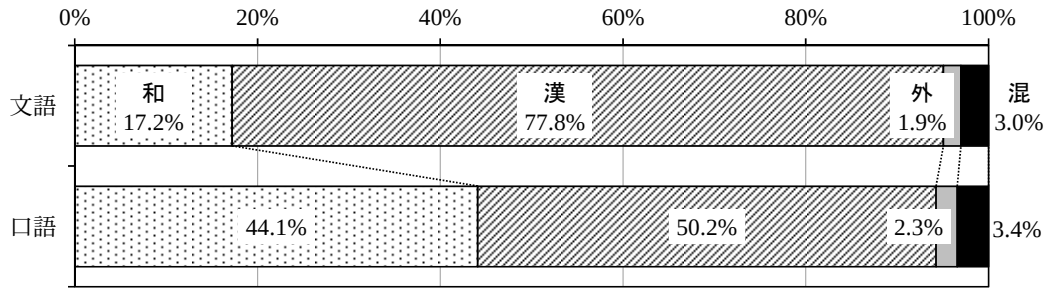

図 3　文体別語種比率（異なり語数）