言語モデルのふるまいと多重実現

坪井祥吾 [1]　菅原朔 [2]
[1] 一橋大学　[2] 国立情報学研究所
tsuboishogo98@gmail.com　saku@nii.ac.jp

概要

大規模言語モデルが人間らしい言語的なふるまいを示すようになってきたことを受けて、言語モデルから人間の言語について何らかの示唆を引き出そうとする研究が増えてきている。他方、言語処理のメカニズムが言語モデルと人間とで大きく異なることを根拠に、この手の研究の妥当性が疑問視されることもある。本稿では、この議論の文脈に、科学哲学における「多重実現」という考え方を導入する。本稿は、この考え方を導入することで、この種の研究手法の妥当性を評価する際に考慮するべき、哲学的・概念的な論点を浮き彫りにすることを目指す。

1　はじめに

大規模言語モデルが人間らしい言語的なふるまいを示すようになってきたことを受けて、言語モデルから人間の言語獲得や処理の仕方について何らかの示唆を引き出そうとする研究が増えてきている [1, 2, 3]。こうした動向について、例えば Warstadt and Bowman [4] は、言語モデルを用いた実験的研究を行うことで、言語学上の対立（e.g., Nature vs. Nurture）を調停するような証拠を提供しうると主張している。他方で、こうした主張に対しては、そもそも既存のベンチマークでは言語モデルと人間の類似性を正確に測れていないとする反対意見 (Vázquez Martínez et al. [5]) や、人間と言語モデルの言語処理メカニズムがあまりにも異なるとするような反対意見 (Cuskley et al. [6]) などがある。

以上のようにして、言語モデルから人間の言語への示唆を引き出そうとする研究は、言語モデルと人間の相違を根拠にして批判される。だが、そうした相違点が正確なところどのようにして問題になるのかは、それほど明らかではない。裏を返せば、どのような条件が満たされれば、人間とは様々な点で異なる言語モデルから何らかの示唆を引き出せるのか（あるいはそのような条件などないのか）に関して、概念上の不明瞭さが残っている。

そこで本稿では、言語モデルから言語学への貢献可能性に見通しを与えるために、哲学的な観点から議論する。本稿で着目するのは、科学哲学という領域において 1960 年代頃 [7, 8] から取り上げられてきた、**多重実現 (multiple realization)** なる概念である[1)]。このやや古い概念を持ち出す理由は二つある。一つの理由は、多重実現という概念を適用できる事象の範囲が非常に広いということである。よって、この概念を用いることにより、言語モデルについての議論をより一般的な観点から捉え直せるようになる。もう一つの理由は、それが古い概念であるがゆえに、多重実現なる概念についての科学哲学的な議論に十分な蓄積があるということである。これにより、言語モデルについて論ずる際に、そうした蓄積から何らかのヒントを得ることが期待できる。

2　言語モデルを言語学的探求に使う

本節では、議論の前提として、言語モデルを用いた言語学的研究の例とその特徴を確認する。

まずは、そもそもなぜ、言語学的研究に言語モデルを用いるのかという点に触れておく。端的に言えばそれは、人間に対する介入的な実験が、技術的に、あるいは倫理的に困難な場合が多いからである [4, 10]。例えば Leong and Linzen [10] は、実験参加者の学習環境を完全に統制することが現実的に難しいという問題を指摘している。また Warstadt and Bowman [4] は、実験参加者の子供に対して、（生得的能力の有無を確かめるために）生まれてから言語に全く触れないように介入することなどは倫理的に許されないという問題を指摘している。対して言語モデルを用いた実験的研究は、こうした問題を回避できる。それゆえ、仮にそれが方法論的に適切なものであるのならば、言語モデルを用いた実験的研究

1)　類似した論点として、McGrath and Russin [9] は、深層学習モデルが認知科学において重要な役割を果たしうることを多重実現に関連づけて論じている。

は有益な手法と言えるのである。

続いて、この手の実験的研究の具体的な例を見よう。一例として、Leong and Linzen [10] は、動詞の受動態化可能性に関する研究を行っている。そこで行われたテストの一つは、容易に受動態化されうる動詞を含む文に対する言語モデルの容認性判断を、そのモデルをあるコーパスで学習させた場合と、そのコーパスから問題の動詞の受動態での出現頻度を大きく減らしたデータセットで学習させた場合とで比較する、というものである。これにより、動詞の受動態化可能性をモデルがいかに学習しているのかの因果的な要因を調べている。Leong and Linzen はさらにこの結果が、受動態化可能性の人間による学習の仕方にも一般化できると示唆している。

本稿での議論を他の研究にも適用できるものとするために、この例を、より一般的な形で記述し直しておこう。大まかに言えば、この種の実験的な研究は、(i) 言語モデルに関して因果的な推論を行った上で、(ii) その推論の結果を人間に外挿している。まず (i) のステップにおいて、興味がある言語的ふるまいの測定結果を変数 B で表すことにすれば、実験では、そのふるまいを生じさせているかどうかを知りたい要因 F に介入して、B の値が変化するかを調べることになる。すなわち、実験対象となる言語モデルにおいて、F が B の原因であるかどうかを調べようとしているわけである。続いて、(ii) のステップにおいて、F が B を引き起こしている（あるいは、F と B は因果的に独立である）という言語モデルについての事実を、人間に外挿する[2)]。

本稿で主に取り上げるのは、(ii) の、言語モデルから人間への外挿にかかわる論点である。次節で導入する多重実現というものが、この外挿の段階で重要な役割を果たすと 4 節で主張する。

3　多重実現

本節では、多重実現という概念を導入する。はじめに断っておくと、多重実現の正確な定義についてはいまだ哲学者らは係争中である。本稿では Polger and Shapiro [11] による学説に依拠して議論を進めるが、これが唯一の学説というわけではない[3)]。

多重実現とは何であるかを言うためには、まず実現 (realization) とは何であるかを述べておく必要があろう。実現関係とは、一つの機能と、それを成り立たせているところの基盤となるシステムとの間の関係である [11]。例えば、コルク抜きという道具を考えよう。この道具は、ワインボトルからコルクを引き抜くという機能を持つ。そしてこの機能が成立するのは、螺旋状の針があり、それに持ち手がついており……といった具体的な構成のおかげである。すなわち、コルクを引き抜くという機能が、コルク抜きの構成要素の配列によって実現されているのである。

多重実現というのは、最も素朴に特徴づけるならば、上記のような実現関係が「多重に」成立することである。例えばコルク抜きには、ウイング式やソムリエナイフ式といった複数の様式がある。それらの様式は、どのような部品から構成されるか、部品がどのように配置されるかという点で大きく異なる。すなわち、コルクを抜くという機能は、ウイング式の機構とソムリエナイフ式の機構の両方によって多重実現される、ということである。以上のことを Polger and Shapiro [11] は、「**同じだが違う (same but different)**」と特徴づける。同じ機能だが違う基盤を持つ、というのが、多重実現の基本的な要件の一つだということである。

だが、この素朴な特徴づけだと、ある種の興味深い仕方での多重実現と、そうではない仕方での多重実現が区別できない。例えば、赤色のウイング式コルク抜きと青色のウイング式コルク抜きを考えよう。これらは、色の点では異なるが、両者ともコルクを抜くという同じ機能を実現しているため、上の素朴な意味での多重実現の実例となる。だが、両者の違いは、問題の機能にとって関係のないものである（何色であろうとコルクは抜ける）。よってこの違いは、コルクを抜くという機能の実現という観点からは、無視すべきであろう。これに対して、ウイング式かソムリエナイフ式かという違いは、その機能の実現という観点から注目に値すると言うべきである。Polger and Shapiro [11] はこの点を、「**関連する仕方で違う (relevantly different)**」と特徴づける。多重実現のためには、機能を実現する基盤の違いが、その機能に関連したものでなければならないのである。

ただし、機能に関連する違いがありさえすれば多重実現が成立すると言うべきかというと、そう

2)　1 節で触れた、既存のベンチマークではモデルの能力を正確に測れていないとするような批判 [5] は、(i) に対するものと言える。他方で、人間と言語モデルのメカニズム的な相違点を強調するような批判 [6] は、(ii) に対するものと言える。

3)　論争の概略については Stanford Encyclopedia of Philosophy の記事 [12] を参照。太田 [13] や山崎 [14] も見よ。

ではない。さらなる細かい注意が必要である。例として、ハンドルの長さが 10cm のコルク抜きと、15cm のコルク抜きを考えよう。ハンドルの長さの違いは、コルクに対してかかる力の違いに影響するため、コルクを抜くという機能に関連している。だが、やはりこの違いも、当該の機能に関しては無視すべき違いと言えよう。ハンドルの長さが一定の範囲に収まってさえいれば、コルクを抜くという機能の実現は同様に果たされるからである。言い換えれば、コルク抜きにとって、ハンドルの長さは個体差に過ぎないのである。Polger and Shapiro [11] はこの点を、「**違う仕方で同じ (differently the same)**」と特徴づける。同じ機能が異なる仕方で実現されるということもまた、多重実現の重要な要件として挙げられるのである。

多重実現は以上のようにして、(1) 同じだが違う、(2) 関連する仕方で違う、(3) 違う仕方で同じ、という三つの要件によって特徴づけられる。ここで、これらの要件に沿って、以下のようにして、興味ある機能が実際に多重実現されているかどうかを判断するための大まかな方針を与えることもできよう[4)]。

まず (1) の「同じだが違う」については、二つの異なるシステムが与えられたときに、そもそもそれらが何らかの意味で同じ機能を果たしているのかどうかを調べる必要がある。それらシステムが同じ機能を果たしていれば、それらは多重実現の候補となるし、逆に果たしていなければその時点で多重実現の候補からは外れる。

次に (2) の「関連する仕方で違う」については、二つの、同じ機能を果たす異なるシステムが与えられたときに、それらの違いが、当の機能を果たすことに関連しているかどうかを調べる必要がある。そこに関連性があれば、それらは多重実現の候補となるし、逆に関連性がなければ、多重実現の候補からは外れることになる。

(3) の「違う仕方で同じ」については、二つの、同じ機能を関連する違う仕方で果たすシステムが与えられたときに、それらの違いが、単なる個体差ではないことを調べる必要がある。もしその違いが単なる個体差ではなければ、それは多重実現のよい候補となるし、逆に単なる個体差であれば、それは多重実現の候補ではない。

4) Polger and Shapiro [11] は、具体的な分類体系 (taxonomic systems) を参照することで、下記の方針における「同じ」や「違う」や「関連する」といった語の内実が確定するとしている。

4 言語的なふるまいの多重実現

前節で導入された多重実現は、本稿で問題としているような、言語学的な仮説を検証するために言語モデルを利用するというタイプの研究においても重要である。というのも、まず、言語的なふるまいは多重実現される可能性があり、そして仮に前節の三つの要件を満たすような仕方で実際に多重実現されているとすれば、その種の研究から引き出せる結論が制約されるからである。

それでは、この種の研究において、多重実現がどのように関わってくるのかをもう少し正確に見てみよう。この種の研究では、例えば、与えられた文の文法性を判断するというタスクを、何らかの言語モデルが人間と遜色ない水準でこなすということから、そのタスクをこなすためにはしかじかの要因が必要／不要である、というような議論がなされる。ここで、言語的なふるまいは一種の機能として捉えられる。ここでの例では、その言語的なふるまいは、「文が与えられたときにそれが文法的であるかを判断する」という機能である。そしてこの機能は、人間の認知メカニズムによっても実現されているし、言語モデルによっても実現されているように見える。従ってこの言語的なふるまいは、少なくとも素朴な意味では、多重に実現される可能性がある。

だが、素朴な意味での多重実現の可能性があるだけでは、現実の研究に対してはそれほど大きな影響は生じない。問題となるのは、前節の三つの要件を満たすような仕方で実際に多重実現が生じる場合である。だが、具体的にはどのような影響があるのだろうか。この点をはっきりさせるために、何らかの言語的なふるまい B が、人間と言語モデルとで三要件を満たす仕方で多重実現されているとしよう。加えて、言語モデルを用いた実験的な研究によって、要因 F がふるまい B と因果的に関連していることが判明したとする。だがこのとき、この F と B の間に因果関係があるという結果が人間についても成り立つ、ということは保証されない。というのも、三要件が満たされているということは、ふるまい B は言語モデルと人間とで本質的に異なる仕方で実現されているということであり、したがって人間に関しても F が B の原因となるかどうかは全く明らかではないからである。以上のようにして、三要件を満たす多重実現は、もし実際に成り立っているとする

と、言語モデルから人間への外挿の段階で大きな障壁となるわけである。

したがって、言語モデルを用いた言語学的研究にとって、三要件を満たす仕方での多重実現が成り立っているかどうかが問題となる。そして、この手の研究に対する、1 節で触れたような批判は、まさしくこの多重実現に着目したものとして捉え直せる。さらに、やはり多重実現に着目することによって、そうした批判に対抗して、言語モデルから人間の言語習得等への示唆を引き出すためにどのようなことを示すべきであるかの指針も見えてくる。大筋としては、人間と言語モデルの間に大きな違いがあることは当然のこととして認めつつ、両者の違いが本質的には無視できるようなものであるということを示すのを目指すべきである。以下では、多重実現の三つの要件に沿って、この点を詳述する。

まず (1) の、「同じだが違う」という基準が満たされて**いる**ことをチェックする必要がある。特に、人間と言語モデルが「違う」ものであることは当然であるから、両者が重要な意味で「同じ」ふるまいをしていることを確かめなければならない。ここでは、人間と言語モデルの言語的ふるまいが「同じである」というための基準を設定することが問題となりうる。例えば、言語モデルのふるまいが入力する文の形式や言語に依存して変動するということを鑑みれば、恣意的に高いハードルを設定して、言語モデルと人間とでふるまいに関して不一致があると結論するのは妥当ではないだろう。

(1) が満たされたとして、続いて (2) の「関連する仕方で違う」という基準が満たされて**いない**ことを示すことができれば、言語モデルから何らかの示唆を引き出すことがある程度は正当化される。すなわち、人間と言語モデルでたしかに同じ言語的ふるまいが実現されているとして、その実現のされ方の違いが、当のふるまいに関連していないということを確かめればよい[5)]。この段階では、人間と言語モデルの内部機構の間の単なる違いを超えて、当該の言語的ふるまいに関連するような違いが問題となる。このことを踏まえると、例えば、言語モデルの内部の重みは不変であるのに対して、人間の内部的なメカニズムは動的であるという違いを指摘し、従って言語モデルは人間の言語のモデルとしては不適当だとするような批判 [6] にも、再反論の余地があることが分かる。というのも、重みの不変性は当の言語的ふるまいにとって関連しない可能性が残っているからである。前節での例との類比で言えば、その違いが、様式の違い（ウイング式かソムリエナイフ式か）ではなくコルク抜きの色の違い（赤色か青色か）の方により似ていることを示すのを目指すべきだということである。

望みに反して (2) が満たされてしまっているとしても、(3) の「違う仕方で同じ」という基準が満たされて**いない**ことが示されれば、やはり言語モデルから人間の言語習得等に関して一定程度正当な仕方で何らかの主張をすることができる。つまり、人間と言語モデルの間の違いが、単なる個体差と言ってよいような違いだということを確かめられればよい。ここで問題となるのは、人間と言語モデルの間の、ふるまいに関連するような単なる違いではなく、ふるまいに関連しつつ、しかもただの個体差とは言い難いような重要な違いが両者の間にあるかどうかである。前節の例になぞらえれば、その違いが、様式の違いではなくハンドルの長さの違いの方に似ていると主張する必要があるのである。

以上で、言語モデルを言語学的研究に応用するような手法を正当化するために検討するべき論点がいくらか明確になったはずである。つまり、同じ言語的ふるまいが人間と言語モデルという違うシステムによって実現されているとして、その違いがふるまいに関連しているのか、単なる個体差以上のものであるのか、といったことに注意を払うべきなのである。ただし、そうした「同じさ」や「違い」や「関連性」については、さらなる議論が必要であろう。

5 おわりに

本稿では、言語モデルを用いた言語学的研究の正当性についての議論を、多重実現という観点から論じた。まず、その手の研究手法は、言語モデルから得られた知見を人間に外挿するというタイプの手法であることが確認された（2 節）。その上で、多重実現という科学哲学の概念が導入された（3 節）。そしてその概念を用いて、問題の研究手法が正当化されるための条件を検討した（4 節）。検討を通じて、言語モデルと人間のふるまいが「同じである」ということや、そのふるまいの実現のされ方が「違う」ということ、その違いが「関連的である」ということなどについて、哲学的な議論が不可欠であることが浮き彫りになった。

5） さらに、人間と言語モデルそれぞれにおける実現のされ方が、その同じ言語的ふるまいに関連する仕方で同じであることまで言えるのが理想的である。

謝辞

本研究は JST 創発的研究支援事業 JPMJFR232R の支援を受けたものです。

参考文献

[1] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 625–641, 2019.

[2] Shammur Absar Chowdhury and Roberto Zamparelli. RNN simulations of grammaticality judgments on long-distance dependencies. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, **Proceedings of the 27th International Conference on Computational Linguistics**, pp. 133–144, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[3] Isabel Papadimitriou and Dan Jurafsky. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6829–6839, Online, November 2020. Association for Computational Linguistics.

[4] Alex Warstadt and Samuel R. Bowman. What artificial neural networks can tell us about human language acquisition. In Jean-Philippe Bernardy Shalom Lappin, editor, **Algebraic Structures in Natural Language**. Taylor & Francis Group, 2022.

[5] Héctor Javier Vázquez Martínez, Annika Heuser, Charles Yang, and Jordan Kodner. Evaluating neural language models as cognitive models of language acquisition. In Dieuwke Hupkes, Verna Dankers, Khuyagbaatar Batsuren, Koustuv Sinha, Amirhossein Kazemnejad, Christos Christodoulopoulos, Ryan Cotterell, and Elia Bruni, editors, **Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP**, pp. 48–64, Singapore, December 2023. Association for Computational Linguistics.

[6] Christine Cuskley, Rebecca Woods, and Molly Flaherty. The limitations of large language models for understanding human language and cognition. **Open Mind**, Vol. 8, pp. 1058–1083, 08 2024.

[7] Hilary Putnam. Psychological predicates. In William H. Capitan and Daniel Davy Merrill, editors, **Art, mind, and religion**, pp. 37–48. University of Pittsburgh Press, 1967.

[8] Jerry A. Fodor. Special sciences (or: The disunity of science as a working hypothesis). **Synthese**, Vol. 28, No. 2, pp. 97–115, 1974.

[9] Sam Whitman McGrath and Jacob Russin. Multiple realizability and the rise of deep learning. **Proceedings of the Annual Meeting of the Cognitive Science Society**, Vol. 46, , 2024.

[10] Cara Su-Yi Leong and Tal Linzen. Testing learning hypotheses using neural networks by manipulating learning data. arXiv preprint 2407.04593, 2024.

[11] Thomas W. Polger and Lawrence A. Shapiro. **The Multiple Realization Book**. Oxford University Press, 06 2016.

[12] John Bickle. Multiple Realizability. In Edward N. Zalta, editor, **The Stanford Encyclopedia of Philosophy**. Metaphysics Research Lab, Stanford University, Summer 2020 edition, 2020.

[13] 太田紘史. 経験科学における多重実現と多様性探求. 哲学論叢, Vol. 33, pp. 1–12, 2006.

[14] 山崎かれん. 実現関係を論じる二つの立場とその対立. 哲学・科学史論叢, Vol. 22, pp. 73–94, 02 2020.