クレオールは計量的に峻別できるか？

川崎義史 [1]　永田亮 [2]　高村大也 [3]　大谷直輝 [4]
[1] 東京大学　[2] 甲南大学　[3] 産業技術総合研究所　[4] 東京外国語大学
ykawasaki@g.ecc.u-tokyo.ac.jp　nagata-nlp2025 @ ml.hyogo-u.ac.jp.
takamura.hiroya@aist.go.jp　otani@tufs.ac.jp

概要

本稿は、テキストデータから求めた指標の大小でクレオールと非クレオールを峻別できるか検証した。データとして、いずれの言語においても同一内容が保証されている聖書を用いた。実験の結果、両者を峻別こそできないが、複数の指標で大別することができた。これは、クレオールが非クレオールとは異なる性質を保有している可能性を示唆している。また、部分的に単純説を支持する結果を得た。

1　はじめに

意思疎通が困難な複数言語の接触により成立した即席の意思伝達手段をピジンと呼ぶ。ピジンは限られた語彙と単純な文法を特徴とする。ピジンが第二世代により母語として獲得されたものを**クレオール**と呼ぶ [1]。ピジン・クレオールの多くは、大航海時代以降、主に西欧諸語（語彙を供給する上層言語）と現地語（基層言語）との接触により生じた。

クレオールについては、相反する二つの説がある。一つは、クレオールはそれ以外の自然言語と同等に複雑だという説である（**同等説**）[2]。もう一つは、クレオールはより単純であるという説である（**単純説**）。様々な言語学的特徴に基づき両者の差異を分析した複数の研究が単純説を支持している [3, 4, 5]。しかし、テキストデータに基づいた包括的な実証的研究は管見の限り存在しない。

そこで、本稿では、テキストデータから求めた指標の大小でクレオールと非クレオールを峻別できるか検証した。両者を峻別できれば、何らかの観点で両者は異なると言える。本稿で採用した指標を表 1 に示す。データとして聖書を利用した。聖書は全言語で同一内容が保証されており、言語間の対照分析に適している。実験の結果、両者を峻別こそできないが、複数の指標で大別することができた[1)]。これは、クレオールが非クレオールとは異なる性質を保有している可能性を示唆している。また、部分的に単純説を支持する結果を得た。

1）　完璧な判別を「峻別」、およその区別を「大別」と呼ぶ。

2　関連研究

McWhorter は音韻・形態・統語・文法の観点から質的対照分析を行い、クレオールの文法が他の言語より単純であると主張している [3]。Parkvall は WALS [6] の質的データを数値化し、クレオールが他の言語に比べて単純であると主張している [4]。Bakker らは Comparative Creole Syntax [7] のデータに回帰分析と系統樹分析を適用し、クレオールが他の言語と類型論的に異なると主張している [5]。Koplenig らは聖書を利用して英語系クレオールを計量的に分析し、クレオールでは形態変化の乏しさを固定語順が補完していると主張している [8]。

3　データ

本稿では聖書をデータとして用いた。聖書はデータとしては小さいものの、全言語で同一内容が保証されており、言語間の対照分析に適している [8]。聖書のうち、BibleNLP[2)] において、下記の分析対象言語のデータが入手可能な『新約聖書』のみを使用した。ただし、ラテン語のデータは欠落していたため、別のコーパス [9] から入手した。

本稿で扱うクレオールは、メタデータにおいて family（語族）が Creole と分類されている 13 言語である：Lesser Antillean French Creole (acf), Belize English Creole (bzj), Chavacano (cbk), Aukan (djk), Sea Island English Creole (gul), Haitian Creole (hat), Kupang Malay (mkn), Kriol (rop), Saramaccan (srm), Sranan Tongo (srn), Torres Strait Creole (tcs), Tetun Dili (tdt), Tok Pisin (tpi)[3)]。ピジンは Hiri Motu (hmo) のみ利用可能だった[4)]。比

2）　https://github.com/BibleNLP/ebible
3）　以降、言語名は ISO 639 のコードに従い表記する。
4）　以降、簡単のために、ピジンも含めてクレオールと呼ぶ。

表 1　本稿で採用した指標：チェックマークは、クレオールと非クレオールを大別できることを表している。

順位頻度分布		要素の大きさ		語彙の豊富さ		パープレキシティ		意味の強さの分布	スペクトル
単語	文字	単語長	節長	TTR	Yule の K	単語	文字		
✓		✓	✓	✓		✓	✓	✓	✓

較対象として、これらのクレオールの上層言語を分析に含めた[5]：English (eng), French (fra), Spanish (spa), Malay (zlm)。上層言語の情報は Ethnologue [10] から補完した。eng, fra, spa については、レンマ化により形態変化を単純化した仮想言語（eng_lemma など）も導入した。更に、通言語的分析のために様々な語族の言語や人工言語も含めた。ただし、表音文字を使用し、単語が空白で区切られる言語に限定した。分析対象言語の詳細は付録の表 2 を参照されたい。

テキストデータの前処理として、単語の構成要素以外の記号（¿¡”’ ¶ など）や注釈を削除した。文末の.?!;: はピリオドに統一した。テキストは全て小文字に変換した。全言語において、ピリオドで区切られた単位を文、文中で空白で区切られた単位を単語とした[6]。eng, fra, spa のレンマ化は `spaCy 3.7.2`[7]で行った。eng には `core_web_lg`、fra と spa には `core_news_lg` のモデルを使用した。

4　指標

テキストデータの性質を表す指標として下記のものを採用した[8]。指標の大小でクレオールと非クレオールを峻別できるか吟味する。クレオールのデータは少なく、利用可能な最先端の言語モデルも限定的なため、主に古典的な手法を利用した。

順位頻度分布　単語と文字の出現頻度について、順位 r を横軸、相対頻度 f を縦軸とした両対数グラフに対する回帰直線 $\log f = -\eta \log r + \log a$ の傾き $\eta > 0$ を指標として用いた（Zipf の法則） [13, 14]。回帰直線は最小二乗法で求めた。

要素の大きさ　指標として単語長と節長の平均を用いた。節とは聖書本文の最小単位であり、全言語でアライメントが取られている。節は複数の文を含むこともある。文分割の方法は言語ごとに異なる可能性があるため、文長の代わりに節長を用いた。単語長は文字数、節長は単語数とした。

語彙の豊富さ　指標として、TTR（Type-Token Ratio）と Yule の K（以下、K） [15] を用いた。TTR は大きいほど、K は小さいほど語彙が豊富となる。

パープレキシティ　系列の複雑さの指標として、単語と文字の n-gram 言語モデルのパープレキシティを用いた。値が小さいほど、複雑さが低いと言える。$n \in \{2, 3\}$ とした。出現回数が 10 回以下の単語、1000 回以下の文字は未知語とした。平滑化には、ラプラス法を採用した。全言語共通で、最初の 9 割の節の文を訓練データ、最後の 1 割をテストデータとした。実装は `nltk` [16] で行った。

意味の強さの分布　「意味の強さ」 [17] の分布の指標として、単語の出現頻度と分散表現のノルムとの順位相関係数 ρ を用いた。分散表現は `gensim` の `Word2Vec` [18] で学習した。次元数は 100、エポック数は 50 とし、その他はデフォルトの設定を用いた。

スペクトル　単語の共起関係を無向グラフで表し、その隣接行列の固有値の列（スペクトル）を指標とした。スペクトルにはグラフ構造に関する情報が含まれている。スペクトルはグラフフーリエ変換 [19] により求めた。隣接行列の要素は、各言語の頻度上位 100 語の間の共起回数とした。共起は文内の前後 3 語以内とした。

5　実験結果

語彙サイズとトークン数の分布　以下の分析の前提として、各言語の語彙サイズとトークン数の分布を図 1 に示す。橙色はピジン、赤はクレオール、灰色は非クレオールの自然言語、緑はレンマ化した仮想言語、青は人工言語を表している。以下、同様の色分けを行う。クレオールは語彙サイズが小さく、トークン数が大きい傾向が見られる。語彙サイズ 4000、トークン数 24 万付近でクレオールと非クレオールを大別できる。採用した指標には、この差異が多かれ少なかれ反映されていると考えられる。

順位頻度分布　図 2 は、各言語の回帰直線の傾き η と切片の分布を示している。クレオールでは値が大きくなり、$\eta \approx 1.7$ でクレオールと非クレオールを大別できる。一方、文字レベルでは両者を大別できなかった（図は非掲載）。

要素の大きさ　平均単語長と平均節長の分布を図 3 に示す。エラーバーは標準偏差を表す。両変数には有意な負の相関が見られた（$\rho = -0.694$, $p <$

5）　データが入手不可の tdt の上層言語 Tetun (tet) を除く。
6）　空白での分割が単語認定に適切かどうかは要検討である。
7）　https://spacy.io/
8）　このほか、Taylor 解析 [11, 12] により単語の出現回数のゆらぎを吟味したが、クレオールを峻別できなかった。

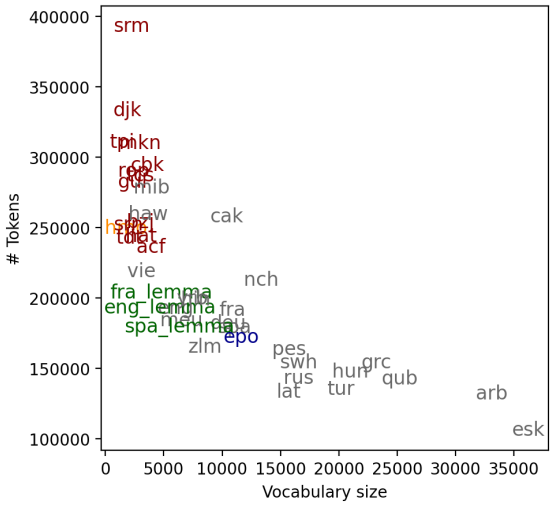

図 1　語彙サイズとトークン数の分布

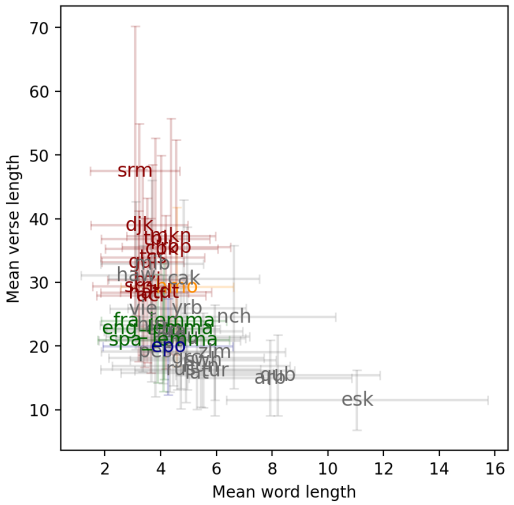

図 3　平均単語長と平均節長の分布

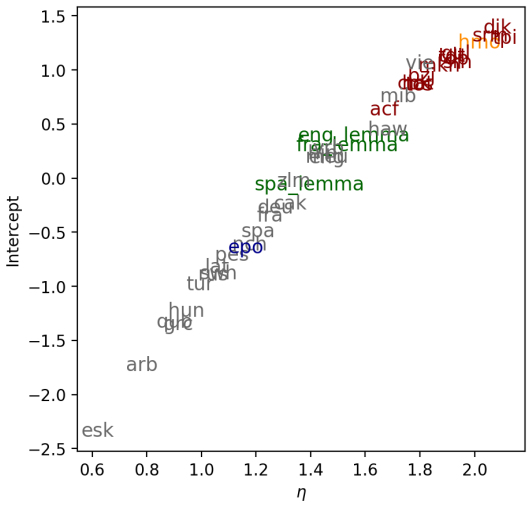

図 2　傾き η と切片の分布（単語）

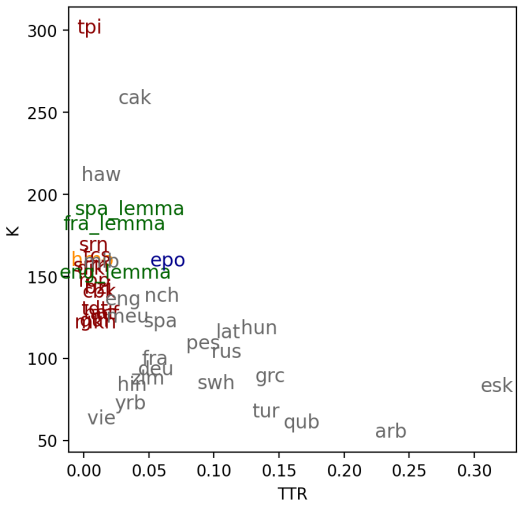

図 4　TTR と K の分布

0.01）。クレオールは単語長が小さく[9]、節長が大きくなる傾向が見られた。平均単語長 4.5、平均節長 27 付近で、クレオールと非クレオールを大別できる。単語長の短さは、各語の持つ情報が相対的に小さいことを示唆している。節長の大きさは、同一内容を述べる際に、各語の持つ情報が相対的に少ないために、多くの語を用いる必要があることを示唆している[10]。これは、クレオールの構成性 [20] や分析的性格の高さを表している可能性がある。

なお、付録の表 3 のとおり、rop を除き、上層言語に比してクレオールの平均単語長は有意に短くなった（Welch の t 検定で $p < 0.01$）。これは、上層言語の語形の単純化を示唆している[11]。反対に、全クレオールで平均節長は有意に長くなった（$p < 0.01$）。

語彙の豊富さ　図 4 は、TTR と K の分布を示している。クレオールは非クレオールよりも TTR が低くなり、0.016 付近で両者を大別できる。一方、K では両者は混在し、区別は難しい。

パープレキシティ　図 5 は、横軸が文字種類数、縦軸が文字 3-gram モデルのパープレキシティを表している。文字種類数に顕著な差がない場合でも、クレオールでは値が小さくなり、5.3 付近でクレオールと非クレオールを大別できる。2-gram でも同様の傾向が見られた。これは、クレオールの音韻配列や語構成が、より単純であることを示唆してい

9）　頻度が高い単語ほど単語長が短いという傾向（Zipf 短縮）[13, 12] が、クレオールでは僅かに強く見られた。

10）　なお、クレオールでは文数が多くなった。非クレオールが複文や重文を多用する一方で、クレオールが単文を好む傾向を表している可能性がある。

11）　例えば、多くのクレオールにおいて、子音連続や語末子音を含む単語種類数の割合が上層言語よりも有意に少なくなった（$p < 0.01$）。

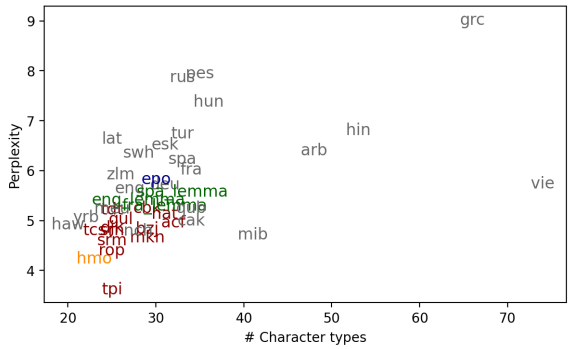

図 5　文字 3-gram モデルのパープレキシティ

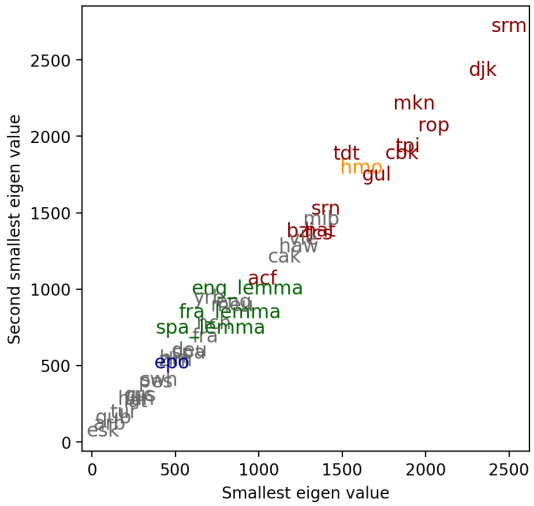

図 6　隣接行列のスペクトルの分布

る。クレオールの語彙サイズの小ささを反映して、単語レベルでは更に顕著な差異が見られた（図は非掲載）。

意味の強さの分布　付録の図 7 は、単語の出現頻度と分散表現のノルムの分布を示している。クレオールでは両者の順位相関係数 ρ が小さく、非クレオールでは大きな値となった。両者は $\rho \approx 0.3$ で大別できる。この差異は、クレオールの高頻度語のノルムが相対的に小さいことに起因する。クレオールは語彙サイズが小さいため、高頻度語の中に多義性が高いものが多く、ノルムが小さくなると推測される。平均ノルムには顕著な違いは見られなかった。

スペクトル　図 6 は、隣接行列のスペクトルのうち、小さい方から順に 2 つの固有値の分布を示している。クレオールでは値が小さくなり、いずれも 1000 付近でクレオールと非クレオールを大別できる。最小固有値が大きいほど密なグラフと言えるので、クレオールでは最頻語同士がより共起しやすいことを示唆している。

6　考察

複数の指標により、クレオールと非クレオールを連続的に位置付け、目視でも大別することができた[12)]。少なくともクレオールとその上層言語は峻別が可能である。クレオール内でも振る舞いに揺れが見られ、その性質が段階的であることが示唆された [21]。多くの指標で、djk, srm, tpi が最も非クレオールから乖離し、acf が最も近い振る舞いを示した。前者はクレオール性が高く、後者は低いと言える。興味深いことに、ピジンの hmo はクレオールから峻別できなかった。

非クレオールの中では、cak, haw, mib, vie がクレオールに近い振る舞いを見せた。このうち、haw, mib, vie は分析的性質が強い言語である。これは、クレオールの高い分析的性質 [22] を示唆していると考えられる。ただし、これらの言語よりも高いクレオール性を示すクレオールが複数存在する一方で、どのクレオールよりも高いクレオール性を示す非クレオールが一つも存在しない点は注目に値する。

上層言語をレンマ化した仮想言語は、一貫して元の言語とクレオールとの中間的な振る舞いを見せた。上層言語に比べ仮想言語は単純性が高いことから、クレオールは更に単純性が高いと解釈できる。よって、間接的に単純説が支持される。しかし、語彙サイズの小ささ（トークン数の多さ）を根源的な要因として、クレオールの指標の値が非クレオールから乖離している可能性も否定できない。

7　おわりに

本稿は、テキストデータから求めた指標の大小でクレオールと非クレオールを峻別できるか検証した。データとして、いずれの言語においても同一内容が保証されている聖書を用いた。実験の結果、両者を峻別こそできないが、複数の指標で大別することができた。これは、クレオールが非クレオールとは異なる性質を保有している可能性を示唆している。また、部分的に単純説を支持する結果を得た。

今後の課題として、(i) 語彙サイズに依存しない指標の考案、(ii) 基層語との比較対照、(iii) 類型論的差異の影響の分析、(iv) 語順等の分析に不可欠なクレオールの解析器の作成 [23]、(v) 数理モデル [24, 20] との対応付け、が挙げられる。

12)　人工言語 epo が非クレオールから大きな乖離を示さなかった点は興味深い。

謝辞

本研究の一部は、JSPS 科研費 JP23K12152 の助成を受けたものです。

参考文献

[1] Jeff Sigel. **The Emergence of Pidgin and Creole Languages**. Oxford University Press, 2008.

[2] Jean Aitchison. **Language Change: Progress or Decay?** Cambridge University Press, 4 edition, 2012.

[3] John H McWhorter. The world's simplest grammars are creole grammars. **Linguistic Typology**, Vol. 5, pp. 125–166, 2001.

[4] Mikael Parkvall. **The simplicity of creoles in a cross-linguistic perspective**, pp. 265–285. John Benjamins Publishing Company, 2008.

[5] Peter Bakker, Aymeric Daval-Markussen, Mikael Parkvall, and Ingo Plag. Creoles are typologically distinct from non-creoles. **Journal of Pidgin and Creole Languages**, Vol. 26, pp. 5–42, 2 2011.

[6] Matthew S. Dryer and Martin Haspelmath. **WALS Online (v2020.4)**. Zenodo, 2013.

[7] John Holm and Peter L. Patrick, editors. **Comparative Creole Syntax: Parallel Outlines of 18 Creole Grammars**. Battlebridge, 2007.

[8] Alexander Koplenig, Peter Meyer, Sascha Wolfer, and Carolin Müller-Spitzer. The statistical trade-off between word order and word structure – large-scale evidence for the principle of least effort. **PLoS ONE**, Vol. 12, p. e0173614, 3 2017.

[9] Christos Christodouloupoulos and Mark Steedman. A massively parallel corpus: the bible in 100 languages. **Language Resources and Evaluation**, Vol. 49, pp. 375–395, 2015.

[10] David M Eberhard, Gary F. Simons, and Charles D. Fennig, editors. **Ethnologue: Languages of the World**. SIL International, 27 edition, 2024.

[11] Tatsuru Kobayashi and Kumiko Tanaka-Ishii. Taylor's law for human linguistic sequences. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1138–1148. Association for Computational Linguistics, 6 2018.

[12] 田中久美子. 言語とフラクタル：使用の集積の中にある偶然と必然. 東京大学出版会, 2021.

[13] George K Zipf. **Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology**. Addison-Wesley Press, 1949.

[14] Steven T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. **Psychonomic Bulletin and Review**, Vol. 21, pp. 1112–1130, 10 2014.

[15] George Udny Yule. **The Statistical Study of Literary Vocabulary**. Cambridge University Press, 1944.

[16] Steven Bird, Edward Loper, and Ewan Klein. **Natural Language Processing with Python**. O'Reilly Media Inc, 2009.

[17] Momose Oyama, Sho Yokoi, and Hidetoshi Shimodaira. Norm of word embedding encodes information gain. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 2108–2130. Association for Computational Linguistics, 2023.

[18] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In **Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks**, pp. 45–50. European Language Resources Association, 2010.

[19] 佐藤竜馬. グラフニューラルネットワーク. 講談社, 2024.

[20] 加藤大地, 上田亮, 宮尾祐介. 簡素な創発言語接触モデルで生じる言語のクレオール単純性と構成性. 第 37 回人工知能学会 全国大会論文集, pp. 1–4. 人工知能学会, 2023.

[21] John Holm. **An Introduction to Pidgins and Creoles**. Cambridge University Press, 2000.

[22] Gary Lupyan and Rick Dale. Language structure is partly determined by social structure. **PLoS ONE**, Vol. 5, p. e8559, 1 2010.

[23] Heather Lent, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard. On language models for creoles. In **Proceedings of the 25th Conference on Computational Natural Language Learning**, pp. 58–71. Association for Computational Linguistics, 11 2021.

[24] 村脇有吾. クレオール形成に対する混合モデル. 言語処理学会 第 22 回年次大会 発表論文集, pp. 853–856, 2016.

A 付録

表 2 分析対象言語：最上段がピジンとクレオール、第二段が上層言語、第三段が仮想言語、最下段がそれ以外の言語である。最上段のカッコ内は上層言語を示している。

Code	Language	Family
hmo	Hiri Motu	Pidgin (meu)
acf	Lesser Antillean French Creole	Creole (fra)
bzj	Belize English Creole	Creole (eng)
cbk	Chavacano	Creole (spa)
djk	Aukan	Creole (eng)
gul	Sea Island English Creole	Creole (eng)
hat	Haitian Creole	Creole (fra)
mkn	Kupang Malay	Creole (zlm)
rop	Kriol	Creole (eng)
srm	Saramaccan	Creole (eng)
srn	Sranan Tongo	Creole (eng)
tcs	Torres Strait Creole	Creole (eng)
tdt	Tetun Dili	Creole (tet)
tpi	Tok Pisin	Creole (eng)
eng	English	Indo-European
fra	French	Indo-European
meu	Motu	Austronesian
spa	Spanish	Indo-European
zlm	Malay	Austronesian
eng_lemma	English (lemma)	Counterfactual
fra_lemma	French (lemma)	Counterfactual
spa_lemma	Spanish (lemma)	Counterfactual
arb	Arabic	Afro-Asiatic
cak	Kaqchikel	Mayan
deu	German	Indo-European
epo	Esperanto	Constructed language
esk	Inupiatun	Eskimo-Aleut
grc	Ancient Greek	Indo-European
haw	Hawaiian	Austronesian
hin	Hindi	Indo-European
hun	Hungarian	Uralic
lat	Latin	Indo-European
mib	Mixtec	Otomanguean
nch	Nahuatl	Uto-Aztecan
pes	Persian	Indo-European
qub	Quechua	Quechuan
rus	Russian	Indo-European
swh	Swahili	Niger-Congo
tur	Turkish	Turkic
vie	Vietnamese	Austro-Asiatic
yor	Yoruba	Niger-Congo

表 3 上層語に対するピジンとクレオールの平均単語長と平均節長の比率

	上層言語	平均単語長比	平均節長比
hmo	meu	**0.996**	**1.366**
acf	fra	**0.821**	**1.248**
bzj	eng	**0.859**	**1.336**
cbk	spa	**0.932**	**1.696**
djk	eng	**0.790**	**1.719**
gul	eng	**0.820**	**1.463**
hat	fra	**0.767**	**1.270**
mkn	zlm	**0.734**	**1.959**
rop	eng	1.113	**1.565**
srm	eng	**0.752**	**2.093**
srn	eng	**0.787**	**1.296**
tcs	eng	**0.907**	**1.493**
tpi	eng	**0.933**	**1.623**

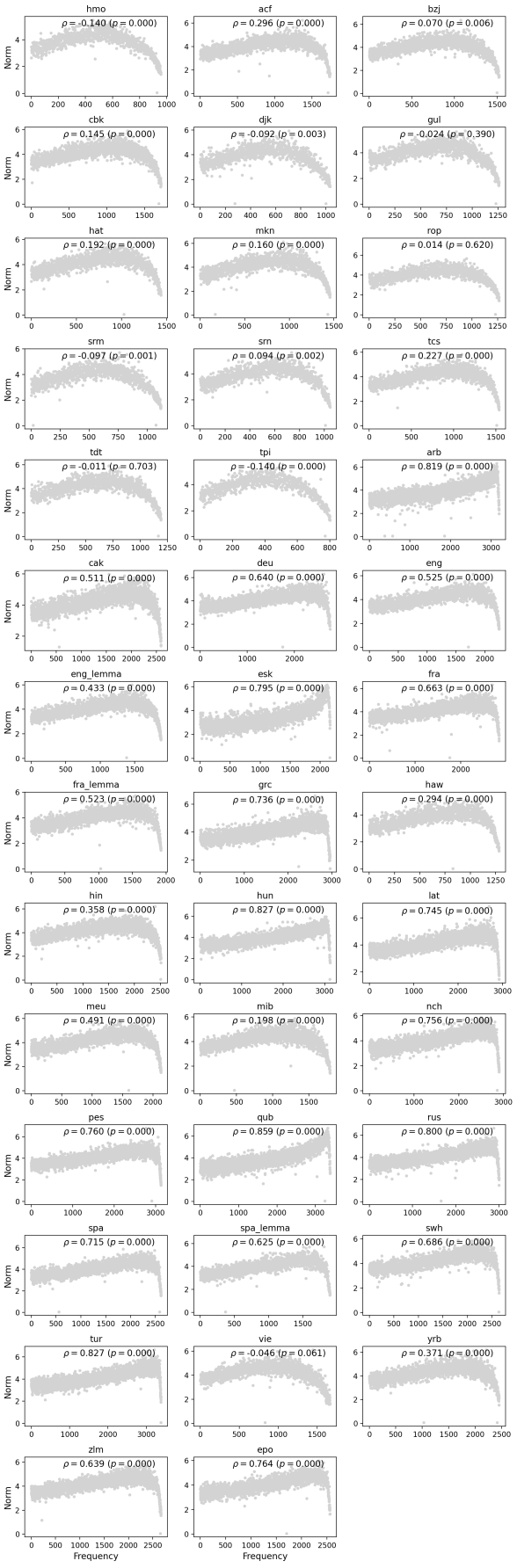

図 7 単語の出現頻度と分散表現のノルムの分布