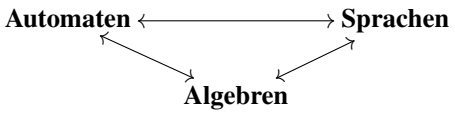
Ein kategorientheoretischer Zugang zur algebraischen Sprachtheorie

Henning Urbat[1]

Abstract: In der algebraischen Theorie der formalen Sprachen werden Automaten und ihre Sprachen klassifiziert, indem man ihnen algebraische Strukturen zuordnet. Der algebraische Ansatz wurde für verschiedene Typen von Sprachen erfolgreich umgesetzt und führte zu einer Vielzahl von wichtigen Ergebnissen (z.B. Entscheidbarkeitsaussagen), die häufig auf strukturell sehr ähnlichen Ideen basieren. In der Dissertation [Ur18] wird nachgewiesen, dass zahlreiche Schlüsselkonzepte der algebraischen Sprachtheorie einen kategorientheoretischen Hintergrund haben. Die Hauptidee des kategoriellen Ansatzes besteht zum einen darin, formale Sprachen und die ihnen zugeordeten algebraischen Strukturen durch Monaden auf einer algebraischen Kategorie zu modellieren, und zum anderen in einer Interpretation der algebraischen Spracherkennung durch Betrachtung von prädualen Kategorien. Auf diese Weise kann gezeigt werden, dass zentrale Elemente der algebraischen Sprachtheorie einen generischen Charakter haben und strukturell unabhängig vom Sprachtyp sind.

1 Einleitung

Automaten als formale Modelle zustandsbasierter Systeme gehören zu den etabliertesten Werkzeugen der Theoretischen Informatik. Sie existieren in Hunderten Varianten und haben vielfältige praktische Anwendungen, etwa im Compilerbau, in der Künstlichen Intelligenz und in der Verifikation sicherheitskritischer Systeme. Die Untersuchung von Automaten als mathematische Strukturen führt zu tiefliegenden Ergebnissen und überraschenden Querverbindungen, die häufig auf der Interpretation von automatentheoretischen Fragestellungen aus einem algebraischen oder topologischen Blickwinkel basieren. Die Grundidee des algebraischen Zugangs zur Automatentheorie besteht darin, mathematische Beziehungen zwischen Automaten, den von ihnen repräsentierten Sprachen, und geeigneten algebraischen Strukturen herzustellen, wie im folgenden Diagramm angedeutet:



Auf diese Weise erhalten mächtige algebraische Methoden, etwa aus der Halbgruppen- und Gruppentheorie, Einzug in die Welt der Automaten und formalen Sprachen. Das im obigen Diagramm beschriebene Prinzip ist sehr allgemein und auf viele Klassen von Sprachen (z.B. reguläre Sprachen, ω-reguläre Sprachen oder Baumsprachen) anwendbar.

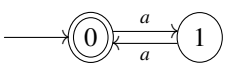
[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl Informatik 8, henning.urbat@fau.de
Englischer Titel der Dissertation: "A Categorical Approach to Algebraic Language Theory"

In der Dissertation [Ur18] und den zugrunde liegenden Arbeiten [Ad14, Ad15, AMU15, Ch16, CU16, Ur17] wird ein kategorientheoretischer Zugang zur algebraischen Sprachtheorie entwickelt, der zahlreiche Elemente dieser Theorie *generisch* – also unabhängig vom konkreten Sprachtyp – präsentiert. Ziel dabei ist (i) eine vereinheitlichende Sicht auf Konzepte, die zuvor für verschiedene Sprachtypen separat untersucht wurden, und (ii) eine modulares und robustes Rahmenwerk, das generische und anwendungsspezifische Aspekte der Theorie voneinander isoliert und damit die Herleitung neuer Ergebnisse vereinfacht. Zur Motivation werden im folgenden Abschnitt zunächst einige klassische Ergebnisse der algebraischen Sprachtheorie diskutiert. Anschließend werden die in der Arbeit entwickelten kategoriellen Methoden mit einigen ihrer Anwendungen vorgestellt.

2 Hintergrund: Algebraische Spracherkennung

Der Ausgangspunkt der algebraischen Sprachtheorie liegt in der Beobachtung, dass sich reguläre Sprachen nicht nur durch die klassischen endlichen Automaten, sondern auch rein algebraisch durch Monoide beschreiben lassen. Diese Korrespondenz basiert auf dem fundamentalen Konzept der *algebraischen Erkennung* von Sprachen: Eine Sprache $L \subseteq \Sigma^*$ über einem Alphabet Σ wird von einem Monoid M erkannt, wenn ein Monoidmorphismus $h\colon \Sigma^* \to M$ und eine Teilmenge $S \subseteq M$ mit $L = h^{-1}[S]$ existiert. Dabei bezeichnet $h^{-1}[S] = \{\, w \in \Sigma^* : h(w) \in S \,\}$ das Urbild von S bzgl. h.

Beispiel. Zur Illustration betrachten wir die Sprache $L = (aa)^*$ aller Wörter gerader Länge über dem Alphabet $\Sigma = \{a\}$. Sie wird von dem abgebildeten endlichen Automaten erkannt. Um diese Sprache algebraisch zu erfassen, verwendet man das additive Monoid $\mathbb{Z}_2 = \{0,1\}$ der ganzen Zahlen modulo 2 und den Monoidmorphismus $h\colon \{a\}^* \to \mathbb{Z}_2$, der ein Wort auf seine Länge modulo 2 abbildet. Dann ist $L = h^{-1}[\{0\}]$, d.h. L wird vom Monoid $\mathbb{Z}_2$ erkannt.



Endliche Automaten und endliche Monoide sind in Bezug auf die Erkennung von Sprachen äquivalente Strukturen, wie das folgende klassische Ergebnis zeigt:

Satz (Myhill-Nerode)**.** *Eine Sprache ist genau dann regulär, wenn sie von einem endlichen Monoid erkannt wird.*

Das in der Einleitung beschriebene Diagramm lässt sich also wie folgt instanziieren:

Die Charakterisierung der regulären Sprachen durch Monoide ermöglicht es, für wichtige Konzepte aus der Theorie der endlichen Automaten ein algebraisches Gegenstück zu identifizieren. Das betrifft etwa die Existenz minimaler Strukturen: Genau wie jede reguläre Sprache einen eindeutigen Minimalautomaten hat, gibt es auch ein eindeutiges minimales Monoid, das sie erkennt – das *syntaktische Monoid*. Im obigen Beispiel ist der abgebildete Automat der Minimalautomat der Sprache $(aa)^*$, und $\mathbb{Z}_2$ ist ihr syntaktisches Monoid.

Eines der Hauptziele des algebraischen Zugangs zur Automatentheorie ist die Klassifizierung von Eigenschaften regulärer Sprachen durch Eigenschaften ihrer syntaktischen Monoide. Das bekannteste Ergebnis dieser Art ist der Satz von Schützenberger [Sc65]. Er betrachtet die Klasse der *sternfreien Sprachen*, also Sprachen, die sich durch einen regulären Ausdruck wie $(\overline{a}+b)\overline{ab}$ beschreiben lassen, der den Komplementoperator $\overline{(-)}$, aber keinen Kleene-Stern $(-)^*$ verwenden darf. Schützenbergers Satz besagt, dass eine reguläre Sprache genau dann sternfrei ist, wenn ihr syntaktisches Monoid *aperiodisch* ist, d.h. wenn es die Gleichung $x^{n+1} = x^n$ für hinreichend großes n erfüllt. Äquivalent dazu ist die Aussage, dass die eindeutige idempotente Potenz x^ω jedes Elements x die Gleichung $x^\omega = x \cdot x^\omega$ erfüllt. In unserem obigen Beispiel sieht man leicht, dass das Monoid $\mathbb{Z}_2$ nicht aperiodisch ist. Folglich ist die Sprache $(aa)^*$ nicht sternfrei. Da das syntaktische Monoid effektiv aus dem Minimalautomaten einer gegebenen Sprache berechnet werden kann, impliziert der Satz von Schützenberger, dass Sternfreiheit eine entscheidbare Eigenschaft regulärer Sprachen ist. Diese Entscheidbarkeitsaussage wäre unter alleiniger Verwendung von Standardergebnissen über Automaten und reguläre Ausdrücke schwierig zu beweisen, und illustriert daher die Relevanz algebraischer Methoden in der Automatentheorie.

Neben den sternfreien Sprachen gibt es unzählige weitere Klassen von regulären Sprachen, die im Stil von Schützenbergers Satz durch Eigenschaften ihrer syntaktischen Monoide charakterisierbar sind. Um einen systematischen Zugang zu derartigen Korrespondenzergebnissen zu erhalten, führte Eilenberg [Ei76] zwei fundamentale Konzepte ein: *Varietäten von regulären Sprachen* und *Pseudovarietäten von Monoiden*. Eine Varietät von regulären Sprachen ist eine Klasse $\mathscr{V}$ von regulären Sprachen, die unter den booleschen Operationen, Ableitungen und homomorphen Urbildern abgeschlossen ist. Das heißt, (i) für alle Sprachen $K, L \subseteq \Sigma^*$ in $\mathscr{V}$ ist $\emptyset, \Sigma^*, K \cup L, K \cap L, \Sigma^* \setminus L \in \mathscr{V}$, (ii) für alle Sprachen $L \subseteq \Sigma^*$ in $\mathscr{V}$ und alle Wörter $y \in \Sigma^*$ ist

$$y^{-1}L = \{x \in \Sigma^* : yx \in L\} \in \mathscr{V} \quad \text{und} \quad Ly^{-1} = \{x \in \Sigma^* : xy \in L\} \in \mathscr{V},$$

und (iii) für alle Sprachen $L \subseteq \Sigma^*$ in $\mathscr{V}$ und alle Monoidmorphismen $h\colon \Delta^* \to \Sigma^*$ ist $h^{-1}[L] \in \mathscr{V}$. Eine Pseudovarietät von Monoiden ist eine Klasse $\mathbb{V}$ von endlichen Monoiden, die unter homomorphen Bildern, Untermonoiden und endlichen Produkten abgeschlossen ist. Die Beziehung zwischen diesen Konzepten wird durch den folgenden Satz hergestellt:

Satz (Eilenberg-Korrespondenz)**.** *Varietäten von regulären Sprachen stehen in bijektiver Korrespondenz zu Pseudovarietäten von Monoiden.*

Eilenbergs Satz liefert eine generische Beziehung zwischen Eigenschaften von Sprachen und Eigenschaften von Monoiden. Beispielsweise bilden die sternfreien Sprachen eine Varietät von regulären Sprachen, und die aperiodischen endlichen Monoide bilden eine Pseudovarietät von Monoiden, und somit kann der Satz von Schützenberger (ebenso wie viele verwandte Ergebnisse) als Instanz der Eilenberg-Korrespondenz interpretiert werden.

Eine wichtige Ergänzung zu Eilenbergs Satz ist die von Reiterman [Re82] bewiesene modelltheoretische Charakterisierung von Pseudovarietäten: diese entsprechen genau den Klassen von endlichen Monoiden, die durch *proendliche Gleichungen* (eine topologische Verallgemeinerung von klassischen Gleichungen zwischen Termen) axiomatisierbar sind.

Die Identität $x^\omega = x \cdot x^\omega$, die aperiodische endliche Monoide definiert, ist ein Beispiel für eine proendliche Gleichung. Reitermans Satz ist nicht spezifisch für Monoide, sondern gilt allgemeiner für Pseudovarietäten endlicher Algebren über einer finitären Signatur.

Die fundamentalen Ergebnisse von Eilenberg und Reiterman stellen eine enge Verbindung zwischen formalen Sprachen, algebraischen Strukturen und proendlichen Gleichungstheorien her, und gehören zu den Grundpfeilern der algebraischen Theorie regulärer Sprachen. In den vergangenen Jahrzehnten wurden zahlreiche Verallgemeinerungen und Erweiterungen des Konzeptes der algebraischen Spracherkennung und insbesondere der Eilenberg-Reiterman-Theorie erforscht, die sich in zwei Richtungen klassifizieren lassen:

(I) Trotz der Allgemeinheit von Eilenbergs Satz wurde schnell realisiert, dass viele interessante Klassen von regulären Sprachen keine Varietäten bilden, weil einige der erforderlichen Abschlusseigenschaften nicht erfüllt sind. Daher wurden diverse Abschwächungen des Varietätenkonzeptes untersucht. Auf der algebraischen Seite erfordert das die Betrachtung von Monoiden mit zusätzlicher Struktur. Beispielsweise betrachtete Pin [Pi95] die Erkennung von Sprachen durch *geordnete* Monoide und bewies eine Korrespondenz zwischen *positiven Varietäten von regulären Sprachen* (die nicht notwendig unter Komplement abgeschlossenen sind) und Pseudovarietäten von geordneten Monoiden.

(II) Eine weitere wichtige Forschungsrichtung befasst sich mit der Erweiterung der klassischen algebraischen Sprachtheorie, die sich auf reguläre Sprachen endlicher Wörter bezieht, auf andere Typen von formalen Sprachen. Hierfür ist es nicht länger ausreichend, Monoide als erkennende algebraische Strukturen zu verwenden. Beispielsweise können ω-reguläre Sprachen (d.h. die von Büchi-Automaten repräsentierten Sprachen unendlicher Wörter) algebraisch durch *ω-Halbgruppen* beschrieben werden, eine Verallgemeinerung von Monoiden mit einer unendlichen Multiplikation. Darüber hinaus existieren algebraische Zugänge für zahlreiche weitere Sprachtypen, darunter rationale Potenzreihen, Sprachen von Wörtern über linearen Ordnungen, Baumsprachen und Kostenfunktionen.

Alle oben beschriebenen Erweiterungen der Eilenberg-Reiterman-Theorie folgen dem Pfad der klassischen algebraischen Theorie regulärer Sprachen und operieren in fünf Schritten:

(1) Identifiziere eine algebraische Theorie T, so dass die betrachteten Sprachen genau den von endlichen T-Algebren erkannten Sprachen entsprechen.

(2) Untersuche die Existenz und Konstruktion von *syntaktischen T-Algebren*, also den minimalen algebraischen Erkennern von Sprachen.

(3) Führe das Konzept einer *Varietät von Sprachen* ein, also einer Klasse von Sprachen, die unter einer Teilmenge der booleschen Operationen, einem geeigneten Konzept von *Ableitungen* und Urbildern unter T-Algebra-Morphismen abgeschlossen ist.

(4) Führe das Konzept einer *Pseudovarietät von T-Algebren* ein, also einer Klasse von endlichen T-Algebren mit geeigneten Abschlusseigenschaften.

(5) Beweise eine Eilenberg-Korrespondenz zwischen Varietäten von Sprachen und Pseudovarietäten von T-Algebren.

Die Implementierung dieser Schritte erfordert jeweils die Adaption von Definitionen, Konstruktionen und Beweisen aus dem klassischen Fall der regulären Sprachen und Monoide. In der Folge erhält man eine Vielzahl von Ergebnissen, die strukturell sehr ähnliche Aussagen für verschiedene Sprachtypen etablieren und oft einen Ad-hoc-Charakter haben. Zum Beispiel existieren in der Literatur mehr als 20 Varianten von Eilenbergs Varietätensatz. Diese Situation motiviert die Suche nach einem allgemeineren Zugang zur algebraischen Sprachtheorie, mit der Zielsetzung, zuvor separate Ergebnisse unter ein gemeinsames Dach zu stellen und als Spezialfälle generischer Prinzipien zu interpretieren. In unserer Arbeit wird dies durch Verwendung kategorientheoretischer Methoden erreicht.

3 Kategorientheoretische Perspektive

In diesem Abschnitt stellen wir die Kernelemente unseres kategorientheoretischen Zugangs zur algebraischen Sprachtheorie vor. Seine zentrale Erkenntnis besteht in der Beobachtung, dass die im vorherigen Abschnitt beschriebenen Schritte (1)–(5) wesentlich vereinfacht oder sogar vollständig automatisiert werden können. Der kategorielle Ansatz wird durch die folgende "Gleichung" zusammengefasst:

Algebraische Sprachtheorie = Monaden + Unäre Präsentationen + Dualität.

Im Folgenden beschreiben wir genauer, wie diese Konzepte verwendet werden.

3.1 Monaden und Sprachen

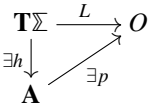
Die Voraussetzung für einen algebraischen Zugang zu einem beliebigen Typ von Sprachen ist eine Beschreibung dieser Sprachen durch geeignete algebraische Strukturen. In unserem Rahmenwerk verwenden wir hierfür das Konzept einer *Monade*, ein etablierter Formalismus der Kategorientheorie, mit dem man algebraische Strukturen abstrakt (unter Weglassung syntaktischer Konzepte wie Terme und Operationen) erfassen kann. Beispielsweise sind Monoide genau die Algebren der Monade $\mathbf{T}\Sigma = \Sigma^*$ auf **Set**, der Kategorie der Mengen, die jeder Menge Σ das freie Monoid Σ^* zuordnet. Allgemeiner betrachten wir eine beliebige Grundkategorie $\mathscr{D}$ von (evtl. geordneten, mehrsortigen) algebraischen Strukturen wie Mengen, halbordneten Mengen oder Vektorräumen, sowie eine Monade $\mathbf{T}$ auf der Kategorie $\mathscr{D}$. Sprachen werden als Morphismen $L\colon \mathbf{T}\mathbb{\Sigma} \to O$ in $\mathscr{D}$ modelliert, wobei $\mathbb{\Sigma}$ und O Objekte von $\mathscr{D}$ sind, die Eingaben und Ausgaben repräsentieren. Auf diese Weise lassen sich zahlreiche Typen von formalen Sprachen kategoriell beschreiben, z.B.:

(a) Für die klassischen regulären Sprachen wählt man die Monoid-Monade $\mathbf{T}\Sigma = \Sigma^*$ auf $\mathscr{D} = \mathbf{Set}$ und das Ausgabeobjekt $O = \{0,1\}$.

(b) Rationale Potenzreihen über einem Körper $\mathbb{K}$ werden durch die Monade $\mathbf{T}$ auf der Kategorie $\mathscr{D}$ aller $\mathbb{K}$-Vektorräume repräsentiert, die freie $\mathbb{K}$-Algebren konstruiert. Als Ein- bzw. Ausgabeobjekt wählt man $\mathbb{\Sigma} = \mathbb{K}^\Sigma$ für ein endliches Alphabet Σ und $O = \mathbb{K}$. Weil die freie $\mathbb{K}$-Algebra $\mathbf{T}\mathbb{\Sigma}$ vom Vektorraum mit Basis Σ^* getragen wird, entspricht eine lineare Abbildung $L\colon \mathbf{T}\mathbb{\Sigma} \to \mathbb{K}$ einer Potenzreihe $L_0\colon \Sigma^* \to \mathbb{K}$.

(c) Für ω-reguläre Sprachen betrachtet man die ω-Halbgruppen-Monade $\mathbf{T}(\Sigma,\Gamma) = (\Sigma^+, \Sigma^\omega + \Sigma^* \times \Gamma)$ auf der Kategorie $\mathscr{D} = \mathbf{Set}^2$ der zweisortigen Mengen, zusammen mit $\mathbb{\Sigma} = (\Sigma, \emptyset)$ für ein endliches Alphabet Σ und $O = (\{0,1\}, \{0,1\})$.

Die Interpretation von Sprachen als Morphismen erlaubt die Einführung eines allgemeinen Konzeptes algebraischer Spracherkennung: eine Sprache $L\colon \mathbf{T}\mathbb{\Sigma} \to O$ ist $\mathbf{T}$-*erkennbar*, wenn sie durch eine endliche $\mathbf{T}$-Algebra $\mathbf{A}$ faktorisiert (siehe Abbildung rechts). Dieses Konzept umfasst u.a. die Erkennung regulärer Sprachen durch Monoide, die Erkennung ω-regulärer Sprachen durch ω-Halbgruppen, und die Erkennung von Potenzreihen durch $\mathbb{K}$-Algebren.



Die Verwendung von Monaden zur generischen Modellierung algebraischer Spracherkennung wurde zuerst von Bojańczyk [Bo15] für den sehr eingeschränkten Fall $\mathscr{D} = \mathbf{Set}$ untersucht. Der Ansatz von *op. cit.* basiert auf speziellen Eigenschaften dieser Grundkategorie, insbesondere der Tatsache, dass alle Monaden durch Operationen und Gleichungen präsentierbar sind. Die Verallgemeinerung auf beliebige Grundkategorien $\mathscr{D}$ wie im oben beschriebenen Rahmenwerk ist daher nichttrivial und erfordert neue Techniken.

3.2 Proendliche Monaden

In der algebraischen Theorie regulärer Sprachen ist es oft hilfreich, eine topologische Perspektive einzunehmen. Das wichtigste Werkzeug in diesem Zusammenhang ist der Stone-Raum $\widehat{\Sigma^*}$ aller *proendlichen Wörter* über dem Alphabet Σ, der als inverser Limes aller endlichen Quotientenmonoide des freien Monoids Σ^* konstruiert wird. Reguläre Sprachen entsprechen genau den zugleich abgeschlossenen und offenen Teilmengen von $\widehat{\Sigma^*}$. Eines der Kernkonzepte unseres kategorientheoretischen Ansatzes ist die *proendliche Monade* $\widehat{\mathbf{T}}$, die der Monade $\mathbf{T}$ zugeordnet wird und die Konstruktion des Raumes der proendlichen Wörter verallgemeinert. Die Monade $\widehat{\mathbf{T}}$ "lebt" in der Kategorie $\widehat{\mathscr{D}}$, die als freie Vervollständigung der Kategorie der endlichen $\mathscr{D}$-Algebren unter inversen Limites entsteht. Beispielsweise ist $\widehat{\mathbf{Set}}$ die Kategorie **Stone** der Stone-Räume, und für die Monoid-Monade $\mathbf{T}\Sigma = \Sigma^*$ auf **Set** ist die proendliche Monade gegeben durch $\widehat{\mathbf{T}}\Sigma = \widehat{\Sigma^*}$ auf **Stone**.

Als erste Anwendung der proendlichen Monade ergibt sich eine kategorielle Verallgemeinerung des Satzes von Reiterman. Hierfür wird das Konzept einer *Pseudovarietät von* $\mathbf{T}$*-Algebren* sowie einer *proendlichen Theorie* über $\widehat{\mathbf{T}}$ eingeführt. Pseudovarietäten von $\mathbf{T}$-Algebren verallgemeinern Pseudovarietäten von Monoiden, und proendliche Theorien bilden eine kategorielle Abstraktion von proendlichen Gleichungen. Wir erhalten:

Satz (Verallgemeinerte Reiterman-Korrespondenz [Ch16, Ur17])**.** *Proendliche Theorien über* $\widehat{\mathbf{T}}$ *stehen in bijektiver Korrespondenz zu Pseudovarietäten von* $\mathbf{T}$*-Algebren.*

Wählt man als $\mathbf{T}$ eine Monade auf der Kategorie der Mengen, die finitäre algebraische Strukturen (z.B. Monoide) repräsentiert, so erhält man den klassischen Reiterman-Satz als Spezialfall dieses Ergebnisses. Darüber hinaus lässt sich der Satz auf Situationen anwenden, für die bisher keine Reiterman-Korrespondenz bekannt war, etwa auf nicht finitäre algebraische Strukturen wie ω-Halbgruppen.

3.3 Unäre Präsentationen und syntaktische Algebren

Der Schlüssel zu einem kategoriellen Verständnis von syntaktischen Monoiden (und anderen synktaktischen Algebren) liegt im Konzept einer *unären Präsentation* einer **T**-Algebra. Anschaulich beschreibt eine solche Präsentation, wie sich eine **T**-Algebra durch unäre Operationen beschreiben lässt. Beispielsweise kann ein Monoid M durch die Translationen $x \mapsto yx$ und $x \mapsto xy$ mit $y \in M$ präsentiert werden. Der Zusammenhang zwischen unären Präsentationen und syntaktischen Algebren ist durch den folgenden Satz gegeben:

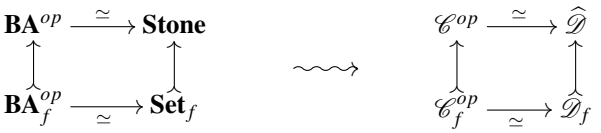
Satz ([Ur17])**.** *Wenn die freie* **T***-Algebra* $\mathbf{T}\mathbb{\Sigma}$ *eine unäre Präsentation hat, dann hat jede* **T***-erkennbare Sprache* $L\colon \mathbf{T}\mathbb{\Sigma} \to O$ *eine syntaktische* **T***-Algebra.*

Dieses Ergebnis vereinheitlicht nicht nur die Konstruktion zahlreicher bekannter Typen von syntaktischen Strukturen, sondern es erklärt auch die Rolle, die diese in der algebraischen Sprachtheorie einnehmen: aus kategorieller Sicht bedeutet das Arbeiten mit syntaktischen Algebren in Wirklichkeit das Arbeiten mit unären Präsentationen. Es ergeben sich viele Vorteile, das Konzept einer unären Präsentation explizit herauszustellen; unter anderem liefert es eine klarere Sicht auf viele Definitionen und Beweise in der Literatur.

3.4 Dualität

Die finale Zutat unseres kategoriellen Zugangs zur algebraischen Sprachtheorie liegt in der Interpretation algebraischer Spracherkennung durch *Dualisierung*. Wie bereits erwähnt, lassen sich die regulären Sprachen topologisch als abgeschlossene offene Mengen im Stone-Raum $\widehat{\Sigma^*}$ aller proendlichen Wörter beschreiben. Pippenger [Pi97] konnte nachweisen, dass sich dieses Ergebnis als Instanz der klassischen *Stone-Dualität* zwischen der Kategorie **BA** der booleschen Algebren und der Kategorie **Stone** der Stone-Räume auffassen lässt: Die boolesche Algebra aller regulären Sprachen über Σ (mit den mengentheoretischen booleschen Operationen) ist dual zum Stone-Raum $\widehat{\Sigma^*}$. Die Korrektheit von Pippengers Ergebnis fußt implizit auf der Tatsache, dass sich die Stone-Dualität zu einer dualen Äquivalenz zwischen der Kategorie $\mathbf{BA}_f$ der endlichen booleschen Algebren und der Kategorie $\mathbf{Set}_f$ der endlichen Mengen (= endlichen Stone-Räume) einschränken lässt.

Eine der zentralen Erkenntnisse unserer Arbeit besteht darin, dass die topologischen Details der Stone-Dualität unerheblich sind: Man kann mit einem beliebigen (!) Paar $\mathscr{C}$ und $\mathscr{D}$ von algebraischen Kategorien arbeiten, die auf der Ebene von endlichen Algebren dual zueinander sind. Zwei solche Kategorien heißen *prädual*. Das bedeutet, dass man die auf der linken Seite des folgendes Diagramms abgebildete Stone-Dualität durch eine abstrakte duale Situation wie auf der rechten Seite ersetzen kann.



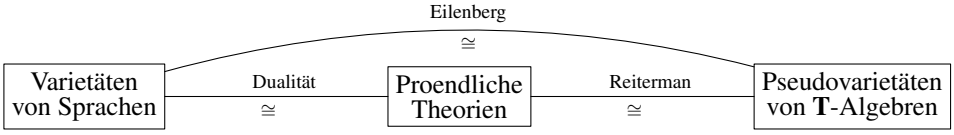
Neben der Stone-Dualität ($\mathscr{C}$ = boolesche Algebren und $\mathscr{D}$ = Mengen) gibt es viele weitere Beispiele von prädualen Kategorien, etwa die Birkhoff-Dualität ($\mathscr{C}$ = distributive Verbände und $\mathscr{D}$ = halbgeordnete Mengen) oder die Selbstdualität von endlichdimensionalen Vektorräumen ($\mathscr{C} = \mathscr{D}$ = Vektorräume). Für jedes präduale Paar $\mathscr{C}/\mathscr{D}$ kann man die Menge $\mathsf{Rec}(\Sigma)$ aller $\mathbf{T}$-erkennbaren Sprachen über dem Eingabeobjekt Σ in natürlicher Weise mit der Struktur einer $\mathscr{C}$-Algebra versehen. Das ermöglicht eine Verallgemeinerung von Pippengers dualer Charakterisierung der booleschen Algebra aller regulären Sprachen auf die Ebene einer beliebigen Monade $\mathbf{T}$:

Satz (Verallgemeinerte Pippenger-Korrespondenz [Ur17])**.** *Die Objekte* $\mathsf{Rec}(\Sigma) \in \mathscr{C}$ *und* $\widehat{\mathbf{T}}\Sigma \in \widehat{\mathscr{D}}$ *sind dual zueinander.*

Dieses Ergebnis ermöglicht es, das sehr allgemeine Konzept einer *Varietät von* $\mathbf{T}$*-erkennbaren Sprachen* einzuführen. Diese Varietäten werden in der Kategorie $\mathscr{C}$ gebildet, und ihre Definition beinhaltet den Begriff einer *Ableitung* von Sprachen, der auf einer unären Präsentation für $\mathbf{T}$-Algebren basiert. Im klassischen Fall (d.h. für die Monoid-Monade $\mathbf{T}\Sigma = \Sigma^*$ auf $\mathscr{D} = \mathbf{Set}$) reflektiert die Abgeschlossenheit von Varietäten von regulären Sprachen unter den booleschen Operationen die Tatsache, dass Varietäten in der Kategorie $\mathscr{C}$ der booleschen Algebren gebildet werden, und die Definition der Ableitungen $y^{-1}L$ und Ly^{-1} ist der unären Präsentation von Monoiden durch die Operationen $x \mapsto yx$ and $x \mapsto xy$ geschuldet. Möchte man modifizierte Konzepte von Varietäten betrachten, die nur unter einer Teilmenge der booleschen Operationen abgeschlossen sind, so muss lediglich die unterliegende Dualität angepasst werden. Beispielsweise modelliert man Pins positive Varietäten von regulären Sprachen durch das präduale Paar $\mathscr{C}$ = distributive Verbände / $\mathscr{D}$ = halbgeordnete Mengen und die Monade $\mathbf{T}$ auf $\mathscr{D}$, die geordnete Monoide repräsentiert. Auf diese Weise erhält man zahlreiche Varietätenbegriffe aus der Literatur als Instanzen unserer allgemeinen Definition. Darüber hinaus ergibt sich eine starke Verallgemeinerung von Eilenbergs Varietätensatz, die eines der Hauptergebnisse unserer Dissertation darstellt:

Satz (Verallgemeinerte Eilenberg-Korrespondenz [Ur17])**.** *Varietäten von* $\mathbf{T}$*-erkennbaren Sprachen stehen in bijektiver Korrespondenz zu Pseudovarietäten von* $\mathbf{T}$*-Algebren.*

Dank der abstrakten kategoriellen Modellierung der verwendeten Begriffe ist der Beweis dieses Satzes konzeptionell erstaunlich einfach. Er basiert auf zwei Beobachtungen: (i) Varietäten von $\mathbf{T}$-erkennbaren Sprachen können, unter Verwendung der verallgemeinerten Pippenger-Korrespondenz, als das *duale* Konzept zu proendlichen Theorien interpretiert werden, und (ii) nach dem verallgemeinerten Reiterman-Satz stehen proendliche Theorien in bijektiver Korrespondenz zu Pseudovarietäten von $\mathbf{T}$-Algebren. Die zentralen Ergebnisse unserer Arbeit werden somit durch das folgende Diagramm in Beziehung gesetzt:



Die kategorielle Eilenberg-Reiterman-Theorie liefert ein abstraktes und parametrisches Rahmenwerk für die algebraische Theorie der formalen Sprachen. Die erzielten Ergebnisse demonstrieren, dass die Schritte (2) bis (5) des "klassischen" Fünf-Punkte-Plans (siehe

Abschnitt 2) vollständig generisch sind: Nach einer anwendungsspezifischen Wahl der Monade $\mathbf{T}$ und ihrer unären Präsentation erhält man die Konstruktion von syntaktischen Algebren, die Konzepte einer Varietät von Sprachen und einer Pseudovarietät von Algebren sowie den Varietätensatz direkt aus unseren allgemeinen Ergebnissen.

4 Anwendungen

Die entwickelten kategoriellen Methoden sind auf verschiedene Typen von Sprachen und auf vielfältige Fragestellungen anwendbar. Als Instanzen des verallgemeinerten Eilenberg-Satzes erhalten wir rund ein Dutzend wichtige Ergebnisse aus der Literatur, darunter fünf Eilenberg-Sätze für reguläre Sprachen, zwei Eilenberg-Sätze für ω-reguläre Sprachen, zwei Eilenberg-Sätze für Sprachen über linearen Ordnungen, einen Eilenberg-Satz für Baumsprachen, und einen Eilenberg-Satz für Kostenfunktionen. Darüber hinaus konnten mehrere neue Eilenberg-Korrespondenzen abgeleitet werden, zum Beispiel eine Erweiterung des lokalen Varietätensatzes von Gehrke, Grigorieff und Pin [GGP08] von endlichen Wörtern auf unendliche Wörter, Bäume und Kostenfunktionen. Als weitere direkte Anwendung unserer Techniken haben sich mehrere neue Beiträge zur Theorie der klassischen regulären Sprachen ergeben, unter anderem eine Verallgemeinerung der Äquivalenz zwischen endlichen Automaten und Monoiden durch Interpretation dieser Konzepte über monoidalen Kategorien [AMU15], eine neue, rein automatentheoretische Interpretation des klassischen Eilenberg-Satzes via Algebra-Koalgebra-Dualität für endliche Automaten [Ad14, Ad15], sowie ein neuer algebraischer Zugang zur Konkatenation regulärer Sprachen durch Betrachtung monoidaler Adjunktionen [CU16]. Die kategorielle Perspektive hat somit auch zu neuen Einsichten über bereits umfangreich erforschte Strukturen geführt.

5 Fazit

Die in der Dissertation [Ur18] entwickelten Techniken tragen substanziell zu einem kategorientheoretischen Verständnis der algebraischen Theorie der formalen Sprachen bei. Durch die Allgemeinheit des eingeführten Rahmenwerks konnte demonstriert werden, dass die Schlüsselelemente dieser Theorie nicht inhärent algebraischer oder topologischer Natur sind, sondern durch abstrakte kategorielle Prinzipien beschrieben werden können. Diese Einsicht führt zu einer konzeptionellen Vereinfachung, Verallgemeinerung und Vereinheitlichung zahlreicher Ideen und Ergebnisse, die zuvor für verschiedene Sprachtypen separat betrachtet wurden. Die klare Trennung zwischen den generischen Aspekten der Theorie und ihrem anwendungsspezifischen Teil ermöglicht sowohl eine frische Sicht auf bekannte Strukturen als auch eine stark vereinfachte Herleitung neuer Ergebnisse.

Literatur

[Ad14] Adámek, J.; Milius, S.; Myers, R.; Urbat, H.: Generalized Eilenberg theorem I: Local Varieties of languages. In: Proc. 17th International Conference on Foundations of Software Science and Computation Structures. Jgg. 8412 in LNCS. Springer, S. 366–380, 2014.

[Ad15] Adámek, J.; Myers, R.; Milius, S.; Urbat, H.: Varieties of languages in a category. In: Proc. 30th Annual ACM/IEEE Symposium on Logic in Computer Science. IEEE, S. 414–425, 2015.

[AMU15] Adámek, J.; Milius, S.; Urbat, H.: Syntactic Monoids in a Category. In: Proc. 6th Conference on Algebra and Coalgebra in Computer Science. LIPIcs. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2015. Best Paper Award.

[Bo15] Bojańczyk, M.: Recognisable languages over monads. In: Proc. 19th International Conference on Developments in Language Theory, Jgg. 9168 in LNCS, S. 1–13. Springer, 2015.

[Ch16] Chen, L.-T.; Adámek, J.; Milius, S.; Urbat, H.: Profinite Monads, Profinite Equations and Reiterman's Theorem. In: Proc. 19th International Conference on Foundations of Software Science and Computation Structures. Jgg. 9634 in LNCS. Springer, S. 531–547, 2016.

[CU16] Chen, L.-T.; Urbat, H.: Schützenberger Products in a Category. In: Proc. 20th International Conference on Developments in Language Theory. Jgg. 9840 in LNCS. Springer, S. 89–101, 2016.

[Ei76] Eilenberg, S.: Automata, Languages, and Machines Vol. B. Academic Press, 1976.

[GGP08] Gehrke, M.; Grigorieff, S.; Pin, J.-É.: Duality and equational theory of regular languages. In: Proc. 35th International Colloquium on Automata, Languages and Programming, Part II. Jgg. 5126 in LNCS. Springer, S. 246–257, 2008.

[Pi95] Pin, J.-É.: A variety theorem without complementation. Russ. Math., 39:80–90, 1995.

[Pi97] Pippenger, N.: Regular languages and Stone duality. Th. Comp. Sys., 30(2):121–134, 1997.

[Re82] Reiterman, J.: The Birkhoff theorem for finite algebras. Algebra Universalis, 14(1):1–10, 1982.

[Sc65] Schützenberger, M. P.: On finite monoids having only trivial subgroups. Inform. and Control, 8:190–194, 1965.

[Ur17] Urbat, H.; Adámek, J.; Chen, L.-T.; Milius, S.: Eilenberg Theorems for Free. In: Proc. 42nd International Symposium on Mathematical Foundations of Computer Science. LIPIcs. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, S. 43:1–43:14, 2017. EATCS Best Paper Award.

[Ur18] Urbat, H.: A Categorical Approach to Algebraic Language Theory. Dissertation, Technische Universität Braunschweig, 2018.

Henning Urbat ist seit 2017 Postdoc am Lehrstuhl für Theoretische Informatik der Friedrich-Alexander-Universität Erlangen-Nürnberg. Zuvor studierte er Mathematik und Informatik an der Technischen Universität Braunschweig und promovierte am Institut für Theoretische Informatik von Prof. Dr. Jiří Adámek. Sowohl seine Diplomarbeit als auch seine Dissertation wurden von der Carl-Friedrich-Gauß-Fakultät der TU Braunschweig ausgezeichnet. Sein Forschungsschwerpunkt liegt in der Entwicklung kategorieller und (ko-)algebraischer Methoden in der Informatik, insbesondere in der Automatentheorie und ihren Anwendungen.