# 『BCCWJ図書館サブコーパスの文体情報』を利用した語の文体差研究の可能性


| | |
|---|---|
| 著者 | 馬場 俊臣 |
| 雑誌名 | 言語資源活用ワークショップ発表論文集 |
| 巻 | 3 |
| ページ | 241-256 |
| 発行年 | 2018 |
| URL | http://doi.org/10.15084/00001658 |

# 『BCCWJ 図書館サブコーパスの文体情報』を利用した語の文体差研究の可能性

馬場 俊臣（北海道教育大学札幌校）

# The Possibility of Studies of Stylistic Features of Words using "Writing Style Annotation for the Library Subcorpus of the BCCWJ"

Toshiomi Baba (Hokkaido University of Education, Sapporo Campus)

**要旨**

本稿では、『BCCWJ 図書館サブコーパスの文体情報』のアノテーションデータを利用し、『現代日本語書き言葉均衡コーパス』の「図書館サブコーパス」内の語(語彙素)を対象として各語の文体差を数値化する試みを行い、この試みが文体研究において有効性・可能性があることを示す。まず、『文体情報』の専門度、硬度などの文体に関する 5 種類の指標別に、各語の平均値を算出する方法を示す。次に、各語の平均値については専門度、客観度、硬度、くだけ度の 4 指標は相互に強い相関があること、品詞別では感動詞と接続詞の 4 指標の平均値とそのばらつきは他の品詞に比べて特異であり品詞の特徴が表れていること、語種別では和語、漢語、外来語の特徴の違いが平均値の違いに表れていることなどの全体的傾向を示す。さらに、各語の平均値が、語の文体差に関する内省判断と強い相関があることを先行研究の調査結果と比較して示す。

## １．はじめに

本稿では、『BCCWJ 図書館サブコーパスの文体情報』[1](以下、『文体情報』)のアノテーションデータを利用して、語の文体差を数値化する試みを行い、文体研究[2]における有効性・可能性を検討する。

『文体情報』は、後述のように『現代日本語書き言葉均衡コーパス』[3](以下、BCCWJ)内の「図書館サブコーパス」の全サンプルに対する文体情報のアノテーションデータである。このアノテーションデータのうちの「専門度」「客観度」「硬度」「くだけ度」「語りかけ性度」(以下、「5 指標」)の値を利用して、「図書館サブコーパス」に出現する全語彙素[4](以下、「語」)の 5 指標のそれぞれの平均値を算出し、この平均値が各語の文体差を表すものと仮定して、その有効性・可能性を検討する。

語の文体差に関しては、「文体的特徴からする単語の分類は、連続的であり、程度の差によるものである。」(宮島 1977：873)と指摘されているように「連続的」である。ただし、実用的には「離散的」に段階差を設定して扱われることが多い[5]。本稿では、語の文体差を「連続的」な姿のまま捉えようとする試みでもある。

---

[1] 国立国語研究所(2015)。

[2] 本稿では、「文体」を、硬さ・柔らかさの違い、書き言葉・話し言葉の違いなどの「類型的な文体」に限定する。

[3] http://pj.ninjal.ac.jp/corpus_center/bccwj/ 参照。

[4] 後述のように、本稿では「長単位」の語彙素のみを対象とする。

[5] 井上(2009)、柏野(2016)など参照。

以下、2 節で、『文体情報』の概要を紹介するとともに、5 指標の平均値の算出方法を示す。その上で、本稿で検討対象とする語の範囲を示す。3 節で、5 種類の文体指標の平均値の全体的傾向・特徴を、5 指標の相互の相関、品詞別の特徴、語種別の特徴の観点から検討する。4 節で、各語の文体指標の平均値がどの程度、語の文体差に関する内省判断と一致するかを先行研究の調査結果を利用しながら検討する。5 節で、全体のまとめと今後の展望を述べる。

## ２．各語の 5 指標の平均値

### ２.１　『BCCWJ 図書館サブコーパスの文体情報』について

まず、『BCCWJ 図書館サブコーパスの文体情報』の概要を紹介する[6]。

『文体情報』は、BCCWJ に収録されている「図書館サブコーパス」の書籍サンプル(10,551 サンプル)を対象として、「内容・表現の文体的特徴を表す分類指標」及び「形式・内容・表現に文体判断が単純にいかない特徴をもつものの分類指標」を付与したデータである。「内容・表現の文体的特徴を表す分類指標」には、「対象読者に想定される読解レベル(難易度)」に関わる「専門度」、「テキストの作成意図」に関わる「客観度」、「さまざまな文体情報」に関わる「硬度」「くだけ度」「語りかけ性度」の 5 種類の文体的特徴を表す分類指標(5 指標)が設けられている。なお、「さまざまな文体情報」のうち、「硬度」「くだけ度」は「形式性、親疎性を問う」指標であり、「語りかけ性度」は「口語性を問う」指標である。これらの 5 指標は、それぞれ「言語データ構築経験有のおおよそ 20～50 代の女性、延べ 9 名。」の作業者によって付与され、それぞれ次の 3 段階～5 段階のいずれかの段階が付与されている。

(a)専門度　1 専門家向き、2 やや専門的な一般向き、3 一般向き、4 中高生向き、5 小学生・幼児向き

(b)客観度　1 とても客観的、2 どちらかといえば客観的、3 どちらかといえば主観的、4 とても主観的

(c)硬度　1 とても硬い、2 どちらかといえば硬い、3 どちらかといえば軟らかい、4 とても軟らかい

(d)くだけ度　1 とてもくだけている、2 どちらかといえばくだけている、3 くだけていない

(e)語りかけ性度　1 とても語りかけ性がある、2 どちらかといえば語りかけ性がある、3 特に語りかけ性はない

5 指標の平均値の算出に当たっては、この各指標の段階の番号を数値として扱った。

なお、5 指標の付与対象となったサンプル数は、「図書館サブコーパス」全 10,551 サンプルのうちの、8,887 サンプル[7]である。

---

[6] 『文体情報』に関する説明は、国立国語研究所(2015)の添付文書「概要」及び柏野(2013)に基づく。

[7] 柏野(2013)には 8,887 サンプルと記載されているが、『文体情報』内のデータでは、専門度、硬度、くだけ度、語りかけ性度が付与されているのはそれぞれ 8,821 サンプル、客観度が付与されているのは 5,901 サンプルである。

## ２.２　平均値の算出方法

『文体情報』の対象である「図書館サブコーパス」内の長単位[8]の全語彙素(語)を対象として、それぞれの語が用いられている全サンプルの専門度、客観度、硬度、くだけ度、語りかけ性度別に平均値を求め、その値を、その語の「専門度平均値」「客観度平均値」「硬度平均値」「くだけ度平均値」「語りかけ性度平均値」とする。

ただし、ある語が同一のサンプルに 2 回以上用いられている場合、そのまま平均値を求めるとそのサンプルの指標の値の影響が相対的に強く出てしまうため、同一のサンプルに 2 回以上用いられている場合は 1 回用いられているものとして平均値を求めた[9]。

実際の算出作業は、データベース管理システム MySql を用い、次の手順で行った。

① 下準備として、『文体情報』アノテーションデータ[10]の「専門度」「客観度」「硬度」「くだけ度」「語りかけ性度」の各列を指標の段階の数値だけに置き替えた[11]アノテーションデータを作成した。

② BCCWJ の DVD 版公開データ(BCCWJ-DVD 版 Version 1.1)の長単位データ[12]及び①で作成した修正版『文体情報』アノテーションデータのそれぞれをインポートしたテーブルを作成した。

③ BCCWJ 長単位データの各行(レコード)(用いられているすべての語彙素)を、サンプル ID[13]の一致する『文体情報』アノテーションデータと結合したテーブル(「延べサンプル方式テーブル」)を作成した。行(レコード)数は 30,273,796 行[14]である。

④ 異なり語数を確認するために、この「延べサンプル方式テーブル」に「語彙素」「語彙素読み」「品詞」「語種」[15]を結合した列(「品詞語彙素等」)を挿入したテーブル(「延べ語テーブル」)を作成し、この「延べ語テーブル」から「品詞語彙素等」が同一の

---

8 「長単位」は「複合語を把握する」ことができ「サンプルの言語的特徴の解明に適した」単位である(国立国語研究所コーパス開発センター2015)とされている。

9 接続詞のみを対象として「硬度」「くだけ度」の平均値を扱った馬場(2018)では、ある語が同一のサンプルに 2 回以上用いられている場合そのまま平均値を求める方式を「延べサンプル方式」、2 回以上用いられている場合でも 1 回用いられているものとして平均値を求める方式を「異なりサンプル方式」と呼んでいる。「異なりサンプル方式」である本稿の算出方法を具体例で示しておく。仮に「御昼」という語が、サンプル A(専門度 1、硬度 1)(使用頻度 3 回)、サンプル B(専門度 3、硬度 3)(使用頻度 1 回)、サンプル C(専門度 3、硬度 2)(使用頻度 2 回)の 3 サンプルで用いられていれば、専門度平均値は「(1+3+3)÷3＝2.3333」、硬度平均値は「(1+3+2)÷3＝2.0000」となる。

10 LB_all.csv。

11 元データは、例えば「3 一般向き」のように指標の段階及び選択肢表現が入っている。これを「3」のように数値のみに置き換えた。この作業は Excel を用いて行った。

12 Disk2(NumTrans 版)の「TSV_LUW_NT」(長単位データ)の LB.zip(「図書館サブコーパス」)のデータを用いた。

13 『文体情報』アノテーションデータの列名は「SampleID」である。

14 「『現代日本語書き言葉均衡コーパス』語彙表 ver.1.1 解説」及び「BCCWJ 品詞構成表(Version 1.1)」では、長単位の延べ語数は 25,031,768 語である。「品詞」フィールドに「URL、カタカナ文、方言、未知語、漢文、空白、英単語、補助記号、言いよどみ、記号」のタグが付けられている(本来の「品詞」以外の)計 5,242,027 語は、この語彙表の長単位の延べ語数からは除かれている。「BCCWJ 品詞構成表(Version 1.1)」と比べると本稿のデータは「名詞」が 3 語多い。なお、語数に 2 語の差がある理由については不明である。

15 「『現代日本語書き言葉均衡コーパス』語彙表 ver.1.1 解説」及び「BCCWJ 品詞構成表(Version 1.1)」では、長単位は「語彙素」「語彙素読み」「品詞」「語種」の 4 つの組を用いて同一の見出し語を特定している。これに倣った。

行(レコード)の重複を削除したテーブルを作成した。行(レコード)数は 821,510 語[16]である。

⑤　④の「延べ語テーブル」から、「サンプル ID」及び「品詞語彙素等」が同一の行(レコード)の重複を削除したテーブル(「異なりサンプル方式テーブル」)を作成した。行(レコード)数は 7,600,397 行である。

⑥　この「異なりサンプル方式テーブル」に基づき、(本稿での)すべての異なり語 821,510 語それぞれの専門度平均値、客観度平均値、硬度平均値、くだけ度平均値、語りかけ性度平均値及びそれぞれの標準偏差の一覧表を作成した。

表 1 に、語別の 5 指標ごとの平均値のリストの一部を示す。

表 1　語別の 5 指標ごとの平均値のリスト(一部)

| 語彙素 | 品詞 | 専門度頻度 | 専門度平均値 | 専門度標準偏差 | 客観度頻度 | 客観度平均値 | 客観度標準偏差 | 硬度頻度 | 硬度平均値 | 硬度標準偏差 | くだけ度頻度 | くだけ度平均値 | くだけ度標準偏差 | 語りかけ性度頻度 | 語りかけ性度平均値 | 語りかけ性度標準偏差 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| だ | 助動詞 | 8821 | 2.9747 | 0.5909 | 5901 | 2.3965 | 0.9239 | 8821 | 2.5912 | 0.7349 | 8821 | 2.5871 | 0.5913 | 8821 | 2.6548 | 0.6442 |
| は | 助詞-係助 | 8821 | 2.9747 | 0.5909 | 5901 | 2.3965 | 0.9239 | 8821 | 2.5912 | 0.7349 | 8821 | 2.5871 | 0.5913 | 8821 | 2.6548 | 0.6442 |
| が | 助詞-格助 | 8821 | 2.9747 | 0.5909 | 5901 | 2.3965 | 0.9239 | 8821 | 2.5912 | 0.7349 | 8821 | 2.5871 | 0.5913 | 8821 | 2.6548 | 0.6442 |
| に | 助詞-格助 | 8821 | 2.9747 | 0.5909 | 5901 | 2.3965 | 0.9239 | 8821 | 2.5912 | 0.7349 | 8821 | 2.5871 | 0.5913 | 8821 | 2.6548 | 0.6442 |
| の | 助詞-格助 | 8821 | 2.9747 | 0.5909 | 5901 | 2.3965 | 0.9239 | 8821 | 2.5912 | 0.7349 | 8821 | 2.5871 | 0.5913 | 8821 | 2.6548 | 0.6442 |

さて、この「語別の 5 指標ごとの平均値のリスト」の異なり語(行数)は 821,510 語であるが、「品詞」フィールドに「URL、カタカナ文、方言、未知語、漢文、空白、英単語、補助記号、言いよどみ、記号」のタグが付けられている語が 492 語含まれている。これらは本来の「語」ではないため、これを除いた 821,018 語が「図書館サブコーパス」の異なり語数となる。さらに、この 821,018 語のうち 101,209 語は 5 指標の付与対象サンプルに出現していない。この 101,209 語を除く 719,809 語が 5 指標の付与対象サンプルに出現している異なり語である。

この 719,809 語のうち、本稿では、専門度、硬度、くだけ度、語りかけ性度の 4 指標の付与対象サンプル数が 100 以上[17]の 7,877 語をこれ以降の分析対象とする。

参考までに、表 2 にこの 7,877 語の品詞別語数を示す。

表 2　本稿で分析対象とする 7,877 語の品詞別語数

| 品詞 | 語数 | 品詞 | 語数 | 品詞 | 語数 |
|---|---|---|---|---|---|
| 名詞 | 3,967 | 形容詞 | 194 | 感動詞 | 59 |
| 動詞 | 2,294 | 助詞 | 137 | 接続詞 | 48 |
| 形状詞 | 524 | 助動詞 | 86 | 連体詞 | 29 |
| 副詞 | 456 | 代名詞 | 79 | 接尾辞 | 4 |

16 「『現代日本語書き言葉均衡コーパス』語彙表 ver.1.1 解説」及び「BCCWJ 品詞構成表(Version 1.1)」では、長単位の異なり語数は 821,025 語である。「品詞」フィールドに「URL、カタカナ文、方言、未知語、漢文、空白、英単語、補助記号、言いよどみ、記号」のタグが付けられている(本来の「品詞」以外の)計 492 語は、この語彙表の長単位の異なり語数からは除かれている。「BCCWJ 品詞構成表(Version 1.1)」と比べると本稿のデータは「名詞」が 7 語少ない。

17 専門度、硬度、くだけ度、語りかけ性度の 4 指標が付与されているサンプルは一致するため、この 4 指標の付与対象サンプル数はどの語も同じである。しかし、この 4 指標が付与されていても客観度が付与されていないサンプルがあるため、客観度の付与対象サンプル数は若干少なくなる。客観度の付与対象サンプル数が 100 未満の語も、以下の分析対象に含まれている。なお、「100 以上」としたのはある程度の大きさのサンプル数を確保するためであり、特に理論的根拠はない。

## ３．5 指標の平均値の全体的傾向・特徴

### ３.１　5 指標の相互の相関

5 指標の平均値の全体的傾向・特徴を見るために、5 指標の各平均値の相互の相関、品詞別の特徴、語種別の特徴を分析する。

まず、語別の 5 指標の各平均値の相互の相関分析を行う。

図 1 は、語別の 5 指標ごとの平均値を二組ずつセットにし、それぞれの散布図、相関係数、無相関検定結果[18]を示したものである。

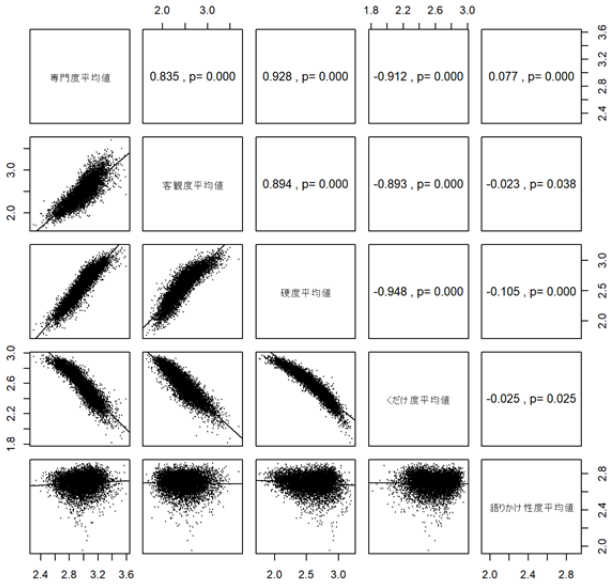


図 1　語別 5 指標各平均値相互の散布図、相関係数、無相関検定結果

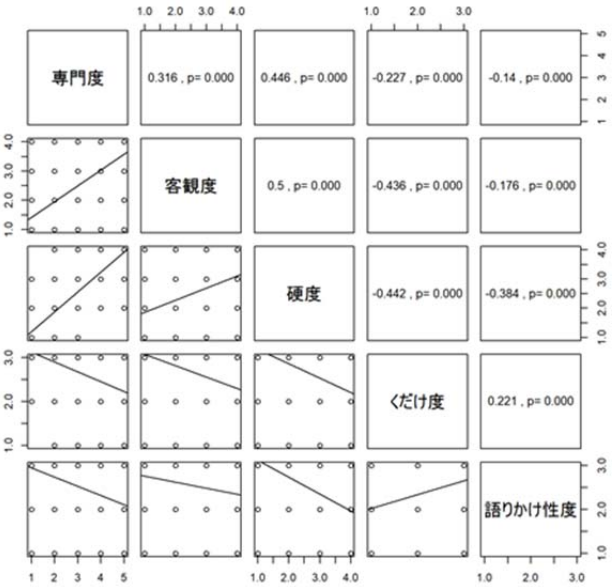


図 2　『文体情報』5 指標相互の散布図、相関係数、無相関検定結果

[18] 統計 R(ver.3.4.3) cor.test 使用。ピアソンの積率相関係数を求めた。相関係数、無相関検定結果は、小数点第 4 位以下四捨五入。

図 1 のとおり、専門度、客観度、硬度、くだけ度の 4 指標の平均値は相互に強い相関があるのに対し、これら 4 指標と語りかけ性度とはいずれも相関がない。

『文体情報』では、もともと 5 指標には相関があるのであろうか。5 指標がともに付与されている 5,901 サンプルの 5 指標の段階をそのまま用いて相関分析を行った。図 2 は、『文体情報』の 5 指標を二組ずつセットにし、それぞれの散布図、相関係数、無相関検定結果[19]を示したものである。図 2 のとおり、専門度、客観度、硬度、くだけ度の 4 指標は相互に中程度の相関ないし弱い相関があるのに対し、語りかけ性度は専門度・客観度とは相関はなく、硬度・くだけ度とは弱い相関がある。

このように、もともと『文体情報』の専門度、客観度、硬度、くだけ度の 4 指標は相互にある程度の相関はあるが、ある語がどのような「内容・表現の文体的特徴」を持つサンプルに使われやすいかという、語別の 5 指標の平均値を比べた本稿の分析方法では、専門度、客観度、硬度、くだけ度の 4 指標で表される文体的特徴は共通性が高いことが示されていると考えられる。

### ３.２　品詞別の特徴

本節では、品詞別に 5 指標の平均値の特徴を分析する。

図 3 は、品詞別に、5 指標ごとの全体の平均値を図示したものである。

図 4 は、品詞別に、硬度のみの語別平均値の分布を示した箱ひげ図である。

図 5 は、品詞別に、5 指標ごとの語別平均値の標準偏差を図示したものである。

図 3 を見ると、専門度、客観度、硬度、くだけ度の 4 指標で、品詞によって若干の傾向の違いがあることが分かる。特に、感動詞は他の品詞に比べて専門度、客観度、硬度の全体平均値が高く、くだけ度の全体平均値が低くなっており[20]、異なった傾向にあることが分かる[21]。感動詞は会話で使われやすいということが表されているものと思われる。図 4 は語別の硬度平均値の分布のみを、品詞別に示したものであるが、感動詞は他の品詞と分布が異なっている。図は示さないが、客観度、硬度、くだけ度についても同様である。

次に、図 5 を検討するが、接尾辞は対象とする語が 4 語で少ないためここでの検討対象からは除く。図 5 のとおり、専門度、客観度、硬度、くだけ度の 4 指標で、接続詞の語別平均値の標準偏差が他の品詞に比べて最も大きい。これら 4 指標で語別平均値のばらつきが大きいということであり、接続詞は語の違いによる文体差が他の品詞に比べて大きいこ

---

[19] 統計 R(ver.3.4.3) cor.test 使用。段階の値をそのまま用いスピアマンの順位相関係数を求めた。なお、ピアソンの積率相関係数であっても傾向は同じである。相関係数、無相関検定結果は、小数点第 4 位以下四捨五入。

[20] 専門度、客観度、硬度は、それぞれより専門的、客観的、硬いほど数値が低くなる。くだけ度は、よりくだけているほど数値が低くなる。数値が逆方向になることに注意が必要である。

[21] 5 指標のそれぞれについて、品詞を群とする等分散性の検定(Bartlett 検定)を行った結果、5 指標すべてにおいて $p < 0.001$ で各群の分散が等しくないと判断された(専門度 $\chi^2 = 75.922$、客観度 $\chi^2 = 42.742$、硬度 $\chi^2 = 84.655$、くだけ度 $\chi^2 = 51.265$、語りかけ性度 $\chi^2 = 121.45$)。そのため、5 指標のそれぞれについて、品詞を群とする Kruskal-Wallis 検定を行った。その結果、5 指標すべてにおいて群の効果は $p < 0.001$ で有意であった(専門度 $\chi^2 = 458.61$、客観度 $\chi^2 = 432.9$、硬度 $\chi^2 = 430.99$、くだけ度 $\chi^2 = 481.02$、語りかけ性度 $\chi^2 = 131.69$)。多重比較(Steel-Dwas 法)を行った結果、専門度、客観度、硬度、くだけ度の 4 指標に関しては、感動詞と(接尾辞を除く)他の各品詞との間で $p < 0.001$ で有意差があった。なお、感動詞の次に専門度等の全体平均値が高い代名詞は、(接尾辞を除くと)客観度(代名詞と副詞との組み合わせ)、硬度(代名詞と接続詞との組み合わせ)で $p < 0.05$ で有意差のない組み合わせがあった。

とが表されている。一方、この 4 指標で、感動詞が他の品詞に比べて標準偏差が最も小さい。感動詞は、(図 3、図 4 の結果と合わせると)専門度、客観度、硬度の値が高いテキスト(サンプル)、くだけ度の値が低いテキスト(サンプル)に集中して使われており、語の違いによる文体差が他の品詞に比べて小さいことが表されている。

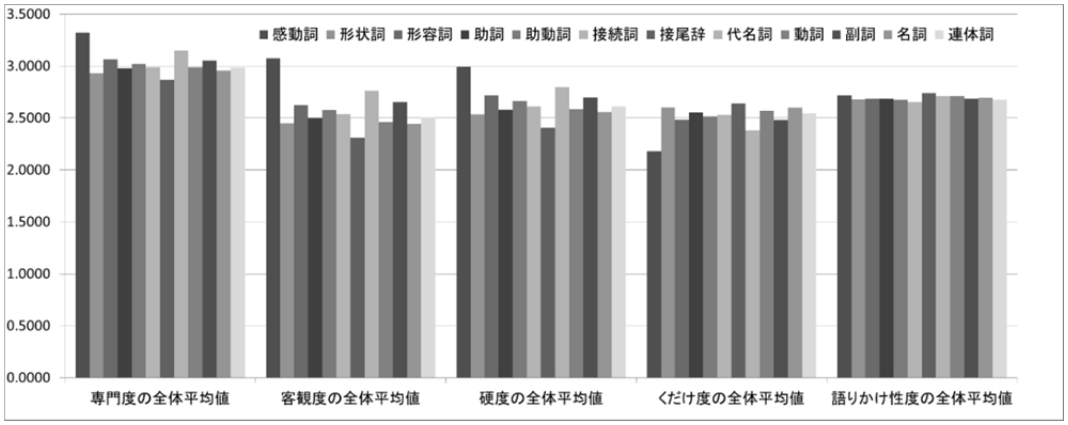


図 3　品詞別の 5 指標ごとの全体平均値

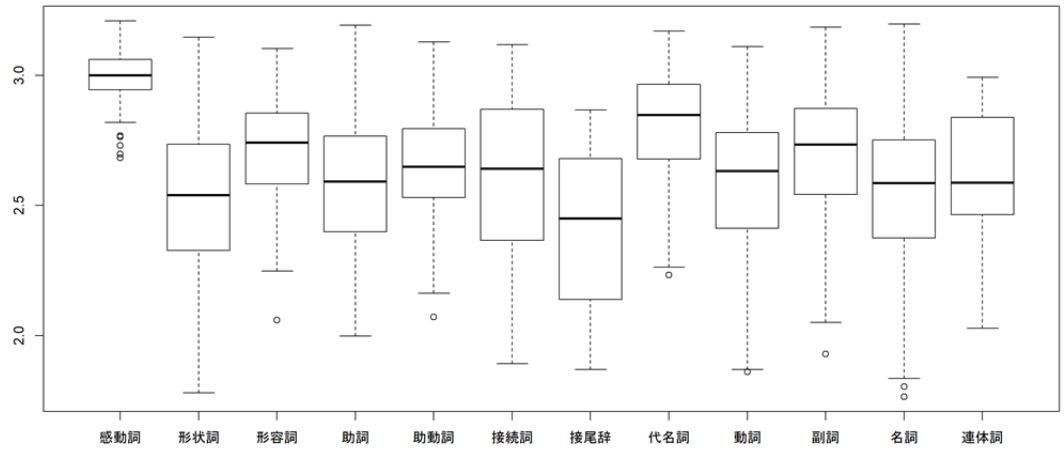


図 4　品詞別の語別硬度平均値の分布を示す箱ひげ図

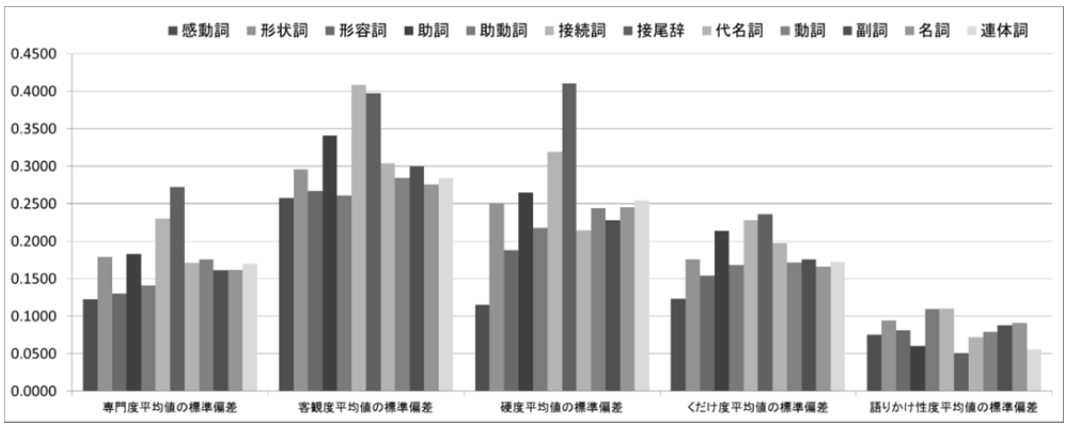


図 5　品詞別の 5 指標ごとの語別平均値の標準偏差

### ３.３　語種別の特徴

本節では、語種別に 5 指標の平均値の特徴を分析する。

図 6 は、語種別に、5 指標ごとの全体の平均値を図示したものである。

図 7 は、語種別に、硬度のみの語別平均値の分布を示した箱ひげ図である。

図 8 は、語種別に、5 指標ごとの語別平均値の標準偏差を図示したものである。

和語、漢語、外来語の 3 種類に絞って分析していく。

図 6 を見ると、専門度、客観度、硬度、くだけ度の 4 指標で、語種によって傾向の違いがあることが分かる。和語と外来語は、専門度、客観度、硬度の全体平均値が高く、くだけ度の全体平均値が低くなっている[22]。和語と外来語は漢語に比べて、硬くないくだけた文体のテキストで使われやすいということが表されている。図 7 は語別の硬度平均値の分布のみを語種別に示したものであるが、和語と外来語は漢語よりも硬度平均値の分布が全体的に高くなっている。図は示さないが、専門度、客観度、くだけ度についても分布の傾向は同じである。

次に、図 8 であるが、和語、漢語、外来語について、専門度、客観度、硬度、くだけ度の 4 指標に共通した特徴を指摘することはできない。専門度、客観度、硬度の 3 指標に限れば、漢語の語別平均値の標準偏差が最も大きい。すなわち、語別平均値のばらつきが大きいということであり、漢語は和語や外来語に比べて文体差が大きい傾向にあるということが表されていると見られる。

参考として、表 3 に、和語、漢語、外来語の硬度平均値の上位と下位の各 20 語(昇順)を示す。

[22] 5 指標のそれぞれについて、語種を群とする等分散性の検定(Bartlett 検定)を行った結果、5 指標すべてにおいて $p < 0.001$ で各群の分散が等しくないと判断された(専門度 $\chi^2 = 123.89$、客観度 $\chi^2 = 67.549$、硬度 $\chi^2 = 231.48$、くだけ度 $\chi^2 = 31.422$、語りかけ性度 $\chi^2 = 89.552$)。そのため、5 指標のそれぞれについて、語種を群とする Kruskal-Wallis 検定を行った。その結果、5 指標すべてにおいて群の効果は $p < 0.001$ で有意であった(専門度 $\chi^2 = 1846.9$、客観度 $\chi^2 = 1238.7$、硬度 $\chi^2 = 1622.3$、くだけ度 $\chi^2 = 1611.8$、語りかけ性度 $\chi^2 = 110.08$)。多重比較(Steel-Dwas 法)を行った結果、専門度、客観度、硬度、くだけ度、語りかけ性度の 5 指標で、和語と漢語との間及び外来語と漢語の間で $p < 0.001$ で有意差があった。また、客観度、硬度、くだけ度の 3 指標で、和語と外来語の間で $p < 0.05$ で有意差がなかった。全体として、和語と漢語との間及び外来語と漢語との間で有意差があり、和語と外来語との間では有意差がないという傾向が見られた。

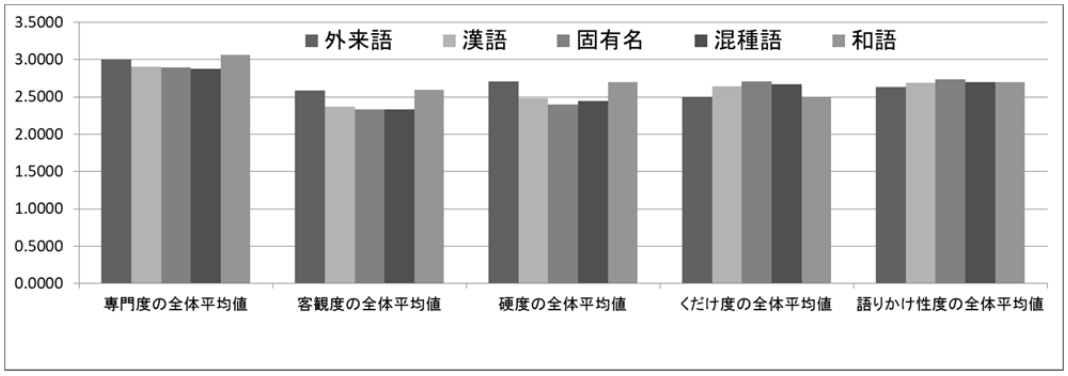


図 6　語種別の 5 指標ごとの全体平均値

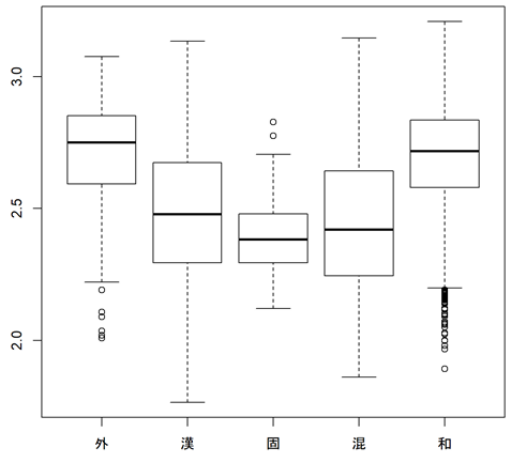


図 7　語種別の語別硬度平均値の分布を示す箱ひげ図

(外:外来語、漢:漢語、固:固有名、混:混種語、和:和語)

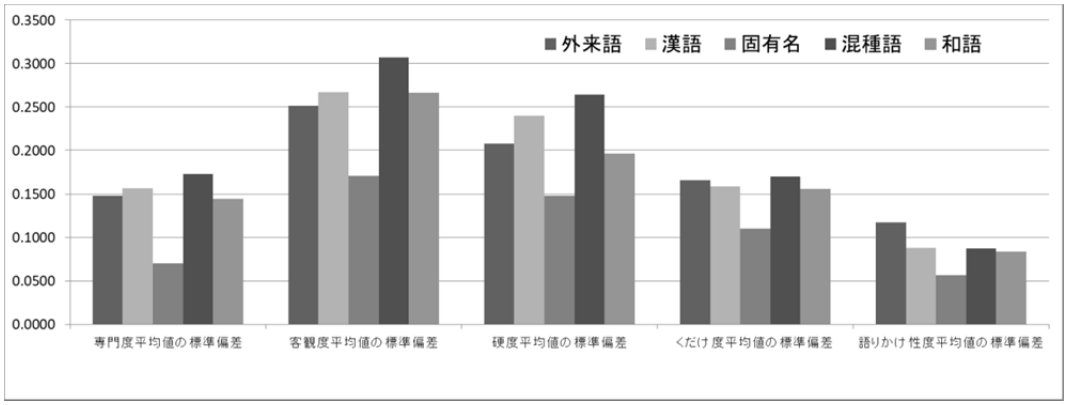


図 8　語種別の 5 指標ごとの語別平均値の標準偏差

表 3　和語、漢語、外来語の硬度平均値の上位と下位の各 20 語(昇順)

| 語彙素 | 語種 | 硬度平均値 | 語彙素 | 語種 | 硬度平均値 | 語彙素 | 語種 | 硬度平均値 |
|---|---|---|---|---|---|---|---|---|
| 並びに | 和語 | 1.8929 | 事項 | 漢語 | 1.7658 | オブ | 外来語 | 2.0088 |
| 見直し | 和語 | 1.9667 | 広範 | 漢語 | 1.7801 | イン | 外来語 | 2.0186 |
| 枠組み | 和語 | 1.9807 | 要件 | 漢語 | 1.8033 | アンド | 外来語 | 2.0351 |
| とともに | 和語 | 1.9980 | 増大 | 漢語 | 1.8358 | イデオロギー | 外来語 | 2.0893 |
| 定め | 和語 | 2.0000 | 形成 | 漢語 | 1.8462 | デ | 外来語 | 2.1081 |
| 及び | 和語 | 2.0244 | 適用 | 漢語 | 1.8492 | ザ | 外来語 | 2.1912 |
| のみならず | 和語 | 2.0279 | 成立 | 漢語 | 1.8561 | メカニズム | 外来語 | 2.2208 |
| 基づく | 和語 | 2.0282 | 規定 | 漢語 | 1.8600 | ニーズ | 外来語 | 2.2214 |
| 其れ故 | 和語 | 2.0502 | 等(接尾辞) | 漢語 | 1.8703 | アプローチ | 外来語 | 2.2276 |
| 著しい | 和語 | 2.0591 | 条約 | 漢語 | 1.8807 | コスト | 外来語 | 2.2423 |
| をめぐる | 和語 | 2.0635 | 紛争 | 漢語 | 1.8889 | プロセス | 外来語 | 2.2576 |
| 異 | 和語 | 2.0650 | 上記 | 漢語 | 1.8987 | スローガン | 外来語 | 2.2700 |
| こととなる | 和語 | 2.0714 | 中核 | 漢語 | 1.8992 | ネットワーク | 外来語 | 2.3333 |
| 基 | 和語 | 2.0909 | 利害 | 漢語 | 1.9000 | プロジェクト | 外来語 | 2.3333 |
| にわたり | 和語 | 2.0964 | 移行 | 漢語 | 1.9027 | メディア | 外来語 | 2.3352 |
| 担い手 | 和語 | 2.0991 | 実施 | 漢語 | 1.9034 | システム | 外来語 | 2.3354 |
| における | 和語 | 2.1022 | 体系 | 漢語 | 1.9118 | ピーク | 外来語 | 2.3643 |
| 押し進める | 和語 | 2.1159 | 輸出 | 漢語 | 1.9252 | キーワード | 外来語 | 2.3661 |
| 欠く | 和語 | 2.1167 | 従来 | 漢語 | 1.9290 | シナリオ | 外来語 | 2.3679 |
| 盛り込む | 和語 | 2.1168 | 契機 | 漢語 | 1.9305 | モデル | 外来語 | 2.3681 |
| : | : | : | : | : | : | : | : | : |
| で(助詞) | 和語 | 3.1193 | 元気(名詞) | 漢語 | 2.9588 | バッグ | 外来語 | 2.9510 |
| すっ(副詞) | 和語 | 3.1250 | 魔法 | 漢語 | 2.9611 | トイレ | 外来語 | 2.9518 |
| 御風呂(名詞) | 和語 | 3.1262 | 洗濯 | 漢語 | 2.9649 | コップ | 外来語 | 2.9560 |
| 御昼(名詞) | 和語 | 3.1275 | 暢気 | 漢語 | 2.9690 | チーズ | 外来語 | 2.9619 |
| ちゃう(助動詞) | 和語 | 3.1284 | 元気(形状詞) | 漢語 | 2.9778 | ポケット | 外来語 | 2.9724 |
| ふん(感動詞) | 和語 | 3.1341 | 結構 | 漢語 | 2.9799 | キッチン | 外来語 | 2.9737 |
| ふうん(感動詞) | 和語 | 3.1390 | 頂戴 | 漢語 | 2.9870 | ベランダ | 外来語 | 2.9741 |
| ずーっと(副詞) | 和語 | 3.1397 | 二匹 | 漢語 | 2.9902 | カップル | 外来語 | 2.9758 |
| すうっと(副詞) | 和語 | 3.1441 | 去年 | 漢語 | 3.0000 | ドレス | 外来語 | 2.9778 |
| こら(感動詞) | 和語 | 3.1463 | 一杯(副詞) | 漢語 | 3.0034 | オーケー | 外来語 | 2.9802 |
| や(形状詞) | 和語 | 3.1471 | 御主人 | 漢語 | 3.0044 | デート | 外来語 | 3.0000 |
| ううん(感動詞) | 和語 | 3.1572 | 変 | 漢語 | 3.0061 | クリスマス | 外来語 | 3.0075 |
| お(感動詞) | 和語 | 3.1681 | 一生懸命(形状詞) | 漢語 | 3.0080 | スカート | 外来語 | 3.0110 |
| 私達(代名詞) | 和語 | 3.1707 | 餓鬼 | 漢語 | 3.0254 | ケーキ | 外来語 | 3.0179 |
| とっても(副詞) | 和語 | 3.1741 | 御免 | 漢語 | 3.0455 | バケツ | 外来語 | 3.0190 |
| じゃん(助詞) | 和語 | 3.1818 | 御飯 | 漢語 | 3.0633 | ピンク | 外来語 | 3.0361 |
| 思いっ切り(副詞) | 和語 | 3.1852 | 内緒 | 漢語 | 3.0672 | プレゼント | 外来語 | 3.0368 |
| ねん(助詞) | 和語 | 3.1923 | 一生懸命(副詞) | 漢語 | 3.0683 | ママ | 外来語 | 3.0382 |
| 御ばあちゃん(名詞) | 和語 | 3.1972 | 本当 | 漢語 | 3.1148 | キス | 外来語 | 3.0435 |
| ん(感動詞) | 和語 | 3.2089 | 一杯(名詞) | 漢語 | 3.1346 | パパ | 外来語 | 3.0765 |

## ４．内省判断に基づく語の文体差との比較

### ４.１　柏野(2016)の文体 4 段階との比較

本節では、各語の文体指標の平均値(「硬度平均値」の結果を主に示す)がどの程度、語の文体差に関する内省判断と一致するかを先行研究の調査結果と比較しながら検討する。

比較する先行研究は、語の文体差に関する多数の文献の記述に基づいて主に接続詞や副詞などの文体差をまとめた柏野(2016)及び語の文体差に関するアンケート調査結果を示している井上(2013)である。

まず、柏野(2016)の「使用目安の分類」(4 段階)との比較を行う。

柏野(2016)は、大学教育や日本語教育における学術的文章作成に役立てるために、「作文技術に関する文献」及び「書き言葉と話し言葉の相互関係に関する文献」に記載された様々な語の文体に関する記述を広く調査し、「書き言葉的」「話し言葉的」として示されている文体差のある語や表現を抽出して一覧にして示し、学術的文章作成の際の「使用目安の分類」(4 段階)を行っている。この「使用目安の分類」(4 段階)は、内省判断に基づく語の文体

差に関するこれまでの文献の知見を総合したものであり極めて有意義なものである。5 指標の語別平均値とこの「使用目安の分類」(4 段階)とがどの程度一致するかを見ることによって、5 指標の語別平均値の妥当性を見ることができる。

「使用目安の分類」(4 段階)は次の a～d のように分類されている(柏野 2016:38)。この a～d をそれぞれ 1～4 の値に置き替えて比較を行う。

a. 「話し言葉的」な語と比較的はっきり位置づけられるため、学術的文章には避けるべき語
b. 「書き言葉的」な語ともいえるが、学術的文章では避けた方が望ましい語
c. 「書き言葉的」な語と比較的はっきり位置づけられるが、学術的文章の文脈・内容によっては使用に注意が必要な語
d. 「書き言葉的」な語として学術的文章での使用に特に問題のない語

比較の対象とする語は、柏野(2016)の「表 1　a.接続の「書き言葉的・話し言葉的」な語」及び「表 2　b.副詞と c.文末の「書き言葉的・話し言葉的」な語」に記載されている語のうち、接続詞と副詞(BCCWJ の長単位の接続詞と副詞のみ)の計 51 語である。

図 9 に 51 語の硬度平均値と柏野の 1(=a)～4(=d)の段階(「柏野 4 段階」)の散布図を示す。

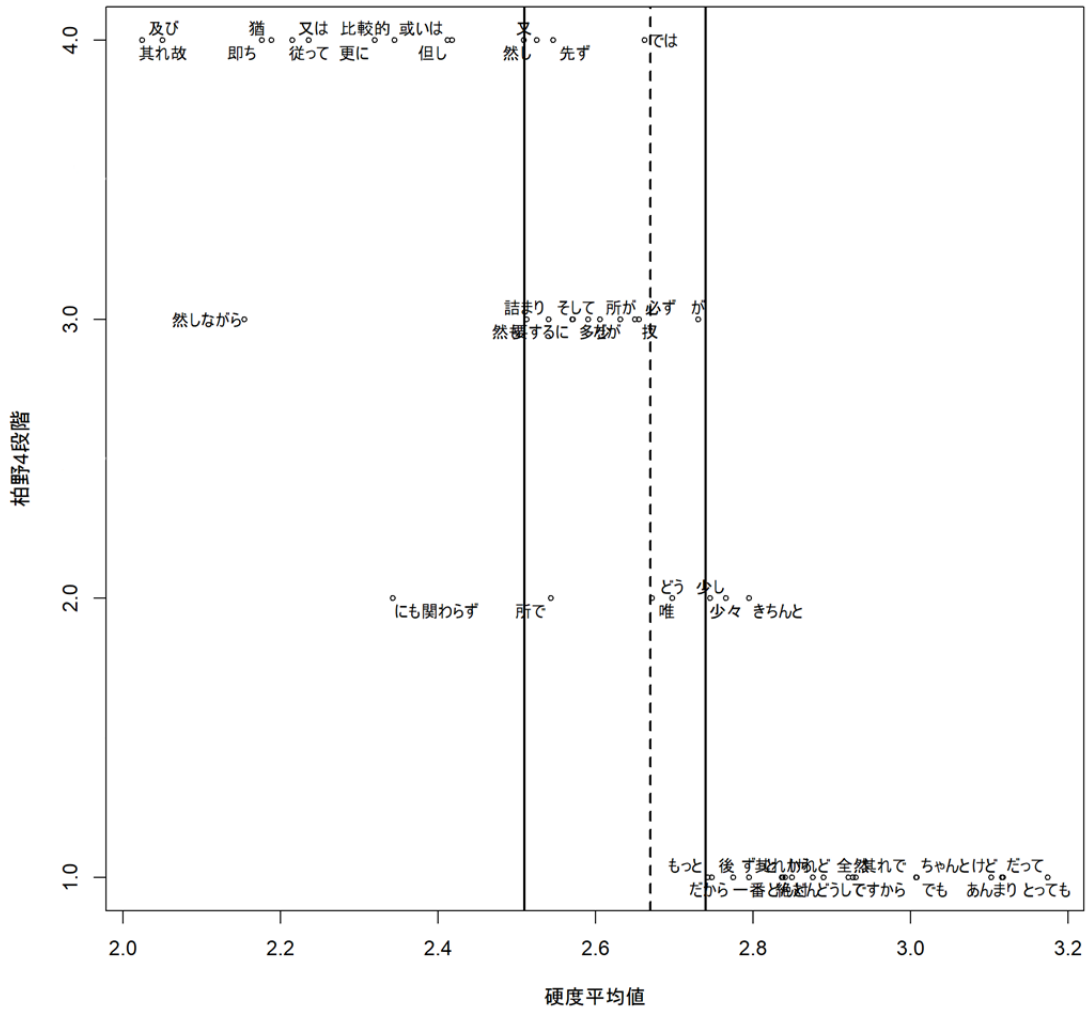


図 9　硬度平均値と柏野 4 段階との散布図(接続詞・副詞 51 語)

硬度平均値と柏野 4 段階の相関分析を行った。スピアマンの順位相関係数は $r_s$ = -0.886（$p$ < 0.001）であり、硬度平均値と柏野 4 段階とには強い相関がある[23]。

図 9 の散布図に基づいて、硬度平均値と柏野 4 段階との語ごとの対応を見る。

「然し、先ず、では」「然しながら、が」「にも関わらず、所で、少し、少々、きちんと」の 10 語を除くと、残りの 41 語は次のとおり対応している。ただし、この硬度平均値による区分けは一応の目安を示しただけのものである。柏野の 4(=d)の段階と 1(=a)の段階とは明確に区分けすることができる。しかし、3(=c)の段階と 2(=b)の段階との区分けは可能なのか、また、3(=c)の段階と 4(=d)の段階、2(=b)の段階の 1(=a)の段階の区分けの基準をどこに設けるのかという点は慎重に検討する必要がある[24]。

柏野の 4(=d)の段階：硬度平均値 2.5100 未満
柏野の 3(=c)の段階：硬度平均値 2.5100 以上 2.6700 未満
柏野の 2(=b)の段階：硬度平均値 2.6700 以上 2.7400 未満
柏野の 1(=a)の段階：硬度平均値 2.7400 以上

このように、硬度平均値は、内省判断による語の文体差とある程度一致しており、硬度平均値(専門度平均値、客観度平均値、くだけ度平均値)を、語の文体差の目安として用いることは可能である。

### ４.２　井上(2013)の「アンケート文体値」との比較

次に、井上(2013)のアンケート調査結果と比較する。

井上(2013)は、調査対象語(64 語)が「単語の文体 5 分類案」のどの分類に位置付けられるかの判断を求める調査を計 381 名に対して行い、その結果に基づいて、「アンケート調査により求めた単語の文体値(アンケート文体値)」[25]を示している。「単語の文体 5 分類案」とは、「レベル 1　卑俗体」「レベル 2　口頭体」「レベル 3　汎用体」「レベル 4　書記体」「レベル 5　文章体」の 5 分類である。

図 10 に、調査対象語のうちの 43 語[26]の硬度平均値と「アンケート文体値」との散布図を示す。

硬度平均値と「アンケート文体値」の相関分析を行った。ピアソンの積率相関係数は $r$= -0.817（$p$ < 0.001）であり、硬度平均値と「アンケート文体値」とには強い相関がある[27]。

---

[23] 統計 R(ver.3.4.3) cor.test 使用。ちなみに、ピアソンの積率相関係数は $r$ = -0.837（ $p$ < 0.001）である。専門度平均値($r_s$ = -0.827, $p$ < 0.001)、客観度平均値($r_s$ = -878, $p$ < 0.001)、くだけ度平均値($r_s$ = 0.859, $p$ < 0.001)の場合もそれぞれ強い相関があり同様の結果であった。語りかけ性度平均値については、スピアマンの順位相関係数は $r_s$ = 0.193（ $p$ = 0.175）、ピアソンの積率相関係数は $r$ = 0.254（ $p$ = 0.072 ）であり、有意な相関があるとは言えない。

[24] 散布図は示さないが、専門度平均値、客観度平均値、くだけ度平均値の場合も、ここで指摘した 3(=c)の段階と 2(=b)の段階との区分けの困難さが表れている。

[25] 「調査語の各文体レベルの判断人数にそのレベル値を乗じ、総和を求めた後、判断人数で除し」(井上 2013:303)て求めた平均値である。なお、井上(2013)は、「アンケート文体値」と「コーパス調査により求めた文体値(コーパス文体値)」とを比較し「コーパス文体値の有効性について検討」しているが、本稿では、「アンケート文体値」のみを利用する。

[26] 「アンケート文体値」が示されている 64 語のうち、本稿で対象とする語(語彙素)と対応させることのできるのは 43 語である。

[27] 統計 R(ver.3.4.3) cor.test 使用。なお、専門度平均値($r$ = -0.796, $p$ < 0.001)、客観度平均値($r$ = -749, $p$ < 0.001)、くだけ度平均値($r$ = 0.797, $p$ < 0.001)の場合もそれぞれ強い相関があり同様の結果であっ

図 10 のとおり、「少々」など若干の語がやや外れた位置にあるが、語の文体差の内省判断を求めるアンケート調査の結果と硬度平均値とはある程度対応しており、やはり、硬度平均値(専門度平均値、客観度平均値、くだけ度平均値)を、語の文体差の目安として用いることは可能である。

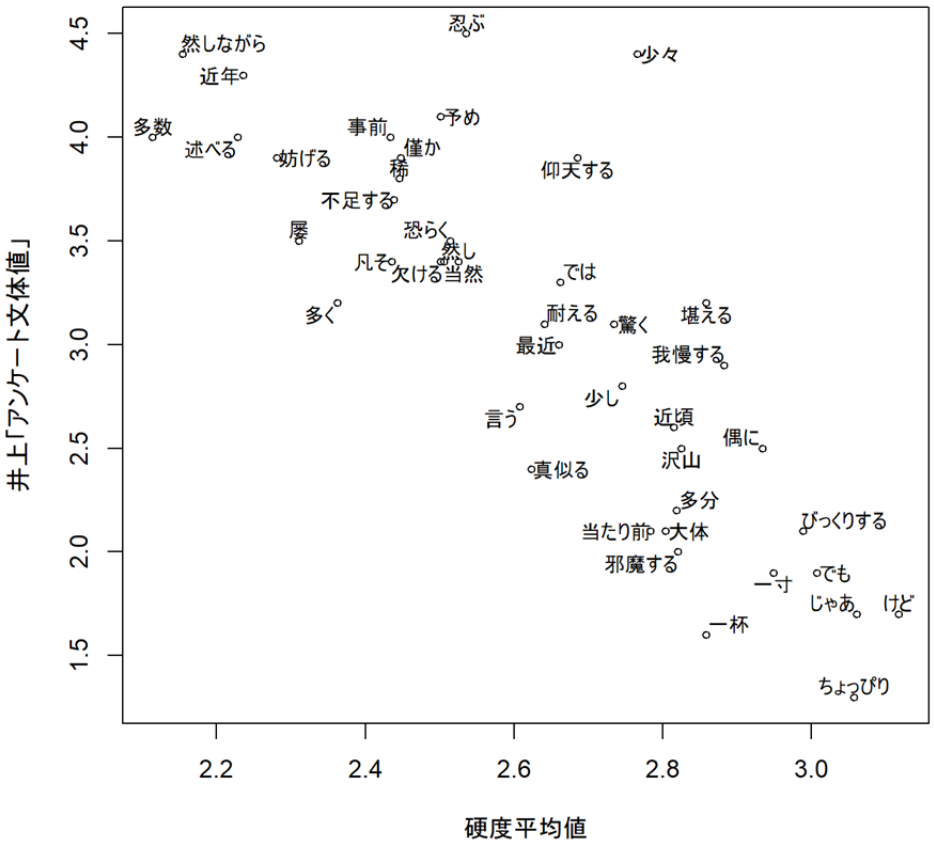

図 10　硬度平均値と井上「アンケート文体値」との散布図

## ５．おわりに

本稿では、『現代日本語書き言葉均衡コーパス』(BCCWJ)内の「図書館サブコーパス」のサンプルに対して文体情報を付与した『BCCWJ 図書館サブコーパスの文体情報』のアノテーションデータを利用し、「図書館サブコーパス」内の語(語彙素)を対象として各語の文体差を数値化する試みを行い、この試みが文体研究において有効性・可能性があることを示した。

本稿の概略は次のとおりである。

まず、『文体情報』の専門度、硬度などの文体に関する 5 種類のアノテーションデータを利用して、「図書館サブコーパス」内の記号類等を除く異なり語 719,809 語それぞれの専門度平均値、客観度平均値、硬度平均値、くだけ度平均値、語りかけ性度平均値を算出する方法を示した。本稿では、この 719,809 語のうち、文体指標の付与対象サンプル数が 100 以上の 7,877 語を対象として種々の観点から分析を行った。

---

た。語りかけ性度平均値については $r = 0.210$（ $p = 0.176$ ）であり、有意な相関があるとは言えない。

次に、語別の 5 指標の各平均値の相互の相関分析を行い、専門度、客観度、硬度、くだけ度の 4 指標の平均値は相互に強い相関があるのに対し、これら 4 指標の平均値と語りかけ性度の平均値とはいずれも相関がないことを示した。ある語が使われるサンプルの「内容・表現の文体的特徴」は専門度、客観度、硬度、くだけ度の 4 指標では共通性が高いと考えられる。品詞別の分析では、感動詞と接続詞は 4 指標の平均値とそのばらつきが他の品詞に比べて特異であり、感動詞は会話のような硬くないくだけた文体のテキストに偏って使われやすいことや接続詞は語の違いによる文体差が他の品詞に比べて大きいことという特徴が表されていることを示した。語種別の分析では、和語・外来語と漢語とは 4 指標の平均値の傾向が異なっており、和語と外来語は漢語に比べて相対的に硬くないくだけた文体のテキストで使われやすいという特徴が表されていることを示した。

さらに、各語の文体指標の平均値が、語の文体差に関する内省判断と強い相関があることを、柏野(2016)の学術的文章作成の際の「使用目安の分類」(4 段階)と井上(2013)のアンケート調査に基づく語の「アンケート文体値」と比較して示した。硬度等の 4 指標の平均値は内省判断による語の文体差とある程度一致しており、硬度等の 4 指標の平均値を語の文体差の目安として用いることが可能であることを示した。

さて、本稿の硬度平均値などの 4 指標の平均値は、語の文体差を「連続的」に捉えることができるものであり、さらに、文体指標付与対象サンプル数 100 以上に限っても 7,877 語という多数の語が対象となっている。今後、この平均値を用いて、語の文体差に関する従来の知見を検証するとともに新たな知見を得るための様々な分析を行いたい。品詞別の詳細な分析、語種別の詳細な分析、あるいは複合辞の文体差などさまざまな観点からの興味ある課題が多数ある。国語教育や日本語教育への応用も課題として挙げられる。

ただし、4 指標の平均値は相互に強い相関があったが、各平均値が果たして同一の性質の文体差を表しているものなのかなど、その性質の検討も必要である。話し言葉・書き言葉、硬・軟、フォーマル・インフォーマルなど文体差は多次元的に捉えることができるが、それとの整合性も検討する必要がある。また、「書き言葉らしさ」「話し言葉らしさ」の観点から語の文体差を一次元の連続的なものとして数値化した「語彙密度平均値」(佐野ほか 2009、佐野 2009、佐野 2016)がある。4 指標の平均値とこの「語彙密度平均値」との異同を検討する必要がある。さらに、平均値以外にも、最頻値や標準偏差などの値に着目して、その特徴や活用の可能性についても検討する必要がある[28]。

このように、今後、『BCCWJ 図書館サブコーパスの文体情報』を利用した様々な研究の進展が期待される。

## 謝　辞

本研究は、国立国語研究所共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」「コーパスアノテーションの基礎研究」による成果の一部である『BCCWJ 図書館サブコーパスの文体情報』を利用して行われたものである。この成果を利用させていただいたことに感謝申し上げる。また、柏野和佳子氏から、『BCCWJ 図書館サブコーパスの文体情報』に関する貴重な情報をいただくとともに、氏のご論考をお送りいただいた。記し

[28] 『文体情報』を用いた研究としては、語の文体差に関わる柏野ほか(2014)などの一連の研究、各指標に関する統計的分析を加えた浅原ほか(2014,2015)や浅原・加藤(2015)の研究、特に「語りかけ性度」に関わる加藤ほか(2014)などの一連の研究などがある。

て感謝申し上げる。
(本研究は JSPS 科研費 JP16K02715 の助成を受けたものである。)

## 文　献

浅原正幸・加藤祥(2015)「文体指標を特徴づける係り受け部分木の抽出」『第 8 回コーパス日本語学ワークショップ予稿集』,国立国語研究所言語資源研究系・コーパス開発センター,pp.171-178.

浅原正幸・加藤祥・立花幸子・柏野和佳子(2014)「文体指標と語彙の対応分析」『第 6 回コーパス日本語学ワークショップ予稿集』,国立国語研究所言語資源研究系・コーパス開発センター,pp.11-20.

浅原正幸・加藤祥・立花幸子・柏野和佳子(2015)「文体指標と語彙系列の対応分析」『第 7 回コーパス日本語学ワークショップ予稿集』,国立国語研究所言語資源研究系・コーパス開発センター,pp.7-16.

井上次夫(2009)「論説文における語の文体の適切性について」『日本語教育』(141),日本語教育学会,pp.57-67.

井上次夫(2013)「単語の文体判断について(3)―話しことばと書きことば―」『全国大学国語教育学会　国語科教育研究　第 125 回広島大会研究発表要旨集』,全国大学国語教育学会,pp.303-306.

柏野和佳子(2013)「書籍サンプルの文体を分類する」『国語研プロジェクトレビュー』4(1),国立国語研究所,pp.43-53.

柏野和佳子(2016)「学術的文章作成時に留意すべき「書き言葉的」「話し言葉的」な語の分類」『計量国語学会第六十回大会予稿集』,計量国語学会,pp.37-42.

柏野和佳子・立花幸子・保田祥・飯田龍・丸山岳彦・奥村学・佐藤理史・徳永健伸・大塚裕子・佐渡島紗織・椿本弥生・沼田寛(2012)「書籍テキストへの文体情報付与の試み―『現代日本語書き言葉均衡コーパス』の収録書籍を対象に―」『第 2 回コーパス日本語学ワークショップ予稿集』,国立国語研究所言語資源研究系・コーパス開発センター,pp.155-164.

柏野和佳子・中村壮範(2014)「BCCWJ 図書館サブコーパスの文体情報検索ツールによるテキスト分析」『第 5 回コーパス日本語学ワークショップ予稿集』国立国語研究所言語資源研究系・コーパス開発センター,pp.171-180.

加藤祥・柏野和佳子・立花幸子・丸山岳彦(2014)「語りかける書きことばの表現」『国立国語研究所論集』(8),国立国語研究所,pp.85-108.

国立国語研究所(2015)『BCCWJ 図書館サブコーパスの文体情報』(第 1 版)(http://pj.ninjal.ac.jp/corpus_center/anno/) (BCCWJ_LB_Stylistics-1.0.zip).

国立国語研究所コーパス開発センター(2015)『『現代日本語書き言葉均衡コーパス』利用の手引　第 1.1 版』(http://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html).

佐野大樹(2009)「「話し言葉らしさ・書き言葉らしさ」の計測―語彙密度の日本語への適用性の検証―」『機能言語学研究』5,日本機能言語学会,pp.89-102.

佐野大樹(2016)「語彙密度から見た語彙シラバス」森篤嗣(編)『ニーズを踏まえた語彙シラバス』,くろしお出版,pp.79-93.

佐野大樹・丸山岳彦・山崎誠・柏野和佳子・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子(2009)『語彙密度を利用した『現代日本語書き言葉均衡コーパス』テクスト分類の試み』(特定領域研究「日本語コーパス」平成 20 年度研究成果報告書),文部科学省科学研究費特

定領域研究「日本語コーパス」データ班.
馬場俊臣(2018)「接続詞の文体差の計量的分析の試み―『BCCWJ 図書館サブコーパスの文体情報』を用いて―」『北海道教育大学紀要　人文科学・社会科学編』69(1),北海道教育大学,pp.1-14.
宮島達夫(1977)「単語の文体的特徴」松村明教授還暦記念会(編)『松村明教授還暦記念　国語学と国語史』,明治書院,pp.871-903.

## 関連 URL

「『現代日本語書き言葉均衡コーパス』語彙表 ver.1.1 解説」 http://pj.ninjal.ac.jp/corpus_center/bccwj/bcc-chu.html (BCCWJ 語彙表解説_1.1.pdf)
国立国語研究所(2015)『BCCWJ 図書館サブコーパスの文体情報』(第 1 版) http://pj.ninjal.ac.jp/corpus_center/anno/ (BCCWJ_LB_Stylistics-1.0.zip)
国立国語研究所共同研究プロジェクト「コーパスアノテーションの基礎研究」成果物配布サイト http://pj.ninjal.ac.jp/corpus_center/anno/
「BCCWJ 品詞構成表(Version 1.1)」 http://pj.ninjal.ac.jp/corpus_center/bccwj/bcc-chu.html (BCCWJ_frequencylist_pos_ver1_1.zip)