語彙の多様性，密度，洗練性から見た語彙の豊富さ

鄭弯弯[1]
[1]名古屋大学人文学研究科附属人文知共創センター
zheng.wanwan.v6@f.mail.nagoya-u.ac.jp

概要

語彙は言語の根幹として認識されている中で，その豊かさを量的に示す課題は，計量言語学分野において依然として難題であり続けている．語彙の豊富さは，言語習得，言語変化，心理学および認知科学などの理論的および実践的な分野で幅広く適用されている．しかし，語彙は複雑であり，その豊かさは通常，語彙の多様性，密度，洗練度といった三つの側面から測定できる．その中で，語彙の多様性は広く研究されているが，テキストの長さに依存しない指標が依然として不足している．一方で，語彙の密度と洗練性の特性は十分に扱われていない．本研究は，日本語における語彙の多様性，密度，および洗練性の特徴を描き出し，それらの相互関係を探求することを目的とする．

1 はじめに

語彙は長い間，言語学習の基盤として認識されてきた．Lewis （1993）[1]は語彙を「言語の核心」と定義し，Long and Richards （2007）[2]は語彙を「すべての言語スキルの核心」と呼んでおり，斎藤（2015）[3]は「語彙力ことが教養力である」と主張している．一般的に，語彙が多いほど物事を正確に理解でき，また適切な言葉遣いを使い分けて豊かに表現できるとされている．そのため，語彙の豊富さ（lexical richness）を量的に評価することは，教育学（言語の発達指標），計量言語学（著者識別，文章分類），心理学と認知科学（認知機能や情報処理においての役割）などのさまざまな学問領域や実用的な応用において重要であり，昔から幅広く研究されてきた．

語彙の豊富さ指標の研究において，Type-Token-Ratio（TTR, Templin, 1957[4]）をはじめ，文書の長さに依存しない指標の構築が求められている．多くの指標が提案されたが，完全に文章の長さの影響を受けない指標はまだ存在しないため，これらの指標の有効性に疑念が抱かれている．また，主に EFL（English as a Foreign Language）学習者の言語発達指標として研究されているため，日本語における研究例は限られている．本研究は，日本語を対象とする．

語彙の豊富さは, 語彙の多様性（lexical diversity），語彙の密度（lexical density）と語彙の洗練性（lexical sophistication）によって反映されている（Lei & Yang, 2020[5]）．その中，語彙の多様性は語彙の豊富さの最も明白な尺度であると言われており，現段階語彙の豊富さの計測はほとんど語彙の多様性を測っている．しかし，語彙の多様性は，同じ単語であるかどうかしか見られず，どのような語彙が使用されているかという点を全く考慮しないことは問題として問われている．さらに，語彙の多様性指標は文書の長さに依存する課題は以前から存在しており，解決できていないままである．語彙の密度は，情報の詰め込みの程度を表し，Ure（1971）[6]は「発表語彙の発達した学習者は，一定量のテキストでより多くの情報を伝達することができるため，語彙の密度は高くなると考えられている」と述べている．しかし同時に，語彙の密度は，統語構造や結束性の影響を受け，必ずしも語彙の豊かさを測っているとは言えないと指摘されている（Read, 2000[7]）．語彙の洗練性は，産出する語彙がいかに低頻度語を含み，洗練されたものであるかを量的に示す指標である．発表語彙の発達した学習者は，基本語以外の技術的・専門的用語も使用することができ，場面に応じてより適切に語彙を使用することができるため，より低頻度で洗練された語彙を産出すると考えられる ．Lei and Yang （2020）[5]は，語彙の豊富さは，語彙の難易度を考慮すべきと述べている．語彙の洗練性は多面的であり，難易度の高いは簡単に決めることはできないが，歴史的に頻度に基づいて研究が発展してきた．そこで，頻度リストの作成に使用したコーパスが語彙の難易度を左右する問題が存在する．また，複数の頻度レベルに対して複数の値が与えられ

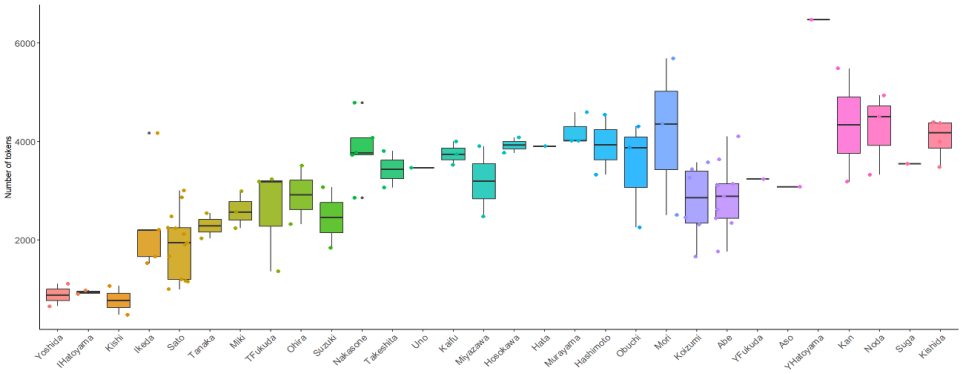


図 1 総理ごとの各演説の延べ語数

るため，語彙が豊富であるかどうか直観的にわかりにくい．さらに，近年は，頻度という観点のみでは定義することは難しいと議論されている（Kyle & CrossLey, 2015[8]）．

そこで，現在の語彙の豊富さ指標の妥当性は十分に検討されていないまま使用されているという実態に対して，本研究は，日本語における語彙の多様性，密度，および洗練性の独自の特徴を描き出し，語彙の豊かさ測定の精度向上の基盤を築くことである．具体的に以下の三つの問いに焦点を当てる．

1) 内容語と機能語は，言語要素の二つの異なるカテゴリである．内容語は open class として，新しい語が加え続けされ，機能語は closed class として語が限られる．そこで，テキストの長さが増加すると，内容語の比率が上昇し，それにより語彙の密度が上昇する傾向があるか？
2) 理論的な仮定は一貫した語彙の洗練度を示唆しているが，テキストの長さに依存するかはさらなる検証が必要とされる．実際のテキストを使用する際の語彙の洗練度はどれほど一定であるか？
3) 語彙の多様性，密度，および洗練度はどのように互いに影響し合うか？

2 使用データ

日本の歴代総理大臣の所信表明演説のテキストデータ（期間：1953～2023）91 件を使用した．そのうち，1953 年から 2014 年までの 82 件は公開されている「所信表明演説コーパス」[i] であり，2015 年から 2023 年までの演説は，首相官邸ホームページで検索して作成した．所信表明演説を用いた理由は主に二つある：1) 語彙は異なるテーマ（例えば，映画，文学，科学など）に基づいて導入される原則に基づている．所信表明演説は特定のテーマに従い政権の運営方針や取り組む重点課題などに関する説明であるため，テーマが共通している；2) 個々の人の語彙の豊かさは，年齢，けが，健康状態などの要因によって異なる可能性がある（Shi & Lei, 2021[9]）．この観点から，これらの演説を行った総理たちは安定期の語彙の豊富さにあると考えられ，かつ演説の発表年度が近いため，語彙の豊富さは一定であったと仮定するのは合理的である．

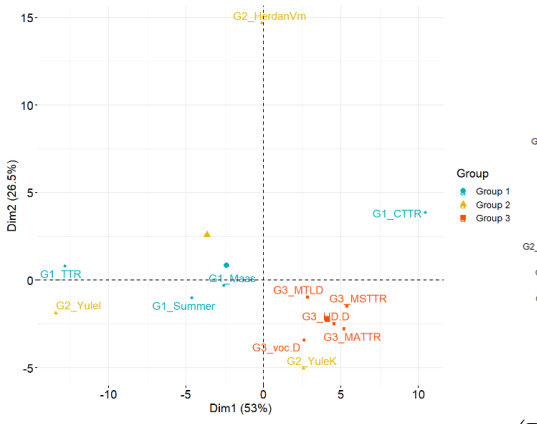
91 件の演説に対して，Juman++を用いて形態素解析を行った．前処理として，（拍手），（ひらがな）などは，補足情報として付け加えたものであるため，削除した．記号は語彙でないため，数えないことにした．すべての単語を普通形に直した．また，演説の中に，二〇二〇年度，五か月，三十億円，第二次補正予算案など数値が含まれるため，数字を#にして，数字だけの形態素を数えないことにした．例えば，「二〇二〇年度」は「#年度」，「五か月」は「#か月」になるが，「第二次補正予算案」は「第＜接頭辞＞次＜接尾辞＞補正＜名詞＞予算＜名詞＞案＜名詞＞」になる．

図 1 では，総理ごとの延べ語数が示されている．演説の数に関して，トップ 3 は佐藤氏が 13 件，安倍氏が 9 件，小泉氏が 6 件であった．総語数に関しては，トップ 3 は安倍氏が 26,028 語，佐藤氏が 24,986 語，中曽根氏が 19,220 語であった．スピーチの平均長に関しては，トップ 3 は鳩山氏が 6,478 語，菅氏が 4,332 語，野田氏が 4,259 語であった．

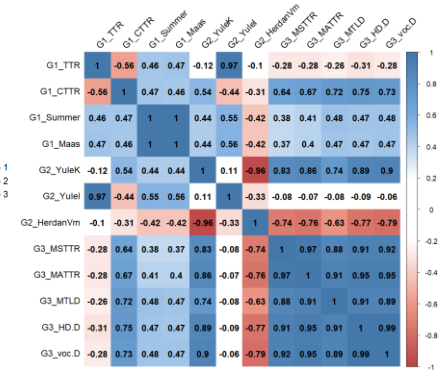
3 分析

3.1 語彙の多様性，密度と洗練性の特徴

[i] https://github.com/yuukimiyo/GeneralPolicySpeechOfPrimeMinisterOfJapan



（A）相関行列を用いた PCA



（B）ピアソン相関を用いたヒートマップ

図 2　語彙の多様性指標の類似性

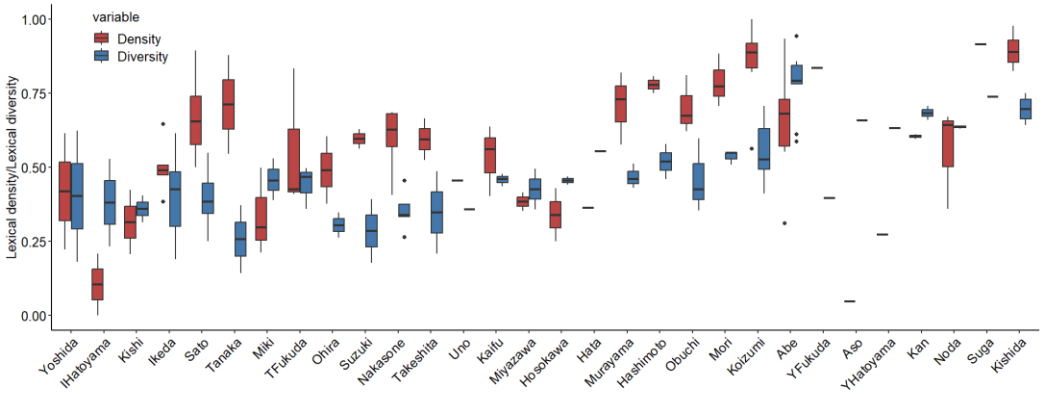


図 3　総理ごとの語彙の多様性と密度

3.1.1 語彙の多様性

各演説の語彙の多様性を，TTR（ベースライン），CTTR, Summer, Maas（異なり語数に基づく指標），YuleK, YuleI, HerdanVM（出現頻度スペクトルに基づく指標），MSTTR, MATTR, MTLD, HD-D, and voc-D（統計的処理に基づく指標）といった 12 の指標を用いて計算した．その結果は主成分分析（PCA）およびヒートマップを用いて視覚化され，図 2 に示した．また，すべての指標に対して最小最大正規化を行った．YuleK および Maas の値が多様性の度合いを示す際，値が小さいほど多様性が高まることを示唆しているため，これらの値は 1 - (YuleK/Maas) に変換され，他の尺度との整合性が図られた．

主成分分析は 12 の語彙の多様性指標から得られる指標内のパターンと指標間の関係をより明確に理解することができる．ヒートマップは類似性行列の視覚的表現を提供し，12 の指標はどのようにクラスタ化または分岐しているかに関する洞察を提供する．その結果，グループ 3 の指標はより高い相関水準を示し，一貫した結果をもたらした．さらに，グループ 2 の YuleK はグループ 3 と類似した．また，HerdanVM は他の尺度と比較して異なる結果を示し（すべてのピアソン係数が負であった），特徴的な尺度として特定できた．相対的に，TTR と YuleI, Summer と Maas の相関はそれぞれ 0.97, 1.00 に達し，高い類似度が示された．

3.1.2 語彙の密度

上記の結果により，特定の語彙の多様性指標が一貫して有効であると断言することは難しいため，各グループを代表する指標を選出し，平均化して語彙の多様性を示すアンサンブルの多様性値を算出した．グループ 1 では，TTR と CTTR が不安定であり，Summer は Maas と高い相関があったため，Maas を選択した．グループ 2 では，HerdanVM が他の尺度と異なる結果を示し，YuleI が TTR と高い相関を示したため，YuleK を選択した．グループ 3 では，MTLD が一貫性のある結果を示し，先行研究で広く推奨されているため，選択された．

表 1　語彙の洗練性における語彙の多様性と密度の比較（安倍氏と小泉氏のデータを例として）

	Average ratio (%)			Average lexical diversity			Average lexical density		
	Abe	Koizumi	*p*-value	Abe	Koizumi	*p*-value	Abe	Koizumi	*p*-value
Lower elementary	**11.3517**	10.1026	0.0419	**0.5920**	0.5020	0.2984	0.7661	**0.8307**	0.0512
Upper elementary	**13.8114**	9.6938	0.0003	**0.6700**	0.4690	0.0495	**0.5844**	0.5744	0.7366
Lower intermediate	**21.0181**	20.6239	0.5628	0.4696	**0.4779**	0.9347	0.7328	**0.7778**	0.0429
Upper intermediate	41.7066	**45.515**	0.0020	**0.5400**	0.5045	0.7277	0.8699	**0.8896**	0.0589
Lower advanced	11.1491	**12.5592**	0.0336	0.4261	**0.5426**	0.1631	0.9686	**0.9691**	0.9068
Upper advanced	0.9630	**1.5055**	0.0969	0.3998	**0.4699**	0.3110	0.5674	**0.7127**	0.0115

図 3 は，総理ごとの語彙密度と多様性を示している．個々の間で密度と多様性の変動が見られます．例えば，田中氏は著しい密度を示していたが，同時に最も低い多様性であった．対照的に，麻生氏は最も高い多様性を示しているが，その密度は最も低かった．高い密度は内容語の大量利用を示唆し，低い多様性は単語の反復を意味する．この反復が内容語に関連しているかどうかは更なる分析が必要である．

さらに，語彙密度と多様性の間に一貫した関係がないことがわかった．例えば，HerdanVM を除くすべての指標が示すところによれば，安倍氏の語彙多様性は小泉氏を上回った．しかし，語彙密度は安倍氏と小泉氏がそれぞれ 0.6635 および 0.8478 であり，小泉氏の方が語彙の密度が高い事例が存在した．

3.1.3 語彙の洗練性

この節では，安倍氏と小泉氏のデータを用いて，「日本語教育語彙表」（Sunakawa et al., 2012[10]）による初級前半，初級後半，中級前半，中級後半，上級前半，上級後半といった 6 つの洗練度レベルにわたり，語彙の多様性，密度，および洗練性に対する包括的な検討を行う．結果は表 1 に示す．

共通点としては，安倍氏と小泉氏は初中級レベルの語彙を主に使用していることがわかった．しかし，安倍氏は初級および中級前半の単語の頻出が見られる一方，小泉氏は中級後半から上級の単語をより広範に組み込んでいた．さらに，安倍氏と小泉氏の上級前半および上級後半での語彙多様性の平均値はそれぞれ 0.4261 および 0.5426，0.3998 および 0.4699 であり，安倍氏の語彙多様性は初級レベルであることがわかった．

言語は量だけでなく，質と複雑さも重要とされている．複雑な言葉を理解し使用することで，より正確で具体的な表現が可能となり，単語一つで複雑な概念や感情を伝える能力は，より豊かで深いコミュニケーションを可能にする．さらに，語彙多様性の指標はテキストの長さに依存しないのが多くの研究者が疑問を呈している．したがって，安倍氏の優れた語彙の多様性は，演説が最も長いことに起因する可能性が考えられる．全体的には，安倍氏の語彙がより豊かであると結論することは妥当でない．

まとめ

本研究は，語彙多様性，密度，および洗練度の本質的な特性を探求し，その相互関係とテキスト長の影響に関する安定性（付録を参照）を検証した．実験の結果から得られた結果は以下の通りである．

1) 語彙多様性の尺度を評価する際には，単に達成された収束率に焦点を当てるのではなく，特定のコーパスに対する適切なサンプルサイズを考慮することが不可欠である．
2) 語彙の密度と洗練性はテキストの長さの影響を受けていない．
3) 語彙多様性，密度，および洗練性の間には一貫した関係がない．

語彙の豊富さを評価するためには，語彙多様性，密度，および洗練度の三つの尺度が利用可能であるが，通常，各尺度は独立して適用されている．近年では，語彙の洗練性が語彙の豊富さを評価する際の重要な要素として注目されている．また，内容語が重要な意味を伝えることを考慮すると，語彙の密度を完全に語彙の豊富さの指標から排除するのは適切でないと考える．さらに，ほとんどの尺度について、最終的なスコアをどのように解釈すべきかが不明確である．そこで，今後の研究として，複数の視点から語彙豊かさを測定することと，最終的なスコアの違いの解釈可能性を提供することを挙げる．

謝辞

本研究は JSPS 課題設定による先導的人文学・社会科学研究推進事業学術知共創プログラム「人間・社会・自然の来歴と未来：「人新世」における人間性の根本を問う」（JPJS00122674991）の委託を受けたものです.

参考文献

[1] Lewis, M. **The lexical approach: The state of ELT and the way forward**. Hove, England: Language Teaching Publications, 1993.

[2] Long, M., Richards, J. **Modelling and assessing vocabulary knowledge** (pp. xii–xiii). In H. Daller, J. Milton & J. Treffers–Daller (Eds.). Cambridge University Press, 2007.

[3] 齋藤孝. **語彙力こそが教養である**. 角川新書, 2015.

[4] Templin, M.C. **Certain language skills in children: Their development and interrelationships**. pp. 1–212, The University of Minnesota Press, 1957.

[5] Lei, S., Yang, R.Y. Lexical richness in research articles: Corpus-based comparative study among advanced Chinese learners of English, English native beginner students and experts. J**ournal of English for Academic Purposes**, Vol. 47, 2020. https://doi.org/10.1016/j.jeap.2020.100894

[6] Ure, J. Lexical density and register differentiation. **Contemporary Educational Psychology**, Vol. 5, pp. 96–104, 1971.

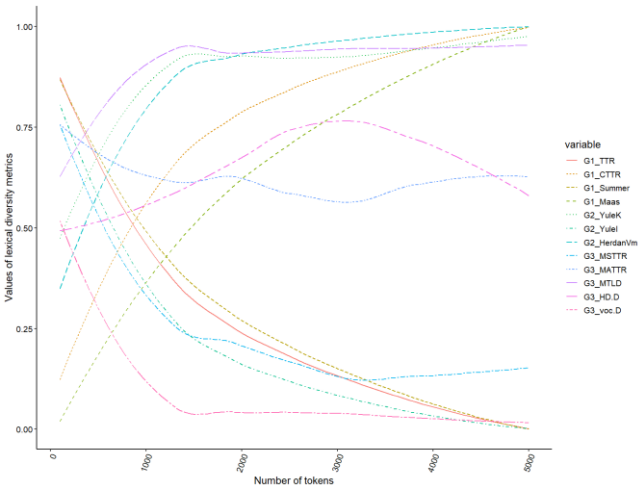
[7] Read, J. **Assing vocabulary**. Cambridge: Cambridge University Press, 2000.

[8] Kyle, K., Crossley, S.A. Automatically assessing lexical sophistication: Indices, tools, findings, and application. **TESOL Q**, Vol. 49, pp. 757–786, 2015. https://doi.org/10.1002/tesq.194

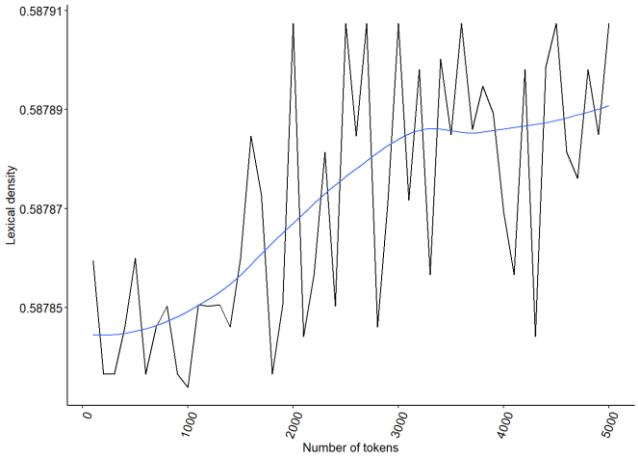
[9] Shi, Y.Q., Lei, L. Lexical use and social class: A study on lexical richness, word length, and word class in spoken English. **Lingua**, Vol. 262, 2021. https://doi.org/10.1016/j.lingua.2021.103155

[10] Sunakawa, Y., Lee, J.H., Takahara, M. The construction of a database to support the compilation of Japanese learners dictionaries. **Acta Linguistica Asiatica**, Vol. 2, No. 2, pp. 97–115, 2012.

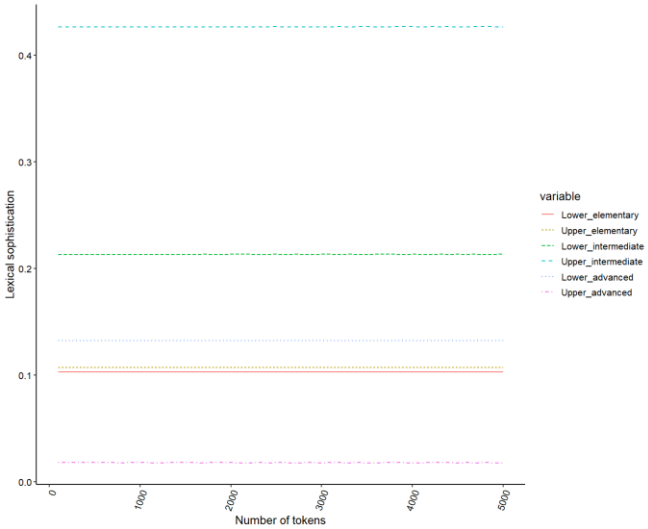
A 付録



（A）語彙の多様性指標とテキストの長さの関係



（B）語彙の密度とテキストの長さの関係



（C）語彙の洗練性とテキストの長さの関係