様式 C-19

科学研究費補助金研究成果報告書

平成２３年　５　月　２８　日現在

機関番号：１６１０１
研究種目：基盤研究（C）
研究期間：　2008　～　2010
課題番号：２０５２０３８９
研究課題名（和文）　文長にみる言語の確率分布

研究課題名（英文）　On probability distributions of Japanese sentence lengths.

研究代表者
石田　基広（ISHIDA MOTOHIRO）
徳島大学・大学院ソシオ・アーツ・アンド・サイエンス研究部・准教授
研究者番号：４０２３２３１８

研究成果の概要（和文）：本研究では，文長の分布が対数正規分布やパスカル分布，あるいは負の二項分布によって表現できるかどうかを検討した．その結果，文長の分布を確率現象としてだけ説明することは不可能であった．さらに文長に一般化線形モデルをあてはめたところ，書き手などの固定的な要因が影響している可能性が示唆された．また本研究において，日本語文章を文字や形態素，あるいは品詞に分解し，これを頻度表として出力するプログラムを開発し，これらを公表した．

研究成果の概要（英文）：
The first aim of this study was to see if any of these probability distributions, log-normal distribution, Pascal distribution, and negative binomial distribution, could be really fit to Japanese sentence lengths.　These had long been hypothesized as the best fit to some European languages (log-normal distribution to Japanese), but none of these proved to be appropriate.　And applying a generalized linear model showed some possibility that sentence lengths might be affected by some other fixed factors than only probability phenomena.
The second aim of developing the software that calculate Japanese sentence lengths automatically has been achieved, and two original software packages are now open to the public.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,100,000	330,000	1,430,000
2009 年度	500,000	150,000	650,000
2010 年度	500,000	150,000	650,000
年度			
年度			
総　計	2,100,000	630,000	2,730,000

研究分野：言語学
科研費の分科・細目：言語学・言語学
キーワード：計量言語学，テキストマイニング，日本語，英語，独語

１．研究開始当初の背景

(1) 欧米語では文章を構成する単位の頻度が，ある確率分布に従うのではないかという仮説がたてられている．すなわち文の長さとして，単語の数や節のなどを数えた場合，その全体平均と個別の文長の差（誤差）が，対数正規分布やパスカル分布，あるいは負の二項分布などの確率分布で表現できるという仮定である．しかし，これを検証しようとした研究が対象としたデータ数は非常に少ない．

また検証手法としては，ほとんどの場合にカイ自乗分布による適合度の検定が用いられている．
(2)ひるがえって日本語研究の分野では，文長の研究はほとんど行われておらず，2 名の研究者による試行的な研究成果が報告されているだけであった．それによれば日本語文長は対数正規分布に従うという．ただし，これらの研究のほとんどは，複数のテキストから，ごく一部の文章を抽出し，文長として単語の数を手作業でカウントしただけの小規模なデータによる検証にすぎなかった．
(3)しかし現在では，大規模なテキストデータが自由に利用出来る環境が整い，また日本語や欧米語のテキストを形態素解析などにかけて，文の構造を自動的に解析する技術も大きく進歩している．そこで，過去の研究成果は，もう一度総括されてしかるべきであろう．

２．研究の目的
(1)本研究では，過去の欧米語および日本語テキストについての研究成果をまとめる．
(2)次に，日本語テキストおよび欧米語テキストの電子データを収集あるいは整備し，これらを対象に文長の分布と，その確率分布の検定手法を検討する．その際，適合度の検定だけではなく，一般化線形モデルなどの適用を検討する．
(3)同時に，大量のテキストデータから文長を一括計算するためのソフトウェア開発を行ない，完成した場合は，これを広く公開する．

３．研究の方法
(1)日本語文章において長さとは何を意味するかの検討から始めた．欧米語の研究では，「節」や「単語」,「音節」などの単位とすることが検討されていた．日本語でこれらに直接対応する単位として近いのは形態素や句であろう．しかし日本語においては，形態素や句への分割は必ずしも一意には定められない．しかし大量のテキストについてそのすべての文章を手作業で分析するのは．当然ながら実際的ではない．また単位の認定が作業者の主観に左右されることも多い．例えば郵便局は一語なのか，あるいは「郵便」と「局」の二つに分割すべきなのかは，議論が分かれるであろう．この議論は最終的に決着しないと予想されるので，少なくとも，データを分析する個人の主観が排除される形で文長を計測する方法が必要である．
(2)そこで本研究では，現状において最良の形態素解析器として知られる MeCab と CaBoCha の機械的解析結果をそのまま利用することとした．また，これらの解析器と本研究で利用する解析ソフトウェアのインターフェイスを独自に作成した．さらに本研究では文長の単位として，形態素や句の他に，文字数も検討した．
(3)公開されているテキストデータベースなどに加え，新たにテキストを取得し，これを電子化する．こうしてデータを取得ないし整備し，作品情報やルビなどの余計な情報を削除する作業をアルバイトに依頼する．精製したテキストを解析データとして，データベースにまとめる．結果として日本語作品 100 本と，比較のための欧米語テキスト 50 本が収集された．
このデータベースから文長を抽出する．文長の抽出では日本語については，先に言及した形態素解析器とインターフェイスを，欧米語テキストについては TreeTagger を利用した．
(4)次に解析結果に，理論分布をあてはめる．欧米語での研究にならい，頻度の実分布に，対数正規分布およびパスカル分布，負の二項分布を仮定し，その期待値との差についてカイ二乗検定を行った．またカイ二乗検定の統計量を修正した検定についても適用を試みた．
(5) さらには適合度の検定として，カイ自乗分布以外を仮定する手法を検討した．たとえばコルモゴロフの検定である．これらの検定手法は，文長の変動を単に平均からの誤差として説明するモデルであるが，本研究では，文長の誤差変動を説明する要因として，書き手やジャンルなどが検討できないかを追加で検討した．
(6) なお言語テキストを解析し，ここから抽出した統計量に検定を適用するには，独自のプログラムを開発するなどの手間が必要になる．本研究では，分析と並行して，言語解析と統計的検定をシームレスに実行することができるソフトウェアパッケージの開発を行った．

４．研究成果
(1)100 本の日本語テキストについて，文長を，「文字」,「形態素」,「句」で計測した結果について，まずは先行研究にならい，これらの文長データに対数正規分布やパスカル分布，さらには負の二項分布をあてはめてみた．そして，実測値と理論値のズレについても，先行研究にならい，カイ自乗分布による適合度の検定を行った．帰無仮説は「実測値と理論値のズレは偶然であり，二つの分布に矛盾はない」だが，この仮説は，文章の数の少ない一部のテキストを除いて棄却された．同時に文の単位として,「節」を取る場合と「単語」を取る場合で確率分布が異なるという一部研究者の報告についても，そのような結果を認められなかった．
(2)そもそも適合度の検定はデータ数に敏感な手法である．先行研究では，テキストから抽出された文章は数十個，せいぜい 100 個程度であったが，本研究ではテキストごとに数

百から数千の文章を抽出している．有意水準は１％としたが，ことごとく仮説は棄却されている．ところが，テキストの文章を一部だけを取り出し，サンプル数を数十程度におさえると，カイ二乗検定で仮説が保留されることがしばしば起こった．すなわち適合度の検定では，仮定する確率分布に関わらず，対象とする文長のデータ数に，検定結果が大きく左右されるのである．これは，日本語テキストに限らず，本研究で再検討した欧米語テキストについても同様であった．そこで，言語テキストの文長の分布については，文長を文字，単語（形態素），句のいずれを単位としてカウントするにせよ，その分布の誤差，あるいは平均からのズレを，対数正規分布やパスカル分布，負の二項分布などで説明することには無理があると結論付けられる．

(3)そこで次に，文長の平均からの誤差を一般化線形モデルによって説明することを試みた．すなわち文長を目的変数とし，説明変数を無しとする NULL Model をあてはめてみた．この場合，あてはめの程度は遊離度を基準とした．すると，日本語テキスト及び欧米語テキストとも，ポアソン分布あるいは負の二項分布によって，文長の誤差をあてはめることができるテキストが複数確認された．また特定の書き手について，複数のテキストで良好なあてはめ結果がえられたが，その書き手の全テキストに該当するわけではなく，同じ書き手でもあてはめの悪いテキストも多数あった．結果として，一般化線形モデルでのあてはめは有望ではあったが，一般化するには至らなかった．

(4)そこで仮説を変更し，文長の平均からの誤差を確率分布だけで説明するのではなく，書き手やジャンル，あるいは年代などを説明変数として加えることで，文長の分布を説明できるかの検討に移った．一般科線形モデルを NULL モデルから拡張する試みを行ない，単独の説明変数を追加した試行と，複数の説明変数を導入したモデルによる解析を行った．これらの解析結果からは，やはり一般化できる結果はえられなかったが，しかし説明変数としての「書き手」が有意と判定される結果も複数えられた．

(5)ここでアプローチを変えて，逆に文長を説明変数とすることで，書き手を推定できるかを検討した．この場合には，一般化線形モデルに加えて，最近の分類手法を試みた．たとえば判別分析やクラスター分析，さらにはニューラルネットワーク・モデルや自己組織化マップなどの分類手法である．また因子分析などの次元圧縮手法も適用してみた．これらの分析結果のいくつかから，文長によってテキストの書き手を判別する可能性のあることが示唆された．ただし，この成果を一般化できるまでには至っていない．この問題については，いずれテーマを改めて，再度検討したい．

(6)最終的な結論としては，文長の平均からの誤差に対して，一般的にあてはめることのできるような確率分布は明らかにならなかった．そもそもテキストの文長のようなケースでは，一部の抽出標本だけから確率分布を導き出すのは不適切であり，また検定手法としてカイ二乗検定をあてはめることも適切ではない．しかしこの結果は，文長の分布を確率的に説明することが不適切だという結論にはならないと思われる．最終的に一般化するまでには至らなかったが，個別の文の長さを目的変数とした一般化線形モデルは，少なからずケースで適合することが確認されたからである．すなわち文長がまったくランダムに決まっているとも考えにくい．文長では，何らかの要因が分布に影響を与えていることが充分に予測される．逆にいえば，このことからも，文長の分布を単なる平均からの誤差とする過去の研究は適切ではないと言える．

(7)最後に，本研究を通して，日本語統計解析用に二つのソフトウェアパッケージを独自に開発した．RMeCab パッケージは，R という統計解析ソフトと，日本語形態素解析器 MeCab をつなぐインターフェイスであり，指定された日本語テキスト（群）を対象に形態素解析を実行し，その結果を行列として返す．また文字，形態素，品詞を単位とした N-Gram を抽出して返すことができる．一方，RCaBoCha は，R と係り受け解析器 CaBoCha を結ぶインターフェイスである．こちらは形態素解析の他，文を形態素や句に分けた場合の頻度を出力することができる．いずれも，開発後インターネット上に公開しており，誰でも自由にダウンロードして利用することができる．すでに国内外の多くの研究報告や論文，また著書において，この二つのパッケージの利用について言及があり，各種教育機関などでも積極的に利用されている．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計　２　件）

①石田 基広,「言語データの統計解析」，日本語学会 2011 年度春季大会予稿集，54-57 頁，2011 年 5 月 28 日，大阪大学

②石田 基広,「RMeCabとRCaBoChaによるテキストマイニング」，統計連合学会連合大会講演報告集，292-293 頁，2010 年 9 月 8 日，早稲田大学

〔図書〕（計　１ 件）
①石田 基広,『 Rによるテキストマイニング入門』, 森北出版, 全173 頁, 2008 年

〔その他〕
ホームページ等

言語解析用ソフトの開発と公開

形態素解析インターフェイス RMeCab
http://groups.google.com/group/rmecab

係り受け解析インターフェイス RCaBoCha
http://groups.google.com/group/rcabocha

６．研究組織
(1)研究代表者
石田　基広（ISHIDA MOTOHIRO）
徳島大学・大学院ソシオ・アーツ・アンド・サイエンス研究部・准教授
研究者番号：４０２３２３１８

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：