作家の句読点の関係について：句点と読点の間に明確な関係はあるか

山沢元成

はじめに

日本語の句読点は書き手によってその位置を比較的自由に決定できる。新国(2016)の実験によると、同じ文章中でも句読点を挿入する位置は個人間で差異が見られたが、中には共通する部分もある。また、江口ら(2020)の研究では、ツイートに含まれる句読点を分析することにより、ある程度年代を推定することに成功していた。これらの研究から、個人が挿入する句読点には特定のパターンが存在すると予想できる。本調査ではそれらが1次式で表されるのではないかと仮説を立てて検証する。

方法

青空文庫から著作権の切れた作品のテキストデータを入手し、それから注釈などを取り除いたものを実験データとして用いる。Mecabを用いて形態素解析をし、その中から句点「。」、読点「、」を取り出し、その個数を集計し、その関係を一次関数で近似してその適合度を調査する。

結果

表１のようになった。調査したすべての作者について。句点と読点の間の関係を高い精度で1次式で近似することができた。ただし、芥川龍之介のデータは他と比較すると当てはまりが悪かった。

結論: 句点と読点の間には１次式で近似できる関係がある。

表1　作者とその作品の中の句点と読点の数を集計し、それらの関係を1次関数で近似した際の傾きと切片、R二乗値を計算した。

名前	句点	読点	データ数	傾き	y切片	R二乗値
芥川龍之介	54865	73989	363	1.4	-7.47	0.662
夏目漱石	88786	88694	106	0.926	61.5	0.937
有島武夫	14519	12774	49	0.851	8.6	0.946
江戸川乱歩	151418	296711	106	1.85	162	0.937
菊池寛	46921	89973	87	1.85	35	0.961
小林多喜二	13197	17555	16	1.34	-10.8	0.946
石川啄木	15827	22739	65	1.17	66	0.868

考察

表１から分かるように、芥川龍之介のR二乗値が比較的低かった原因の一つは、他の作者と比較して、句点がほとんどない短文ばかりで構成された文章や対話形式の文章、雑記のようなものが多く含まれてるなど、文体・書式が多様であったことが考えられる。そのため、句点と読点の関係をより正確に表すには、作品の「種類」を考慮する必要があるかもしれない。

おわりに

ある作家が使用する句点と読点の間には１次関数の関係があることが分かった。しかし、中には比較的適合度が低いものもあった。そのためより正確に句点と読点の関係を述べるためには、どのような文章であるということを考慮したうえで、今回検証した１次関数のモデルを使用する必要がある。また使用したデータ内に不必要な部分が多少含まれていたことなど改善の余地がありそうだ。

文献

新国佳祐 ,「日本語文理解における読点の役割についての認知心理学的研究」,東北大学,2016年8月17日 ,博士論文,pp.9-14.
江口大賀ほか,「ツイートの長さと句読点に基づく年齢・性別の推定」,第82回全国大会講演論文集,2020巻1号,2020年2月20日,pp.431-432.