『太陽』コーパスの作成

『太陽』コーパスの作成

国語辞典編集室　　田中牧郎
(E-mail　mtanaka@kokken.go.jp)

要旨

（1）国立国語研究所国語辞典編集室では、20世紀はじめを代表する総合雑誌『太陽』のコーパス作成を進めており、現在、1901年分が完成しつつある段階である。

（2）『太陽』コーパスを利用した調査研究の一例として、「国語」「日本語」という語を取り上げると、コーパス上の多様な用例を検索・整理することを通して、次の二点について明らかにすることができる。

・出現する記事分野の特徴から、文体的価値を解明できる。

・共起する語の特徴から、意味を分析できる。

（3）『太陽』コーパスを作成する場合に生じる重要な問題点のうち次の三点を紹介する。

・『太陽』の漢字をＪＩＳコードの漢字と同定する作業のなかで、符号位置の包摂規定をどのように行うのがよいか。

・『太陽』の原文テキストをコーパスデータに移し換える作業のなかで、本文をどのレベルまで修訂するのがよいか。

・『太陽』コーパスの利用と公開にあたって、著作権問題をどのように扱ったらよいか。

キーワード　　『太陽』コーパス　「国語」と「日本語」　文字コード　本文批判

1　『太陽』コーパスの概要

1.1『太陽』の資料性

1895年（明治28）～1928年（昭和３）刊行の月刊の総合雑誌。博文館刊行。創刊号を例にとれば、本文四六倍判200余頁、５号活字２段組原則、記事の内容は網羅的、発行部数は28万5千部（永嶺 1997)。『太陽』は、記事量の多さや幅の広さにおいても、よく読まれた点においても、当時を代表する言語資料として一級のものである。

1.2『太陽』コーパス作成の経緯

1983年：国語辞典編集準備室の用例採集文献の選定にあたって、国語辞典編集準備調査会委員10名全員が『改造』『文芸春秋』『婦人公論』『子供の科学』『帝国文学』『アララギ』『ホトトギス』などとともに、『太陽』を選定した（国立国語研究所国語辞典編集準備室 1983)。

1988年：『太陽』のうち、1895,1901,1909,1917,1925,1928（２号で終刊）の６年分62冊（臨時増刊を除く）を、「スカウト式用例採集事業」の最初の対象とした。

1995年：スカウトされた語にプログラムによって文脈をつけるために、全文入力を開始

した（国立国語研究所国語辞典編集室 1995）。

1998 年：「国研コーパス」（仮称）構築の一環として、『太陽』上記 62 冊の全文をコーパスデータ化することに重点を置くこととした。

現在：1901 年 12 冊分の本文のコーパスデータ化が完成しつつあるところである。本稿では、1901 年分『太陽』のコーパスデータを指して、「『太陽』1901」と称することとし、以下に示す数値や用例は、1998 年 10 月 23 日現在の『太陽』1901 に基づく。

1.3『太陽』コーパスの規模と特徴

『太陽』コーパスの規模を文字数を指標として見ると、『太陽』1901 の総字数は 330 万字余り（振り仮名は除く）。これまでの国立国語研究所の用字調査のうち、「郵便報知新聞」（1877 ～ 1878）が 17 万字（国立国語研究所 1959、進藤 1967）、雑誌九十種（1956）が 28 万字（国立国語研究所 1963）であるから、これらに比較してかなり大規模なものである。62 冊全文のコーパスデータ化が完成すれば、1700 万字程度の規模となることが予想され、朝日新聞CD-ROM（1993 年朝日新聞の 1 年分）調査の 1712 万字に（横山他 1998）匹敵する。

現代語のコーパスは情報処理の学界等で整備されてきており、近世までの文献資料のコーパスも国文学研究資料館等を中心に作成が進められている（安永1998）。そのはざまに位置する近代文献のコーパスについては、管見の限り本格的なプロジェクトは立ち上がっておらず、『太陽』コーパスはそこを埋めるものともなろう。また、コーパス作成が、歴史辞典編集を目標とする事業のなかで始まったものである点も、他のコーパスとは異なる特徴となっている。

『太陽』1901の記事分野別の文字数は、次の通りである。記事分野は、『太陽』の記事配列の枠組みを指標に分類したものである。このように非常に広範囲にわたる内容をもっている点も、『太陽』コーパスの特徴の一つである。

1 論説	417,939	9 歴史地理	427,327	17家庭	187,657
2 政治	106,302	10伝記	82,184	18投書	20,824
3 経済	272,909	11随筆	38,947	19編集	20,101
4 法律	74,408	12文芸	146,813	20談話	115,906
5 教育	60,551	13社会	60,064	21小説	290,175
6 宗教	55,514	14海外	162,961	22韻文	16,760
7 農工業	301,777	15世論	98,415	23漢文	77,705
8 科学	106,618	16彙報	176,100	24その他	385
				合　計	3,318,342

1.4『太陽』1901のコーパスデータの一例

```
【本文ファイル】
01-01-01017 B 01,　文明批評家としての
01-01-01017 B 02,文學者
01-01-01017 B 03,　（本邦文壇の側面評）
01-01-01017 B 04,　文學士　高山林次郎
```

01-01・01017 B 05,　一
01-01・01017 B 06,　予、人のニーツエを語るを聞く毎に其の書に接するの暇無
01-01・01017 B 07,きを恨みしや久し、頃來閑に乗じて彼が二三の著書を讀み其
01-01・01017 B 08,の要領を會するを得しが、初めは其説の太だ意表に出づるも
01-01・01017 B 09,のあるに驚きぬ。抑々歴史を尚ぶ進化論と平等を旨とする社

2『太陽』コーパス利用の一例

(『太陽』1901を用いた「国語」「日本語」等の分析)

2.1用例数　(文字列検索プログラムを利用した後、人手で確認)

国語　83例　　日本語　19例　　邦語　4例

言語　120例　　言葉　117例　　(単独用法のみの数)

2.2記事分野別分布

	国語	日本語	邦語	言語	言葉		国語	日本語	邦語	言語	言葉
1論説	20	11	1	25	4	13社会		1		6	
2政治						14海外		1		4	
3経済				3		15世論				5	
4法律			2	1	1	16彙報	2	1	1	1	
5教育	46			1	1	17家庭	2			16	5
6宗教	1			5		18投書					
7農工業	1	1				19編集					
8科学		3		24	4	20談話				1	
9歴史地理	5	1		18	20	21小説				1	67
10伝記				4		22韻文					1
11随筆				1	6	23漢文 注	−	−	−	−	−
12文芸	6			4	8	24その他					
						合　計	83	19	4	120	117

(注)　23漢文は、語の認定が困難な例があるので、算入しない。

＊「国語」は、教育に集中する点で、「日本語」「言語」と異なる。
＊「日本語」は、科学・社会・海外に見られる点で「国語」と異なる。
＊「言語」は、科学・歴史地理・家庭に目立つ点で、「国語」「日本語」と異なる。
＊「言葉」は、他の語がほとんど見られない小説に集中し、この語のみが日常語であったと考えられる。

2.3共起語

○対比語・並列語

・一人の教員にして朝に國語を教へ、夕に漢文を教へて可也。　01-01・03054 B 12
・そが歴史文明及國語宗教は彼等の自由に任せざるべからずと論じたり。　01-01・

13116 A 03

国語　漢文 21・漢文科１・漢学４・字音４・国文２・国字１・国史１・和歌１・修身１・宗教１・外国語１・西洋諸国の国語

日本語　イギリス語２・ラテン語１

邦語　外国語２・英文１

＊「国語」は、漢文との対比・並列として言われることが多い。また、他の文化概念と対比・並列することもあって、文化目録の一つとして意識されていたようである。

＊「日本語」「邦語」は他国語との対比・並列で言われる。

○修飾語

・然れども漢文は外國語也。純然たる國語に非ず。　01-01･04043 A 16

・英佛獨等著名の國語は少しも知らず、　01-01･02019 A 06

国語　日本の・我が・ポーランドの２・英仏独等著名の・そが２・単純の・純粋の・純然たる・荒怠したる・習用の病

日本語　我が

＊「国語」は、日本に限らず国の言語を指す。

＊「国語」には、価値評価を伴う修飾語がつくことがあり、〈純粋な状態〉と〈乱れて荒れた状態〉が対照的に問題にされている。

○動詞

・波蘭［ポーランド］の國語は學校及官廳より驅逐せられて、　01-01･13114 A 13

・國語の假名遣を改めるだけの勇氣はなかりき。勇氣はなかりしかど、國語もそれに準じて改めて可なるべしとの意見は有せしに似たり、　01-01･12045 B 08

国語　使用する・なす・本とする・作る・課程に入れる・教える２・独立する２・移用速進する・駆逐される・改める・改良する・簡便にする・自由に任す・兼ねる・素養ない・書く・国書を製する

日本語　使う・主とする・伝わる・教授する・包む・著述する・写す３・記す・記載する

邦語　著す２・記す

＊「国語」は、「独立」「移用」「駆逐」などの語とともに用いられ、他者との対立的な関わりの文脈にあらわれやすいことがわかる。

＊「国語」は、「改良する」「簡便にする」「自由に任す」などとともに用いられ、人為的に構築したものであるかのように見られていることがわかる。

＊「国語」は、「兼ねる」（自分のものとする）「素養ない」のように、教養として身につけるものとしてとらえられる面があったのではないか。

＊「日本語」に多い「写す」「記す」「記載する」は、「国語」には見られず、「日本語」が、音声や文字として言語が意識される場合に用いられやすかったことがわかる。

＊「邦語」は、「著す」「記す」に限られ、文字言語を表すことが普通であったことが思われる。

2.4まとめ

『太陽』1901 を検索して用例を分析するだけでも、上記のように、語の調査研究として、かなりのことを明らかにすることができる。2.2 記事分野別から見た語の文体的価値の解明や、2.3 共起語の整理による意味分析は、多様な用例を容易に検索・整理できるコーパスデータによって、多角的かつ効率的に進めることができるものであろう。ここで明らかにしたような諸点は、「国語」の詳細な語誌をまとめた京極（1993）でも触れられていないものが多い。

3　コーパス作成上の問題点

3.1　文字コードの問題

3.1.1　JISコード適用の問題点

現時点でのコーパスデータ作成においては、文字コードはＪＩＳコード（JISX 0208:1997）第一・第二水準を用いるのが通常である。『太陽』コーパスの場合も、古い時代の資料を扱う例に漏れず、ＪＩＳコードをどのように運用するのかが大きな問題となる。『太陽』には、ＪＩＳが示す字形とは異なる字形をもつ活字や、ＪＩＳには含まれない文字を用いている場合が多いからである。これについては、木村・田中・飯島 1997 で問題にしたことがあるが、その後の調査を踏まえて、さらに考察を進めることにする。ここでは、漢字に限って取り上げる。

3.1.2　字体の同定

『太陽』の字形とＪＩＳの字形とが全く一致する場合は、ひとまず問題はない。

Ａ　　娃　　阿　　哀　　愛

字形が異なる場合でも、ＪＩＳが示す「包摂規準」（芝野1997）によることで同定できる漢字が約700 字ある。例えば、次のようなものである。左側が『太陽』の字形、右側がＪＩＳの字形、括弧内は「包摂規準」の「連番」である。

Ｂ　　羽・羽　(18)　　青・青　(146)　　神・神　(161)

さらに、「包摂規準」に準じて字形の整理を行うと、次の漢字をはじめとして約 140 字が同定可能となる。ＪＩＳは示していない規準であるが、明治大正期の活字文献を扱う場合には、明示的な規準として立てることができると考えられる（飯島・大塚 1998）。

Ｃ　　監・監　　匁・匁　　熱・熱　　剳・剳

ＢＣは、字形の類似に基づいて規準を定めることができるものである。これら以外に、字形の類似は明示できないが、意味・機能の面から見て同定が可能な漢字として、つぎのようなものをあげることができる。

Ｄ　　欵・款　　煞・殺　　窻・窓　　明・明　　慙・慚

詳細は別に報告したいが、これらは字形から見れば包摂し難いものであるが、文脈的な意味や機能からは、『太陽』1901 に関する限り等価と見られるものである。この類の漢字は約 180 字ある。

このように、ＪＩＳの「包摂規準」（Ｂ）とそれを拡大した規準（Ｃ）、および意味・

機能上の等価という規準（D）を適用することによって、一見したところでのＪＩＳにない漢字の多さを、かなり解消することができる。これらの規準のいずれも適用できない約730字が、「外字」として残ることになる。

E　　丰　　𠂉　　份　　佔　　佾　　侔　　倘　　傖

『太陽』1901では、「外字」位置に＝（下駄記号）を入力し、出典コードによって外字表（画像または印字）と対照することで、字形を指示するようにしている。その際、UNIコード番号と『大漢和辞典』（大修館）の検字番号を添えるようにしている。

【外字表】

出典コード	用例	UNIコード	文字	大漢和巻〈頁〉検字番号
01-01･01001D15	＝より醇に赴かしむ	91A8	醕	1〈0391〉400010

【当該本文】

01-01･01001 D 15,人の社會を化して＝より醇に赴かしむ、此間一點の疑を容る

3.1.3　今後の問題点

現在、ＪＩＳ第三・第四水準の文字選定が進められており、これが一般化すれば、１万数千字の文字集合の利用が普通になる可能性がある。また、東京大学の漢字プロジェクト等、数万字の文字集合の提供を目指すプロジェクトもいくつか進められている。これだけ大規模な文字集合の利用が可能な環境でのコーパスのあり方を考えると、作成時の入力ミスや利用時の検索漏れを防ぐためには、文字の同定作業に今まで以上の精度が求められることになろう。ここで述べたような規準の明示化が不可欠のものとなる。

3.2　本文批判の問題

3.2.1　誤植･誤用と正用･通用

原文にある誤植・誤用等は、修訂してコーパスデータを作成することに、異論はなかろう。しかし、誤植・誤用であるのか正用・通用であるのかについて判断が難しい場合が少なくない。つまり、本文批判の問題である。例えば、次の（１）用字、（２）語法について、ａは誤植または誤用と認められ、ｃは正用あるいは通用と判断されるものであるが、ｂは判断が難しいものである。

（１）　ａ　裁培するを　　01-01･09187 A 09
　　ｂ　商家の内政に就き商家主人に注告すべき事項　01-01･02167 A 20
　　ｃ　假想的直段を以て賣り　　01-01･12184 B 25

（２）　ａ　政府も亦更迭しと雖も　　01-01･14224 A 09
　　ｂ　靜かに宇宙の眞理を考ふ時は　01-01･07206 B 09
　　ｃ　長太息するを禁じ能はざりき　　01-01･07041 A 23

ａの「裁培」や、「と雖も」の上接語が動詞連用形になる例は、『太陽』1901では一箇所にしか見られず、「栽培」「更迭すと雖も」の誤植または誤用と認めて問題ないだろう。

反対に、ｃの「ねだん」を「直段」と表記したり、「能はず」の上接語が動詞連用形になったりするのは、現代人の目から見ると奇異であるが、『日本国語大辞典』（小学館）では、こうした用字や語法を認めており、当時としては正しい用法であったことがわかる。実際、『太陽』1901には、「値段」４例に対して「直段」８例があり、このほか「高直」「安

直」という例もあって、当時は「直」という用字が一般的であった状況がうかがわれる。同じように、「能はず」に上接する動詞が連用形になる場合は 49 例見られ、連体形になる場合 768 例、連体形＋コトの場合 413 例に比べれば少数ではあるが、連用形＋「能はず」も通用していたことが確かめられる。これらは、『日本国語大辞典』等既存の辞書を規準に処理を行うことができる。

それでは、ｂはどうであろうか。「注告」という漢字語は『日本国語大辞典』には掲出しないが、字義通りの意味で文脈に整合するように考えられる。上記の「注告」の用例は、佐野善作（高等商業学校教授）・祖山鍾三（同講師）著の「商業世界」欄に見られるが、『太陽』1901 には、別の号にもう一箇所「注告」が使われており、それも同じ著者達による同じ欄の記事の中である。こうして見れば、「注告」は偶然の誤用とはいい難く、この用字のままとするのがよいと思われる。また、「考ふ」は、通常は下二段にしか活用せず（『日本国語大辞典』にも下二段のみ）、「考ふ時」という四段形は異例である。しかし、『太陽』1901 には、もう一例「考ふもの」という四段形の例が見え、単なる「る」の脱字とは認めにくい。同じような、二段活用と四段活用のゆれが見られる語に、「伺ふ」「挟む」（四段が正用）等が拾われる。とすれば、語法のゆれの範囲内と判断して、原文のままの形とするのが適当であると判断される。

このようにｂは、資料の内部徴証に照らした判断が求められるわけであるが、その際によるべき（よった）規準を明示することが、コーパス作成時においても利用時においても必要になる。

この種の問題は、（３）清濁においていっそう頻出する。

（３）ａ期し得可きか如し　01-01･01170 B 20

ｂ不適任者の採用を防かんとするの　01-01･03218 B 17

ｃ最も安全にして、又最も缺ぐべからざるものなるべし。　01-01･05038 B 02

ａのような格助詞または接続助詞「が」が期待される文脈で「か」が用いられている箇所は 200 箇所以上あり、ｂ「防く」も 10 例ある。ただし、一方に「防ぐ」が 120 例あり、助詞「が」もおそらく数千例規模の例があると思われることからすれば、清濁が異例となる比率は必ずしも高くなく、語形上の問題ではなく表記上（印刷上）のゆれではないかと思われる。一方、ｃは、「缺く」16 例に対して「缺ぐ」７例で、「缺ぐ」は決して例外ではない。これは表記（印刷）のゆれというよりも、語形のゆれを反映していると見てよいのではないか。「缺ぐ」と同じような様相を示す語に、「招ぐ」「若しくば」等をあげることができる。ａｂは修訂すべきもの、ｃは原文通りとすべきものと判断される。

3.2.2　仮名遣い

明治30年代ころまでは、歴史的仮名遣いは徹底して用いられていたわけではないという（築島1986）。『太陽』1901（明治 34）でも、語によって不統一が目立つものがある。

（４）ａ今日はどふしても八通關まで進まうと　01-01･13132 B 08

ｂ目新しく見へたのは　01-01･01136 B 15

ｃ大ひに注意すべき事件の在て　01-01･13221 A 14

ｄ或ひは立て居るもあれば、或ひは坐て居るのもあり　01-01･08095 B 14

下線部は、すべて歴史的仮名遣いに合致せず、原文の表記を修訂するか否かの判断が求められることになる。上記の４語について、歴史的仮名遣いの正用と、それに外れる例との

出現回数をまとめると次の通りである。左側が正用である。

a どう　148 例　　どふ　12 例　　b 見え　371 例　　見へ　80 例

c 大いに　12 例　　大ひに　7 例　　d 或いは　0 例　　或ひは　5 例

いずれの語においても正用に外れる例は少なくなく、a から d へとその比率は上昇していき、d では正用が 1 例も見つからない。これだけの振幅の大きさに対処できる現実的な処理は、すべて原文通りとするか、すべて正用に改めるかのいずれかであろう。入力ミスと検索漏れを防ぐことに配慮して、正用に統一することとしている。

3.3　著作権の問題

『太陽』1901 記事の著者の著作権について、没後50年を経過しているか否かでその状況を整理すると、次の通りである。

署名記事	著者没後 50 年以上経過（1998 年現在）		６７４本
	著者没後 50 年未満	2002 年までに没後 50 年	２３７本
		2003 年以後に没後 50 年	７１本
	著者没年等の情報不明		１１８本
無署名記事　（匿名記事・編集部記事含む）			５２４本
全記事			１６２４本

網掛け部分（200本弱）がコーパスを公開するには問題となる。この状況のままでは、全体の一割強が著作権をクリアできないことになる。没後 50 年を経過していない場合は、著作権者（継承者）に承諾を求めていきたい。また、署名されていてもその著者の没年が不明なものについては、調査文献や調査機関を広げることによって、解明に努力したい。

文献

京極興一 1993 『「国語」とは何か』（東宛社）改訂新版1996

国立国語研究所 1959 『明治初期の新聞の用語』国立国語研究所報告15

国立国語研究所 1963 『現代雑誌九十種の用語用字』国立国語研究所報告21

芝野耕司 1997『ＪＩＳ漢字字典』日本規格協会

進藤咲子 1967 「明治初期の新聞の用字」『ことばの研究』３

築島裕 1986 『歴史的仮名遣い　その成立と特徴』中公新書

永嶺重敏 1997 『雑誌と読者の近代』日本エディタースクール出版部

安永尚志 1998『国文学とコンピュータ』勉誠社

横山詔一・笹原宏之・野崎浩成・エリクロング 1998 『新聞電子メディアの漢字　朝日新聞ＣＤ－ＲＯＭによる漢字頻度表』国立国語研究所プロジェクト選書１、三省堂

国立国語研究所国語辞典編集準備室 1983 『用例採集のための主要雑誌目録』国語辞典編集準備資料３

国立国語研究所国語辞典編集室 1995 『スカウト式用例採集処理の手引き』国語辞典編集準備資料11

飯島満・大塚みさ 1998 「近代活字文献の電子テキスト化における字形の整理　ＪＩＳ包摂規準の活用と提案」語彙辞書研究会第14回研究発表会

木村睦子・田中牧郎・飯島満 1997 『『太陽』コーパスの作成と活用』文部省科学研究費「国際社会の日本語についての総合的研究」研究班４梶原チーム研究報告書