# 『現代日本語書き言葉均衡コーパス』書籍サンプルの読み時間の分析 *

浅原正幸 †　　加藤 祥
国立国語研究所　目白大学

## 1. はじめに

本研究では、書籍のジャンル・文体が読み時間にどのような影響を与えるのかについて分析を行う。

『現代日本語書き言葉均衡コーパス（以降 BCCWJ）』(Maekawa et al. 2014) の白書・教科書・書籍サブコーパスについて、自己ペース読文法により収集された 100 人規模の読み時間データ BCCWJ-SPR2 が公開されている (浅原 2021)。また、BCCWJ の書籍サンプルには、日本十進分類法 (NDC)(日本図書館協会分類委員会 2018) による書籍の分野情報 BCCWJ-NDC (加藤ほか 2021) が付与されている。さらに『BCCWJ 図書館サブコーパスの文体情報』(柏野 2013) に準じた文体情報（専門度・硬度・くだけ度・語りかけ性度・客観度）を今回新たに付与した。

BCCWJ の書籍・生産実態サブコーパスコアデータ（以降 PB コアデータ）83 サンプルについて、浅原 (2021) にならい、文節の読み時間を分野情報・文体情報を固定要因とした一般化線形混合モデルにより検討を行った。結果、分野情報においては「7 類 芸術・美術」が、文体情報においては「語りかけ性度」が読み時間に影響を及すことがわかった。

以下、2 節において、分析に用いたデータの情報と統計処理手法について示す。3 節に結果を示し、4 節に考察を示す。5 節にまとめと今後の研究の方向性について示す。

## 2. 方法

### 2.1 BCCWJ-NDC: BCCWJ 書籍サンプルの分野情報

BCCWJ-NDC(1) (加藤ほか 2021) は、BCCWJ の書籍サンプルに付与されている日本十進分類法 (NDC) (日本図書館協会分類委員会 2018) を国立国会図書館サーチ(2) に基づき、修正・増補したものである。書籍の分野情報として NDC の類目（1 次区分：NDC の上位一桁）を用いる。表 1 に今回分析に用いた PB コアデータの分野情報の分布を示す。

| 0 類 総記 | 1 類 哲学 | 2 類 歴史 | 3 類 社会科学 | 4 類 自然科学 | |
|---|---|---|---|---|---|
| 6 | 6 | 7 | 18 | 6 | |
| 5 類 技術・工学・工業 | 6 類 産業 | 7 類 芸術・美術 | 8 類 言語 | 9 類 文学 | 分類なし |
| 8 | 6 | 4 | 2 | 19 | 1 |

表 1　BCCWJ-NDC: PB コアデータ分野情報の分布（数値は分析に用いたサンプル数）

### 2.2 『BCCWJ 図書館サブコーパスの文体情報』に準じた文体情報付与

『BCCWJ 図書館サブコーパスの文体情報』(3) (柏野 2013) は、BCCWJ の書籍・図書館サブコーパスの文体情報データである。専門度・硬度・くだけ度・語りかけ性度・客観度の情報が人手により整備

* 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および科研費 JP17H00917, JP18H05521, 18K00634 によるものです。
† masayu-a@ninjal.ac.jp
(1) https://github.com/masayu-a/BCCWJ-NDC
(2) https://iss.ndl.go.jp/information/api/
(3) http://doi.org/10.15084/00003109

されている。しかしながら、BCCWJ の PB コアデータは、図書館サブコーパス (LB) からではなく生産実態サブコーパス (PB) からサンプリングされており、上記データは利用できない。今回新たに同様の文体情報を　PB コアデータに付与した。以下に各文体情報の定義を示すとともに、PB コアデータの文体情報の分布を示す。

専門度は、テキストを理解するうえでの「高度な知識の必要性」の有無に基づき判定し、想定読者のスケールで測るもので、「1 専門家向き」から「5 小学生・幼児向き」の 5 段階に分類されている。表 2 に専門度の分布を示す。

| 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 | (未定義) |
|---|---|---|---|---|---|
| 1 | 13 | 60 | 2 | 6 | (1) |

表 2　専門度の分布（数値は分析に用いたサンプル数）

硬度は、テキストの文体の形式性・親疎性をとらえるために「硬いか軟らかいか」を判断したものである。かしこまっている堅苦しい感じの「1 とても硬い」から、かしこまっていない親しみやすい感じの「4 とても軟らかい」まで 4 段階に分類されている。表 3 に硬度の分布を示す。

| 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 | (未定義) |
|---|---|---|---|---|
| 2 | 35 | 37 | 8 | (1) |

表 3　硬度の分布（数値は分析に用いたサンプル数）

くだけ度は、テキストの形式性・親疎性をとらえるためのもう一つの指標で、「くだけているか」を判断したものである。「くだけているか」の逆の概念として「改まっているか」を想定するが、「改まっている」程度を判断しにくいとして、「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（＝改まっている）」の 3 段階に分類されている。表 4 にくだけ度の分布を示す。

| 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 | (未定義) |
|---|---|---|---|
| 4 | 38 | 40 | (1) |

表 4　くだけ度の分布（数値は分析に用いたサンプル数）

語りかけ性度は、テキストの文体の口語性を問うものである。「あなた」や「みなさん」などの呼びかけ表現や、「でしょう」「ではないでしょうか」といった問いかけや相づちを求めるような文末表現など、読み手に直接的に語りかけているような表現があるものを、「語りかけ性がある」ものと定義されている。「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」の 3 段階に分類されている。表 5 に語りかけ性度の分布を示す。

| 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 | (未定義) |
|---|---|---|---|
| 14 | 27 | 41 | (1) |

表 5　語りかけ性度の分布（数値は分析に用いたサンプル数）

客観度は、小説以外のテキストの書き手の態度が「客観的」か「主観的」かの区別を付与したものである。「1 とても客観的」から「4 とても主観的」の 4 段階の分類が設定されているが、PB コアデータには「1 とても客観的」は含まれなかった。表 6 に客観度の分布を示す。

| 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 | (未定義) |
|---|---|---|---|
| 19 | 31 | 11 | (22) |

表 6　客観度の分布（数値は分析に用いたサンプル数）

なお、PB コアデータの中に「対話」が含まれており、これらは文体指標の付与対象外とした。さらに客観度においては、小説は付与対象外となっている。これらは表中「(未定義)」としているが、次に示す読み時間分析の対象外とした。

### 2.3　BCCWJ-SPR2: BCCWJ に対する読み時間情報付与

| レジスタ | サンプル | 文 | 文節 |
|---|---|---|---|
| PB　書籍 | 83 | 10,075 | 84,736 |
| | (平均) | 121.4 | 1,020.9 |

表 7　刺激文の統計

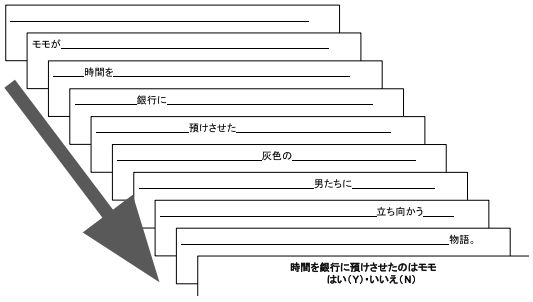


図 1　移動窓方式による自己ペース読文法

| 摘要 | PB |
|---|---|
| (1) 不同意 | 199 |
| (2) 重複回答 | 64 |
| (3) <150 or 2,000< | 2,875 |
| (4) 誤答 | 2,090 |
| 適切なサンプル | 11,325 |
| 合計 | 16,553 |

表 8　不適切なデータの排除（サンプル・人数）

| 摘要 | PB |
|---|---|
| 適切なサンプル | 11,309,783 |
| (5) <100 or 3,000< | 540,403 |
| 分析対象 | 10,769,380 |

表 9　不適切なデータの排除（データポイント数）

BCCWJ-SPR2 [(4)] (浅原 2021) は、BCCWJ の白書・教科書・書籍を刺激文とした読み時間データである。本研究では書籍 (PB) コアデータ 83 サンプルを用いた。文・文節数の情報を表 7 に示す。

この刺激文に対して、自己ペース読文法を用いて、読み時間を収集した。自己ペース読文法は、移動窓を用いて疑似的な視線移動環境を設定しながら、部分的に呈示された単語や文節の表示時間により読み時間を計測する手法である。図 1 に例を示す。スペースキーを押すたびに逐次的に文節が表示され、スペースキーを押した時間間隔をミリ秒単位で記録することで読み時間を計測する。1 文章を読んだ最後に「はい」「いいえ」の 2 択で答える内容把握の設問を設定した。ブラウザ上で自己ペース読文法による被験者実験を行うため、ibexfarm [(5)] を用いた。日本語対応は中谷 (2019) に倣った。

被験者は Yahoo! クラウドソーシング[(6)]により募集した。2020 年 10 月に、あらかじめ「単語親密度」の評定実験 (浅原 2020) を行い、単語親密度の回答において、回答の分散が大きい 2,092 人を対象に募集した。適切に作業に取り組む方をあらかじめ選別するために行った。また、単語親密度推定を行うことにより、被験者毎の語彙力も推定できるため、今後、語彙力が読み時間に与える影響を検討する。被験者は、1 サンプルあたり 200 人募集した。内訳は、内容確認質問の正解が YES である実験 100 人、

(4) https://github.com/masayu-a/BCCWJ-SPR2
(5) https://spellout.net/ibexfarm/
(6) https://crowdsourcing.yahoo.co.jp/

NO である実験 100 人である。

本データの収集はオンラインで実施したため、対面で収集した際よりもデータの品質が悪いことが想定される。そこで、不適切なデータの排除を試みた。まず、(1) 実験開始時に実験データの取り扱いおよび謝金の支払方法に同意していないものを排除した。次に、(2) 同じ被験者が同じサンプルを複数回実施した場合、2 回目以降の試行データを削除した。さらに、(3) サンプル単位の平均読み時間が 150ms 未満もしくは 2,000ms 超過のもの、(4)YES/NO 質問を誤答しているものを排除した。表 8 にサンプル・人単位の不適切なデータ排除件数を示す。

最後にデータポイント単位に不適切なデータを排除した。表 9 の「適切なサンプル」の行に、適切なサンプルに含まれるデータポイント数を示す。これらから、さらに (5) 100ms 未満もしくは 3,000ms 超過のものを排除した。結果、表 8 中「分析対象」に示す件数を適切なデータポイントとして扱うことにした。基準として、英語における Natural Stories Corpus (NSR) (Futrell et al. 2018) が 100ms 未満もしくは 3,000ms 超過のデータポイントを削除しているのを参考にした。NSR のデータポイント数が 848,857 と比しても、大規模な読み時間データが構築できたといえる。

### 2.4 分析手法

本稿では、頻度主義的な分析手法 (Baayen 2008, Vasishth et al. To Appear) による結果を示す。文節読み時間 `SPR_reading_time` の検討を一般化線形混合モデル (R(R Core Team 2020), lme4(Bates et al. 2015), stargazer(Hlavac 2018)) を用いて行う。

固定効果として、**[分野情報 or 文体情報]** を一次式でモデル化した。**[分野情報]** は NDC の値を因子化したものを用い、**[文体情報]** は程度の数値を用いた。また、文体情報の分析にあたっては、小説が情報付与対象外である「客観度」のみ別実験でモデル化した。呈示順の情報である `SPR_sentence_ID` (実験時の文 ID)・`SPR_bunsetsu_ID` (実験時の文節 ID) と、表層形の文字数である `SPR_word_length` を考慮した。また、BCCWJ-DepPara (浅原・松本 2018) の情報に基づいた当該文節の係り受けの数 `DepPara_depnum` を考慮した。同一の被験者が複数のサンプルを読んだ際の試行順序 `SPR_trial` も固定効果としてモデル化した。また、被験者間の個人差をモデル化するために `SPR_subj_ID` (被験者 ID) をランダム効果として考慮した。サンプル間の個体差をモデル化するために `BCCWJ_Sample_ID` (BCCWJ のサンプル ID) もランダム効果として考慮した。分析式は次のとおり：

```
SPR_reading_time ~ SPR_sentence_ID + SPR_bunsetsu_ID + SPR_word_length
 + SPR_trial + DepPara_depnum + BCCWJ_OT_school_type + [分野情報 or 文体情報]
  +(1|SPR_subj_ID_factor)+(1|BCCWJ_Sample_ID).
```

次節では、一度モデルを推定したうえで、3SD よりも外側の値のデータポイントを排除し、再推定を行った結果を示す。

## 3. 結果

分析は、浅原 (2021) に倣い、書籍各サンプルの文節の読み時間を、分野情報・文体情報を固定要因とした、一般化線形混合モデルにより行った。呈示順序（文 ID・文節 ID）・文節の文字列長・文節に対する係り受けの数・被験者の実験参加回数を基本的な固定要因とし、被験者 ID とサンプル ID をランダム要因とした分析を行う。一度回帰をおこなったうえで、3 標準偏差より外側のデータポイントを排除したものの結果を示す。

表 10 に分野情報 (NDC) に対して推定された固定要因の結果を示す。表中カッコなしの値が推定係数で、カッコつきの値が標準誤差である。まず、呈示順が進むにつれて読み時間が短くなる効果と文字長が長くなるにつれて読み時間が長くなる効果が確認された。係り受けの数が多いものほど予測が効く

| | | | | | |
|---|---|---|---|---|---|
| BCCWJ_NDC_1 | 7.538 | (11.109) | SPR_sentence_ID | -0.194*** | (0.001) |
| BCCWJ_NDC_2 | 9.878 | (10.704) | SPR_bunsetsu_ID | -1.114*** | (0.010) |
| BCCWJ_NDC_3 | -9.300 | (9.070) | SPR_word_length | 13.801*** | (0.023) |
| BCCWJ_NDC_4 | -12.776 | (11.107) | DepPara_depnum | -10.343*** | (0.055) |
| BCCWJ_NDC_5 | -7.432 | (10.391) | SPR_trial | 0.409*** | (0.010) |
| BCCWJ_NDC_6 | -9.253 | (11.108) | Constant | 306.933*** | (9.667) |
| BCCWJ_NDC_7 | -24.697** | (12.419) | Observations | 10,870,175 | |
| BCCWJ_NDC_8 | 4.716 | (15.711) | Log Likelihood | -71,170,562 | |
| BCCWJ_NDC_9 | -7.332 | (9.010) | Note: | *p<0.1; **p<0.05; ***p<0.01 | |
| BCCWJ_NDC_undef | -12.147 | (20.778) | | | |

表 10　分野情報に対して推定された固定要因 (NDC 0 総記 が基準)

| | | | | | |
|---|---|---|---|---|---|
| 専門度 | -0.046 | (3.497) | SPR_sentence_ID | -0.194*** | (0.001) |
| 硬度 | -2.256 | (4.262) | SPR_bunsetsu_ID | -1.112*** | (0.010) |
| くだけ度 | -5.939 | (4.263) | SPR_word_length | 13.797*** | (0.023) |
| 語りかけ性度 | -5.557* | (3.272) | DepPara_depnum | -10.373*** | (0.055) |
| Observations | 10,751,178 | | SPR_trial | 0.408*** | (0.010) |
| Log Likelihood | -70,384,828 | | Constant | 304.860*** | (21.877) |

表 11　文体情報に対して推定された固定要因 (客観度以外)

| | | | | | |
|---|---|---|---|---|---|
| 客観度 | 0.267 | (5.413) | SPR_sentence_ID | -0.188*** | (0.001) |
| Constant | 279.607*** | (16.965) | SPR_bunsetsu_ID | -1.312*** | (0.012) |
| | | | SPR_word_length | 13.699*** | (0.027) |
| Observations | 7,627,734 | | DepPara_depnum | -10.159*** | (0.066) |
| Log Likelihood | -49,961,568 | | SPR_trial | 1.050*** | (0.014) |

表 12　文体情報に対して推定された固定要因 (客観度)

ために読み時間が短くなることも確認された。また試行順が進むにつれて読み時間が長くなる効果も確認された。分野情報においては、「7. 芸術・美術」の読み時間が短くなること ($p < 0.05$) が確認された。

表 11, 12 に文体情報に対して推定された固定要因の結果を示す。文体情報の分析においては、専門度 (1-5)・硬度 (1-4)・くだけ度 (1-3)・語りかけ性度 (1-3)・客観度 (1-4) の各指標の値を固定要因として分析を行ったところ、語りかけ性がないと読み時間が短くなる（つまり、語りかけ性があると読み時間が長くなる）有意傾向 ($p < 0.1$) が確認された。

## 4. 考察

分野情報においては「7 芸術・美術」の読み時間が短くなることが確認されたが、主に 76 音楽・78 スポーツのサンプルから構成されていた。いずれも「3 一般向き」の「3 特に語りかけ性がない」サンプルであった。

文体情報においては、語りかけ性があると読み時間が長くなる有意傾向 ($p < 0.1$) のみが確認された。浅原 (2019) では、新聞記事における述語項構造と視線走査法の読み時間の対照分析において、主語のゼロ代名詞が外界 2 人称を指す場合に述語要素において、1 回目の読み時間は有意ではないが長く、2

回目の読み時間 (Second Pass Time) が有意に短くなる傾向を確認している。これらのことから、語りかけ性や読者への言及など読者自身に強く注意を向ける表現であり、1 回目の読み時間が長くなり、強い印象が残ることで、2 回目の読み時間が短くなる可能性が確認された。

## 5. おわりに

本研究では、書籍のテキストに対して、分野・文体情報が読み時間に対してどのような影響を及ぼすのかについて、計量的な分析を行った。分野情報においては「7 芸術・美術」の読み時間が短くなることが確認された。文体情報においては語りかけ性があると読み時間の長くなることが確認された。語りかけ性については、過去の研究 (浅原 2021) において、主語のゼロ代名詞が外界 2 人称を指す場合に、述語要素において 1 回目の読み時間が長くなる傾向が観察されたことに整合性があった。語りかける表現が、読み手の読み時間に影響を与えることが示唆された。今後、書籍サンプルにおける述語項構造・共参照アノテーションと読み時間の対照分析を進め、より詳細な検討を行う。分析に用いたデータは `https://github.com/masayu-a/BCCWJ-SPR2/tree/master/analysis/`日本語学会春季大会 `2021/` を参照されたい。

## 文 献


Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced Corpus of Contemporary Written Japanese.” *Language Resources and Evaluation*, 48, pp. 345–371.

浅原正幸 (2021).「クラウドソーシングによる大規模読み時間データ収集」 言語処理学会第 27 回年次大会発表論文集.

日本図書館協会分類委員会 (2018).『日本十進分類法進呈 10 版』 日本図書館協会.

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, 21, pp. (To Appear).

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, pp. 44–53.

中谷健太郎 (2019).「自己ペース読文課題を使った実験：ウェブ編」 中谷健太郎 (編)『パソコンがあればできる！ことばの実験研究の方法 容認性調査、読文・産出実験からコーパスまで』 ひつじ書房, 東京, 第 4 章 pp. 81–106.

浅原正幸 (2020).「Bayesian Linear Mixed Model による単語親密度推定と位相情報付与」 自然言語処理, 27:1, pp. 133–150.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko (2018). “The Natural Stories Corpus.” *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

R. Harald Baayen (2008). *Analyzing Linguistic Data: A practical Introduction to Statistics using R.*: Cambridge University Press.

Shravan Vasishth, Daniel Schad, Audrey Bürki, and Reinhold Kliegl (To Appear). *Linear Mixed Models in Linguistics and Psychology: A Comprehensive Introduction.*

R Core Team (2020). *R: A Language and Environment for Statistical Computing.*, R Foundation for Statistical Computing Vienna, Austria.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker (2015). “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*, 67:1, pp. 1–48.

Marek Hlavac (2018). *stargazer: Well-Formatted Regression and Summary Statistics Tables.*, Central European Labour Studies Institute (CELSI) Bratislava, Slovakia. R package version 5.2.2

浅原正幸・松本裕治 (2018).「『現代日本語書き言葉均衡コーパス』に対する文節係り受け・並列構造アノテーション」 自然言語処理, 25:4, pp. 331–356.

浅原正幸 (2019).「読み時間と述語項構造・共参照情報について」 言語処理学会第 25 回年次大会発表論文集, pp. 249–253.