数学で国語を斬る！

中学１年 杉本修嗣

① 研究をするに至った理由

僕の学校では、毎月一冊の本が配布されます。これは必須として読み切らなければならないのですが、読みやすい本と読みにくいものがあり、読みにくい時にはこの一冊が苦痛以外のなにものでもありません。

そう思っていた時、「計量文献学」という学問があることを知りました。これは、文献の特徴を数値化して、比較検証を行う学問です。「源氏物語は本当に紫式部が書いたのか？」「シェークスピアは一人の人物なのか？」といった昔の作家の特定に利用されるばかりでなく、アメリカなどでは最近まで、誘拐事件などの脅迫文を解析するのに使われていました。

具体的には、「人の記す文章には、その人の個性や癖が現れる。」ということに着目して、品詞分類をした時のそれぞれの使用率、読点と読点の間の文字数、単語または品詞の使用頻度 などを詳しく調査し、個人を特定していくのです。「必ずこれ！」といった決まった解析方法はなく、状況に応じて使い分けているそうです。

もしかすると、僕にとっての読みやすさはこの「計量文献学」で分析できるのではないか？ ここから、今回の研究は始まりました。

② 僕の文章には傾向があるのか？

まず、この「計量文献学」がどの程度信じられるのかを実感するために、自分の文を分析することにしました。

1) データの収集

収集方法…　僕が書いた小5から今に至るまでの4種類の文章から、それぞれ 5 文を任意に選び出し、その品詞の使用率を求めます。
品詞分けについては、自力のみでは危険なので、インターネットの品詞分解ソフトを利用しました。

文の種類…
a)　小5の時の作文
b)　小6の時の作文
c)　僕が書いてきたお城探訪記＆日記
d)　中学に入ってからの文章

2) データ

それぞれのデータです。下段 2 段は、5文の合計、および、単語数にするそれぞれの品詞の割合を記入しました。

a) 小5の作文　　表１

	動詞	名詞	形容詞	形容動詞	助詞	助動詞	連体詞	副詞	感動詞	接続詞	接尾語	total
い	2	4	0	1	7	4	0	0	0	0	0	18
ろ	2	5	1	0	6	1	0	1	0	1	0	17
は	2	4	0	0	8	5	0	0	0	1	0	20
に	4	5	2	0	8	2	0	1	0	0	1	23
ほ	1	3	0	1	5	5	1	0	0	0	1	17
合計	11	21	3	2	34	17	1	2	0	2	2	95
%	11.58%	22.11%	3.16%	2.11%	35.79%	17.89%	1.05%	2.11%	0.00%	2.11%	2.11%	100.00%

b) 小6の作文　　表2

	動詞	名詞	形容詞	形容動詞	助詞	助動詞	連体詞	副詞	感動詞	接続詞	接尾語	total
い	1	6	1	1	6	6	0	1	0	0	0	22
ろ	4	4	0	1	10	4	0	0	0	0	1	24
は	2	4	0	0	4	3	0	0	0	1	0	14
に	3	6	2	0	8	2	0	0	0	0	1	22
ほ	2	3	0	1	9	4	1	1	0	0	0	21
合計	12	23	3	3	37	19	1	2	0	1	2	103
%	11.65%	22.33%	2.91%	2.91%	35.92%	18.45%	0.97%	1.94%	0.00%	0.97%	1.94%	100.00%

c) お城探訪記＆日記　　表３

	動詞	名詞	形容詞	形容動詞	助詞	助動詞	連体詞	副詞	感動詞	接続詞	接尾語	total
い	1	5	1	0	5	2	0	0	0	0	0	14
ろ	5	9	0	1	9	4	0	0	0	0	1	29
は	1	6	0	1	5	5	1	1	0	0	1	21
に	3	7	2	1	10	2	0	1	0	1	1	28
ほ	2	5	0	0	10	5	0	0	0	0	0	22
合計	12	32	3	3	39	18	1	2	0	1	3	114
%	10.53%	28.07%	2.63%	2.63%	34.21%	15.79%	0.88%	1.75%	0.00%	0.88%	2.63%	100.00%

d) 中1の文章　　表４

	動詞	名詞	形容詞	形容動詞	助詞	助動詞	連体詞	副詞	感動詞	接続詞	接尾語	total
い	3	5	0	1	8	3	0	0	0	0	0	20
ろ	3	9	1	1	9	9	0	1	0	0	0	33
は	2	5	0	1	7	3	0	1	0	1	0	20
に	3	5	2	0	10	4	0	0	0	0	0	24
ほ	3	4	1	1	8	3	0	1	0	0	1	22
合計	14	28	4	4	42	22	0	3	0	1	1	119
%	11.76%	23.53%	3.36%	3.36%	35.29%	18.49%	0.00%	2.52%	0.00%	0.84%	0.84%	100.00%

3）考察・分析

正直なところ、ここまで似たような数値が現れるとは思っていませんでした。品詞の性質を考えると「名詞や助詞は頻出度が高いだろう」とか「感動詞の頻出度は低いだろう」という予想はできましたが、書いた年代や性質の違う種類の文章で、これだけ似たような数字が並ぶとは本当に驚きです。

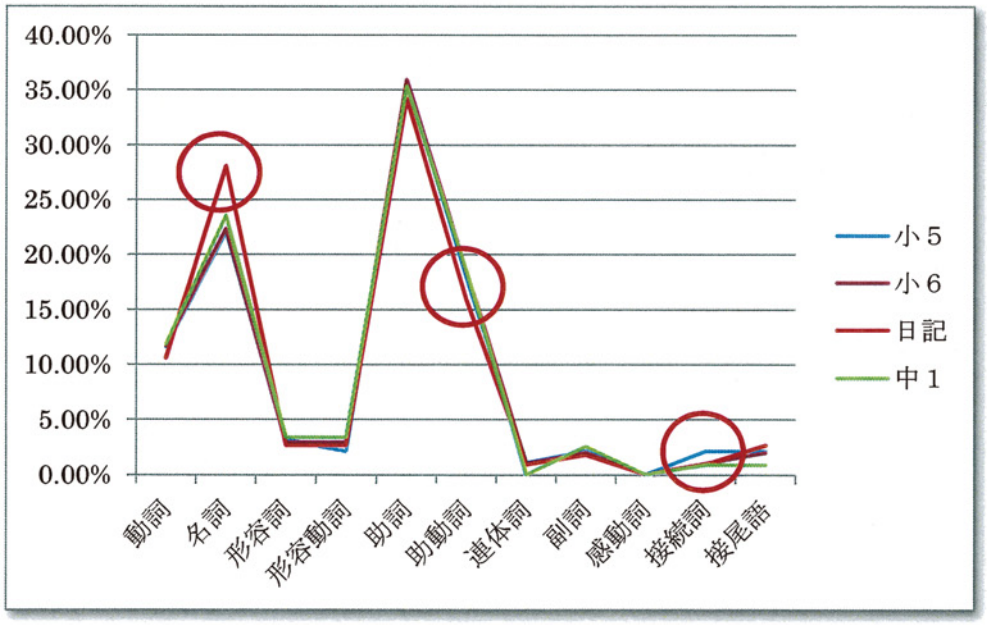
「品詞の頻出度合」と「種類別 1センテンスあたりの単語数」についてグラフ化してみました。

グラフ化した中で、注目するべき、と考えた個所は赤線で囲みました。

＜品詞別頻出度合＞

表5

	動詞	名詞	形容詞	形容動詞	助詞	助動詞	連体詞	副詞	感動詞	接続詞	接尾語
小5	11.58%	22.11%	3.16%	2.11%	35.79%	17.89%	1.05%	2.11%	0.00%	2.11%	2.11%
小6	11.65%	22.33%	2.91%	2.91%	35.92%	18.45%	0.97%	1.94%	0.00%	0.97%	1.94%
日記	10.53%	28.07%	2.63%	2.63%	34.21%	15.79%	0.88%	1.75%	0.00%	0.88%	2.63%
中1	11.76%	23.53%	3.36%	3.36%	35.29%	18.49%	0.00%	2.52%	0.00%	0.84%	0.84%



グラフ１

ア）「名詞」について…

他の種類では 22～24％であるのに対し、「日記」では 28％を超えています。この約 5％の差は誤差範囲外として捉えるべきと考えます。原因は「日記」＝「お城探訪記」なので、城の場所、施設の説明などがメインであるからだと考えました。誤差範囲外と判断はしましたが、今回はデータとしてそのまま使用しました。 …（※1）

イ）「助動詞」について…

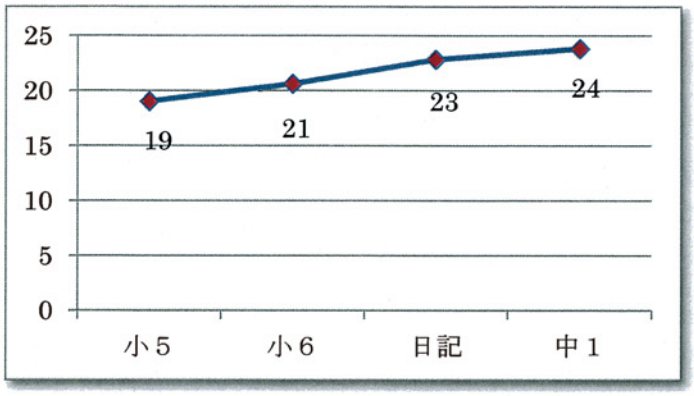
他の種類では 17～19％であるのに対し、「日記」では 16％を割っています。これは、「堀である」「この山は実は土塁」といったように、単純な断定文が多く用いられたり、「体言止め」が多く利用されているからだと考えました。差が 2％程度であるので、傾向はあるものの、一応誤差範囲内と考えました。

ウ) 「接続詞」について…

この品詞は「小5」の文で 2%を超えて使われています。全体として決して多い数字ではありませんが、文と文のつなぎに「接続詞」を多用する幼い子の文の特徴です。恥ずかしながら、この結果は、僕の幼さからきているのかもしれないと判断せざるを得ないようです。

僕の1センテンスにおける単語数の推移を表したものが、グラフ2です。

＜種類別 1センテンスにおける単語数の推移＞



グラフ 2

これをみると、年齢が上がるに従って、1センテンスごとの単語数は確実に増えています。知ってる言語数が増え、それとともに、書く文における単語数が増える … これは成長によるものと考えます。

グラフ1とグラフ2より、1センテンスあたりの単語数が多くなっても、各品詞の頻出度合はあまり変わらず、ほぼ同じ形状のグラフになっています。これは、成長はしているものの、結局書く文の癖にはあまり変化がないということであり、これが「計量文献学」でいうところの「個人を特定できる特徴」になるのだと思います。「計量文献学」の秘めている力を実感する瞬間です。

③ 僕の「読みやすい文章」と「読みにくい文章」に特徴はあるのか？

実際に僕が「読みやすい」「読みにくい」と思う文について、②と同様に分析してみました。

1) データの収集

収集方法… 図書館で適当に取り出した本の中から、適当に開いた数ページを読み、読みやすい本4冊、読みにくい本4冊を選びだしました。それに学校から配布された本二冊を加え、計 10 冊を対象とします。それぞれの本から任意の 3 文を選びだし、その品詞の使用率を求めます。

品詞分けは上記と同様の品詞分解ソフトを利用しました。

選んだ本… 読みやすい本

- 小僧の神様
- 魂にメスはいらない
- 食える数学
- 天地人
- 思城居

読みにくい本

- 詩のこころを読む
- ソシュールを読む
- 1Q84
- ピーター流外国語習得術
- 日本の神秘

2) データ

＜読みやすい文章＞　　表 6

	動詞	名詞	形容詞	形容動詞	助詞	助動詞	連体詞	副詞	感動詞	接続詞	接尾語	total
い	1	5	0	1	4	3	0	0	0	0	0	14
ろ	2	5	3	2	8	2	0	0	0	0	0	22
は	3	3	1	1	8	4	0	3	0	0	1	24
に	2	7	1	1	6	0	2	0	0	0	1	20
ほ	3	9	2	1	6	7	1	0	0	0	1	30
い	5	8	2	2	10	3	0	0	0	0	0	30
ろ	6	8	1	0	15	3	0	0	0	1	1	35
は	3	4	2	0	7	1	0	2	0	0	0	19
に	0	9	4	0	6	8	0	0	0	0	2	29
ほ	3	7	1	1	8	1	0	0	0	0	1	22
い	4	5	2	0	7	5	0	0	0	0	2	25
ろ	3	13	1	2	14	5	0	0	0	0	2	40
は	2	5	1	0	7	3	0	1	0	0	0	19
に	3	5	2	1	11	2	0	0	0	1	0	25
ほ	3	5	0	1	8	2	0	2	0	0	0	21
合計	43	98	23	13	125	49	3	8	0	2	11	375
%	11.47%	26.13%	6.13%	3.47%	33.33%	13.07%	0.80%	2.13%	0.00%	0.53%	2.93%	100%

＜読みにくい文章＞　　表 7

	動詞	名詞	形容詞	形容動詞	助詞	助動詞	連体詞	副詞	感動詞	接続詞	接尾語	total
い	4	11	0	1	14	7	0	1	0	0	0	38
ろ	3	12	0	0	9	5	1	0	0	1	2	33
は	4	7	0	1	9	5	0	0	0	0	1	27
に	4	6	0	0	7	4	2	0	0	0	1	24
ほ	2	7	0	0	9	2	1	2	0	0	1	24
い	4	10	0	1	9	6	1	0	0	0	1	32
ろ	3	7	0	0	9	6	0	1	0	0	0	26
は	5	9	1	0	14	5	0	0	0	0	0	34
に	2	7	0	0	7	4	0	1	0	1	2	24
ほ	3	7	1	1	7	3	1	0	0	0	1	24
い	5	7	0	1	13	7	0	1	0	1	0	35
ろ	4	10	0	0	17	5	2	2	0	0	3	43
は	6	7	1	0	9	6	0	1	0	0	0	30
に	4	14	0	0	14	5	1	1	0	0	1	40
ほ	2	6	0	0	6	2	0	1	0	0	0	17
合計	55	127	3	5	153	72	9	11	0	3	13	451
%	12.20%	28.16%	0.67%	1.11%	33.92%	15.96%	2.00%	2.44%	0.00%	0.67%	2.88%	100%

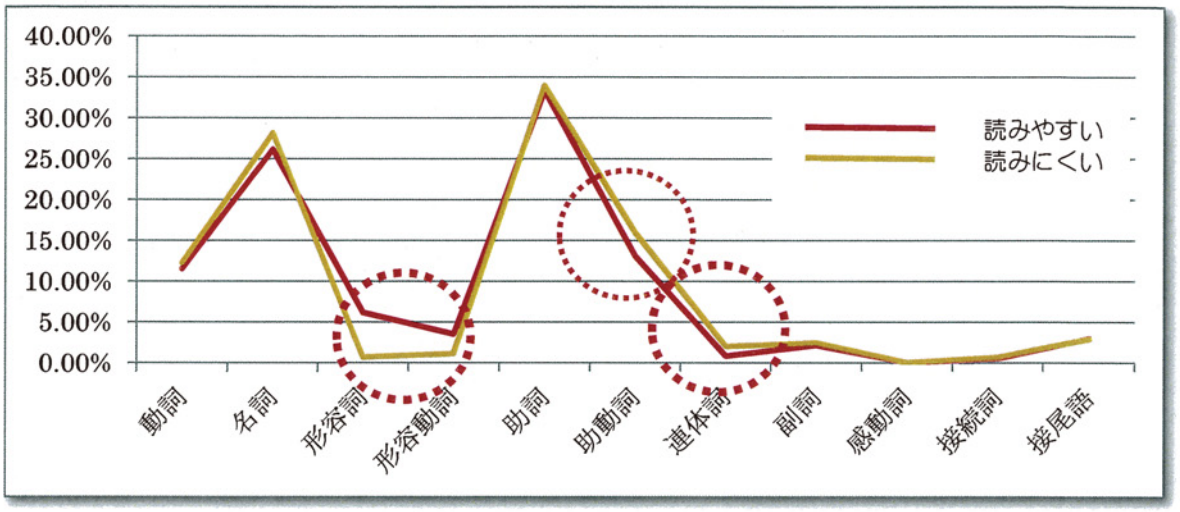
3) 考察と分析

僕が「読みやすい文」と「読みにくい文」とを比較すると、表8のようになりました。2 点が明らかに違うことが見て取れます。グラフ3を見ると、その相違点は出ていますが、グラフの概形は同じです。

＜読みやすい・読みにくい別、 品詞の頻出率＞

表8

	動詞	名詞	形容詞	形容動詞	助詞	助動詞	連体詞	副詞	感動詞	接続詞	接尾語
読みやすい文	11.47%	26.13%	6.13%	3.47%	33.33%	13.07%	0.80%	2.13%	0.00%	0.53%	2.93%
読みにくい文	12.20%	28.16%	0.67%	1.11%	33.92%	15.96%	2.00%	2.44%	0.00%	0.67%	2.88%
自分の文	11.37%	24.13%	3.02%	2.78%	35.27%	17.63%	0.70%	2.09%	0.00%	1.16%	1.86%



グラフ 3

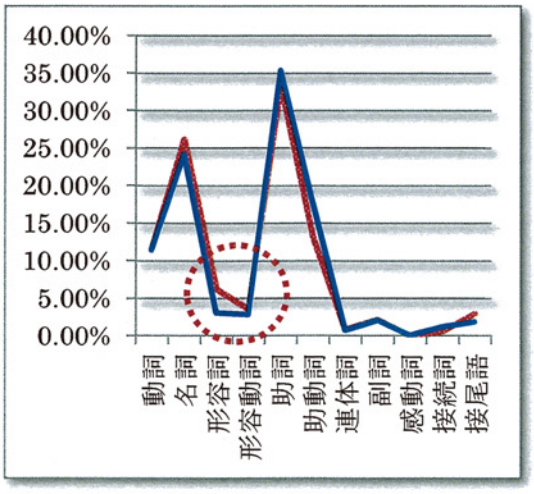
表 8 及びグラフ3より次の 2 点が分かります。

ア) 僕の好き嫌いにかかわらず、物語など読むことを目的とする文章は、似たような品詞の頻出度である(グラフの形より)

イ) 僕の「読みやすい」と「読みにくい」の差は、形容詞・形容動詞・連体詞・助動詞にある

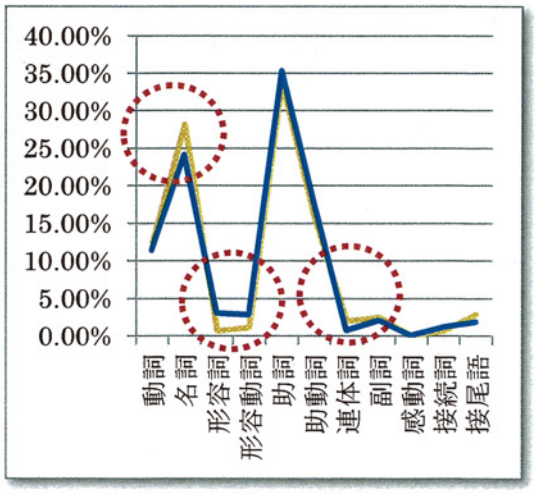
「僕の書く文」と「読みやすい文」、「読みにくい文」についても、比較しグラフにしてみました。

＜読みやすい文 vs 僕の文＞



グラフ 4

＜読みにくい文 vs 僕の文＞



グラフ 5

グラフ4、グラフ5より、

ウ) 表8で、読みやすい文・読みにくい文と僕の値を比較したとき、その差が小さい方に色をつけました。黄色は、その差が明らかに小さい(相手の数値差が2倍以上ある)ものです。近い数値を示す品詞の個数としては 5 対 5 ですが、「読みやすい文」には黄色のデータが4つある為、僕の書く文章は、僕の「読みやすい文」に似ている

ということがわかりました。

ア) グラフの形について…

前にも述べましたが、日本語である限り、「助詞・助動詞」が多くなることは当然です。「感嘆詞・接続詞・接尾語」などは、人により「使うか使わないか」という差が明らかに出て、その値の大きさが一つの個人特定の材料になると思われます。しかし、その数が助詞を上回ったりすることは一般の文章では考えられません。従って、グラフの形としては、二つ山のある、この形が一般的なのだと思います。

この形が違ってくると思われるのは、本の種類によるのではないでしょうか？ 例えば、

「幼児向きの本」→「感嘆詞」が多くなる。 「助動詞」が少なくなる。

「科学や数学などの学術書」→「形容詞・形容動詞・副詞・連体詞」などはは少ない。

などとなるような気がします。これは、また実際の本で確認する必要があります。 ･･･(※2)

イ)「読みやすい」「読みにくい」の差について…

まさに、僕の好みが数値にて表された個所です。違いが現れた「形容詞、形容動詞、連体詞」。これらの共通点は、体言を修飾するときに使われる単語です。つまり、体言を修飾するときに、どの品詞で修飾するかです。僕は、グラフ3からわかるように、「形容詞」や「形容動詞」で修飾された文を読みやすいとし、「連体詞」で修飾されたものは読みにくいと判断しています。

体言を修飾する場合、「形容詞・形容動詞」を利用すると「可愛い人」「静かな教室」というように、直接そのイメージが表現されています。それに対して、「連体詞」を利用すると「色んな家」「あらゆる先生」のように同様にイメージしやすいものもありますが、「あの様子」「その場合」という「こそあど」言葉なども含まれ、実際のイメージについて、自分で考える必要が出てきます。つまり、常に頭を働かせながら本を読む必要がでてきます。その煩わしさが、僕が感じる「読みにくさ」なのかもしれません。 ･･･(※3)

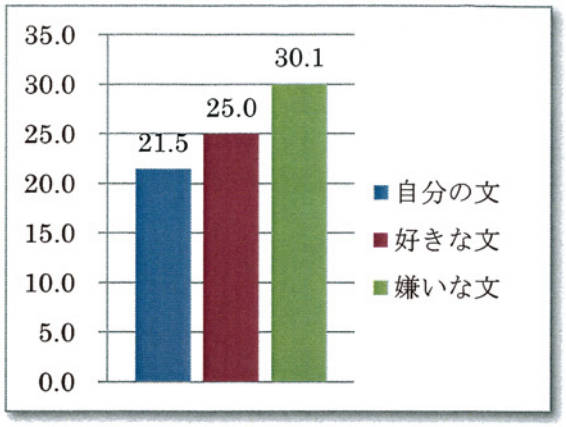
これは「助動詞」の頻出度にも表れていると思います。一般に「助動詞」の使用頻度が増えると、文章の末尾が複雑になってきます。例えば「られる」などの場合、それが受け身なのか使役なのかなどを考える必要がでてきます。一瞬でも、「どっちかな？ どういう意味かな？」と考える事を僕は本能的に避けているのかもしれません。

名詞の使用頻度については、「読みにくい」「読みやすい」とも僕のデータより、多い値を示していますが、あまり多いものは好まないようです。

ウ)「僕の書く文」について…

完全に一致したわけではありませんが、「読みやすい」からそのタイプの文を「僕が書く」のか、「僕が書く文」に似たタイプだから「読みやすい」と思うのか。いずれにしても、「読みやすい」と感じる文は、比較的自分の書く文に似ていることがわかります。

グラフ 6 の<1センテンスあたりの単語数>を見ると、「僕の書く文」と「読みにくい文」の「一文あたりの単語数」には 9 文字近い差があります。単語数が多ければ、全体的に文は長くなるわけで、「文が短い方が読みやすい」というのは、誰でもが感じることです。しかも、一文で 9 単語違うと、文章となった場合のリズム感が変わってくるのは当たり前です。僕の「短いセンテンスのほうが読みやすい」というのは、結局、理解しやすいということなのかもしれません。感覚的に、当たり前の結果ではありますが、それが数学的に証明されたのですから、一つの成果であったと思います。



グラフ6

エ) 表8、グラフ4,5,6 を総合的に考えると…

「僕が書く文」と「読みやすい文」の単語数を比較すると、「読みやすい文」の方が単語数が多く、僕はもう少し長めの文がよいと感じていることが分かります。同様のことは、「形容詞」「形容動詞」の使用頻度にも表れていて、いずれも「僕の書く文」の値が「読みやすい文」と「読みにくい文」の中間に位置しています。つまり、僕の感覚としては、もっと形容する言葉を多用し、もう少し長めの文が良い、と思っているのに書けていない、ということです。

文を書く時、色々なイメージを伝える言葉を使った方が良いとか、文が長い方がレベルは高い…というわけではありませんが、とりあえず僕の場合、自分で好きな文を書けるレベルまでは達していない、ということが総合的に分かった事実のようです。

④ 研究を通して

今回の実験をするにあたり、一番気を使ったのはデータの数でした。1冊の好きな本と嫌いな本のデータをとって判断することは危険だという事は、感覚的にもすぐに分かります。でも、100冊もの本からデータを取ることはできません。では、いくつが適当なのか？！

支持率調査は 1600 人程度のデータ数で行っていますし、テレビ視聴率調査などは 300 件程度のデータ数で行っていることが知られています。だから、今回の必要データ数も、簡単に計算で割り出せると思っていました。ところが調べてみると、「基準の対象がない場合は、まず少ないデータ数で予備調査を行い、更にもう少しデータ数を増やして予備調査を行う、ということを繰り返し、データの分布形などを検討し、データのばらつきや変化が許容範囲(小数点何桁レベルの誤差)に納まるようなデータ数を算出する。」という実に人間臭い作業をして必要データ数を決めなければならないということがわかりました。つまり、未知の調査項目では必要なデータ数を導き出すためには試行錯誤で適切と思われるデータ数をひたすら探すしかないということなのです。この事は、「統計学」を用いれば、もっと簡単に必要データ数が求められると思っていた僕にとって、かなりの驚きでした。

また、データ値として、違う値が出てきた場合、その差は誤差範囲内なのか、誤差範囲外…つまり異常値なのか？…これについても、悩みました。調べてみましたが、どの程度を異常値とするかとい

うのも、分析する側が決めるという事を知りました。勿論「統計学」では、分布状態、母集団、データ抽出方法、などが分かれば、それなりのデータ数を計算する方法はあるようですし、正規分布の場合は上下2.5%(全体の5%)を異常値とするという一般的な数値は存在するようですが、とても僕の手におえるものではありませんでした。
そこで今回は、データ数については、「正規分布を示すサンプルであれば、20～30 個、正規分布でない場合は最低6～8個、厳密さを要求される場合は 50 個位のデータ数が必要」という経験則上指標的数字を利用して 15～20 で行うことにしました。異常値については、今回はグラフにしたとき、あきらかに離れているものについて考察しました。

最初から、こうした新たな発見があった研究ですが、当初の目的としていた、僕の「読みやすい・読みにくい」は、僕の場合、一文当たりの単語数の多さと体言の修飾方法にあり、それらを突き詰めると僕にとって「読みやすさ」=「理解しやすさ」であるということになりました。
実験を始める前、文章には書く人の癖が現れるとは思っていましたが、それらを数値化するのは、かなり大変な作業で、膨大なデータ数が必要であろうと思っていました。正直なところ、自分の文章の傾向を調べた段階で、この研究は終わる…つまり諦めることになるのだろう、と予想していたのです。それは、「計量文献学」のなかでも、一番単純な品詞ごとの頻出度合という手法を使ったからでもあります。ですから、自分の文章が数値化された時は、本当に驚きました。その上、僕の感じる「読みやすさ・読みにくさ」の原因が、数値でもって表されたのには感動しました。
今回の簡単な実験で、「読みにくい」「読みやすい」…というとても曖昧な人間の感情すらも、数学の手法によって傾向が数値化できたのは、驚くべき確認であると同時に、数学の魅力ある未来をみるような気がしました。
人の感情は、その日の天気ででも変わるように僕は思っています。でも数学の前では、嘘がつけないということなのです。今回のデータも、結果的に、数日間に渡ってとることになったのですが、僕の気分によってだけで「読みやすい・読みにくい」の傾向が変わらない事を見事に示してくれました。また、成長しているはずなのに、「僕の書く文」がここ3年ほど、傾向が変わっていないことも確認できました。ある意味恐ろしいことですが、数学の世界は、一見かけ離れている文学の世界までにも入り込めるという新たな発見となりました。

⑤ 今回の研究を終えて

今回の僕の実験での反省点は※1、※2、※3の点などいくつかあります。例えば※3「読みやすさ」の傾向が分かり「連体詞」に着目することになった段階で、すぐに「連体詞」を 4 種類ほどに種類分けして分析しなかったことです。それをすれば「こそあど」言葉が苦手なのかもしれない、という仮説が正確に判定されていたはずです。※1や※2についても、もう少し時間をかければ、僕の予想だけでなく、より正確に判定できたはずです。また、全体的な意味では、僕自身だけでなく、他の人にも協力してもらい、同様の実験をすれば、本当に個人を特定できるかどうかまで、分かったかもしれません。ただこの場合、やはり最低 15 人程度のデータ数が必要になるわけで、とても協力者がでてくるとは思えませんでしたが…。
早めに方向性を考え時間をとり、協力者を探しておけば、もう少し、数学的に深い意味のある結果が出せたのではないかと思いました。

それでも、今回の研究で僕は、一番数学とは遠いところにあると思われがちな人間の感覚すらも、数学的に分析できるということを知りました。昔、レオナルド・ダヴィンチが、黄金比を多く用いる事で「モナリザ」など美しい作品を作り出したと聞いたとき、黄金比について調べ、数字の世界が美術や音楽まで入り込んでいることを知りました。ピタゴラスが「数」「図形」「音楽」「天文学」を四大学問としたのは有名な話です。つまり、1000 年以上昔に彼らはすでに、「数学によって世界の全てが決まり、美しさですらも比で表せる」という事に気が付いていたのです。しかし人間の感覚にまで数学が入り込める,ということにも気づいていたのでしょうか？ 僕は文についての読みやすさという分析を簡単に行ったまでですが、これを一歩一歩深めていくと、人間同士の好き嫌いまでを数値化できるのかもしれません。「数」「図形」「音楽」「天文学」ついには、「人間の感覚」…。もしかしたら、「社会を発展させてきた基礎となるものは数学」なのかもしれません。
そういえば、三角関数を駆使して作られた地形データを基に、コンピューターの上で動く地理情報システム(GIS)が作られ、この GIS を基に警察の中に貯蓄された巨大なデータと組み合わせて犯罪予知システムなどが動いているそうです。つまり、三角関数という数学が、周りの数学の世界を取り込んで、犯罪予測…という人間の行動予測をしはじめている。そう考えていたら、数学の進歩により、数学が世界を動かし始めたような奇妙な感覚になりました。

まだ、僕は将来、どんな道に進むかは決めていません。しかし、もし、数学の道に進むとしたら、数学の一番の基礎となる「1」(僕は、それを数学の真髄と名付けました)を発見したいと思っています。それが、どんなものであるか分かりませんが、多分、レオナルド・ダ・ヴィンチだってピタゴラスだって発見していないはずです。ここまできたら、数学が哲学みたいな話になってしまいますが、人間が数学に乗っ取られないためにも、数学は世界に貢献する力を持つだけでなく、恐ろしい力を秘めているかもしれないという事を感じながら、これからも数学に挑戦していきたいと思いました。