Japio

コーパスにみる特許関連文章の特徴

大学共同利用機関法人人間文化研究機構国立国語研究所
言語資源研究系准教授
山崎　誠

PROFILE
1957年生。1980年埼玉大学教養学部卒。1984年筑波大学大学院文藝言語研究科単位取得退学。現在、大学共同利用機関法人人間文化研究機構国立国語研究所言語資源研究系准教授。1984年より国立国語研究所で語彙調査、シソーラスの編纂、コーパスの構築に従事。2007年度から特許版産業日本語委員会。
yamazaki@ninjal.ac.jp　TEL

1 はじめに

国立国語研究所は2006年度から5年計画で『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。このプロジェクトの目標はこれからの日本語研究の共通基盤となる1億語の書き言葉コーパスを構築し、コーパス日本語学の確立を目指すことである。本稿は構築中のBCCWJを使って、特許関連の文章の特徴を計量言語学的に調査した結果である。

2 BCCWJの基本構成と特徴

BCCWJは図1に示すような3つのサブコーパス(SC)から構成される。生産実態SCと流通実態SCは、ランダムサンプリングを基本とするコーパスである。

生産実態（出版）SC	流通実態（図書館）SC
書籍、雑誌、新聞 2001年～2005年 約3500万語 固定長＋可変長	書籍 1986年～2005年 約3000万語 固定長＋可変長
非母集団（特定目的）SC	
白書、法律、国会会議録、ベストセラー、検定教科書、Webの掲示板、広報紙など 期間はまちまち（最長1976年～2005年） 約3500万語 可変長（一部、固定長＋可変長）	

図1　BCCWJの構成

BCCWJの特徴の一つに、均衡のとれた書籍のサンプルを大量に収録した点がある。従来コーパスとして使われてきた書籍のデータは、「新潮文庫の100冊」や「青空文庫」であったが、それぞれ問題がある。「新潮文庫の100冊」はそれぞれの作品の言語量がまちまちであり、検索結果を統計的に評価する際に工夫が必要である。また、ほとんどが文学であることから内容的な多様性に欠ける。「青空文庫」も文学が中心で、かつ、著作権が切れた作品が主であるため、現代語としてはやや古い作品が多くなっている。BCCWJでは、出版リストあるいは図書館の所蔵リストに基づき、書籍をランダムに選定しているため、多様なジャンルの書籍を収録している。また、対象となる年代は1986年～2005年であり、現代語のデータとして申し分ないものである。

3 特許に関する文章の特徴

BCCWJを利用して特許関連の文章の特徴を見てみよう。BCCWJにはさまざまなジャンル・内容のテキストが収録されている。特許関連の文章も例外ではない。ただし、BCCWJの書誌管理情報であるNDC（日本十進分類法）には「特許」という分類は存在しないため、それにいちばん近い分類を以下の手順で求めた。

BCCWJの一部である「書籍」のサンプルから「特許」「知的財産」という言葉を検索した。その結果以下のような出現状況が観察された（注2）。「特許」とい

う言葉は全体で 476 回、同じく「知的財産」は 86 回出現している。それを各サンプルが抽出された元の書籍の NDC 別に整理したのが表 1 である。

NDC	「特許」	「知的財産」
0. 総記	20	11
1. 哲学	1	0
2. 歴史	14	0
3. 社会科学	105	21
4. 自然科学	23	0
5. 技術，工学	282	54
6. 産業	8	0
7. 芸術，美術	4	0
8. 言語	0	0
9. 文学	18	0
（記載なし）	1	0
合計	476	86

表 1　「特許」「知的財産」の NDC 別出現状況

表 1 から、「特許」の使用例のうち約 59%、「知的財産」では約 63% が NDC の「5. 技術、工学」に集中していることが分かる。詳しく見てみると、「技術、工学」の中でも「507」（研究法、 指導法、 技術教育）がもっとも頻出する分類となっている。「特許」では、282 例中 234 例が、「知的財産」では 54 例全例が NDC「507」の分類をもつ書籍からのサンプルである。

そこで、NDC「507」を特許関連の文章とみなし、この特徴をさぐることにする。抽出された書籍を見ると『特許法概説』（吉藤幸朔著、有斐閣刊、1996）、『知的財産法』（田村善之著、有斐閣刊、1999）、『工業所有権法逐条解説』（特許庁編、発明協会刊、1986）などのタイトルが並び、抽出に成功していることが分かる。また、比較する対象として NDC「913」（日本文学の小説、物語）を同コーパスから選んだ。

4 比較の結果

今回調査したデータ中には NDC507 の文献は 42 タイトルあったため、比較対象とする NDC913 についても 42 タイトルを無作為に抽出した。抽出したタイトルの一部を挙げると、『理由』（宮部みゆき著、新潮社刊、2004）、『徳川家康』（山岡荘八著、講談社刊、1988）、『木曜島の夜会』（司馬遼太郎著、文芸春秋刊、1993）などである。

4.1　文の長さ

文の長さを正確にはかることは難しいため、形態素解析結果で出力される文境界の情報を利用することにした。そのため、文の長さは 1 つの文に含まれる形態素数で計測する。その結果を表 2 に示したが、特許関係の文章における文は、小説の 2 倍であることが分かる。特許の明細書は一文が長くなりがちであると言われるが、特許関連文章も一般の小説に比べると長いことが分かる。

分類	平均形態素数
NDC507（特許）	27.9
NDC913（小説）	14.4

表 2　文の長さの比較

4.2　品詞構成比

次に品詞構成比を見てみよう。品詞構成比はそれぞれ抽出した 42 タイトルを 1 つのデータとして扱い、延べ語数、異なり語数をベースに算出したものである。図 2 は、延べ語数、図 3 は異なり語数による品詞構成比の比較である（注 3）。

延べ語数と異なり語数とでは若干傾向が異なるが、概略特許関連の文章は小説と比べると名詞と接尾辞が多く、動詞、形容詞、副詞、助詞、助動詞が少ないことが

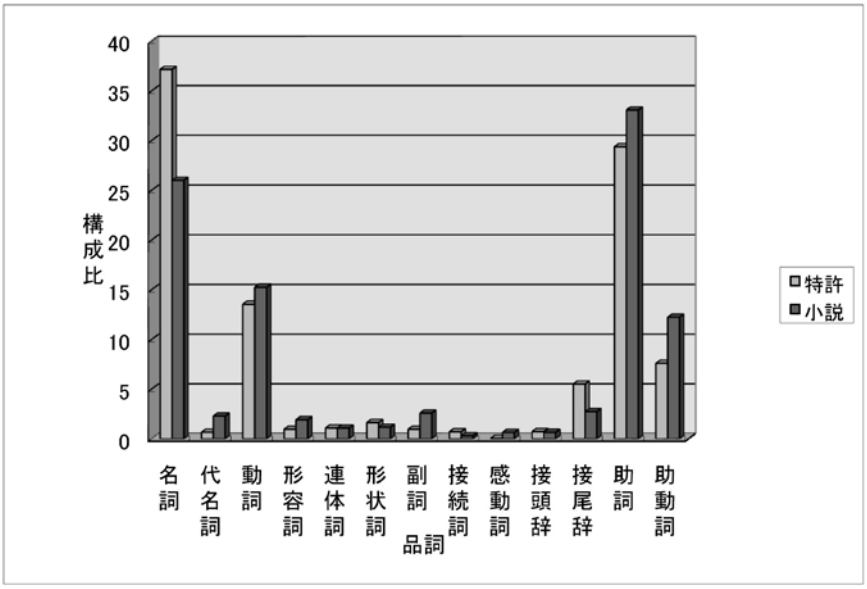

図２　品詞構成比の比較（延べ語数）

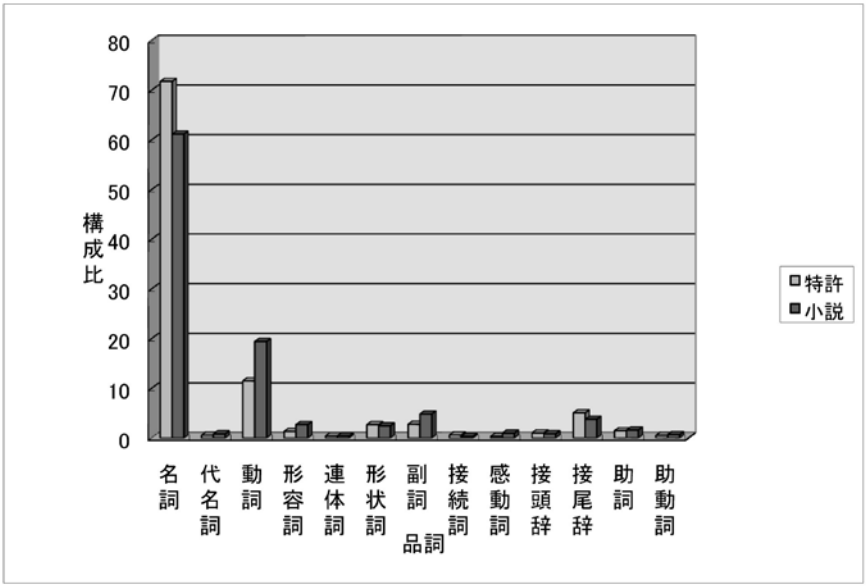

図３　品詞構成比の比較（異なり語数）

分かる。これはもっぱら情報量を担う名詞を多用することが特許関連の文章の特徴の一つであることを示唆していると言えよう。

4.3　語種構成比

語種とは、和語、漢語、外来語など語の出自の区別のことである。結果を表３に示す。特許関連の文章は漢語の比率が高く、とくに異なり語数では、和語の比率を大きく超えていることが分かる。

4.4　形容語句の比較

ここでは、副詞に占めるオノマトペの割合と形容詞に占めるシク活用の割合を調査した。どちらも口語的あるいは感情表現などと結び付く語句である。結果を表４

	延べ語数		異なり語数	
語種	特許	小説	特許	小説
和語	60.9	80.9	29.0	50.0
漢語	33.8	13.0	55.2	34.7
外来語	2.8	1.2	9.7	4.7
混種語	1.0	0.9	2.2	2.3
固有名	1.1	3.8	3.2	7.9
不明	0.1	0.1	0.2	0.3
記号	0.3	0.0	0.6	0.0

（数値は%）

表３　語種構成比の比較

に示す。表４によると、特許関連の文章では、オノマトペの割合は小説の 1/3 程度になっている。シク活用は、延べ語数では小説より少ないが、異なり語数ではほぼ同じ値となっている。具体的な語例で見ると、特許の文章で使われたシク活用の形容詞を使用度順に挙げると「新しい」「難しい」「美しい」「詳しい」「優しい」となり、一方、小説では、「新しい」「可笑しい」「怪しい」「嬉しい」「恐ろしい」となり、同じシク活用でも特許のほうがより客観的な意味を持つものが多いことが分かる。

	延べ語数		異なり語数	
形容語句	特許	小説	特許	小説
オノマトペ	4.5	17.5	10.6	34.8
シク活用	13.1	18.3	37.8	38.7

表３　語種構成比の比較

5 おわりに

本稿では特許関連の文章の特徴を計量的に概観した。比較対象として小説を選んだが、法律や他の専門的な文章とはどう異なるのか、今後の課題としたい。

注１　BCCWJ の概要については、山崎 (2009) を参照。

注２　使用したデータは、「BCCWJ 領域内公開データ（2009年度版）」における書籍（生産実態サブコーパス＋流通実態サブコーパス）の固定長サンプルである。このデータは特定領域「日本語コーパス」の内部で利用しているデータであり、「固定長サンプル」とは空白及び補助記号を除く 1,000 字からなるサンプルである。対象としたサンプル数は 18,558 で、その全体の語数は約 1100 万語と推測される。なお、ここでいう「語数」は短単位という形態素の 1 回結合までを許す言語単位である。

注３　品詞構成比の集計では、未知語・空白・補助記号・記号は除外した。

参考文献

山崎誠 (2009)「代表性を有する現代日本語書籍コーパスの構築」「人工知能学会誌」24-5、pp.623-631.