研　究　報　告　書

「言語の計測可能な不変量の探求」

研究タイプ：通常型
研究期間：　平成26年 10 月～平成30年3月
研　究　者：　田中（石井）　久美子

1.　研究のねらい

実データを対象とした工学処理では、数理モデルを仮定して、種々の手法を考案する。しかしながら、数理モデルでは成り立つ数学的性質が、実データでは成立しないことがよくある。このことは、数学的性質の吟味により数理モデルを評価しうることを示唆する。本研究では、複雑さに関する統計量を自然言語の実データとその数理モデルにおいて考えることを通して、自然言語の性質を論じ、実対象の数理モデルを吟味するための方法論を探求する。

昨今では、莫大量の自然言語データがあるとはいえ、その複雑さについては未解明の事項も多く、エントロピーレートやフラクタル次元などの大域的な統計量が探求されてきた。統計量は、データに対して定常性やエルゴード性など、数学的に安定で良い性質を仮定し、しかもデータ量が無限大である時の極限などとして定式化されていることが多い。しかし、実際はデータ量は有限で一回限りであるため、前述の齟齬が生じ、データの性質が明らかとはならない場合がある。

本研究では、実データにおいて量やデータ種に対して同じ値、つまり不変・普遍な値となる統計量を題材とする。特に、ある一定量以上のデータについて収束する統計量を、計測可能な不変量と定義し、手がかりとする。通常は、データ量に対する統計量の収束性は数理モデルにおいてしか吟味されないが、本研究では、有限で一回きりの実データに対しても収束性が成り立つような関数を、それを構成するための数理モデルを含めて探す。得られる計測可能な不変量は、一定量以上の場合には量に依存せず同じ値をとるため、この収束値をもって安定してデータを特徴付ける事が期待される。

本研究では第一に、量や(データ)種に対して不変あるいは普遍な統計量としてどのようなものがあるのか理論的な整理を行う。第二に、人間の自然言語の数量的考察を行い、音楽やプログラムといった関連データとの比較を通して、種々の言語的な実データの複雑さの相貌を考察する。以上を通して、自然言語とはどのような系であるのかを探求し、また、数理的観点から、数理モデルと実データの乖離を考察する方法論を模索する。

2.　研究成果

（1）概要

成果は理論的整理と諸実験に分かれる。本研究のテーマである統計量は自然言語の大域的な性質を表す。このため、自然言語の経験則を網羅的に吟味することを通して研究を進めた。自然言語には、無限性（どんなに大きなデータでも一回限りの語彙が多数含まれる性質）、ならびに自己再帰性(離れた二部分が類似する性質)があるが、いずれも冪として経験則が知られ、それぞれについて理論的・実験的考察を行った。

理論的整理では以下の二つのことを行った。第一に、経験則に関わる複数の統計量の計

測方法を改良し、データ量に対する安定性や言語種にまたがる普遍性を吟味した。まず、無限性については、複雑さを表現する Zipf 則, Heaps 則、ならびに符号化レートの冪減衰の冪指数のふるまいを考察した。また、自己再帰性については、長相関ならびに Taylor 則の、自然言語文書での新しい計量方法を考案した。

第二に、経験則を満たす数理的生成モデルを吟味し、経験則の要因を考察した。自然言語の生成モデルとしては、マルコフモデル、Poisson/生成過程、文法的生成モデル、ニューラル言語モデル、Simon/Pitman-Yor 過程、ランダムウォークなどがある。自然言語が満たす経験則を、以上のどれが満たすのかを吟味し、満たさない場合には満たすようにする方法を探求した。

実験的な研究では、さきがけを通して多量多種の言語的なデータを収集し、理論的整理を通して得られた統計量の値を吟味した。人の言語の複雑さは、これまで定性的な分類がなされ、たとえばチョムスキー階層などとして知られる。本研究で考案した諸統計量を用い、自然言語とプログラムの差など、人の時系列の複雑さを定量的に考察することにつながった。

本さきがけ研究は、JST-RISTEX の「人と情報のエコシステム領域」の「冪則からみる実社会の共進化研究」へと引き継がれ、今後は得られた成果の社会実装へ向かう地点に到達している。特に、理論面で行った生成モデル研究を進めて、よりよい数理モデルを得ることで、より性能の高い工学応用を開拓する。また、本研究の焦点は言語にあったが、研究期間中は統計物理学、哲学、認知科学、プログラミング言語など分野横断的な学術活動を行い、複雑系科学の一環としての数理言語研究のアウトリーチに努めた。

（2）詳細

研究テーマ A　理論的整理：計測可能な不変量の模索と生成モデル

A-1. 自然言語の経験則と普遍・不変量の探求

自然言語には以下の二つの特徴がある。

X.　どのような大きなデータでも、語彙の大部分は1回限りの語である（無限性）

Y.　自然言語のテキストは離れた部分でも比較的類似している(自己再帰性)

このような性質はこれまでに経験則として捉えられてきた。さきがけでは自然言語の以下の経験則を再吟味することから始めた。

a.　Zipf 則とそこから派生する Heaps 則

b.　符号化レート冪減衰

c.　長相関、ならびに Taylor 則

まず、a については、そもそも冪であることが粗い近似に過ぎない。特に子供の発話データや文字データなどでは、順位頻度分布は冪ではなく両対数軸で凸性を示す。このため、数理モデルの改良を行った。Zipf 則に代わるモデルを物理学の相転移を記述する関数を元に提案した。

b については当初 Yule の提案としてのデータ量に依存しない統計量—計測可能な不変量—を手掛かりとして研究を開始した。Plugin エントロピー(要素単位独立とみなし、相対頻度を測度として計測したエントロピー)はデータ量に対して収束し、これが Yule の K と等価であ

ることが見出された。続いて、エントロピーレートの算出に挑んだ。自然言語のエントロピーレートはシャノン以来のテーゼであるが、未だ自然言語のレートが存在するのか、するならば正であるのかは未解決問題である。また Plugin ではない高次エントロピーレートは結局はシャノンレートを計測する問題に帰着される。本研究では、符号化レートをユニバーサル符号を用いて計測した。符号化レートを無限遠点に補完して得られる統計量は、エントロピーレートの上限となっているが、データ量に対して安定な値をとることがわかり、未解決問題に一石を投じた。また、冪指数は、自然言語の種類に依存せず一定の値を示し、その普遍性が実証された。

c については、そもそも自然言語で計量する既存方法にさまざまな問題があった。そこで、自然言語を単語の間隔の時系列に変換し、さらに極値解析を適用することにより、文書に対して長相関を値として計測する新しい方法を考案した。また、研究期間の終盤に、長相関と関係のある Taylor 則を吟味した。Taylor 則の自然言語における計測方法を改良し、既存方法との対比から提案手法を用いる利点を明らかにした。

A-2. 経験則を満たす生成モデルに関する研究

研究の後半から、経験則を満たす生成モデルに関する研究を行った。自然言語の生成モデルには、マルコフモデル、ポワソン・再生過程、文法的生成モデル、ニューラル言語生成モデル、Simon/Pitman-Yor 過程、ランダムウォークが考えられてきた。これらと経験則の関係を解析的ならびに実験的に吟味した。

A-1 の X の無限性を満たすモデルは、ニューラル言語モデル(文字単位)と Simon/PY 過程、ランダムウォークのみである。そのうち、自己相似性をもっているものは、複雑系科学分野で研究されている後二者だけとなり、うち、定性的にすべての冪則を満たすのは、ある特殊の条件下にあるランダムウォークだけであることがわかった。それも、経験則を満たす様相に問題がある。既存モデルの改良が今後課題として残されている。この取り組みを通して、経験則を用いて数理モデルを吟味することができることを示した。

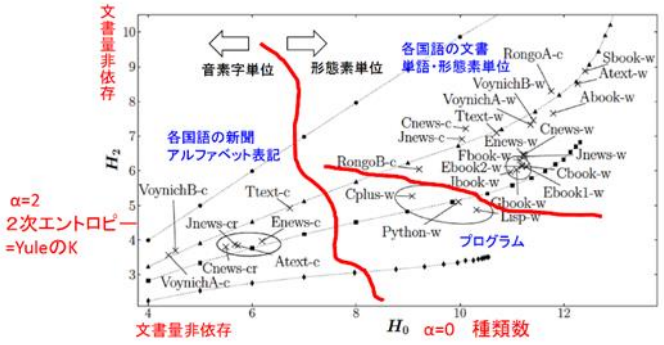

図 1 高次エントロピーによる諸文書の位置づけ

研究テーマ B 実験的吟味:
種々の言語関連データでの
不変量の算出
—言語の複雑さの地図—

テーマ B では、さまざまなデータを整備し、A で整理した統計量を測した。データ種は、実データとランダムデータに分かれる。実データとしては、世界各国語の新聞データ、Gutenberg の長い単著、幼児の長い発話データなど、また、自然言語ではないデータとして、複数の異なるプログラム言語で書かれたプログラムソース、音楽、最終年度には、長相関の比較を行うために、金融市場データも得て

整備した。ランダムデータとしては、A-2 で述べた生成モデルを実装し、擬似データを作成した。　諸データにおいて、A で整理された統計量を計量し、データにおける相貌を考察した。以下に、本報告書では、上記 A-1-b ならびに A-1-c で報告した統計量を用いて得られた結果に限って報告するが、同種の地図は、別の統計量を用いて他にも得られている。

図1は A-1-b で説明した Plugin 2 次エントロピーを、さまざまな自然言語で計測した結果である。横軸は1万要素の中の異なり数(0 次エントロピー)、縦軸が 2 次エントロピーである。文字、単語単位で計測した場合の文書の複雑さが位置付けられている。自然言語の種類では値に差はあまりなく、このため縦横の位置関係が近接している。一方で、文字種で大きな差が見られ、また、プログラムと自然言語の峻別がみられる。

図2は A-1-c で探求した Taylor 指数を利用して、さまざまな文書の揺らぎを調べている。揺らぎは自己再帰性と深く関係することが知られる。縦軸に Taylor 指数、横軸にさまざまなデータ種を示している。Gutenberg や青空文庫は長い単著の文書集合で、Taylor 指数は約 0.58 を平均とする一方で、幼児の発話、プログラム、音楽などでは、まったく異なる平均値をとっている。文法の複雑さに応じて、Taylor 指数の値が大きく変わり、Taylor 指数を用いて複雑さが定量的に峻別できていることがわかる。

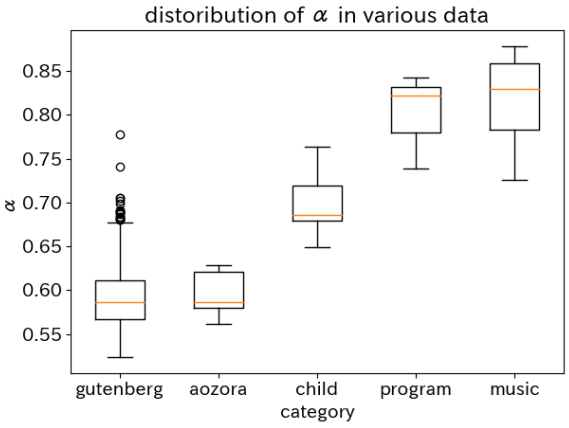

図 2　Taylor 則の指数に基づく地図

総じて、本研究で対象とした統計量は、個別自然文書や言語、ジャンルを峻別するものではなく、より大域的に、自然言語とそれ以外、自然言語内では表音文字と表意文字を峻別するといった粗い区分となる。本研究を通して、人の言葉に関連する時系列をこのように定量的に位置づける試みは過去に例もほとんどなく、独自の結果と思われる。

3.　今後の展開

本さきがけ研究では、現段階では日本ではほとんどまとまった研究がない『自然言語の経験則』についての基礎研究に関するものである。本研究の波及効果としての応用研究は今後見出していくことになる。

一つの大きな方向性として、経験則を満たす数理モデルを考えることで、数理モデルを改良し、応用につなげることがある。実対象を工学処理するうえでは、数理モデルが性能の鍵を握る。そして、現在の数理モデルは、実対象が満たす経験則を満たしていないことが多い。このため、経験則を網羅し、それを満たす数理モデルを考えることは、さまざまな応用システムの性能を向上させる可能性を秘めている。さきがけにおいて行った生成モデルの基礎研究はその布石となるだろう。

さきがけの研究をさらに推し進めるために、2017 年 10 月より、JST－RISTEX の「人と情報のエコシステム」領域において「冪則に基づく実社会の共進化研究」と題するプロジェクトを新たに開始した。複雑系では、分野を俯瞰して経験則を整理することが、対象に関する新しい理解と、効果的な応用につながる。新プロジェクトでは、自然言語に限らず、金融データ、コミュニケーション

ネットワークにおいて、冪則を満たす生成モデルを探求し、社会実装へつなげる。

4. 評価

（1）自己評価

（研究者）

研究目的の達成状況としては、当初、予定していた研究内容を大きく超えて、計測可能な統計量だけでなく、経験則を俯瞰して研究を進めることとなった。経緯として、アドバイザを担当してくださった楠岡成雄先生のご助言が大きい。特に、生成モデルを精査することは、当初は明確には予定に含まれていなかったため、この領域の一員として研究を進められた意義は自分としては非常に大きい。

研究の進め方については、実施体制の構築も研究費執行も順調であった。実施体制に関しては、海外協働も複数件行い、分野の最前線の研究者と未来に向けて協働するきっかけをいただいた。また、さきがけ研究を通して、さまざまなデータを徹底整備し、研究の基盤を構築することができた。研究費の執行は、JST の担当の方々が常に迅速に指示くださり、滞りなく順調であった。

社会・経済への波及効果については、もともと本さきがけでは基礎研究を目指していたため、今後の発展を模索中である。特に 3 に記載したように、現在はさきがけを基礎として社会実装を行うプロジェクトへと移行途中にあり、対象として金融も含まれている。

学術的な波及効果としては、本研究は、対象が言語であるため人文の言語学と、方法論が統計物理学であるため物理学との間の、折衷分野に位置づけられ、その意味であまり例のない研究とはなっている。一方で、既存の学会が無いことが大変な困難であり続けた。世界には同じ目的をもつ研究者が少数いるが、同様の困難を抱えている。しかし、さきがけ研究を通して、これらの研究者らと親交を深め、コミュニティ構築に向けて一歩踏み出すことができた。論文は現在も成果の大きな部分が査読中・執筆中であり、またさきがけ全体の成果を和文・英文書籍として準備中である。さらに、期間中には、統計物理学、哲学、プログラミング言語、認知科学とさまざまな分野の会議からさきがけの内容について講演に招いていただき、分野として確立していないテーマではあるが、諸分野の関心を呼ぶ内容に結実しつつあると考える。

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

（研究総括）

自然言語の性質を統計的普遍性という観点から解明することを目指し，自然言語の経験則に対する理論的研究と，自然言語データやそれに関連するデータの解析と比較による自然言語の特徴の抽出において優れた成果を挙げた．自然言語の経験則に関しては，従来知られていた様々な統計量を再吟味し，統計量の計量方法の改良や冪則からのずれ，種々の統計量の間の関係を明らかにし，自然言語の長相関などの多くの新しい知見を得たことは重要な理論的成果である．またそれらの理論的成果を基に，自然言語データの解析と，様々な他の言語的データとの比較により，Taylor 指数の識別力の発見などの，自然言語に潜む興味深い新しい知見を得たことは高く評価される．

これらは自然言語というユニークな対象に数理的方法によって切り込み，膨大な言語データに対する実直な解析によって得られた，他に類のない独自性をもったインパクトのある研究成果で

あると思われる. さきがけ研究により海外の共同研究者との連携もできて，今後の研究の発展への足がかりができたことは喜ばしい.

研究成果は発表予定の論文の他，様々な分野の国際会議や研究会で発表された. 成果の発信については今後に待つところが大きいが，和文と英文の著書としてまとめる計画とのことで期待される.

5. 主な研究成果リスト

（1）論文（原著論文）発表

1. Kumiko Tanaka-Ishii and Shunsuke Aihara. Text Constancy Measures. Computational Linguistics, 41(3): 481--502, 2015.
2. Kumiko Tanaka-Ishii and Armin Bunde.Long-range memory in literary texts: On the universal clustering of the rare words. PLoS One, online journal, 2016.
3. Ryosuke Takahira, Kumiko Tanaka-Ishii, and Debowski Lukasz. Large scale verification of entropy of natural langauge. Entropy, online journal, 2016.
4. Shuntaro Takahashi and Kumiko Tanaka-Ishii. Do neural nets learn statistical laws behind natural langauge?. PLoS One, online journal, 2017.

現在他 3 件さきがけの成果としての原著論文が査読中.

（2）特許出願

研究期間累積件数：0 件（公開前の出願件名については件数のみ記載）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

招待講演

[1] ``Computational Constancy Measures of Texts'' in Workshop of Language Modeling at IPIAN, 2015, Warsaw.

[2]「言語に内在する大域的な性質---物理的観点から」統計物理学懇談会, 2017, 慶応大学, 日吉.

[3]「プログラムはどうフラクタルか」第 59 回プログラミングシンポジウム招待講演. 2018, 箱根.

国際会議

[1] Kumiko Tanaka-Ishii and Armin Bunde. Rare words appear in clusters. In Conference on Complex Systems, 2016.

[2] Kumiko Tanaka-Ishii and Shunsuke Aihara. Quantitative verification of text constancy measures. In Quantitative Linguistics Conference, 2014.