

国立国語研究所学術情報リポジトリ

『現代日本語書き言葉均衡コーパス』新聞記事情報 を用いたジャンル別語彙分布

| | |
|-------|---|
| メタデータ | 言語: Japanese 出版者: 公開日: 2022-01-07 キーワード (Ja): キーワード (En): Word List by Semantic Principles (WLSP), Balanced Corpus of Contemporary Written Japanese (BCCWJ) 作成者: 加藤, 祥, 森山, 奈々美, 浅原, 正幸, MORIYAMA, Nanami メールアドレス: 所属: |
| URL | https://doi.org/10.15084/00003489 |

『現代日本語書き言葉均衡コーパス』新聞記事情報を用いた ジャンル別語彙分布

加藤 祥 (目白大学) †

森山 奈々美

浅原 正幸 (国立国語研究所)

The Lexical Distribution by the Topic annotation data on the Newspaper Articles in the Balanced Corpus of Contemporary Written Japanese

Sachi Kato (Mejiro University)

Nanami Moriyama

Masayuki Asahara (National Institute for Japanese Language and Linguistics)

要旨

コーパスに付与されたジャンル情報を用いることにより、ジャンル毎の語彙分布の傾向が確認される。しかし、レジスタによる文体差の影響や、ジャンルの分類基準の問題が考えられる。そこで、本稿は、文章内容情報が付与された文体的な影響の少ないコーパスを用い、品詞分布・語彙分布・語義分布に内容別の傾向が見られることを確認する。具体的には、『現代日本語書き言葉均衡コーパス』の新聞サブコーパス (PN, 1,473 サンプル) に含まれるサンプルを記事単位 (5,585 記事) に分割し、記事ごとの内容情報や種別情報を付与した (加藤ほか 2020) データを用いる。分類語彙表番号の付与された BCCWJ-WLSP (加藤ほか 2019) と重ね合わせるにより語義分布も調査する。

1. はじめに

『現代日本語書き言葉均衡コーパス』 (以降 BCCWJ) に対する様々な情報付与を進める中で、文脈に関わる情報として、文脈的な意味の判断や文章情報の付与を行っている (加藤・浅原・山崎 2019, 加藤・森山・浅原 2020, 加藤・森山・浅原 2021 など)。読み手が意味を読み取るに際しては、言語化されている情報はもちろん、広く想起されたと考えられる言外の情報、読み手の有する百科事典的な知識などの様々な文脈の影響が考えられるが、特に、文章内容によって語彙の分布に傾向のあること、さらに語義の分布は文脈と関連のあることが期待される。よって、これまでに進めてきた「文脈」に関する情報付与コーパスを用い、文脈情報による語彙分布の傾向を調査し、また、文脈情報が語義の絞り込みに有用であるのか検証したい。

但し、文体が語彙分布に及ぼす影響が考えられる。そこで、文体情報の影響が少なく多様な内容分布の得られる文章として、新聞記事を用いる調査を行うこととした。しかし、BCCWJ の新聞サブコーパス (以降 PN) には、ジャンル (全国紙・ブロック紙・地方紙の別) と朝刊夕刊の別、掲載年月日が付与されているものの、記事内容の情報はない。また、設計上、1つのサンプルには複数の記事が含まれている場合があり、複数記事は複数面にわたって掲載されている場合もある。そのため、BCCWJ の PN ファイル全サンプル (非コアも含む) について、サンプル内に含まれる記事単位の分割を行うとともに、記事ごとの掲載

† s.kato@mejiro.ac.jp

紙面・記事分類(記事の内容, 記事の種別)の情報を付与する作業を行った(加藤ほか 2020)。

本稿は, BCCWJ に付与した新聞記事情報のうち文脈的な情報として文章内容に該当する記事内容情報を用い, ジャンル別の品詞・語彙の分布傾向を確かめる。さらに, BCCWJ-WLSP を重ね合わせることで, ジャンルと語義の分布を調査する。また, ジャンル横断的に出現すると考えられる一般的な動詞などについても, 語義的な分布傾向を確かめる。

2. BCCWJ 新聞サンプルへの情報付与

2.1 記事分割

まず, サンプルを記事に分割した。具体的には, 記事間の境界情報となる各記事の開始位置と, サンプル内の記事数, 記事の完結・未完結情報を付与した。BCCWJ 内に付与された記事タグを参照し, 既存の記事タグに準じたが, 作業者が記事種を付与するにあたって別記事の扱いが適切と考えられた際には, さらに細かく記事を分割したことがある。サンプル内の最後の記事は, 取得サンプルの最後が可変長(文章のまとまりをもとに長さが決まる)の最後ではなく, 固定長(サンプルの開始点を含む文の文頭からサンプルの終了点を含む文の文末までが収録されている)の最後であるとき, 完結していない場合がある。そのため, サンプル内の最後の記事が完結しているか未完結であるかという情報とサンプル内の記事数も取得した。

2.2 記事分類情報の付与

ジャンル情報にあたる記事の分類情報として, 掲載紙面情報と記事内容情報を付与した。各記事の掲載紙面情報は, 掲載面の名称等を新聞紙面から取得した。また, 作業者が各記事を読み, 国内外の分類情報と記事内容の大分類と小分類を付与した。記事の掲載されている新聞紙面を人手で確認し, 掲載面の一般名称(総合・国際・社会など)の記載があれば取得した。掲載面に新聞固有の名称(「経済がわかる 企業がみえる」「いぶにんぐスペシャル」など)があれば参考として取得した。名称の記載がない場合でも, 一面や明らかなラテ欄であれば「なし(一面)」「なし(ラテ欄)」などと「なし」に()で備考を付した。また, 各記事が国内・海外・国際のいずれの内容であるのかを付与した。

記事内容情報として, 表1の分類(大分類・小分類)を付与した。記事内容の分類にあたっては, 紀伊國屋書店『CD-HIASK』の分類を援用した。作業者が新聞記事を読み, 大分類を選定した後, 小分類を選定した。なお, BCCWJ の設計上, 見出しのみが記事認定されている場合があり, 記事内容が判別しにくい例もあるため, 分類に「見出し」を加えた。

記事に記事分類が複数該当する場合, 一記事に複数の内容が含まれる場合は, 作業者に主となると判断された分類が選ばれている。別の小分類が考えられた場合は備考欄にその旨が付されている。記事内容に重複が考えられた場合については, 作業基準を統一してある。たとえば, 訴訟に関する記事は, 事件(7)か司法(16)で作業者による判断揺れが生じたため, 事件等取扱い内容であれば「事件」とし, 「訴訟」「国会(代表質問)」などを備考欄に付すようにした。

また, 地方紙において政治記事が全て「地方行政(15)」となるなど, 小分類よりもさらに下位の分類が必要と考えられる場合もあったため, たとえば地方行政(15)であれば「訴訟」「財政」などの小分類に該当する情報を備考欄に付した。また, 小分類の「諸競技(88)」などは, 「ボクシング」「ゴルフ」のような具体的な競技名を付与するなどさらに下位の分類情報を備考欄に記述した。「野球(86)」のように, 「高校」「プロ」のような分類を付与した場合もある。

表1 記事内容の分類

| 大分類 | 小分類 | | | | | | | | | | |
|--------|---------|---------|---------|---------|---------|----------|----------|---------|---------|------------|--|
| 0_総類 | 00_総類 | 01_皇室 | 02_言論報道 | 03_新聞 | 04_放送 | 05_出版 | 06_新メディア | 07_大戦 | | | |
| 1_政治 | 10_政治 | 11_国会 | 12_選挙 | 13_政党 | 14_行政 | 15_地方行政 | 16_司法 | 17_財政 | 18_外交 | 19_防衛 | |
| 2_経済 | 20_経済 | 21_金融 | 22_貿易 | 23_商業 | 24_鉱工業 | 25_エネルギー | 26_農林漁業 | 27_運輸交通 | 28_情報通信 | 29_土木建設 | |
| 3_労働 | 30_労働 | 31_雇用 | 32_労働条件 | 33_賃金 | 34_労働運動 | 35_労働組合 | 36_官公労組 | 37_民間労組 | 38_職業 | | |
| 4_文化 | 40_文化 | 41_文学 | 42_教育 | 43_宗教 | 44_美術 | 45_芸能 | 46_演劇 | 47_映画 | 48_音楽 | | |
| 5_科学 | 50_科学 | 51_地球 | 52_天文 | 53_宇宙 | 54_原子力 | 55_生物 | 56_医療 | 57_生理疫学 | 58_保健衛生 | 59_公害 | |
| 6_社会 | 60_社会 | 61_福祉 | 62_世代 | 63_家庭 | 64_生活 | 65_趣味娯楽 | 66_旅行観光 | 67_世相風俗 | 68_行事 | | |
| 7_事件 | 70_事件 | 71_災害 | 72_火災 | 73_事故 | 74_交通事故 | 75_凶悪犯 | 76_人質犯 | 77_経済犯 | 78_公安事件 | 79_その他犯罪 | |
| 8_スポーツ | 80_スポーツ | 81_総合競技 | 82_陸上 | 83_水上 | 84_冬季 | 85_球技 | 86_野球 | 87_武道 | 88_諸競技 | 89_キャンブル競技 | |
| 9_国際 | 90_国際 | 91_国際政治 | 92_国際軍事 | 93_平和運動 | 94_被爆 | 95_国際経済 | 96_国連 | | | | |
| 見出し | | | | | | | | | | | |

2.3 記事種別情報の付与

新聞記事はいずれも文体差の少ないことが期待されるが、新聞には一般報道記事のほか、様々な種類の記事が掲載されている。そのため、記事の種別として、2種類の分類情報を付与した。記事種別情報についても、紀伊國屋書店『CD-HIASK』の種別を援用し、以下の基準によって情報を付与した。

記事種別1

- A_評論：硬めの囲み記事（社説や新聞社コラムを含む）
- B_投書：囲み記事で「投書」とあるか、投書の引用が明示的な場合
- C_インタビュー：発言がそのまま記述されている場合（※討論は備考に付す）
- D_ハウツー：料理レシピや手芸指南等に付与
- E_催し案内：日時が紹介されている場合（イベント情報・テレビ番組等含む）
- F_人：人物紹介、評伝など（囲み記事であってもインタビューではない場合）
- G_色：家庭、文化、娯楽などの囲み記事（囲碁将棋の枠も含む）
- H_人事
- I_募集
- J_賞

記事種別2

- 1_連載：連載番号または「上中下」「前後」等が確認できた記事（小説も含む）
- 2_解説：報道記事の説明部分、事件事象等の囲み説明・催しの囲み説明等
- 3_テキスト：一般報道記事
- 4_名簿：名前の羅列
- 5_略歴：人物の略歴が別途囲み記事になっている場合
- 6_日誌：当該日付の出来事紹介、記者日記等
- 7_用語：用語解説・用語説明
- 8_データ：日時のみをはじめ数値情報・式辞発表・判決などの全文掲載等（データ表等はBCCWJではサンプリング対象外のため該当なし）
- 9_死亡：死亡記事（訃報、おくやみ・解説・略歴等の囲み記事形式も含む）

2.4 語義情報の利用

語の単位で文脈的に意味情報の付与されたコーパスデータとして、BCCWJの短単位に分類語彙表の意味分類が付与されたBCCWJ-WLSP(加藤・浅原・山崎, 2019)がある。各短単位に、読み手が文脈上どのような意味として読んだのか、手作業で意味分類が付与されている。当該文脈における当該語の意味情報が適宜付与されているデータである。なお、読み手が文脈上読み取った意味は、分類語彙表に当該語の語義として掲載された番号に限定されない場合もある。BCCWJ-WLSPのPNデータは115,876短単位であり、付属語を除く65,344短単位に語義が、助動詞6,532短単位に用法が付与されている。本稿では、BCCWJ-WLSPのPNデータを使用する。

3. 語彙分布

まず、PNデータの固定長を用い、ジャンル(記事内容の大分類)ごとの語彙分布を調査した。ここでは品詞分布に見られるジャンルの特徴を示す。また、BCCWJ-WLSPと重ね合わせることで、ジャンルごとの語義分布も調査した。

3.1 品詞分布

ジャンル別の品詞分布から、名詞と修飾語句に特徴が見られる傾向がわかった。表2にジャンル別の体・用・相分布(固定長・相対頻度)とMVRを示す。体はUniDicの品詞における名詞と代名詞、用は動詞、相は形容詞・形状詞・副詞・連体詞として集計した。

経済記事で特に名詞(体)率が高い傾向が見られる。また、文化・社会・科学記事において修飾語句(相)率が高いという分布傾向が確認されるが、科学記事は、動詞(用)率が高く、特にMVRの高い文化・社会記事よりもMVRが低いという点で差異が見られ、ジャンルによる違いが現れている。但し、ジャンルによって記事種別分布¹が異なるため、記事種別による影響を考える必要がある。たとえば、このほか、事件・国際記事で類似した傾向が見られるが、いずれもMVRが低く、記事種が報道中心であるというジャンルの特徴が影響しているものと考えられる。記事種別情報を用いた検証が求められよう。

表2 ジャンル別体用相分布(BCCWJ全体・固定長・相対頻度)

| 品詞 | 相対頻度 | 総類 | 社会 | 政治 | 事件 | スポーツ | 文化 | 科学 | 国際 | 経済 | 労働 |
|-----|----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 体 | 403151 | 391770 | 395821 | 412472 | 409887 | 405738 | 370747 | 381198 | 406158 | 445826 | 404179 |
| 用 | 91400 | 92795 | 90804 | 91509 | 93841 | 85982 | 96575 | 101654 | 94125 | 83892 | 91499 |
| 相 | 26193 | 27205 | 29689 | 22941 | 19297 | 26434 | 32233 | 31152 | 22905 | 21624 | 27738 |
| MVR | 100万語当たり | 293169 | 326958 | 250692 | 205638 | 307441 | 333760 | 306449 | 243350 | 257756 | 303150 |

また、表3に品詞がどのような記事で出現するかを検討したカイ二乗検定結果を示す。品詞間のジャンルごとの頻度を係数し、カイ二乗検定を行い、標準化残差の値により検討する。標準化残差は±1.96より外側の場合 $p < 0.05$ 水準で有意(表中↗↘で示す)、±2.56より外側の場合 $p < 0.01$ 水準で有意(表中↑↓で示す)とされる。

名詞は経済記事で特に多く、文化記事で特に少ないことが確認できるが、助動詞で反対の傾向(文化で多く、経済で少ない)がある。また、文化記事においては、代名詞が多いという特徴が見られる。感動詞や副詞の多さからも、連載小説やインタビュー等、記事種別の影響が考えられる。なお、事件記事で特に接尾辞が多く示されるのは、「署」「警」をはじめ、「者」「官」「等」のような特定の集団や職種など(主に人物)を示す語の影響である。品詞

¹ 本稿では詳述しないが、たとえば、記事種別1のG(色)の分布は新聞全体で5.3%ながら、社会記事では25.6%と突出している。総類記事は49.0%がE(催し案内)であるという特徴もある。また、記事種別2の3(一般報道)は新聞全体で73.9%である。一般報道記事の割合が高いのは、スポーツ記事92.5%、事件記事90.8%、国際記事84.3%である。しかし、文化記事では報道記事が53.3%にとどまり、28.8%が2(解説)、15.2%が1(連載)である。文化記事では概ね連載小説が1に該当する。

の分布にはジャンルによって異なる傾向があるといえよう。

さらに、品詞分布の詳細を見ても、ジャンルに関わる傾向が見られる。そこで、表4では例として助詞の種類別分布を示す。表5は、助詞と記事ジャンルの出現を評価した結果を可視化した表である。社会記事で接続助詞、副助詞、準体助詞、終助詞が多く、格助詞が少ない。スポーツ記事では副助詞が取り立てて少ないが、係助詞が多い。頻度(表4)においてスポーツと総類に次いで副助詞が少ない国際記事では、格助詞が多いという異なるジャンル別の傾向が見られる。

なお、副助詞がスポーツ記事で低頻度となるのは、付加や並列の用例が少ないためである。また、終助詞が文化記事で高頻度となるのは、記事種(連載小説)の影響がある。

表3 ジャンル別品詞分布 (BCCWJ 全体・固定長 カイ二乗検定の標準化残差)

| | 総類 | 社会 | 政治 | 事件 | スポーツ | 文化 | 科学 | 国際 | 経済 | 労働 |
|-------|---------|---------|----------|----------|----------|----------|----------|---------|----------|---------|
| 名詞 | ↓ -7.11 | ↓ -7.72 | ↑ 8.97 | ↑ 5.61 | ↑ 2.82 | ↓ -30.18 | ↓ -11.32 | ↔ 2.50 | ↑ 35.11 | ↔ 0.34 |
| 助詞 | ↓ -3.78 | ↔ -2.52 | ↑ 5.20 | ↔ 0.35 | ↓ -9.37 | ↑ 3.92 | ↑ 9.15 | ↑ 7.79 | ↓ -7.06 | ↔ -0.25 |
| 補助記号 | ↑ 18.80 | ↑ 10.84 | ↓ -15.95 | ↓ -13.44 | ↑ 12.27 | ↑ 13.26 | ↓ -8.24 | ↓ -9.13 | ↓ -10.57 | ↓ -2.81 |
| 動詞 | ↔ 1.20 | ↔ -1.13 | ↔ 0.03 | ↑ 2.65 | ↓ -7.94 | ↑ 7.46 | ↑ 9.14 | ↑ 2.65 | ↓ -10.17 | ↔ 0.00 |
| 助動詞 | ↔ -2.14 | ↔ -0.94 | ↓ -8.50 | ↑ 3.63 | ↑ 7.64 | ↑ 14.82 | ↑ 2.67 | ↓ -0.06 | ↓ -17.54 | ↔ 0.55 |
| 接尾辞 | ↓ -8.82 | ↓ -5.68 | ↑ 14.62 | ↑ 18.74 | ↓ -10.52 | ↓ -2.90 | ↔ 2.23 | ↔ -1.70 | ↓ -6.57 | ↑ 4.28 |
| 空白 | ↔ -0.15 | ↑ 6.93 | ↓ -3.27 | ↓ -2.59 | ↔ 1.85 | ↔ 2.23 | ↔ -1.53 | ↔ -1.43 | ↓ -4.43 | ↔ 0.28 |
| 形容詞 | ↔ -1.50 | ↑ 8.36 | ↓ -7.42 | ↓ -8.43 | ↑ 5.17 | ↑ 6.76 | ↑ 7.86 | ↓ -7.68 | ↓ -5.90 | ↔ 0.60 |
| 接頭辞 | ↓ -7.54 | ↔ -0.58 | ↑ 5.71 | ↔ 0.64 | ↓ 2.43 | ↔ -4.18 | ↔ -0.85 | ↔ -4.08 | ↑ 5.94 | ↔ -1.50 |
| 形状詞 | ↔ -0.97 | ↔ 1.10 | ↑ 0.95 | ↓ -4.94 | ↓ -6.04 | ↑ 4.76 | ↑ 2.82 | ↑ 3.13 | ↔ -1.07 | ↔ 1.43 |
| 副詞 | ↔ 1.90 | ↑ 6.84 | ↓ -6.62 | ↓ -9.26 | ↑ 2.56 | ↑ 11.19 | ↑ 3.15 | ↓ -4.27 | ↓ -8.63 | ↔ 2.05 |
| 代名詞 | ↑ 8.74 | ↑ 6.46 | ↓ -7.00 | ↓ -8.21 | ↓ -3.68 | ↑ 19.01 | ↔ -1.61 | ↓ -4.88 | ↓ -10.79 | ↔ 0.07 |
| 連体詞 | ↑ 5.59 | ↑ 4.12 | ↓ -4.11 | ↓ -5.07 | ↓ -0.89 | ↑ 10.42 | ↔ 1.01 | ↓ -3.21 | ↓ -7.60 | ↔ -2.15 |
| 記号 | ↑ 11.57 | ↔ -1.47 | ↓ -10.40 | ↓ -9.15 | ↑ 20.32 | ↔ -1.88 | ↓ -3.69 | ↓ -3.88 | ↔ 0.42 | ↔ -4.72 |
| 接続詞 | ↔ -0.88 | ↔ -1.61 | ↑ 3.67 | ↓ -2.71 | ↓ -2.90 | ↔ -0.35 | ↔ 1.15 | ↑ 3.67 | ↔ 0.39 | ↔ 0.77 |
| 感動詞 | ↔ 2.00 | ↔ 1.87 | ↓ -4.07 | ↔ -1.71 | ↔ -0.17 | ↑ 9.93 | ↔ -2.34 | ↓ -2.59 | ↓ -4.41 | ↔ -0.33 |
| 未知語 | ↑ 4.88 | ↑ 4.33 | ↓ -5.09 | ↓ -3.86 | ↔ -1.57 | ↑ 4.13 | ↔ -1.05 | ↔ 0.41 | ↔ -1.44 | ↔ -1.79 |
| URL | ↔ 1.88 | ↑ 6.57 | ↓ -3.01 | ↓ -2.61 | ↔ -1.62 | ↔ -1.36 | ↔ -0.56 | ↔ -0.95 | ↔ 0.55 | ↔ 0.66 |
| 英単語 | ↔ -0.07 | ↔ -2.24 | ↔ -2.36 | ↔ -1.84 | ↔ -1.37 | ↑ 11.70 | ↔ -1.47 | ↔ -1.64 | ↔ -2.18 | ↔ 0.48 |
| カタカナ文 | ↔ -0.80 | ↔ -1.44 | ↔ -1.26 | ↔ -0.98 | ↔ 1.51 | ↔ -1.28 | ↑ 8.69 | ↔ -0.88 | ↔ -1.17 | ↔ -0.43 |
| 言いよどみ | ↔ -0.25 | ↔ -0.45 | ↔ -0.40 | ↔ -0.31 | ↔ -0.40 | ↔ 2.47 | ↔ -0.25 | ↔ -0.28 | ↔ -0.37 | ↔ -0.13 |

表4 ジャンル別助詞種分布 (全体・固定長・相対頻度)

| 品詞 | 相対頻度 | 総類 | 社会 | 政治 | 事件 | スポーツ | 文化 | 科学 | 国際 | 経済 | 労働 |
|------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 格助詞 | 156579 | 157438 | 147590 | 164199 | 159648 | 151913 | 155217 | 164147 | 171597 | 154153 | 151863 |
| 係助詞 | 32791 | 28365 | 32831 | 33417 | 30607 | 35357 | 31648 | 33932 | 33100 | 33378 | 34531 |
| 接続助詞 | 23971 | 22805 | 26579 | 22687 | 24577 | 22053 | 26968 | 27029 | 22738 | 19914 | 24856 |
| 副助詞 | 11461 | 10020 | 13781 | 11224 | 12330 | 7108 | 12213 | 14202 | 10490 | 11640 | 12660 |
| 準体助詞 | 3420 | 3315 | 4644 | 2881 | 2510 | 2586 | 4758 | 4139 | 2542 | 2527 | 3963 |
| 終助詞 | 1769 | 2200 | 2517 | 1020 | 958 | 1668 | 3269 | 1580 | 1027 | 863 | 1544 |

表5 ジャンル別助詞種分布 (BCCWJ 全体・固定長 カイ二乗検定の標準化残差)

| | 総類 | 社会 | 政治 | 事件 | スポーツ | 文化 | 科学 | 国際 | 経済 | 労働 |
|---------|---------|----------|---------|---------|----------|---------|---------|---------|---------|---------|
| 助詞-格助詞 | ↑ 5.77 | ↓ -16.21 | ↑ 7.22 | ↑ 3.81 | ↑ 3.15 | ↓ -7.79 | ↓ -3.01 | ↑ 9.09 | ↑ 4.70 | ↓ -2.73 |
| 助詞-係助詞 | ↓ -5.70 | ↑ 0.98 | ↓ -0.34 | ↓ -4.38 | ↑ 9.85 | ↓ -4.29 | ↔ -1.49 | ↔ -2.23 | ↑ 3.90 | ↔ 1.54 |
| 助詞-接続助詞 | ↔ -1.04 | ↑ 9.14 | ↓ -5.25 | ↔ 1.16 | ↓ -2.77 | ↑ 7.33 | ↔ 2.53 | ↓ -4.72 | ↓ -8.78 | ↔ 0.90 |
| 助詞-副助詞 | ↓ -2.92 | ↑ 11.12 | ↔ -1.97 | ↑ 2.64 | ↓ -15.73 | ↔ 2.25 | ↑ 4.83 | ↓ -4.16 | ↔ 2.12 | ↔ 1.67 |
| 助詞-準体助詞 | ↔ -0.10 | ↑ 10.28 | ↓ -4.43 | ↓ -5.15 | ↓ -5.09 | ↑ 9.22 | ↔ 2.13 | ↓ -5.13 | ↓ -5.29 | ↔ 1.33 |
| 助詞-終助詞 | ↑ 3.04 | ↑ 8.70 | ↓ -7.78 | ↓ -6.31 | ↔ -0.29 | ↑ 14.74 | ↔ -1.83 | ↓ -5.61 | ↓ -7.92 | ↔ -0.74 |

3.2 語義分布

記事のジャンルによって文脈の違いがあると考えられるため、ジャンル別の語彙の違いや語彙分布の特徴が期待されよう。実際に、記事ジャンル別の名詞においてジャンルの特徴が確認できる。表6に上位頻度名詞を示す。「県」は国際記事で低く、事件記事で高い頻度となっていた。「米」「政府」は国際記事、「選手」「代表」はスポーツ記事に突出する。「首相」は政治記事、「参加」「子供」は社会記事、「事業」は経済記事で高い。各ジャンルにおける記事内容が、特徴的な語彙分布と関係すると考えられる。しかし、内容語の上位頻度語

全般を見ると、「為る」「居る」「有る」のような動詞や「其の」「此の」のような連体詞、「多い」「大きい」のような形容詞等も含まれる。そこで、BCCWJ-WLSPを用い、記事ジャンル別の語義分布を確認することにしたい。

なお、BCCWJ-WLSPのPNは前節で確認したPNの一部サンプルであるため、表7と表8にBCCWJ-WLSPにおけるPNの「類」分布を示す。表2とほぼ同様の傾向が確認できている。

表6 PNの記事ジャンルと上位頻度名詞語彙素分布 (BCCWJ 全体 固定長・相対頻度)

| 名詞 | 頻度 | 総類 | 社会 | 政治 | 事件 | スポーツ | 文化 | 科学 | 国際 | 経済 | 労働 |
|----|------|------|------|------|------|------|------|------|------|------|------|
| 事 | 3271 | 2948 | 3089 | 3935 | 3176 | 2211 | 3588 | 4060 | 3967 | 2901 | 4426 |
| 市 | 2188 | 1176 | 2720 | 2454 | 2926 | 1246 | 2470 | 2164 | 860 | 2481 | 3294 |
| 県 | 1298 | 901 | 1079 | 1741 | 2562 | 1012 | 936 | 1722 | 218 | 1511 | 1801 |
| 会 | 1237 | 596 | 1427 | 2181 | 958 | 596 | 1430 | 1248 | 1053 | 1038 | 1235 |
| 人 | 781 | 886 | 1667 | 420 | 479 | 228 | 1047 | 1453 | 231 | 344 | 1235 |
| 区 | 778 | 153 | 1128 | 1027 | 1146 | 368 | 982 | 506 | 449 | 557 | 1081 |
| 問題 | 740 | 565 | 486 | 1300 | 875 | 201 | 312 | 806 | 2427 | 473 | 1184 |
| 同 | 715 | 214 | 641 | 660 | 1229 | 509 | 780 | 711 | 578 | 992 | 772 |
| 物 | 700 | 932 | 892 | 500 | 417 | 429 | 1099 | 900 | 552 | 588 | 669 |
| 米 | 617 | 504 | 123 | 560 | 573 | 643 | 228 | 237 | 2632 | 855 | 875 |
| 世界 | 599 | 794 | 428 | 207 | 125 | 1139 | 760 | 853 | 706 | 580 | 412 |
| 選手 | 589 | 336 | 112 | 20 | 83 | 3785 | 85 | 0 | 77 | 30 | 0 |
| 政府 | 534 | 351 | 107 | 1441 | 458 | 13 | 130 | 189 | 2414 | 397 | 412 |
| 関係 | 517 | 382 | 278 | 740 | 854 | 214 | 370 | 395 | 1207 | 565 | 721 |
| 代表 | 515 | 351 | 283 | 774 | 186 | 1527 | 247 | 142 | 629 | 214 | 103 |
| 首相 | 486 | 214 | 11 | 2454 | 260 | 0 | 52 | 15 | 1361 | 61 | 0 |
| 参加 | 478 | 260 | 834 | 420 | 271 | 442 | 455 | 300 | 758 | 244 | 669 |
| 子供 | 469 | 382 | 1213 | 140 | 354 | 174 | 702 | 363 | 295 | 61 | 823 |
| 事業 | 464 | 168 | 267 | 900 | 312 | 13 | 143 | 363 | 167 | 1611 | 566 |

表7 ジャンル別 分類語彙素「類」分布 (BCCWJ-WLSP・固定長・相対頻度)

| 類 | 相対頻度 | 見出し | 総類 | 社会 | 政治 | 事件 | スポーツ | 文化 | 科学 | 国際 | 経済 | 労働 |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| - | 441694 | 312500 | 469537 | 448177 | 424749 | 432542 | 467561 | 470902 | 446328 | 441361 | 395940 | 459559 |
| 1:体 | 425482 | 687500 | 414109 | 407216 | 440134 | 445220 | 392161 | 391938 | 403551 | 431287 | 482907 | 397978 |
| 2:用 | 91809 | 0 | 80623 | 99244 | 92308 | 93743 | 93760 | 93964 | 105327 | 90287 | 81689 | 92831 |
| 3:相 | 37449 | 0 | 31837 | 41439 | 38499 | 26285 | 44868 | 38581 | 42373 | 33454 | 36054 | 41360 |
| 4:他 | 3566 | 0 | 3894 | 3924 | 4311 | 2210 | 1650 | 4615 | 2421 | 3611 | 3410 | 8272 |

表8 ジャンル別 分類語彙素「類」分布 (BCCWJ-WLSP・固定長 カイ二乗検定の標準化残差)

| | 総類 | 社会 | 政治 | 事件 | スポーツ | 文化 | 科学 | 国際 | 経済 | 労働 |
|-----|---------|---------|---------|---------|---------|---------|---------|---------|----------|---------|
| - | ↑ 3.80 | ⇒ 1.42 | ↓ -4.34 | ⇒ -1.82 | ↑ 5.45 | ↑ 7.59 | ⇒ 0.47 | ⇒ -0.06 | ↓ -11.11 | ⇒ 1.19 |
| 1:体 | ⇒ -1.55 | ↓ -4.02 | ↑ 3.79 | ↑ 3.93 | ↓ -7.05 | ↓ -8.74 | ⇒ -2.23 | ⇒ 0.90 | ↑ 14.01 | ⇒ -1.84 |
| 2:用 | ↓ -2.64 | ↑ 2.80 | ⇒ 0.20 | ⇒ 0.64 | ⇒ 0.70 | ⇒ 0.95 | ⇒ 2.36 | ⇒ -0.40 | ↓ -4.24 | ⇒ 0.11 |
| 3:相 | ⇒ -2.01 | ⇒ 2.29 | ⇒ 0.69 | ↓ -5.77 | ↑ 4.09 | ⇒ 0.76 | ⇒ 1.31 | ⇒ -1.58 | ⇒ -0.89 | ⇒ 0.68 |
| 4:他 | ⇒ 0.37 | ⇒ 0.65 | ⇒ 1.58 | ↓ -2.23 | ↓ -3.37 | ⇒ 2.27 | ⇒ -0.97 | ⇒ 0.06 | ⇒ -0.32 | ↑ 2.62 |

以下の表9と表10に分類語彙素の部門分布を示す。「-」は部門がない語(分類語彙素番号がない主に付属語など)であることを示す。「1:関係」が経済記事に高頻度であるほか、「2:主体」が科学に低く国際記事に高い、「4:生産物」が社会で高くスポーツ記事に低い、「5:自然」が科学記事に突出するなどの特徴的な分布が確認される。ジャンルごとに語義分野の分布が異なることがわかる。

なお、部門以下の中項目を見ることで、ジャンルの特徴語義も確認できる。表 11 に上位頻度の中項目分布を示す。たとえば科学記事では作用 (.15) のほか表外では資材 (.41) 機械 (.46) 自然 (.50) 動物 (.55)，経済記事では量 (.19) 事業 (.38)，社会記事では社会 (.26) のほか表外の食料 (.43) 住居 (.44)，文化記事で言語 (.31) 表外の芸術 (.32) のように、記事ジャンルの文脈が明らかとなっている。

表 9 ジャンル別 分類語彙表「部門」分布 (BCCWJ-WLSP・固定長・相対頻度)

| 部門 | 相対頻度 | 総類 | 社会 | 政治 | 事件 | スポーツ | 文化 | 科学 | 国際 | 経済 | 労働 |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| - | 441694 | 312500 | 469537 | 448177 | 424749 | 432542 | 467561 | 470902 | 446328 | 441361 | 395940 |
| 1:関係 | 263958 | 375000 | 214384 | 257249 | 277295 | 253664 | 276534 | 239345 | 260291 | 235507 | 315063 |
| 3:活動 | 164943 | 125000 | 155978 | 156283 | 188183 | 150965 | 146364 | 160886 | 139629 | 179434 | 175152 |
| 2:主体 | 93990 | 156250 | 117270 | 84697 | 91862 | 114329 | 89015 | 91656 | 67797 | 120509 | 81770 |
| 4:生産物 | 18847 | 31250 | 16491 | 30816 | 9736 | 25936 | 9283 | 19182 | 27845 | 11025 | 23873 |
| 5:自然 | 16568 | 0 | 26340 | 22777 | 8175 | 22563 | 11243 | 18028 | 58111 | 12165 | 8201 |

表 10 ジャンル別 分類語彙表「部門」分布 (BCCWJ-WLSP・固定長 カイ二乗検定の標準化残差)

| | 総類 | 社会 | 政治 | 事件 | スポーツ | 文化 | 科学 | 国際 | 経済 | 労働 |
|-------|---------|---------|---------|---------|---------|---------|---------|---------|----------|---------|
| - | ↑ 3.80 | ⇒ 1.42 | ↓ -4.34 | ⇒ -1.82 | ↑ 5.45 | ↑ 7.59 | ⇒ 0.47 | ⇒ -0.06 | ↓ -11.11 | ⇒ 1.19 |
| 1:関係 | ↓ -7.63 | ⇒ -1.66 | ↑ 3.85 | ↓ -2.28 | ↑ 3.00 | ↓ -7.21 | ⇒ -0.42 | ↓ -4.83 | ↑ 13.98 | ↓ -4.10 |
| 3:活動 | ⇒ -1.64 | ⇒ -2.56 | ↑ 7.94 | ↓ -3.70 | ↓ -5.26 | ↓ -1.42 | ⇒ -3.45 | ↑ 2.92 | ↑ 3.31 | ↑ 2.92 |
| 2:主体 | ↑ 5.42 | ↓ -3.48 | ⇒ -0.92 | ↑ 6.84 | ⇒ -1.78 | ⇒ -1.02 | ↓ -4.53 | ↑ 6.82 | ↓ -5.03 | ⇒ 0.39 |
| 4:生産物 | ⇒ -1.17 | ↑ 9.64 | ↓ -8.50 | ↑ 5.11 | ↓ -7.37 | ⇒ 0.32 | ↑ 3.35 | ↓ -4.31 | ↑ 4.46 | ⇒ -1.91 |
| 5:自然 | ↑ 5.20 | ↑ 5.32 | ↓ -8.35 | ↑ 4.60 | ↓ -4.38 | ⇒ 1.47 | ↑ 16.45 | ↓ -2.59 | ↓ -7.90 | ⇒ 2.14 |

表 11 ジャンル別 分類語彙表「中項目」分布例 (BCCWJ-WLSP・固定長・相対頻度)

| 中項目 | 相対頻度 | 総類 | 社会 | 政治 | 事件 | スポーツ | 文化 | 科学 | 国際 | 経済 | 労働 |
|-----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| -- | 441694 | 469537 | 448177 | 424749 | 432542 | 467561 | 470902 | 446328 | 441361 | 395940 | 459559 |
| 19:関係-量 | 113659 | 86349 | 104029 | 112375 | 111072 | 124807 | 105070 | 103309 | 82874 | 153553 | 76287 |
| 30:活動-心 | 40574 | 29776 | 42684 | 51208 | 36753 | 34760 | 41537 | 43180 | 39346 | 35891 | 54228 |
| 15:関係-作用 | 33883 | 28859 | 34166 | 36046 | 28728 | 37648 | 30432 | 30670 | 32883 | 39383 | 25735 |
| 25:主体-公私 | 31849 | 37792 | 26127 | 26830 | 39195 | 24446 | 29134 | 29863 | 61585 | 33049 | 12868 |
| 34:活動-行為 | 30807 | 25653 | 24500 | 39093 | 37218 | 27334 | 23004 | 33091 | 41247 | 31019 | 32169 |
| 16:関係-時間 | 27940 | 30463 | 28137 | 26384 | 26518 | 32904 | 28196 | 19774 | 25661 | 27771 | 31250 |
| 12:関係-存在 | 23920 | 20843 | 24883 | 28911 | 21866 | 22486 | 20625 | 31477 | 24330 | 23873 | 18382 |
| 11:関係-類 | 21849 | 14201 | 21820 | 29803 | 20935 | 15369 | 17307 | 18563 | 31173 | 23711 | 19301 |
| 31:活動-言語 | 20698 | 23591 | 18184 | 19844 | 19888 | 12274 | 27764 | 14528 | 34594 | 16484 | 29412 |
| 37:活動-経済 | 18381 | 9620 | 22873 | 20290 | 11049 | 10108 | 8221 | 10089 | 12735 | 43281 | 12868 |
| 10:関係-事柄 | 15686 | 13284 | 16461 | 17540 | 15469 | 15781 | 15144 | 16546 | 15396 | 14373 | 17463 |
| 27:主体-機関 | 13676 | 10994 | 6316 | 27350 | 16167 | 13512 | 7211 | 6457 | 16917 | 11774 | 12868 |
| 36:活動-待遇 | 12977 | 9391 | 13686 | 23857 | 14887 | 5673 | 8942 | 3228 | 17297 | 10394 | 18382 |
| 35:活動-交わり | 12941 | 15117 | 6508 | 18357 | 5583 | 25890 | 8726 | 4843 | 21859 | 9176 | 13787 |
| 20:主体-人間 | 12830 | 19240 | 15408 | 9810 | 24075 | 9386 | 15360 | 12510 | 3421 | 6253 | 30331 |
| 17:関係-空間 | 12499 | 8016 | 13781 | 11297 | 15818 | 12790 | 10673 | 20985 | 8934 | 14048 | 8272 |
| 38:活動-事業 | 12426 | 11223 | 16078 | 8621 | 12910 | 4229 | 11755 | 14528 | 8554 | 22655 | 5515 |
| 24:主体-成員 | 11972 | 15117 | 7465 | 14270 | 13492 | 17638 | 12331 | 3228 | 14636 | 7065 | 10110 |
| 33:活動-生活 | 11691 | 16491 | 10432 | 6243 | 12212 | 25580 | 14062 | 12914 | 3231 | 5035 | 26654 |

3.3 動詞分布の問題

前節で見た通り、ジャンルと語義の分布には関連性が考えられる。記事の内容に応じた意味の語彙が分布するためである。名詞 (表 6 参照) の分布でも同様の傾向が見られているが、特に頻度の高い語として「為る」「居る」「有る」をはじめ、「言う」「来る」「見る」「行く」「つく」などのようにジャンルに横断的な動詞があり、このような動詞に関しては、ジャンルによる分布が見えにくいという問題がある。しかし、語義においては分布

傾向が確認できる可能性がある。そこで、語義の分布を用い、動詞の語義分布についてもジャンルごとの傾向を確かめておきたい。

まず、表 12 に動詞のジャンル別分布を確認しておく。特定動詞が特定記事ジャンルに出現すれば、名詞同様、ジャンルごとの特徴語により、文脈の定まる傾向が期待される。しかし、新聞は特に名詞の出現頻度が高く²サ変動詞の頻出する傾向が見られるため「為る」がいずれのジャンルでも高頻度で出現する。また、「ている」形を含む「居る³」も頻度が高いということになる。「成る」「言う」「有る」が頻出し、「因る」「見る」「出来る」などが次ぐ。なお、BCCWJ 全体の動詞頻度を見ると、上位頻度動詞は順に、「為る」「居る」「有る」「言う」「成る」「来る」「思う」「見る」「行く」「出来る」「因る」「つく」「考える」「仕舞う」「持つ」である。新聞記事においては、「受ける」「開く」などが特徴的な上位頻度動詞であるといえる。

表 12 PN の記事ジャンルと上位頻度動詞（語彙素）分布
(PN 全体・固定長・相対頻度)

| 動詞 | 相対頻度 | 総類 | 社会 | 政治 | 事件 | スポーツ | 文化 | 科学 | 国際 | 経済 | 労働 |
|-----|---------|-------|--------|--------|--------|-------|--------|--------|--------|-------|--------|
| 為る | 214721 | 18605 | 171690 | 264951 | 258894 | 16118 | 194021 | 235850 | 276044 | 24143 | 215624 |
| 居る | 7715239 | 66293 | 73902 | 75224 | 99350 | 5902 | 83523 | 98574 | 80502 | 72054 | 95204 |
| 有る | 3799185 | 32841 | 42054 | 37012 | 3760 | 2392 | 48684 | 4945 | 32612 | 3687 | 4220 |
| 成る | 3289 | 30702 | 35428 | 32944 | 20203 | 31083 | 3471 | 45811 | 2594 | 3626 | 43742 |
| 言う | 2903 | 33452 | 3420 | 23741 | 2739 | 22442 | 3744 | 38071 | 26064 | 2168 | 31392 |
| 出来る | 1136 | 7790 | 12504 | 11804 | 9164 | 11321 | 10400 | 1754 | 9373 | 1168 | 1441 |
| 因る | 1110264 | 9623 | 5931 | 1234 | 2427 | 389 | 8190 | 15007 | 2093 | 12060 | 12351 |
| 来る | 1107 | 1711 | 13252 | 954 | 646 | 10852 | 12870 | 1106 | 1066 | 916 | 6690 |
| 見る | 971 | 1100 | 930 | 774 | 12913 | 563 | 1125 | 1106 | 13353 | 916 | 12351 |
| 行く | 8190023 | 1100 | 1021 | 5602 | 552 | 871 | 12480 | 6793 | 591 | 5114 | 978 |
| つく | 793 | 596 | 481 | 1441 | 8331 | 3952 | 527 | 885 | 1695 | 8243 | 3602 |
| 行う | 6893498 | 5193 | 583 | 734 | 5311 | 13934 | 5200 | 585 | 9501 | 2900 | 6690 |
| 思う | 590 | 703 | 866 | 347 | 3020 | 6900 | 10400 | 411 | 360 | 2214 | 5661 |
| 受ける | 5642625 | 367 | 433 | 6002 | 10310 | 315 | 4420 | 806 | 7062 | 6641 | 8234 |
| 開く | 5012623 | 474 | 476 | 607 | 3124 | 3550 | 8190 | 427 | 6933 | 298 | 4632 |

ジャンル差は、事件記事で頻度の高い「因る」がスポーツ記事で頻度が低いこと、政治と国際記事で頻度が高い「つく」がスポーツと社会記事で頻度が低いことなどにおいて見られている。一部の動詞の分布については、文脈的な特徴が現れていると考えられる。このほか、科学に頻度が高い「有る」は「てある」、言うは「という」、事件や国際の「見る」は「とみる」などのように、用法の影響が見られる例もある。しかし、上位頻度語の分布は一部に特徴が確認できるものの、ほとんどの記事ジャンルの差が見えにくく、ジャンル横断的な分布であるといえる。また、「つく」「行く」「受ける」なども含め、高頻度語には多義語が頻出しているという問題がある。文脈に特有と考えられる名詞や、頻度の低い語に着目する場合は、ジャンルごとの特徴語が有用と考えられるが、動詞の高頻度語は多義的に用いられているため、特に意味的な分布を考慮する必要があると考えられ

² BCCWJ の名詞頻度（延べ）は 35%であるが、PN の名詞頻度は 46.5%である。OP（広報誌）61%、OW（白書）50.7%、OL（法律）47.5%に次いで高い。

³ 事件や国際記事に関して、「居る」の頻度が高めとなる傾向は確認できる。「ている」の影響が見られている。品詞分布（表 2・3）で見た記事種別（報道記事）が多いというジャンル特徴とも関わると考えられる。

る。

3.4 動詞の語義分布

それでは、BCCWJ-WLSP と重ね合わせることで、動詞の語義分布に着目したい。動詞の語義分布であれば、ジャンル別の特徴が確認できる可能性がある。

表 13 にジャンル別の動詞の語義分布を示す。動詞の語義についても記事内容というジャンル、大まかな文脈による分布傾向が推測される。但し、動詞の語義分布は、ジャンルの特徴語の多い名詞を含むジャンル全般の語義分布（表 9～11 を参照）とは異なる分布が見られた。たとえば「心」(.30) は政治・労働記事に頻出する傾向がある（表 11）が、動詞の語義分布（表 13）に限定すると、文化・スポーツ記事で頻度の高い傾向にある。「時間」(.16) は科学記事で最も頻度の低い傾向がある（表 11）が、動詞（表 13）では、科学記事がスポーツ記事に次ぐ高い頻度となっている。ジャンル別で頻出する語義が、動詞では同様の傾向を示すのではない場合もあるといえ、ゆえに動詞の語義については大まかな文脈では限定されにくい可能性が考えられる。どのような要素が語義の判定に関わるのか確認する必要がある。

表 13 記事ジャンルと動詞語義（中項目）分布（固定長・相対頻度）

| 中項目 | 相対頻度 | 総類 | 社会 | 政治 | 事件 | スポーツ | 文化 | 科学 | 国際 | 経済 | 労働 |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 34:活動-行為 | 24288 | 19240 | 20959 | 30100 | 27565 | 21454 | 19038 | 24617 | 32313 | 24442 | 30331 |
| 12:関係-存在 | 16948 | 14659 | 18375 | 18432 | 16981 | 16813 | 15649 | 26231 | 13115 | 16890 | 10110 |
| 15:関係-作用 | 16825 | 15117 | 19332 | 13973 | 14655 | 20320 | 19831 | 19371 | 13115 | 14779 | 18382 |
| 30:活動-心 | 9865 | 6871 | 12537 | 8919 | 9537 | 11140 | 12476 | 9282 | 7033 | 7471 | 8272 |
| 31:活動-言語 | 5943 | 6642 | 7273 | 4905 | 6047 | 5054 | 7500 | 5246 | 6463 | 4547 | 5515 |
| 11:関係-類 | 4706 | 3436 | 4498 | 6912 | 5583 | 1754 | 3606 | 2018 | 7983 | 4710 | 8272 |
| 37:活動-経済 | 3456 | 1832 | 4307 | 2081 | 3024 | 5467 | 3678 | 3228 | 2851 | 3492 | 4596 |
| 33:活動-生活 | 2426 | 2290 | 2775 | 1709 | 3140 | 2991 | 2380 | 7668 | 2091 | 1056 | 3677 |
| 38:活動-事業 | 1458 | 1603 | 2584 | 297 | 1744 | 825 | 2452 | 2421 | 570 | 1218 | 0 |
| 36:活動-待遇 | 1373 | 1145 | 1531 | 1338 | 1163 | 1754 | 1226 | 1211 | 2281 | 1056 | 919 |
| 35:活動-交わり | 1299 | 2520 | 957 | 743 | 698 | 3713 | 1298 | 1614 | 570 | 487 | 1838 |
| 57:自然-生命 | 772 | 1832 | 1435 | 668 | 1163 | 722 | 937 | 403 | 0 | 0 | 0 |
| 16:関係-時間 | 588 | 916 | 383 | 372 | 698 | 413 | 649 | 807 | 760 | 731 | 919 |
| 13:関係-様相 | 539 | 229 | 766 | 595 | 930 | 413 | 577 | 807 | 380 | 244 | 0 |
| 32:活動-芸術 | 441 | 1832 | 479 | 149 | 0 | 0 | 1442 | 0 | 0 | 81 | 0 |
| 17:関係-空間 | 343 | 0 | 191 | 595 | 116 | 413 | 433 | 0 | 570 | 325 | 0 |
| 51:自然-物質 | 184 | 0 | 287 | 149 | 582 | 0 | 216 | 404 | 0 | 81 | 0 |
| 50:自然-自然 | 184 | 0 | 191 | 297 | 0 | 206 | 433 | 0 | 0 | 81 | 0 |
| 19:関係-量 | 147 | 229 | 287 | 74 | 116 | 309 | 144 | 0 | 190 | 0 | 0 |
| 14:関係-力 | 12 | 229 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 56:自然-身体 | 12 | 0 | 96 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

表 14 では、記事ジャンルと動詞語義の関係性について、語義（中項目）単位にどのような記事で出現するかを評価している。

どの記事でも頻度の高い「為る（語彙素）」を含む行為（.34）の語義は、政治記事で特に出現しやすく、文化や社会記事では出現しにくい可能性がある。また、経済（.37）の語義は、経済そのものを扱う記事では用いられにくく、スポーツ記事と社会記事に関係が見られるという特徴がある。類（.11）の語義は、特に国際記事、次いで政治や経済記事において関わりが見られるが、特にスポーツ記事、そして文化記事においては関わりにくいという傾向がある。

動詞の語義は、記事ジャンルと直結するような語義分布でない場合もあり得るが、文脈によってある程度限定されている可能性が考えられる。すなわち、ジャンルにおいて頻度差の少ない上位頻度語（概ね形式的な意味の語や多義語）においても、いずれかの語義の頻度が高いという傾向が期待される。

表 14 ジャンル別 分類語彙表動詞「中項目」分布
(BCCWJ-WLSP・固定長 カイ二乗検定の標準化残差)

| | 総類 | 社会 | 政治 | 事件 | スポーツ | 文化 | 科学 | 国際 | 経済 | 労働 |
|----------|-----------|---------|---------|---------|---------|---------|---------|---------|---------|-------|
| 34:活動行為 | ⇒ -1.13 ↓ | -4.20 ↑ | 5.38 ↗ | 2.01 ↓ | -2.61 ↓ | -5.58 ⇒ | -1.15 ↑ | 4.77 ↑ | 2.68 ⇒ | 1.43 |
| 12:関係存在 | ⇒ -0.14 ⇒ | 0.05 ⇒ | 1.50 ⇒ | -0.27 ⇒ | -0.44 ⇒ | -1.85 ↑ | 2.73 ⇒ | -2.28 ⇒ | 1.95 ⇒ | -1.97 |
| 15:関係作用 | ⇒ 0.21 ⇒ | 1.03 ↓ | -3.18 ↘ | -2.09 ↑ | 2.78 ↑ | 2.85 ⇒ | 0.03 ⇒ | -2.21 ⇒ | -0.21 ⇒ | 0.39 |
| 30:活動心 | ⇒ -1.38 ↗ | 2.11 ⇒ | -1.35 ⇒ | -0.55 ⇒ | 1.18 ↑ | 3.25 ⇒ | -1.03 ⇒ | -2.15 ⇒ | -1.76 ⇒ | -0.60 |
| 31:活動言語 | ⇒ 1.38 ⇒ | 1.21 ⇒ | -1.82 ⇒ | -0.03 ⇒ | -1.42 ↗ | 2.43 ⇒ | -1.00 ⇒ | 0.63 ⇒ | -1.26 ⇒ | -0.22 |
| 11:関係類 | ⇒ -0.75 ⇒ | -0.93 ↑ | 4.13 ⇒ | 1.13 ↓ | -4.75 ↘ | -2.32 ↘ | -2.39 ↑ | 3.80 ⇒ | 0.99 ⇒ | 1.74 |
| 37:活動経済 | ⇒ -1.51 ⇒ | 1.05 ↓ | -3.06 ⇒ | -0.85 ↑ | 3.49 ⇒ | 0.31 ⇒ | -0.60 ⇒ | -0.72 ⇒ | 0.91 ⇒ | 0.63 |
| 33:活動生活 | ⇒ 0.24 ⇒ | 0.33 ⇒ | -1.90 ⇒ | 1.32 ⇒ | 1.10 ⇒ | -0.27 ↑ | 4.75 ⇒ | -0.46 ↓ | -2.87 ⇒ | 0.83 |
| 38:活動事業 | ⇒ 0.62 ↑ | 2.82 ↓ | -3.91 ⇒ | 0.66 ⇒ | -1.82 ↑ | 3.24 ⇒ | 0.93 ⇒ | -1.72 ⇒ | -0.27 ⇒ | -1.29 |
| 36:活動待遇 | ⇒ -0.12 ⇒ | 0.14 ⇒ | -0.15 ⇒ | -0.63 ⇒ | 0.99 ⇒ | -0.62 ⇒ | -0.47 ⇒ | 1.91 ⇒ | -0.57 ⇒ | -0.42 |
| 35:活動交わり | ↑ 2.78 ⇒ | -1.32 ↘ | -1.99 ⇒ | -1.71 ↑ | 6.93 ⇒ | -0.11 ⇒ | 0.16 ⇒ | -1.49 ↘ | -2.36 ⇒ | 0.48 |
| 57:自然生命 | ↑ 3.01 ↗ | 2.30 ⇒ | -0.49 ⇒ | 1.32 ⇒ | -0.25 ⇒ | 0.68 ⇒ | -0.82 ↘ | -2.07 ↓ | -3.14 ⇒ | -0.93 |
| 16:関係時間 | ⇒ 1.19 ⇒ | -1.11 ⇒ | -1.15 ⇒ | 0.39 ⇒ | -0.81 ⇒ | 0.25 ⇒ | 0.26 ⇒ | -1.72 ⇒ | 1.08 ⇒ | 0.44 |
| 13:関係様相 | ⇒ -0.76 ⇒ | 0.84 ⇒ | 0.29 ⇒ | 1.59 ⇒ | -0.62 ⇒ | 0.14 ⇒ | 0.39 ⇒ | -0.49 ⇒ | -1.29 ⇒ | -0.78 |
| 32:活動芸術 | ↑ 4.98 ⇒ | 0.01 ⇒ | -1.78 ↘ | -2.09 ↘ | -2.23 ↑ | 6.06 ⇒ | -1.14 ⇒ | -1.56 ⇒ | -1.88 ⇒ | -0.70 |
| 17:関係空間 | ⇒ -1.18 ⇒ | -1.03 ⇒ | 1.71 ⇒ | -1.23 ⇒ | 0.35 ⇒ | 0.56 ⇒ | -1.01 ⇒ | 0.95 ⇒ | 0.13 ⇒ | -0.62 |
| 51:自然物質 | ⇒ -0.86 ⇒ | 0.69 ⇒ | -0.34 ↑ | 2.82 ⇒ | -1.44 ⇒ | 0.27 ⇒ | 0.67 ⇒ | -1.01 ⇒ | -0.77 ⇒ | -0.45 |
| 50:自然自然 | ⇒ -0.86 ⇒ | -0.06 ⇒ | 1.05 ⇒ | -1.35 ⇒ | 0.14 ↗ | 2.31 ⇒ | -0.74 ⇒ | -1.01 ⇒ | -0.77 ⇒ | -0.45 |
| 19:関係量 | ⇒ 0.60 ⇒ | 1.12 ⇒ | -0.77 ⇒ | -0.27 ⇒ | 1.37 ⇒ | -0.07 ⇒ | -0.66 ⇒ | 0.28 ⇒ | -1.37 ⇒ | -0.41 |
| 14:関係力 | ↑ 4.50 ⇒ | -0.40 ⇒ | -0.45 ⇒ | -0.35 ⇒ | -0.37 ⇒ | -0.46 ⇒ | -0.19 ⇒ | -0.26 ⇒ | -0.39 ⇒ | -0.12 |
| 56:自然身体 | ⇒ -0.22 ↗ | 2.50 ⇒ | -0.45 ⇒ | -0.35 ⇒ | -0.37 ⇒ | -0.46 ⇒ | -0.19 ⇒ | -0.26 ⇒ | -0.39 ⇒ | -0.12 |

5. まとめ

BCCWJのPNに付与した記事情報を用い、ジャンルによる語彙の分布傾向を確認した。品詞の分布をはじめ、助詞種類の分布などにも異なる傾向が見られる。しかし、内容に応じた語の使用が期待されるものの、名詞に限定すれば分布傾向が見られやすいが、ジャンル横断的で不明瞭な動詞が多いという問題がある。そこで、BCCWJ-WLSPと重ね合わせることで、ジャンルによって語義の分布に傾向が見られることを確かめた。また、上位頻度の一般動詞についても、語義分布を見ることでジャンルによる傾向が確認できた。

今後、ジャンル情報を用いた語義の絞り込みの可能性が期待される。現在進行中の文脈情報の付与作業を進め、語義との関係を整理したい。

謝 辞

本研究は、国立国語研究所コーパス開発センター共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」、科研費基盤(C)「文体分析を目的としたコーパスの文書情報拡張及びその利用」による。

文 献

加藤祥, 浅原正幸, 山崎誠 (2019a) 「分類語彙表番号を付与した『現代日本語書き言葉均衡コーパス』の書籍・新聞・雑誌データ」『日本語の研究』15(2): 134-141.

加藤祥, 浅原正幸, 山崎誠 (2019b) 「『現代日本語書き言葉均衡コーパス』新聞・書籍・雑誌データの助動詞に対する用法情報付与」『日本語学会 2019 年度春季大会予稿集』, 161-166.

加藤祥, 森山奈々美, 浅原正幸 (2020) 「『現代日本語書き言葉均衡コーパス』新聞サブコーパスに対する新聞記事情報の付与」『日本語学会 2020 年度秋季大会予稿集』

加藤祥, 森山奈々美, 浅原正幸 (2021) 「『現代日本語書き言葉均衡コーパス』書籍サンプルのNDC 情報増補—NDC 情報を用いた随筆の抽出と文体調査—」『国立国語研究所論集』21:65-84.

国立国語研究所 (2004) 『分類語彙表増補改訂版』大日本図書

国立国語研究所『現代日本語書き言葉均衡コーパス』Version 1.1.

関連 URL

コーパス検索アプリケーション『中納言』

<https://chunagon.ninjal.ac.jp/>