公開

研 究 終 了 報 告 書

「測度論的な概念を用いた形式言語理論への新たなアプローチ」

研究期間: 2021 年 10 月～ 2024 年 3 月
研 究 者: 新屋 良磨

1. 研究のねらい

形式言語・オートマトン理論は，種々の言語クラスを対象に各言語クラスの性質，例えば

(1) 閉包性(その言語クラスがどのような演算に閉じているか？)

(2) 決定可能性(その言語クラス上でどのような問題が決定可能・不能か？)やその計算量

(3) 別表現での等価性(その言語クラスを代数や論理の言葉を使って特徴づけられるか？
例:言語 L が正規 ⇔ L は有限モノイドで認識可能 ⇔ L は狭義単項二階述語論理で定義可能，等)

や，それぞれの言語クラス間の関係を調べる学問である．オートマトン理論はその誕生(Kleene 1951)から多数の結果や理論が創出されており，プログラミング言語・自然言語の構文論の基礎付けやモデル検査などプログラム検証の分野にも大いに貢献している．

本研究では，ある**言語クラス$\mathcal{C}$で極限的に近似できる言語全体のクラス**(およびその拡張・制限)という，従来の文法的・計算モデル的な定義を持つクラスとは異なる新たな言語クラスの性質の解析を行った．ここで，アルファベット A 上の言語 L が「$\mathcal{C}$ で極限的に近似できる」あるいは$\mathcal{C}$ **可測**である，とは図 1 用な状況 $-$ Lの密度にいくらでも近い密度を持つ$\mathcal{C}$ に属す言語の下位集合および上位集合が存在する $-$ すなわち L に「収束」する$\mathcal{C}$ に属す言語の内側・外側からの無限列が存在することを言う．言語 L の密度とは図 2 で定義される極限 $\delta_A(L)$ であり，直感的には「ランダムに取ってきた文字列が L に属する確率」すなわち L の「大きさ」を表した 0 以上 1 以下の実数値である．例えばランダムに取った文字列の長さが偶数になる確率は 1/2 であるため偶数長の文字列全体の集合$(AA)^*$の密度は 1/2 となる(図 2-(1))．また，2 つの文字 $a,b\in A$ に対して，ランダムに取ってきた文字列の先頭が a であるか b であるかは等確率であるため，「aで始まる文字列全体の集合」$aA*$ の密度は $1/\#(A)$ となる(図 2-(2))．任意の文字列 w に対して「文字列 w を部分に含む文字列全体の集合」A^*wA^* の密度が 1 になるという古典的な事実はしばしば無限の猿定理と呼ばれている(図 2-(3))．

公開

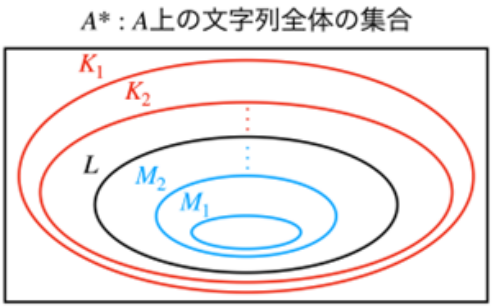

図 1 $\mathcal{C}$ 可測性(各 n で $M_n, K_n \in \mathcal{C}$)

$$\delta_A(L) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \frac{\#(L \cap A^i)}{\#(A^i)}$$

ここでA^iは長さiの文字列全体を，
$\#(X)$は集合Xの要素数を表す．

例 (1) $\delta_A((AA)^*) = 1/2$
(2) $\delta_A(aA^*) = 1/\#(A)$
(3) $\delta_A(A^*wA^*) = 1$

定理 任意の正則言語は密度が有理数に収束する．

図 2 密度の定義と例

$\mathcal{C}$ 可測性は測度論的なアイディアに基づく概念であり，既知の言語クラス$\mathcal{C}$（正規言語など）についての$\mathcal{C}$ 可測性の様々な性質（閉包性，決定可能性，別表現での等価性，etc.）を調べることで，形式言語理論の新たな側面や密度に基づいた**近似理論**の発見・応用提案を行うことが本研究の大きな目標である．

2. 研究成果

（1）概要

$\mathcal{C}$ 可測性は 2021 年に提案者が導入した概念であり，ACT-X 研究開始時点ではまだまだ未解明な部分が多いものであった．研究期間中は，$\mathcal{C}$ として特に正規言語やその部分クラスを対象とした場合の$\mathcal{C}$ 可測性に対するの諸性質を数理的に解明し，形式言語のクラスの分類や解析に用いられる新たな理論体系を創造することを目標とした．具体的には，以下の 2 つの課題(A)(B)に取り組み一連の成果を上げることに成功した．

(A) 正規可測な言語の特徴付けおよび必要条件・十分条件の解析

$\mathcal{C}$ 可測性は「言語クラス $\mathcal{C}$ で極限的に近似可能」という自然な概念であり，数学的にも**カラテオドリ条件**と呼ばれる測度論の言葉を用いて自然に特徴づけることができる．測度論においては実数の集合 X がルベーグ可測であることの同値な特徴付けが多く与えられているが，同様に，正則可測性においても多様な特徴付けを与えられるかは興味深い．一般的に，（例えば正規言語それ自身のように)多様な同値な特徴付けを持つ概念は豊かな理論を持つことが多い．正規可測性は現時点では図 1 の元定義以外はカラテオドリ条件による特徴付けのみが与えられており，それ以外の特徴付けを解明していく．また，同値よりも弱い，正規可測性の必要条件や十分条件が得られれば，与えられた言語が正規可測・非可測であることの便利な補題となる可能性もあるためこれらについても考察を行う．

(B) 正規言語の部分クラスでの可測性の解析

正規可測な言語は非可算無限個存在し，また多くの複雑な文脈自由言語が正規可測となる．あ

る言語が正規可測であるかどうかの判定は一般的に難しく，特に，

「与えられた文脈自由言語が生成する文脈自由言語が正規可測な言語を生成するかどうか」

という決定問題が(文脈自由言語の密度に関するある予想のもとで)決定不能であることが提案者によって示されている．正規言語全体のクラスに比べ，文脈自由言語全体のクラスははるかに複雑であり，現時点では理論的な解析が困難である．例えば「任意の文脈自由言語は密度を持つか？」という素朴な問題も未解決である．そのため，本課題(B)では，「正規可測な文脈自由言語」の“理論的なミニチュア”についての解析を行い，そこで得られた知見や技術を正規可測性の理論に拡張することを目標とする．具体的には，**局所多様体**と呼ばれる「いくつかの演算での閉包性」で定義される正規言語の多様な部分クラス$\mathcal{C}$について，「$\mathcal{C}$可測な正規言語」の特徴付けやアルゴリズム的側面(与えられた正規言語が$\mathcal{C}$可測であるかどうかを判定するアルゴリズムが存在するか？)を考察する．局所多様体は代数的に豊かな構造を持ち，理論的な解析の道具が豊富に揃っている．そのため「正規可測な文脈自由言語全体のクラス」よりも「$\mathcal{C}$可測な正則言語全体のクラス」のほうが解析がしやすく，後者の理解を深めることで，前者の理論的解析(課題 A)に役立つと予想される．

（2）詳細

正規言語の部分クラス（局所多様体）として「一階述語論理で定義可能」という論理的な特徴付けを持つ**星無し言語(Star Free, SF)**は歴史的に最も有名かつ深く研究されてきた部分クラスである．星無し言語のさらなる部分クラスとして，同様に論理的な特徴付け（詳細は省略）を持つ言語クラスとして**無曖昧多項式（Unambiguous Polynomial, UPol）**，**区分検査可能言語（Piecewise Testable, PT）**，**文字検査可能言語(Alphabet Testable, AT)**も代数的言語理論や符号理論のにおいて深く研究されてきた部分クラスもある．また，より計算機的な視点から，「有限長の接頭辞と接尾辞の検査のみで所属判定が決定できる」という特徴付けを持つ**一般化有限確定言語（Generalised Definite, GD）**や**局所検査可能言語(Locally Testable, LT)**も歴史的に良く研究されてきた正規言語の部分クラスである．これらのクラスは SF ⊇ UPol ⊇ PT ⊇ AT ⊆ GD ⊆ LT ⊆ SF という包含関係にあり，いずれの包含関係も真の包含関係を成す（図 3）．また，論理的な特徴付けを持つ UPoL や PT と計算機的な特徴付けを持つ GD や LT はそれぞれ比較不能な言語クラスであり，特別な関係などはこれまで特に発見されていなかった．

公開

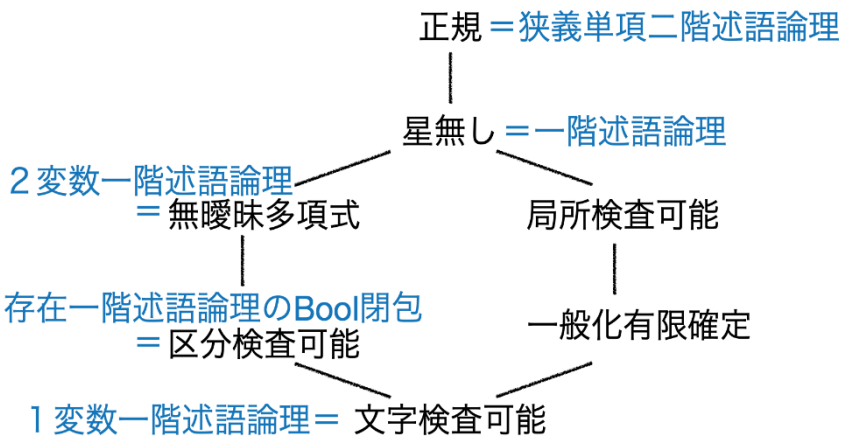

図 3 言語と論理の階層

ACT-X 期間中，課題(B)の具体的な対象として UPol, PT, AT, GD, LT に着目し，それぞれの可測性について以下の一連の成果を得た．

定理 1：任意の言語Lについて，Lは AT 可測 ⇔ Lかその補集合が $\bigcap_{a\in A} A^* a\, A^*$ という形の言語（全ての文字を含む語全体）を部分に含む．

定理 2：任意の言語Lについて，Lは PT 可測 ⇔ ある語 wが存在して Lかその補集合が「wを部分列（scattered subword）に含む語全体」を部分に含む．

定理 3：任意の言語Lについて，Lは UPoL 可測 ⇔ Lは LT 可測 ⇔ Lは GD 可測．

定理 1--2（「代表的な論文」3 の成果）は論理的な特徴付けを持つ AT と PT に対してその可測性（≃近似可能性）を言語的・組合せ論的に特徴付けたものであり，定理 3（「代表的な論文」1,3 の成果）は論理的な特徴付けを持つ UPol と計算的な特徴付けを持つ GD と LT が可測性の観点からは同じ表現力を持つというこれまで発見されていなかった非自明な関係を明らかにしたものである．また，入力となる言語 Lが決定性オートマトンで与えられた場合における AT 可測性・PT 可測性・GD 可測性がそれぞれ **coNP 完全・線形時間可解・決定可能**であるという成果も得られた．可測性の決定可能性は「文脈自由言語に対する正規可測性は決定不能」という否定的な結果のみが知られており，これらの決定可能性は初めての肯定的な結果と言える．AT 可測性・PT 可測性の計算量の解析は山口勇太郎氏（大阪大学）・中村誠希氏（東京工業大学）との共同研究成果（「代表的な論文」2 の成果）であり，GD 可測性の決定可能性は「代表的な論文」1 の成果である．

3. 今後の展開

可測性はもともと「非正規言語を，正規言語を用いて分類する」ために申請者が導入した概念であり，いくつかの具体的な言語に対する正規可測性・非可測性の例や閉包性および「文脈自由

公開

言語に対する正規可測性は決定不能」という計算的に否定的な結果が知られているのみであった. 一方, 正規言語の部分クラスでの可測性を考えることで, ACT-X 期間中に AT・PT・GD 可測性の非自明な決定可能性を与えることに成功し, また, AT 可測性・PT 可測性についてはタイトな計算量のバウンドも与えられている. 期間中に得られた$\mathcal{C}$可測性の決定可能性は「正規言語Lと許容誤差 $\varepsilon \geq 0$」を入力として「Lとの密度の差が ε 以下の $\mathcal{C}$に属する内側(外側)からの近似言語」を出力する**近似アルゴリズム**を伴うものであり, この近似アルゴリズムを正規言語に関連する決定問題(正規表現マッチングやモデル検査)の高速化などに応用できないかどうかは興味深い展開である.

4. 自己評価

ACT-X 申請時に目標とした課題(B)については当初の設定目標以上の成果を挙げることができた. 特に, 計算量の解析について分野内外の専門家らとも期間中に共同研究を行えたことが大きい. 課題(A)についても, 課題(B)で得られた知見を基にいくつかの具体的な予想や結果(未発表)が得られている. ACT-X 期間中の研究を通じて可測性という萌芽的な概念について理論的な理解は大きく深まったと言える.

その一方, 申請時に別目標として掲げていた課題「可測性の別分野への応用」については, いくつかのアイディアは得られたものの具体的な成果と言えるものはまだ出ておらず, 当初の計画通りには進まなかったと感じている. 可測性の理論的な理解を深めることは形式言語理論においては重要であるが, より広範な学術的インパクトを与えるためには別理論(例:ω 正規言語や木正規言語, 確率オートマトンや制約オートマトンなどの理論)との新たなつながりや応用(例:近似アルゴリズムを用いた高速化の提案)の提案が今後の必要課題であると考えている.

5. 主な研究成果リスト

(1)代表的な論文(原著論文)発表

研究期間累積件数: 4 件

1. Ryoma Sin'ya. Measuring Power of Generalised Definite Languages. In Proceedings of the 27th International Conference on Implementation and Application of Automata. 2023, LNCS Vol, 14151. pp. 278--279.
A language L is said to be C-measurable, where C is a class of languages, if there is an infinite sequence of languages in C that "converges" to L. In this paper, we investigate the measuring power of GD of the class of all generalised definite languages. Although each generalised definite language only can check some local property (prefix and suffix of some bounded length), it is shown that many non-generalised definite languages are GD-measurable. Further, we show that it is decidable whether a given regular language is GD-measurable or not.
2. 新屋 良磨, 山口 勇太郎, 中村 誠希. 部分語の出現情報の検査のみで近似できる正規言語について. コンピュータソフトウェア. 2022, 40 巻 2 号, pp. 49--60.
言語 L が正規可測であるとは, L に「収束」する正規言語の対の無限列が存在することを言う. 本論文では, 正規言語の代わりに正規言語の部分クラスである区分検査可能(Piecewise

公開

Testable (PT): 部分語の出現情報の Bool 演算で記述可能)言語および文字検査可能(Alphabet Testable (AT): 文字の出現情報の Bool 演算で記述可能)言語に焦点を当てその可測性を考察する. 特に, 正規言語に対する AT 可測性は co-NP 完全である一方, PT 可測性は線形時間で決定できることを示す.
3. Ryoma Sin'ya. Measuring Power of Locally Testable Languages. In Proceedings of the 26th International Conference on Developments in Language Theory. 2022, LNCS Vol. 13257, pp 274–285.
A language L is said to be C-measurable, where C is a class of languages, if there is an infinite sequence of languages in C that converges to L. In this paper we investigate the measuring power of LT the class of all locally testable languages. Although each locally testable language only can check some local property (prefix, suffix, and infix of some bounded length), it is shown that many non-locally-testable languages are LT-measurable. In particular, we show that the measuring power of locally testable languages coincides with the measuring power of unambiguous polynomials. We also examine the measuring power of some fragments of unambiguous polynomials.

（2）特許出願

研究期間全出願件数:0 件（特許公開前のものは件数にのみ含む）

1	発　明　者	
	発明の名称	
	出　願　人	
	出　願　日	
	出願番号	
	概　　要	

（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）

学会発表:

Ryoma Sin'ya. Measuring Power of Generalised Definite Languages. The 27th International Conference on Implementation and Application of Automata (CIAA' 23). 2023 年 9 月 22 日.

学会発表:

Ryoma Sin'ya. Measuring Power of Locally Testable Languages. The 26th International Conference Developments in Language Theory (DLT' 22). 2022 年 5 月 12 日.

招待講演:

新屋良磨. 正規言語族の無限内部階層における分離問題および可測性について. 2022 日本数学会 年会 特別講演(数学基礎論および歴史分科会). 2022 年 3 月 29 日.