Algebraic Language Theory = Monads + Duality

Henning Urbat*

Institut für Theoretische Informatik
Technische Universität Braunschweig, Germany

Algebraic language theory studies machine behaviors by relating them to algebraic structures. For example, regular languages (the behaviors of finite automata) can be characterized purely algebraically as the languages recognized by finite monoids. A key result for the algebraic approach to regular languages is Eilenberg's *variety theorem*. It states that *varieties of languages* (classes of regular languages closed under boolean operations, derivatives[1], and preimages of monoid morphisms) correspond bijectively to *pseudovarieties of monoids* (classes of finite monoids closed under quotients, submonoids, and finite products).

In the past decades, the variety theory has developed into two orthogonal directions. First, Eilenberg-type results were discovered for classes of regular languages with weaker closure properties, e.g. dropping closure under complement. Second, the theory has been extended beyond languages of finite words, e.g. to weighted languages, languages of infinite words, languages of words on linear orderings, tree languages, and cost functions. This zoo of results has raised interest in categorical approaches to algebraic language theory, with the goal of deriving all these results from *one* general variety theorem. Two steps towards this, corresponding to the two orthogonal directions indicated above, were recently made by Bojańczyk [2] and by Adámek et. al. [1]. In [2] monoids are replaced by algebras for a *monad* on sorted sets, making it possible to model e.g. infinite words. In contrast, in [1] we keep monoids but consider them in categories $\mathscr{D}$ of (ordered) algebras such as posets, semilattices, and vector spaces. For example, monoids in the category $\mathbf{Vec}_K$ of vector spaces over a field K are K-algebras.

In this note we present our candidate for the *one* variety theorem. It puts a common roof over [1, 2] and covers many additional applications, including all the extensions of Eilenberg's theorem mentioned above. Our categorical framework is based on two parameters: (i) a pair of varieties $\mathscr{C}/\mathscr{D}$ of (ordered) algebras whose full subcategories on finite algebras are *dual*, and (ii) a *monad* $\mathbf{T}$ on the product category $\mathscr{D}^S$ for some finite set S of sorts. Varieties of languages live in $\mathscr{C}$, while the monad $\mathbf{T}$ represents their recognizing algebras. For example, for the original Eilenberg theorem for regular languages, one takes the pair $\mathbf{BA}/\mathbf{Set}$ (boolean algebras and sets) and the monad $\mathbf{T}\Sigma = \Sigma^*$ on $\mathbf{Set}$ representing monoids. For regular ∞-languages, one takes $\mathbf{BA}/\mathbf{Set}$ and the monad $\mathbf{T}(\Sigma, \Gamma) = (\Sigma^+, \Sigma^\omega + \Gamma)$ on $\mathbf{Set}^2$ representing ω-semigroups. Weighted languages over a finite field K are handled by $\mathbf{Vec}_K/\mathbf{Vec}_K$ and the monad $\mathbf{T}$ on $\mathbf{Vec}_K$ constructing free K-algebras.

---

* Based on joint work with Jiří Adámek, Liang-Ting Chen, and Stefan Milius. Preprint: `arxiv.org/abs/1602.05831`

[1] Recall that the *derivatives* of $L \subseteq \Sigma^*$ are the languages $a^{-1}L = \{w \in \Sigma^* : aw \in L\}$ and $La^{-1} = \{w \in \Sigma^* : wa \in L\}$ with $a \in \Sigma$.

We model languages categorically as morphisms $L\colon T\mathbb{\Sigma} \to O$, where O is a fixed finite "object of outputs" in $\mathscr{D}^S$, and $\mathbb{\Sigma}$ is the free $\mathscr{D}^S$-object on a finite S-sorted alphabet Σ in $\mathbf{Set}^S$. A language is called $\mathbf{T}$-*recognizable* if there is a homomorphism $h : \mathbf{T}\mathbb{\Sigma} \to A$ into a finite $\mathbf{T}$-algebra and a morphism $p\colon A \to O$ with $L = p \cdot h$. For the three monads above, this gives precisely the classical notions of recognizability for languages of finite words, ∞-languages, and weighted languages. Generalizing Eilenberg's concepts, a *pseudovariety of* $\mathbf{T}$*-algebras* is a class of finite $\mathbf{T}$-algebras closed under quotients, subalgebras, and finite products. A *variety of* $\mathbf{T}$*-recognizable languages* is a class of $\mathbf{T}$-recognizable languages closed under $\mathscr{C}$-algebraic operations (e.g. union, intersection and complement for $\mathscr{C} = \mathbf{BA}$), derivatives, and preimages of $\mathbf{T}$-homomorphisms. Here the *derivatives* of a language depend on a *unary presentation* of $\mathbf{T}$, viz. a set of unary operations that fully determine the structure of finite $\mathbf{T}$-algebras. For example, algebras for the monad $\mathbf{T}\Sigma = \Sigma^*$ on $\mathbf{Set}$ (= monoids) can be presented by the operations of left and right multiplication with fixed elements. We obtain the following

Variety Theorem. Varieties of $\mathbf{T}$-recognizable languages correspond bijectively to pseudovarieties of $\mathbf{T}$-algebras.

Our categorical framework is flexible and easily allows for introducing additional parameters. For example, the ω-semigroup monad $\mathbf{T}(\Sigma, \Gamma) = (\Sigma^+, \Sigma^\omega + \Gamma)$ on $\mathbf{Set}^2$ captures ∞-languages as subsets of $\mathbf{T}(\Sigma, \emptyset) = (\Sigma^+, \Sigma^\omega)$, so one needs to restrict to alphabets of the form $(\Sigma, \emptyset)$. One can make our notions of variety and pseudovariety, and the variety theorem, parametric in a subclass $\mathbb{A}$ of alphabets. Also, instead of considering languages recognized by *arbitrary* finite $\mathbf{T}$-algebras, one may need to restrict to a subclass $\mathcal{Q}$ of finite $\mathbf{T}$-algebras in some applications. Again, all concepts can be made parametric in $\mathcal{Q}$. This enables us e.g. to cover the recent variety theorem for cost functions of Daviaud, Kuperberg, and Pin [3].

The proper instantiation of the parameters is an application-specific task. For example, one needs to verify for a candidate monad $\mathbf{T}$ that the $\mathbf{T}$-recognizable languages coincide with the class of languages one has in mind (e.g. regular word languages, tree languages, or cost functions), and find a "small" unary presentation of $\mathbf{T}$. However, once the parameters are clear, our variety theorem does the rest of the work. By choosing different sets of parameters, it specializes to more than a dozen variety theorems in the literature. In addition, it produces several new Eilenberg-type correspondences, including an extension of the local variety theorem of Gehrke, Grigorieff and Pin [4] from finite to infinite words.

References

1. Adámek, J., Myers, R., Milius, S., Urbat, H.: Varieties of languages in a category. In: 30th Annual ACM/IEEE Symposium on Logic in Computer Science. IEEE (2015)
2. Bojańczyk, M.: Recognisable languages over monads. In: DLT'15, LNCS, vol. 9168, pp. 1–13. Springer (2015), extended version: `arxiv.org/abs/1502.04898`
3. Daviaud, L., Kuperberg, D., Pin, J.E.: Varieties of cost functions. In: STACS (2016)
4. Gehrke, M., Grigorieff, S., Pin, J.E.: Duality and equational theory of regular languages. In: ICALP'08, Part II. LNCS, vol. 5126, pp. 246–257. Springer (2008)