『太陽』コーパスの作成とその活用

国立国語研究所　木村睦子
加藤安彦
田中牧郎
日本女子大学　藤原浩史

はじめに

日本語の言語資料を時代別に見渡すと、本文批評の歴史があり、索引化の行き届いた古い時代と、機械可読データの多い現代語とのはざまで、近世・近代語の研究資料は乏しく、国語史を研究する上で、大きな欠陥となっている。近年国文学研究資料館において、岩波書店の古典文学大系の機械可読化が進められていると聞くが、それも近世までであって、明治大正期の、しかも文学以外の言語資料については、索引も機械可読データもほとんど存在しない。その穴を埋めるべく、総合雑誌『太陽』の機械可読化と索引化を実施したい。索引化の方は、国立国語研究所国語辞典編集室の業務として 1988 年から進めているので、ここではコーパス作りをめざす。

１．資料の性格

『太陽』は明治 28 年から昭和 3 年まで 33 年間にわたって博文館から刊行された月刊の総合雑誌である。出版マスプロの先駆といわれる博文館が、当時刊行中だった５種の雑誌（それぞれ評論・商業・農業・法律・婦女の専門誌）をことごとく廃刊し、「太陽」一本にまとめたもので、廃刊され統合された雑誌の中には発行部数が数十万を数えるものもあったという。それだけに「太陽」は質量ともに卓越したものであり、内容が広範囲にわたるのみならず、各方面に一流の執筆陣を揃えていた。したがってこれを機械可読化することは、単に国語史の研究だけでなく、社会経済史、政治外交史、自然科学史などの研究にも貢献するところが大きいと思われる。

２．データベースの概要

２．１不採択部分

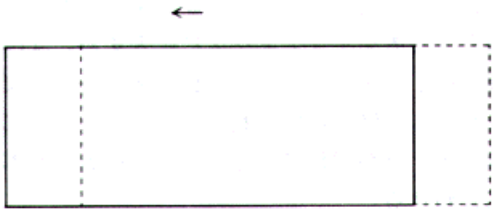
次のものは日本語研究資料として役に立たないと思われるので除外する。

（1）英文
（2）漢詩
（3）広告

ただし、広告は用例採集の対象になっているので、本来はコーパスに加えるべきものであるが、毎号同じものが掲載されたりするので、整理選択が必要である。したがって当面除外する。

２．２入力フォーマット

入力本文の表記や割付けは、なるべく原典に忠実な形にし、１行１レコードとする。各ページの先頭に、☆印を先頭にして資料名・刊行年月・ページを示す出典番号を入れる。何段かに分かれている場合には、代り目ごとに「@上」「@中」「@下」が入る。「原典に忠実」といっても、それによって不都合が生じる場合には変更を加える。たとえば、パラグラフの先頭以外で行頭字下げがある場合には、全体を行頭に移動する（破線位置→実線位置）。また外字にはゲタ記号を用い、微妙な異体字がある場合には、校正の際に不等号で囲んだりしている。



２．３テキスト・フォーマット

入力本文の各行をすべて次の形に変換する。〈出典番号〉〈段記号〉〈行番号〉,〈本文文字列〉〈改行〉

ただし、この形のものがつねに使い勝手がよいとは限らない。量がふえるのは別としても、各行の先頭に必ず出典番号等が入ることによって、既製のソフトでは行をまたぐ文字列検索ができなくなるという不都合が生じる。元の形式でも行末に改行コードが入るので、ソフトによっては同じ結果になるが、改行だけならば無視しうる検索ソフトもある。検索の便という観点からすれば、１センテンス１レコードという形式が望ましいと思うが、句点などで機械的に区切れない部分も少なくないので、実現には手間がかかる。

３．コーパスの利用法

３．１索引への文脈付加

国立国語研究所国語辞典編集室では、国語辞典編集のための用例採集の一環として『太陽』をとりあげ、1985,1901,1909,1917,1925,1928の各年における臨時増刊号を除く全６２冊（最終年のみ２冊）について、用例採集と索引作成作業を行ってきた。この用例採集は１行あたり約１個という高密度のものであり、採集部分は文節単位としてルビ付で入力しているが、これだけでは用例として不十分であり、原典を参照する必要が生じる。機械可読本文があれば、これに文脈を付加してＫＷＩＣにすることができる。その作業を１月号について試験的に実施した。作業手順は下記の通りである。

（１）エラーチェック

インデックス・ファイルと本文ファイルとを突き合せ、インデックスで指定された行に指定された文字列があるかどうかを調べ、あればその文字位置をインデックスに記入する。なければ前後１行ずつについて同様の検査をおこない、それでも見付からなければエラーファイルに書き出す。行がずれている場合は、行番号を本文に合わせる。

（２）複数チェック

指定された文字列が同じ行に二つ以上ないかどうかを調べ、あれば複数ファイルに書き出す。

（３）エラー箇所の探索および修正（人手による）

（４）複数ある場合の選定（人手による）

（５）ＫＷＩＣ作成

エラーチェックの際に付加した文字位置（バイト番号）に従ってＫＷＩＣを作成する。また、コーパスにはルビを入れていないが、インデックスの出現形部分にルビが付いている場合には、それを生かしてＫＷＩＣに取り込む。

３．２文字頻度調査

『太陽』コーパス 1901 年１～５月号の５冊分について、すべての文字の頻度調査をおこなった。各月および全体について、コード順文字度数表、頻度順文字度数表を作成し、さらに字種別延べ字数、異なり字数を数えた。

延べ字数、異なり字数は「表１.『太陽』コーパスにおける延べ字数と異なり字数」に掲げる通りである。英字など、号によるばらつきの大きいものもあるが、全体としてはわりあい平均しているようである。ただし外字はすべてゲタ記号で入力してあり、表面的には異なり字数１となっているが、むろん実際はもっと多いはずである。内訳はまだ調べていない。

４．今後の課題

索引のＫＷＩＣ化という第一の目的は今後も続くが、それ以外にも利用法は考えられる。索引が選択採集方式なので、索引中に求める用例がみつからなくても、本文中にみつかる可能性がある。しかしながら、求める用例がどのような文字列として存在しうるかを想像し、あますところなく列挙するのは非常に困難である。漢字の字種・字形の多様性、かなづかいを含む文字づかいの多様性、語尾変化等にどのように対処するかがコーパスを活用する上での大きな問題になるであろう。

［入力フォーマット］
☆01－01・01001
@上
太陽
明治三十四年
人生には限りありて◆冀望には際涯なし、有限の生を以て、◆無限の望を懷く、其發して理想となる者、千萬年後の光景を豫測すべし、◆四圍暗黒の社會に、哲人が黄金世界を豫言するは之が爲めなり、春秋の時代は、臣其君を弑し子其父を◆弑する暗黒世界なりしも、孔子は別に道義の世界を◆開拓して、苟も道を踐み義を行はゞ、四海の内皆兄弟なりと◆道破せり、プラトの理想的社會を説けるは、希臘國民が大哲ソクラテスを◆刑殺せし昏昧時代に在り、トーマスモーアのユトピアを◆草せるは、英國の◆暴王が無罪の彼を刑せる時代に於てせり、哲人の

［テキスト・フォーマット］
01－01・01001 C 01, 太陽
01－01・01001 C 02, 明治三十四年
01－01・01001 C 03, 人生には限りありて冀望には際涯なし、有限の生を以て、無
01－01・01001 C 04, 限の望を懷く、其發して理想となる者、千萬年後の光景を豫
01－01・01001 C 05, 測すべし、四圍暗黒の社會に、哲人が黄金世界を豫言するは
01－01・01001 C 06, 之が爲めなり、春秋の時代は、臣其君を弑し子其父を弑する
01－01・01001 C 07, 暗黒世界なりしも、孔子は別に道義の世界を開拓して、苟も
01－01・01001 C 08, 道を踐み義を行はゞ、四海の内皆兄弟なりと道破せり、プラ
01－01・01001 C 09, トの理想的社會を説けるは、希臘國民が大哲ソクラテスを刑
01－01・01001 C 10, 殺せし昏昧時代に在り、トーマスモーアのユトピアを草せる
01－01・01001 C 11, は、英國の暴王が無罪の彼を刑せる時代に於てせり、哲人の

［インデックス・ファイル（エラーチェック後）］
きぼう, 01－01・01001, C, 03, 19, 冀望
むげん, 01－01・01001, C, 03, 53, 無限の
あんこく, 01－01・01001, C, 05, 11, 四圍暗黒の
しいあんこく, 01－01・01001, C, 05, 11, 四圍暗黒の
しいする, 01－01・01001, C, 06, 49, 弑する
かいたくする, 01－01・01001, C, 07, 41, 開拓し
どうはする, 01－01・01001, C, 08, 41, 道破し
けいさつする, 01－01・01001, C, 09, 53, 刑殺せし
そうする, 01－01・01001, C, 10, 49, 草せる
ぼうおう, 01－01・01001, C, 11, 11, 暴王
くうき, 01－01・01001, C, 12, 25, 空氣
あんちゅう, 01－01・01001, C, 13, 5, 暗中
てつじん, 01－01・01001, C, 14, 17, 哲人

［KWIC］
きぼう, 01－01・01001, A, 03, 19, 冀望
　人生には限りありて▼冀望には際涯なし、有限の生を以て、無
むげん, 01－01・01001, A, 03, 53, 無限の
　冀望には際涯なし、有限の生を以て、▼無限の望を懷く、其發して理想となる者、
あんこく, 01－01・01001, A, 05, 11, 四圍暗黒の
　る者、千萬年後の光景を豫測すべし、▼四圍暗黒の社會に、哲人が黄金世界を豫
しいあんこく, 01－01・01001, A, 05, 11, 四圍暗黒の
　る者、千萬年後の光景を豫測すべし、▼四圍暗黒の社會に、哲人が黄金世界を豫
しいする, 01－01・01001, A, 06, 49, 弑する
　春秋の時代は、臣其君を弑し子其父を▼弑する暗黒世界なりしも、孔子は別に道
かいたくする, 01－01・01001, A, 07, 41, 開拓し
　界なりしも、孔子は別に道義の世界を▼開拓して、苟も道を踐み義を行はゞ、四
どうはする, 01－01・01001, A, 08, 41, 道破し
　み義を行はゞ、四海の内皆兄弟なりと▼道破せり、プラトの理想的社會を説ける
けいさつする, 01－01・01001, C, 09, 53, 刑殺せし
　けるは、希臘國民が大哲ソクラテスを▼刑殺せし昏昧時代に在り、トーマスモー
そうする, 01－01・01001, A, 10, 49, 草せる
　に在り、トーマスモーアのユトピアを▼草せるは、英國の暴王が無罪の彼を刑せ
ぼうおう, 01－01・01001, A, 11, 11, 暴王
　モーアのユトピアを草せるは、英國の▼暴王が無罪の彼を刑せる時代に於てせり、

	1月号		2月号		3月号		4月号		5月号		全体	
字種	延べ	異なり	延べ	異なり	延べ	異なり	延べ	異なり	延べ	異なり	延べ	異なり
記号	23895	43	20526	47	22186	48	24932	47	21599	48	113138	66
算用数字	2	1	48	9	37	10	79	10	235	10	401	10
英字	1121	41	112	32	686	45	99	29	74	28	2092	50
ひらがな	115629	75	116299	74	115739	73	118251	71	116797	72	582715	76
カタカナ	5878	80	5870	81	3812	78	3694	78	4730	82	23984	84
第一水準漢字	107233	2359	111636	2352	113758	2409	117022	2147	112523	2330	562172	2752
第二水準漢字	16076	940	16478	955	18257	1190	18381	1159	17260	919	86452	1941
ゲタ（外字）	46	1	91	1	145	1	195	1	79	1	556	1
計	269880	3450	271060	3551	274620	3854	282653	3812	273297	3490	1371510	4980

表1．『太陽』コーパスにおける延べ字数と異なり字数

1	一	10377	34	學	2405	67	至	1622
2	其	7239	35	八	2400	68	世	1611
3	國	6375	36	等	2377	69	道	1608
4	人	6074	37	今	2373	70	要	1607
5	二	4805	38	方	2359	71	合	1580
6	大	4795	39	々	2321	72	自	1580
7	十	4585	40	地	2300	73	金	1552
8	事	4311	41	同	2248	74	當	1552
9	日	4273	42	業	2228	75	内	1525
10	三	3963	43	然	2223	76	圓	1522
11	之	3901	44	法	2167	77	是	1508
12	似	3756	45	政	2148	78	又	1499
13	者	3660	46	六	2105	79	教	1487
14	如	3618	47	用	2052	80	前	1482
15	所	3521	48	云	1980	81	長	1437
16	此	3307	49	多	1943	82	月	1465
17	會	3267	50	實	1934	83	入	1431
18	上	3148	51	下	1895	84	發	1450
19	年	3141	52	間	1880	85	力	1437
20	五	3110	53	家	1856	86	無	1435
21	於	3021	54	物	1856	87	利	1400
22	時	2968	55	生	1854	88	後	1397
23	出	2893	56	分	1844	89	不	1396
24	行	2888	57	外	1839	90	社	1371
25	中	2815	58	七	1834	91	名	1371
26	見	2810	59	九	1825	92	立	1367
27	四	2736	60	有	1810	93	千	1365
28	○	2707	61	居	1740	94	彼	1353
29	爲	2653	62	百	1730	95	知	1341
30	來	2532	63	議	1699	96	言	1330
31	何	2498	64	我	1697	97	關	1329
32	本	2451	65	文	1668	98	第	1321
33	得	2406	66	及	1635	99	成	1309

表2．頻度順漢字リスト