科学論文における要旨の構造とその学術的引用および社会的注目との関連

本那 真一 [1]　橋本 敬 [1]
[1] 北陸先端科学技術大学院大学 先端科学技術研究科
{honna-shinichi, hash}@jaist.ac.jp

概要

本研究は，テキスト構造が科学論文の引用やソーシャルメディアを通じた普及にどのように関連しているかを検証する．4 つのオープンアクセスジャーナルから得られた 40,000 件のアブストラクト（要旨）を対象に，統語木および談話木を用いて分析を行った．構造だけで普及の結果すべてを説明できるわけではないが，一貫したパターンが明らかになった．学術的引用においては，受動態や過去分詞を多用する統語構造や，対比と詳述を組み合わせた談話構造が正の影響を与えていた．一方，社会的普及（公的言及）においては，正の影響を与える構造的特徴は見出されず，むしろ学術的に好まれる「対比構造」が負の影響を与えるケースが確認された．これらの結果は，専門家コミュニティと一般社会とでは，評価されるテキスト構造が質的に異なることを示唆しており，引用分析やオルトメトリクスに関する先行研究に新たな定量的証拠を提供するものである．

1　序論

科学的知識の普及は，知見が共有され，再利用されることで初めて達成される [1, 2]．この普及プロセスには，学術コミュニティ内での「学術的受容」と，社会一般への「社会的受容」という二つの側面がある．学術界において被引用数は知識蓄積への貢献度を示す指標となる一方 [1]，科学的知見は政策や技術開発を通じて社会にも影響を与えるため，オルトメトリクス（公的言及）を通じた社会的普及の理解も不可欠である [3, 4].

これらの普及において，論文のテキスト構造は重要な役割を果たす．修辞学やジャンル分析の研究は，語彙選択や議論の構成が読者の理解や説得力に影響することを示唆している [5]．しかし，科学的ライティングの構造分析には課題が残されている．従来の物語論的アプローチ [6, 7] や近年の NLP による物語分析 [8, 9] は，フィクション特有の要素に依存しており，論理性を重んじる科学論文には適用しにくい．一方で，科学論文に特化した既存研究の多くは，語彙難易度や文長といった表層的特徴の分析に留まるか [10, 11]，事前に定義されたラベル（背景・目的等）に依存する手法 [12] であり，データ駆動的に微細な構造パターンを捉えるには至っていない．

そこで本研究では，媒体固有のラベルや表層的特徴への依存を避け，データ駆動的に構造的組織化を比較できる枠組みを提案する．具体的には，文が内部的にどう組織化されているかを示す「統語レベル」[13] と，文がどのようにつながって意味を形成するかを示す「談話レベル」[14] の双方からテキスト構造を抽出する．このアプローチにより，以下のリサーチクエスチョン（RQ）に取り組む．

RQ1: 科学論文のテキスト構造は，学術的普及（引用）とどのように関連しているか？

RQ2: 科学論文のテキスト構造は，社会的普及（公的言及）とどのように関連しているか？

2　データ

分析対象は，自然科学分野の主要なオープンアクセスジャーナル 4 誌（eLife，Scientific Reports，Nature Communications，PLOS ONE）からランダムに抽出した計 40,000 件のアブストラクトである．これらは幅広い分野を網羅し，かつ一般読者へのアクセス性が高いため，社会的普及の分析に適している [15].

普及の指標として，以下の 2 変数を使用する．

1. **被引用数**: 学術的普及の指標．OpenAlex API より取得 [1].
2. **公的言及**: 社会的普及の指標．Altmetric データベースの `cohort_pub`（ニュース，ブログ，SNS 等での言及数）を使用 [16].

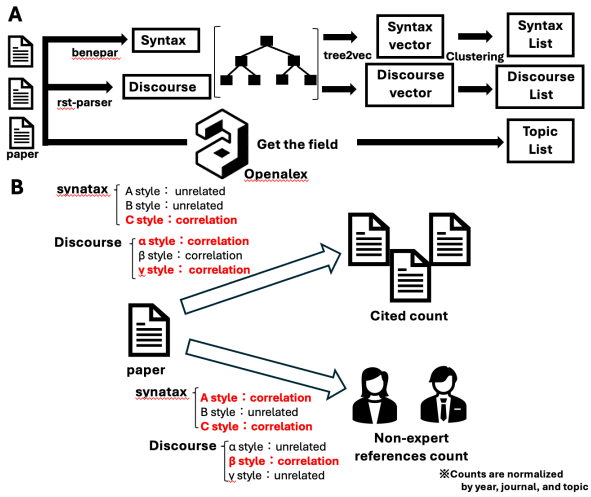

図 1　分析のフレームワーク．(A) アブストラクトから統語・談話構造を抽出しベクトル化する．(B) クラスタリングと回帰モデルを用いて普及との関係を検定する．

なお，普及数は研究分野，出版年，ジャーナルの影響を強く受けるため [17]，本分析ではこれらをモデルのオフセット項および固定効果として組み込み，構造的要因の影響を厳密に分離して評価する．データセットの詳細な分布については，付録の図 4 に示す．多くの論文は注目度が低いが，ごく一部が高い値を占めるロングテール分布が確認できる．

3　手法

本研究では，前処理による構造特徴の抽出と，統計モデルによる普及要因の分析という 2 段階のアプローチをとる（図 1 参照）．

3.1　前処理：構造特徴の抽出

まず，spaCy [18] を使用してテキストのトークン化を行い，以下の 2 つのレベルで構造を抽出した．

統語構造 (Syntactic Structure)　benepar [19] を使用して各文の句構造木を取得した．本研究では，具体的な単語の影響を排除し，純粋な構造的特徴のみを抽出することを目的としている．そのため，解析された木の葉ノードにある具体的な単語を一般的なトークンに置換し，句ラベル（NP，VP など）の階層構造のみを保持した．各文の木を文書レベルとして統合し，Tree2Vec を用いて構造的配置を捉えた密なベクトル表現を学習した．Tree2Vec は，木構造内の部分木をトラバースし，文脈に応じた構造の分散表現を獲得するため，頻出する構文パターンを効率的に捉えることができる．

談話構造 (Discourse Structure)　IsaNLP RST パーサー [20] を使用して，因果関係，詳述，対比といった修辞関係を特定し，談話木を構築した．これにより，文がいかに修辞的に接続され，局所的な記述が全体の議論にどう寄与しているかを捉えた．これも同様に Tree2Vec によりベクトル化した．

最後に，各アブストラクトに対して OpenAlex API [21] から取得した分野および書誌的メタデータを付与した．

3.2　モデリング

得られた統語的・談話的埋め込みを，ディリクレ過程混合ガウスモデル（DP-GMM）を用いてクラスタリングした．DP-GMM はノンパラメトリックな手法であり，データの複雑さに応じてクラスター数 K を自動的に決定できる利点がある．これにより，人為的な設定に依存せず，各アブストラクトの支配的な構造傾向を捉えるカテゴリカル変数を生成した．

普及の結果（被引用数と公的言及数）は，負の二項分布を用いた一般化線形モデル（NB-GLM）を使用してモデル化した．引用データは一般に平均よりも分散が大きい「過分散（overdispersion）」の性質を持つため，ポアソン分布ではなく負の二項分布が適している．各文書 i の観測された普及数 y_i は，平均 μ_i と過分散パラメータ α を持つ負の二項分布に従うと仮定される：

$$y_i \sim \mathrm{NegBin}(\mu_i, \alpha) \tag{1}$$

対数リンク関数により，期待度数を以下の線形予測子に関連付ける：

$$\log(\mu_i) = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \log(\mathrm{offset}_i) \tag{2}$$

ここで，$\mathbf{x}_i$ には統語および談話クラスターのダミー変数に加え，ジャーナルおよび研究分野の固定効果が含まれる．時間的露出（temporal exposure）の違いを調整するため，記事年齢に基づくオフセット項 $\log(\mathrm{offset}_i) = \log(y_{\max} - y_i + 1)$ を導入した．

モデルの評価には対数尤度，AIC，BIC を用い，構造情報を含まないモデルと比較することで，テキスト構造の説明力を検証した．

表 1　統語クラスターの特徴と解釈

ID	言語学的特徴と解釈
0	**標準的構文 A**: 名詞句・動詞句中心．形容詞による修飾が顕著.
1	**副詞修飾**: 名詞句・動詞句に加え，副詞句による論理的な修飾を含む.
2	**標準的構文 B**: Cluster 0 と類似．形容詞修飾を含む基本的な文構造.
3	**従属節**: SBAR や WHNP が頻出し，複文構造による詳細記述を行う.
4	**複雑な名詞句**: 関係詞や括弧を用いた高度に階層的な名詞句修飾が顕著.
5	**受動態・過去分詞**: VBN や受動態構文が顕著．客観的な記述.
6	**前置詞句**: 名詞句への前置詞句（PP）による修飾が支配的.
7	**過去形**: 過去形（VBD）が中心．実験手続きや観察結果の記述.
8	特徴的な規則性は検出されず.
9	特徴的な規則性は検出されず.
10	特徴的な規則性は検出されず.

表 2　談話クラスターの特徴と解釈

ID	解釈
0	**対比と並列**: 明示的な対比を中心に組織され，並列構造と追加の詳述を含む.
1	**対比と詳述**: 対比的な記述が一般的であり，その後に説明や追加の詳細が続く.
2	**詳述と帰属**: 論点の詳述に加え，出典や主体への帰属を含む.
3	**結合と要約**: 複数の並列要素と，それらの要約.
4	**結合と詳述 A**: 並列的な記述が優勢で，補足的な詳述を伴う.
5	特徴的な関係性は検出されず.
6	特徴的な関係性は検出されず.
7	**詳述のみ**: 詳述（詳細追加）関係によって特徴付けられる.
8	特徴的な関係性は検出されず.
9	**結合と詳述 B**: リスト化されたセグメントとその詳述.
10	特徴的な関係性は検出されず.

4　結果

4.1　構造クラスターの特徴

抽出された全ての統語・談話クラスター（ID 0〜10）の言語的解釈を表 1 および表 2 に示す.

4.2　テキスト構造が学術的普及に与える影響

被引用数（学術的普及）に対する完全モデルの対数尤度は -1.8684×10^5 であった．構造変数を投入することでモデルの適合度は有意に改善し（尤度比検定 $p < .001$），AIC も低下した．これは，テキスト構造が引用に対して一定の説明力を持つことを示している.

統語レベルの影響: 統語クラスターにおいては，**Cluster 5** が有意に正の影響を示した．表 1 より，Cluster 5 は過去分詞（VBN）や受動態構文を多用する特徴を持つ．この結果は，科学論文において「客観性」や「手続きの記述」を重視する受動態的な文体が，学術コミュニティ内での信頼性や評価に寄与していることを示唆している．一方で，特徴的な規則性を持たない **Cluster 8** および **Cluster 9** は，有意に負の影響を示した.

談話レベルの影響: 談話クラスターにおいては，**Cluster 1** が有意に正の影響を示した．表 2 より，Cluster 1 は「対比」と「詳述」によって構成されている．これは，先行研究とのギャップや対立点を明確にし（対比），その上で自らの研究内容を詳しく説明する（詳述）という論理展開が，学術的なインパクトを高める上で有効であることを意味している．対照的に，**Cluster 5, 6, 7, 8, 9** は有意に負の影響を示した．特に Cluster 7（詳述のみ）や Cluster 9（結合と詳述）が負であることは，単に情報を羅列したり詳細を追加したりするだけでは，引用を集めるには不十分であることを示唆している.

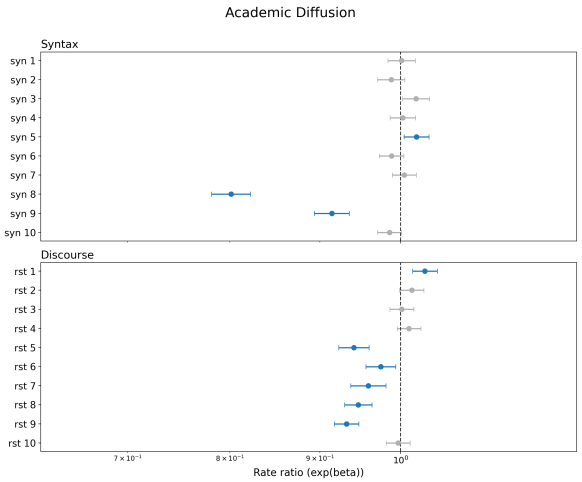

図 2　引用モデルにおける各クラスターの推定レート比（$\exp(\beta)$）．1.0 を超える値は引用への正の影響を示す.

4.3　テキスト構造が社会的普及に与える影響

公的言及（社会的普及）に対するモデルの説明力は低く（擬似 $R^2_{\mathrm{CS}} \approx 0.03$），構造的要因の影響は学術的普及に比べて限定的であった.

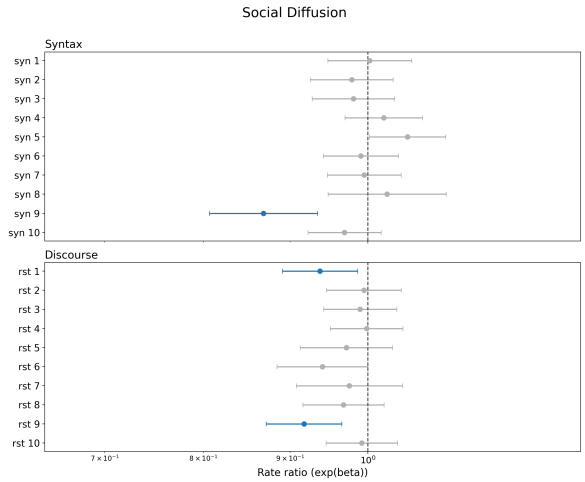

図 3　社会的普及（公的言及）モデルにおける各クラスターの推定レート比（$\exp(\beta)$）．ベースラインは $10^0 (= 1)$ である．

統語レベルの影響: 学術的普及で正の影響を示した Cluster 5 を含め，公的言及に対して有意に正の影響を与える統語クラスターは存在しなかった．一方で，**Cluster 9** は有意に負の影響を示した．このことは，SNS 等の社会的メディアにおいて，特定の統語スタイルが拡散を促進するわけではないが，不明瞭または不規則な構造（Cluster 9）は敬遠される可能性があることを示している．

談話レベルの影響: ここでも，有意に正の影響を与えるクラスターは存在しなかった．特筆すべきは，引用モデルにおいては正の影響を持っていた **Cluster 1（対比と詳述）** が，社会的普及モデルにおいては有意に **負** の影響を示した点である．加えて，**Cluster 9（結合と詳述）** も負の影響を示した．学術界で評価される「対比による論証」の構造は，一般社会向けのコミュニケーションにおいては，むしろ複雑さや難解さとして受け取られ，普及を阻害する要因となっている可能性がある．

5　考察

本研究の結果は，学術的普及と社会的普及という二つの異なるコンテキストにおいて，テキスト構造が果たす役割が質的に異なることを明らかにしている．

RQ1（学術的普及）に対する回答: 科学論文のテキスト構造は，学術的引用と明確に関連している．特に，受動態や過去分詞を用いた客観的な記述スタイル（統語 Cluster 5）や，既存の知見との対比を明確にした上で詳述を行う論理構成（談話 Cluster 1）は，被引用数を高める要因となる．これは，科学コミュニティが「客観性」と「新規性の明確化」という二つの規範を，テキスト構造の手がかりを通じて評価していることを示唆している．先行研究 [5] が指摘した「明確な目的記述の重要性」は，本研究における対比構造の有効性として，データ駆動的に裏付けられたと言える．

RQ2（社会的普及）に対する回答: 対照的に，社会的普及においては，正の影響を与える特定の構造は見出されなかった．むしろ，学術的に有効であった「対比と詳述（Cluster 1）」の構造は，社会的普及においては負の効果を示した．この逆転現象は，専門家と一般大衆の読み方の決定的な違いを反映している．専門家は論理的な対比やギャップの提示を好むが，一般の読者や SNS ユーザーにとって，そのような修辞的複雑さは認知的な負荷となり，共有への障壁となる可能性がある．社会的普及を決定づけるのは，アブストラクトの内部構造よりも，トピック自体の社会的関心度や，掲載ジャーナルのブランド力といった外的要因が支配的である．

5.1　実践的示唆

以上の知見は，研究者に対して「文体戦略の使い分け」を示唆するものである．純粋な学術的インパクト（引用）を最大化したい場合，受動態を用いて手続きを淡々と記述しつつ，「何が従来と違うのか」という対比構造をアブストラクトに明示的に組み込むことが推奨される．しかし，その戦略をそのまま社会的普及（プレスリリースや SNS での発信）に適用することは避けるべきである．社会的注目を集めるためには，学術的な論証構造に拘泥せず，トピックの重要性を平易に伝えることに注力すべきであり，あるいはアブストラクト構造以外の要素（タイトルや図表など）の工夫が必要となるだろう．

限界と今後の方向性: 本研究の限界として，アブストラクトのみを対象とした点，普及を静的なものとして扱った点，対象ジャーナルが英語圏の自然科学分野に限定されている点が挙げられる．今後は全文解析やマルチモーダル特徴の統合，時間的ダイナミクスを考慮した分析が求められる．しかしながら，本研究は構造的組織化と普及との関係を大規模かつ定量的に示した点で，科学的コミュニケーションの理解に新たな視座を提供するものである．

参考文献

[1] Eugene Garfield. Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science*, Vol. 178, No. 4060, pp. 471–479, 1972.

[2] Henk F Moed. *Citation analysis in research evaluation*. Springer, 2005.

[3] Jason Priem, Paul Groth, and Dario Taraborelli. The altmetrics collection. *PloS one*, Vol. 7, No. 11, p. e48753, 2012.

[4] Cassidy R Sugimoto, Sam Work, Vincent Larivière, and Stefanie Haustein. Scholarly use of social media and altmetrics: A review of the literature. *Journal of the association for information science and technology*, Vol. 68, No. 9, pp. 2037–2062, 2017.

[5] Ken Hyland. *Disciplinary discourses, Michigan classics ed.: Social interactions in academic writing*. University of Michigan Press, 2004.

[6] Vladimir Propp. *Morphology of the Folktale*. University of Texas press, 1968.

[7] Roland Barthes. Image-music-text. *Hill and Wang*, 1977.

[8] Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. The emotional arcs of stories are dominated by six basic shapes. *EPJ data science*, Vol. 5, No. 1, pp. 1–12, 2016.

[9] Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1534–1544, 2016.

[10] Sandeep Soni, David Bamman, and Jacob Eisenstein. Predicting long-term citations from short-term linguistic influence. *arXiv preprint arXiv:2210.13628*, 2022.

[11] Thomas S Jacques and Neil J Sebire. The impact of article titles on citation hits: an analysis of general and specialist medical journals. *JRSM short reports*, Vol. 1, No. 1, pp. 1–5, 2010.

[12] Simone Teufel and Marc Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, Vol. 28, No. 4, pp. 409–445, 2002.

[13] Noam Chomsky. *Aspects of the Theory of Syntax*. No. 11. MIT press, 2014.

[14] William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, Vol. 8, No. 3, pp. 243–281, 1988.

[15] Heather Piwowar, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, and Stefanie Haustein. The state of oa: a large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, Vol. 6, p. e4375, 2018.

[16] Mohammad Javed Ali. Understanding the altmetrics. In *Seminars in Ophthalmology*, Vol. 36, pp. 351–353. Taylor & Francis, 2021.

[17] Vincent Larivière, Stefanie Haustein, and Philippe Mongeon. The oligopoly of academic publishers in the digital era. *PloS one*, Vol. 10, No. 6, p. e0127502, 2015.

[18] Matthew Honnibal. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *(No Title)*, 2017.

[19] Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. *arXiv preprint arXiv:1805.01052*, 2018.

[20] Elena Chistova. Bilingual rhetorical structure parsing with large parallel annotations. *arXiv preprint arXiv:2409.14969*, 2024.

[21] Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

A　データセットの詳細

本研究で使用したデータセットの詳細な分布を図 4 に示す.

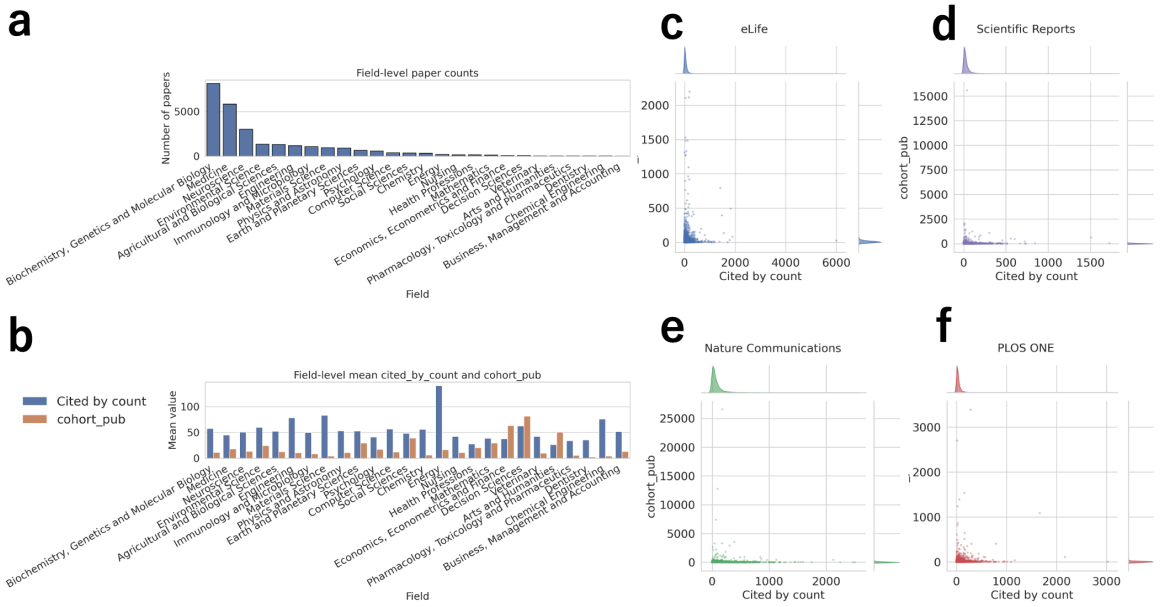

図 4　データセットの概要。(a) 分野別論文数。(b) 分野別平均被引用数・公的言及数。(c–f) 各ジャーナルの被引用数と公的言及数の分布。