「昭和・平成書き言葉コーパス」の語彙統計情報の公開

相田 太一 [1,2]　近藤 明日子 [3]　小木曽 智信 [1]
[1] 国立国語研究所　[2] 東京都立大学　[3] 東京大学
aida-taichi@ed.tmu.ac.jp　akondo@l.u-tokyo.ac.jp　togiso@ninjal.ac.jp

概要

本研究では、1933 年から 2013 年までの間を 8 年おきにカバーする「昭和・平成書き言葉コーパス」に関して、*n*-gram 頻度形式と SVMlight 形式の共起情報を公開した[1)]。以前公開された統計情報に比べて昭和・平成の時代に限定されているが、今回は雑誌に加えて新たに**ベストセラー書籍**と**新聞**のレジスター（ドメイン）を追加し、複数の時期だけでなく複数のレジスターを跨いだ分析が可能になった。本稿では、**自然言語処理**と**日本語学**の二つの側面において、この統計情報を用いた研究の可能性を示す。

1　はじめに

2023 年に「昭和・平成書き言葉コーパス」[2)]が公開された [1]。このコーパスは、1933 年から 2013 年までの 80 年間を 8 年おきに雑誌・新聞・ベストセラー書籍を収録した通時的なコーパスであり、「日本語歴史コーパス」[3)]と「現代日本語書き言葉均衡コーパス」[2] などの現代語コーパスの間をつなぐものである。これにより、現代日本語がどのような変化を経て成立してきたのかを計量的に探るための基礎となるデータが整備されたことになる。

自然言語処理の分野では近年、通時的なコーパスを用いた研究が注目を集めている。特に、単語の通時的な意味変化を検出・分析することは、言語学・社会学だけでなく情報検索や大規模言語モデルの効率的な追加訓練に有用 [3, 4] であり、盛んに取り組まれている。歴史的な言語変化に関するワークショップである LChange[4)] も 2023 年で 4 回目を迎え、関心の高まりを見せている。しかし、様々な言語で歴史的な言語変化に関する研究が進められている一方で、日本語では利用可能な大規模コーパスが存在しないため、日本語におけるこの分野の研究は十分な進展を遂げていない。

そこで、本研究では、構築された「昭和・平成書き言葉コーパス」をもとに、研究利用可能なデータを公開する。上記のコーパスは権利上、生のテキストデータを配布できないため、今回は *n*-gram 頻度形式と SVMlight 形式の共起情報を提供する[5)]。

2　関連研究

幅広い年代の文書が含まれる通時的なコーパスは、様々な言語で作成・公開されている。特に Google Books Ngram コーパス [5] はその代表例であり、1800 年代から 2000 年代にかけて英語・フランス語・ドイツ語・イタリア語・スペイン語・中国語・ヘブライ語・スペイン語の *n*-gram 頻度データを提供している[6)]。その後、幅広い年代を含むコーパスの生データが英語 [6, 7]・ラテン語 [8]・ドイツ語[7)8)]・スウェーデン語[9)]で整備・公開された。これらは近年開催された単語の通時的な意味変化を捉える共有タスク SemEval-2020 Task 1 [9] でも採用されている。

しかし、日本語ではいまだに利用可能である通時的なコーパスが不足している。以前我々が「近現代雑誌通時コーパス」の一部の統計データを公開 [10] したが、作成途中のコーパスを用いていたため、十分とはいえない。さらに、雑誌だけに限られていたため、レジスター間の比較は困難である。

3　データの構築

3.1　雑誌データ

「昭和・平成書き言葉コーパス」は昭和期から平成期までの書き言葉の通時的な変化を研究することを目的として構築されたコーパスである。この期間

1) https://github.com/a1da4/shc-data
2) https://clrd.ninjal.ac.jp/shc/
3) https://clrd.ninjal.ac.jp/chj/
4) https://languagechange.org/events/
5) このデータは CC BY-SA 4.0 ライセンスの下で公開される。
6) https://storage.googleapis.com/books/ngrams/books/datasetsv3.html
7) https://www.deutschestextarchiv.de/
8) https://zefys.staatsbibliothek-berlin.de/
9) https://spraakbanken.gu.se/en/resources/kubhist2

表 1　今回対象とした「昭和・平成書き言葉コーパス」の各時期・各レジスターにおける延べ語数。

刊行年	延べ語数			計
	雑誌	書籍	新聞	
1933	373 万	20 万	14 万	407 万
1941	274 万	26 万	16 万	316 万
1949	114 万	31 万	12 万	157 万
1957	353 万	52 万	11 万	416 万
1965	231 万	43 万	32 万	306 万
1973	266 万	40 万	44 万	350 万
1981	303 万	35 万	42 万	380 万
1989	314 万	36 万	37 万	387 万
1997	290 万	35 万	32 万	357 万
2005	288 万	37 万	28 万	353 万
2013	307 万	41 万	27 万	375 万
計	3113 万	396 万	295 万	3804 万

表 2　単語の *n*-gram 頻度形式と語彙素 ID の SVMlight 形式で出力した例。

出力形式	出力例
n-gram 頻度（$n = 5$）	た の で ある 。　1697
SVMlight	1571　0:13 1:7 6:10 ...

に刊行された雑誌から 8 年おきに各年代を代表する雑誌・ベストセラー書籍・新聞記事を選定し、テキストを収録している。このコーパスは現在、オンライン検索ツール「中納言」[10]による検索サービスが提供されており、前後 30 語の範囲で前後文脈が参照可能となっている。

「昭和・平成書き言葉コーパス」に収録されているデータの刊行年と各レジスターにおける述べ語数を表 1 に示す。

3.2　データの加工

今回対象にしたコーパスは著作権保護の観点から生データを公開できないため、前回 [10] と同様に、*n*-gram 頻度と SVMlight[11] の 2 つの形式に加工し、元のコーパスを再現できない形で公開する。各形式で加工した例を表 2 に示す。

***n*-gram 頻度**　Google Books Ngram と同様の形式であり、各文書における単語および語彙素 ID の *n*-gram とその頻度を出力した。語彙素 ID は、形態素解析用辞書 UniDic [11, 12] に収録され、辞書見出しに相当する「語彙素」ごとに与えられた識別子である。辞書アーカイブ（現代語用 UniDic のフルパッケージ）[12]の語彙素一覧（lex.csv）に含まれる語彙素 ID 列に対応付けることで、語彙素（代表表記）はもちろんのこと、語彙素読みや語種、品詞の上位概念を示す類などの情報を取り出すことが可能である。今回も単語/語彙素 ID 1-gram（単語/語彙素 ID の頻度）から 5-gram（連続する 5 つの単語/語彙素 ID の頻度）まで集計した。

SVMlight　ここでは、単語を語彙素 ID に変換し、語彙素 ID 同士の共起情報を SVMlight の形式で出力した。この形式では、対象の語彙素 ID（例：1571）に対して、共起する語彙素 ID（例：0, 1, 6）とその共起回数（例：13, 7, 10）が

　　共起する ID：対象の ID との共起回数

の形式で表現されている（表 2）。SVMlight は全て ID で記述されるため、*n*-gram 頻度形式と比べると解釈が難しい。しかし、一行について一つの対象 ID に関する共起情報が出力されており、共起回数 1 回以上の共起 ID に関する情報だけが表示されるため、スパースな情報の表現に優れている。今回は、各時期・各レジスターで頻度が 20 回以上の語彙素 ID で語彙を形成し、前後 4 単語で共起頻度を集計した。

3.3　データの公開

公開データは UTF-8 形式のテキストファイルであり、単語 *n*-gram（surface_ngram）、語彙素 ID *n*-gram（lemmaID_ngram）、語彙素 ID SVMlight（lemmaID_svmlight）の 3 つのフォルダに分け、zip 形式で圧縮した。各形式のフォルダには、さらにレジスターごとに雑誌（magazine）、書籍（bestseller）、新聞（newspaper）の 3 つのフォルダに分かれており、刊行年ごとの統計情報が保存されている。

4　使用用途

ここでは、今回作成した統計情報から期待される研究の可能性について、自然言語処理と日本語学の二つの側面から言及する。

4.1　自然言語処理

今回の統計情報により、**通時的なコーパスが揃っている英語やドイツ語、中国語などで盛んに行われていた単語の通時的な意味変化に関する研究を日本語でも行えるようになる**。前回公開した統計情

10）https://chunagon.ninjal.ac.jp/
11）https://www.cs.cornell.edu/people/tj/svm_light/
12）https://clrd.ninjal.ac.jp/unidic/back_number.html

表 3　自然言語処理における応用例。対象のデータセットを昭和 (1933-1981) と平成 (1989-2013) に分割し、それぞれの時期で単語分散表現を学習した後、対象単語に対する近傍単語をそれぞれの時期で算出した。

単語	時代	近傍単語 (cos 類似度)
操作	1933-1981	算定 (0.816), 實行 (0.787), 是正 (0.786), 充實 (0.777), 取り扱い (0.776), 吊り上げ (0.775), 遷延 (0.773), 測定 (0.770), 騰貴 (0.768), 実効 (0.766)
	1989-2013	接続 (0.795), 遠隔 (0.780), 制御 (0.765), 入力 (0.764), 保管 (0.758), 誘導 (0.758), 入手 (0.757), 表示 (0.753), 改良 (0.748), 診断 (0.747)
通信	1933-1981	無線 (0.768), 国営 (0.766), 東北 (0.746), 製糖 (0.742), 協會 (0.742), タス (0.739), 電氣 (0.731), 紡績 (0.717), 工学 (0.715), 文藝 (0.712)
	1989-2013	化学 (0.763), 印刷 (0.759), 航空 (0.748), 製薬 (0.732), メディア (0.732), 系列 (0.732), 流通 (0.718), 新華 (0.712), 住友 (0.706), 各社 (0.704)
流出	1933-1981	含有 (0.821), 流入 (0.819), 飲料 (0.808), 收 (0.807), 滞船 (0.802), 鉱石 (0.799), 小口 (0.799), 添加 (0.797), 灯油 (0.787), 多量 (0.782)
	1989-2013	高騰 (0.835), 放出 (0.814), 頭打ち (0.811), 蒸気 (0.798), 浄化 (0.795), 汚染 (0.789), 除染 (0.786), 制御 (0.786), 補給 (0.784), 供給 (0.783)

報 [10] と比べて時代の範囲は短くなるが、今回は完成したコーパスから算出した統計情報であり、雑誌だけでなく書籍と新聞が追加されているため、複数の時期・レジスターを横断した研究が可能になる。

例として、単語 5-gram 形式の全てのレジスターを結合し、昭和（1933-1981）と平成（1989-2013）の二つの時期に分けて分析を行う。各時期で単語分散表現を学習した後、同じ単語に関して近傍単語を計算して、時期間で比較した。学習では、次元数を 50、文脈窓幅を前後 4 単語、5 回以上出現する単語を対象とし、Gensim[13] のパッケージを使用して単語分散表現 Skip-Gram with Negative Sampling を訓練した[14]。表 3 より、「操作」は平成で「接続」や「遠隔」といった近傍単語が出現しており、技術革新による変化が現れている。また、「通信」は昭和では軍事関係の近傍単語が多かったが、平成では「メディア」や「流通」といった近傍単語が出現しており、一般化した様子を捉えている。さらに、「流出」は平成で「汚染」「除染」といった近傍単語が現れており、2011 年の震災の影響を捉えていることを示す。今回はレジスターを結合して昭和と平成の二つの時期で分析を行ったが、あるレジスターにおける時間変化やレジスター間の違い、複数の時期で分析など、様々な用途での活用が期待できる。

13）https://radimrehurek.com/gensim/models/word2vec.html

14）今回訓練したモデルは統計情報と合わせて https://github.com/a1da4/shc-data で公開する。昭和・平成の各時期において、全てのレジスターを結合して訓練したモデルだけでなく、各レジスターで訓練したモデルも公開する。

また、この統計情報を用いることで、**他言語で効果的だった手法が日本語でも効果的に動作するのかといった検証**が可能になる。SemEval-2020 Task 1 [9] では英語・ラテン語・ドイツ語・スウェーデン語の 4 言語で意味変化検出の性能評価が行われた。そのほかにも、ロシア語 [13] や中国語 [14] でも評価用データの構築が進み、様々な言語での性能評価が可能になった。しかし、日本語では大規模な訓練用の通時的なコーパスが存在しなかったため、日本語における意味変化検出の性能を評価することが困難であった。そこで、凌ら [15] が構築した昭和と平成の時代間で意味変化検出の性能を評価する単語リストと今回作成した統計情報を合わせることで、日本語における性能評価が可能となる。

4.2　日本語学

日本語学では、コーパスから作成した単語 *n*-gram を用いて、複合辞やコロケーション表現といった特定の意味・機能を有する単語連続を抽出する研究 [16, 17]、接続詞を抽出し明治・大正期の通時的変化を考察する研究 [18]、室町時代の 2 作品間を比較し表現の共通点・差異を考察する研究 [19] 等が行われてきた。同様に今回の統計情報を利用することで、**昭和・平成期の書き言葉の通時的変化やレジスター間の差異の計量的研究が可能**になる。

例として、雑誌・新聞の語彙素 ID の 4-gram のデータを使用して、4-gram とレジスター・刊行年と

表 4　コレスポンデンス分析第 1 次元スコア。4-gram 列の語彙素 ID は語彙素に置き換えて示す。レジスター・刊行年列は m が雑誌、n が新聞、数字が刊行年を示す。

4-gram	スコア
居る の だ 有る	-1.95
事 と 成る た	-1.94
た 物 だ 有る	-1.81
た の だ 有る	-1.31
と 言う の だ	-0.88
て 居る の だ	-0.66
と 為る て は	-0.59
ない ば 成る ない	-0.55
て 居る た が	-0.22
と 成る て 居る	0.02
と 為る て 居る	0.10
と 言う の は	0.27
に 成る て 居る	0.31
の だ は 無い	0.46
て 居る た の	0.46
為る れる て 居る	0.48
為る て 居る た	0.56
れる て 居る た	0.61
に つく て は	1.57
て 居る ます た	2.56
為る て 居る ます	2.56

レジスター・刊行年	スコア
m1941	-1.51
n1941	-1.35
m1933	-1.17
n1933	-1.01
m1949	-0.98
m1957	-0.78
m1965	-0.28
m1973	-0.23
n1949	-0.07
m1981	-0.03
n1965	0.16
n1973	0.57
m1989	0.66
n1957	0.68
m1997	0.68
n1981	0.94
m2005	1.08
n1997	1.13
n1989	1.14
n2005	1.51
n2013	1.53
m2013	1.86

表 5　コレスポンデンス分析第 2 次元スコア。4-gram 列の語彙素 ID は語彙素に置き換えて示す。レジスター・刊行年列は m が雑誌、n が新聞、数字が刊行年を示す。

4-gram	スコア
て 居る ます た	-2.63
為る て 居る ます	-2.20
居る の だ 有る	-0.84
た の だ 有る	-0.80
て 居る た の	-0.71
事 と 成る た	-0.47
と 言う の だ	-0.37
て 居る の だ	-0.35
た 物 だ 有る	-0.32
の だ は 無い	-0.19
と 言う の は	-0.18
為る て 居る た	0.23
れる て 居る た	0.25
ない ば 成る ない	0.44
に 成る て 居る	0.56
と 為る て は	0.65
為る れる て 居る	0.70
て 居る た が	0.98
と 為る て 居る	1.22
と 成る て 居る	2.09
に つく て は	3.28

レジスター・刊行年	スコア
m2013	-1.42
m2005	-0.65
m1941	-0.59
m1933	-0.48
m1989	-0.20
m1957	-0.10
m1949	0.01
m1965	0.12
m1997	0.38
m1973	0.39
m1981	0.52
n1941	0.68
n1933	0.76
n1965	2.88
n2013	3.04
n2005	3.43
n1973	3.54
n1981	3.57
n1989	3.62
n1949	4.05
n1957	4.32
n1997	4.56

の対応関係を分析する。数詞・記号類を含む 4-gram は除いたうえで各レジスター・刊行年での頻度上位 5 位までの 4-gram を抽出すると、異なりで計 21 種が得られた。この 4-gram21 種とレジスター・刊行年の組み合わせ 22 種の粗頻度[15]のクロス表を作成し、コレスポンデンス分析を行った。4-gram とレジスター・刊行年それぞれの第 1 次元（寄与率 60.4%）スコアを表 4 に、第 2 次元（寄与率 17.1%）スコアを表 5 に示す。

表 4 のレジスター・刊行年列から、第 1 次元はおおよそ刊行年の新旧の区分を示していると考えられる。昭和初期は文末辞「である」を含む語連続、平成期は文末辞「ます」を含む語連続により特徴づけられることが分かる。一方、表 5 のレジスター・刊行年列から、第 2 次元は雑誌と新聞のレジスターの区分を示していると考えられる。雑誌は「である」「ます」を含む語連続、新聞は「については」「となっている」「としている」といった語連続により特徴づけられることが分かる。加えて、第 1 次元の寄与率は 60.4%と第 2 次元の 17.1%と比較して非常に高く、刊行年は語彙素 ID の 4gram と特に重要な対応関係をもつことが分かる。このように *n*-gram 頻度情報を使用することで、昭和・平成期の日本語の書き言葉における、特定の機能を有する語連続をはじめとする語彙と使用時期・レジスターとの関係性を明らかにすることができる。

15）　ただし今回の分析では、各レジスター・刊行年において頻度 5 以上の 4-gram を集計対象とした。

5　おわりに

本研究では、「昭和・平成書き言葉コーパス」の公開に伴い、コーパスの統計情報を公開した。前回公開した統計情報と比べて昭和・平成の時代に限定されるが、今回は新たにベストセラー書籍と新聞が追加され、複数の時代・レジスター（ドメイン）を跨いだ比較・分析が可能となった。また、本稿では自然言語処理・日本語学の二つの側面からこのデータを用いた研究の可能性を示した。今回発表したデータを皮切りに、日本語における言語変化の研究が盛んになることを期待している。

謝辞

本研究は、国立国語研究所共同研究プロジェクト「多様な語彙資源を統合した研究活用基盤の共創」および「開かれた共同構築環境による通時コーパスの拡張」による成果の一部であり、JSPS 科研費 19H00531 の助成を受けたものです。

参考文献

[1] 小木曽智信, 近藤明日子, 髙橋雄太, 間淵洋子. 『昭和・平成書き言葉コーパス』の設計・構築・公開. 情報処理学会誌, Vol. 65, No. 2, 印刷中.

[2] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written japanese. **Language Resources and Evaluation**, Vol. 48, No. 2, pp. 345–371, June 2014.

[3] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, **Proceedings of the 27th International Conference on Computational Linguistics**, pp. 1384–1397, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[4] Zhaochen Su, Zecheng Tang, Xinyan Guan, Lijun Wu, Min Zhang, and Juntao Li. Improving temporal generalization of pre-trained language models with lexical semantic change. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 6380–6393, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[5] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. **Science**, Vol. 331, No. 6014, pp. 176–182, 2011.

[6] Mark Davies. Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. **Corpora**, Vol. 7, No. 2, pp. 121–157, 2012.

[7] Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. CCOHA: Clean corpus of historical American English. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 6958–6966, Marseille, France, May 2020. European Language Resources Association.

[8] Barbara McGillivray and Adam Kilgarriff. Tools for historical corpus research, and a corpus of latin. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, **New Methods in Historical Corpus Linguistics**, Tubingen, 2013. Narr.

[9] Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In **Proceedings of the Fourteenth Workshop on Semantic Evaluation**, pp. 1–23, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[10] 近藤明日子, 相田太一, 小木曽智信. 単語分散表現の結合学習による単語の意味の通時的変化の分析. 言語処理学会第 28 回年次大会 発表論文集, pp. 1695–1698, アクトシティ浜松 コングレスセンター (オンライン開催), 2022.

[11] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源–形態素解析用電子化辞書の開発とその応用. 日本語科学, Vol. 22, pp. 101–123, 2007.

[12] 小木曽智信, 小町守, 松本裕治. 歴史的日本語資料を対象とした形態素解析. 自然言語処理, Vol. 20, No. 5, pp. 727–748, 2013.

[13] Andrey Kutuzov and Lidia Pivovarova. Three-part diachronic semantic change dataset for Russian. In **LChange**, pp. 7–13, Online, August 2021. Association for Computational Linguistics.

[14] Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection. In Nina Tahmasebi, Syrielle Montariol, Haim Dubossarsky, Andrey Kutuzov, Simon Hengchen, David Alfter, Francesco Periti, and Pierluigi Cassotti, editors, **Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change**, pp. 93–99, Singapore, December 2023. Association for Computational Linguistics.

[15] 凌志棟, 相田太一, 岡照晃, 小町守. 日本語意味変化検出の評価セットの拡張と検出手法の評価. 言語処理学会第 30 回年次大会 発表論文集, 神戸国際会議場, 2024.

[16] 近藤泰弘. BCCWJ 複合辞リストについて. 青山語文, No. 42, pp. 10–15, 2012.

[17] 李在鎬, 佐々木馨. 教科書コーパスを利用した難易度別コロケーション辞書の提案. 第 8 回コーパス日本語学ワークショップ予稿集, 2015.

[18] 近藤明日子. 明治・大正期の書き言葉における文体と語彙―順接の接続詞を例に―. コーパスによる日本語史研究 近代編, pp. 115–136, 2021.

[19] 渡辺由貴. 短単位 n-gram からみた『虎明本狂言集』と『天草版平家物語』の表現の特徴. 日本語文法史研究, No. 5, pp. 149–172, 2020.