日本語の計量的分析

—教養ゼミナールでの試みから—

田　貝　和　子

Statistical Analysis of Japanese : At the Trial in the Culture Seminar Class

Kazuko Tagai

（平成23年12月 6 日受理）

This is the presentation about the statistical analysis of Japanese language, and analysis which was introduced at the "Culture Seminar" Class of the 3rd grade of 2011th. In this class, the students selected two kinds of texts exceeding 1000 Japanese characters on the web etc.. After analized by a morphological method on the "Chasen, Windows version" and checked on the Windos Excel, the data was compared with the analysis of "science of a style." The text chosen by the students was mostly their favorite novels from "Aozora Bunko", and some students chosed news paper texts. Although it seems difficult for the student to give part-of-speech information correctly, they advanced gradually by checking the data briskly through my guidance. As a result, we can point out that news paper texts shows high noun ratios and juvenile texts do not show the correlation between noun ratio and Chinese character words.

1. はじめに

国語学の分野においても，計量的手法に基づき，ことばを分析する方法がある。この手法のメリットは，誰が調査を行っても同じ結果が得られ，客観的な主張ができるということにある。

日本語の文章を客観的にとらえる指標の一つとして，樺島忠夫･寿岳章子『文体の科学』綜芸社（1965）がある。50年近くも前に出版されたものであるが，これ以降，国語学的文体研究の方法論は出ておらず，個々の事象にしたがって，その時々の方法で調査を行っている状況である。

計量的な手法には，工学分野の技術が必要となる。そこで，平成23年度前期の教養ゼミナールにおいて，日本語も統計的手法を用いることにより，計量的に観察することができることを実践した。本稿は，この教養ゼミナールでの授業の報告であるが，取り上げたデータに関しては，学生の調査ではなく，改めて調査を行っている。

2. 『文体の科学』

『文体の科学』における調査は，当時の「現代小説」100作品から各作品80文を無作為抽出して分析している。調査項目は，

①名詞の比率（%）
②MVR
③指示詞の比率（%）
④字音語の比率（%）
⑤文の長さ（自立語数）
⑥引用文の比率（%）
⑦接続詞をもつ文の比率（%）
⑧現在どめの文の比率（%）
⑨表情語の比率（‰）
⑩色彩語の比率（‰）

である。「MVR」とは，「形容詞・形容動詞・副詞・連体詞の組の比率Mに100をかけたものを動詞の比率Vでわった値」のことである。つまり，様態を表す語が多いか，動作を表すが多いかであり，数値が高ければ様態を表す語が多く，低ければ動作を表す語が多いということである。「表情語」とは，オノマトペなどのことである。

この調査を80文ずつではあるが，100作品すべてについて行い，統計的特性値の大きさを評価するために，平均値からどれだけ離れているかを，5段階尺度を用いている。「現代小説」100作品において全作品の数値の少ない方の10％以下は「極めて小」，30％以下は「小」，多い方の10％以下は「極めて大」，30％以下は「極めて大」となり，それ以外の40％を「普通」としている。以下に『文体の科学』における調査の統計的特性値の分布である。[1)]

6．文体の統計的観察

評語 / 出現率	極めて小 ← 10％以下	小 ← 30％以下	普通	大 → 30％以下	極めて大 → 10％以下
名詞％	45	48	54	56	
ＭＶＲ	34	41	55	65	
指示詞％	2.1	2.8	5.0	6.0	
字音語％	13	16	26	31	
文長	7	9	14	18	
引用文％	1	8	30	70	
接続詞をもつ文％	3	7	21	27	
現在止％	3	13	47	76	
表情語‰	0.4	3.5	13.5	24.5	
色彩語‰	／	1.0	7.5	17.0	

表１『文体の科学』（綜芸舎）より

この数値は，小説の調査であり，ジャンルの異なる文章の調査においては，偏った結果となることが予想されるが，小説以外のジャンルにおいて，他に比較となるような調査データは管見では見当たらないため，小説との違いとして捉えたい。

3. 学生の作品選択

教養ゼミナールの授業においては，拙稿（2009）[2)]の田邊花圃著『藪の鶯』で行った調査を紹介し，前章２で述べたそれぞれの調査項目における数値の意味を押さえた。また，日経オンライン（www.nikkei.co.jp）トップ記事をコピー＆ペーストし，日本語解析ソフトによるデータ分析を紹介した。その上で，ウェブ上で閲覧可能なニュース，青空文庫（www.aozora.gr.jp）等から好きな文章1000字以上を２種類コピー＆ペーストし，調査及び比較を行うようにした。

学生21名の選んだ文章は以下の通りである。[3)]

・ニュース：５名（時事，スポーツ等）
・近代小説：８名（太宰，漱石，鷗外，芥川等）
・児童文学：２名（宮澤賢治等）
・翻訳作品：２名（ポー，リットン等）
・近代小説と現代小説：２名（太宰等）
・現代推理小説：１名（綾辻行人，有栖川有栖）
・同作家別ジャンル：１名（ジョナサン）

分析をする上では，ニュース記事等を選択した方が後の作業が容易であるが，意外にも青空文庫からの小説作品を選んだ学生が多かった。中でも太宰作品が顕著である。

4. 日本語解析ソフト

日本語をある単位によって分ける処理を「形態素解析」という。この「形態素解析」は「自然言語処理のために研究，開発されてきたもの」[4)]であるが，この「形態素解析」が行われたデータは，日本語研究においても有効である。教養ゼミナールにおいて『文体の科学』と同様の調査を行う際に，調査項目を一つ一つカウントする手間が省ける部分も出てくるというメリットがある。

4.1 形態素解析ツール

形態素解析システムは，無料でダウンロードできるものが，いくつかある。

・JUMAN（http://nlp.ist.i.kyoto-u.ac.jp/index.php?日本語形態素解析システムJUMAN）
・茶筌（http://chasen-legacy.sourceforge.jp/）
・MeCab（http://mecab.sourceforge.net/）

上記いずれもUNIX系のOS用とWindows用が公開されている。Windows用といっても，コマンドプロンプト上で処理を行う必要があるものがほとんどである。また，解析用辞書を別にダウンロードする必要もあり，パーソナルコンピュータの利用度に差がある学科混合型の授業においては，利用が困難であった。

「茶筌」には，GUI版で視覚的にも簡便な「wincha（chasen-2.1-ipadic-2.4.4-sp5-exe）」がある。これは，解析用辞書も同梱されており，ダブルクリック一つでといっていいほど，ダウンロードも簡単である。膨大なコーパスデータを一度に解析する必要がない場合には，有効である。

4.2 Windows版茶筌

「WinCha」の初期画面は以下のようになっている。

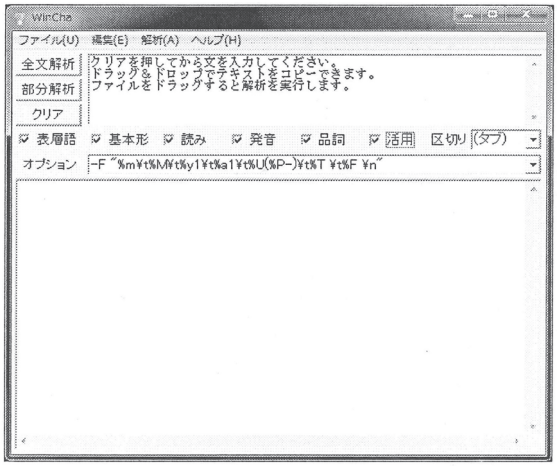

図 1　WinChaの初期画面

上の小さな四角の中に，直接文を入力するか，あるいは，コピー＆ペーストも可能である。そして，ほしい解析情報の欄にチェックを入れ，「全文解析」を押すと下の大きな四角の中に，文すべてを解析した結果を表示する。

授業で示した日経オンライン上のトップ記事（2011.04.26，626文字）は，「全文解析」により解析を行うと以下のような状態になる。

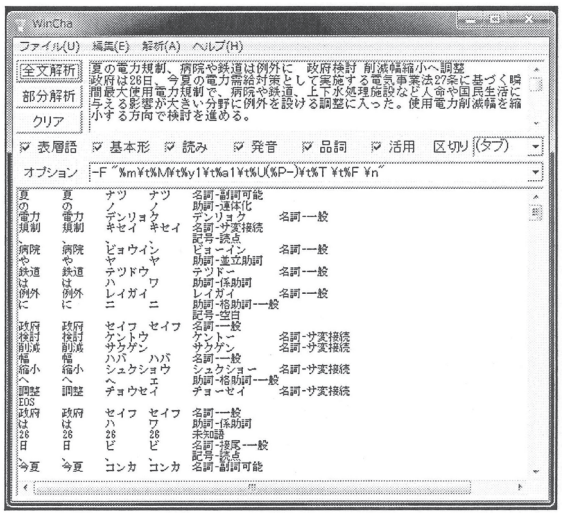

図 2　ニュース記事の解析

上記は区切りを「タブ」にしている。他に「空白」「改行」「コンマ」を選択できるが，タブ区切りにすると，そのままExcelに貼り付けることができ，データの扱いが簡便である。

図2を全選択し，Excelに貼り付けたのが，図3である。

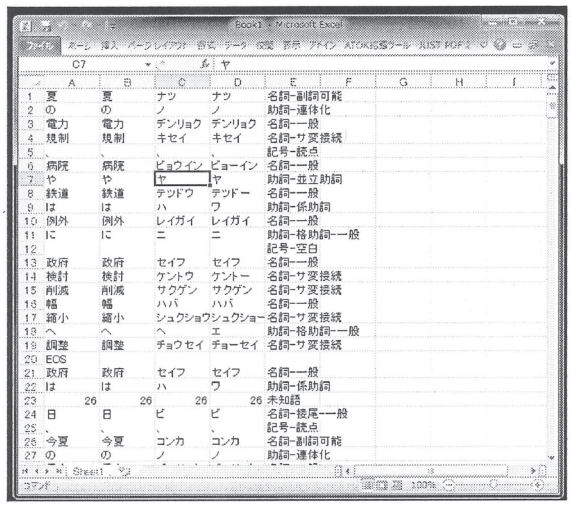

図 3　ニュース記事のExcelデータ

以上の作業を，学生がそれぞれ2種類の文章について行った。その二つのExcelファイルのデータで，もともと改行や空白であった部分の行を削除し，それぞれのファイルデータをナンバリングした。

5.　短単位情報の結合

「茶筌」は，単語よりも細かい単位で形態素解析をするため，情報を結合させる必要がある。例えば，接頭辞，接尾辞等である。特に，形容動詞は，形容動詞語幹と活用語尾が分かれてしまう。例えば，「お利口さんだ」は「お／利口／さん／だ」と四分割されてしまうため，これを一つにする必要がある。「お利口な」も「お／利口／な」と解析される。

お	接頭詞-名詞接続
利口	名詞-一般
さん	名詞-接尾-人名
だ	助動詞

お	接頭詞-名詞接続
利口	名詞-形容動詞語幹
な	助動詞

「接頭詞」というのが，品詞よりも細かい単位であることや，複合語など，意味のまとまりのあるものを一つにする作業は，学生にとって困難であったようである。形容動詞に関しては，「名詞」の後にある「形容動詞語幹」という情報を手がかりに，ウェブ上にある辞書を丁寧に引いていく作業をしていく必要があるが，形容動詞の認識は，国文系の大学生

にとっても難しいものであるため，授業内においては，ある程度作業が終了した時点で，Excelファイルをこちらでチェックする方法をとった。

Excelファイルの情報が品詞単位でほぼ正確になった段階で，『文体の科学』の項目を加えた。つまり，指示詞，字音語，引用文，現在止め，色彩語，表情語など，品詞情報だけでは，把握できない情報を付与する。Excelのフィルタ機能を有効に使うことで作業効率が良くなるが，Excelを使い慣れていない学生も多いようであった。

6. 日本語の計量的分析

『文体の科学』での調査項目における図式を際立った特徴のある作品のみ以下に示す。

6.1 同作家 2 作品

太宰治の作品を採択した学生は非常に多くいた。その中で，「秋」と「朝」の比較を図 4 に示した。

まず，MVRの値が大きく異なる。「秋」は50であり，「朝」は71.43である。「秋」の方が動詞が多く用いられている「動き描写的」な文章であり，「朝」の方は，様態を示す「ありさま描写的」である。指示詞，文長，接続詞も大きく違う。同作家でも作品により，文体が違うことがわかる。

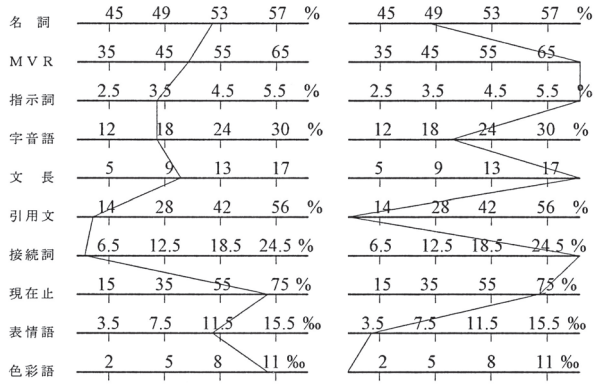

図 4 太宰治「秋」（左）「朝」（右）

6.2 同名２作家

図 5 は，どちらも「竹青」という作品であり，田中貢太郎と太宰治である。これは，中国文学の翻訳であり，内容はほとんど同じであるが，文体としては，大きく異なる結果となっている。

田中の方が，MVRが30.39であり，「動き描写的」であり，太宰はMVR80と「ありさま描写的」である。また，太宰の方が文長が大きいのは，会話文が多いということである。同じ題材をもとにしても，作家によって，文体が変わることを表している。

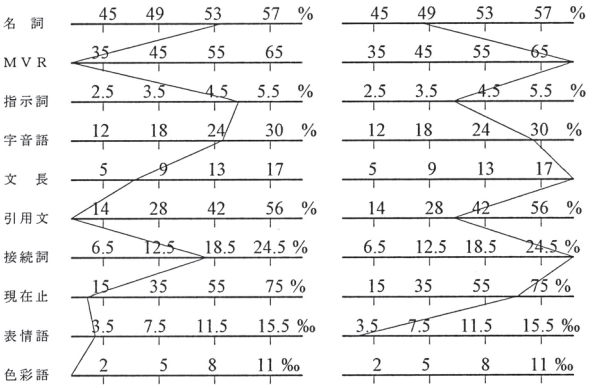

図 5 「竹青」田中貢太郎（左）太宰治（右）

6.3 近代作家

図 6 は，森鷗外「沈黙の塔」と夏目漱石「琴のそら音」である。どちらも明治期に活躍した作家であるが，文体は大きく異なる。

まず，名詞の比率が異なる。そして，鷗外は字音語も多く，漢語を多く用いていることがわかる。漱石は会話が多く，引用文の比率が高くなっている。

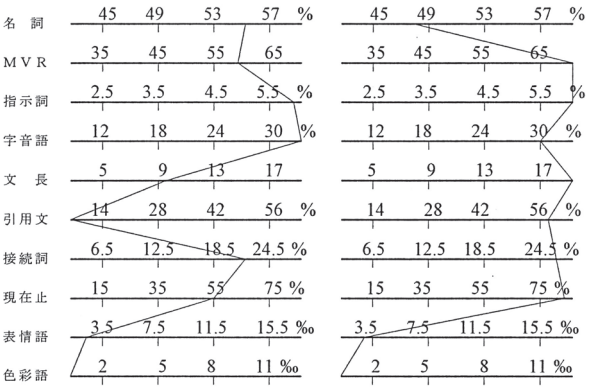

図 6 森鷗外「沈黙の塔」（左）夏目漱石「琴のそら音」（右）

6.4 児童文学

図 7 は宮澤賢治「どんぐりと山猫」と佐野洋子「百万回生きたねこ」を比較したものである。どちらも児童文学としての特徴がみられる。つまり，名詞の比率と字音語の比率が相関関係をなしていないということである。子ども向けの易しい文章では，漢字はほとんど用いないのである。また，どちらもMVRが高く「ありさま描写的」であり，「どんなだ」という表現を多く用い，様子を記述しながら展開していく。表情語，色彩語も多い。

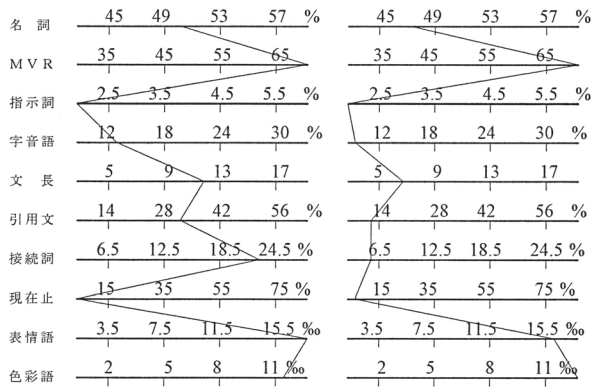

図７　宮澤賢治「どんぐりと山猫」（左）
佐野洋子「百万回生きたねこ」（右）

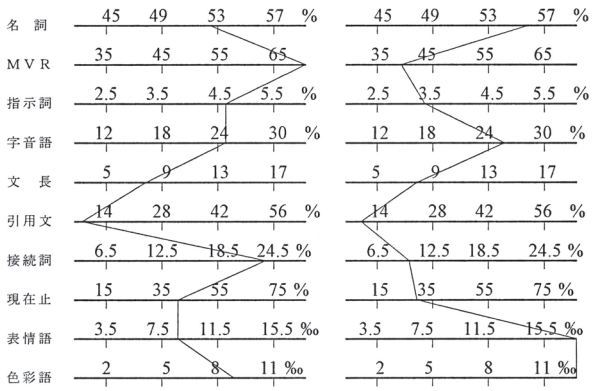

図９　綾辻行人「十角館の殺人」（左）
有栖川有栖「46番目の密室」（右）

6.5　近代作家と現代作家

図８は太宰治の「人間失格」と現代作家である乙一の「天帝妖狐」の比較である。数値は相反しているが，どちらも名詞比率と字音語に相関関係がないことが特徴的である。太宰は漢語の多い固い文章，乙一は和語の多い読みやすい文章であることがわかる。また太宰は，作品毎にさまざまな文体を用いていることがわかる。

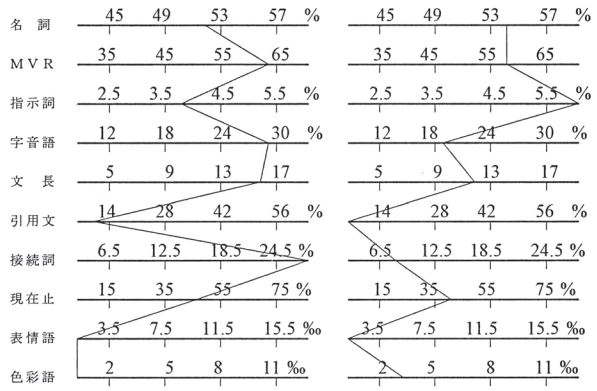

図８　太宰治「人間失格」（左）乙一「天帝妖狐」（右）

6.6　現代推理小説

図９は綾辻行人の「十角館の殺人」と有栖川有栖の「46番目の密室」を比較したものである。どちらも現代の推理作家であるが，ジャンルが同じであっても，文体が異なっていることがわかる。MVRに注目すると，綾辻行人は「ありさま描写的」であり，何がどうであるかを詳しく述べている。それに対し，有栖川有栖は「動き描写的」で何がどうしたということを主に進めていく文章であることがわかる。

6.7　児童文学と評論

図10はスウィフト・ジョナサンの「ガリバー旅行記」と評論のような文章の「穏健なる提案」である。児童文学は字音語が少なく，表情語となっているオノマトペを多く用いることがよくわかる。反対に，評論文は字音語を多く用い，表情語を用いていない。

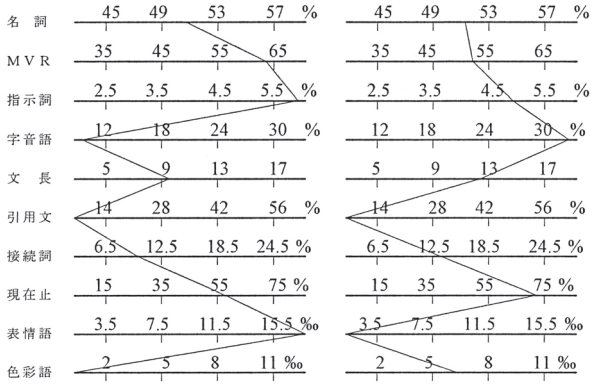

図10　スウィフト・ジョナサン
「ガリバー旅行記」（左）「穏健なる提案」（右）

6.8　スポーツ記事

図11はどちらも野球記事である。小説と比べると，名詞の比率が多く，MVRが少ない。つまり，「要約的文章」であることがわかる。また，指示詞も少なく，他の要素も少なく，できる限り名詞（字音語）を多くして，情報量を増やしていることがわかる。

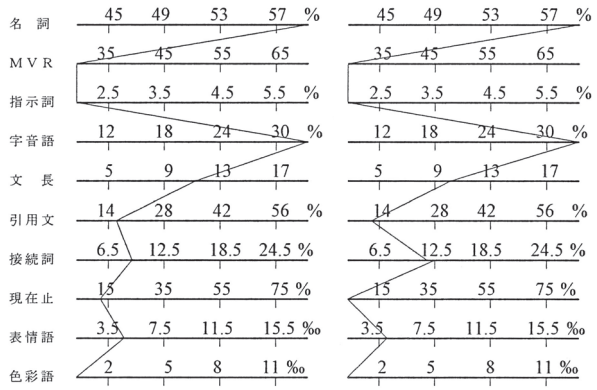

図11　野球記事「サンケイスポーツ」（左）
「日刊スポーツ」（右）

7. おわりに

以上，教養ゼミナールでの日本語の計量的分析を紹介した。『文体の科学』の指標は，小説がベースとなっているが，他のジャンルにおいても有効であることがわかった。しかしながら，数値が振り切ってしまうこともあるため，特性地の見直しが必要かもしれない。

教養ゼミナールの授業では，文法知識やExcelの使用など，慣れないことも多く，学生にとって困難な作業であったようだが，日本語も統計的手法により，客観的な分析ができることを是非知っておいていただきたい。

注

1）樺島忠夫･寿岳章子『文体の科学』綜芸舎（1965），p.130
2）田貝和子「花圃著『藪の鶯』の文体―統計的分析から―」『言語文化研究』（2009）
3）具体的には以下のような採択である。
〈ニュース：5名〉
「サッカー／サッカー」「野球／野球」「風力発電／BDプレーヤー」「電力／食中毒」「ロイヤルウェディング／ロイヤルウェディング」
〈近代小説：8名〉
同作家・太宰治「パンドラの匣／フォスフォレッスセンス」「秋／朝」
同名作品「嘘（太宰治／渡辺温）」「竹青（太宰治／田中貢太郎）」
「沈黙の塔（森鷗外）／琴のそら音（夏目漱石）」「羅生門（芥川龍之介）／我が輩は猫である（夏目漱石）」「こころ（夏目漱石）／桜の森の満開の下（坂口安吾）」「街上スケッチ（牧野信一）／買食ひ（片山廣子）」
〈児童文学：2名〉
「銀河鉄道の夜（宮澤賢治）章違い」「ドングリと山猫（宮澤賢治）／百万回生きたねこ（佐野洋子）」
〈翻訳作品：2名〉
「盗まれた手紙（エドガー・アラン・ポー）／イエスとペテロ（片山廣子）」「貸家（エドワード・リットン）／牡丹灯記（瞿佑）」
〈近代小説と現代小説：2名〉
「人間失格（太宰治）／空の境界（奈須きのこ）」「人間失格（太宰治）／天帝妖狐（乙一）」
〈現代推理小説：1名〉
「十角館の殺人（綾辻行人）／46番目の密室（有栖川有栖）」
〈同作家別ジャンル：1名〉
スウィフト・ジョナサン「ガリバー旅行記／穏健なる提案」
4）柏野和佳子「形態素解析」『講座ITと日本語研究2　アプリケーションソフトの基礎』明治書院（2011）p.132

参考文献

荻野綱男・田野村忠温編『講座ITと日本語研究1　コンピュータ利用の基礎知識』明治書院（2011）
荻野綱男・田野村忠温編『講座ITと日本語研究2　アプリケーションソフトの基礎』明治書院（2011）
荻野綱男・田野村忠温編『講座ITと日本語研究5　コーパスの作成と活用』明治書院（2011）
荻野綱男・田野村忠温編『講座ITと日本語研究7　ウェブによる情報収集』明治書院（2011）
樺島忠夫・寿岳章子『文体の科学』綜芸舎（1965）
計量国語学会編『計量国語学事典』朝倉書店（2009）