東京大学 大学院総合文化研究科・教養学部
The University of Tokyo, Komaba
Graduate School of Arts and Sciences, College of Arts and Sciences

2022 年 5 月 27 日

言語と非言語の境界は存在するか？
～自然言語の数理モデルに相転移がないことを証明～

１．発表者：

中石　海（東京大学　大学院総合文化研究科　広域科学専攻　博士課程 1 年）
福島　孝治（東京大学　大学院総合文化研究科　広域科学専攻　教授）

２．発表のポイント：

◆ Random Language Model（注 1）と呼ばれる自然言語（注 2）の数理モデルで相転移（注 3）が起こり得ないことを証明しました。
◆ 先行研究では言語獲得（注 4）を相転移と解釈し、状況証拠に基づいてこの相転移が存在することを予想していましたが、本研究はそれを明確に否定しました。
◆ 本研究の結果は、幼児の無秩序な「言語」から成熟した言語への変化は連続的なものであることを示唆します。本研究は、数理モデルの解析を通して自然言語を理解する研究の可能性を切り拓くものです。

３．発表概要：

東京大学大学院総合文化研究科の中石 海（博士課程 1 年）と福島 孝治教授が、Random Language Model（RLM）と呼ばれる自然言語の数理モデルで相転移が起こり得ないことを証明しました。

DeGiuli は、自然言語をある決まったルールに従って確率的に文字列を生成する系として単純化し、RLM という数理モデルを導入しました。また、彼はこのモデルで相転移が起こると予想しました。RLM はこの相転移によって、無秩序な文字列を生成する領域から秩序だった文法構造を持つ文字列を生成する領域へと転移します。彼はこれを、幼児が秩序だった言語を身につける言語獲得に対応するものとして解釈しました。

しかし本研究は数学的議論に基づいて、この相転移が実際にはあり得ないことを証明しました。この結果を DeGiuli の研究にならって解釈するならば、幼児が母語を身につける過程では、あくまでも連続的な変化しか起こっていないということになります。

RLM は、自然言語を数理モデルの解析と物理現象との対比で理解するという新たな科学的アプローチの可能性を示しました。本研究が相転移の有無という重要な問題に明確な答えを出したことは、このアプローチのもとでなされた最初の具体的成果と言えるでしょう。

４．発表内容：

【背景】

地球上にはさまざまな言語が存在し、そのそれぞれが多様な振る舞いを見せますが、どんな言語であっても必ず文法というルールに従っています。そこで、言語をある決まった規則に従って文字列を生成するものとして単純化し、あらゆる言語を統一的に記述することを試みるのが、形式言語理論です。さらにこの枠組みを拡張し、確率的に文字列を生成する数理モデルを

考えることもできます。このモデルは多数の文字や文法規則が互いに相互作用しながら確率的に振る舞う物理系と見なすことができ、その振る舞いは統計力学（注 5）と呼ばれる物理学の一分野の立場から解析できます。

DeGiuli が導入した RLM は、このような自然言語の数理モデルのひとつです。彼はシミュレーションによってこのモデルを解析し、相転移と呼ばれる現象が起こることを予想しました。相転移とは、なんらかのパラメータを動かしていったときにある点で系の振る舞いが不連続に変わる現象を指します。例えば、日常生活でも見られる相転移としては、温度を下げていくと水が氷になることが挙げられます。水と氷はどちらも水分子がたくさん集まった物質ですが、液体と固体というようにその性質は全く異なります。液体の水では水分子はバラバラですが、固体となった氷では水分子は整然と並ぶ結晶になります。このように、多くの場合新たな秩序の出現を伴うというのも、相転移の大きな特徴です。

DeGiuli の予想によれば、RLM においてモデルの乱雑さを制御するパラメータを連続的に動かしていくと、ある点よりも前では無秩序な文字列が生成され、それよりも後では秩序だった文法構造を持つ文字列が生成されます。DeGiuli はこの相転移が、無秩序で意味をなさない幼児の「言語」が文法に基づく成熟した言語になる言語獲得に対応すると解釈しました。しかし、彼の先行研究はこのような相転移の証拠を明確には示していませんでした。

【研究内容】

DeGiuli の先行研究を受けて、本研究は相転移の有無を確かめるべく RLM をより詳細に調べました。そして、RLM を特徴づける多くの性質が文法構造における文字の出現確率から導けることに気づき、さらに、この出現確率を数学的に解析できることを見出しました。私たちの解析から、先行研究の予想に反して相転移は存在しないことが証明されました。つまり、RLM には確かに無秩序な領域と秩序を持つ領域が存在するのですが、それらのあいだに明確な境界はなく、両者は徐々に移りかわるのです。先行研究にならい、言語獲得に対応づけてこれを解釈するならば、幼児の無秩序な「言語」と成熟した言語のあいだに質的な違いはなく、前者から後者への変化は連続的なものであるということになります。

【今後の展望】

数理モデルの解析と物理現象との対比によって自然言語を理解する DeGiuli の研究は、自然言語に対する新たな科学的アプローチの可能性を示すものとして、物理学者たちのあいだで大きな注目を集めました。RLM について、相転移という物理学的に重要な現象の有無を厳密に確かめた本研究は、このアプローチからの研究として初めて具体的な科学的成果を挙げたと言えます。

もちろん、RLM は単純化されたモデルであり、自然言語の全ての側面を捉えているわけではありません。より多くの側面を反映する RLM よりも複雑なモデルで相転移などの興味深い現象が現れるかどうかは、依然として未知です。今後、より複雑なモデルの提案と解析が蓄積されることで、自然言語の物理学とも言うべき研究が発展していくことが期待されます。

本成果は 2022 年 5 月 27 日（米国東部時間）に米国物理学協会発行の学術雑誌『Physical Review Research』のオンライン版で掲載されました。

本研究は、東京大学 先進基礎科学推進 国際卓越大学院教育プログラムの支援により実施されました。

５．発表雑誌：
雑誌名：「Physical Review Research」（オンライン版：5 月 27 日）
論文タイトル：Absence of Phase Transition in Random Language Model
著者：Kai Nakaishi*、 Koji Hukushima
DOI 番号：10.1103/PhysRevResearch.4.023156

６．問い合わせ先：
東京大学　大学院総合文化研究科　広域科学専攻／附属先進科学研究機構
福島　孝治（ふくしま　こうじ）
E-mail: k-hukushima（末尾に“@g.ecc.u-tokyo.ac.jp”をつけてください）
東京大学　大学院総合文化研究科　広域科学専攻
博士課程 1 年　中石　海（なかいし　かい）
E-mail: nakaishi-kai787（末尾に“@g.ecc.u-tokyo.ac.jp”をつけてください）

７．用語解説：
（注 1）「Random Language Model」
E. DeGiuli, Random language model, Physical Review Letters 122, 128301 (2019).
Philip Ball, Learning Language Requires a Phase Transition, Physics 12, 35 (2019).

（注 2）「自然言語」
日本語、英語など、人間が日常において使う言語。プログラミング言語や数式といった人工的な言語と区別する目的で、この用語を用いる。

（注 3）「相転移」
ランダムに振る舞う多数の要素が集まった系のなんらかのパラメータを連続的に動かしていったときに、ある点で系の振る舞いが質的に変化する物理現象。日常的にみられる相転移には水が氷になる現象がある。多くの場合、相転移には新たな秩序の出現が伴う。例えば、液体の水はランダムに運動する多数の水分子の集まりだが、温度を下げていくとある温度で水分子が秩序だった構造をなし、固体の氷になる。

（注 4）「言語獲得」
幼児が体系的に教わることなく母語を身につけること。

（注 5）「統計力学」
ランダムに振る舞う多数の要素が集まったときに系全体としてどのような性質を持つかを考える、物理学の一分野。