近現代雑誌通時コーパスの語彙統計情報の公開

近藤 明日子 [1]　相田 太一 [2]　小木曽 智信 [1]
[1] 国立国語研究所　[2] 東京都立大学
{kondo,togiso}@ninjal.ac.jp　aida-taichi@ed.tmu.ac.jp

概要

1874 年から 2013 年までの 140 年間に刊行された日本語の雑誌をほぼ 8 年おきにカバーする「近現代雑誌通時コーパス」を構築し、これをもとに *n*-gram 頻度形式と SVMlight 形式の共起語情報をオープンデータとして公開した。これは、長期間にわたって比較的均質な資料に基づく日本語の歴史的な変化を研究することができる初めての大規模データとなる。これにより、近年注目を集めている言語変化に対する計算言語学的なアプローチによる研究が日本語においても可能となった。

1　はじめに

国立国語研究所では、2005 年公開の「太陽コーパス」[1] 以来、明治から大正期の総合雑誌コーパスの構築を行ってきた。また、発表者等は 2019 年より『昭和・平成書き言葉コーパス』 の構築 [1)] の一環として総合雑誌のデータ整備を行っている。双方をあわせた「近現代雑誌通時コーパス」は 1874 年から 2013 年までの 140 年間をおおよそ 8 年おきにカバーする通時的なコーパスとなる。

一方で、自然言語処理の分野では近年、通時的な言語変化に関する研究が盛んになっている。単語分散表現を活用した意味変化の検出、意味変化の計算モデルの構築などが高い関心を集め、2019 年以降開催されている歴史的言語変化を扱うワークショップ LChange [2)] も 2022 年に 3 回目を迎える。また、SemEval2020 のシェアドタスク [2] も注目された。しかし、これまで日本語においては通時的な大規模データが存在しなかったため、この分野での日本語を対象とした研究は十分に進んでいない。

そこで、本発表では「近現代雑誌通時コーパス」をもとに、自然言語処理の手法を用いて日本語の通時的な言語変化の研究に利用できるデータを公開する。同コーパスは権利関係上フルテキストデータを配布することはできないため、*n*-gram 頻度形式と SVMlight 形式の共起語情報をオープンデータ（CC BY-SA 4.0）として提供するものである 。

2　データの構築

2.1　雑誌データ

「近現代雑誌通時コーパス」は、明治期から平成期までの書き言葉の通時的変化を研究することを目的として構築したコーパスである。この期間に刊行された雑誌から 8 年おき（一部、6-7 年おき）に各年代を代表する総合雑誌 1 誌を選定し、該当年の全号（特集号は除く）の全テキスト（目次・刊記・広告等は除く）を収録したコーパスである。明治・大正期のデータは『日本語歴史コーパス 明治・大正編 I 雑誌』[3]（短単位データ 1.2）[3)] に基づき、昭和・平成期のデータは発表者等が 2019 年より構築中の『昭和・平成書き言葉コーパス』の 2021 年 11 月時点のデータに基づく。『昭和・平成書き言葉コーパス』は昭和・平成期の新聞・雑誌・ベストセラー書籍を収録するコーパスで、2023 年よりオンライン検索ツール「中納言」[4)] による検索サービスを提供する予定である。

コーパスには短単位 [4] による形態論情報を付与した。形態論情報は MeCab[5)] 用の形態素解析用辞書 UniDic[5] [6] による形態素解析結果を一部人手修正して作成した。UniDic は「明治・大正期の文語文」「明治・大正期の旧仮名遣いの口語文」「昭和期の旧仮名遣いの口語文」「昭和・平成期の現代仮名遣いの口語文」のそれぞれの時代・文体・表記に適合す

1）　JSPS 科研費・基盤研究 (A)「昭和・平成書き言葉コーパスによる近現代日本語の実証的研究」（19H00531）による。このコーパスは平成 30 年の改正著作権法に基づいて著作権処理は行わず、統計情報と検索サービスの提供という形で公開する。
2）　Workshop on Computational Approaches to Historical Language Change. https://languagechange.org/events/
3）　https://ccd.ninjal.ac.jp/chj/meiji_taisho.html#zasshi
4）　https://chunagon.ninjal.ac.jp/
5）　https://taku910.github.io/mecab/

るよう再学習して新たに作成したものを利用した。

「近現代雑誌通時コーパス」に収録した雑誌名と刊行年、および各年の延べ語数を表 1 に示す。

表 1　収録雑誌名・刊行年と延べ語数

刊行年[6]	雑誌名	延べ語数
1874	明六雑誌	18 万
1881	東洋学芸雑誌	21 万
1887	国民之友	109 万
1895	太陽	224 万
1901	太陽	219 万
1909	太陽	209 万
1917	太陽	199 万
1925	太陽	230 万
1933	中央公論	373 万
1941	中央公論	274 万
1949	中央公論	114 万
1957	中央公論	353 万
1965	文芸春秋	231 万
1973	文芸春秋	266 万
1981	文芸春秋	304 万
1989	文芸春秋	315 万
1997	文芸春秋	290 万
2005	文芸春秋	289 万
2013	文芸春秋	309 万
計		4349 万

2.2　データ処理・統計情報

著作権の観点から、2.1 節で収集した雑誌データを再現できない形で公開するために、*n*-gram 頻度と SVMlight[7] の 2 つの形式に加工した。それぞれの形式における出力例を表 2 に示す。

表 2　単語 *n*-gram と SVMlight の形式で出力した例

出力形式	出力例
n-gram 頻度（$n = 5$）	た の で ある 。　1075
SVMlight	0　0:274 1:64 10:10 ...

_n_-gram 頻度　Google Books Ngram[8] と同様に、単語（表層形）の *n*-gram とその頻度をタブ区切りで出力した。今回は 1-gram（単語の頻度）から 5-gram（連続する 5 単語の頻度）まで集計を行った。また、各単語を UniDic に収録されている語彙素 ID に変換し、語彙素 ID *n*-gram の形式も用意した。語彙素 ID は、UniDic の辞書見出しに相当する「語彙素」ごとに与えられた ID で、異表記や異語形をまとめ上げた見出し語（lemma）を一意に示す。UniDic の辞書アーカイブ[9] 中の語彙表（lex.csv）に含まれる語彙素 ID 列と対応付けることで、語彙素（代表表記）・語彙素読み・語種・類（品詞の上位概念）などの情報を取り出すことができる。これにより、表層形とは異なる見出し語の *n*-gram が利用できる。データ作成にあたり、*n* の値が大きくなるほど低頻度の *n*-gram の数が膨大になる（図 1）ため、5 回以上出現する *n*-gram のみ出力した。

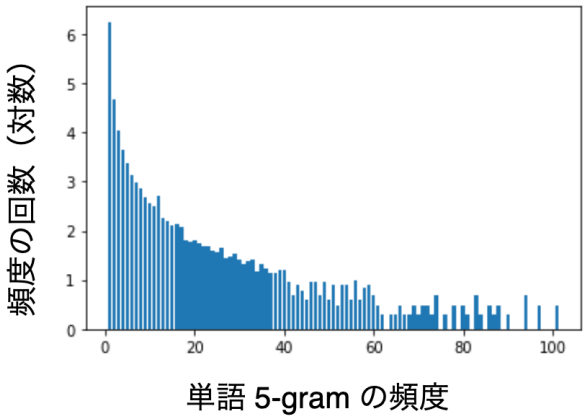

図 1　文芸春秋（1965 年）における上位 100 件の単語 5-gram 頻度および各頻度の出現回数

SVMlight　ここでは、各単語を ID に変換し、共起情報を SVMlight 形式で出力した。SVMlight 形式では、表 2 より、対象の ID（0）に共起する ID（0, 1, 10）とその共起回数（274, 64, 10）を「共起する ID：共起回数」の形で表現する。共起回数が 1 以上の ID のみ出力するため、スパースな情報を表現するのに優れている。今回は *n*-gram 頻度と同様に、各単語を語彙素 ID に変換した。頻度 5 回以上の語彙素 ID を集計対象にし、前後 5 つの語彙素 ID の共起情報を獲得した。

2.3　公開データ

公開データは UTF-8 形式のテキストファイルで、表層形 *n*-gram、語彙素 ID*n*-gram、SVMlight の 3 つのディレクトリに分けて zip 形式で圧縮し

6）『明六雑誌』『東洋学芸雑誌』『国民之友』は言語量確保のため、2 年分を収録している。表中の刊行年はその 2 年の初年を表す。

7）https://www.cs.cornell.edu/people/tj/svm_light/

8）https://storage.googleapis.com/books/ngrams/books/datasetsv2.html

9）https://ccd.ninjal.ac.jp/unidic/download

た。SVMlight 形式は年ごとに 19 ファイル、各々の *n*-gram は 1〜5gram までを年ごとにまとめた 95 ファイル（5 × 19）からなる。アーカイブは下記の URL からダウンロードできる。公開ライセンスは CC BY-SA 4.0 である。https://bit.ly/3HT96Ii

3 使用用途

ここでは、本データの期待される用途について、自然言語処理および日本語学の側面から言及する。

3.1 自然言語処理

自然言語処理では、Google Books Ngram から統計情報や単語分散表現を獲得し、異なる時期間で意味の変化した単語を検出・分析する研究が数多く行われている [7, 8]。しかし、Google Books Ngram に収録されている言語は英語、中国語、フランス語、ドイツ語、ヘブライ語、イタリア語、ロシア語、スペイン語であり、日本語での分析を行うことができない。

そこで、本データを使用することで、日本語でも同様の分析が可能になる。現在、本データを戦前と戦後に分割して単語分散表現を学習し、意味の変化した単語の検出・分析が行われている [9, 10]。

3.2 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（以下、「連語」と呼ぶ）を抽出する研究が行われている [11, 12]。また、『日本語歴史コーパス 明治・大正編 I 雑誌』の単語 *n*-gram を用いて接続詞を抽出し、明治・大正期の接続詞の通時的変化を考察する研究もある [13]。ただし、これまで明治期から平成期までの書き言葉を連続して収録するコーパスが存在しなかったため、近代語から現代語への通史の実証的な実態の把握は困難であった。それが本データを使用することで可能になる。

例として、表層形の 4-gram のデータを使用して、4-gram と刊行年との対応関係を分析してみよう。データから記号類を含むものを除いたうえで各刊行年での頻度上位 5 位の 4-gram を抽出すると、異なりで計 31 種が得られる（表 3 参照）。そのほとんどが何らかの意味・機能を有する連語と見なされる。この 4-gram31 種×刊行年 19 種の頻度[10)]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・刊行年それぞれの第 1 次元（寄与率 53.2%）スコアを表 3・表 4 に示す。

表 3 コレスポンデンス分析による 4-gram の第 1 次元スコア

4-gram	スコア
に至ては	-2.684
時に於て	-2.681
以て之を	-2.597
ざるを得ず	-2.508
を以て之	-2.496
に至りては	-2.482
今日に於て	-2.206
何となれば	-2.191
ものにして	-2.039
に於ては	-1.831
せんとする	-1.662
に於ても	-1.570
ねばならぬ	-0.656
なければならぬ	-0.595
てゐるので	-0.424
してゐた	-0.422
てゐたの	-0.405
のであるが	-0.384
ないのである	-0.038
としては	0.024
たのである	0.100
においては	0.309
なければならない	0.453
ではないか	0.557
のではない	0.653
ているので	0.858
されている	0.907
というのは	0.918
ていたの	0.937
していた	0.938
ていました	1.019

表 3 と表 4 のスコアの高低は対応関係にある。表 4 においてスコア昇順で刊行年が古い年から新しい年の順に並んでいる。それに対応して、表 3 ではスコア昇順で、刊行年の古い雑誌に使用されやすいものから新しい雑誌に使用されやすいものへと 4-gram が並んでいることになる。ここから明治期から平成期にかけて使用される連語の消長を概観できる。このように本データを使用することで、連語をはじめとする語彙の通時的変化の実態を明らかにすることができる。

10） データに収録されない頻度 5 未満の 4-gram は頻度 0 と見なした。

表 4 コレスポンデンス分析による刊行年の第 1 次元スコア

刊行年	スコア	刊行年	スコア
1874	-2.839	1949	0.372
1881	-2.721	1957	0.535
1887	-2.491	1965	0.673
1895	-2.351	1973	0.715
1901	-1.967	1981	0.743
1909	-1.240	1989	0.814
1917	-0.868	1997	0.837
1925	-0.529	2005	0.839
1933	-0.396	2013	0.920
1941	-0.325		

4 おわりに

「近現代雑誌通時コーパス」の整備により、日本語においても大規模なデータに基づいて通時的な言語変化を研究することが可能になった。しかし、このコーパスは未だ整備途上であり、今回公開するデータにも不完全な点がある。コーパスは 2023 年度に完成予定である。

また、今回提供する *n*-gram 頻度形式と SVMlight 形式のデータはコーパスか得られる統計情報の一例に過ぎない。研究利用上、別の形式のデータが必要となる場合にはぜひ発表者にコンタクトしてほしい。コーパスの完成に合わせて有用な形式のデータを公開することを検討している。

今後、これらのデータを用いることで、日本語においても計算言語学的なアプローチによる言語変化の研究が盛んに行われることを期待している。

謝辞

本研究は JSPS 科研費 19H00531、20K20411、国立国語研究所共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」の助成を受けたものです。

参考文献

[1] 国立国語研究所. 太陽コーパス—雑誌『太陽』日本語データベース. 博文館新社, 2005.

[2] 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.

[3] 国立国語研究所（近藤明日子・間淵洋子・服部紀子・南雲千香子ほか）. 日本語歴史コーパス 明治・大正編 I 雑誌（短単位データ 1.2）. 国立国語研究所, 2019.

[4] 小椋秀樹, 冨士池優美. 『現代日本語書き言葉均衡コーパス』の言語単位. 『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 4 版（上）, pp. 1–10, 2011.

[5] 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.

[6] 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.

[7] Kristina Gulordava and Marco Baroni. A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In **Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics**, pp. 67–71, Edinburgh, UK, July 2011. Association for Computational Linguistics.

[8] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1489–1501, Berlin, Germany, August 2016. Association for Computational Linguistics.

[9] 相田太一, 小町守, 小木曽智信, 高村大也, 坂田綾香, 小山慎介, 持橋大地. 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 26 回年次大会 発表論文集, 2020.

[10] 相田太一, 小町守, 小木曽智信, 高村大也, 持橋大地. 通時的な単語の意味変化を捉える単語分散表現の同時学習. 言語処理学会第 27 回年次大会 発表論文集, 2021.

[11] 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.

[12] 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.

[13] 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.