合成データによる事前学習から見る言語と音楽の関係性

稲葉 達郎 [1]　乾 健太郎 [1,2,3]　栗林 樹生 [1]
[1]MBZUAI　[2] 東北大学　[3] 理化学研究所
{tatsuro.inaba,tatsuki.kuribayashi,kentaro.inui}@mbzuai.ac.ae

概要

本研究は，言語と音楽に共通する構造的性質を明らかにすることを目的とし，Transformer を擬似的な学習者として用いた計算論的検討を行う．異なる規則や構造をもつ合成データによる事前学習を施したモデルを，言語または音楽データでファインチューニングし，各構造的バイアスが学習性能に与える影響を定量的に評価した．その結果，言語と音楽の双方に共通して有効または無効となる構造と，ドメイン固有に有効となる構造が存在することが示された．特に，可変長の反復構造は両ドメインに有効である一方，拍節構造は音楽に対してのみ有効であることが明らかとなった．

1　はじめに

“Language and music define us as human.”

— *Aniruddh D. Patel (2007)*

言語と音楽が構造的・表現的にどのような共通点を持ち，それが人間のコミュニケーションや感情表現にいかなる影響を与えているのかという問いは，古くから哲学 [1]，生物学 [2, 3]，文学 [4]，認知科学 [5, 6] など多岐にわたる分野において議論されてきた．歴史的・進化的観点からも，両者は共通の起源をもつ可能性が指摘されており，プラトンからルソー，ダーウィンに至るまで，多くの思想家が言語と音楽の連関を論じてきた．言語と音楽の関連性を理解することは我々人間を理解する上で重要な要因の一つである．

近年，言語と音楽を人間に固有のコミュニケーション体系として比較する研究が進展している．両者は音声を主要な媒体とし，文化普遍的に存在する点に加え，階層的構造や時間的展開といった形式的特徴を共有することが指摘されてきた [5, 6, 7, 8]．一方で，意味の指示性や統語規則の厳密さにおいては重要な差異も存在し，言語と音楽は完全に同一の体系ではなく，部分的に共有されつつ分岐した認知システムとして理解されている [6, 9, 10, 11]．

しかし，これらの知見の多くは行動実験や神経科学的手法に基づいており，構造の学習や一般化の過程そのものを直接検証することには限界がある [6, 11]．そこで本研究では，この点を補うために計算論的モデルを用い，言語と音楽に共通する構造的バイアスを操作可能な形で検討する．具体的には Transformer を擬似的な学習者として用い，異なる構造をもつ合成データによる事前学習を施した複数のモデルを構築する．これらのモデルをテキストデータまたは楽譜データでファインチューニングすることで，どのような構造的バイアスが言語および音楽の習得に有効に寄与するのかを定量的に評価することを目的とする．

実験の結果，構造的事前学習の効果は一様ではなく，導入される構造の性質および対象ドメインに依存することが示された．可変長の反復構造に基づく事前学習は言語と音楽の双方において一貫した性能向上を示した一方で，拍節構造や階層構造に基づく事前学習はドメイン依存的な結果を示した．特に，周期的な拍節構造は音楽に対してのみ有効であり，言語と音楽に有効な構造的バイアスが部分的に共有されつつも，その適合性が構造の種類によって異なることが示唆された．本研究は，言語と音楽における構造の学習可能性を直接比較可能な形で検討し，従来の行動実験や神経科学的研究では捉えにくかった学習過程に新たな視点を提供する．

2　関連研究

本研究は，言語と音楽の関連性を探る学際的研究の系譜の中に位置づけられる．言語と音楽の関係については，構造的類似性に加え，両者の処理に関与する認知・神経基盤が部分的に共有されていることが多くの研究によって示されている [5, 6, 9, 10, 12, 13]．このような処理基盤の共有は，音楽経験が言語能力に波及的な効果をもたらす可能性とも整合的であり，実際に，歌唱やリズム活

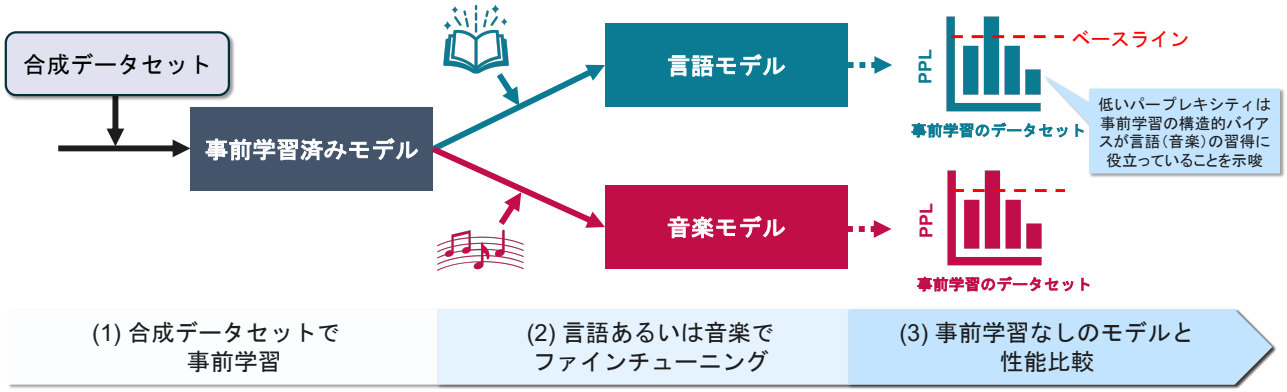


図 1　実験の全体像．はじめに合成データセットを用いてモデルを事前学習し，その後，言語または音楽データによるファインチューニングを行う．最後にテストデータを用いてモデルの性能を評価し，事前学習によって獲得された機能的バイアスが各モダリティの学習にどの程度寄与するかを測定する．

動が発音，語彙学習，読解流暢性などに寄与することが報告されている [14, 15, 16, 17, 18]．

第二に，本研究は，深層学習モデルを擬似的な学習者として用い，人間の言語獲得や処理を検証する研究の流れにも位置づけられる．計算モデルによる言語知識の獲得を，仮定された学習環境の下での学習可能性を示す証明概念として用いる立場が提案されており [19, 20, 21]，この枠組みでは，人間では実施困難な因果的操作を通じた検証が可能となる．

これらの視点は，言語や音楽とは何かという根源的な問いとも密接に関係している．どのような構造的性質が言語や音楽を成立させているのか，また，それらの境界はどこにあるのかという問題に対し，計算論的学習の観点から新たな示唆を与えることを本研究は目指す．

3　手法

本章では，本研究の手法の全体像を概観するとともに，実験で用いる合成データの設計について詳述する．

3.1　概要

本研究では，異なる種類の合成データによる事前学習が言語と音楽の学習に与える影響を測る．図 1 に手法の全体像を示す．初めに合成データセットの中から一つを選択し，デコーダ型の Transformer を事前学習する．その後，テキストデータあるいは楽譜データでファインチューニングを行い，テストデータにおけるパープレキシティを評価する．その結果から，言語，あるいは音楽の学習に役立つ機能バイアスとなる合成データを推定し，言語と音楽でこの傾向がどう異なるかを分析する．

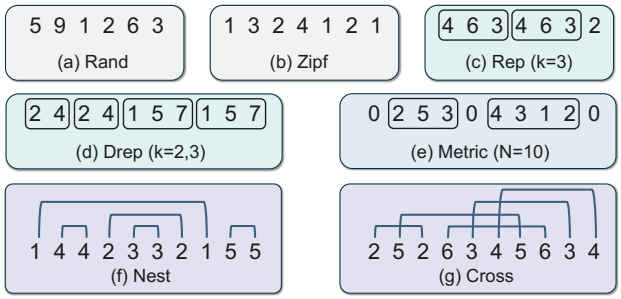


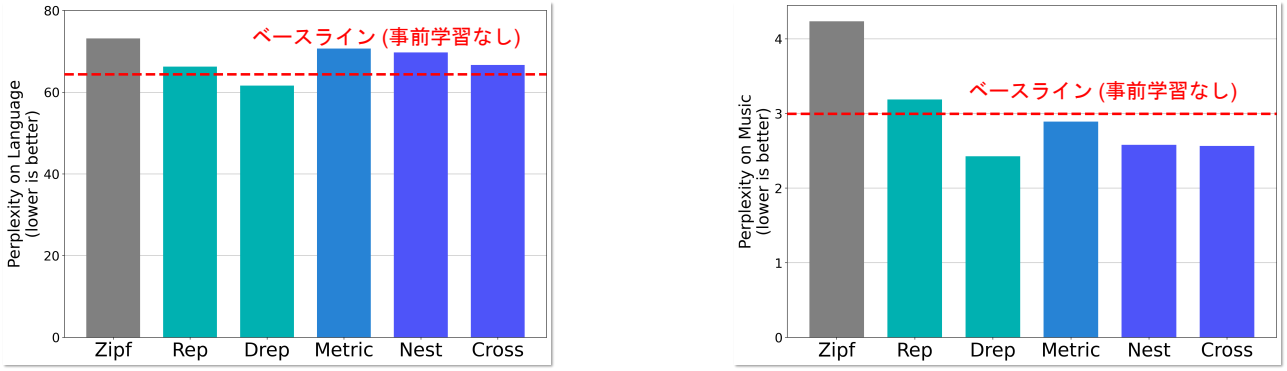
図 2　事前学習に使用した合成データセット．各データセットは異なる種類の規則または構造をもつ．

3.2　合成データ

本研究では，六種類の合成データセットを構築し，モデルの事前学習に用いる．各合成データセットは，特定の規則性あるいは構造的特徴を意図的に付与することで設計されている．図 2 に，各データセットのトークン列の例を示す．

ジップの法則（Zipf）　トークンの出現頻度がジップの法則（Zipf’s law）に従うように生成された合成データセットである．図 2 (b) に例を示す．ジップの法則によれば，頻度順位が r のトークンの出現頻度は，最頻トークンの出現頻度に対しておおよそ $1/r$ に比例する．自然言語において語彙の出現頻度分布がジップの法則に従うことは広く知られているが，音楽の分野においても同様の性質が観察されている．具体的には，音符単体に限らず，複数の音符から構成される和音やコード進行といった高次の音楽単位においても，ジップ的な頻度分布が成り立つことが報告されている [22, 23]．

反復構造（Rep）　長さ k のトークン列（以下，ブロックと呼ぶ）が，直後に必ず同一の順序で繰り返される構造を持つ合成データセットである．図 2



(a) 言語のパープレキシティ　　(b) 音楽のパープレキシティ

図 3　各種構造的事前学習を施した Transformer モデルを，言語および音楽タスクでファインチューニングした際の性能比較．縦軸はパープレキシティを表す（低いほど性能が高い）．

(c) の例では，長さ $k = 3$ のブロック `4,6,3` が，連続して二回出現している．本研究の実験では，反復長を $k = 10$ に固定してデータを生成した．

可変長反復構造（Drep）　Drep は Rep と同様に反復構造を含むが，繰り返しの長さ k が各ブロックごとに一様ランダムにサンプリングされる点が異なる．図 2 (d) の例では，長さ $k = 2$ のブロック `2,4` の反復に続いて，長さ $k = 3$ のブロック `1,5,7` が反復されている．本研究の実験では，k を 1 から 100 の範囲で一様にサンプリングすることで，可変長の反復構造を導入した．

拍節構造（Metric）　周期的・拍節的な構造を捉えることを目的として設計された合成データセットである．トークン列の累積和が所定の閾値 N の倍数に達するたびに，境界を示すトークン `0` を挿入することで，周期的な区切りを明示的に導入する．図 2 (e) の例では，`2+5+3=10` において一区切りとなり，続く `4+3+1+2=10` において再び区切りが挿入されている．このような周期的な境界構造は，固定長の小節や拍に基づく音楽のリズム構造を単純化して模倣したものであり，音楽に特有の時間的・階層的組織を反映した帰納バイアスをモデルに与えることを意図している．本研究の実験では $N = 1000$ と設定した．

階層構造（Nest）　このデータセットは Dyck 言語としても知られる対応する入れ子括弧列から構成されている．図 2 (f) に例を示す．各開き記号は正しい順序で対応する閉じ記号によって閉じられなければならず，階層的で文脈自由な依存関係となる．先行研究 [24] に従い，以下の単純な確率的手続きを用いてトークン列を生成する．各位置において，未閉鎖の開き括弧が存在しない場合には必ず開き括弧を選択し，それ以外の場合には，開き括弧を選択する確率を $p = 0.49$，直前に開かれている括弧を閉じる確率を $p = 0.51$ としてランダムに選択する．

交差構造（Cross）　このデータセットは Nest が交差しないように制限されていたのと異なり，依存関係が交差することを許すように拡張したものである．Semi-dyck 言語とも呼ばれる．図 2 (g) に例を示す．非文脈自由な交差依存を除いて Nest とできる限り同一となるよう，その他の点では Nest と極力一致させている．具体的には，開き記号と閉じ記号との期待距離を Nest データセットの経験的分布からサンプリングし，各開きトークンをどの距離で閉じるか決定する際に使用している．

4　実験設定

本研究では，次元数 256，層数 6，自己注意機構のヘッド数 8，最大系列長 1,024 を持つ，デコーダ型 Transformer モデルを用いた．事前学習には合成データセットを使用し，語彙数を 500 に統一した上で，各データセットにつき 10 億（1B）トークンを生成し，それぞれ独立にモデルを学習した．

自然言語のファインチューニングには English WikiText コーパス [25] を用いた．トークナイザには GPT-2 の BPE トークナイザ（語彙数 50,257）を使用し，最終的な総トークン数は 117M となった．音楽のファインチューニングには，ポピュラー音楽の楽譜データセットである POP909 [26] を使用した．楽譜データは，イベント単位のトークナイズ手法である REMI+ [27] に基づいてトークン化し，総トークン数は 53M となった．また，楽譜データの前処理

および MIDI 操作には MusPy [28] を用いた．

事前学習とファインチューニングでは使用する語彙および語彙数が異なるため，ファインチューニングの開始前に，トークン埋め込み層および出力側のデコード層（LM head）を初期化する．先行研究では，これらの層を完全にランダム初期化するよりも，事前学習済みの埋め込みベクトルをランダムにサンプリングして初期値として用いる方が有効であることが示されており [29]，本研究でも同手法を採用した．

さらに，事前学習の効果を明確に評価するため，事前学習を行わずにファインチューニングのみを施したモデルをベースラインとして比較に用いた．なお，先行研究 [21] と同様にランダムなトークン列からなるデータセット（図 2 (a) に例）を用いた事前学習についても検討したが，パープレキシティが極端に大きく，モデル間の詳細な比較が困難であったため本研究では除外した．

5　結果・考察

図 3 に結果を示す．なお，Rand を用いて事前学習を行った場合，パープレキシティが極端に大きくなったため，可視性の観点から図中では省略した．

Zipf は有効な効果を示さなかった．　Zipf 分布に基づく事前学習は，言語および音楽のいずれにおいても，明確な性能向上を示さなかった．事前学習によって獲得できるのはあくまで頻度分布の形状に関する規則であり，言語や音楽において具体的にどの語彙や要素が頻出するかといった知識までは含まれない．その結果，モデルは Zipf 的な頻度分布という規則を，実際の言語・音楽の語彙，さらにはより高次の構造的要素へと一般化することができず，性能向上に結びつかなかったと考えられる．

複雑な反復構造は高い効果を示した．　固定長の反復構造のみを含む Rep は性能向上に繋がらなかった．単一の反復長に基づいて事前学習されたモデルは，過度に特化した帰納バイアスを内部化し，新たなパターンの学習を阻害している可能性がある．一方，より複雑な反復構造を持つ Drep による事前学習は，言語および音楽において一貫した性能向上を示した．この結果は，可変長の反復構造への曝露が，言語と音楽の双方に広く存在する，長さの異なる反復フレーズやモチーフを捉えるための柔軟な帰納バイアスを誘導することを示唆している．

拍節構造は音楽には有効だが言語には有効でなかった．　Metric による事前学習は，音楽において多少の性能向上を示した一方で，言語に対しては有効性を示さなかった．この結果は予想と整合的である（3.2 節参照）．すなわち，周期的な拍節構造は音楽のリズムにおいて本質的な要素であり（例：4/4 拍子における固定長の小節構造），音楽の構成原理と強く結びついている．一方，自然言語にはこのような厳密な時間的周期性は存在しない．したがって，Metric は音楽の構造と整合した帰納バイアスを導入する一方で，言語構造とは適合しないバイアスとなっており，この不整合が言語タスクにおける性能向上が見られなかった要因であると考えられる．

階層構造 / 交差構造は言語に有効でなかった．　予想に反して，Dyck 言語に類似した階層構造を持つ Nest および，文脈自由文法では表現できない交差依存を含む Cross のいずれも，言語タスクにおける性能向上をもたらさなかった．一方で，音楽タスクにおいては両構造が一定の性能向上を示しており，階層的あるいは長距離依存的な構造が，音楽の構成要素（フレーズやモチーフ）の学習には有効に機能した可能性がある．これらの構造はいずれも統語理論において重要な役割を果たすものであることから，言語タスクにおける結果は一見すると直観に反する．一つの可能な要因として，本研究で用いた Transformer モデルが比較的小規模（層数 6，次元数 256）であったため，階層構造や交差依存といった高度な統語的制約を，言語データ上で十分に活用できなかった可能性が考えられる．

6　おわりに

本研究では，合成データを用いた事前学習を通じて，どのような構造的規則が言語および音楽の学習に転移するのかを体系的に検証した．その結果，多様な反復構造は言語および音楽の双方において有効である一方で，拍節構造や階層構造・交差構造は，音楽学習時にのみ有効であるという差異が観察された．今後，異なるモデルサイズや，合成データにおける構造の難易度をより細かく制御した設定での検証を通じて，言語と音楽の関係性をさらに明らかにしていきたい．また，音楽ジャンルの違いが言語学習や言語能力に与える影響を調べるなど，より多様な設定での実験も行っていきたい．

謝辞

本研究は AMED JP25wm0625405 の助成を受けたものです.

参考文献

[1] Jean-Jacques Rousseau. **Essai sur l'origine des langues**. Chez Pissot, Paris, 1781.

[2] Charles Darwin. **The Descent of Man, and Selection in Relation to Sex**, Vol. 1. John Murray, London, 1871.

[3] Steven Mithen. **The Singing Neanderthals: The Origins of Music, Language, Mind, and Body**. Harvard University Press, Cambridge, MA, 2006.

[4] Roman Jakobson. Linguistics and poetics. In Thomas A. Sebeok, editor, **Style in Language**, pp. 350–377. Massachusetts Institute of Technology Press, Cambridge, MA, 1960.

[5] Aniruddh D. Patel. Language, music, syntax and the brain. **Nature Neuroscience**, Vol. 6, No. 7, pp. 674–681, 2003.

[6] Aniruddh D. Patel. **Music, Language, and the Brain**. Oxford University Press, 12 2007.

[7] Fred Lerdahl and Ray Jackendoff. **A Generative Theory of Tonal Music**. MIT Press, Cambridge, MA, 1983.

[8] Ray Jackendoff. Parallels and nonparallels between language and music. **Music Perception**, Vol. 26, No. 3, pp. 195–204, 2009.

[9] Mireille Besson and Daniele Schön. Comparison between language and music. **Annals of the New York Academy of Sciences**, Vol. 930, pp. 232–258, June 2001.

[10] L. Robert Slevc. Language and music: Sound, structure, and meaning. **Wiley Interdisciplinary Reviews: Cognitive Science**, Vol. 3, No. 4, pp. 483–492, July 2012.

[11] David Temperley. Music and language. **Annual Review of Linguistics**, Vol. 8, No. 1, pp. 153–170, January 2022. Available at SSRN.

[12] Evelina Fedorenko, Aniruddh Patel, Daniel Casasanto, Jonathan Winawer, and Edward Gibson. Structural integration in language and music: Evidence for a shared system. **Memory & Cognition**, Vol. 37, No. 1, pp. 1–9, 2009.

[13] Richard Kunert, Roel M. Willems, Daniel Casasanto, Aniruddh D. Patel, and Peter Hagoort. Music and language syntax interact in broca's area: An fmri study. **PLOS ONE**, Vol. 10, No. 11, pp. 1–16, 11 2015.

[14] Charlotte P. Mizener. Enhancing language skills through music. **General Music Today**, Vol. 21, No. 2, pp. 11–17, 2008.

[15] Robert Legg. Using music to accelerate language learning: An experimental study. **Research in Education**, Vol. 82, No. 1, pp. 1–12, 2009.

[16] Dacian Dorin Dolean. The effects of teaching songs during foreign language classes on students' foreign language anxiety. **Language Teaching Research**, Vol. 20, No. 5, pp. 638–653, 2016.

[17] Brittany A. McCormack and Christopher Klopper. The potential of music in promoting oracy in students with english as an additional language. **International Journal of Music Education**, Vol. 34, No. 4, pp. 416–432, 2016.

[18] Angela Salmon. Using music to promote children's thinking and enhance their literacy development. **Early Child Development and Care**, Vol. 180, No. 7, pp. 937–945, 2010.

[19] Tal Linzen. What can linguistics and deep learning contribute to each other? Response to Pater. **Language**, Vol. 95, No. 1, pp. e99–e108, 2019.

[20] Alex Warstadt and Samuel R. Bowman. What artificial neural networks can tell us about human language acquisition. In **Algebraic Structures in Natural Language**, pp. 17–60. CRC Press, 2022.

[21] Isabel Papadimitriou and Dan Jurafsky. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6829–6839, Online, November 2020. Association for Computational Linguistics.

[22] Bill Manaris, Juan Romero, Penousal Machado, Dwight Krehbiel, Timothy Hirzel, Walter Pharr, and Robert B. Davis. Zipf's law, music classification, and aesthetics. **Computer Music Journal**, Vol. 29, No. 1, pp. 55–69, 03 2005.

[23] Juan I. Perotti and Orlando V. Billoni. On the emergence of zipf's law in music. **Physica A: Statistical Mechanics and its Applications**, Vol. 549, p. 124309, 2020.

[24] Isabel Papadimitriou and Dan Jurafsky. Injecting structural hints: Using language models to study inductive biases in language learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 8402–8413, Singapore, December 2023. Association for Computational Linguistics.

[25] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In **International Conference on Learning Representations**, 2017.

[26] Ziyu Wang, Ke Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, Guxian Bin, and Gus Xia. Pop909: A pop-song dataset for music arrangement generation. In **Proceedings of 21st International Conference on Music Information Retrieval**, 2020.

[27] Dimitri von Rütte, Luca Biggio, Yannic Kilcher, and Thomas Hofmann. FIGARO: Controllable music generation using learned and expert features. In **The Eleventh International Conference on Learning Representations**, 2023.

[28] Hao-Wen Dong, Ke Chen, Julian McAuley, and Taylor Berg-Kirkpatrick. Muspy: A toolkit for symbolic music generation. In **International Society for Music Information Retrieval Conference**, 2020.

[29] Zhengxuan Wu, Alex Tamkin, and Isabel Papadimitriou. Oolong: Investigating what makes transfer learning hard with controlled studies. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 3280–3289. Association for Computational Linguistics, 2023.