DISCUSSION PAPER No.200

プレプリントとジャーナル論文の差異：bioRxiv を用いた試行

Differences between preprints and journal articles: a trial using bioRxiv

2021 年 8 月

文部科学省　科学技術・学術政策研究所

データ解析政策研究室

小柴　等　　林　和弘

【執筆者】

小柴　等　　文部科学省科学技術・学術政策研究所
データ解析政策研究室・主任研究官

林　和弘　　文部科学省科学技術・学術政策研究所
データ解析政策研究室・室長

【Authors】

KOHISBA Hitoshi　　Senior Research Fellow, Research-Unit for Data Application, National Institute of Science and Technology Policy (NISTEP), MEXT

HAYASHI Kazuhiro　　Senior Research Fellow, Research-Unit for Data Application, National Institute of Science and Technology Policy (NISTEP), MEXT

本報告書の引用を行う際には、以下を参考に出典を明記願います。
Please specify reference as the following example when citing this paper.

小柴等・林和弘（2021）「プレプリントとジャーナル論文の差異：bioRxiv を用いた試行」, NISTEP DISCUSSION PAPER, No.200, 文部科学省科学技術・学術政策研究所.
DOI: https://doi.org/10.15108/dp200

KOSHIBA Hitoshi and HAYASHI Kazuhiro (2021) “Differences between preprints and journal articles: a trial using bioRxiv,” NISTEP DISCUSSION PAPER, No.200, National Institute of Science and Technology Policy, Tokyo.
DOI: https://doi.org/10.15108/dp200

プレプリントとジャーナル論文の差異：bioRxiv を用いた試行

文部科学省 科学技術・学術政策研究所 データ解析政策研究室

要旨

エビデンスに基づく政策立案 (以下，EBPM：Evidence-based Policy Making) 機能の強化や，オープンサイエンスの潮流を踏まえ，本稿では，プレプリントを用いた分析を活用することで，原著論文を基とした研究力の分析に対して相補的に新しい知見を得ることができるのではないかと考えた．特に，プレプリントとそれが最終的に出版された OA ジャーナル論文とを比較することで知見を得られないか試行した．

まず，近年のオープンジャーナルなどの流れもあって，bioRxiv のプレプリントと OA ジャーナル論文の全文 XML を一定量確保することができ，プレプリントとジャーナル論文の比較に関する技術的な可能性は検証できた．他方，参考文献数や単語数などの外形的な基準や，簡単な文書類似度から両者の差分を明らかにしようとした今回の試行の範囲では，プレプリントとジャーナル論文，ジャーナル論文になったプレプリントとそうではないプレプリントの間で明確な違いを見いだすことはできなかった．機械学習手法を用いても分類精度は 47% 程度と高くなかった．

プレプリントとジャーナル論文の間に大きな差はないという結果は先行研究でも示されている知見と一致するものであるが，より大規模かつ比較的最近の状況でも再現性が確認できた．これらに加えて本稿のもたらした新たな知見としては，著者数などの多くの外形的的基準においても差は小さい，ジャーナル論文になっていないプレプリントとの違いも大きくない，といった点が挙げられる．

Differences between preprints and journal articles: a trial using bioRxiv

Research-Unit for Data Application, National Institute of Science and Technology Policy (NISTEP), MEXT

ABSTRACT

In light of the strengthening of the evidence-based policy making (EBPM) function and the trend toward open science, this paper considers whether analysis using preprints can provide new knowledge in a complementary manner to the analysis of research power based on original papers. In particular, the preprint and the final version of the preprint are used as the basis for the analysis. In particular, we attempted to see if we could gain insights by comparing preprints with OA journal articles in which they were finally published.

Translated with www.DeepL.com/Translator (free version)

First, due to the recent trend of open journals, we were able to secure a certain amount of full-text XML of bioRxiv preprints and journal articles, and verified the technical feasibility of comparing preprints and journal articles. On the other hand, within the scope of this trial, which attempted to identify differences between preprints and journal articles, and between preprints that became journal articles and those that did not, we could not find any clear differences. Even using machine learning

methods, the classification accuracy was not high at about 47%.

The result that there is no significant difference between preprints and journal articles is consistent with the findings of previous studies, but it is also reproducible in a larger and relatively recent context. In addition to these findings, the new findings of this paper are that the differences are small in many external criteria such as the number of authors, and the differences between preprints and journal articles are not large.

目次

1　はじめに

我が国の科学技術イノベーション（以下，STI とする）政策立案において，エビデンスに基づく政策立案 (以下，EBPM：Evidence-based Policy Making) 機能の強化が求められている．「第 6 期科学技術・イノベーション基本計画 (令和 3 年 3 月 26 日閣議決定)」（以下，第 6 期基本計画という）においても，科学技術・イノベーション行政において，客観的な証拠に基づく政策立案を行う EBPM を徹底することとされ，内閣府などにおいてもエビデンス等を整備する関連する取組が進められてきた．

その中でも研究力については，計量書誌学や科学計量学を基礎とした学術ジャーナルの査読を通った原著論文（以下，原著論文とする）に着目した定量的な分析と，政策づくりへの反映が試みられている．例えば，科学技術・学術政策研究所（以下 NISTEP とする）においては，サイエンスマップ，国別ベンチマーキング，大学ベンチマーキング，などの調査分析を行い，その内容が STI 政策づくりの一助となっている．

一方，原著論文の分析においては，研究成果が生まれてから，査読・編集・出版を経て公開されるまでのタイムラグが含まれるため，全体の傾向（trends）をレビューすることには向いているが，新興領域を早く押さえることは構造的に難しい．また，原著論文においては原則それぞれの領域で確立された科学的判断基準によって査読が行われるため，学際領域や融合領域，あるいは全く新しい概念の論文が不利になりやすい．

そこで，論文原稿の草稿であり，査読による選別もされていない「プレプリント」に着目し，その分析を行うことで，原著論文を基とした研究力の分析に対して相補的に新しい知見を得ることができるのではないかと考え，NISTEP では一連の報告を行っている[1]．

ここで，ジャーナル論文として投稿する前の段階と最終的にジャーナル論文になった段階とでどのような違いがあるのか，ジャーナル論文になっているプレプリントと現状においてジャーナル論文になっていないプレプリントの間にはどのような違いがあるのかをみることで，プレプリントや，査読，あるいは原著論文そのものの信頼性に関する議論が行える．

本稿では，以上の背景と問題意識のもと，生物系のプレプリントサーバである bioRxiv を対象に，前掲のジャーナル論文とそのプレプリントの差異，後にジャーナルに掲載されたプレプリントと未掲載のプレプリントの差異について調査した結果を報告する．

2　位置づけ

類似する先行研究としては [Klein19, Carneiro20, Akbaritabar21] などが挙げられる．

Klein[Klein19] はまさにジャーナル誌に掲載された論文とそのプレプリント版の比較を行ったも

[1] 第 6 期基本計画においても，“論文のオープンアクセス化や研究成果の迅速な公開の場の一つとしてのプレプリントの活用も一層加速しており，研究データの公開・共有を含め，オープンサイエンス等の世界的な知の共有を目指した研究成果のオープン化が進みつつある” との認識が示されている．

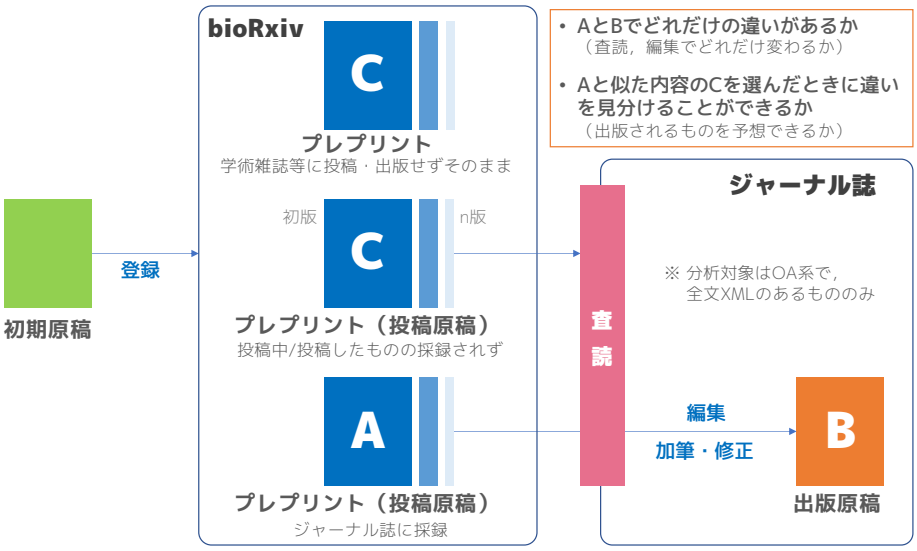

図 1　概要図

ので，本稿の内容と極めて類似する．ここでは arXiv，bioRxiv を対象に 12,000 件を超えるプレプリントとそのジャーナル版を比較して，その間に差はほとんどなかったことを報告している．ただし主体は arXiv で，bioRxiv については 2013 年 11 月のサイト開設から 2016 年 11 月までの期間に投稿されたものを解析対象にしており，全体で 7 千件，論文フルテキストは 220 件と比較的少数にとどまっている．また，内容については単語それぞれを独立次元とする Bag-of-Word の Cos 類似度を採用している．後述するとおり，本稿では bioRxiv に限定し期間は 2019 年．プレプリント件数は全体で 2 万件超，論文フルテキストのペア比較には約 7 千件を用いていることから，bioRxiv の比較的最近の状況について，より広範に調査していることになる．また，内容については分散表現ベースで類似度を判定している．さらに，プレプリントとジャーナル論文だけでなく，ジャーナル論文に採録されているプレプリントと採録されていないプレプリントの比較も行っている．

Carneiro[Carneiro20] もやはり，bioRxiv に搭載され，後に PubMed に搭載されたジャーナル論文を対象に質の変化を分析し，ピアレビューを経てジャーナル誌に掲載されたものはプレプリントよりわずかに質が高いもののその差は大きくないと報告している．この論文では bioRxiv は 2016 年に投稿されたものを対象にしており，主にアンケートに基づいて質の変化を調査している．アンケートであるので分析用に用いられたプレプリントとジャーナル論文のペアは 56 件である．調査期間が 2016 年と比較的古い点，相対的に少数のデータについて，アンケートベースで調査している点，に差異がある．

Akbaritabar[Akbaritabar21] はプレプリントとジャーナル誌の間での参考文献の差を調査している．6000 件を超えるプレプリントとジャーナル論文のペアについて定量的に調べているほか，100 件程度について人間が内容を読み込んで評価し，追加される文脈の種類を分類したり，分野ごとの違いを明らかにしている．当該論文が参考文献に焦点を当て，精密な分析を試みていることに対して，本稿では様々な指標について網羅的に概観している点に差異がある．

3　データ

本章では分析データについて述べる

3.1　プレプリントサーバ：bioRxiv

今回はプレプリントサーバとして生物系の分野で用いられる bioRxiv を設定した．理由を以下に述べる．

プレプリントサーバとしては物理・情報系の分野でよく用いられる arXiv が有名で，もっとも長い歴史を有し，投稿やダウンロードなどの利用数も多い[2)]．この点からは arXiv が適切と考えられる．しかしながら，情報系の分野では必ずしもジャーナル論文が重視されずトップカンファレンスのプロシーディングスがジャーナル論文同等の価値を持っているという特性があり，“ジャーナル論文との比較” を考えた際に議論が複雑になる．また，原稿のソースが TeX 形式で公開されているものの，カスタマイズの自由度が高いために解析は必ずしも容易ではない．また，ジャーナル論文やトップカンファレンスのプロシーディングスのソースが閲覧できるか，解析しやすい HTML や XML の形式で取得できるかも明らかではない．

他方，情報系以外の分野ではトップカンファレンスのプロシーディングスがジャーナル論文同等の価値，との議論は観察されない．また，bioRxiv は JATS(Journal Article Tag Suite) 形式の全文 XML を提供している．bioRxiv を経由して採録されたジャーナル論文についてもオープンアクセス形式のものが上位に並んでおり [林 21]，それらの多くについて後述の通り JATS 形式での全文 XML 取得が可能である．以上より，最終成果がジャーナル論文と結びつくと期待できること，プレプリントとジャーナル論文のデータを同じ形式で取得できること，などから bioRxiv を選定した．欠点としては普及率の差に起因して解析対象の数が 1 万件に満たない点にある．

3.2　分析の対象（期間）・データ数

今回，分析の対象（期間）としては 2019 年に投稿されたプレプリントとした．理由は以下の通りである．

まず，生物系の分野も COVID-19 には密接に関連しており，2020 年以降 COVID-19 関連の投稿が一定数観測されている [小柴 20]．bioRxiv 全体から考えるとそれらの数は大きなものではない [林 21] が外乱として作用する可能性があり考察を複雑化させる．次に，古いデータを使うとこれらの影響を排除できるが，あまり古すぎるとトレンドの変化等で現状には適用できない可能性が高まる．

ここで COVID-19 関連のプレプリントは 2019 年 12 月末ごろから増加し始めており，2019 年のデータを使うと比較的最近で，かつ COVID-19 の影響を考慮しなくて良い．また，既報

2) 例えば，投稿数やダウンロード数は以下で確認できる． https://arxiv.org/help/stats

[Abdill19, 林 21] のとおり bioRxiv に投稿されたもののうち，ジャーナル論文として採録されているものはプレプリント登録から平均的に 6 ヶ月から 8 ヶ月でジャーナル論文化されている．本稿の分析は 2021 年 5 月に行っていることから，2019 年 12 月末からみて 1 年半近く 17 ヶ月の期間が開いており 2019 年登録のプレプリントについてジャーナル論文化はほぼ完了して，今後ジャーナル論文化するものは多くないと期待できる．

以上の背景から 2019 年に投稿されたプレプリントを対象とする．結果，対象となったプレプリントの数，うちジャーナル論文の記載があるものの数，さらに今回の分析に用いる全文 XML のあるジャーナル論文の数はそれぞれ以下の通りとなった．（なお全文 XML が取得できても，中身の値が正常ではないケースがあることから有効数は下回る．）

プレプリント数　28,805 件
うちジャーナル論文数　13,450 件
うち全文 XML 保有数　7,985 件

ここで，プレプリントは何度でも改訂可能であるが今回の分析では初版のみを対象とする．これは，投稿されたすべてのプレプリントについて初版は確実に存在し，統制が容易であることに起因する．ジャーナル論文から最も近い版を採用すると，ジャーナル掲載されていないものも最新だけ見るのか，ジャーナル掲載後に更新があった場合の扱いはどうするのか，初版から最新までの差異も見るか，など議論が複雑化する．

3.3　データ取得方法

データはそれぞれ以下の手法で取得した．

bioRxiv については，過去（2021 年 4 月）に bioRxiv の API [3] を通じて，各プレプリントの詳細情報を取得済みである．この詳細情報に “jats xml path” が記載されていることから，単にそれらの URL から適切に情報を取得した．この際，当該 URL にアクセスしてもエラーが返ってくるものも数百件ほど存在しており，これらは分析の際に除外する．

ジャーナル論文については何段階かのステップを踏む．

まず，プレプリントとジャーナル論文の紐付けは以下の通りである．前述の各プレプリントの詳細情報には “doi” の記載があり，CrossRef の API [4] を通じてプレプリントの DOI 詳細を取得する．すると，この中に “is-preprint-of” という属性があり当該プレプリントに基づくジャーナル論文等があった場合に値が設定される．今回は “is-preprint-of” の “id-type” が “doi” で具体の ID(DOI) が設定されている場合をジャーナル論文として採用した．従って，本稿で言うジャーナル論文はすべて DOI を有している．

次に，ジャーナル論文の全文 XML を 2 段階に分けて取得する．一般にジャーナル論文のデータは各出版社のプラットフォームで提供され，API もそれぞれに用意されることからそれらに対

3) https://api.biorxiv.org/
4) https://api.crossref.org/

応することは煩雑である．ここで多くのオープンアクセスジャーナルについて PubMed Central (PMC) が PMC Open Access Subset [5]において JATS 形式の全文 XML を一元的に提供してくれている．そこで，ここでは PubMed Central OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) [6]を用いてデータを収集する．

PMC OAI-PMH では PMC が各論文に対して独自に割り当てた ID である PMCID をベースにデータを取得するため，PMC の ID Converter API [7]を通じて DOI を PMCID に変換する．その上で，OAI-PMH からデータを取得する．すべての DOI が PMCID に変換できるわけではないし，PMCID があるからと言って全文 XML が存在するわけではないが，事前に全文 XML の存在を確認することは困難であるので，今回は PMCID が取得できた場合はすべて OAI-PMH にかけてデータを取得した．

4　分析方法

本章では分析方法について述べる

4.1　比較対象

比較の対象は 3 パタンを設定する．

まず ジャーナルになったプレプリント（ジャーナルプレプリント）とジャーナルの比較，次にジャーナルプレプリントと類似する別のプレプリントの比較，最後に ベースラインとしてプレプリントとジャーナルをランダムに組み合わせたペアの比較．

“類似する別のプレプリント” については説明を要するため，後で改めて説明する．

4.2　特徴量の選定

差分を観測するための特徴量として今回は以下を設定した．著者数，参考文献数，段落数，単語数，章タイトル，内容の類似度．

著者，参考文献は単純な数の比較に加えて，名前やタイトルが一致するものがどの程度あるかも見る．また，章タイトルは一致に加えて片方だけに存在する単語は何かも調査する．

4.3　類似度の考え方

特徴量で述べた内容の類似度については以下の通りとする．

基本的には 2 つの文書（プレプリントやジャーナル論文）について段落単位で比較する．この際，段落の位置が変わっても検出できるように以下のようにする．

---

5) https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
6) https://www.ncbi.nlm.nih.gov/pmc/tools/oai/
7) https://www.ncbi.nlm.nih.gov/pmc/tools/id-converter-api/

1. 2 つの文書 A,B の全段落間で類似度を検出する．
2. A 側の各段落から見て最も高い類似度をもつ B 側の段落をペアにする．
3. A,B を入れ替えて 2. の処理を行い，B 側から見たペアを作成する．
4. 2.，3.，でできたペアを全体として，類似度が一定以上のもののみを 2 つの文書の共通ペアとする．
5. Jaccard 係数を計算し，2 つの文書 A,B 間の類似度とする．

なお 4. で用いる閾値については，一旦 3. までで算出した段落間の全類似度の分布を見て定める．

ここまでの説明では “類似度” で説明してきたが，今回の具体の作業においては各段落を分散表現を用いたベクトルに変換し，ベクトル間の距離で計算している．距離は近い方が似ている，類似度は大きい方が似ている，ということで方向性は異なるが，その点を除くと基本的な手続きに違いはない．

4.4 分散表現

先述した類似度算出においては単語を基準にしてコサイン類似度を求める方法が一般的であるが，今回は分散表現を用いて算出した．これは，査読等の過程で単語や言い回しが変化した場合でもそれらの差異をある程度吸収して類似度を算出するためである．

分散表現獲得の具体的な手法としては FastText [8] を用いた．データには分析対象であるプレプリント，ジャーナル論文のタイトル，概要，全文データを用い，skip-gram で 300 次元の埋め込みを実施した．作業に先立ってデータは NLTK と WordNet をベースに Lemmatize やストップワード除去も行った．

結果，230,020 語の 300 次元分散表現を獲得した．段落等の分散表現はそれらの含まれる分散表現を線形加算して正規化したものを使用する．

4.4.1 分散表現の獲得手法について

生物系分野においては例えば BioBERT [9] が公開されており，これを使うことで，より高い精度の分散表現獲得が期待できる．ただし，処理速度の面から今回は BioBERT の使用は見送り，独自に構築した FastText ベースの分散表現を利用した．

具体的に，今回利用した全文データから適当に 5 件を抜き出して処理速度を計算したところ，BioBERT では 5 件の合計で 27.8sec，平均 5.6sec，FastText ベースで処理したものでは 5 件の合計で 0.1sec である [10]．

FastText ベースでの処理は分散表現辞書の構築や，そのための Lemmatize，ストップワードの除去などの処理コストがかかるため，一概にこの計算結果のみを持ってコストの比較を行ってはなら

8) https://fasttext.cc/
9) https://github.com/dmis-lab/biobert
10) FastText ではコンテキストに応じて座標値が変わることがないため，あらかじめ計算した単語単位の分散表現を DB に格納し，都度読み出して計算している．

ないが，単純に前述した 5 文章あたりの処理速度を比較すると 278 倍の差がある．1 文書辺り 20 段落あるとして仮に 2 万件の文書を処理する場合，40 万回の算出を要することになり，BioBERT では 622 時間（約 26 日）が見込まれる．他方，FastText ベースでは 2 時間程度と見込まれる．

今回，分散表現を用いる理由は単語や言い回しの揺らぎの吸収にあって，厳密な精度を追求するものではないため，処理速度の速さを重視して FastText ベースでの処理を採用した．BioBERT を用いた場合と FastText ベースでの処理において精度にどの程度差が出るかの検証は今後の課題である．

4.5 “類似するプレプリント” の考え方

ジャーナルプレプリントと，それ以外の一般のプレプリントを比較する際，単にジャーナルプレプリントと，それ以外のプレプリントについて特徴量の統計値を算出して比較するのはわかりやすいが，一方でジャーナルプレプリントとジャーナルを比較した結果との対比は困難になる．そこで，ジャーナルプレプリントと類似するプレプリントを探索し，この類似するプレプリントとの比較結果を利用する．

“類似するプレプリント” は概要の分散表現値の距離が最も小さいもの（類似度が最大のもの）を設定する．この際，類似するプレプリントはジャーナルプレプリントよりもタイムスタンプが古いもので，かつジャーナルになっていないものに限定する．これにより，ジャーナルになっていないことはもちろん，ジャーナルやジャーナルプレプリントから影響を受けた可能性も排除できる．

ここで，分散表現値の距離が最も小さいものについては，計算量削減のため NGT [11] を用いて近似的に上位 100 件を取得することで行った．類似するプレプリントの上位 100 件中に前述の条件（時間的に古く，かつジャーナルになっていないもの）を満たすものがないケースも千件程度存在し，結果，ジャーナルプレプリントと，類似するプレプリントのペアは 6,905 件である．また，同一のプレプリントが複数のジャーナルプレプリントに結びつくケースも多く，“類似するプレプリント” のユニークな件数は 3,874 件である．

5 結果

5.1 初期設定

まず，類似度算出の際の閾値を定めるために用いる，ジャーナルプレプリント - ジャーナル，ジャーナルプレプリント - 類似するプレプリント，ランダムマッチングペア，それぞれの全段落組み合わせにおける類似度（距離）の分布を図 2 に示す．

図 2 を見ると，ランダムマッチングの場合では距離 0.3 以下のものはほとんど見られない．一方でジャーナルプレプリントと，ジャーナル・類似するプレプリントの組み合わせでは 0.3 以上のものも観察される．そこで，文書間の Jaccard 係数算出に用いる閾値は距離 0.3 を採用する．

11) https://github.com/yahoojapan/NGT

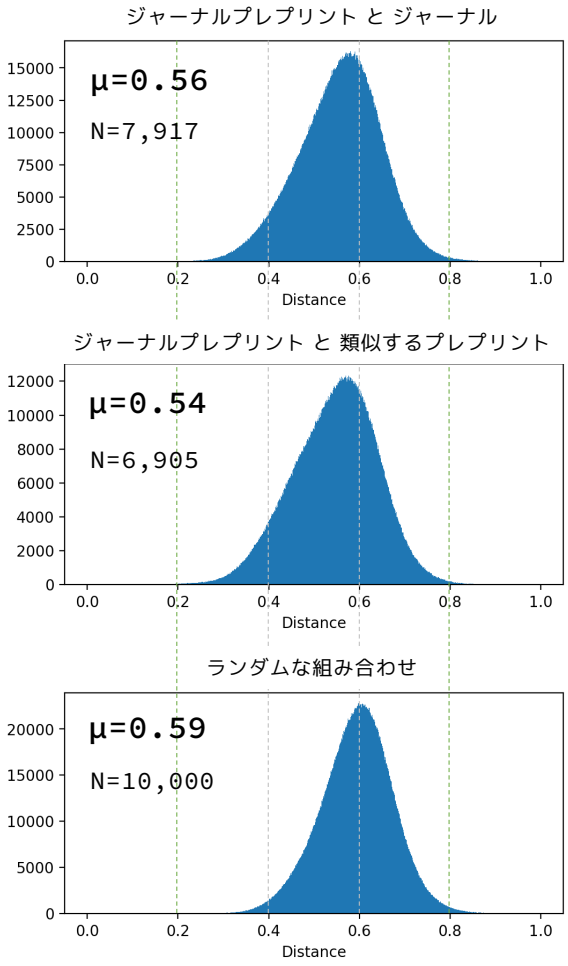

図 2　全段落組み合わせにおける距離の分布

5.2　内容の類似度

前述の閾値に基づいて算出した文書間の Jaccard 係数の分布を図 3 に示す.

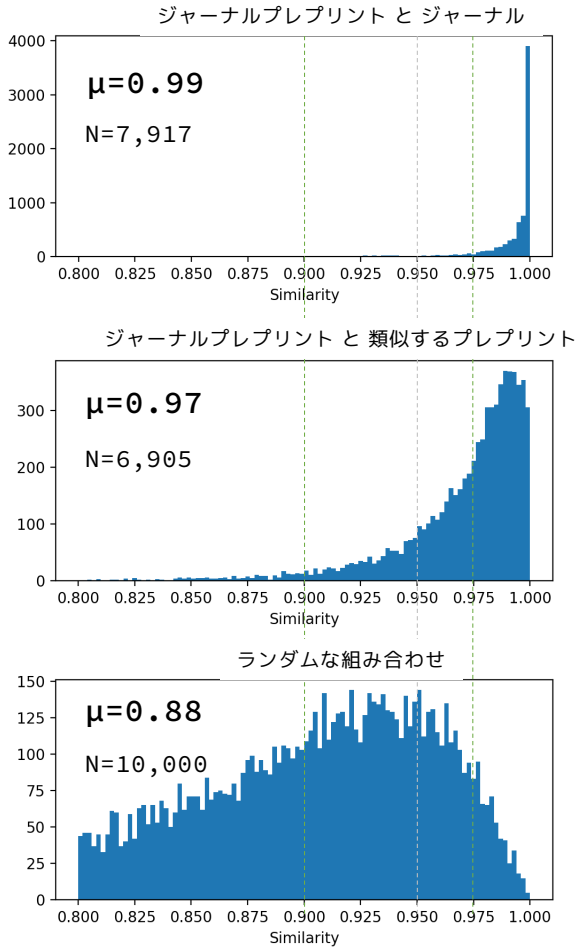

図 3　文書間類似度の分布

図 3 を見ると，まず，ベースラインとなるランダムマッチでは 0.95 以上の類似度を持つものは少なく，0.93 あたりをピークにする分布を描いている．ジャーナルプレプリントとジャーナル論文は当然ながら類似度が高く 1.0 にピークを有する．ジャーナルプレプリントと類似するプレプリントでは類似度 1.0 など高い類似度を示すものもあるが なだらかに分布しており，基本的に著者自体が異なることを考えると自然である．

5.3　その他，外形的な基準

著者数や参考文献数など，内容の類似度以外の外形的な基準について以下に示す．

著者数，参考文献数の差および実数分布を図 4,5 に示す．参考文献について “ジャーナル論文” とは綺麗に分布がシフトしており，ジャーナル論文になる際に参考文献が増える傾向が確認できる．

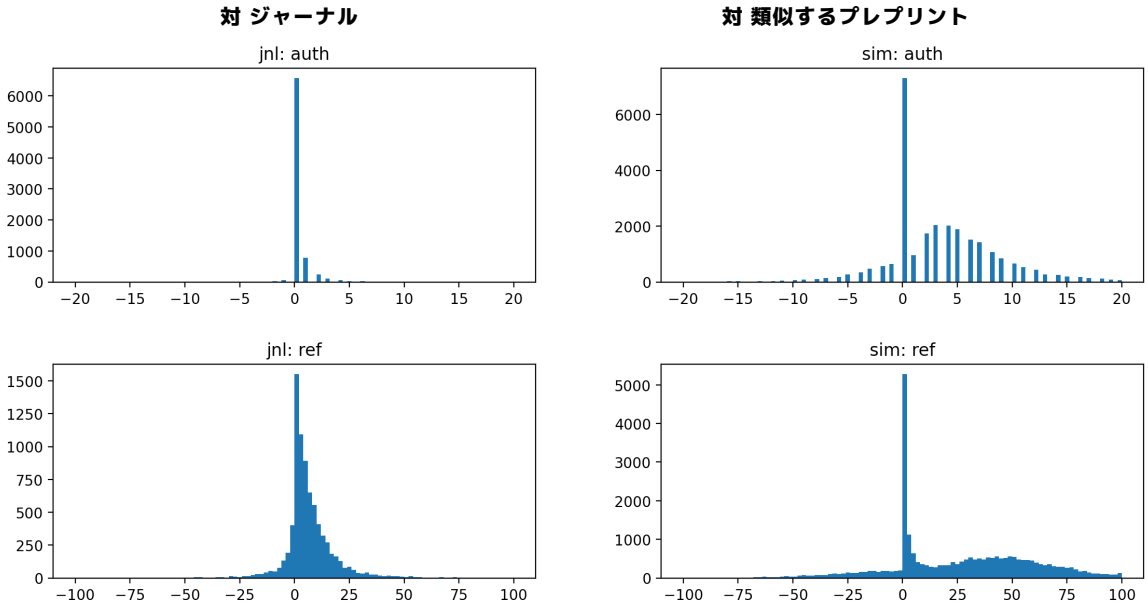

図 4　著者数，参考文献数の差の分布

図表数の差および実数分布を図 6,7 に示す．図 6,7 中，tab は JATS-XML の table タグを，tabw は table-wrap タグを意味する．プレプリント側では table タグがあまり使われず，table-wrap が多用されている傾向があるため，table, table-wrap の取り扱いには留意を要する．

段落数・単語数の差および実数分布を図 8,9 に示す．図 9 をみるとジャーナル論文と比較した際に，特にわかりやすく分布が右にシフトしておりジャーナル論文化する際に段落数・単語数が増加していることが分かる．図 8 からも段落数・単語数が減るケースは希であることが分かる．

章のタイトルについて，同一のもの，異なるものの数の分布を図 10 に示す．

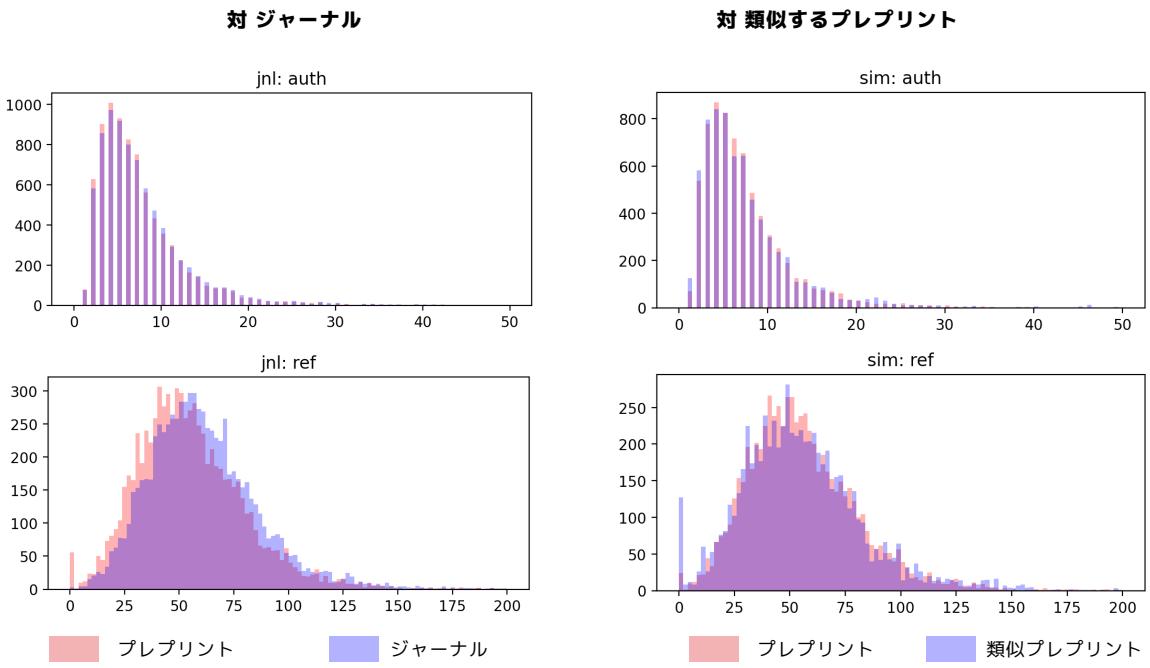

図 5　著者数，参考文献数の分布

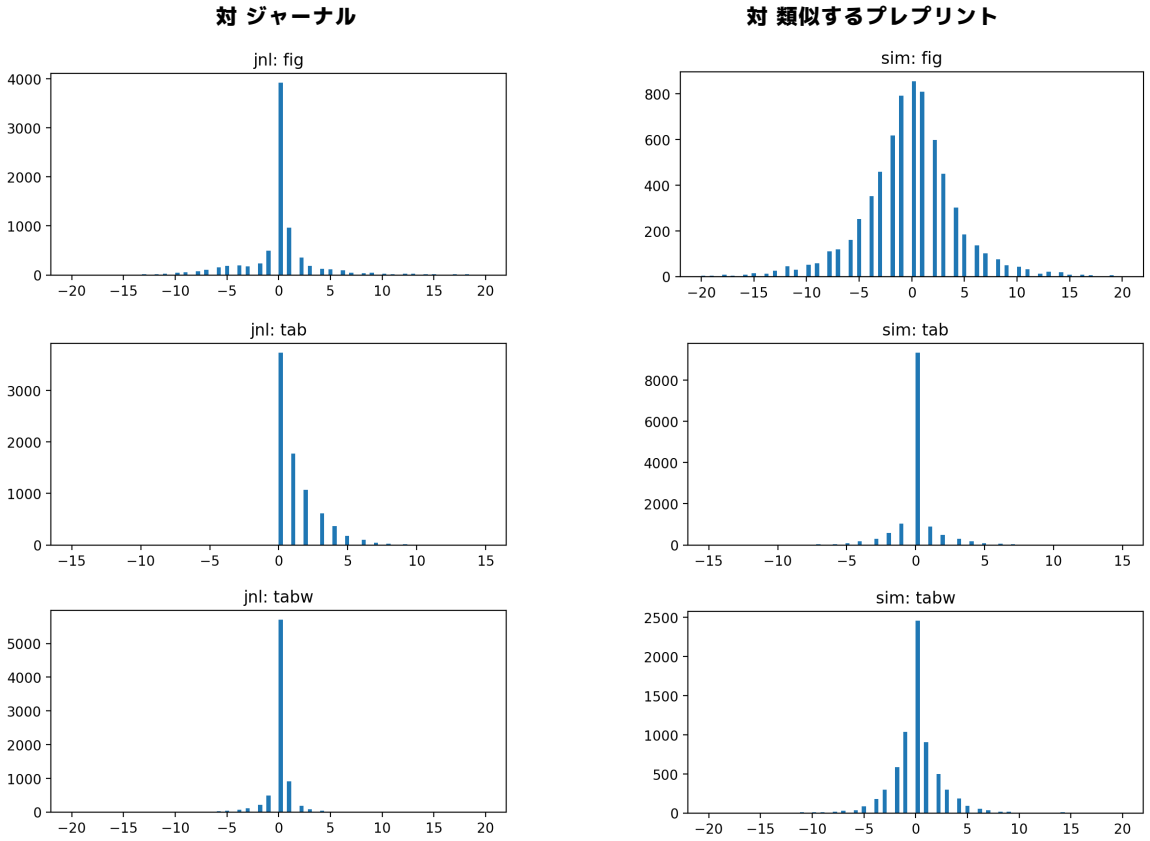

図 6　図表数の差の分布

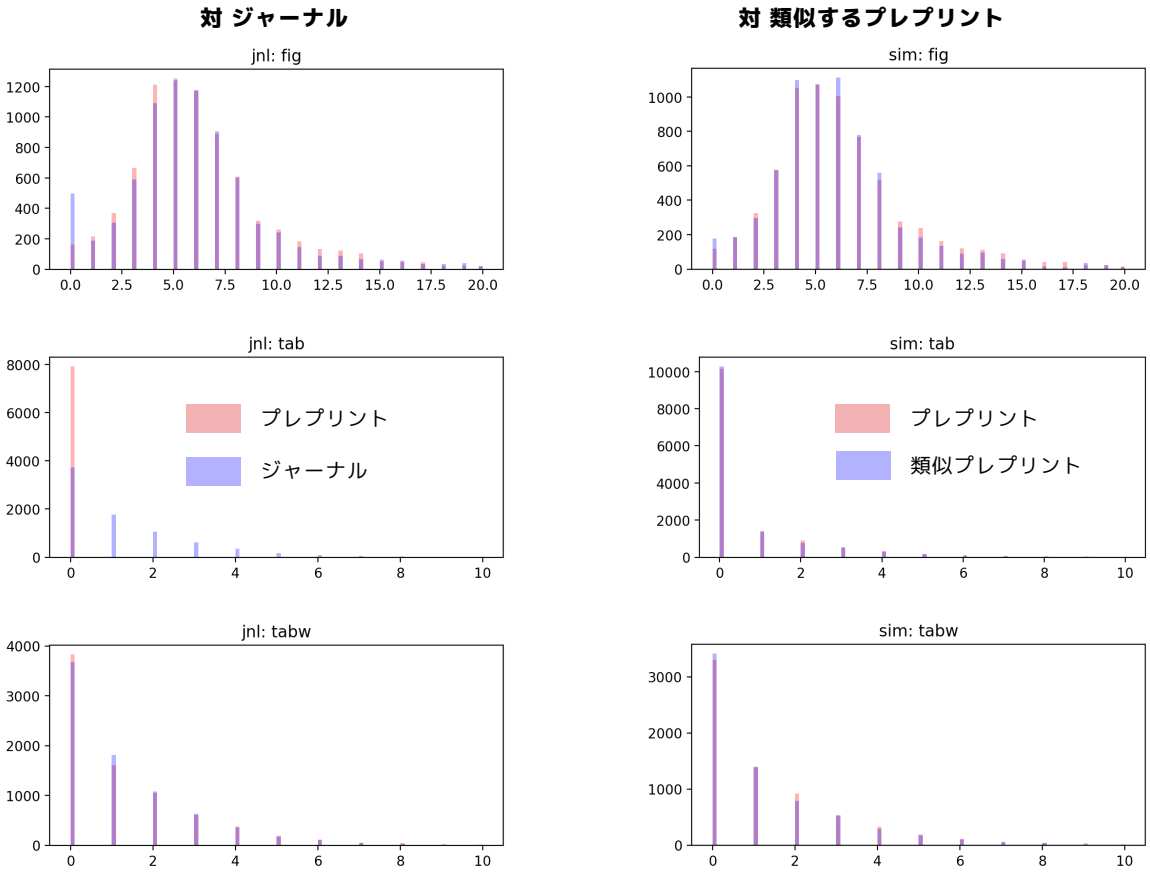

図 7　図表数の分布

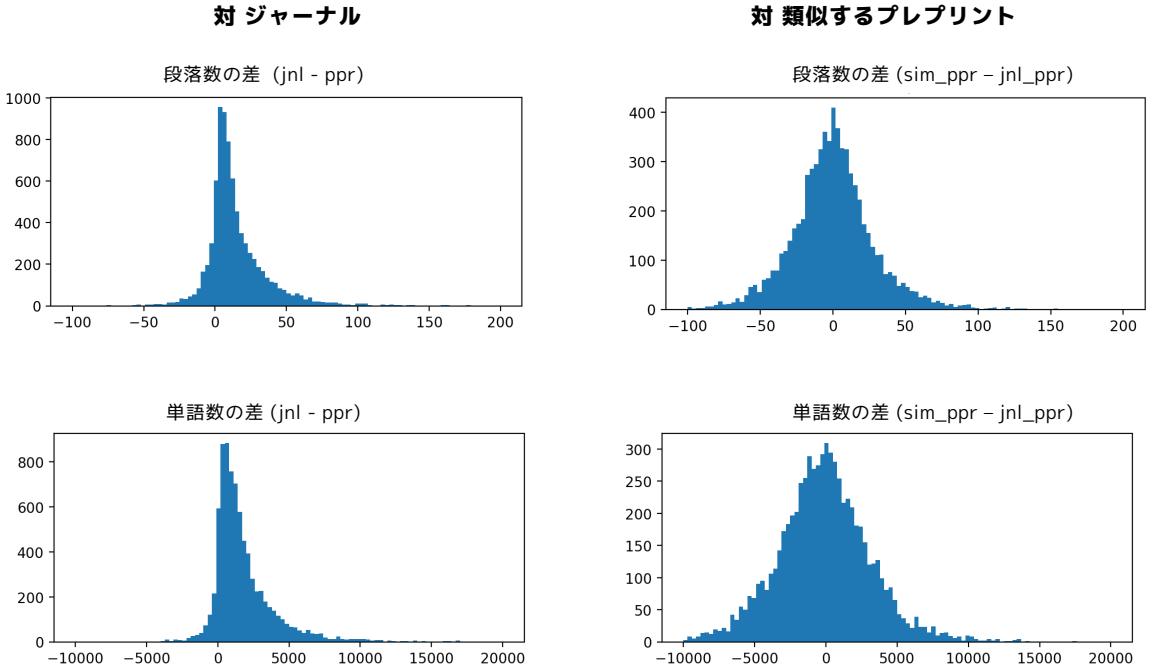

図 8　段落数・単語数の差の分布

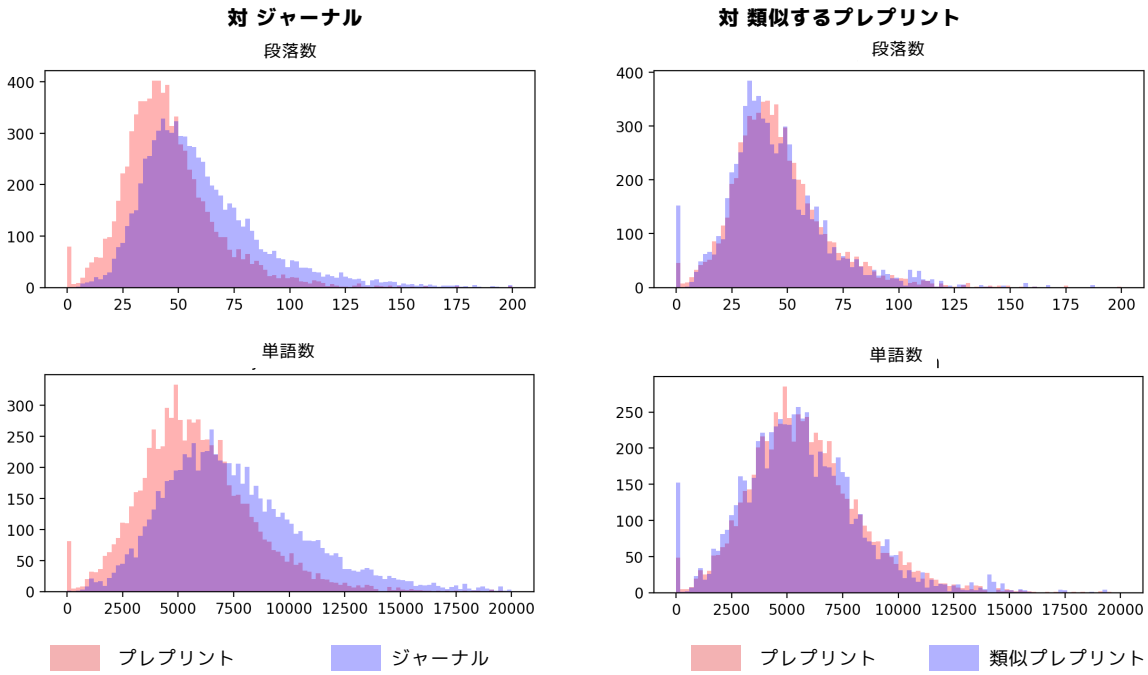

図 9　段落数・単語数の分布

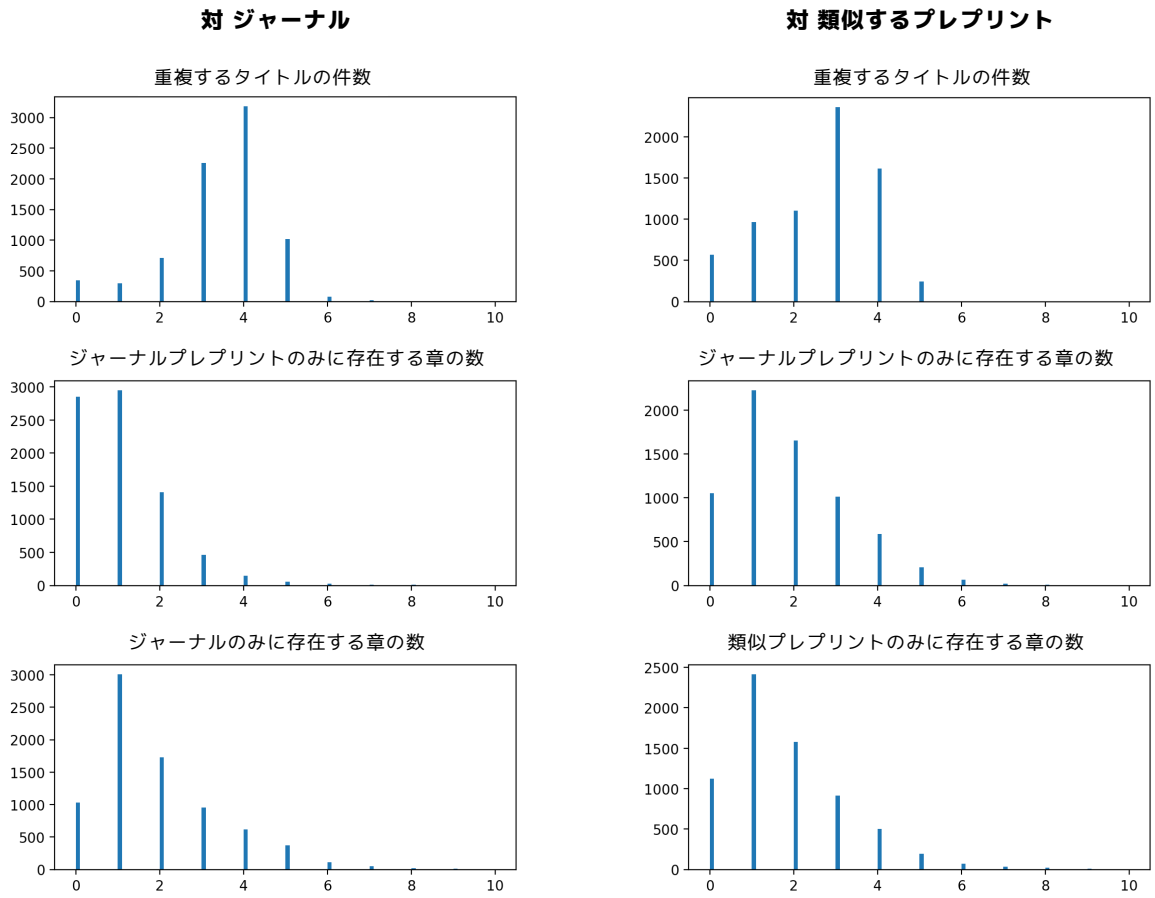

図 10　章のタイトルの和・差集合の分布

5.4　タイトルの差に関する調査

タイトルについては単純に数を見るだけでなく，重複するタイトルや異なるタイトルについても調べた．

5.4.1　章タイトルの差

章タイトルの差については，共通するもの，ジャーナルプレプリントのみに存在するもの，ジャーナル論文もしくは類似するプレプリントのみに存在するもの，についてそれぞれワードクラウドとして整理した．

この際，タイトルの前の “1. ” などの番号やタイトル最後のピリオドや空白は除去し，大文字小文字についても無視して比較している．

ジャーナルプレプリント - ジャーナル についての調査結果を図 11,12,13 に示す．

図 11　重複するもの（ジャーナル）

図 12　ジャーナルプレプリントのみ（ジャーナル）

図 12,13 を比較すると頻出語は似通っており，例えば Material が消されるケースも足されるケースもあることが分かる．

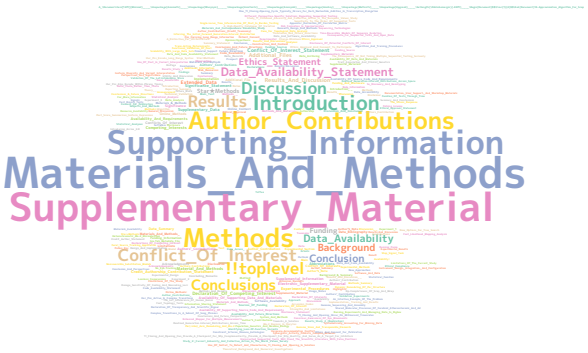

図 13　ジャーナルのみ（ジャーナル）

図 14　重複するもの（類似）

図 15　ジャーナルプレプリントのみ（類似）

ジャーナルプレプリント - 類似するプレプリント についての調査結果を図 14,15,16 に示す．

図 15,16 を比較するとやはり頻出語は似通っており，例えば Material が消されるケースも足されるケースもあることが分かる．ただし図 14 は特定の単語への偏重を示しており，プレプリントに共通する章立ての構造の存在は伺える．

図 16　類似するプレプリントのみ（類似）

5.4.2 タイトルの差

ジャーナルプレプリント - ジャーナル については，章タイトルではなく，論文タイトルそのものに差があるかも調査した．

ここでは，査読や校正を経てタイトルにどういう変化があるかを観察したいこと，一般にタイトルのバリエーションは多様であり重複は意味が薄そうなことから，タイトルの単語セットがどの程度一致しているか (Jaccard) と具体の単語の差分のみを調査した．

結果を図 17,18,19 に示す．

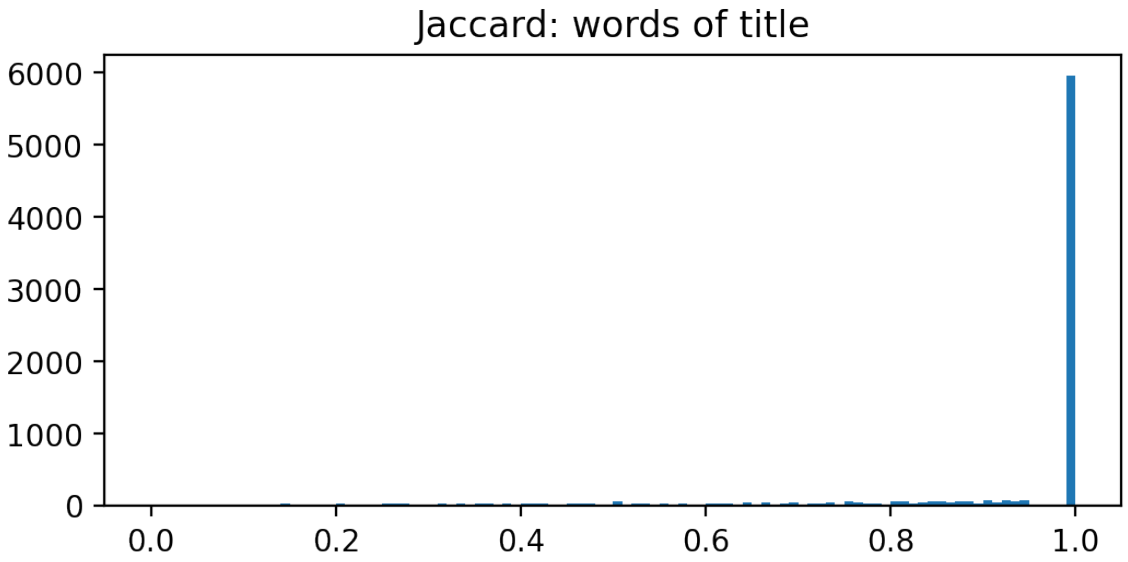

図 17　Jaccard 係数の分布

図 17 によると，単語セットの比較では多くのケースで差はない (Jaccard 係数が 1.0)，また具体の単語の差についても章タイトルと同様，図 18,19 の間に大きな変化は見られない．つまり，定冠詞がつくケースもあれば消されるケースもあるなど，ケースバイケースで，一定の傾向はこのレベルでは観察できなかった．

図 18　ジャーナルプレプリントのみ

図 19　ジャーナルのみ

5.5　外形的な基準に関する重要度の推定

ここまでで，ジャーナルプレプリントとジャーナル論文の間には，単語数や参考文献数に多少の差が見られたものの大きな差は見られなかった．

ただし，一つの指標で差がないとしてもいくつかの条件と複数の指標の組み合わせで差が見られることもある．そこで，機械学習を用いてジャーナル論文，ジャーナルプレプリント，その他のプレプリントを判別できるか調査した．

ここでは，ジャーナル論文はこれまで用いてきたものをそのまま用いるが，ジャーナルプレプリントは範囲を変更し，更に “その他のプレプリント” という新たな区分を導入する．これらはサンプルのサイズを増やすための処置である．

まず，これまでジャーナルプレプリントはそのペアとなるジャーナル論文の全文 XML が取得できたもののみを対象としていたが，ここではペアは必要ないため，全文 XML の有無にかかわらずジャーナル論文と紐付いておりかつプレプリントの全文 XML が正常に取得できたもの 12,925 件をジャーナルプレプリントとした．また，ジャーナル論文に紐付いておらずプレプリントの全文 XML が正常に取得できたもの 14,673 件をプレプリントとした．

これらの種別について，外形的な基準から分類する分類機を生成し，各基準（特徴量）の重要度を調べる．基準についてはこれまでの分析内容を参考に以下を設定した．

変数名	詳細
auth	著者数
ref	参考文献数
word	総単語数
fig	図の数
tab	表 (table) の数
intro	章タイトルに “intro” を含む章の単語数が全体に占める割合
metho	章タイトルに “method” を含む章の単語数が全体に占める割合
resul	章タイトルに “result” を含む章の単語数が全体に占める割合
discu	章タイトルに “discuss” を含む章の単語数が全体に占める割合
concl	章タイトルに “conclusion” を含む章の単語数が全体に占める割合

表 1　分類に用いる特徴量

分類手法には Python の機械学習パッケージである scikit-learn の RandomForest を用い，全データからランダムに抽出した 70% を学習データ，30% をテストデータとして処理した．

結果を以下に示す．

Accuracy は 0.47，特徴量の重要度は図 20 の通りである．

事前知識がなにもない場合，分類数が 3 なので単純に 1/3 の確率で当たるとすると Accuracy

		Exact		
		Journal	Journal-PPr	PPr
Predict	Journal	1,156	595	524
	Journal-PPr	1,003	1,922	1,863
	PPr	834	2,695	3,487

* PPr: PrePrint

表 2　RandomForest によるテストデータの分類結果

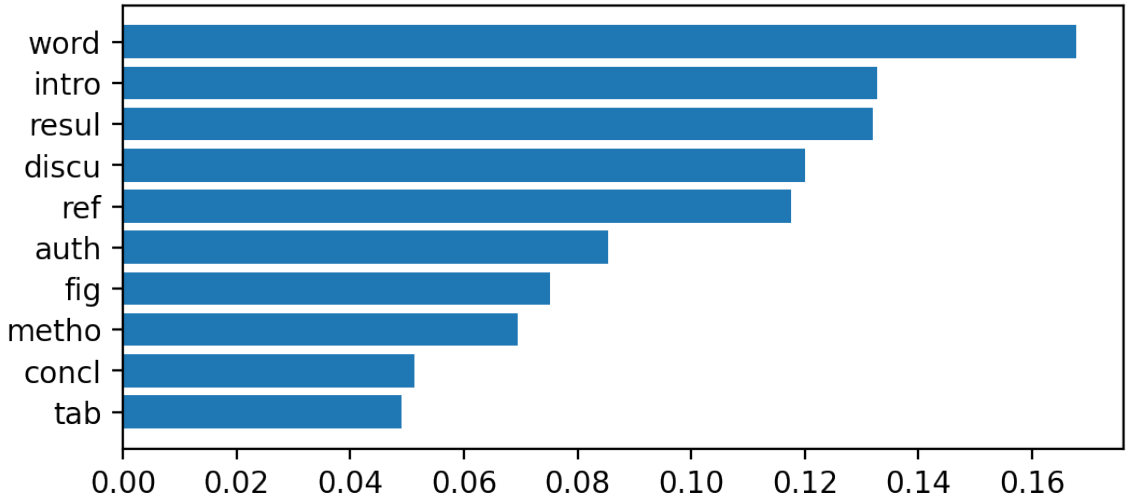

図 20　特徴量の重要度

0.47 はそれを上回るが決して良い精度ではない．また，判別には単語数が最も効いており，次いで導入部，結論部が全体に占める割合が大きい，という結果になっている．

6　考察

今回の分析から表現上の大筋ではジャーナルプレプリントとジャーナル論文，類似するプレプリントの間で差は比較的小さいことが読み取れる．つまり，分野の専門知識を有しないような人間がジャーナルプレプリントとジャーナル論文，類似するプレプリントを示されたときに，章の構成や参考文献の数などを手がかりとして，どれがジャーナル論文かを見分けることは難しいと考えられる．

また，ジャーナルプレプリントとジャーナル論文の内容的な異なりも査読・出版の前後の比較として相対的に小さいことも分かる．ただし，今回の手法ではランダムマッチした場合でも内容の類似度分布においてピーク値は 0.93 程度と高い．また，類似するプレプリントとの比較でも類似度 1.0 のものが少なくない．このように今回の基準は比較的類似度を高めに見積もる傾向がある．さらに，今回の類似度算定手法では数式や数値が切り捨てられるが，論文においてはそれらのわずかな差が大きな意味を持つこともある．

このほか，結果の読み取りには後述する種々の留意点が存在するが，単純化して議論するならば，今回分析した範囲ではプレプリントとジャーナル論文を外形的に識別することは難しいことが分かった．

ジャーナルプレプリントとジャーナル論文の間の差はほとんど見られないという結果は，Klein[Klein19] や Carneiro[Carneiro20] の主張を支持するものである．

従って，当初の疑問点である「ジャーナル論文はどのようにできあがっていくのか」に関してはより踏み込んだテキストマイニングを行うとともに，Carneiro[Carneiro20] や Akbaritabar[Akbaritabar21] と同様に分野の専門家を加えてプレプリントとジャーナル論文の差異を検討する必要がある．

6.1　留意点

今回はプレプリントサーバに bioRxiv を選定した．従って分野としては生物学系のみに限定される．さらに，生物系は他分野に比較すると活発とはいえ分野の研究者数や物理・情報系の arXiv の利用率を考えると生物系分野内での bioRxiv の普及率は高いとはいえない．この点と相まって，今回調査できたジャーナル論文は 7 千件程度とごく少数で，さらにオープンアクセス論文のみになっている．したがって，生物系の中でもプレプリントを使う比較的先進的なユーザ層の，オープンアクセス論文に限った議論になっている点に十分な留意が必要である．

また，一部内容も分析しているものの機械的に処理しており，具体の中身や文脈は考慮できていない．現実には単語のセットが同一であったとしても，並びが変われば意味が全く異なることもあり，文章がほとんど同じでも細かい数値の違いが大きなな意味を持つこともある．こうした点にも留意が必要である．

プレプリントとジャーナル論文の “質” についても慎重な議論を要する．ジャーナル論文になっ

ていないから質・価値が低いといえるかどうかは分からない．単にジャーナル論文に投稿していないだけで投稿すれば採録されたものもあるかもしれないし，投稿したが査読者の専門分野がたまたま異なっていて正当に評価されずプレプリントのままなのかもしれない．特に社会課題解決研究において，社会的にはインパクトを持つ論文が科学的な新規性や速報性を認められずに査読が通らないことが起こりうる．あるいは，現状は評価されていないが，数十年後に何かの成果と合わさると大きな価値を生み出すものもあるかもしれない．ジャーナル論文になっていない理由にも多様なものが考えられる．本稿ではプレプリントとジャーナル論文，プレプリント間の違いを調べたが，単に違いを調べただけであって，それらが研究の本質にどのように効いているのかまでは調べられていない．追加された参考文献は内容にどの程度クリティカルに効いているのか，章のタイトルや章立ての変更は内容にどのように影響したのか，その結果，研究の価値はどの程度変化したのか，明らかにはしていない．単に「こんな違いが，この程度ある」と示したにとどまっている点に留意が必要である．

7　まとめ

本稿では，研究はどのように行われるのか，特にジャーナル論文はどのようにできあがっていくのか，という疑問に対して，プレプリントとそれが最終的に出版されたジャーナル論文とを比較することで知見を得られないか試行した．

まず，近年のオープンジャーナルなどの流れもあって，bioRxiv のプレプリントとジャーナル論文の全文 XML を一定量確保することができ，プレプリントとジャーナル論文の比較に関する技術的な可能性は検証できた．

他方，参考文献数や単語数などの外形的な基準や，簡単な文書類似度から両者の差分を明らかにしようとした今回の試行の範囲では，プレプリントとジャーナル論文，ジャーナル論文になったプレプリントとそうではないプレプリントの間で明確な違いを見いだすことはできなかった．機械学習手法を用いても分類精度は 47% 程度と高くなかった．

プレプリントとジャーナル論文の間に大きな差はないという結果は先行研究 [Klein19, Carneiro20] でも示されている知見であり，より大規模かつ比較的最近の状況でも再現性が確認できた．これらに加えて本稿のもたらした新たな知見としては，著者数などの多くの外形的的基準においても差は小さい，ジャーナル論文になっていないプレプリントとの違いも大きくない，といった点が挙げられる．

以上より，ジャーナル論文はどのようにできあがっていくのか，プレプリントの段階からどういった点がブラッシュアップされて論文になるのか，ジャーナル論文として採録されるものと採録されないものの決定的な違いはなにか，といった点について検証を進めるためには，分野の専門家も加えた上でより高度なテキストマイニング等も駆使して分析する必要性が示唆された．

派生的な示唆としては，プレプリントは分量をはじめとして外形的にはジャーナル論文と同等であって，プレプリントの中からジャーナル論文に採択されそうなものをその外形的な基準から見つけるようなことは難しそうであることが伺える．また，本研究のアプローチを突き詰めていくこと

で，査読というシステムの必要性や意義を改めて見直せる可能性がある．

参考文献

[Abdill19] Abdill, R.J and Blekhman, R.: Meta-Research: Tracking the popularity and outcomes of all bioRxiv preprints. *eLife*. Apr 2019. https://doi.org/10.7554/eLife.45133

[Akbaritabar21] Akbaritabar, A, and Stephen, D： A Disciplinary View of Changes in Publications' Reference Lists After Peer Review. *arXiv* . Feb 2021. arXiv:2102.03110

[Carneiro20] Carneiro, C.F.D., Queiroz, V.G.S., Moulin, T.C. et al.： Comparing quality of reporting between preprints and peer-reviewed articles in the biomedical literature. *Research Integrity and Peer Review* Vol.5, Article number 16. Dec 2020. https://doi.org/10.1186/s41073-020-00101-3

[Klein19] Klein, M., Broadwell, P., Farb, S.E. et al.：Comparing Published Scientific Journal Articles to Their Pre-Print Versions. *International Journal on Digital Libraries* Vol.20, pp.335 – 350. Dec 2019. https://doi.org/10.1007/s00799-018-0234-1

[MEXT20] MEXT–NISTEP プレプリント調査・検討チーム：プレプリントをめぐる近年の動向及び今後の科学技術行政への示唆. **文部科学省 科学技術・学術審議会 情報委員会 ジャーナル問題検討部会 第 *7* 回 配布資料** 資料 1-別添, Oct 2020. https://www.mext.go.jp/content/20201026-mxt_jyohoka01-000010684_2.pdf

[小柴 20] 小柴　等，林　和弘，伊藤裕子： COVID-19 / SARS-CoV-2 関連のプレプリントを用いた研究動向の試行的分析． *NISTEP Discuttion Paper,* No.186, June 2020. http://doi.org/10.15108/dp186

[林 20] 林和弘, 小柴等： arXiv に着目したプレプリントの分析. *NISTEP DISCUSSION PAPER* Vol.187, Aug 2020. https://doi.org/10.15108/dp187

[林 21] 林和弘, 小柴等： bioRxiv に着目したプレプリントの分析. *NISTEP DISCUSSION PAPER* Vol.1XX, Aug 2021. https://doi.org/10.15108/dp1XX

付録 A　被引用数の比較

被引用数はどのタイミングで計測するかで評価が異なるなど，指標として用いるには課題が多いため本編では使用していないが，2021 年 4 月末時点での被引用数についてジャーナルプレプリントとジャーナル掲載のないその他のプレプリントを比較すると図 21 の様になる．

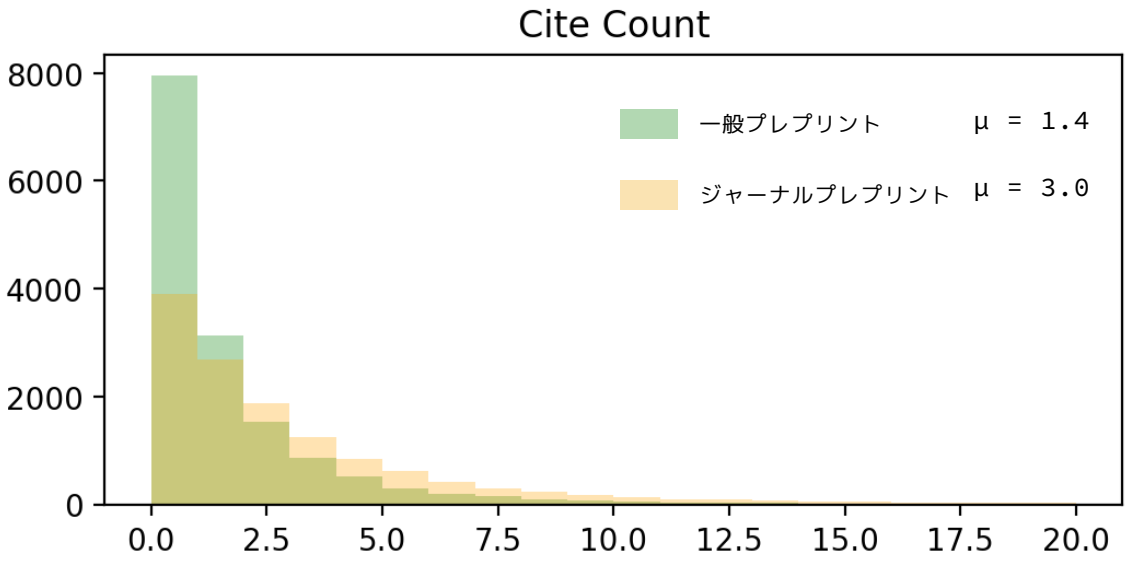

図 21　被引用数の比較

ジャーナルプレプリントの方が引用数が大きいが，ジャーナルに掲載されたことで注目を受け引用されているのか，よく引用されているものがジャーナルに掲載されやすいのか，といった理由については不明である．

付録 B　版数の比較

版数（更新回数）も被引用数と同様，指標として用いるには課題が多いため本編では使用していないが，2021 年 4 月末時点での更新回数についてジャーナルプレプリントとジャーナル掲載のないその他のプレプリントを比較すると図 22 の様になる．

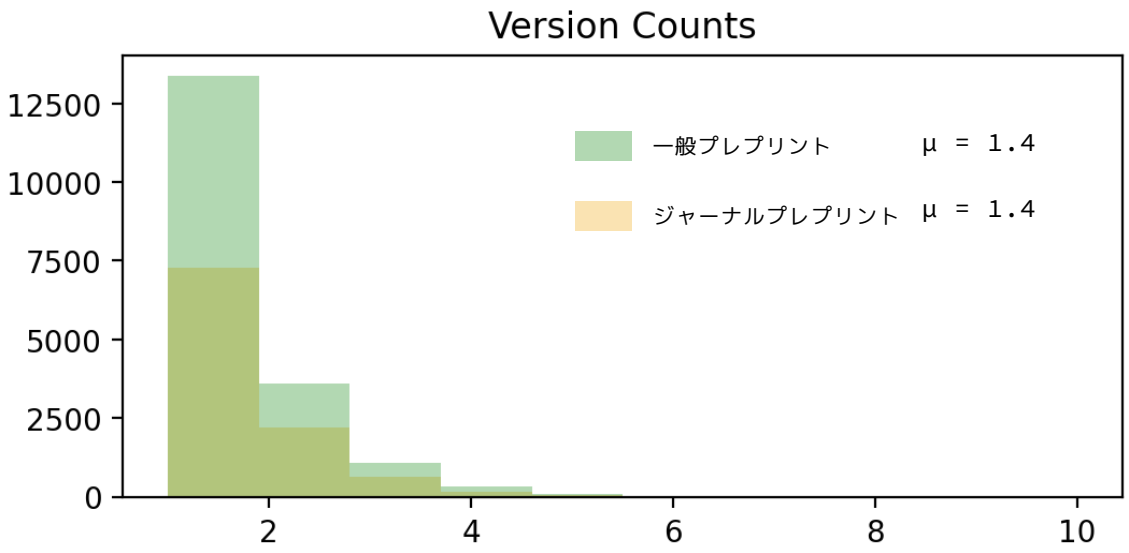

図 22　版数（更新回数）の比較

ジャーナルプレプリント，その他のプレプリントの双方で平均 1.4 回と差は見られず，傾向も一致している．

DISCUSSION PAPER No.200

プレプリントとジャーナル論文の差異：bioRxiv を用いた試行

2021 年 8 月

文部科学省 科学技術・学術政策研究所 データ解析政策研究室
小柴　等,　林　和弘

〒100-0013　東京都千代田区霞が関 3-2-2 中央合同庁舎第 7 号館 東館 16 階
TEL: 03-3581-2393

Differences between preprints and journal articles: a trial using bioRxiv

Aug 2021

KOSHIBA Hitoshi and HAYASHI Kazuhiro

Research-Unit for Data Application
National Institute of Science and Technology Policy (NISTEP)
Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan

https://doi.org/10.15108/dp200

https://www.nistep.go.jp