

N_033

複雑ネットワークの手法を用いた文章の解析

The Investigation of the internal structure of expression in a literature by using a Complex networks method

望月 朝香†
Tomoka Mochizuki

鈴木 泰博†
Yasuhiro Suzuki

1. 概要

文章には著者の書き方により醸し出される「雰囲気」がある。この雰囲気の特定方法としては文章をどのような視点から解析するかにより様々な方法があるが、本実験では、文章中に使用される単語のみに着目し、そこから雰囲気分析を行った。単語のみに着目した理由は、著者の背景知識により語彙の豊富さが異なり、また愛用語や言い回しに特徴が出るが、それが単語の使用頻度に現れて来るからである。

また文章の内容に捕われず、単語自体から喚起される雰囲気から文章全体へ及ぼす雰囲気を考える。これは似た内容の文章を書いた事なる著者が存在した場合に、両者の文章の違いを比較する際に役に立つ。加えて単語と単語の連想的繋がりを見る事で著者の思考パターンを明らかにする事が出来る。

2. 対象とした文章

「The Little Prince」は多くの人に読まれている。しかし、内容が何を伝えようとしていたのかが分からず漠然としか覚えていないという意見が多数きかれる。そこで本実験ではこの作品に対して実験を行い、雰囲気の特定を試みた。

3. 実験

3.1 単語抽出方法

単語抽出は、次の手順で行われる。

手順1 文章中に使用されている単語を全て切り出す。

手順2 切り出した各単語の使用頻度をとりヒストグラムを作成する。使用頻度は、文章の始めから終わり迄を一括したもの（以下、「作品全体」とする。）と各章毎にとる。

本研究では以上から得られた結果より、対象文章の醸し出す雰囲気について考察を行う。

3.2 ネットワークの構築

語彙知識を用いて単語や文書の類似性や関連性を判断するため、カーネギーメロン大(米)の英語版類語辞典シソーラス Lexical FreeNet [2.3]を利用した。この辞典では語や概念間の様々な関係を調べる事が出来るが、本研究では2つの語間の「類似語」と「連想語」を介した関係を用いてネットワークを構築した(Fig 1)。

ネットワーク構築に際しては、3.1で示した手順1と手順2で得られたヒストグラムから使用頻度の高い名詞を抽出し、それらの名詞についてシソーラスを用い単語間の距離を測った。例えば、night と earth (Fig1)の場合、両者の同義語、連想語を介した距離は4となる。

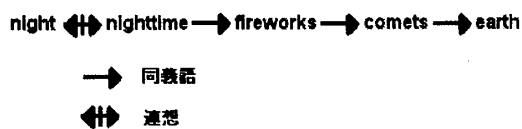


Fig1 単語間の距離

各名詞に対して以上の作業を行い距離のあるネットワークを構築する。今回は、27章(最終章)の特定の単語についてネットワークの構築を試みた。その理由として、他の章と比較して27章では意味が掴み難い部分が多いためである。

4. 実験結果と考察

単語の使用頻度については、全体を通して高い頻度で否定語と疑問詞が使用されていることを確認した(Fig2, Fig3)。最も多く用いられていた否定語は not で160回、以下 but 152回、no 75回であった。

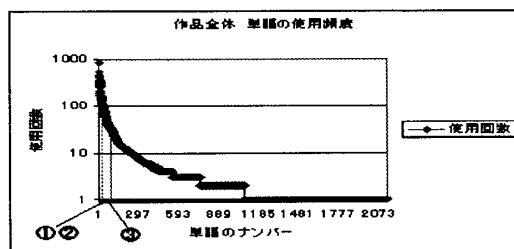


Fig2 否定形の使用頻度 図中の①は what, ②は who, ③は why にそれぞれ対応している (対数スケール)

否定語は負のイメージがあるが、使用頻度が高い事で作品全体に漂うやや暗い雰囲気と関係性があると考えられる。また、疑問詞については what の使用頻度がもっとも高く87回、以下 who 51回、how 29回、why 19回であった(Fig 3)。質問や疑問の多い作品は殆どが子供の登場する物語であるが、子供らしい可愛らしさと共に質問等があまりにも多いと戸惑いや不穏な雰囲気を纏う。疑問詞の多用に戸惑いや不穏な雰囲気の醸成に関係していると思われる。

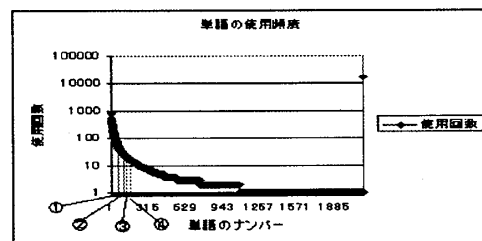


Fig3 疑問詞の使用頻度 図中で①は what, ②は who, ③は how, ④は why にそれぞれ対応している。

また他に使用頻度が高かった語として、planet と earth がある(Fig4)。一般に物語では現実世界では不可能な、時と場所との瞬時移動が可能であり読者はそこに心を惹かれるが、「The Little Prince」では王子さまは小惑星間と地球とを瞬時に移動可能となっている。それが、planet と earth の使用頻度を高めている要因と考えられる。読者はこの2つの単語から何度も王子さまの訪れた惑星と地球を思い浮かべ思い出しながら、空想の中で各星々の移動を繰り返す事になる。言い換えれば手軽な宇宙空間の旅であり、作品の不思議さ、可愛らしさといった雰囲気形成する要因の一つである。これは前述した否定形や疑問詞など負の雰囲気を作り出す単語とは相反する。

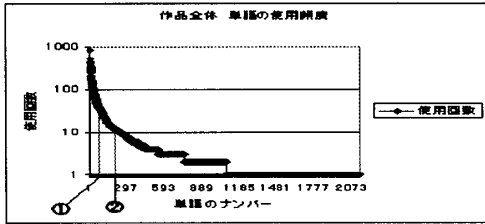


Fig4 planet と earth の使用頻度 図中で①は planet, ②は earth にそれぞれ対応している。

次に、作品全体と 27 章の単語使用頻度を比較してみると、作品全体(Fig 5 上)の場合と比べ、27 章では(Fig. 5 下)、一度しか使用しない単語数(図中囲み部)が多い。27 章で使用されている名詞は作品全体で使用頻度の高い名詞が多く、作者が作品全体を通して特徴的な単語を用い語の相互の結び付きを用いることで最終章としての雰囲気を醸成している可能性がある。そこで、会話者である prince と narrator (pilot) を省いた名詞のうち、作品全体から見た使用頻度の高い単語と照らし合わせて抽出した 8 つの単語 sheep、flower、night、planet、stars、desert、earth、rose を使用し語間のネットワークを構築した。

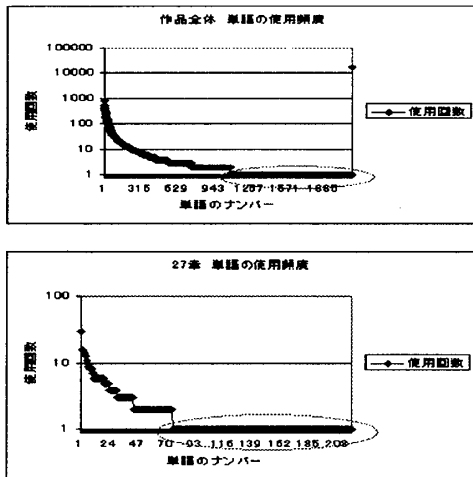


Fig5 作品全体と 27 章の単語使用頻度 上図は作品全体の単語使用頻度、下図は 27 章。囲み部は使用頻度が 1 の語をそれぞれ示している。

† 名古屋大学大学院

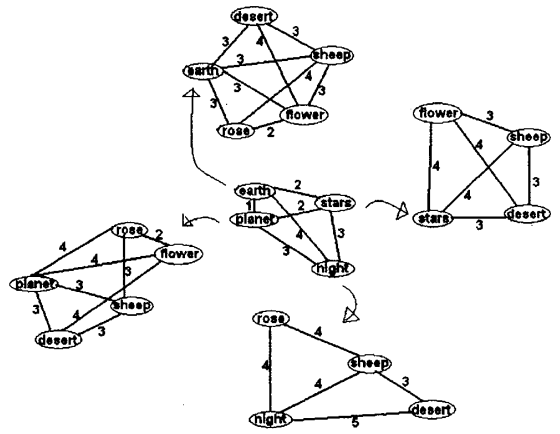


Fig6 27 章の 8 単語によるネットワーク

ネットワークの構築にあたっては、まず sheep、flower、night、planet、stars、desert、earth、rose のすべてに共通する関連語 night、planet、stars、earth を抽出し、まずこの 4 つの単語についてネットワークを作成した(以下「中心」、Fig6 の中心にあるネットワーク)。この 4 つの単語からなるネットワークでは単語間の平均距離は 2.8 であった。次にこの 4 つの単語と他の語との関係をそれぞれ調べてネットワークを作成した (Fig6、中心部の周りの 4 つのネットワーク)。作成したネットワークの単語間の距離の平均は、4.0, 3.3, 3.1, 3.5 (Fig 6 の上から時計回りに)であり、4 つの単語からなるネットワークに比べて単語間の距離が大きくなっていることを確認した。また planet と earth と関連する語間の距離は比較的長く、作者は作品の雰囲気を作り出す planet と earth からやや連想し難い意外性のある単語を多く使用する傾向がある。それが本作品に単調でない奇抜な雰囲気を醸している要因の一つと考えられる。本研究は準備的なものであり、データ解析には厳密さが欠ける部分も多い。今後はコントロールデータ(新聞や他のジャンルの書籍など)を用いて明確に比較を行いたい。また、全ての章においてネットワークを構築し、そこから雰囲気の特定を試みたいと考えている。

参考文献

- [1]山崎庸一郎 著、『星の王さまの秘密』(彌生書房 1984 年 8 月 15 日)
- [2]金田知子 岩崎晴美 齊藤兆古 宮沢賢治 堀井清之「文学作品における文体構造の可視化 宮沢賢治「銀河鉄道の夜」の解析」
- [3]岩崎晴美 齊藤兆古 宮沢賢治 堀井清之 金田知子「文学作品の線形空間論による解析」(『法政大学計算科学研究科センター研究報告』11 巻 1998 年)