Regular Languages Defined by Generalized First-Order Formulas with a Bounded Number of Bound Variables

Howard Straubing\(^1\) and Denis Thérien\(^2\)

\(^1\) Computer Science Department, Boston College
Chestnut Hill, Massachusetts, USA 02467
\(^2\) School of Computer Science, McGill University
Montréal, Québec, Canada H3A2A7

Abstract. We give an algebraic characterization of the regular languages defined by sentences with both modular and first-order quantifiers that use only two variables.

1 Introduction

One finds in the theory of finite automata a meeting ground between algebra and logic, where difficult questions about expressibility can be classified, and very often effectively decided, by appeal to the theory of semigroups. This line of research began with the work of McNaughton and Papert [7], who showed that the languages definable by first-order sentences over ‘<’ are precisely the ‘star-free’ regular languages, and thus, by a theorem of Schützenberger, the languages whose syntactic monoids are aperiodic—that is, contain no nontrivial groups. Let us give an example of the kind of first-order formulas we are considering:

\[
\exists x \exists y (Q_x \land Q_y \land x < y \land \neg \exists z (x < z \land z < y))
\]

This sentence is meant to be interpreted in words over a specified finite alphabet \(\Sigma\) that contains the letter \(\sigma\). The variables in the sentence denote positions in the word (that is, integers between 1 and the length of the word, inclusive) and the subformula \(Q_x\) means ‘the letter in position \(x\) is \(\sigma\)’. The subformula \(x < y \land \neg \exists z (x < z \land z < y)\) says that position \(x\) is to the left of position \(y\), and that there is no position strictly between them (i.e., that \(y = x + 1\)), and thus the whole sentence says ‘there are two consecutive occurrences of \(\sigma\’). We say that the sentence defines a language over \(\Sigma\), namely the set of all strings that contain the factor \(\sigma\sigma\).

Since McNaughton and Papert’s work, researchers have investigated the expressibility of regular languages in various restrictions and extensions of first-order logic over <. For example, we can replace the predicate \(x < y\) by the (weaker) predicate \(y = x + 1\) (Beauquier and Pin [2]). We can permit the use of modular quantifiers, which count, modulo a fixed period, the number of positions of a string satisfying a given condition (Straubing, Thérien and Thomas [16]). For example, the sentence \(\exists^1 \mod 2xQ_x\) is interpreted to mean ‘the number of positions containing \(\sigma\) is congruent to 1 modulo 2’, and thus defines the set of strings containing an odd number of occurrences of \(\sigma\).
In all the cases considered, the family of regular languages obtained can be characterized in terms of the syntactic monoids or the syntactic morphisms of its members, and in most cases this characterization gives rise to an algebraic algorithm for deciding membership of a given language in the family. The book by Straubing [14] provides an exhaustive catalogue of such results.

Kamp [6] and later, Immerman and Kozen [5] showed that every first-order sentence over $<$ is equivalent to such a sentence in which only three variables are used. The number of bound variables that occur in a formula can be considered as a kind of expressibility resource, along with the kinds and depth of the quantifiers and the set of available atomic formulas.

Thérien and Wilke [17] considered the regular languages defined by sentences in which only two variables are used, and found that these, too, could be characterized in algebraic terms: A language $L$ is definable by a sentence with two variables if and only if its syntactic monoid belongs to a particular family $\mathbf{DA}$ of finite aperiodic monoids. (We will give the definition of $\mathbf{DA}$ in the next section.) It was already known that the two-variable definable languages are precisely those definable in the fragment of propositional temporal logic that includes both the past and future versions of the Next and Eventually operators, but excludes the Until operator (Etessami, Vardi and Wilke [4]). Since it is possible to determine from the multiplication table of a finite monoid whether it belongs to $\mathbf{DA}$, the Thérien-Wilke result provides an algorithm for determining whether a given regular language is definable in this fragment of temporal logic.

In the present paper we investigate the effect of bounding the number of variables in sentences that include the modular quantifiers $\exists r \mod n$ as well as ordinary first-order quantifiers, and we characterize, again in algebraic terms, the regular languages that are thereby defined.

We have shown that the three-variable property continues to hold for formulas that include modular quantifiers. That is,

**Theorem 1.** Let $\phi$ be a sentence over $<$ containing first-order and modular quantifiers. Then $\phi$ is equivalent to such a sentence with only three variables.

For formulas that contain only modular quantifiers, we have an even stronger result:

**Theorem 2.** Let $\phi$ be a sentence over $<$ in which only modular quantifiers appear. Then $\phi$ is equivalent to such a sentence with only two variables.

Our main theorem is that the languages $L$ defined by two-variable sentences are characterized by membership of their syntactic monoids $M(L)$ in the pseudovariety $\mathbf{DA} * \mathbf{G}_{sol}$, defined in Section 2:

**Theorem 3.** Let $\Sigma$ be a finite alphabet. A regular language $L \subseteq \Sigma^*$ is defined by a two-variable sentence over $<$ containing first-order and modular quantifiers, if and only if $M(L) \in \mathbf{DA} * \mathbf{G}_{sol}$.

It is important to remark that while our main theorem permits us in many individual cases to show that a language is, or is not, two-variable definable, the general problem of determining membership in $\mathbf{DA} * \mathbf{G}_{sol}$ is not known to be decidable.