『日本語話し言葉コーパス』における品詞分布の分析

『日本語話し言葉コーパス』における品詞分布の分析

山口昌也
国立国語研究所研究開発部門

1 はじめに

品詞分布には，テキストの種類ごとに特徴があることが古くから指摘されている [1]。ここで言う品詞分布とは，あるまとまったテキストに含まれる語の品詞別割合のことを指す。本研究では，日本語話し言葉コーパス (以後，CSJ コーパス) の品詞分布を分析することにより，CSJ コーパスの言語的な位置づけを明らかにすることを目的とする。

品詞分布の分析は，大きく分けて次の二つの側面から行う。

- 書き言葉との比較を行い，大局的な側面から CSJ コーパスの位置づけを探る。本研究では，書き言葉のコーパスとして，京都大学テキストコーパス (以後，京大コーパス) を使う。
- CSJ コーパスを構成する講演を単位として，品詞分布を比較し，CSJ コーパス内の構成要素の品詞分布を明らかにする。

2 分析対象のデータ

本論文では，CSJ コーパスと京大コーパスを分析対象のデータとする。本節では，二つのコーパスの内容，および，規模について簡単に説明する。

2.1 CSJ コーパス

CSJ コーパスに含まれるデータのうち，本研究で分析対象とするのは，人手修正済みの形態論情報を付与された 359 講演である。これらは，学会講演 (内部の研究会での講演は除く)，模擬講演からなっている。CSJ コーパスにおいて付与されている形態論情報には，短単位と長単位があるが，本研究では，短単位を用いる。なお，CSJ コーパスに関する全体的な内容については前川 [2] を，形態論情報の詳細については小椋ら [3] を参照していただきたい。

分析対象とする講演数，総発話時間，延べ単位数，異なり単位数，講演ごとの延べ単位数の平均・標準偏差を表 1 に示す。

表 1 分析対象のデータ

	CSJ			京大
	学会講演	模擬講演	合計	社説
講演・記事数	141	218	359	609
総発話時間 (min)	2306	2381	4687	—
延べ単位数	445506	436409	881915	424396
異なり単位数	9769	13502	18655	18985
単位数 [平均]	3160	2001	2457	697
単位数 [標準偏差]	1887	488	1364	101

2.2 京都大学テキストコーパス

書き言葉のコーパスとして分析対象とする京大コーパス (ver.3.0) は，1995 年の毎日新聞に対して，形態素情報，および，係り受け情報を付与したコーパスである。京大コーパスには，社説 1 年分と全記事 1 カ月が収録されている。本研究では，書き言葉との比較という目的と，言語的な均質性の面を考慮し，社説 1 年分を分析対象とすることにした。

品詞情報については，CSJ コーパスと京大コーパスで体系が異なるため，両者を比較するには，品詞体系の変換が必要となる。粒度の点から見ると，京大コーパスのほうが CSJ コーパスよりも細かいので，京大コーパスの品詞情報 (大分類) を CSJ コーパスの品詞情報 (大分類) に合わせることにした。紙面の都合上，詳細は省くが，主として，京大コーパスのナ/ナノ/タル形容詞，指示詞に対し，品詞変換を行い，活用語尾に対しては再分割を行った。

分析対象とする記事数，延べ単位数，異なり単位数，各記事の延べ単位数の平均，標準偏差を表 1 に示す。これらの値は，すべて品詞体系修正後の値である。なお，句読点，括弧は，計測していない。

3 品詞分布の分析 (書き言葉との比較)

本節では，CSJ コーパスと京大コーパスの品詞分布を比較する。本節の目的は，CSJ コーパスを言語的に違いの大きい話し言葉と比較することにより，CSJ コーパスの特徴を大局的に捉えることである。

比較は，二つの方法で行う。一つは，コーパスを単位として品詞分布を比較する方法である。もう一つの方法は，書き言葉，話し言葉という分類をなくし，個々の講演，記事ごとに品詞分布を測定し，それらを主成分分析する方法である。この後の節では，上記二つの方法で分析した結果を述べることにする。

3.1 コーパス単位での比較

コーパス単位での比較では，品詞分布をコーパス単位でまとめた上で比較する。CSJ コーパス，京大コーパス，それぞれの品詞分布の測定結果を棒グラフにしたものを図 1 に示す。

まず，自立語について見てみると，話し言葉は書き言葉に比べて，名詞が少なく，感動詞，副詞，代名詞，接続詞が多いという結果となった。この結果は，樺島 [1] と「談話語の実態」[4] で得られている結果と大きな違いはない。少し詳しく見てみると，次のことが分かる。(1) 名詞の比率は，CSJ コーパスのほうが京大コーパスよりも 14.3% 小さい。(2) 感動詞，言いよどみは，CSJ コーパス全体の 7.9% にも及ぶ。

次に，樺島 [1] と「談話語の実態」[4] で大規模な調査が行われていない助詞，助動詞について見てみると，助動詞の比率の差

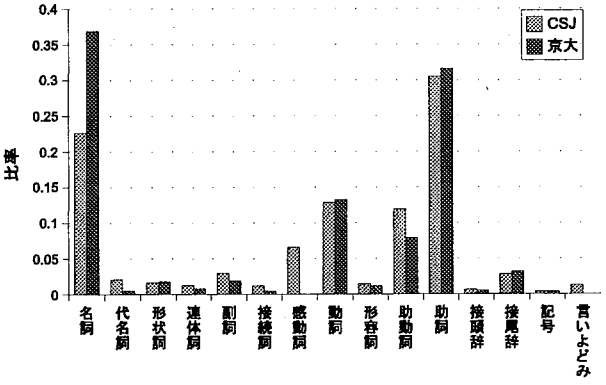

図 1　コーパス別の品詞分布

が顕著であり，CSJ コーパスのほうが 4.1% 大きくなっていることがわかる。一方，助詞については，両者とも全体の 30% 以上を占めるにもかかわらず，その差は 1.1% しかない。本来，名詞の比率の減少に伴い，格マーカである格助詞や係助詞の比率も少なくなるはずであるが，この結果ではそのようになっていない。これは，話し言葉では述語部分の助詞が豊富に存在する [4] ということを反映しているためだと考えられる。

3.2　講演，記事単位での比較

本節では，コーパス全体ではなく，講演，記事単位に品詞分布を比較する (なお，京大コーパスにほとんど現れることのない，感動詞，言いよどみは除外して分析した)。ここでは，各講演，記事の品詞分布に対して，主成分分析を行った。第 1 主成分，第 2 主成分をそれぞれ x，y 軸として，結果を図 2 に示す。第 1 主成分の寄与率は 0.86，因子負荷量は名詞 (0.90)，助動詞 (−0.34) であった。一方，第 2 主成分の寄与率は 0.05 であり，第 1 主成分でほとんどの説明が可能である。なお，第 2 主成分の因子負荷量は助詞 (0.79)，助動詞 (−0.53) であった。

第 1 主成分は，名詞と助動詞の因子負荷量が高いこと，また，図 2 より，正方向が書き言葉，負方向が話し言葉らしさを意味するものと推察できる。図 2 を見ると，社説は第 1，4 象限，CSJ コーパスのうち，模擬講演は第 2，3 象限に集まる。また，学会講演は y 軸を原点中心に約 45 度回転した軸上に集まる傾向にあり，第 1 主成分で見ると，模擬講演から社説と重複する，幅広い範囲に分布する。以上のことから，社説，学会講演，模擬講演の順で話し言葉らしくなっていることがわかる。

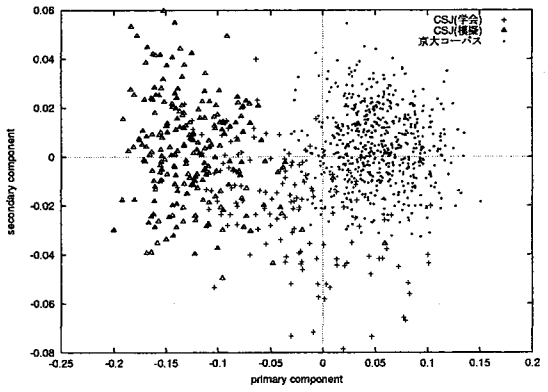

図 2　CSJ コーパスと京大コーパスの比較 (第 1，第 2 主成分)

4　品詞分布の分析 (CSJ コーパス内での比較)

ここでは，CSJ コーパスの構成要素の分析に焦点を当てるため，個々の講演の品詞分布を主成分分析する。第 1 主成分，第 2 主成分をそれぞれ x，y 軸として，結果を図 3 に示す。第 1 主成分の寄与率は 0.56，因子負荷量は名詞 (0.78)，格助詞 (0.34)，副助詞 (−0.26)，助動詞 (−0.25) であった。また，第 2 主成分の寄与率は 0.19，因子負荷量は感動詞 (0.92)，動詞 (−0.24)，格助詞 (−0.15)，副助詞 (−0.12)，接続助詞 (−0.12)，助動詞 (−0.12) であった。

第 1 主成分は，名詞，格助詞，助動詞の因子負荷量が高いことと図 2 の結果を考え合わせると，話し言葉／書き言葉らしさを表す成分だと考えられる。第 1 主成分上，学会講演と模擬講演は，図 3 の y 軸周辺で分割されていることがわかる。第 2 主成分の意味づけには，さらなる調査が必要だが，フィラーとして使われることが多い感動詞の因子負荷量が大きいこと，負値の因子負荷量をとる品詞が節や句を接続する要素 (「〜をし ... 〜して」のように) となりうることから，発話時に句，節をつなぐ表現を表す成分であると予想される。

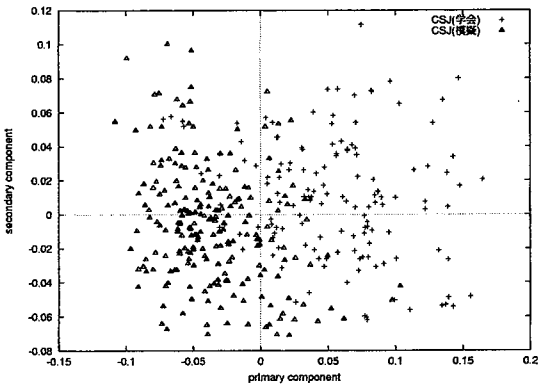

図 3　CSJ コーパス内での比較 (第 1，第 2 主成分)

5　おわりに

本研究では，CSJ コーパスの言語的な位置づけを明らかにするために，品詞分布に関して書き言葉との比較を行った。その結果，従来の研究結果を再確認するとともに，主成分分析により CSJ コーパスと京大コーパスとの関係を示した。また，CSJ コーパス内の講演の品詞分布を主成分分析し，話し言葉／書き言葉らしさの成分，発話時に句，節をつなぐ表現に関する成分を特徴として抽出した。

参考文献

[1] 樺島忠夫：現代文における品詞の比率とその増減の要因について，国語学 18，pp.15–20 (1954)

[2] 前川喜久雄:『日本語話し言葉コーパス』の設計と実装，平成 15 年度国立国語研究所公開研究発表会予稿集 (2003)

[3] 小椋秀樹ら：『日本語話し言葉コーパス』における形態論情報の設計，平成 15 年度国立国語研究所公開研究発表会予稿集 (2003)

[4] 国立国語研究所：談話語の実態 (1955)