自然言語における冪則と統語構造の関係の再考

中石海　吉田遼　梶川康平　福島孝治　大関洋平
東京大学
{nakaishi-kai787,yoshiryo0617,kohei-kajikawa,k-hukushima,oseki}
@g.ecc.u-tokyo.ac.jp

概要

自然言語の系列において，要素間の相関は距離について冪的に減衰することが知られている．これは，系列中のある要素を変更すると，その影響がどれだけ離れた要素にも及び得ることを意味する．先行研究 [1] は，このような特徴的現象を，系列の背後にある階層的な統語構造と結びつけて説明している．しかし，この説明は統語構造に関するいくつかの仮定に依拠しており，それらの仮定が実際に自然言語で成り立つかどうかは検証されていない．そこで，本研究は，ツリーバンクを用いて統語構造の統計的性質を調べ，先行研究の仮定がいずれも成り立たないことを明らかにする．

1　導入

自然言語の系列において，要素間の相関は距離について冪的に減衰することが知られている．例えば，系列上の 2 つの文字の間の相関は，両者の間にある文字の数について冪的に減衰する．このような振る舞いは，文字，音，単語といった要素のうちいずれに注目するか，コーパスや言語としてなにを選択するか，相関の定量としてどのような指標を用いるかなどによらず普遍的に見られることが，先行研究により確かめられている [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]．

多くの自然現象で相関は漸近的には指数関数的に減衰する．そのような場合，相関がどれくらいの距離まで働くかを表す典型的スケールが定義できる．それに対し，冪的減衰は指数関数的減衰よりも質的に遅い．また，冪関数は指数関数とは異なり，スケール変換に対して不変である．これらのことから，相関の冪的減衰は，相関の典型的スケールが発散していることを意味する [11]．直観的には，系列中のある要素を変更すると，その影響がどれだけ離れた要素にも及び得るということである．

Lin と Tegmark [1] は，このような特徴的現象が自然言語において普遍的に見られることを，文の背後にある階層的な統語構造と結びつけて理解することを提案している．統語構造を構文木などの木構造で表すことを考える（図 1）．統語構造の深さに左右で大きな偏りがないと仮定する．さらに，統語構造の各ノードが持つ子の数の典型的な値を c とし，注目する 2 つの葉ノードから共通の先祖ノードまでの距離を d とする．このとき，統語構造上での 2 つの葉ノード間の距離 r_{str} はおよそ $2d$ である．一方，文上での距離は $r_{\mathrm{seq}} \sim c^d$ である．よって，相関 C が統語構造上での距離について $C \sim \exp(-r_{\mathrm{str}}/\xi)$ のように指数関数的に減衰するならば，相関の文上での距離への依存性は $C \sim \exp(-2d/\xi) \sim c^{-2d/(\xi \ln c)} \sim r_{\mathrm{seq}}^{-2/(\xi \ln c)}$ のように冪的になる．

また，同研究は，統語構造を単純化した数理モデルとして，文脈自由文法（context-free grmmar, CFG）を確率的に拡張した確率文脈自由文法（probabilistic context-free grmmar, PCFG）を取り上げ，このモデルで，仮説と同様，相関が木構造上で指数関数的に，葉ノード上で冪的に減衰することを証明した．

この仮説を踏まえると，ある系列において相関が冪的に減衰することは，その系列の背後に階層構造が存在することの証拠と解釈できる．実際，そのような解釈のもと，多くの先行研究が，言語獲得の途

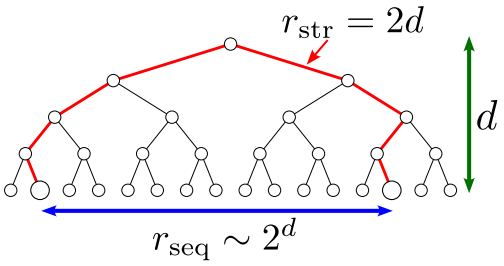

図 1　木構造における 2 つの葉ノードの間の系列上での距離 r_{seq} と構造上での距離 r_{str}．図は各ノードが 2 つの子ノードを持ち ($c = 2$)，葉ノードの深さが全て揃っている場合を示している．

上にある幼児の発話，そして動物の信号や行動などについて相関を測定し，その減衰が冪的かどうかに基づいてそれらが階層構造を持つかどうか検証している [6, 12, 13, 14, 15].

しかし，上述の仮説が妥当であるためには，(i) 相関が統語構造上での距離について指数関数的に減衰すること，そして (ii) 文上での距離が統語構造上での距離について指数関数的に増加することという 2 つの仮定が，自然言語において成り立っていなければならない．これらはいずれも検証されるべきである．特に，自然言語の統語構造は左か右のどちらかに強く偏る傾向があり [16, 17]，このような場合には 2 つめの仮定は必ずしも成り立たないと考えられる．また，自然言語には CFG では記述できない統語現象があること [18, 19]，PCFG のみに基づく素朴な構文解析は十分な性能を達成できないこと [20, 21, 22] が知られており，(iii) 自然言語の統語構造のモデルとして PCFG が妥当かどうかにも疑問が残る．

本研究では，ツリーバンクを用いて統語構造の統計的性質を調べることにより，これら 3 つの問題点を検証する．その結果，いずれの点でも先行研究の仮説は妥当ではないことを明らかにする．具体的な結果は以下である：

(i)相関は統語構造において，指数関数的ではなく冪的に減衰する．
(ii)文における距離は統語構造における距離について，指数関数的ではなく線形に増大する．
(iii)統語構造は PCFG から乖離している．

結果 (i) と (ii) は，先行研究の仮説が必要とする仮定が自然言語において成り立たないことを意味する．また，結果 (iii) から，PCFG において先行研究の仮説が成り立つとしても，自然言語において同じことが起こるとは期待し難い．以上より，自然言語における相関の冪的減衰には新たな説明が必要である．また，この仮説を踏まえておこなわれてきた様々な先行研究の結果の解釈は再考されるべきである．

2 前処理

ツリーバンクとしては BLLIP'99 コーパス [23] を用いる．これは，1987–89 年の Wall Street Journal から抽出された文に対して構文木がアノテーションされたデータであり，単語数は約 3×10^7 単語である．

前処理として，まず，コーパスに含まれる各構文木から，省略や移動の痕跡といった音形のない要素を除去する．

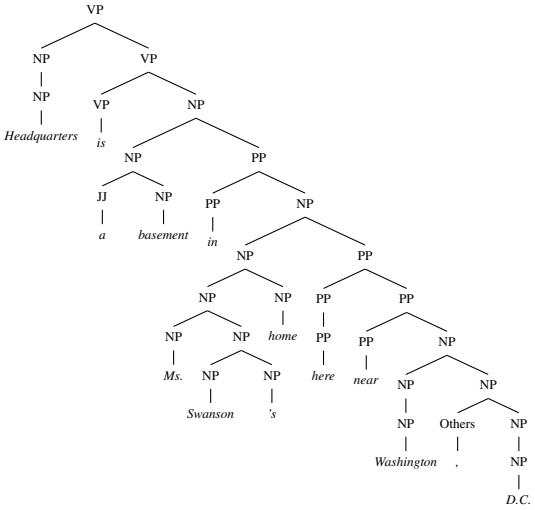

図 2 BLLIP'99 コーパス中の *Headquarters is a basement in Ms. Swanson's home here near Washington, D.C.* の構文木に，二分木化，および品詞と句のタグへの置き換えを施して得られる木構造．

さらに，子の数がノードに依存して異なる場合，異なる構文木間でのノードの対応づけが困難になるため，構文木を二分木化する．具体的には，NLTK ライブラリ [24] の `Tree.chomsky_normal_form` メソッドを用いて，Chomsky 標準形の意味で等価になるように二分木化をおこなう．このとき，品詞でもその親ノードでもないノードが 1 つしか子を持たないような構造が生じ得るが，これは言語学的解釈のつかないアーティファクトであるため，`Tree.collapse_unary` メソッドを用いて除去する．

最後に，各ノードがとり得る状態数が多いと統計量の推定に多くのデータが必要になるため，品詞や句をより少ない種類のタグで置き換える．このとき，統語論的に似た役割を持つ品詞や句は同じタグで表されるようにする．具体的には，表 1 のようにして，単語に対応する葉ノード以外のノードがとる状態数を 7 通りに削減する[1)]．これらの前処理後の構文木の例を図 2 に示す．

3 文における相互情報量

統語構造の解析の前に，まずは文における相関を調べ，減衰が冪的かどうか確認する．先行研究では単語間の相関や文字間の相関などが測られている

1) タグを決める際には，BLLIP'99 コーパスよりもタグの種類数が少ない Universal Dependencies コーパス [25] を参考にしつつ，範疇文法 [26, 27, 28] において振る舞いが似ているものを同じタグにまとめるという方針をとった．

タグ	品詞，句
NP	NP NN NNP NNS CD PRP POS QP NNPS EX NX FW
VP	VP VBD VB VBN VBG VBZ VBP TO AUX MD AUXG S S1 SBAR SINV FRAG SQ SBARQ RRC
PP	IN PP RP PRT RB ADVP RBR RBS
JJ	DT PRP$ PDT JJ ADJP JJR JJS NAC
Others	, . ” “ $: PRN -RRB- -LRB- # X INTJ UH SYM LS LST HYPH NFP
CC	CC UCP CONJP
WH	WHNP WDT WP WHADVP WRB WHPP WP$ WHADJP NML

表 1　本研究で用いる 7 通りのタグと，各タグが代表する品詞および句．

が，本研究では，統語構造についての議論との関係をわかりやすくするため，品詞間の相関に注目する．また，相関の大きさの定量としては，先行研究 [1] にならい相互情報量を用いる．

2 つの単語の間の距離を一方から他方までの単語数で定義し，以下，系列距離と呼ぶ．また，系列距離 r_{seq} だけ離れた単語のペアをランダムに選んだときの，それぞれがどのような品詞であるかという事象の間の相互情報量を I_{POS} とする．実際にコーパスから品詞のペアを N_{data} 個ランダムに抽出して経験分布を計算し，それに基づいて推定された相互情報量 I_{POS} の値を図 3 に示す[2)]．一般に，経験分布に基づく相互情報量の推定値は真の値よりも大きい方へ偏るが，図では異なる N_{data} のもとでの結果がほとんど重なっており，偏りが十分小さいとわかる．そして，相互情報量の系列距離依存性は冪的である．これは系列上で相関が冪的に減衰するという先行研究の結果と整合する[3)]．

4　統語構造における相互情報量

続いて，統語構造におけるタグ間の相互情報量を考える．2 つのノードの間の距離を両者をつなぐ経路を構成するエッジの本数で定め，構造距離と呼ぶ．構造距離が r_{str} であるようなノードのペアをランダムに抽出したときに，それぞれがどのようなタグを割り当てられるかという事象の間の相互情報量を I_{tag} とする．前節と同様にして得られた推定値を

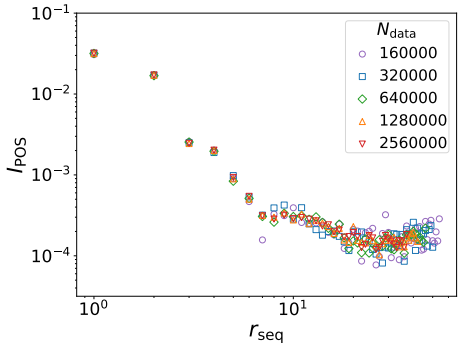

図 3　品詞間の相互情報量 I_{POS} の系列距離 r_{seq} 依存性．マーカーの違いはデータサイズ N_{data} の違いを表す．

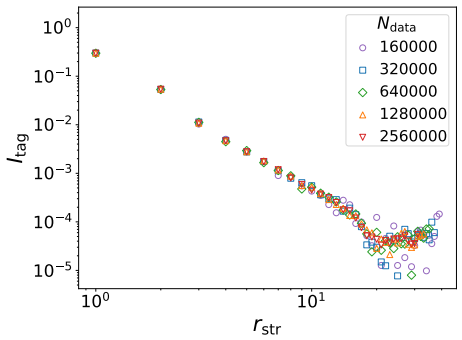

図 4　タグ間の相互情報量 I_{tag} の構造距離 r_{str} 依存性．マーカーの違いはデータサイズ N_{data} の違いを表す．

図 4 に示す．ここでも，データサイズ由来のバイアスは十分小さいことが確認できる．そして，相関は統語構造上で指数関数的に減衰するという先行研究の仮定に反し，相互情報量は構造距離について冪的に減衰している[4)]．

5　構造距離と系列距離の関係

ここでは構造距離と系列距離の関係を調べる．具体的には，コーパスから品詞に対応するノードのペアを抽出し，両者の間の構造距離 r_{str} と系列距離

2)　ある品詞のペアと，その 1 つめの品詞と 2 つめの品詞を入れ替えて得られるペアは，別のものとして数えている．また，経験分布から相互情報量を計算するときには，文献 [29] の手法を用いている．次節以降でも，相互情報量は同様のやり方で計算されている．

3)　系列距離 30 前後では，相互情報量はほとんど減衰せず一定の値にとどまっているようにも見える．しかし，この領域は推定値の揺らぎが大きく，明瞭なことは言えない．また，この領域では，かなり長い文の端同士に位置するペアのような特殊な場合が主に考慮されるため，それに由来する意図しない効果が現れている可能性もある．

4)　距離 20 前後より遠い領域では，相互情報量の値がほぼ一定になっているようにも見える．しかし，品詞間の相互情報量の系列距離 30 前後での振る舞いと同様の理由から，この領域での振る舞いについて明確な結論を下すことは難しい．

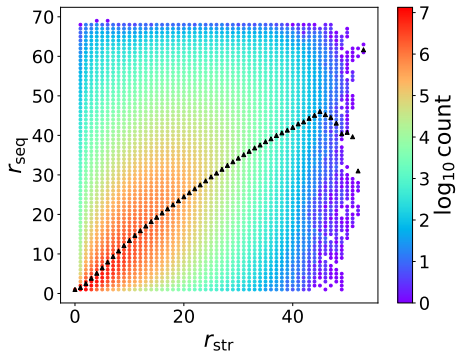

図 5　構造距離 r_{str}，系列距離 r_{seq} であるようなノードのペアの数．黒三角のマーカーは各構造距離ごとの系列距離の平均を示す．

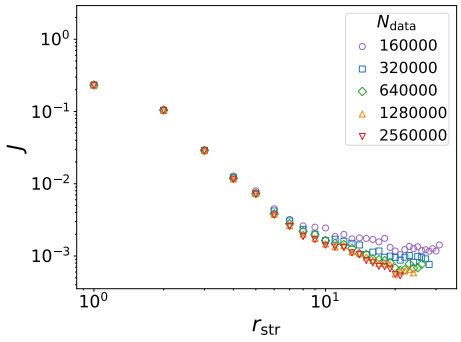

図 6　2 つのノードのタグがともに NP であるときの，PCFG からの乖離の大きさ J の構造距離 r_{str} 依存性．マーカーの違いはデータサイズ N_{data} の違いを表す．

r_{seq} を測定する．各 $(r_{\mathrm{str}}, r_{\mathrm{seq}})$ に対応するペアの数をヒートマップとして表したのが図 5 である．図には各構造距離ごとの系列距離の平均もプロットされている．図からわかるように，系列距離の構造距離への依存性はおおよそ線形である．これは，系列距離が構造距離について指数関数的に増大するという先行研究の仮定が成り立たないことを示している[5)]．

6　確率文脈自由文法からの乖離

最後に，統語構造のモデルとしての PCFG の妥当性を検証する．PCFG が生成する木構造においては，あるノードの状態が決まるとその子の分布も決まる．そのため，木構造中の 2 つのノードの状態を固定すると，それぞれの子は互いに独立である．よって，ある木構造の分布が PCFG からどれだけ乖離しているかは，状態を固定された 2 つのノードの子の間の相互情報量によって定量できる [30]．

構造距離が r_{str} であり，ともにタグ NP を割り当てられているようなノードのペアをランダムに抽出することで推定された，両者の子の間の相互情報量 J の推定値を，図 6 に示す．データサイズ依存性が小さいことから，バイアスが十分小さいことがわかる．そして，PCFG からの乖離の大きさは正の値をとっており，自然言語の統語構造が PCFG から逸脱していることがわかる．さらに，J の減衰は構造距離について冪的であり，スケールをどれだけ大きくしてもこの乖離が無視できないことを示している．

7　議論

自然言語における相関の冪的減衰が統語構造に由来するという先行研究の仮説の妥当性を検証するべく，本研究では，仮説が依拠するいくつかの仮定が現実の統語構造において成り立っているかどうかを定量的に調べた．その結果，いずれの仮定も成り立っていないことが示された．

従って，自然言語における相関の冪的減衰には，この仮説に代わる新たな説明が必要である．1 つの可能性は，自然言語を統計物理学における臨界現象としてとらえることである．実際，大規模言語モデルの温度パラメータを変化させると，ある温度で臨界現象が起こり，この領域で生成される系列は，相関の冪的減衰を含む自然言語の統計的振る舞いを再現することが報告されている [31]．

また，導入で述べたように，様々な先行研究がこの仮説を踏まえ，相関の冪的減衰を系列の背後に階層構造が存在することの証拠だと解釈したうえで，幼児の発話や動物の信号を解析している．しかし，本研究が示したように仮説が妥当ではない以上，これらの先行研究の結果をどのように解釈すべきかは再考されるべきである．

今後の展望としては，まず，今回用いた BLLIP'99 コーパスとはドメインや言語の異なるコーパスにおいて，同様の振る舞いが見られるかどうか調べることが挙げられる．また，本研究で明らかになった統語構造の統計的性質がどのような形式文法によって再現可能かも，重要な問題である．実は，PCFG だけでなく，確率文脈依存文法の統計的性質も，典型的には統語構造に見られるものと異なることが観察されている [30]．そのため，木接合文法 [32]，組合せ範疇文法 [28]，ミニマリスト文法 [33] など，自然言語の統語現象をより正確に記述可能な形式文法に基づく確率モデルが必要だと考えられる．

5)　構造距離がおよそ 45 以上の領域では，データの少なさのために，平均的な傾向は見えづらくなっている．

謝辞

本研究は JSPS 科研費 23KJ0622，24H00087，および JST さきがけ JPMJPR21C2 の支援を受けたものです.

参考文献

[1] H. W. Lin and M. Tegmark. Critical behavior in physics and probabilistic formal languages. **Entropy**, Vol. 19, No. 7, p. 299, June 2017.

[2] W. Ebeling and A. Neiman. Long-range correlations between letters and sentences in texts. **Physica A**, Vol. 215, No. 3, pp. 233–241, 1995.

[3] W. Ebeling and T. Pöschel. Entropy and Long-Range correlations in literary english. **Europhysics Letters**, Vol. 26, No. 4, p. 241, 1994.

[4] W. Li. Mutual information functions of natural language texts. Technical Report 89-10-008, Santa Fe Institute, 1989.

[5] N. Mikhaylovskiy and I. Churilov. Autocorrelations decay in texts and applicability limits of language models. **arXiv preprint arXiv:2305.06615**, 2023.

[6] T. Sainburg, B. Theilman, M. Thielk, and T. Q. Gentner. Parallels in the sequential organization of birdsong and human speech. **Nature Communications**, Vol. 10, No. 1, p. 3636, 2019.

[7] H. Shen. Mutual information scaling and expressive power of sequence models. **arXiv preprint arXiv:1905.04271**, 2019.

[8] S. Takahashi and K. Tanaka-Ishii. Do neural nets learn statistical laws behind natural language? **PloS one**, Vol. 12, No. 12, p. e0189326, 2017.

[9] S. Takahashi and K. Tanaka-Ishii. Evaluating computational language models with scaling properties of natural language. **Computational Linguistics**, Vol. 45, No. 3, pp. 481–513, 2019.

[10] K. Tanaka-Ishii and A. Bunde. Long-range memory in literary texts: On the universal clustering of the rare words. **PLoS One**, Vol. 11, No. 11, p. e0164658, 2016.

[11] 西森秀稔. 相転移・臨界現象の統計物理学 新物理学シリーズ（新物理学シリーズ 35）. 培風館, 2005.

[12] E. Howard-Spink, M. Hayashi, T. Matsuzawa, D. Schofield, T. Gruber, and D. Biro. Nonadjacent dependencies and sequential structure of chimpanzee action during a natural tool-use task. **PeerJ**, Vol. 12, p. e18484, 2024.

[13] T. Sainburg and T. Q. Gentner. Toward a computational neuroethology of vocal communication: from bioacoustics to neurophysiology, emerging tools and future directions. **Frontiers in Behavioral Neuroscience**, Vol. 15, p. 811737, 2021.

[14] T. Sainburg, A. Mai, and T. Q. Gentner. Long-range sequential dependencies precede complex syntactic production in language acquisition. **Proceedings of the Royal Society B**, Vol. 289, No. 1970, p. 20212657, 2022.

[15] M. Youngblood. Language-like efficiency and structure in house finch song. **Proceedings of the Royal Society B**, Vol. 291, No. 2020, p. 20240250, 2024.

[16] J. A. Hawkins. A parsing theory of word order universals. **Linguistic Inquiry**, Vol. 21, No. 2, pp. 223–261, 1990.

[17] M. Dryer. The greenbergian word order correlations. **Language**, Vol. 68, pp. 81–138, 1992.

[18] S. M. Shieber. Evidence against the context-freeness of natural language. **Linguistics and Philosophy**, Vol. 8, p. 333, 1985.

[19] H. Senuma and A. Aizawa. Computational complexity of natural morphology revisited. **Transactions of the Association for Computational Linguistics**, Vol. 12, , 2024.

[20] E. Charniak. Statistical techniques for natural language parsing. **AI Magazine**, Vol. 18, No. 4, pp. 33–33, 1997.

[21] M. Johnson, T. Griffiths, and S. Goldwater. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. In **Advances in Neural Information Processing Systems**, Vol. 19, 2006.

[22] T. J. O'Donnell, J. B. Tenenbaum, and N. D. Goodman. Fragment grammars: Exploring computation and reuse in language. Technical Report MIT-CSAIL-TR-2009-013, Massachusetts Institute of Technology, 2009.

[23] E. Charniak, D. Blaheta, N. Ge, K. Hall, J. Hale, and M. Johnson. BLLIP 1987-89 WSJ corpus release 1, LDC2000T43. Linguistic Data Consortium, 36, 2000.

[24] S. Bird, E. Loper, and E. Klein. **Natural Language Processing with Python**. O'Reilly Media Inc., 2009.

[25] J. Nivre, M. C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman. Universal Dependencies v2: An evergrowing multilingual treebank collection. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 4034–4043, 2020.

[26] K. Ajdukiewicz. Syntactic connexion. In **Polish logic: 1920–1939**, pp. 207–231. Oxford University Press (Original work published in 1935), 1967.

[27] Y. Bar-Hillel. A quasi-arithmetical notation for syntactic description. **Language**, Vol. 29, No. 1, pp. 47–58, 1953.

[28] M. Steedman. **The syntactic process**. MIT press, 2000.

[29] P. Grassberger. Entropy estimates from insufficient samplings. **arXiv preprint physics/0307138**, 2003.

[30] K. Nakaishi and K. Hukushima. Statistical properties of probabilistic context-sensitive grammars. **Physical Review Research**, Vol. 6, p. 033216, Aug 2024.

[31] K. Nakaishi, Y. Nishikawa, and K. Hukushima. Critical phase transition in large language models. **arXiv preprint arXiv:2406.05335**, 2024.

[32] A. K. Joshi. **Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions?**, pp. 206–250. Studies in Natural Language Processing. Cambridge University Press, 1985.

[33] E. Stabler. Derivational minimalism. In **Logical Aspects of Computational Linguistics**, pp. 68–95, 1997.