Possibilities of Discovering Corpus-Based Studies Focusing on Data Citation

データ引用による言語資源活用文献の把握の可能性: BCCWJの分析から

中渡瀬 秀一 (国立情報学研究所) 加藤文彦 (国立情報学研究所)
大向一輝 (国立情報学研究所)

Possibilities of Discovering Corpus-Based Studies Focusing on Data Citation

Hidekazu Nakawatase (National Institute of Informatics)
Fumihiro Kato (National Institute of Informatics)
Ikki Ohmukai (National Institute of Informatics)

要旨

言語資源データの引用情報調査に基づいて、そのデータを活用した研究文献の発見可能性について論じる。このために言語処理学会年次大会発表論文集を対象として「現代日本語書き言葉均衡コーパス」などの引用情報を調査した。本稿ではその結果と今後の課題について報告する。

1. はじめに

著者らの所属機関では学術情報を検索するサービスである学術情報ナビゲータ CiNii(1)を提供している。これまで学術情報として論文、図書・雑誌や博士論文を対象としてきたが、新たにデータセットなども検索対象とする「CiNii Research」の開発が現在進められている (大向一輝 2019)。ここでは「研究データを取り巻く情報」 からデータセットを検索する機能を提供することを目指しており、そのような情報のひとつとしてデータセットとその利用文献の関係も収集する必要がある。

そこで本稿では言語資源データを活用した研究文献を発見するため、その引用情報を利用する方法について調査に基づいた検討を行った。調査では言語資源（コーパス）に「現代日本語書き言葉均衡コーパス」など、文献には言語処理学会年次大会発表論文を用いた。2 節では 引用情報の調査について述べる。

2. 引用情報の調査

引用情報の調査は次のように行った。まずコーパス名（または略称）が記載されている文献を 2007 年から 2018 年の言語処理学会年次大会発表論文集から抽出し、年度毎の件数の推移を集計した。用いたコーパスは現代日本語書き言葉均衡コーパス (BCCWJ)、国語研日本語ウェブコーパス（NWJC)、日本語話し言葉コーパス（CSJ)、日本語歴史コーパス（CHJ）の 4 種

(1) http://ci.nii.ac.jp/

である。次に抽出文献数の多いコーパスについては、抽出された文献がコーパスを活用した研究であるかを確認し、引用情報の記載箇所を集計した。

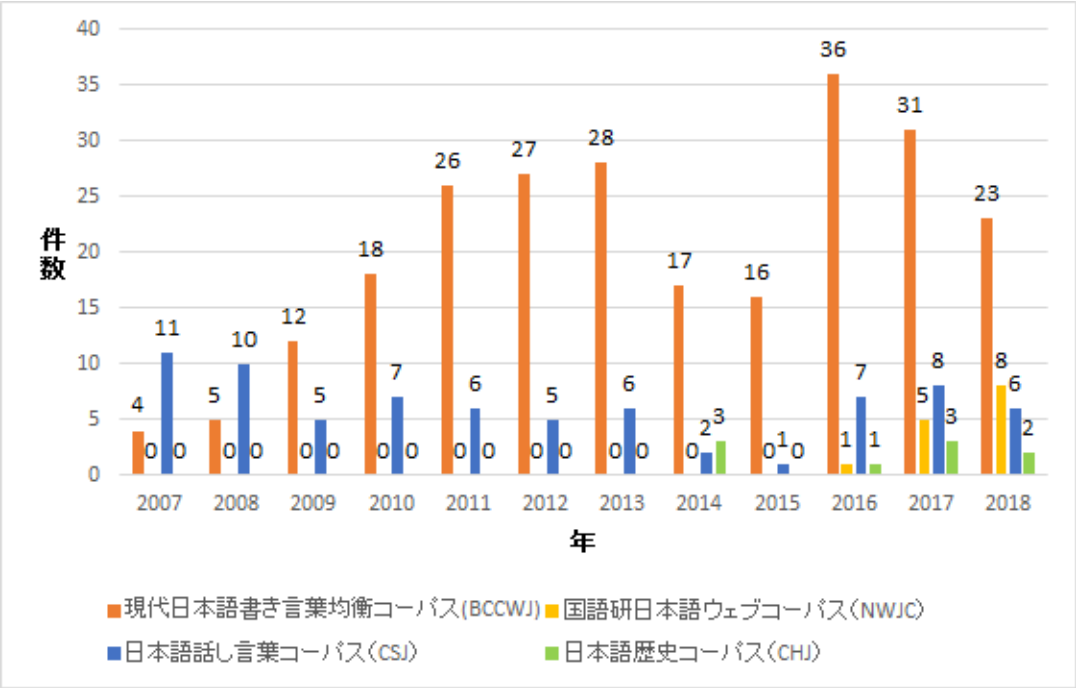

図 1　NLP 年次大会における各コーパス名の出現文献数

2.1　文献に出現する各コーパス名の推移

図 1 に各コーパス名（またはその略称）が出現する文献数の推移を示す。4 コーパスのうち BCCWJ の記載された文献が最も多い。以後の調査・分析は BCCWJ に限定して行った。

2.2　BCCWJ が出現する文献数と成果報告文献数の推移

BCCWJ は有償の DVD または無償のオンラインサービスによって利用することが可能である。ただし有償、無償を問わずこれらを利用して研究成果を公表する場合には、規約によりコーパス開発センターに報告する必要がある（2019 年 3 月時点）。この成果報告文献は公表されている[(2)]。この中から言語処理学会年次大会の文献数と前記の BCCWJ が含まれる文献数とを対比する。図 2 に各年度の「BCCWJ」が含まれる文献数と成果報告文献数の推移を示す。

成果報告文献数は一貫してコーパス名 BCCWJ が含まれる文献数の半数にも満たない。例えば後者で 2012 年度分は 27 件あるが、そのうち前者に含まれるのは 3 件に過ぎない。この 27 件を通読した結果、20 件が BCCWJ を利用した研究であることが確認された。つまり公表された成果報告文献は実際の成果文献の一部であることが分かる。次にコーパスの引用情報から与えられた文献が成果文献であるか判断する方法を検討するために、それらの出現箇所につ

(2) *https : //pj.ninjal.ac.jp/corpus_center/bccwj/list.html*

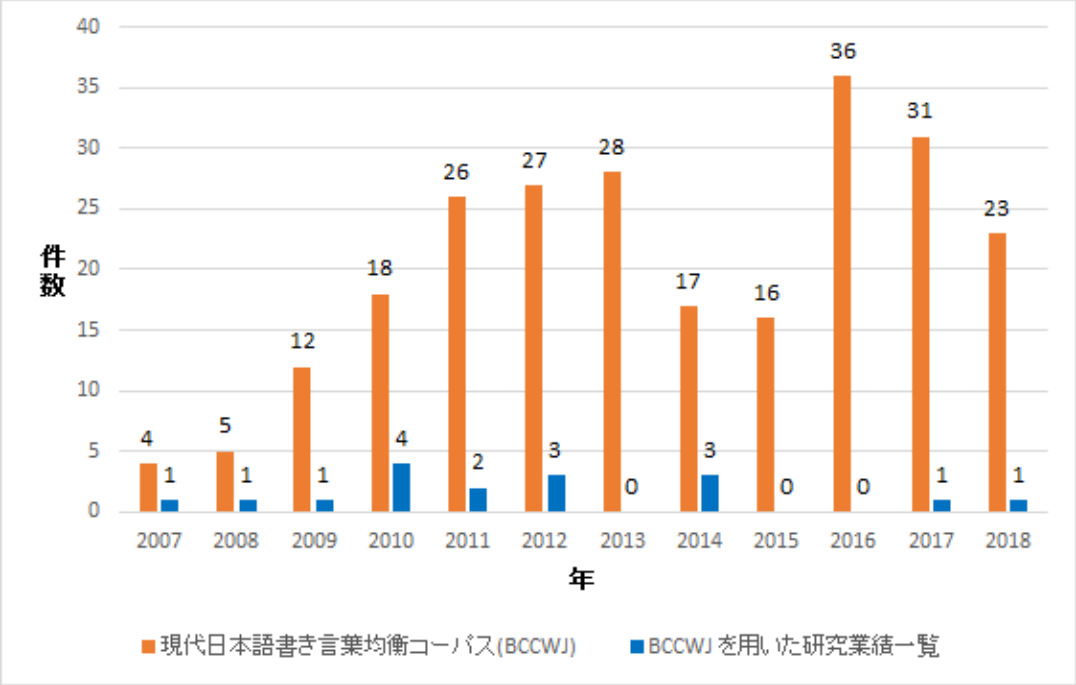

図 2　NLP 年次大会における BCCWJ の出現文献数と成果報告文献数

いて調査した。対象文献は前記の 2012 年度分を用いた[3]。BCCWJ の場合、引用情報の記載方法ついて規約で以下のように指示されている。

- 『著作中に「書き言葉コーパス」を利用した旨を明記』 (DVD または「中納言」の場合)
- 『「現代日本語書き言葉均衡コーパス」の利用による成果である旨を明示』 (「少納言」の場合)

このように具体的な引用情報の記載方法は利用者に委ねられているため、その箇所についての実態を明らかにする。

2.3　引用情報の出現箇所

引用情報が出現しうる箇所には文献欄、謝辞、脚注[4]、本文の 4 種類（およびその組合せ）がある。前記 27 件の文献について出現箇所を調査した結果を表 1 に示す。全体的には少なくとも本文に記載されるケース（23 件）が多く、次に少なくとも文献欄に記載されるケース（12 件）が多い。それらの組合せでは本文のみ（10 件）と文献欄と本文（8 件）が多い。

表 1　引用情報の出現箇所（文献欄・謝辞・脚注・本文とその組合せ）の度数分布

	文献欄のみ	文献欄と本文	謝辞のみ	謝辞と本文	脚注と本文	文献欄と脚注と本文	謝辞と脚注と本文	本文のみ
度数	3	8	1	1	2	1	1	10

[3] BCCWJ は 2011 年に完成したため、それ以降で最も成果報告期間の長い 2012 年とした。
[4] 本文中のコーパス名に付随する脚注も含む

2.4　引用情報の出現箇所と成果文献との関係

前記 27 件の文献のうち 20 件が BCCWJ を利用した研究の成果文献である。次にそれらと引用情報の出現箇所との関係を分析する。まず 4 種類の出現箇所別に該当文献が成果文献である比率を求める。たとえば「文献欄」の場合には少なくとも「文献欄」への記載がある 12 件のうち成果文献は 8 件となるので 67 %である。同様に他の出現箇所について計算した結果を表 2 に示す。引用情報の記載箇所において最も成果文献比率が高いのは「謝辞」である。謝辞にはほぼコーパス利用に対する謝意のみが示されるのに対し、「本文」や「文献欄」にはそれ以外の言及が可能である。このため成果文献比率が低くなっている。

表 2　引用情報の出現箇所別の成果文献比率

	文献欄	謝辞	脚注	本文
成果文献数	8	3	3	18
該当文献数	12	3	4	23
成果文献比率 (%)	67	100	75	78

次に表 1 のようにに分類した出現箇所の場合における成果文献比率を表 3 に示す。

表 3　引用情報の詳細な出現箇所別の成果文献比率

	文献欄のみ	文献欄と本文	謝辞のみ	謝辞と本文	脚注と本文	文献欄と脚注と本文	謝辞と脚注と本文	本文のみ
成果文献数	1	6	1	1	1	1	1	8
該当文献数	3	8	1	1	2	1	1	10
成果文献比率 (%)	33	75	100	100	50	100	100	80

ここでも「謝辞」を含む部分は必ず 100 %となる。それ以外の部分では「文献欄と脚注と本文」なども 100 %であるが表 3 から分かるようにこの部分は度数が 1 である。表 2 において成果文献比率が概ね 7 割以上であるのに比べ、成果文献比率の低さが目立つのは「文献欄のみ」である。このケースでは BCCWJ を利用した他の研究文献を引用していたために本文にはコーパス名が記載されず文献欄のみ BCCWJ が現れていることが原因である。

3.　おわりに

本稿では本文や参考文献欄などに記載された引用情報をもとにして言語資源データを利用した研究文献を発見するための検討を行った。具体的には言語資源に BCCWJ、対象文献に言語処理学会年次大会発表論文を用いた調査を行った。

本調査の結果、成果文献を特定するのに最も有効な引用情報は「謝辞」に存在することが判明した。しかし成果文献全体で謝辞のある文献は 1 割程度である。一方、本文に含まれる引用情報に注目すれば、成果文献比率は 8 割弱に低下するが成果文献全体の 9 割をカバーしてい

る。この状況を改善する方法としては、文献内容に「コーパスを利用している」ことが含まれることを正確に認識することで判断する方法やコーパスの利用規約において成果文献に記載すべき具体的な謝辞内容を指定する方法が考えられる。後者の方法は既に助成金などの利用規約では導入されているものである。また後者には前者のような技術的課題がなく容易に実施できる上に効果も見込めるため導入の検討を期待したい。

謝　辞

本研究の一部は科研費（課題番号：16K12833）の助成を受けて行われたものである.

文　献

大向一輝 (2019), 「研究データディスカバリーにおける 引用情報の利活用」 JOSS2019
https://japanlinkcenter.org/rduf/doc/joss2019_rdc_06.pdf

関連 URL

言語処理学会年次大会発表論文集　https://www.anlp.jp/guide/nenji_proceedings.html
BCCWJ を用いた研究業績一覧　https://pj.ninjal.ac.jp/corpus_center/bccwj/list.html