RUDIMENTARY LANGUAGES AND SECOND-ORDER LOGIC

M. More, F. Olive

Abstract. The aim of this paper is to point out the equivalence between three notions respectively issued from recursivity theory, computational complexity and finite model theory. One the one hand, the rudimentary languages are known to be characterized by the linear hierarchy. On the other hand, this complexity class can be proved to correspond to monadic second-order logic with addition. Our viewpoint sheds some new light on the close connection between these domains: we bring together the two extremal notions by providing a direct logical characterization of rudimentary languages, and a representation result of second-order logic into these languages. We use natural arithmetical tools, and our proofs contain no ingredient from computational complexity.

Subject classifications. 03C13, 03D20.

Key words. Rudimentary sets, monadic second-order logic, linear hierarchy, spectra.

1. Introduction

This paper contains two related parts. The first part gives an arithmetical proof of a descriptive complexity result about rudimentary languages. The second part is devoted to a representation result of second-order logic into rudimentary sets.

Descriptive complexity is interested in providing logical characterizations for various complexity classes of languages, in the following sense: a class B of formulas is said to be a logical characterization of a class A of languages when any element of A corresponds to the set of all finite models of a formula in B and

conversely. For instance, the class of languages in NP is logically characterized by the class of existential second-order formulas (see Fa74). Similarly, the class of regular languages corresponds to monadic existential second-order logic with linear order (see Bü60), the class of non-deterministic linear-time languages is captured by the existential monadic second-order logic with addition (see Ly82) and the languages in Stockmeyer's polynomial hierarchy correspond to second-order logic (see St76 and Im87).

The Rudimentary relations are the relations over words (over some finite alphabet) obtained from concatenation using boolean operations, bounded quantifications and explicit transformations (see Qu46 and Sm61). The class of all rudimentary relations, seen as relations over IN (using the p-adic representation of integers) is independant of the alphabet (see Be62). Also, it was proved (see Sm61, Be62 and Ha73) that a relation over integers is rudimentary iff it is defined by a bounded arithmetical formula, i.e. a formula of signature $\{+, \times\}$, in which all quantifications are bounded by some polynomial of the free variables. For instance, the set of prime numbers is defined by the following bounded arithmetical formula: x is prime if and only if $x > 1 \land [\forall y < x \ \forall z < x \ \neg (x = y \times z)]$

In this paper, we show that the class of rudimentary languages (i.e. unary rudimentary relations) is logically characterized by monadic second-order logic with addition. Our proof uses arithmetical tools, but the result could also be obtained in an indirect way using complexity tools. Indeed, rudimentary languages also have a complexity characterization (see Wr78), and in turn the corresponding complexity class can be proved to correspond to monadic second-order logic with addition, using results of Ly82 and Im87.

The spectrum of a first-order sentence is the set of cardinalities of all finite models of this formula. It is known that rudimentary sets are spectra and that equality would imply $NP \neq co - NP$ and the collapsing of Stockmeyer's hierarchy (see Wo81). On the other hand, a converse representation result is known: we say that a set A of sets of integers is represented in a set B of sets of integers if there is a one-to-one function from IN to IN transforming any element of A into an element of B. If A is the spectrum of a sentence of k-ary signature then 2^{A^k} is a rudimentary set and the replacement of any of the exponential functions (for $k \geq 2$) by some polynomial function would imply that all spectra are rudimentary sets (see Mo94).

In this paper, we generalize the representation theorem above to a larger class of formulas and prove a converse result. The *(generalized) spectrum* of a second order formula is the set of the cardinalities of all finite models of this formula. We show that a set of integers A is the spectrum of a second-order

formula involving quantified relation symbols of arity $k \geq 2$ at most (with no extra-predicate) if and only if 2^{A^k} is a rudimentary set.

Finally, we list some consequences of the two previous results, concerning complexity theory and bounded arithmetics.

2. Definitions and Theorems

The purpose of this paragraph is to set up the main definitions formalizing the connexion between integers, words and finite structures. So doing, we give a precise meaning to the notion of *logical definability* of a language and we precisely settle our results.

Except in last section, we only consider languages over the dyadic alphabet.

DEFINITION 2.1. Let σ be a signature, an let k be a non-zero integer. We call $SO_k(\sigma)$ the set of second-order formulas of arity k and of signature σ . Hence, the formulas in $SO_k(\sigma)$ are of the form $Q_1R_1 \dots Q_pR_p\Psi$, where $Q_i \in \{\forall, \exists\}$, R_i is a k-ary relation symbol, and Ψ is a first-order formula of signature $\sigma \cup \{R_1, \dots, R_p\}$.

DEFINITION 2.2. Let n be a non-zero integer. We denote by ld(n) its dyadic length, i.e. we have $n = \sum_{i=0}^{ld(n)-1} n_i \cdot 2^i$ with $n_i \in \{1,2\}$. Note that $ld(n) \geq 1$ because n is non-zero. We denote by w_n the dyadic notation for n, i.e. the word $n_0 n_1 \dots n_{ld(n)-1}$.

DEFINITION 2.3. Let w be a finite word over $\{1,2\}$ and n be a non-zero integer such that $w = w_n$. We denote by \mathcal{Z}_w the unary relation over $\{0, \ldots ld(n) - 1\}$ defined by: $i \in \mathcal{Z}_w$ iff the digit of weight 2^i of w is 2. Let σ be a finite set of predefined relations. We denote by $\mathcal{R}_w(\sigma)$ the structure $\langle ld(n), \mathcal{Z}_w, \sigma \rangle$.

DEFINITION 2.4. Let k be a non-zero integer and let σ be a finite set of predefined relations. We say that a language $L \subseteq \{1,2\}^*$ is definable in $SO_k(\sigma)$ when there exists a formula Θ in $SO_k(\sigma \cup X)$ (where X is a unary relation symbol) such that: $w \in L$ iff $\mathcal{R}_w(\sigma) \models \Theta$.

DEFINITION 2.5. For each set of integers $A \subseteq IN$, we call dyadic language of A the set of dyadic notations of elements of A (denoted by $L_2(A)$).

DEFINITION 2.6. Let us denote by RUD the class of unary rudimentary relations over integers. The class of rudimentary languages, denoted by $\mathcal{L}_2(RUD)$, is the class of the dyadic languages of the rudimentary sets.

REMARK 2.7. In the sequel, order (denoted by <) and addition (denoted by +) are seen as predefined relations over the sets of the form $\{0, \ldots, n-1\}$.

Let us state our main theorem:

THEOREM 2.8. Let L be a language over $\{1,2\}$. L is definable in $SO_1(+)$ iff $L \in \mathcal{L}_2(RUD)$.

Then, theorem 2.8 is used to prove theorem 2.12. First, we need two more definitions.

DEFINITION 2.9. If F denotes a set of second-order formulas, let us denote by Sp(F) the class of spectra of formulas in this class.

Remark 2.10. First-order spectra correspond to $Sp(\exists SO(\emptyset))$.

DEFINITION 2.11. Let $f : \mathbb{N} \to \mathbb{N}$ and $S \subseteq \mathbb{N}$. We denote by $f(S) = \{f(s); s \in S\}$ the set of images under f of elements of S.

Theorem 2.12. For any set of integers S, the following equivalences hold:

- 1. for any integer $k \geq 1$, $S \in Sp(SO_k(+))$ iff $2^{S^k} \in RUD$;
- 2. for any integer $k \geq 2$, $S \in Sp(SO_k(\emptyset))$ iff $2^{S^k} \in RUD$.

3. Languages

In this section, theorem 2.8 is proved along lemmas 3.1 and 3.2.

LEMMA 3.1. Let A be a set of integers. If $L_2(A)$ is definable in $SO_1(+)$, then $A \in RUD$.

PROOF. The hypothesis of the lemma asserts that there exists a $SO_1(+)$ formula $\Theta(X, +) \equiv Q_1U_1 \dots Q_pU_pq_1x_1 \dots q_sx_s\psi(\overline{x}, \overline{U}, X, +)$ such that $w \in L_2(A)$ iff $\mathcal{R}_w(+) \models \Theta$. From the very definition of w_n , we can translate this hypothesis by: $n \in A$ iff $\langle lg(n), 2_n, + \rangle \models \Theta$.

In using Bennett's characterization of rudimentary sets, we have to construct a bounded arithmetical formula $\Phi(n)$ with one free variable such that $n \in A$ iff $\mathbb{N} \models \Phi(n)$. The main idea to proceed is to encode each unary relation U over ld(n) as an integer u smaller than n. Then we have to translate assertions about U (actually U(x), for a given x < ld(n)) into equivalent arithmetical assertions about u. At last, we can build the formula Φ , taking a specific care to the bounds of its quantifications. Let us describe briefly these three steps:

- Encoding. Let n be an integer and l = ld(n). Each subset U of l defines a single word w of length l such that 2w = U. Now, consider the integer u such that $w_u = w$. It is easy to check that the encoding thus obtained is a bijection from the set of unary relations over l onto the set of integers $[2^l 1; 2^{l+1} 2]$ (note that these integers are smaller than 2n).
- Translation. Let U and u be defined as above. Let x < l. Referring to the definition of u from U, observe that: U(x) holds in n iff $(\exists a, b, c < 2^{l+1})(u = a + b + cb \wedge 2^x 1 \le a < 2^{x+1} 1 \wedge b = 2^{x+1} \wedge c \ne 0)$ is true in IN. We denote by digit(u, x) this last formula. Now we are ready to built the bounded arithmetical formula $\Phi(n)$.
- Construction of Φ . The atomic formulas in Θ have the following four possible forms: $U_i(x_j)$, $X(x_j)$, $x_j = x_k$ and $x_j + x_k = x_\ell$ and $\Phi(n)$ is obtained from Θ by the following substitutions:

```
\begin{array}{ll} Q_i U_i & \text{becomes} \ \ Q_i u_i \in [2^l-1;2^{l+1}-2] \\ q_j x_j & \text{becomes} \ \ q_j x_j < l \\ U_i(x_j) & \text{becomes} \ \ digit(u_i,x_j) \\ X(x_j) & \text{becomes} \ \ digit(n,x_j) \\ x_j = x_k & \text{remains unchanged} \\ x_j + x_k = x_\ell & \text{remains unchanged} \end{array}
```

It remains to deal with l, which appears up to now like a predefined variable, and with exponential, which must not occur in a bounded arithmetical formula. First, note that the relation l = ld(n) is defined by the sentence: $(\exists L < 2n)(L = 2^l - 1 \land L \le n \le 2L)$. Finally, we get rid of the exponential using the fact that the graph of exponentiation is rudimentary (see Be62). \square

LEMMA 3.2. If a set of integers A is in RUD, then $L_2(A)$ is definable in $SO_1(+)$.

PROOF. This proof has the same pattern than the previous one but now we use Smullyan's definition of rudimentary sets instead of Bennett's one. Indeed, we interpret the hypothesis of lemma 3.2 as follows: there exists a bounded $\{*\}$ -formula $\Phi(n)$, such that $n \in A$ iff $\mathbb{N} \models \Phi(n)$.

- Encoding. Let $n \in \mathbb{N}$ and l = ld(n). To each integer u of dyadic length smaller than l, we associate the couple of unary relations $(\mathcal{1}_u, \mathcal{2}_u)$ over l defined by: $\forall x < l$, $\mathcal{1}_u(x)$ (resp. $\mathcal{2}_u(x)$) holds iff the letter of index x of w_u is 1 (resp. 2). Remark that the map $u \mapsto (\mathcal{1}_u, \mathcal{2}_u)$ is a bijection from $[0, 2^{l+1} 2]$ onto the set of couples (U, V) such that $U, V \subseteq l$, $U \cap V = \emptyset$, and $U \cup V$ is an initial segment of l.
- Translation. Now we have to translate the atomic formulas which constitute the formula $\Phi(n)$. Actually, we have to interpret the subformulas $u_i < n$ and $u_i * u_j = u_k$ (some of these u_i 's may be n). To proceed, we first translate without proof some intermediate assertions.
- There exists a formula $WECO_{< max}(U, V)$ such that for every couple (U, V) of unary relations over a domain n we have: $n \models WECO_{< max}(U, V)$ iff there exists an x < n such that $(1_x, 2_x) = (U, V)$.
- There exists a formula (U, U') * (V, V') = (T, T') such that for every l, n such that l = lg(n), for every u, v, t < l, we have $l \models u * v = t$ iff $n \models (\mathcal{1}_u, \mathcal{2}_u) * (\mathcal{1}_v, \mathcal{2}_v) = (\mathcal{1}_t, \mathcal{2}_t)$.
- Construction of Θ . Now, let $\tilde{\Theta}(1_n, 2_n, <, +)$ be the second-order formula obtained from Φ by the following substitutions:

$$\forall u_i < n \quad \text{becomes} \quad \forall \mathcal{1}_i \forall \mathcal{2}_i \ [WECO_{< n}(\mathcal{1}_i, \mathcal{2}_i) \rightarrow \exists u_i < n \quad \text{becomes} \quad \exists \mathcal{1}_i \exists \mathcal{2}_i \ [WECO_{< n}(\mathcal{1}_i, \mathcal{2}_i) \land u_i * u_j = u_k \quad \text{becomes} \quad (\mathcal{1}_i, \mathcal{2}_i) * (\mathcal{1}_j, \mathcal{2}_j) = (\mathcal{1}_k, \mathcal{2}_k)$$

Note that, although it doesn't appear explicitly in our notations, the linear order < and the relations I_n and \mathcal{Z}_n are in the signature of $\tilde{\Theta}$: the relation < is hidden in the formulas $WECO_{< n}$ and I_n and I_n and I_n appear with the translations of the subformulas $u_i * u_j = u_k$ in wich some of the variables are n. But < is trivially expressed with addition, and I_n can be replaced by I_n . All in all, the sought formula I_n is obtained from I_n in expliciting I_n with I_n and I_n by I_n . I_n

4. Proof of the equivalence between exponentiations of spectra and rudimentary sets

In this section, we prove theorem 2.12. Precisely, we first prove the first item and then deduce the second one. The proof consists in two lemmas respectively corresponding to each implication.

LEMMA 4.1. If $S \in Sp(SO_k(+))$, then $2^{S^k} \in RUD$.

PROOF. The main idea is to encode each R_i into an integer c_i smaller than $2^{n^k+1}-2$, and to replace second-order quantification over R_i by a bounded first-order quantification over c_i . We do not give a detailed proof of this lemma, because it is easy and similar to the proof given in Mo94. \square

LEMMA 4.2. If $L_2(2^{S^k})$ is definable in $SO_1(+)$, then $S \in Sp(SO_k(+))$.

PROOF. Let S be a set of integers such that $L_2(2^{S^k})$ is definable in $SO_1(+)$. Let $\Psi \equiv Q_1V_1 \dots Q_pV_pq_1y_1 \dots q_ry_r\Phi$ be a $SO_1(+)$ -formula (where Φ is a first-order open formula of signature $\{=,+,X,V_1,\dots,V_p\}$ and variables $y_1 \dots y_r$) such that $w_m \in L_2(2^{S^k})$ iff $\langle ld(m),=,2_{w_m},+\rangle \models \Psi$.

such that $w_m \in L_2(2^{S^k})$ iff $\langle ld(m), =, 2_{w_m}, + \rangle \models \Psi$. Since $m \in 2^{S^k}$, there exists $n \in S$ such that $m = 2^{n^k}$. Consequently, w_m is $1^{n^k-1}2$. Hence, we have to change a structure of size $ld(m) = n^k$ into a structure of size n, to cancel the second-order unary free variable X and to replace the addition over $\{0, \ldots, n^k - 1\}$ by the addition over $\{0, \ldots, n - 1\}$. Formally, we have to construct a $SO_k(+)$ -formula Δ such that $n \in S$ iff $\langle n, =, + \rangle \models \Delta$.

To change the cardinality of the domain, we encode each integer $t < n^k$ into a k-tuple of integers $t_i < n$ (notation in basis n), and thus naturally translate every unary relation V_i over n^k into a k-ary relation R_i over n. We cancel the free variable X (interpreted by $2_{w_m} = \{n^k - 1\}$) by introducing a new second-order k-ary variable R_0 satisfying the condition $VALR_0$ expressing

the fact that $n^k - 1 = \sum_{i=0}^{k-1} (n-1).n^i$. We replace u + v = w (addition over $\{0, \ldots, n^k - 1\}$) by $ADD'(u_0, \ldots, u_{k-1}, v_0, \ldots, v_{k-1}, w_0, \ldots, w_{k-1})$ (k additions over $\{0, \ldots, n-1\}$), i.e. we compute in basis n, and we deal by hand with the carries.

Then we apply on the formula Ψ the following transformations:

- we add the second-order quantification $\exists R_0$ over a k-ary relation;
- each second-order quantification over a unary relation Q_iV_i is replaced by the same second-order quantification Q_iR_i over a k-ary relation;
- each first-order quantification $q_i y_i$ is replaced by the k first-order quantifications $q_i y_{i,1} \ldots q_i y_{i,k}$;

• the first-order open formula Φ is changed into a first-order formula of signature $\{=,+,R_0,R_1,\ldots,R_p\}$ denoted by Γ and obtained by replacing $y_i=y_j$ by $\bigwedge_{s=1}^k (y_{i,s}=y_{j,s}), \ y_i+y_j=y_l$ by $ADD'(y_{i,1},\ldots,y_{i,k},y_{j,1},\ldots,y_{j,k},y_{l,1},\ldots,y_{l,k}),$ $X(y_j)$ by $R_0(y_{j,1},\ldots,y_{j,k}),$ and $V_i(y_j)$ by $R_i(y_{j,1},\ldots,y_{j,k}).$

Finally, the formula Δ is $\exists R_0 \ Q_1 R_1 \dots Q_p R_p \ q_1 y_{1,0} \dots q_r y_{r,k-1} (VALR_0 \wedge \Gamma)$. \Box

The proof of the first item of theorem 2.12 is now obtained via theorem 2.8. In order to prove the second item, the remaining task is to eliminate the predefined addition. Nothing changes in the proof of lemma 4.1 if one is interested into spectra in $Sp(SO_k(\emptyset))$ instead of $Sp(SO_k(+))$. But, in the proof of lemma 4.2, we actually use the addition over the structure $\langle n, =, + \rangle$. To overcome this difficulty, we just have to invoke the following result:

THEOREM 4.3. (Ly82) The addition is first-order definable with binary relations.

5. Consequences

Let us state first a corollary in the field of complexity.

COROLLARY 5.1. $\mathcal{L}_2(RUD)$ contains many natural NP-complete languages.

PROOF. Many natural NP-complete languages, in which the twenty-one listed by Ka72, are known to belong to NLIN, a complexity class introduced by E. Grandjean to formalize linear time complexity on RAM's. But it is also proved in GrOl95 that all languages in NLIN are definable in $\exists SO_1(+)$, so that the above mentioned languages are in $\mathcal{L}_2(\text{RUD})$. \square

Now, let us turn to descriptive complexity. The particular role played by the dyadic representation of integers in theorem 2.8 can easily be avoided. First, let us note that it is almost straightforward to adjust our definitions to the *p*-adic case. We only precise the following one.

DEFINITION 5.2. For each word w over $\{1, 2, ..., p\}$, we still denote by \mathcal{R}_w the finite structure $\langle l, 1_w, ..., p_w \rangle$, where l is the length of w and $1_w, ..., p_w$ are the unary relations over the domain $l = \{0, ..., l-1\}$ defined by: for all s = 1 to p, for all x < l, $s_w(x)$ holds iff the letter of index x of w is s.

Now, we say that a language $L \subseteq \{1, \ldots, p\}^*$ is definable in $SO_1(+)$ if there exists a formula Θ (with p free unary second-order variables X_1, \ldots, X_p) in $SO_1(+)$ such that for all $w \in \{1, \ldots, p\}^*$ we have: $w \in L$ iff $\mathcal{R}_w(+) \models \Theta(X_1, \ldots, X_p)$ (where X_s is interpreted by s_w in $\mathcal{R}_w(+)$). Let us note that for all integers $p \geq 2$ and $p \geq 1$, there is a bijection between the integers of p-adic length $p \geq 1$ and the p-tuples of unary relations providing a partition of an initial segment of $p \geq 1$. Then, the proof of the following result strickingly follows the one of theorem 2.8.

COROLLARY 5.3. For any integers $p \ge 2$ and $q \ge 2$,

- 1. a language $L \subseteq \{1, \ldots, p\}^*$ is definable in $SO_1(+)$ iff $L \in \mathcal{L}_p(RUD)$.
- 2. Let $A \subseteq \mathbb{N}$. Then $L_p(A)$ is definable in $SO_1(+)$ iff $L_q(A)$ is definable in $SO_1(+)$.

Theorem 2.12 leads to a similar result in the scope of spectra. The proof is straightforward, since RUD is closed under inverse ranges by polynomials.

COROLLARY 5.4. For any integers $p \ge 2$ and $k \ge 1$, for any set of integers A, the following equivalence holds: $A \in Sp(SO_k(+))$ iff $p^{A^k} \in RUD$.

Finally, let us note that, according to theorem 2.12, generalized Fagin's question (see Fa75): does $Sp(SO_k(\emptyset)) = Sp(SO_2(\emptyset))$ for all $k \geq 2$? is equivalent to the following question: for all $k \geq 2$, for all set S of integers, does $2^{S^k} \in \text{RUD}$ iff $2^{S^2} \in \text{RUD}$? Remark that we know no such equivalence for Fagin's question: does $Sp(\exists SO_k(\emptyset)) = Sp(\exists SO_2(\emptyset))$ for all k > 2?

Acknoledgements

The authors wish to express their thanks to E. Grandjean for suggesting the problem and for many stimulating conversations.

References

Be62: J.H. Bennett. *On Spectra*. Doctoral Dissertation, Princeton University, Princeton, N.J., 1962.

Bü60: J.R. Büchi. Weak second order arithmetic and finite automata. Z.Math. Logik Grundlagen Math. 6, 1960, pp.66-92.

Di80: C. Dimitracopoulos. *Matijasevic's Theorem and Fragments of Arithmetic*. Ph. D. Thesis, University of Manchester, 1980.

Fa74: R. Fagin. Generalized First-Order Spectra and Polynomial-Time Recognizable Sets. SIAM-Proceedings, 7, 1974, p. 43-72.

Fa75: R. Fagin. A Spectrum Hierarchy. Math. Logik Grundlagen Math., 21, 1975, p. 123-134.

Gr85: E. Grandjean. Universal Quantifiers and Time Complexity of Random Access Machines. Math. Systems Theory, 18, 1985, p. 171-187.

GrOl: E.Grandjean and F.Olive. *Monadic Logical Definability of NP-Complete Problems*, CSL'94 Proc. LNCS 933 (1995), pp 190-204.

Ha73: K. Harrow. Sub-elementary Classes of Functions and Relations. Doctoral Dissertation, New-York University, Department of Mathematics, 1973.

Im87: N. Immerman. Languages that Capture Complexity Classes. SIAM J. Comput., vol.16, No.4, 1987, pp. 760-777.

Ka72: R.M. Karp. Reducibility among combinatorial problems. IBM Symp.1972, Complexity of Computers Computations, Plenum Press, New York, 1972.

Ly82: J.F. Lynch. Complexity Classes and Theories of Finite Models. Math. Systems Theory, 15, 1982, pp. 127-144.

Mo94: M. More. *Rudimentary representations of spectra*. prépublication du LLAIC1, Numéro 40, 1994.

Qu46: W. Quine. Concatenation as a Basis for Arithmetic from a Logical Point of View. J. Symb. Logic, 11, 1946, p. 105-114. Selected Logic Papers, ch V, Random House, 1966, p. 70-82.

Sm61: R. Smullyan. *Theory of Formal Systems*. Annals of Math. Studies, 47, Princeton University Press, Princeton, N.J., 1961.

St76: L.J. Stockmeyer. *The polynomial time hierarchy*. Theor. Comput. Sci. 3, 1976, p. 1-22.

Wo81: A. Woods. Some Problems in Logic and Number Theory, and their Connections. Doctoral Dissertation, University of Manchester, 1981.

Wr78: C. Wrathall. Rudimentary predicates and relative computation. SIAM J. Comput. vol. 7 no.2, 1978, p. 194-209.

Malika MORE. GRAL, dépt. de mathématiques, Université de Caen, 14032 CAEN cedex, FRANCE. Malika.More@math.unicaen.fr Frédéric OLIVE. GREYC, dépt. d'informatique, Université de Caen, 14032 CAEN cedex, FRANCE. olive@logique.jussieu.fr