N-039

ジャンルごとの語彙多様性を再現するために必要なテキストサイズの推定

Estimating the Required Text Length for Reproducing Genre-Specific Lexical Diversity

鄭　齊齊†
Wanwan Zheng

1. はじめに

語彙の多様性を測定する各種指標は、テキストの長さによって値が左右されるため、一世紀前からそれらの安定性に関する比較評価が行われてきた。一般的な研究手法としては、異なる長さのテキストに対して複数の指標を用いて語彙の多様性を算出し、各指標値がどの程度の速さで収束するか（例えば、サンプルサイズによって生じる変動の大きさ）を比較するアプローチが採用されている。分析対象となるテキスト長は、最小 50 語から 1000 語程度まで幅広く設定されることが多い。例えば、Nan Bernstein et al., (2024)による子どもの表現語彙力を評価した研究では、voc-D (Mckee et al., 2000)が最も高い性能を示した。voc-D は、35 語から 50 語までのテキストをランダムサンプリングし、それぞれのサンプルの TTR をもとに形成された TTR 曲線の減衰特性に基づいて計算される指標である。Bestgen (2024)は、既存指標のテキスト長依存性を検証する際に、最小 60 語のテキスト長を設定している。Kyle & Eguchi (2023)は、50 語単位の MAATR (Covington & McFall, 2010) を用いて語彙の多様性を算出している。しかしながら、語彙の多様性指標に関する安定性の問題は単に計算にとどまらない。仮にある指標が 50 語程度で安定した値を示したとしても、その指標が語彙の多様性を適切に捉えていると即断することの妥当性については再考の余地があり、この点が本研究の出発点となる根本的な問題意識である。

著者識別の分野では、使用されるテキストの長さが不十分である場合、著者特有の文体や語彙的特徴が十分に現れにくいことが、これまで数多くの研究に指摘されてきた。この課題は、語彙の多様性を個人やジャンルの特性として捉える際にも共通しており、語彙的特徴を適切に把握するためには、一定の単語数（テキストサイズ）が必要であることが示唆される。「指標が安定になったかどうか」を検討する前に、「どの程度の語数から、個人やジャンルに固有の語彙的特徴が再現性をもって安定的に捉えられるか」という問いは、極めて重要であると考えられる。一方で、もともとばらつきの大きいデータに対して安定した値を得て解釈するのは特に誤った結論を導く危険性がある。

したがって本研究では、「ジャンルに固有の語彙多様性を安定して再現するために、どの程度のテキストサイズが必要である」という仮説を立て、日本語コーパスに含まれる複数のジャンル（演説、自然会話、小説、ニュース）を対象として、ジャンルごとに語彙多様性を安定的に再現するために必要なテキストサイズの推定を試みた。

2. データと前処理

2.1 テキストデータ

テキストデータとして、日本語の政治演説、自然会話、ニュース記事、小説といった 4 つの異なるジャンルから構成されるコーパスを使用した。これらのジャンルは、それぞれ異なる言語スタイルや使用パターンを有すると同時に、一定の共通性も備えていることから、比較対象として選定された。

政治演説データは、1953 年から 2024 年にかけて日本の歴代首相が国会で行った 93 の演説が使用された。これらの演説には、形式的な言語が用いられ、説得や動機付けを目的とした修辞的かつ抽象的な表現が多用される。語彙の選択には慎重さが求められ、複数のトピックに対応しながら聴衆への影響を意図した修辞的技巧を反映するため、語彙の多様性は一般的に高い傾向がある。

自然会話データは、「BTSJ 日本語一千人会話コーパス」（Usami, 2023）を使用した。本データセットに含まれるテキストは、語彙の計画性が比較的に低く、カジュアルな言語スタイルを特徴とする。会話には、相槌、フィードバック、共感表現といったインタラクティブな要素が本質的に含まれており、話し手は相手の即時的な反応を期待して発話する。そのため、語彙の多様性は比較的低い。

ニュース記事データは、2012 年 9 月に「livedoor News」から収集されたテキストデータを使用した。収録された記事は、時事、スポーツ、テクノロジー、家電、エンタメ、ライフスタイル、文化など、多様なテーマを網羅している。ニューステキストは、正確性、中立性、客観性を重視した語彙使用が特徴とし、情報を効果的かつ簡潔に伝達することを目的としている。そのため、精緻で多様な語彙の使用が求められ、語彙の多様性は中程度から高程度であると考えられる。

小説データは、1964 年から 2019 年の間に発表された 563 作品の近現代日本文学が使用された（李・金，2022）。本データは、各時代を代表すると考えられる受賞作を中心に選定されたデータである。小説は、作者の文体に依存する語彙の多様性が高く、感情的、哲学的、主観的な言語表現に富むのが特徴である。小説の主な目的は、感情的・思想的な表現にあり、より中立的・客観的な情報を伝達を重視するテキストとは明確に一線を画している。

これら 4 つのジャンルは、話し言葉と書き言葉の言語的特性の違いを明確に示している。政治演説と自然会話は、対話性や説得性といった話し言葉に共通する特徴を有し、ニュース記事と小説は、物語的構造や多様な主題表現といった書き言葉に特有の性質を備えている。

2.2 前処理

前処理の段階では、「拍手」や「3 秒間の沈黙」、「軽い笑い声」などの補足情報がすべて除去された。また、句読点などの記号類は語彙項目とは見なされず、すべて削除した。各テキストは形態素解析器 Juman++を用いて処理され、単語が抽出された。

日本語語彙に焦点を当てるため、形態素解析の後に数字やアルファベットを含む単語は“#”に置換され、連続す

†名古屋大学大学院人文学研究科　Graduate School of Humanities, Nagoya University

る“#”は 1 つに統合された。さらに、形式的な表現はすべて原形にした。くわえて、タイトル、発行日、URL、著者情報など、主たる本文以外のメタ情報はすべて削除され、分析対象として本文のみに限定した。

前処理後のコーパスの概要は表 1 に示す。近年の日本語の小型辞書には約 6 万から 7 万語の見出し語が収録されており、日本人成人が理解可能な語彙数は約 5 万語と推定されている。以上を踏まえると、本研究に用いたコーパスは、語彙使用の安定性を分析するために十分な語彙規模および主題的多様性を備えていると判断できる。

表 1　コーパスの概要

ジャンル	ファイル数	延べ語数	異なり語数
政治演説	93	279,836	8,831
自然会話	514	1,116,117	25,519
ニュース記事	7,367	3,572,149	59,388
小説	563	5,067,175	86,013

2.3 サンプルテキストの抽出

各ジャンル内のすべてのテキストを統合した上で、ランダムにシャッフルを行った。このシャッフル処理により、語彙使用における文脈的な偏りが平均化され、各ジャンルにおける全体的な語彙分布に着目した分析が可能となった。

また、同一の語が文脈に応じて異なる品詞で使用される点を考慮し、語彙の文法的多様性に対応するため、品詞情報を区別して語をカウントするアプローチを採用した。具体的には、シャッフル後のテキストにおいて、各トークンは「単語＋品詞（POS）」の組み合わせとして集計された。

異なるテキスト長にわたる包括的な分析を行うために、以下の 2 種類のサンプリング手法を実施した：

- **方法 1：移動ウィンドウサンプリング**

データセット全体を移動ウィンドウ方式で分割し、サンプルの長さを $\boldsymbol{L}$=[100, 200, 300, …, 10000, 10100]のように段階的に変化させながら、重複のないテキストサンプルを作成した。この方法により、データの有効活用とサンプルの多様性の確保が図られた。ただし、ウィンドウ長によって得られるサンプル数は異なり、たとえば 100 語のサンプル数は 10100 語のサンプル数のおよそ 101 倍に達する。

- **方法 2：ランダムサンプリング**

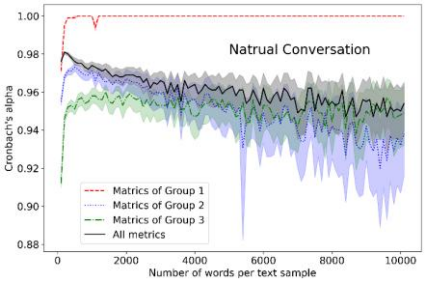
この手法では、同様に $\boldsymbol{L}$=[100, 200, 300, …, 10000, 10100]の各長さに対して、ランダムに 1000 サンプルを抽出した。これにより、すべてのテキスト長においてサンプル数を均等に保つことができ、バランスの取れた比較が可能となった。

3. 実験

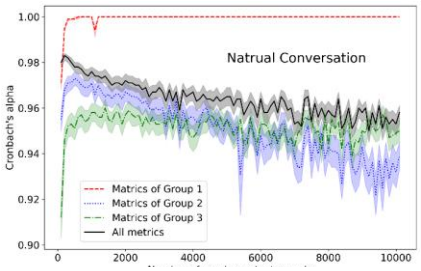
語彙の多様性指標は、さまざまな概念的枠組みと数学的手法に基づいて提案されてきた。本研究では、既存の指標を以下の三つのカテゴリに分類した：①タイプベースの指標（Group1：TTR、Summer、Maas）、②語彙分布ベースの指標（Group2：YuleK、YuleI、HerdanVm）、③統計処理に基づく指標（Group3：MSTTR、MTLD、MATTR、HD-D、voc-D）。

3.1 指標の一致度に基づくテキストサイズの推定

クロンバックのα係数は、テキストやアンケートなどにおける内的一貫性を評価するために広く用いられる信頼性指標である。本研究では、複数の語彙の多様性指標が同じ概念（語彙の多様性）を測定しているかどうかを判断する目的で、クロンバックのα係数を使用した。図 1 は移動ウィンドウサンプリングとランダムサンプリングを用いた自然会話の結果である。



(B)　移動ウィンドウサンプリングの結果



(A)　ランダムサンプリングの結果

図 1　異なるテキストサンプルサイズにおける語彙の多様性指標のクロンバックのα係数

移動ウィンドウサンプリングの場合、Group2 の指標の信頼区間が広かったが、ランダムサンプリングの場合には信頼空間が縮小し、テキストサンプル数の影響がある程度抑制された。また、サンプルテキストが 500 語程度までの範囲では、いずれのグループにおいても各指標の一貫性が低く、不安定であった。500 語から 1000 語の間では、指標値の変動幅が緩やかになり始め、1000 語から 2000 語の範囲で徐々に安定性が確保され、2000 語以降では大きな変化がなかった。Group2 の指標は語彙の集中度を測定対象としているため、テキスト長に最も影響され、テキストが長くなると、クロンバックのα係数が低下する傾向が確認された。こちらの結果は、指標によって異なる特性が示されていることを示唆しており、Group2 の指標を使用する際には、テキストの長さに応じた慎重な選択と解釈が求められる。

4. おわりに

本研究は、日本語コーパスに含まれる 4 つのジャンルを対象として、ジャンルごとに語彙多様性を安定的に再現するために必要なテキストサイズの推定を試みた。その結果、語彙多様性の傾向がジャンルによって異なることが確認され、適切な最小テキストサイズは指標およびジャンルに依存することが明らかとなった。数値実験の詳細、およびジャンル別に求められる目安値とその評価方法については、報告にて詳述する。

参考文献

[1]　Bestgen, Y. (2024). “Measuring Lexical Diversity in Texts: The Twofold Length Problem”, Language Learning, Vol.74, No.3, pp. 638–671.