擬似均衡コーパスの構築と統計的性質

高村大也 [1]　永田亮 [2]　川崎義史 [3]　大谷直輝 [4]
[1] 産業技術総合研究所　[2] 甲南大学　[3] 東京大学　[4] 東京外国語大学
takamura.hiroya@aist.go.jp

概要

言語研究への利用を目指し，LLM を用いて擬似均衡コーパスを構築し，その統計的性質を分析する．特に，単語や文の多様性，単語の多義性などの点から，擬似コーパスが実コーパスとどの程度類似しているかを調べる．また，構築方法が擬似コーパスの質へ与える影響を調べる．実験の結果，擬似コーパスは，多様性や多義性の観点では実コーパスと振る舞いが異なること，ある条件下では語義頻度則に従うことなどがわかった．

1　序論

言語学では，かつては分析者自身が内省を通して考えた作例に基づいた研究が主流であったが，現在は実際に発話された実例から成る言語コーパスに基づく実例主義の重要性も認識されている．しかし，コーパスを用いた分析では，コーパスが未整備の言語使用域や出現頻度が低い現象などを扱うことが難しい．また，様々な条件でのコーパスを構築することは非常にコストが高い作業となる．これらの問題点を乗り越える手段として言語モデルの活用が考えられる．特に近年急速に発展した大規模言語モデル (LLM) は言語理解と言語生成の能力が高く，高性能な LLM により生成された文章は，人間が書いた文章としばしば見分けがつかない．つまり，LLM の出力を**擬似コーパス**として，実例から成る実コーパスのように言語研究に利用できる可能性がある．

一方，LLM や擬似コーパスの言語研究への利用には慎重になる必要もある．第一に，擬似コーパスは言語使用の自然なサンプルであるか不明である．特に均衡コーパスを目指した場合，この点は重要になる．また，LLM の学習データにまったく含まれていない言語使用域の擬似コーパスは信頼できないと考えられる．つまり，LLM は言語研究にどこまで使えるのかに答える必要がある．

この問いに答えるための第一歩として，日本語の均衡コーパスである『現代日本語書き言葉均衡コーパス』(BCCWJ)[1] を参考に，LLM を用いて擬似均衡コーパスを構築し，その統計的性質を分析する．特に，単語や文の多様性，単語の多義性などの点から，擬似コーパスが実コーパスである BCCWJ とどの程度類似しているかを調べる．また，構築方法が擬似コーパスの質へ与える影響を調べる．

2　関連研究

LLM の模倣学習や知識蒸留のために合成コーパスを学習データとして用いる試みは多く行われている [2, 3]．学習における有効性は示されているが，言語使用のサンプルとしての質は不明である．

言語モデルからコーパスを生成する試みとしては，子供向けの物語の単語頻度の分析 [4]，人間とコンピュータの対話データの生成 [5]，ChatGPT と自然言語の単語レベルの比較 [6, 7] などがある．また，合成テキストと実テキストの近さを測る試みもある [8]．これらに対し，本研究は，文書単体でなくコーパスとしての質を，様々な観点から測ることを試みたものである．

3　擬似コーパス生成方法

3.1　BCCWJ の図書館分類分布の活用

BCCWJ[1] は緻密に設計されたサンプリングにより構築され，代表性を有する日本語均衡コーパスである．BCCWJ におけるサンプルの分野などの分布を用い，擬似均衡コーパスを構築する．特に，BCCWJ の図書館サブコーパスに着目し，サンプリング設計時の各時期[1)]と日本十進分類の各第一次区分[2)]に対するサンプル数を利用する (詳細は 3.2 節) [9][3)]．

1）　1986-1990 年，1991-1995 年，1996-2000 年，2001-2005 年
2）　0. 総記，1. 哲学，2. 歴史，3. 社会科学，4. 自然科学，5. 技術，6. 産業，7. 芸術，8. 言語，9. 文学，n. 分類なし．
3）　17 から 20 ページの表に記載されている．

3.2 生成方法

生成は LLM を用いて段階的に行う．日本語で生成するように指示を与える．まず，出版時期 p と日本十進分類第一次区分 c を与え，それに合った架空の書籍名および著者名を生成する．その上で，p，c，書籍名，著者名に合った架空の書籍の要約を生成する．さらに，要約を条件に加えて書籍本文の一部を生成する．このような生成を本稿では三段階生成とよび，要約生成段階を除いたものを二段階生成とよぶ．ただし，同じ書籍名が複数回生成されることを抑制するため，同一の p と c に対してそれまでに生成した書籍名のリストを与え，それと異なる書籍名を生成するように指示した．

1 サンプルの長さは，BCCWJ の設計に従い 3,900 文字を基準とするが，LLM は出力長を完全には制御できないため[4]，足りない場合は続きを生成することで，おおよそ 3,800 文字から 3,900 文字にする．3.1 節で説明したように，出版時期 p と第一次区分 c に対しサンプル数 s が決まるので，その回数だけ生成を繰り返す．詳細は，Algorithm 1 と付録 A のプロンプトを参照されたい．比較のため，BCCWJ のサンプル数を使わず，全ての (p, c) について同じサンプル数を使う一様分布を用いた実験も行った．

LLM として，OpenAI の GPT-4o mini[5] および GPT-5[6]，オープンなモデルである Llama 3.1 Swallow[7] (以下，Swallow) の三種類を用いた．それぞれについて，上記の二段階生成と三段階生成を試した．温度パラメータは $t = 1.0$ としたが，GPT-4o mini については $t = 1.2$ も試した．

4 分析手法

前処理としては，spacy[8] の Sentencizer で文分割を，ginza[9] で単語分割を行った．文脈依存の単語埋め込みの計算には日本語 BERT-large [10] を，文埋め込みの計算には日本語 BERT-base[11] を用いた．

Algorithm 1 擬似コーパス生成手順

```
corpus ← 空集合
P={1986-1990,1991-1995,1996-2000,2001-2005}
NDC={0. 総記, 1. 哲学, . . ., 9. 文学, n. 分類なし }
for p ∈ P and c ∈ NDC do
    s ← BCCWJ での p,c に対するサンプル数
    for i = 0 to s − 1 do
        text ← 空文字列
        [ LLM による書籍名・著者名生成 ]
        [ LLM による要約生成 ]
        while | 本文 | < 3900 文字 do
            [ LLM による本文の続きの生成 ]
        end while
        [ 本文を文分割 ]
        [ 3800 文字を超えるまで text に文を追加 ]
        [ corpus に text を追加 ]
    end for
end for
return corpus
```

分析の観点は主に，擬似コーパスについて，語彙や文集合が十分な多様性を有するか，実コーパスと同様の語彙を持つか，各単語は実コーパスと同様の多義性を持つか，である．

4.1 語彙の分析

語彙の分析は，多様性の分析，頻度順位と頻度の関係，実コーパスとの類似性の三つに大別される．

まず，語彙の多様性は以下の二つの指標で測る．

- **語彙サイズ** $|V|$: コーパス内の単語タイプ数．
- **タイプトークン比** TTR: コーパス内の単語タイプ数を単語トークン数で除算した値であり，TTR が大きいほど語彙の多様性が高い．

語彙の多様性と関連し，単語の頻度と頻度順位の関係を調べる．頻度順位に対する頻度の変化が，実コーパスと擬似コーパスでどのように異なるかを下記の分布を観察することで調査する．

- **順位頻度分布**: 各単語に対し，頻度順位の対数を横軸に，頻度の対数を縦軸にしてプロットしたものである．定義上右下がりのグラフになるが，下がり方が激しい場合は，低頻度帯の単語がより出現しにくいことを表す． 線形回帰 (傾き δ_r) による近似精度が高い場合，Zipf の法則に従うといえる．

実コーパスの語彙と擬似コーパスの語彙の類似性については，以下の 2 つの指標を用いる．

- **ジャッカード係数** J: 両者の語彙の共通部分のサイズを，和集合のサイズで除算したもの．
- **多重集合に対するジャッカード係数** J_{multi}: コー

4) オープンモデルで eos 記号の出力を禁止して実験を行ったが，repetition が発生するなどテキストの乱れが生じた．
5) https://platform.openai.com/docs/models/gpt-4o-mini
6) https://platform.openai.com/docs/models/gpt-5
7) https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5
8) https://spacy.io/
9) https://megagonlabs.github.io/ginza/
10) https://huggingface.co/tohoku-nlp/bert-large-japanese-v2
11) https://huggingface.co/tohoku-nlp/bert-base-japanese

パス C における単語 w の頻度を $f_{w,C}$ で表すと，$\sum_w \min(f_{w,C_1}, f_{w,C_2}) / \sum_w \max(f_{w,C_1}, f_{w,C_2})$ と算出できる．単語の有無だけを考慮する J と異なり，J_{multi} では頻度も考慮される．

4.2 単語の多義性の分析

多義性については，以下の観点から分析を行う．

- **多義度** v: 各単語タイプに関し，BERT で求めた文脈依存埋め込みを正規化し，von Mises-Fisher 分布の集中度パラメータ κ を算出した．その逆数 $v = 1/\kappa$ を多義度と定義する．v が大きいほど多義性が高い [10].
- **語義頻度則** [11]: 多義度 v と単語頻度 f の間に，$\log(v) = \delta_m \log(f) + (\text{定数})$ という関係が実コーパスでは成立するとされている．ここで δ_m は傾き (決定係数は R_m^2) を表す．この関係が擬似コーパスについても成立するか検証する．

4.3 文集合の多様性の分析

コーパスを文集合とみなしたときの多様性は以下の指標で測る．$S_{\min}$ が大きいほど多様性が高い．

- **最小クラスタ内平方和** $S_{\min}$: 各文を日本語 BERT で解析し得られた [CLS] の埋め込み x を，文ベクトルとする．各コーパスを文ベクトル集合 D とみなし，次のように計算する:

$$S_{\min} = \min_{c_1, c_2, \ldots, c_K} \sum_{x \in D} \min_{k \in \{1,\ldots,K\}} ||x - c_k||^2. \quad (1)$$

ここで c_k はクラスタ中心であり，右辺の二つの最小化は k 平均法[12]により近似的に計算する [12]．クラスタ数は $K = 1000$，最大繰り返し回数は 100 とし，初期値の決定には k-means++を用いた．D としては，各コーパスの “9. 文学” 部分の，長さ 20 文字以上の最初の 100,000 文を用いる．

4.4 日本語確率

一般的に LLM は流暢性については優れているが，プロンプトに指定されたものと異なる言語でテキストを生成することがある．この点でのテキストの質を測るため，日本語文を生成する確率を測る．

- **日本語確率** P_{jpn}: コーパスの各文に対し，langdetect[13]により言語判定を行い，日本語と判定された文の割合を算出する．ただし，数字や数式などの文字列により言語判定精度が低下する可能性があるため，日本十進分類 “9. 文学” の長さ 20 文字以上の文のみを計算対象とする．

12) https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
13) https://pypi.org/project/langdetect/

5 結果と考察

実験で得られた統計値を表 1 に示す．*GPT-4o mini は，BCCWJ のサンプル分布でなく一様分布を利用した場合に対応する．以下，結果に対する考察を行う．生成されたテキストの例を付録 B に示す．

5.1 語彙の分析結果

語彙サイズ $|V|$ と TTR の値は，実コーパスでは語彙が豊富であることを示す．実コーパスと同程度以上の値を持つのは $t = 1.2$ の場合だけであるが，これらは日本語確率 P_{jpn} が相対的に低い．これは，プロンプトに反し日本語以外のテキストを生成しており，コーパスとしての質が低下していることを示す．$t = 1.0$ の場合では，GPT-5 の $|V|$ と TTR が他の擬似コーパスより大きいものの，実コーパスとは大きな差がある．また，三段階生成は二段階生成と比較して僅かながら語彙が豊富であることがわかる．

J と J_{multi} については，分野の一様分布を仮定した生成方法の値が低く，均衡コーパスのサンプリング設計の情報の効果が見られる．また，GPT-5 は他のモデルよりも実コーパスに近い語彙分布を生成していることがわかる．

図 1 に，順位頻度分布を示す．見やすさのため，一部のコーパスのみ示す．いずれのコーパスについても決定係数は 0.95 以上であり，Zipf の法則が成り立つことがわかる．図 1 および表 1 の傾き δ_r の値より，GPT-5 が実コーパスに近いことがわかる．一方，擬似コーパスでは実コーパスほどには低順位の語を産出できていない点が明確な差異である．いずれのプロットも低順位帯での傾きの減衰が実コーパスより激しい．LLM の出力を局所的に観察すると人間の産出する文章と遜色ないようにみえるが，大域的な語彙の使用分布は異なることがわかる．

5.2 多義性の分析結果

図 2 に，頻度 100 以上の単語について多義度をプロットした．x 軸は実コーパスの多義度，y 軸は擬似コーパスの多義度である．生成には三段階生成を用いた．いずれの図においても多くのプロット点が $y = x$ の直線の下側にあり，ほとんどの単語につい

表 1: 実験結果

生成モデル	t	文書数	文字数	トークン数	$\|V\|$	TTR	J	J_{multi}	δ_{r}	δ_{m}	$S_{\min}$	P_{jpn}
(実コーパス)	—	10,551	46,731,798	28,801,819	265,684	.0092	NA	NA	−1.5	.106	3539	—
*GPT-4o mini (2)	1.0	12,607	47,819,391	28,571,956	53,267	.0019	.151	.452	−2.0	.038	2686	.9999
Swallow (2)	1.0	12,607	48,340,094	30,346,338	55,331	.0018	.154	.462	−2.0	.043	2791	.9996
GPT-4o mini (2)	1.0	12,607	47,948,858	29,394,260	55,035	.0019	.158	.481	−2.0	.043	2678	.9999
GPT-4o mini (2)	1.2	12,607	48,122,936	28,542,569	287,996	.0101	.167	.503	−1.6	.007	2834	.9840
GPT-5 (2)	1.0	12,607	49,476,372	31,455,953	132,137	.0042	.261	.528	−1.8	.063	3149	.9998
Swallow (3)	1.0	12,607	48,230,896	30,262,732	59,240	.0020	.169	.474	−1.9	.049	2854	.9997
GPT-4o mini (3)	1.0	12,607	47,957,147	29,600,025	57,317	.0019	.164	.488	−2.0	.041	2599	.9999
GPT-4o mini (3)	1.2	12,607	48,135,762	28,683,064	294,731	.0102	.167	.507	−1.6	.005	2789	.9825
GPT-5 (3)	1.0	12,607	49,264,824	31,181,923	147,981	.0047	.261	.515	−1.8	.066	3271	.9991

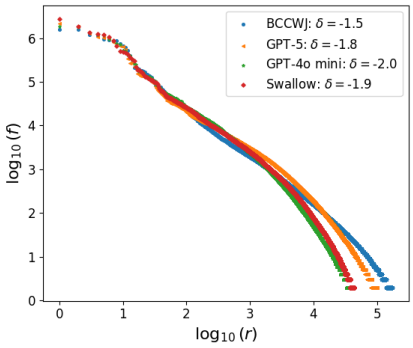
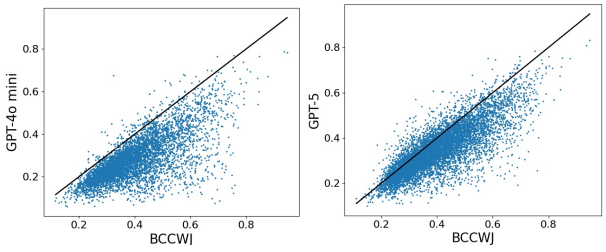


図 1: 各コーパスの順位頻度分布.



(a) GPT-4o mini **(b)** GPT-5

図 2: 多義度のプロット (実コーパス vs 擬似コーパス).

て実コーパスの方が相対的に多義性が高いことを示す．左右の図の比較から，GPT-5 の方が実コーパスに近いことがわかる．紙面の都合上二つの図だけを掲載したが他の擬似コーパスも同様の結果である．

図 3 に単語頻度と多義度をプロットする．x 軸は頻度の対数，y 軸は多義度の対数である．ただし，先行研究 [11] に従いサイズ 100 で binning を施した．青色でプロットした実コーパスは直線回帰の決定係数も高く，語義頻度則が成立している．これに最も近いのが橙色でプロットした GPT-5 であり，決定係数も高い．つまり，他の擬似コーパスと比較して，単語頻度に対する多義度の振る舞いが実コーパスに近い．一方，傾きは依然として差が大きい．GPT-4o mini や Swallow は決定係数がやや落ちる．また，図では省略したが t = 1.2 の擬似コーパスは決定係数

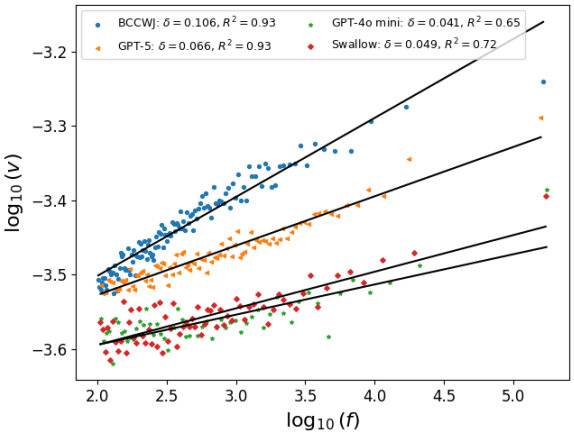


図 3: 単語頻度と多義度の関係.

が 0.05 以下であり，語義頻度則に適合しない．

5.3 文集合の多様性の分析結果

表 1 によると，実コーパスの $S_{\min}$ が最も高く，多様性が高いことがわかる．擬似コーパスの中では GPT-5 の二つが最も高い．それ以外の擬似コーパスは文集合としての多様性がやや低い．つまり，類似した文が多く生成されていることがわかる．また，単語レベルでは優劣が不明瞭であった Swallow と GPT-4o mini を比較すると，文レベルでは Swallow が優れている．文クラスタの例を付録 C に記載する．

6 結論

LLM を用いて擬似均衡コーパスを構築し，その統計的性質を分析した．擬似コーパスは単語や文の多様性が十分でないこと，低頻度語が生成できていないこと，単語の多義性が低いこと，一部の設定において語義頻度則が成立することなどがわかった．

擬似コーパスの活用のためには，人口統計学的情報の活用など，多様性の向上や，低頻度語のカバレッジの向上のための技術開発が望まれる．

謝辞

この成果は，産総研政策予算プロジェクト「フィジカル領域の生成 AI 基盤モデルに関する研究開発」および NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託業務（JPNP25006）の結果得られたものである．

参考文献

[1] 国立国語研究所コーパス開発センター. 『現代日本語書き言葉均衡コーパス』利用の手引き, 第 1.1 版, 2015.

[2] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023.

[3] Youmi Ma, Sakae Mizuki, Kazuki Fujii, Taishi Nakamura, Masanari Ohi, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Koki Maeda, Kakeru Hattori, Takumi Okamoto, Shigeki Ishida, Rio Yokota, Hiroya Takamura, and Naoaki Okazaki. Building instruction-tuning datasets from human-written instructions with open-weight large language models. In *Proceedings of the Second Conference on Language Modeling (COLM)*, 2025.

[4] Job J. Schepens, Hanna Woloszyn, Nicole Marx, and Benjamin Gagl. Can large language models generate useful linguistic corpora?: A case study of the word frequency effect in young german readers. *Open Mind : Discoveries in Cognitive Science*, Vol. 9, pp. 1597 – 1656, 2025.

[5] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. Evaluating large language models in generating synthetic HCI research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[6] Aleksandar Kostić Dragica Ljubisavljević, Marko Koprivica and Vladan Devedžić. Homogeneity of token probability distributions in chatgpt and human texts. In *Proceedings of CELDA*, 2023.

[7] Satoru Uchida. Using early llms for corpus linguistics: Examining chatgpt's potential and limitations. *Applied Corpus Linguistics*, Vol. 4, No. 1, p. 100089, 2024.

[8] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, Vol. 34, pp. 4816–4828. Curran Associates, Inc., 2021.

[9] 丸山岳彦, 山崎誠, 柏野和佳子, 佐野大樹, 秋元祐哉, 稲益佐知子, 田中弥生, 大矢内夢子. 『現代日本語書き言葉均衡コーパス』に含まれるサンプルおよび書誌情報の設計と実装, 2011.

[10] Ryo Nagata, Hiroya Takamura, Naoki Otani, and Yoshifumi Kawasaki. Variance matters: Detecting semantic differences without corpus/word alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15609–15622, Singapore, December 2023. Association for Computational Linguistics.

[11] Ryo Nagata and Kumiko Tanaka-Ishii. A new formulation of Zipf's meaning-frequency law through contextual diversity. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15323–15335, Vienna, Austria, July 2025. Association for Computational Linguistics.

[12] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, Vol. 4. Springer, 2006.

A　プロンプト

日本十進分類 “9. 文学”，出版時期 2001-2005 年に対するプロンプトの例を示す．必要に応じて，それまでに生成された書籍名のリスト，要約，そこまで生成されたテキストが後続する．

書籍名・著者名生成のプロンプト例

日本語で返答してください。挨拶や返事などは不要です。日本十進分類法 (NDC) が「9. 文学」であり、2001 年から 2005 年に出版されたような架空の書籍のタイトルと著者名を 1 個挙げてください。タイトルと著者以外は不要です。一行にタイトルと著者名をタブ区切りで書いてください。例えば、「ノルウェイの森 村上春樹」のように書いてください。ただし、下に挙げるタイトルと異なるものを挙げてください。

要約生成のプロンプト例

日本語で返答してください。挨拶や返事などは不要です。タイトルが「黄昏にほどける舟」で著者が「矢庭澄人」であり、日本十進分類法 (NDC) が「9. 文学」であり、2001 年から 2005 年に出版されたような架空の書籍の要約を生成してください。長さはおよそ 300 文字にしてください。

本文の続きの生成のプロンプト例

次のテキストは、タイトルが「黄昏にほどける舟」で著者が「矢庭澄人」であり、要約が下記のようなものであり、日本十進分類法 (NDC) が「9. 文学」であり、2001 年から 2005 年に出版されたような架空の書籍に含まれていると思われる架空のテキストです。続きの節あるいは章を 1 個生成してください。

B　生成されたテキストの例

日本十進分類 “9. 文学”，出版時期 2001-2005 年を与え，GPT-5 で三段階生成をしたときの生成テキスト例．スペースの都合上，冒頭部分のみ掲載する．

題名: 黄昏にほどける舟
著者: 矢庭澄人

第六章　継ぎ目の暮色

作業台の上に広げられた板は，父が選んで乾かしていたという古い杉だ．鉋をかけるたび，薄く巻いた削り屑が膝に落ちて，指先が松脂の匂いでぬるりとする．窓の外では，港の水面が夕刻の風に細かく震え，波間に回る光が工房の壁にゆらぎの模様を投げかけている．ここに戻ってから，湊はこのゆらぎに合わせるように腕を動かすのだと知った．父はいつもそうしていたのだろう．鉋の音と潮の音は，どちらが先か分からぬほどに似合っていた．

竜骨に沿って浮き上がったひびを見つけると，父が教えた手順の記憶が胸の底から起き上がる．まず外板を外し，肋骨の歪みを読む．木槌で拍を刻むように打ち込み，固着した釘の錆を揺らす．手の裏で木の脈を探るこの動作を，父は「舟の声を聞く」と言った．湊が子どものころ，暑い午後に眠たげな目をこすりながら，その言葉の意味を半分も分からぬまま頷いていた日のことを思い出す．

父の手は大きく，甲に走る傷はどれも海の形をしていた．春の初めに太い麻綱を手のひらで撚り直していたとき，掌に盛り上がった古い皮膚が，日差しの下で銀色に光った．その手が，嵐の夜にはどれほど確かな指針であったのか，湊は今ようやく想像できる．あの夜，防波堤に打ち付けた波の高さや，巻き上がった潮の音の厚さより，父が鎚を打つ音のほうが，湊の耳の奥には深く残っている．

固く結ばれた約束は，いつでも子どもにとっては重くて甘い．遼とかわした約束もそうだった．「灯が三つ見えたら，沖は越えない」「西の風が湿っていたら，小舟は出さない」．笑いながら，舌打ちしながら，二人は何度もそう言い合った．だが約束は，波に似ている．どれも同じに見えて，一つ一つが違う重さと勢いを持つ．嵐の夜の裂け目に，わずかに外れた結び目がほどけていく，その瞬間まで気づかない．

作業台の下には網が折り畳まれて積まれている．父が最後に手入れしたまま，塩と埃で重くなっていた．外板を持ち上げるための楔を探していた湊の膝に，その網がふっと触れ，乾いた音を立てて崩れた．絡んでいた細い浮子の間から，茶色く波打った紙が一枚，ずるりと滑り出た．指でつまむと，潮に磨かれた魚の鱗のように紙肌が冷たく，角はぼそぼそと砂のように崩れ · · ·

C　クラスタリング結果の例

GPT-4o mini で三段階生成 (t = 1.0) した擬似コーパスに対するクラスタリング結果に関し，あるクラスタ (クラスタ番号 112) の一部を記載する．

- その言葉に、リョウは少しの安心感を得た。
- その言葉に、健二の心は少しだけ温まった。
- その言葉に、健二は再び勇気を与えられた。
- その言葉に、健二は胸が締め付けられる思いだった。
- その言葉に、光は感動で胸がいっぱいになった。
- その言葉に、太一は深い意味を感じ取った。
- その言葉に、太郎の心は少し救われた気がした。
- その言葉に、恵子の心はさらに温かくなった。
- その言葉に、恵子は嬉しい気持ちと共に自信を取り戻した。
- その言葉に、悠太は少し照れくさくなりながらも、嬉しく思った。
- その言葉に、拓海は心が温かくなるのを感じた。
- その言葉に、智也は深い心の安らぎを感じた。
- その言葉に、梓は心が温かくなるのを感じた。
- その言葉に、真理の胸の内が徐々に明るくなっていくのを感じた。
- その言葉に、秋人の心は少しだけ軽くなった。
- その言葉に、陽一は自分の気持ちが解放されるのを感じた。