

構成性の度合いはカオスの縁で最も高くなる —ESN モデルと TopSim 指標を用いたケーススタディ—

上田亮
東京大学

ryoryoueda@is.s.u-tokyo.ac.jp

概要

我々には想像もつかないような帰納バイアスにおいて、想像もつかないような記号体系が最適であったかもしれないのに、構成性が言語の必然的な性質であると信じて良いのだろうか？ この問いに答えるためには、言語の進化（文化進化）のみならず、人類の進化（生物進化）にも思いを馳せながら、構成性と帰納バイアスの関係を考察する必要がある。事例研究として、Echo State Network モデルがカオスの縁と呼ばれる超パラメタ領域において構成的な（高い TopSim スコアをもつ）記号体系に適合・汎化するバイアスをもつことを実験的に示し、構成的な言語が創発する必然性への傍証を試みる。

1 はじめに

大規模言語モデル (LLM) が躍進している裏で、**言語創発** (Language Emergence) 又は**創発コミュニケーション** (Emergent Communication; EmCom, EC) と呼ばれる、計算言語学や進化言語学の派生分野が一部で注目を集めている [1] (以降では **EC** と呼ぶことにする)。EC は、エージェント同士でコミュニケーションをさせたときに生じる言語のようなプロトコル (**創発言語**) を研究対象とし、これが人間の言語らしくなる条件を探求することによって間接的に言語のメカニズムの解明を志す—いわゆる構成論的アプローチをとる分野である。とりわけ、近年の国内コミュニティにおいては、記号創発ロボティクス (Symbol Emergence in Robotics) との融合を通して**生成的 EC** とでも言うべき新たな潮流が生まれつつある [2, 3, 4]。生成的 EC は、これまで主にマルチエージェント強化学習やゲーム理論の範疇として扱われてきた EC の諸問題を、おおそ等価なベイズ的学習・推論として再解釈することによって『言語とコミュニケーションの創発』の問題をより一般的な枠

組みから捉え直そうという構想・立場である。あるいは、EC とはエージェントが協調して言語という表現を作る表現学習であって、そのうち特にベイズ的なアプローチをとる立場とも言える¹⁾。

生成的 EC の構想の下、ベイジアンネットワークのようなフレームワークで EC の問題を捉え直すことには一定の意義がある一方で、計算言語学において伝統的に研究対象とされてきた**構成性 (Compositionality)** のような性質を説明するにはやや不十分に思われる。構成性は明らかに帰納バイアスに関わる問題であり、その性質は具体的なモデルの定義に依存するため、グラフィカルモデル上の頂点や矢印の類では（少なくとも簡潔には）表せないだろう。ベイズ的なアプローチをとるといふ基本的な方針からはそれほど乖離することなく、この問題に対する良い道具立てを用意することはできないだろうか？

ここで、従来の意味での EC や、それに関連する進化的言語学の先行研究に視点を戻してみると、言語において構成性が創発するのは「構成性の低い言語（らしきもの）に比べて学習しやすい（適合しやすく汎化もしやすい）から」だと考えられているようである。有名な例として Iterated Learning Model (ILM) [6, 7] という、親世代から子世代への言語継承を考慮した言語創発モデルがある。ILM の枠組みでは、言語を継承する過程がボトルネックとなり、言語はそのボトルネックを潜り抜けることができるように“進化”していく (**文化進化**)。文化進化の帰結として、言語の構成性が向上していくことが人間や LSTM エージェントを用いた実験で確認されている [8, 9] (**実験記号論**)。しかしながら、ここでもやはり構成性は帰納バイアスに依存した性質であることに注意する必要がある。即ち、高い構成性をもつ

1) 谷口 [5] はこの立場に立脚し、言語とは感覚運動機能を有する人間の集団が作る世界モデルであり、LLM はこの世界モデルの模倣しているという仮説を唱えている (CPC 仮説)。

言語がそうでない言語に比べて（我々の直感においても実験記号論の結果においても）学習しやすいのは、文化進化よりも遥かに緩やかで長い時間スケールにおいて起こってきた人類の進化（**生物進化**）の帰結としてもたらされた特定の帰納バイアスにおいて偶然そうなのであって、必然的にもたらされたものではないかもしれない。構成的な言語で思考する我々人類には最早想像もつかない帰納バイアスにおいて、想像もつかない記号体系が学習しやすいと見做されることもあり得たかもしれない。これに対して適切に反論するためには、生物進化の帰結として必然的にもたらされる帰納バイアス（パラメタ領域）は如何なるものであるのか？ その帰納バイアスのもとで本当に構成性は最適であり得るのか？ という問いに答えなければならない。

また、近年の EC の発展は深層学習に依るところが大きい [10]。深層学習はエージェントの学習能力を劇的に向上させ、面白い実験結果をもたらしてくれるようになった一方で、複雑な非線形性によってその解釈は難しく、帰納バイアスと構成性についての議論を一層困難なものにしている。

以上のような背景と問題意識から、理論的に扱いやすく、“技巧倒れ”しておらず（下手に複雑化することによって解釈性や認知的妥当性を損なってしまうことなく）、生物進化の帰結に関わる議論をするのに必要な知見がある程度蓄積しているものであって（例えば、進化計算による超パラメタ探索の結果どのような領域に収斂しやすそうかなどの議論ができるとうまい）、一方である程度意味のある議論ができるように最低限の学習能力を備えたモデルを人工エージェントとして採用し、議論をすることが望ましい。本稿では、そのようなモデルとして **Echo State Network (ESN)** [11, 12, 13] を採用する。ESN は、回帰型ニューラルネットワーク (RNN) の一種であり、所謂 Simple RNN と同様の構造をもち、特に線形出力層以外をランダム初期化時の値のまま固定（凍結）したモデルである。著しい制約がある一方、（直感に反して）それなりの学習性能をもつことが時系列データを扱う研究で知られている [14]。興味深いことに、秩序相とカオス相と呼ばれる超パラメタ領域の間にある極々狭い領域—**カオスの縁**—において、ESN は特に高い性能を発揮することが知られている [15]。カオスの縁は、複雑系科学や人工生命、生命進化の文脈で古くから用いられ、カオスの縁こそが知能や生命（らしさ）の創発する領域な

のではないかと目されてきた [16]。従って、この性質は生物進化の収斂先を議論するのに相応しいといえよう。また、ESN はガウス過程 (GP) 回帰モデルとして用いることもできるため、ベイズ的な立場をとる生成的 EC の価値観ともそれほど乖離しない。

実験においては、構成性の度合いを測るための指標として **Topographic Similarity (TopSim)** [17] を採用することとした。人工的に構築したデータセットに対し、ESN に基づく GP 回帰モデルを適用した結果、カオスの縁においてモデルの TopSim が特に高くなることが明らかになった。

2 問題設定

\mathbb{M} をメッセージ空間、 \mathbb{Z} を意味空間とする。メッセージ $\mathbf{m} \in \mathbb{M}$ からそれに対応する意味 $\mathbf{z} \in \mathbb{Z}$ を予測するタスクを考え、これを ESN GP 回帰モデルで解く。直感的には ILM における親から子への継承に相当し、親を教師とするベイズ教師あり学習をする格好となる²⁾。アルファベットサイズを N_{in} とし、メッセージの最大長を L とする。このときメッセージ空間 \mathbb{M} を次のように定義する³⁾：

$$\mathbb{M} := \left\{ m_1 \cdots m_\ell \mid \begin{array}{l} m_1, \dots, m_{\ell-1} \in \{2, \dots, N_{\text{in}}\}, \\ m_\ell = 1, 1 \leq \ell \leq L \end{array} \right\}. \quad (1)$$

一方で、意味空間 \mathbb{Z} は N_{out} 次元 Euclid 空間とする⁴⁾。ここで、メッセージと意味のペアから成るデータセット $\mathbb{D} = \{(\mathbf{m}_i, \mathbf{x}_i)\}_{i=1}^{N_{\text{data}}}$ を考える。 \mathbb{D} の TopSim スコア（以降の節で定義する）は所与のパラメタ ρ_{data} におおよそ一致するようになっているものとする。これを 1:1 の大きさに訓練データ $\mathbb{D}_{\text{train}}$ と評価データ \mathbb{D}_{eval} に分割して学習と評価に使う。

3 ESN

入力系列 $\{\mathbf{u}_t \in \mathbb{R}^{N_{\text{in}}}\}_{t=1}^T$ に対して、時刻 t における ESN の（隠れ）状態 $\mathbf{h}_t \in \mathbb{R}^{N_{\text{rec}}}$ および出力 $\mathbf{o}_t \in \mathbb{R}^{N_{\text{out}}}$ は典型的には以下のように定義される [13]：

$$\mathbf{h}_t := (1 - a)\mathbf{h}_{t-1} + af(\mathbf{W}_{\text{rec}}\mathbf{h}_{t-1} + \mathbf{W}_{\text{in}}\mathbf{u}_t) \quad (2)$$

$$\mathbf{o}_t := \mathbf{W}_{\text{out}}\mathbf{h}_t. \quad (3)$$

2) ここで、GP 回帰モデルには「不確実性を考慮する 1 人の子エージェントによる推論」と「点推定する子エージェント集団の分布」という 2 通りの解釈があり得るように思われるが、本稿ではその解釈を保留しておく。

3) 1 をメッセージの終端を意味する特別な記号として扱う。

4) そもそも意味とは何か？という問題に立ち入ることは避けるが、当分野がニューラルネットワークのパラダイムに従い続ける限り、データの変換過程で必ず経由する実数ベクトルを便宜的に意味と呼ぶのは不自然ではないと思う。

ここで、 $f(\cdot)$ は要素ごとの（非線形）写像であり、 \tanh や ReLU などが用いられる。 $a \in (0, 1]$ は漏れ率と呼ばれる超パラメタであり、機械学習分野における残差接続に類似した機構を提供する。3つの行列 $\mathbf{W}_{\text{rec}} \in \mathbb{R}^{N_{\text{rec}} \times N_{\text{rec}}}$, $\mathbf{W}_{\text{in}} \in \mathbb{R}^{N_{\text{rec}} \times N_{\text{in}}}$, $\mathbf{W}_{\text{out}} \in \mathbb{R}^{N_{\text{out}} \times N_{\text{rec}}}$ のうち $\mathbf{W}_{\text{rec}}, \mathbf{W}_{\text{in}}$ はランダム初期化をしたのち固定（凍結）し、 \mathbf{W}_{out} だけを最適化する。 $\mathbf{W}_{\text{rec}}, \mathbf{W}_{\text{in}}$ の初期化は $\sigma_{\text{rec}}, \sigma_{\text{in}} > 0$ を超パラメタとして以下のようにして行うこととする⁵⁾：

$$(\mathbf{W}_{\text{rec}})_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{\sigma_{\text{rec}}^2}{N_{\text{rec}}}\right), \quad (\mathbf{W}_{\text{in}})_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_{\text{in}}^2\right). \quad (4)$$

入力に関する記号の濫用 入力 $\mathbf{u}_t \in \mathbb{R}^{N_{\text{in}}}$ は N_{in} 次元であるとしたが、本稿で実際に考えることになる入力記号は有限の自然数であり、行列演算をそのまま適用することができない。入力を one-hot 符号化することでこの問題に対処するが、これは言語処理分野における一般的なプラクティスであるため以降の議論ではわざわざ明示しない⁶⁾。

初期状態の決め方 \mathbf{h}_0 をゼロベクトルではなくランダム系列を 100 ステップ入力して得られたベクトルとする。これは RC 分野における一般的なプラクティスである（便宜的に“warm up”と呼ぶ）。

4 構成性の指標 TopSim

メッセージ空間 \mathbb{M} , 意味空間 \mathbb{Z} にはそれぞれ距離関数 $d_{\mathbb{M}}(\cdot, \cdot), d_{\mathbb{Z}}(\cdot, \cdot)$ が備わっているとす。また、メッセージと意味のペアから成るデータセット $\mathbb{D} = \{(\mathbf{m}_i, \mathbf{z}_i)\}_i^N$ が与えられているとし、距離関数とデータセットを用いて以下のデータ

$$\times := \{(d_{\mathbb{M}}(\mathbf{m}_i, \mathbf{m}_j), d_{\mathbb{Z}}(\mathbf{z}_i, \mathbf{z}_j)) \mid 1 \leq i < j \leq N\}, \quad (5)$$

を考える。このとき、 \mathbb{D} に関する TopSim $\rho(\mathbb{D})$ を、データ \times の Pearson 相関係数とする⁷⁾⁸⁾。本稿では、 $d_{\mathbb{M}}$ を Levenshtein 距離、 $d_{\mathbb{Z}}$ を Euclid 距離とする。

5 ESN を用いる意義

LSTM [18] や Transformer [19] のような比較的最近のモデルを使わず、わざわざ ESN を採用する意義

- 5) \mathbf{W}_{rec} の初期化はスペクトル半径を制御する形で行われることの方が多い [13] が、各成分を独立に初期化する方が“創発”に関する研究としては自然と考えこのようにしている。
- 6) あえて明示的に書くなら $\mathbf{h}_t = (1 - a)\mathbf{h}_{t-1} + af(\mathbf{W}_{\text{rec}}\mathbf{h}_{t-1} + \mathbf{W}_{\text{in}}\text{OneHot}(\mathbf{u}_t))$ 。あるいは $(\mathbf{W}_{\text{in}})_{:,u_t}$ などとしてもよい。
- 7) [17] に倣って Pearson 相関係数を用いているが、Spearman 相関係数を用いることも多い。
- 8) これが構成性の指標であることは直感的ではないではないかもしれないが、構成性が「写像の構造保存性」として特徴づけられるならば、TopSim は「写像がどの程度距離を保存するか」を測る指標であるため、そこまで不自然ではない。

については第 1 節でも触れたが、本節で少し補足的な説明を加えておく。

5.1 理論的な扱いやすさ

出力層 \mathbf{W}_{out} 以外を凍結していることから、回帰層は古典的機械学習の意味での素性関数 $\phi(\mathbf{m})$ と見做せる。故に ESN は線形回帰モデルである：

$$\mathbf{o} = \mathbf{W}_{\text{out}}\phi(\mathbf{m}). \quad (6)$$

実際、ESN を用いた研究ではリッジ回帰によって \mathbf{W}_{out} を求めることが多い。さらに、

$$k(\mathbf{m}, \mathbf{m}') := \frac{1}{N_{\text{rec}}} \langle \phi(\mathbf{m}), \phi(\mathbf{m}') \rangle + \sigma_{\text{out}}^2 \delta(\mathbf{m}, \mathbf{m}') \quad (7)$$

などとおけば、 $k(\cdot, \cdot)$ はカーネル関数として使えるため、GP 回帰 [20, 21] を適用することもできる [22, 23]。ここで、 $\langle \cdot, \cdot \rangle$ は通常の内積であり、 $\delta(\cdot, \cdot)$ は 2 引数が同一のとき 1、それ以外のとき 0 を返す関数である。訓練用メッセージデータ $\{\mathbf{m}_i\}_{i=1}^{N_{\text{train}}}$, 訓練用意味データ $\{\mathbf{y}_i\}_{i=1}^{N_{\text{train}}}$, 評価用メッセージデータ $\{\mathbf{m}_i^*\}_{i=1}^{N_{\text{eval}}}$ が与えられたとき、以下の 4 行列

$$\begin{aligned} (\mathbf{K})_{i_1 i_2} &:= k(\mathbf{m}_{i_1}, \mathbf{m}_{i_2}), & (\mathbf{K}_*)_{i_3 i_4} &:= k(\mathbf{m}_{i_3}^*, \mathbf{m}_{i_4}), \\ (\mathbf{Y})_{i_1 i_5} &:= (\mathbf{y}_{i_1})_{i_5}, & (\mathbf{K}_{**})_{i_3 i_4} &:= k(\mathbf{m}_{i_3}^*, \mathbf{m}_{i_4}^*) \end{aligned} \quad (8)$$

を用いれば（ただし $1 \leq i_1, i_2 \leq N_{\text{train}}, 1 \leq i_3, i_4 \leq N_{\text{eval}}, 1 \leq i_5 \leq N_{\text{out}}$ ），評価用メッセージデータに対する意味の予測分布は以下のように表される：

$$\mathcal{N}(\mathbf{K}_* \mathbf{K}^{-1} \mathbf{Y}, \mathbf{K}_{**} - \mathbf{K}_* \mathbf{K}^{-1} \mathbf{K}_*^T). \quad (9)$$

5.2 認知的な妥当性

認知的に妥当と思われるモデルを採用することは EC 分野においても重要であるが、LSTM や Transformer のような比較的最近のモデルが人間らしいといえるかどうかについては未だはっきりとした結論が得られていない。1つの妥協案として、できるだけシンプルなモデルを採用することがあり得る。ESN は、非常に簡素な連続時間微分方程式で記述されるニューロンモデルの自然な離散化として導けることが知られている [24]。

5.3 複雑系科学（カオスの縁）の再訪

ESN の研究ではカオスの縁という概念がしばしば登場する。カオスの縁とは、秩序相とカオス相の間に存在する狭い超パラメタ領域であり、この領域において ESN の性能が最も高くなることが経験的に知られている。ESN が秩序相、カオス相、カオスの

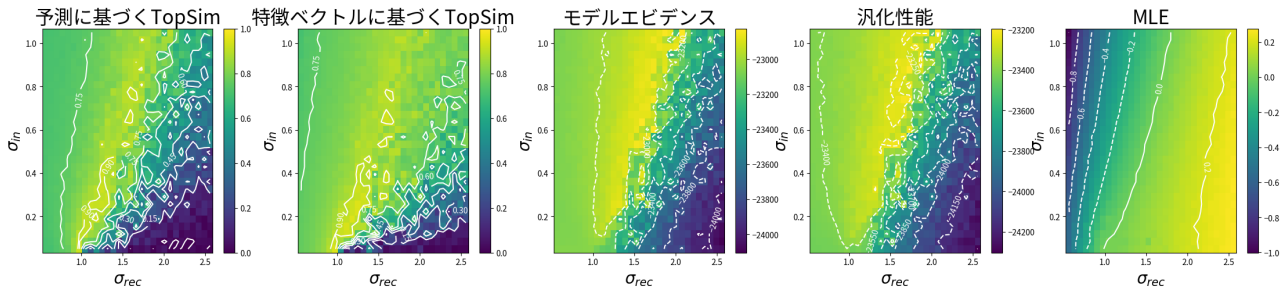


図1 超パラメタ σ_{rec} (横軸) 及び σ_{in} (縦軸) と各指標との関係。

縁のいずれに位置しているのか確認するため**最大リヤブノフ指数 (MLE)** [25] がしばしば用いられる：

$$\lambda_{mle} := \lim_{T \rightarrow \infty} \frac{1}{T} \log \left\| \frac{\partial \mathbf{h}_T}{\partial \mathbf{h}_{T-1}} \frac{\partial \mathbf{h}_{T-1}}{\partial \mathbf{h}_{T-2}} \cdots \frac{\partial \mathbf{h}_1}{\partial \mathbf{h}_0} \right\|. \quad (10)$$

$\|\cdot\|$ は行列のスペクトルノルムとする。 $\lambda_{mle} < 0$ のとき秩序相, $\lambda_{mle} > 0$ のときカオス相, そして $\lambda_{mle} \approx 0$ のときカオスの縁を意味する。 $\lambda_{mle} < 0$ の場合, \mathbf{h}_t が“小さく”なるようとするため, 過去の入力情報の記憶がすぐに失われてしまう。逆に $\lambda_{mle} > 0$ の場合, \mathbf{h}_t が“大きく”なるようとするため, 過去の小さなノイズを拡大して無秩序な状態になってしまう。 $\lambda_{mle} \approx 0$ のときにのみ, 頑健性と柔軟性を両立したちょうどよい状態が保たれる。

6 実験

第2節で述べた問題設定を考え, 超パラメタ $\sigma_{rec}, \sigma_{in}$ と評価指標の関係について議論する (その他の超パラメタは表1を参照)⁹⁾¹⁰⁾。

表1 $\sigma_{rec}, \sigma_{in}$ 以外の超パラメタ。

N_{rec}	N_{in}	N_{out}	N_{data}	ρ_{data}	a	L	σ_{out}	f
512	32	32	1024	0.7	1	64	1	tanh

6.1 評価指標

予測に基づく TopSim 評価用データからの入力 (メッセージ) と, それに対する GP 回帰の推定値 $\mathbf{K}_* \mathbf{K}^{-1} \mathbf{Y}$ の間の TopSim を評価指標として用いる。

特徴ベクトルに基づく TopSim ESN が特徴関数 $\phi(\cdot)$ と見做せることは既に述べたが, 特徴関数から得られた特徴ベクトルを“意味”と見做すことによっても TopSim を定義できる。

モデルエビデンス 訓練データに対する対数周辺尤度 (モデルエビデンス) は, 学習パラメタ \mathbf{W}_{out} を周辺化することによって超パラメタの影響を明らかにするのに役立つ。

9) 人工データセットの作成方法については付録 A を参照。

10) また, 議論を補強するため, 教師データが全く構成的でない場合 ($\rho_{data} = 0$) についても実験を行った (付録 B 参照)。

汎化性能 GP 回帰モデルによる, 評価用データに対する対数尤度, 即ち汎化性能を測定する。

MLE $\lambda_{mle} \approx 0$ となる領域がカオスの縁であり, カオスの縁は生物進化による超パラメタの収斂先として尤もらしい領域であると期待できる。定義から明らかなように厳密な計算は困難であるため, “warm up” で得た \mathbf{h}_0 を初期状態とし, $T = 100$ として T ステップのランダム系列を入力して得た状態列 $\mathbf{h}_0, \dots, \mathbf{h}_T$ から推定した。

6.2 TopSim はカオスの縁で最も高くなる

$\sigma_{rec}, \sigma_{in}$ と各指標の関係を図1に示す。カオスの縁 ($\lambda_{mle} \approx 0$) においてモデルの性能 (モデルエビデンス・汎化性能) が最も高くなると同時に, TopSim も最も高くなることから分かる。この結果は「言語の構成性の創発は学習のしやすさに起因する」という従来の EC の視点に対して「モデルの学習能力が最大となるカオスの縁において特徴ベクトルに基づく TopSim が最も高くなるため, モデルは TopSim の高い予測をしやすくなる」という新視点を与える。

7 今後の課題

本稿の内容は実験結果の報告に留まった。今後, カオスの縁において TopSim が高くなる原理, 特に $\phi(\mathbf{m})$ の距離構造について議論する必要がある。ここで, $\|\phi(\mathbf{m}) - \phi(\mathbf{m}')\| \propto \sqrt{k(\mathbf{m}, \mathbf{m}) + k(\mathbf{m}', \mathbf{m}') - 2k(\mathbf{m}, \mathbf{m}')}$ という関係に注目すると¹¹⁾, カーネル関数の解析解を得て¹²⁾, それを上手く解釈する必要があることが伺える。これに対しては, ESN の構造の簡潔さに加え, 出力層以外全て凍結していることから, Deep Neural Network as Gaussian Process (DNN-GP) [23, 26, 27] の方法論が適用できると考えられるため, 今後検証する予定である。

11) ただし $\sigma_{out} = 0$ の場合に限る。

12) つまり, 初期化に起因するランダム性を極限 $N_{rec} \rightarrow \infty$ で“周辺化”する。

謝辞

本研究は JSPS 科研費 JP23KJ0768, JST ACT-X (JPMJAX24C5) の助成を受けたものです。

参考文献

- [1] 上田亮, 谷口忠大, 鈴木麗瑩, 江原広人, 中村友昭, 岩村入吹, 橋本敬. 言語とコミュニケーションの創発に関する構成論的研究の展開. *認知科学*, Vol. 31, No. 1, 2024.
- [2] Tadahiro Taniguchi, Yuto Yoshida, Yuta Matsui, Nguyen Le Hoang, Akira Taniguchi, and Yoshinobu Hagiwara. Emergent communication through metropolis-hastings naming game with deep generative models. *Adv. Robotics*, Vol. 37, No. 19, 2023.
- [3] Ryo Ueda and Tadahiro Taniguchi. Lewis’s signaling game as beta-vae for natural word lengths and segments. In **The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024**. OpenReview.net, 2024.
- [4] Tadahiro Taniguchi, Ryo Ueda, Tomoaki Nakamura, Masahiro Suzuki, and Akira Taniguchi. Generative emergent communication: Large language model is a collective world model, 2024.
- [5] Tadahiro Taniguchi. Collective predictive coding hypothesis: symbol emergence as decentralized bayesian inference. *Frontiers in Robotics and AI*, Vol. 11, , 2024.
- [6] Simon Kirby. Spontaneous evolution of linguistic structure—an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, Vol. 5, No. 2, 2001.
- [7] Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, Vol. 141, , 2015.
- [8] Simon Kirby, Hannah Cornish, and Kenny Smith. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, Vol. 105, No. 31, 2008.
- [9] Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B. Cohen, and Simon Kirby. Compositional languages emerge in a neural iterated learning model. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.
- [10] Angeliki Lazaridou and Marco Baroni. Emergent multi-agent communication in the deep learning era, 2020.
- [11] Herbert Jaeger. The “echo state” approach to analysing and training recurrent neural networks. Technical report, German National Research Center for Information Technology GMD Technical Report 148, 2001. Erratum note available at <https://www.ai.rug.nl/minds/uploads/EchoStatesTechRepErratum.pdf>.
- [12] Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, Vol. 3, No. 3, 2009.
- [13] Mantas Lukoševičius. **A Practical Guide to Applying Echo State Networks**. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
- [14] Gouhei Tanaka, Toshiyuki Yamane, Jean Benoit Héroux, Ryosho Nakane, Naoki Kanazawa, Seiji Takeda, Hidetoshi Numata, Daiju Nakano, and Akira Hirose. Recent advances in physical reservoir computing: A review. *Neural Networks*, Vol. 115, , 2019.
- [15] Guillermo B. Morales and Miguel A. Muñoz. Optimal input representation in neural systems at the edge of chaos. *Biology*, Vol. 10, No. 8, 2021.
- [16] S.A. Kauffman. **At Home in the Universe: The Search for Laws of Self-organization and Complexity**. Oxford paperbacks. Oxford University Press, 1995.
- [17] Henry Brighton and Simon Kirby. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artif. Life*, Vol. 12, No. 2, 2006.
- [18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, Vol. 9, No. 8, 1997.
- [19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.
- [20] Carl Edward Rasmussen and Christopher K. I. Williams. **Gaussian Processes for Machine Learning**. The MIT Press, 2006.
- [21] 持橋大地, 大羽成征. ガウス過程と機械学習. 機械学習プロフェッショナルシリーズ. 講談社サイエンティフィク, 東京, 2019.
- [22] Sotirios P. Chatzis and Yiannis Demiris. Echo state gaussian process. *IEEE Trans. Neural Networks*, Vol. 22, No. 9, 2011.
- [23] Michiel Hermans and Benjamin Schrauwen. Recurrent kernel machines: Computing with infinite echo state networks. *Neural Comput.*, Vol. 24, No. 1, 2012.
- [24] Herbert Jaeger, Mantas Lukoševičius, Dan Popovici, and Udo Siewert. Optimization and applications of echo state networks with leaky-integrator neurons. *Neural Networks*, Vol. 20, No. 3, 2007. Echo State Networks and Liquid State Machines.
- [25] Holger Kantz and Thomas Schreiber. **Nonlinear Time Series Analysis**. Cambridge University Press, 2 edition, 2003.
- [26] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In **6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings**. OpenReview.net, 2018.
- [27] Greg Yang. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. In **Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019**.

A 人工データセットの構築方法

本節では、本編にて言及した人工データセットの作成方法、即ち、人工メッセージデータ \mathbb{M}_{data} と意味データ \mathbb{M}_{data} について説明する。

データセット全体のサイズを N_{data} とし、所与の TopSim を ρ_{data} とする。まず、メッセージ空間 \mathbb{M} 上の一様 uni-gram モデル（終端記号 1 を生成するまで、あるいは最大長 L の 1 つ手前に達するまで記号を一様分布から生成する）から N_{data} 個になるまでメッセージを繰り返しサンプリングする。このとき、全てのメッセージが互いに異なるようにする（もし同一のメッセージをサンプリングしてしまった場合は棄却し、サンプリングし直す）。さもなくば、グラム行列 \mathbf{K} が逆行列を持たなくなってしまうからである。このようにして得られたメッセージデータを $\mathbb{M}_{\text{data}} := \{\mathbf{m}_i\}_{i=1}^{N_{\text{data}}}$ とする。次に、一時的な意味データ $\mathbf{Z}' := \{\mathbf{z}'_i \in \mathbb{R}^{N_{\text{out}}}\}_{i=1}^{N_{\text{data}}}$ を次のように初期化する：

$$\mathbf{z}'_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (11)$$

このとき、以下の目的関数 J を考える：

$$J := (\rho_{\text{data}} - \rho(\mathbb{X}))^2. \quad (12)$$

ここで、 \mathbb{X} は

$$\mathbb{X} := \left\{ \left(d_{\mathbb{M}}(\mathbf{m}_i, \mathbf{m}_j), d_{\mathbb{Z}}(\mathbf{z}'_i, \mathbf{z}'_j) \right) \mid 1 \leq i < j \leq N_{\text{data}} \right\} \quad (13)$$

とする。ただし、各 i について $\mathbf{z}'_i := (\mathbf{z}'_i - \mathbb{E}_k[\mathbf{z}'_k]) / \sqrt{\text{Var}_k[\mathbf{z}'_k]}$ とする（途中で意味データの中心やスケールが変化してしまうのを防ぐため）。この目的関数 J を \mathbb{Z}' に関して勾配法で最小化する。 J の収束後、意味データ \mathbb{Z}_{data} を $\{\mathbf{z}'_i\}$ ではなく $\{\mathbf{z}_i\}$ の集合として得る。

B 教師データが全く構成的でない場合 ($\rho_{\text{data}} = 0$) の結果

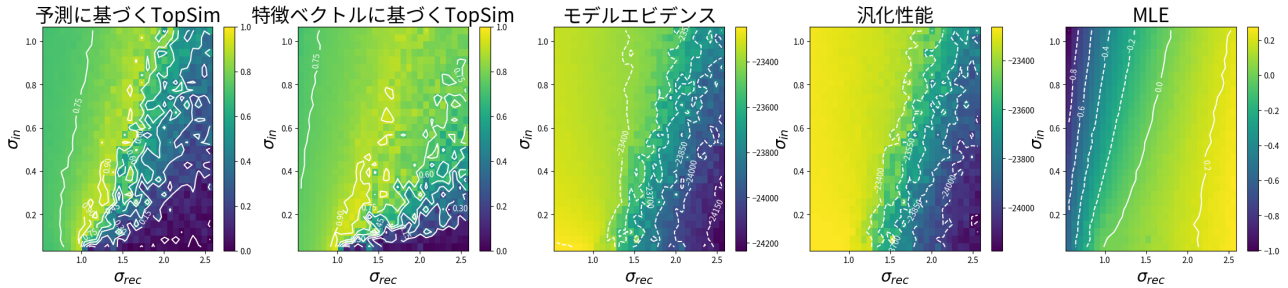


図 2 $\rho_{\text{data}} = 0$ とした場合の、超パラメータ σ_{rec} （横軸）及び σ_{in} （縦軸）と指標の関係。

$\rho_{\text{data}} = 0$ とした場合の結果を図 2 に示す。本編での結果とは異なり、モデルの性能（モデルエビデンス・汎化性能）はカオスの縁ではなく秩序相 ($\lambda_{\text{mle}} < 0$) において最も高くなっている。これは、メッセージと意味の関係が全くの出鱈目（ランダム）であるために、どんな入力に対しても正規分布に近い予測分布を返すのが有利になるからであると考えられる。実際、 $\sigma_{\text{rec}}, \sigma_{\text{in}}$ が小さくなるほど σ_{out} の影響が支配的になり、 $k(\mathbf{m}, \mathbf{m}') \approx \sigma_{\text{out}}^2 \delta(\mathbf{m}, \mathbf{m}')$ となることからこのことが伺えよう。だが、このような設定においてもやはりカオスの縁 ($\lambda_{\text{mle}} \approx 0$) において予測に基づく TopSim が最も高くなる傾向にあることが分かる。このことから、エージェントが初めからカオスの縁に位置していさえいれば、たとえ“原始の言語”が全く構成的でなかったとしても、世代を経るごとに言語の構成性が高くなるような圧力がかかることが示唆される。