文の長さとその文に付与する読点の数の相関性について

坂野 維仁 23B30434
東京工業大学情報理工学院

1. はじめに

本調査は、文の長さと、その文に付与される読点の数に相関はあるといえるのかを明らかにするために行った。

句読点をどこに打つかなどについてはすでに研究が行われているが、そもそも句読点はいくつ打たれるのかについては明らかでなかった。直感的には文の長さに比例するように思われるが、そうであるか確かめるために、データを集め分析した。

2. 方法

次のようなアンケートを実施した。

- Wikipedia・青空文庫から抽出した98の文章から読点を取り除き、日本語自然言語処理ライブラリGiNZAを用いて文節で分かち書きする。
- 回答者は無作為に一文が与えられるので、最も自然に思われる位置に読点を挿入する。

アンケートで得たデータから、文ごとに平均いくつの読点が付与されるかを求め、文の文字数との相関係数を調べた。

3. 結果

合計で1136件の回答が集まった。各文の長さと、それに付与された平均の読点数は図1を参照されたい。
これらに対し相関係数を求めたところ、$\mathrm{r} = 0.867$ と値を得、標本に対しては強い正の相関が認められた。

次に、母集団について、有意差1%での無相関検定を実施したところ、$p < 0.001$ だったことから、母集団においても有意な相関があるだろうということがわかった。

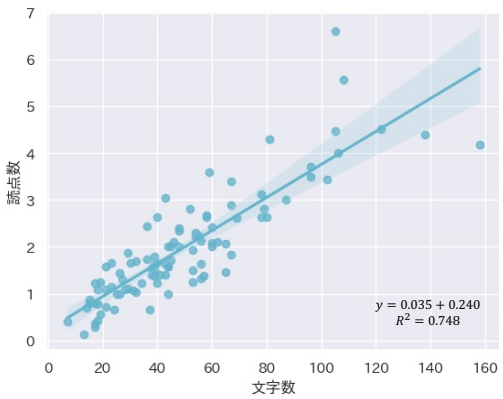

図1 : 文の文字数と挿入される平均の読点数

4. 考察

得たデータに対してさらに単回帰分析を行ったところ、$y = 0.035 + 0.240$ として回帰直線を得た（図1の直線）。これはおよそ 28 文字おきに読点が付与されることを示唆しており、読みやすい文章作成のための1つの指標として用いることができると考える。
しかし、$n \geq 100$ などでの長文においては散らばりが大きく、この指標は適切ではない。実際、100文字以上の文での回答は、図2に示すように、かなりまばらであることがわかる。

また、文例1では、「馬鹿、こんなに晩く（後略）」といったように、感動詞「馬鹿」に読点が続く回答が全てであった。これは岩畑(2004)で「絶対読点」と分類されていたが、この絶対読点の影響を排除して調査し、文の要素・長さに影響する読点、「相対読点」について検討することで、より本質的な分析が可能になると推察できる。

文例1 (斜線は文節の句切れを表す):
「馬鹿 / こんなに / 晩く / 行かなくとも / 明日 / 寝起きに / 行けば / いい / ぢやないか」

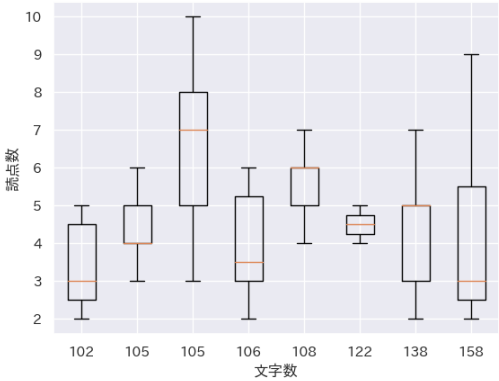

図2 : 100文字以上の長文に対する回答の分布

5. おわりに

本調査では、文に付与される読点の数はその文の長さに相関するか確かめるべく、さまざまな文にどう読点を与えるかをアンケートにより調べ、定量的な分析を行った。

結果として、両者には正の相関があり、文が長くなるほど付与される読点の数も増える傾向にあることがわかった。

加えて、得た結果から、挿入する読点の数の目安や、より効果的な分析のための方法について考察した。

文献:

- 岩畑貴弘, Iwahata, T. (2004). 読点の使用とその決定要素について：「構造」と「長さ」から. 人文研究 : 神奈川大学人文学会誌, 154, A51–A81.
- 拓也岩崎. (2017). 日本語母語話者と日本語学習者の作文コーパスから見た読点と助詞の関係性. 一橋大学国際教育センター紀要, 8, 27–39. https://doi.org/10.15057/28746