言語を複雑さを「測る」ことの正体*

－測度論基盤の目論みと測る対象としての言語－

中山拓人

1. 序論

「言語は等しく複雑である」といういわゆる言語の等複雑性は、一般的な言説だと言えるほど、言語学において広く流布している言説である。その一方で、これは実際に検証を経て提唱された主張という訳ではなく、「その傾向があるようだ」という、経験則に基づく言説に過ぎない。長い間、人の手で全ての言語の複雑さを調べることは不可能であったため、この言説の検証は取り組み得ないものであったが、近年のコンピュータ技術の発達によって、言語学者はようやく検証可能性を手に入れることとなり、これまで複雑さの計測方法が数多く提案されてきた。同時に、「言語の複雑さとは何か」という問いに関しても、広く議論され始めたが、複雑さの定義や、その妥当な計測方法に関しては、未だ確立しているとは言い難い。

障壁となる課題は幾つもあるが、その中の一つとして、「測る」という操作についての厳密な定義、及び議論が行われていないことが指摘できる。先行研究では、「言語の複雑さを測る」という目標に対して、「言語の複雑さをは何か？」についての議論は多く行なってきており、その問いを出発点に、計測手法の考案を行なってきた。その一方で、「測る」という操作そのものに関しては議論がなく、例えば言語の複雑さ自体の計測可能性についてさえ、議論無しに自明のものと考えられている様である。そこで本稿が目的とするのは、先行研究とは異なる方向からのアプローチを試みる。即ち、「測る」という操作とは何であるのかを定義した後、その操作の要求に従う形で、言語そのもの、および言語の複雑さとは、どの様なものであるかを考える。本稿では、この測るという操作を、数学における測度論によってその定義を求め、それを基盤とした言語そのもの定義を考え、言語の複雑さを考察する基盤を構築することを目的とする。

以下は、次のような構成である。2 節では、従来提案されてきた複雑さの計測手法の紹介と批判的検討を行う。3 節では、測るという操作を、素朴に考察を通じて整理した後、標準的な測度論の概念を導入し、それに従う様な言語の理論的モデルを提案する。4 節では、言語の非加法性を失わない形での定義を求めて、Fuzzy 測度論の導入、及びそれに基づく言語そのものの定義、そして言語の複雑さに関して考察する。

2. 「言語の複雑さ」とは何か？

言語の等複雑性の検証を念頭に、これまで多くの計測方法が提案されてきた。特に 20 世紀の終わり頃から、コンピュータ技術の登場を受けて、より大規模なデータの分析が可能に

* 本稿は、JSPS 科研費 JP24KJ1938 の助成を受けたものです。

なってきている。その様な状況下において、多くの先行研究が、「言語の複雑さとは何か？」という問いにはアプローチしてきたと言える。例えば、Çöltekin and Rama (2023) は、そのタイトルが表す通り、これまで提案されてきた形態論的な複雑さの計測尺度に関して、類似点と相違点、及びそれらがどんな類型論的特徴を捉えるものであるかを評価している。彼らは、type/token 比、テキストファイルの圧縮ツールを用いたコルモゴロフ複雑性の近似、単語列のエントロピー、屈折パターンの数、レンマからの屈折予測難易度などを挙げ、各尺度の計測対象について整理、考察を行なった。これらの計測手法は、言語をどのように捉えた計測をするかによって大別可能であり、特に規則基盤尺度を用いる手法と統計基盤尺度を用いる手法の２つに分類できる。

表１　規則基盤尺度と統計基盤尺度

規則基盤尺度	統計基盤尺度
音素、形態素の数 音韻規則の多さ 語彙の変化形の多さ 屈折の多さ …	TTR（type/token ratio） コルモゴロフ複雑性 エントロピー …

規則基盤尺度とは、規則の豊かさ、言い換えると記述の難しさを、複雑さとして解釈する尺度である。例えば、語形変化の豊かさは、代表的な規則基盤尺度の１つである。ある語が、平均して幾つの変化形を有するかを数え、それが多い言語ほど複雑であるとする尺度である。Baechler (2016) は、ドイツ語の地理的、及び時代的変種を対象に、名詞、形容詞、冠詞の変化形の豊かさを計測し比較した。ドイツ語は上記の３つの品詞に関して、文法性と数による変化があり、その変化形が多いほど複雑であるという尺度の定義は、直観的に理解しやすい。結果として、Baechler (2016) は、歴史的に単純化する傾向があること、そして閉鎖的なコミュニティにおける変種の方が複雑さが高い傾向があることを示唆した。同様に、動詞の屈折の豊かさも規則基盤尺度に分類できる。

一方で、統計基盤尺度とは、特定の言語の統語や形態論的な規則に依らない、例えば文字系列としての言語における統計的性質を、複雑さとして解釈する尺度である。TTR（type/token ratio）は、統計基盤尺度の１つであると言える。TTR は、ある文書において、異なる単語の数を、全単語数で割ったものである。この尺度は、文書内の語彙の豊かさを示すものであり、この値が高いほど語彙が豊かである、即ち複雑であるとされる。Çöltekin and Rama (2023) は、この手法は形態論的な複雑さを測る尺度であり、また最も直接的で理解のしやすい尺度の１つであることを指摘している。

その他の統計基盤尺度として、情報理論を背景とした尺度も多く用いられてきた。Juola (1998)、Juola (2008) は、コルモゴロフ複雑性を基盤とした複雑さの計測方法を提案し

た。コルモゴロフ複雑性とは、文字列を実現するための最小アルゴリズム長による尺度であり、より短いアルゴリズムで表せる文字列ほど、複雑さが低いと定義されるものである。具体的な実装として、zip ファイル化した時のバイト数の圧縮率がそれに当たるため、調査ではこの手法が主に使われる。また Juola（1998）、Juola（2008）は、完全な状態の文書と、一部をランダムに削除したファイルの圧縮後のバイト数を比較することで、形態論的、及び統語的複雑さの計測方法も提案した。ランダムに文字を削除したファイルを比較に用いた時、圧縮後のバイト数の差が少ないほど、文字単位の並びに規則性が少ないということが言えるため、形態論的により複雑であると言える。同様に、ランダムに単語を削除したファイルを比較に用いた時、圧縮後のバイト数の差が少ないほど、単語単位の並びに規則性が少ないということが言えるため、統語論的により複雑であると言える。

Ehret and Szmrecsanyi（2016）は、このコルモゴロフ複雑性を用いて、統合と形態論的な複雑さに関する、英語の通時的な複雑さの変化、及び多言語間比較を行った。結果として、その両方の調査において、形態論的複雑さと統語的複雑さの間にトレードオフの関係があることが示唆された。彼女らはこの手法に関して、“we［…］approximate Kolmogorov complexity and thus to assess lin- guistic complexity on the overall, syntactic and morphological plane.”（75）と述べている。

また、コルモゴロフ複雑性とは別の尺度を用いた研究もある。シャノンの情報理論を基盤とした研究では、情報エントロピー（Shannon, 1948）が複雑さの尺度として採用されることが多く見られる。シャノンの情報エントロピーは、ある事象がどれだけ予測しづらいかの尺度である。n 個の事象があり、i 番目の要素が発生する確率が p_i である時、エントロピー H は通常以下の公式で定義される。

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

実際の適用のあり方としては、形式系列（e.g., 文字列、単語列など）の予測不能性を計測することが、多く用いられる。ある形式系列が与えられた時、次に何の要素が続くがの予測しづらさを、情報エントロピーを用いて定量化することで、言語の複雑性を計測することができる。Koplenig et al.（2023）は、コーパスデータを使って、上記の手法にて各言語の情報エントロピーを算出し、さらにそれを複数のコーパスに渡って行った。結果として、あるコーパス内で見られたエントロピー差が、他のコーパスでも同様に見られたことから、言語の複雑さが、言語間で有意なものであるということを示唆した。Bentz et al.（2017）も、同様に情報エントロピーを用いた調査を、単語を対象に行っている。結果として、非常に狭い範囲に単語のエントロピーが収まっていることが観察され、Koplenig et al.（2023）とは反対に、言語の等複雑性を示唆するような結果となった。

これまで挙げてきたように、コンピュータを用いた大規模なデータ処理に基づく複雑さの計測は、多く行われてきた。そしてその中では、「言語の複雑さとは何か？」という問いを出発点として、言語の複雑さへのアプローチがなされてきたことがわかる。測るという操

作自体に目を向けたと言える研究も、数は少ないが存在している。Deutscher (2009) は、一つの提案として、ベクトルを用いる手法を示している。彼は、単一の値による複雑さの評価が困難であり、それに伴い、多次元ベクトルとして複雑さを表現した上で比較する手法を提案している。Bentz et al. (2023) はこの考えに則り、複数に渡って言語の側面 (e.g., 音韻、形態論、統語) に関する複雑さあをワークショップにて多くの人手を使って評価し、それらを多次元ベクトルとして表現することで、言語全体の複雑さの比較を行った。最終的にこの研究は、各言語の間に有意差が見られなかったことから、言語の等複雑性を示唆する結果となった。確かに多次元ベクトルを用いる手法は、複数の側面を同時に評価できる点で、特に言語の多角的な複雑さを評価するのに有効ではある。しかし、各次元に格納される値に関して、測るという操作についての議論を行なっていない以上、上記で指摘してきた問題を回避できるとは言えない。勿論、次元数の増加に伴い、本節が提起する問題を解決できるという望みはあるが、それでもどれだけの次元数があれば十分出るのか自明ではなく、やはりその意味で、測るという操作の厳密な定義が不必要になるわけではないことがわかる。

以上のように、先行研究では、言語のというものの性質や、計算手法といった要素に関して議論を重ねており、「言語の複雑さとは何か？」という問いには多くの関心が向けられてきたことがわかる。その一方で、「測るとはどの様な操作か？」という問いは、ほとんど目を向けられてこなかった。次節では、測るという操作に関する考察を行うと同時に、言語の複雑さを定義するためには、むしろこの概念こそが出発点になるべきであることを、測度論の導入を交えながら示す。

3. 「測る」とはどの様な操作か？

3.1. 素朴な「測る」

「測る」という操作を考えるために、まず、日常にありふれた概念を例に、素朴な思弁を行う。ここでは例として、「鉛筆の長さを測る」という操作が、どの様な性質をもっているかを考える。

- **鉛筆の長さ**を測ると、一方が他方より長い、ということがわかる。
- 鉛筆を削り切ってしまうと、長さは 0 である。
- ある鉛筆を適当な位置で切断してできた 2 本の鉛筆に関しても、長さを測れる。
- その 2 本の**鉛筆の長さ**を足し合わせると、元の**鉛筆の長さ**に一致する。

以上に挙げたものは、全て当然のことの様に思えるが、同様の性質が、例えば「水の体積」でも言い換えが可能である。

- **水の体積**を測ると、一方が他方より大きい、ということがわかる。
- 水を飲み切ってしまうと、体積は 0 である。
- ある容器に入った水を適当な量で別の容器に分けてできた 2 つの水に関しても、体積を測れる。

・ その 2 つの**水の体積**を足し合わせると、元の**水の体積**に一致する。

これも至極当然の命題ばかりであるが、鉛筆の長さと水の体積に関する、「測る」という操作には、何らかの共通項があることが伺え、またそれは他にも、例えば何らかの「面積」にも同様に適用可能である様に考えられる。実際に、測るという操作を一般化させたものが、次節で導入される測度論と呼ばれる数学の理論体系である。

言語の複雑さも、測度論的な定義を施せば、上に見てきた長さや体積と同じ様に、以下の様に扱える素朴な「測る」操作の対象として定義することが可能である。

・ **言語の複雑さ**を測ると、一方が他方より複雑だ、ということがわかる。
・ 言語が空であったら、その複雑さは 0 である。
・ ある言語を適当に分割すると、できた 2 つの言語に関しても、複雑さを測れる。
・ その 2 つの**言語の複雑さ**を足し合わせると、元の**言語の複雑さ**に一致する。

勿論、このままでは「言語の複雑さ」や「言語」そのものが一体何を指しているのか、さらに言うと、「ある言語を適当に分割する」、「言語の複雑さを足し合わせる」とはどういうことなのか、全くの不明瞭である。次節では、まず測度論の標準的な定義を導入した後で、それを基盤とした時に、これらの用語が指す対象が、理論的にどの様なものであることが要求されるかについて、考察していく。

3.2. 測度論の導入

測度論（measure theory）は、前節で挙げたような、長さや体積といった測るという操作を、「測度（measure)」として一般化している。以下ではまず、測度の標準的な定義を与える[1]。

定義 3.1. 標準的な測度論
ある空でない集合 S に対して、その部分集合の集合族 $\mathfrak{M}$ が以下の性質を満たすとき、$(S, \mathfrak{M})$ の対を、可測空間という。

・ $\emptyset \in \mathfrak{M}$
・ $A \in \mathfrak{M} \Rightarrow A^c \in \mathfrak{M}$
・ $A_n \in \mathfrak{M}\,(n = 1, 2, \ldots) \Rightarrow \bigcup_{n=1}^{\infty} A_n = A_1 \cup A_2 \cup \ldots \in \mathfrak{M}$

この時、可測空間 $(S, \mathfrak{M})$ に対して、$\mathfrak{M}$ 上で定義された関数 μ が以下を満たすとき、μ を $(S, \mathfrak{M})$ 上の測度と呼び、$(S, \mathfrak{M}, \mu)$ の組を測度空間と呼ぶ。

・ 任意の $A \in \mathfrak{M}$ に対して、$0 \leq \mu(A) \leq \infty$ 特に $\mu(\emptyset = 0)$
・ $A_1, A_2, \ldots \in \mathfrak{M}\,(\sqcap_{n=1}^{\infty} A_n = \emptyset) \Rightarrow \mu(\sqcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$

[1] 原（2017）

この定義が主張していることは、測るという操作が、特定の条件を持つ集合族上に定義された、実数に対する関数であるということである。具体例として、再度、鉛筆の長さを例に、この定義について考える。

定義 3. 2. 測度としての鉛筆の長さ
「ある空でない集合 S」に相当するものを、0cm から Pcm までの閉区間 $[0, P](\subset \mathbb{R})$cm とし、その部分集合の集合族 $\mathfrak{M}$ に相当するものを、$[0, P]$ のボレル集合族とすると、以下が成り立つ。

- $[0, 0] \in \mathfrak{M}$
- $[0, x] \in \mathfrak{M} \Rightarrow [x, P] \in \mathfrak{M}$
- $A_n \in \mathfrak{M}\,(n = 1, 2, \ldots) \Rightarrow \bigcup_{n=1}^{\infty} A_n = A_1 \cup A_2 \cup \ldots \in \mathfrak{M}$

第 1 点目は、長さが 0 である鉛筆が定義されていることを表しており、第 2 点目は「ある鉛筆の長さを測れる時、それを引いた長さもまた測れる」と言うことを意味している。第 3 点目が表しているのは、個別で測れるものは、無限に足し合わせたとしても、同様に測れるということである。長さだけでなく体積も、上記の定義に従う測度であり、さらに言うと、幾何学的な測度に関わらず、例えば確率に関しても、上記の定義に従う測度であることが知られている。

序論でも述べたとおり、本稿の目的は、「測る」という操作に関しての定義、即ち測度論的定義を元に、言語の複雑さを考察することである。ここで、上記の定義に当てはめた言語の複雑さを考えてみる。

定義 3. 3. 測度と言語
ある空でない集合 U に対して、その部分集合の集合族 $\mathfrak{L}$ が以下の性質を満たす、可測空間 $(U, \mathfrak{L})$ である。

- $\emptyset \in \mathfrak{L}$
- $L \in \mathfrak{L} \Rightarrow L^c \in \mathfrak{L}$
- $L_n \in \mathfrak{L}\,(n = 1, 2, \ldots) \Rightarrow \bigcup_{n=1}^{\infty} L_n = L_1 \cup L_2 \cup \ldots \in \mathfrak{L}$

この時、可測空間 $(U, \mathfrak{L})$ に対して、$\mathfrak{L}$ 上で定義された関数 γ が以下を満たし、$(U, \mathfrak{L}, \gamma)$ は、測度空間となる。

- 任意の $L \in \mathfrak{L}$ に対して、$0 \leq \gamma(L) \leq \infty$。特に $\gamma(\emptyset) = 0$
- $L_1, L_2, \ldots \in \mathfrak{L}\,(\sqcap_{n=1}^{\infty} L_n = \emptyset) \Rightarrow \gamma(\sqcup_{n=1}^{\infty} L_n) = \sum_{n=1}^{\infty} \gamma(L_n)$

以上が測度として言語の複雑さを定義したものである。順に追って検討していく。

関数としての複雑さ

定義より、言語の複雑さは、言語を入力に非負の実数を出力する関数 γ として表現できる[2]。言語に対して非負の実数として複雑さを当てがう関数として、言語の複雑さが定義されていることは、直感的に妥当なものである。但し問題となるのは、「実際にどの様な関数であるのか」という点である。ここではまだ、具体的な関数の実体を確定させることはできない。なぜなら、「入力に相当する言語がどの様なものであるか」という点が、まだ明らかでないためである。そこで、一旦関数そのものは置いておいて、入力に相当する言語に関して、考察を進めることとする。

集合族の元としての言語

定義より、「何らかの元からなる集合 U」と「それからなる集合族 $\mathfrak{L}$」とがあり、測度 γ は $\mathfrak{L}$ の元 L_i を入力にもつため、この L_i こそが、この定義における言語に相当するものでなければならない。よって、次のことが自然と要求される。

- 何らかの元の集合 U が存在し、その中の適当なものからなる部分集合 L があり、それが言語である。
- 空である言語 $L_\emptyset$ と、任意の L に対する補集合 L^c が、言語として必ず存在している。
- 言語 $L_1, L_2, \ldots$ が存在するなら、それらを無限に足し合わせたような言語 $\bigcup_{n=1}^{\infty} L_n$ もまた、存在する。

以上は、仮定の部分から導かれる自然な推論である。以下では、測度の定義から導かれる言語の定義に関して、１つ１つ考察していく。

言語 L が集合族の元であることは、何を意味しているのだろうか。第一に、差異の同質性を意味していることが考えられる。言語 L_1 と L_2 が異なるとすれば、それはそれらを構成する元が異なっていることに他ならない。言語の差異が、元の違いという統一的な基準の下に定義されることで、例えば英語と日本語の差も、方言差も、そして個人間の言語の僅かな差も、質としては全く同じ、元の違いによって説明されることが意味されている。第二に、任意の言語の存在である。この定義上、複数の言語を足し合わせた様な言語に関しても、複雑さを測るすることが可能である。例えば、日本語と英語を完全に足し合わせた様な言語は、現状存在しないが、上で見た様に、その様な言語の存在も $\mathfrak{L}$ に含まれ、γ による複雑さ計測が可能であることが要求される。これは、うまく $\mathfrak{L}$ を作ってやれば、現存しない言語や、想像上の言語であっても、問題なく複雑さを測ってやれることを意味している。

では、言語 L を構成する元は、どの様なものが適格であるだろうか。これが定義されれ

[2] 慣習として、この関数はギリシャ文字で表されるが、Complexity の頭文字 C に相当する文字が存在しない。そのため、C と同じく 3 番目のギリシャ文字である γ を用いた。

ば、測度 γ の実体も、自ずと決定される。というのも、元の正体が確定するということは、複雑さを測るするために、何に注目すべきかが確定するということだからである。そのために、まずは先行研究で用いられていた尺度らを検討していきたい。規則基盤的尺度である、屈折の数の多さを考える。全ての屈折のパターンを含む集合 U があり、その部分集合の集合族 $\mathfrak{L}$ が、以下の条件を満たしている。

- いかなる屈折パターンも含まない $L_\emptyset$ が $\mathfrak{L}$ 内に存在する
- ある屈折パターンの集合 L が $\mathfrak{L}$ 内に存在する時、その補集合となる屈折パターンの集合である L^c もまた、存在する。
- $L_1, L_2, \ldots \in \mathfrak{L}$ となる屈折パターンの集合らがある時、$\bigcup_{n=1}^{\infty} L_n \in \mathfrak{L}$

となり、可測空間 $(U, \mathfrak{L})$ が構成される。こうすると、言語 L は、「屈折パターンの集合」によって定義されていることがわかる。即ち、全く同じ屈折パターンを持つもの同士は同じ言語と見做されることを意味しており、例えば孤立語は全て同一の言語として捉えられてしまうことになる。これは直観に沿わない結果であるため、この屈折の数の多さのみで言語の複雑さを計測することは、難しいことがわかる。しかし、測度 γ 自体は、以下の通り要請される通りに定義されることができる

- 任意の $L \in \mathfrak{L}$ に対して、その屈折パターンの数を返す $\gamma(L)$ は正の実数値を取り、特に $\gamma(L_\emptyset) = 0$ である
- 排反である屈折パターン $L_1, L_2, \ldots \in \mathfrak{L}$ について、$\gamma(\bigsqcup_{n=1}^{\infty} L_n) = \sum_{n=1}^{\infty} \gamma(L_n)$ が成り立つ

次に、コルモゴロフ複雑性に関しても考察していく。$U := x \in \{a, b, \ldots, z\}^* : |x| \leq n$ として、長さ n 以下のすべての文字系列の全体集合を考え、その部分集合族 $\mathfrak{L}$ が、以下の条件を満たして存在する。

- 文字系列の長さが 0 の $L_\emptyset$ が存在する
- ある文字系列の集合 L が $\mathfrak{L}$ 内に存在する時、それ以外の文字系列全てからなる補集合 L^c もまた $\mathfrak{L}$ 内に存在する
- $L_1, L_2, \ldots \in \mathfrak{L}$ となる文字系列の集合らが $\mathfrak{L}$ 内に存在する時、$\bigcup_{n=1}^{\infty} L_n \in \mathfrak{L}$

となり、可測空間 $(U, \mathfrak{L})$ が構成される。この場合、言語 L は n 文字以下の文字系列の集合として定義されていることがわかる。即ち、あり得る文字系列の全体集合が同一である場合、同一言語として見做されることを意味しており、特に $n \to \infty$ とすると、直観的には実際の言語と合致し得ると考えられる。しかし、測度 γ はその加法性に関して、要請される方法では定義できない。

- 任意の $L \in \mathfrak{L}$ に対して、その文字系列のコルモゴロフ複雑性を返す $\gamma(L)$ は正の実数値を取り、特に $\gamma(L_\emptyset) = 0$ である。

・排反である文字系列の集合 $L_1, L_2, \ldots \in \mathfrak{L}$ について、$\gamma(\bigsqcup_{n=1}^{\infty}) = \sum_{n=1}^{\infty}(L_n)$ は、常には成り立たず、$\gamma(\bigsqcup_{n=1}^{\infty}) \leq \sum_{n=1}^{\infty}(L_n)$ である。これは組み合わさった時に圧縮可能性は増すことが考えられるためである。

以上の様に、コルモゴロフ複雑性に関しては、測度の加法性という性質が備わっていないため、現段階では測度としては認められない。

次節では、直観的な言語のあり様に合致しており、且つ測度として要請される条件を満たす様な複雑さの尺度を検討する。その中で、加法性を定義に含めるのは、制限が強すぎて、そのままでは使いづらいため、この条件を緩和しながら、測度としての複雑さを考察していく。

4. 測度論と非加法性

4.1. 三人よれば文殊の知恵・船頭多くして船山登る

加法性の制限を緩めるとは、個の単純な足し算が全体と一致しない場合も考慮に入れることを意味している。例えば、「三人寄れば文殊の知恵」という諺が表すような、全体が個の総和を超えることであったり、反対に「船頭多くして船山登る」という諺が表すような、全体が個の総和を下回ることを、考慮に入れることが可能になる。前節の定義が要求する測度の加法性を、上に提示した２つの場合を含めるように緩めると、測度に関しては以下の３つが必要となる。

$$\gamma(\bigcup_{n=1}^{\infty} L_n) = \sum_{n=1}^{\infty} \gamma(L_n) \text{ (加法性)} \tag{2}$$

$$\gamma(\bigcup_{n=1}^{\infty} L_n) \geq \sum_{n=1}^{\infty} \gamma(L_n) \text{ (三人文殊)} \tag{3}$$

$$\gamma(\bigcup_{n=1}^{\infty} L_n) \leq \sum_{n=1}^{\infty} \gamma(L_n) \text{ (船頭多し)} \tag{4}$$

この３つを扱えるような測度論の枠組みで、言語の複雑さを捉えることを試みる。この様な加法性の制限を緩めた測度論の枠組みは、Fuzzy 測度論と呼ばれており、次節で導入する。

4.2. Fuzzy 測度論の導入

Fuzzy 測度[3] $\mu : 2^N \to \mathbb{R}^+$ は、以下の条件を満たす測度である。

$$\mu(\emptyset) = 0 \tag{5}$$

$$S \subseteq T \to \mu(S) \subseteq \mu(T) \tag{6}$$

$$S \cap T = \emptyset \to \mu(S \cup T) \geq \max\{\mu(S), \mu(T)\} \tag{7}$$

因みに、(6) と (7) は、同値であることが知られている。

[3] 藤本（2008）

さて、では一体どの様な測度が、言語の複雑さを測るのに適格であるのか、また、集合族の元として適宜される言語の姿は、どの様なものだろうか。まず、前節で見たコルモゴロフ複雑性に関しては、この Fuzzy 測度として再度定義することで、測度として機能させることが可能である。前節におけるコルモゴロフ複雑性の、測度としての問題は、加法性が成り立たないことであった。これは文字系列の集合を合わせてから測る場合、個別に測る時には無かった圧縮が存在する可能性があるため、より短いアルゴリズム長による記述が存在する可能性があることに起因している。一方で、Fuzzy 測度としてなら、成り立つことは自明である。単純な加法性ではなくとも、文字系列の集合が大きくなれば、自ずと最小アルゴリズム長は長くなるためである。

では、コルモゴロフ複雑性が複雑さを測る測度であり、言語は文字系列の集合であると考えて良いかというと、そうではないと考えられる。この定義においては、文字系列の集合の元によって、その言語が区別されることになるが、以下の状況では不自然な区別をせざるを得なくなる。

1. 全く同じ文字系列集合を用いる言語が、L_a と L_b の 2 つ存在する。
2. しかし、L_a では、りんごを “apple” という文字系列で表現するのに対し、L_b では “banana” という文字系列で表現するとする。
3. この時、L_a と L_b は、同一の言語としてみなして良いか。

直観的には、答えは否である。即ち、見た目上の文字系列自体が同じだとしても、その用法、ひいては意味が異なる場合、異なる言語として区別できた方が、都合が良いが、コルモゴロフ複雑性を測度として用いた場合、その区別ができないことがわかる。この問題を越えるには、意味の側面を考える必要がある。

ここで提案したいのが、「文字系列と意味のペアリングの集合」として言語を定義することである。これは以下で表せる。

文字系列と意味のペア全ての集合 U があり、その冪集合 2^U を考えると、以下の様に可測空間の条件が、自明に満たされる。

1. $\emptyset \in 2^U$
2. $A \in 2^U \rightarrow A_c \in 2^U$
3. $A_n \in 2^U\ (n = 1, 2, \ldots) \rightarrow \bigcup_{n=1}^{\infty} A_n = A_1 \cup A_2 \cup \ldots \in 2^U$

この場合、見た目上の文字系列に注目した時、集合が同一に見えたとしても、意味の差異があれば、それによって言語を区別する余地を持たせることができる。また、文字系列と意味のペアリングの集合として言語を定義しているため、差異の同質性や、任意の言語の存在も担保できている。

最後に、複雑さを測る測度に関しては、「より大きい集合に対しての測度は、より大きく

なる」という、上記の条件を満たしさえしていればなんでも良い。特にその中で、複雑さを測り得るものとして妥当なものを適当に定義できれば、測度の定義に忠実な複雑さの定義が完成する。

5. 結語

本稿では、言語の複雑さの定義に際して、言語の複雑さを「測る」という操作が何をする操作であるかを理論的に考察し、それを基盤とした言語の複雑さの定義を試みた。従来の研究では、「言語の複雑さとは何か？」と言う問いを起点として、言語の複雑さを定義しようとしてきたが、複雑さという概念は勿論、言語という概念でさえ、共通した定義が存在しない以上、「言語の複雑さとは何か？」を問うことを起点にした議論は、必ずしも意味のあるものとは言えなかった。反対に、この「測るとはどの様な操作か？」という問いを基盤に言語の複雑さを議論した研究は、これまでなされてこなかった一方で、測るという概念は、数学における測度論による厳密な定義を与えることが可能であるため、言語の複雑さを考察する基盤としては、より妥当であるように推定される。結果として、Fuzzy 測度論を用いた定義をすることで、言語の非加法性を失わないモデル化が可能であることが示唆された。

その一方で、言語の複雑さを如何に定義するかに関しては、具体的な提案ができていない。今後の研究の指針として、文字系列と意味のペアリングの集合としての言語が、どの様な性質を持つのか、およびその数理モデリングを行うことで、より具体的な言語の複雑さの測度を定義することを試みる。

参考文献

Baechler, R. (2016). Inflectional complexity of nouns, adjectives and articles in closely related *non*—isolated varieties. In Complexity, isolation and variation. De Gruyter.

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017). The entropy of Words—Learnability and expressivity across more than 1000 languages. Entropy, 19(6), 275. https://doi.org/10.3390/e19060275

Bentz, C., Gutierrez-Vasques, X., Sozinova, O., & Samardžić, T. (2023). Complexity trade-offs and equi-complexity in natural languages: A meta-analysis. Linguistics vanguard : multimodal online journal, 9(s1), 9-25. https://doi.org/10.1515/lingvan-2021-0054

Çöltekin, Ç., & Rama, T. (2023). What do complexity measures measure? correlating and validating corpus-based measures of morphological complexity. Linguistics vanguard: multimodal online journal, 9(s1), 27-43. https://doi.org/10.1515/lingvan-2021-0007

Deutscher, G. (2009). “overall complexity”: A wild goose chase? In G. Sampton, D. Gil, & P. Trudgill (Eds.), Language complexity as an evolving variable (pp. 243-251). Oxford University PressOxford. https://doi.org/10.1093/oso/9780199545216.003.0017

Ehret, K., & Szmrecsanyi, B. (2016). An information-theoretic approach to assess linguistic complexity. In R. Baechler & G. Seiler (Eds.), Complexity, isolation, and variation (pp. 71-94). De Gruyter. https://doi.org/10.1515/9783110348965-004

Juola, P. (2008). Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), Language complexity: Typology, contact, change (pp. 89-108). https://doi.org/10.1075/slcs.94.07juo

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. Journal of quantitative linguistics, 5(3), 206-213. https://doi.org/10.1080/09296179808590128

Koplenig, A., Wolfer, S., & Meyer, P. (2023). A large quantitative analysis of written language challenges the idea that all languages are equally complex. Scientific reports, 13(1), 15351. https://doi.org/10.1038/s41598-023-42327-3

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

原 啓介. (2017). **測度・確率・ルベーグ積分－応用への最短コース**. 講談社.

藤本 勝成. (2008). 入門：ファジィ測度とその周辺 －第１回：ファジィ測度とファジィ積分の概要－. **知識と情報**, 20(2), 218-225.