Conjunctive Grammars Can Generate Non-regular Unary Languages

Artur Jeż*

Institute of Computer Science,
ul. Joliot-Curie 15, 50-383 Wroclaw, Poland
aje@ii.uni.wroc.pl
http://www.ii.uni.wroc.pl/~aje

Abstract. Conjunctive grammars were introduced by A. Okhotin in [1] as a natural extension of context-free grammars with an additional operation of intersection in the body of any production of the grammar. Several theorems and algorithms for context-free grammars generalize to the conjunctive case. Still some questions remained open. A. Okhotin posed nine problems concerning those grammars. One of them was a question, whether a conjunctive grammar over unary alphabet can generate only regular languages. We give a negative answer, contrary to the conjectured positive one, by constructing a conjunctive grammar for the language $\{a^{4^n} : n \in \mathbb{N}\}$. We then generalise this result—for every set of numbers L such that their representation in some k-ary system is regular set we show that $\{a^{k^n} : n \in L\}$ is generated by some conjunctive grammar over unary alphabet.

Keywords: Conjunctive grammars, regular languages, unary alphabet, non-regular languages.

1 Introduction and Background

Alexander Okhotin introduced conjunctive grammars in [1] as a simple, yet beautiful and powerful extension of context-free grammars. Informally speaking, conjunctive grammars allow additional use of intersection in the body of any rule of the grammar. More formally, conjunctive grammar is defined as a quadruple $\langle \Sigma, N, P, S \rangle$ where Σ is a finite alphabet, N is a set of nonterminal symbols, S is a starting nonterminal symbol and P is a set of productions of the form:

$$A \to \alpha_1 \& \alpha_2 \& \ldots \& \alpha_k, \quad \text{where } \alpha_i \in (\Sigma \cup N)^* .$$

Word w can be derived by this rule if and only if it can be derived from every string α_i for $i = 1, \ldots, k$.

We can also give semantics of conjunctive grammars with language equations that use sum, intersection and concatenation as allowed operations. Language

* Research supported by MNiSW grant number N206 024 31/3826, 2006-2008.

T. Harju, J. Karhumäki, and A. Lepistö (Eds.): DLT 2007, LNCS 4588, pp. 242–253, 2007.

generated by conjunctive grammar is a smallest solution of such equations (or rather one coordinate of the solution, since it is a vector of languages).

The usage of intersection allows us to define many natural languages that are not context-free. On the other hand it can be shown [1] that languages generated by the conjunctive grammars are deterministic context-sensitive.

Since in this paper we need only a small piece of theory of conjunctive grammars, we give an example of a grammar, together with language generated by it, instead of formal definitions. The Reader interested in the whole theory of the conjunctive grammars should consult [1] for detailed results or [2] for shorter overview. Also work on the Boolean grammars [3], which extend conjunctive grammars by additional use of negation, may be interesting.

Example 1. Let us consider conjunctive grammar $\langle \Sigma, N, P, S \rangle$ defined by:

$$\begin{aligned} \Sigma &= \{a, b, c\} \ , \\ N &= \{S, B, C, E, A\} \ , \\ P &= \{A \to aA \mid \epsilon, C \to Cc \mid \epsilon, S \to (AE)\&(BC), \\ &\quad\ B \to aBb \mid \epsilon, E \to bEc \mid \epsilon\} \ . \end{aligned}$$

The language generated by this grammar is equal to $\{a^n b^n c^n : n \in \mathbb{N}\}$. The associated language equations are:

$$\begin{aligned} L_A &= \{a\}L_A \cup \{\epsilon\} \ , \\ L_C &= \{c\}L_C \cup \{\epsilon\} \ , \\ L_S &= (L_A L_E) \cap (L_B L_C) \ , \\ L_B &= \{a\}L_B\{b\} \cup \{\epsilon\} \ , \\ L_E &= \{b\}L_E\{c\} \cup \{\epsilon\} \ . \end{aligned}$$

Their smallest solution is:

$$\begin{aligned} L_A &= a^* \ , \\ L_C &= c^* \ , \\ L_S &= \{a^n b^n c^n : n \in \mathbb{N}\} \ , \\ L_B &= \{a^n b^n : n \in \mathbb{N}\} \ , \\ L_E &= \{b^n c^n : n \in \mathbb{N}\} \ . \end{aligned}$$

Many natural techniques and properties generalize from context-free grammars to conjunctive grammars. Among them most important are: existence of the Chomsky normal form, parsing using a modification of CYK algorithm *etc.* On the other hand many other techniques do not generalize—there is no Pumping Lemma for conjunctive grammars, they do not have bounded growth property, non-emptiness is undecidable. In particular no technique for showing that a language cannot be generated by conjunctive grammars is known; in fact, as for today, we are only able to show that languages that are not context sensitive lay outside this class of languages.

A. Okhotin in [4] gathered nine open problems for conjunctive and Boolean grammars considered to be the most important in this field. One of those problem was a question, whether conjunctive grammars over unary alphabet generate only regular languages. It is easy to show (using Pumping Lemma), that this is true in case of context-free grammars. The same result was conjectured for conjunctive grammars. We disprove this conjecture by giving conjunctive grammar for a non-regular language $\{a^{4^n} : n \in \mathbb{N}\}$.

The set $\{4^n : n \in \mathbb{N}\}$ written in binary is a regular language. This leads to a natural question, what is the relation between regular (over binary alphabet) languages and unary conjunctive languages. We prove that every regular language (written in some k-ary system) interpreted as a set of numbers can be represented by a conjunctive grammar over an unary alphabet.

2 Main Result—Non-regular Conjunctive Language over Unary Alphabet

Since we deal with an unary alphabet we identify word a^n with number n and work with sets of integers rather than with sets of words.

Let us define the following sets of integers:

$$\begin{aligned} A_1 &= \{1 \cdot 4^n : n \in \mathbb{N}\} , \\ A_2 &= \{2 \cdot 4^n : n \in \mathbb{N}\} , \\ A_3 &= \{3 \cdot 4^n : n \in \mathbb{N}\} , \\ A_{12} &= \{6 \cdot 4^n : n \in \mathbb{N}\} . \end{aligned}$$

The indices reflect the fact that these sets written in tetrary system begin with digits $1, 2, 3, 12$, respectively and have only 0'es afterwards. We will show that those sets are the minimal solution of the equations:

$$B_1 = (B_2B_2 \cap B_1B_3) \cup \{1\} , \tag{1}$$
$$B_2 = (B_{12}B_2 \cap B_1B_1) \cup \{2\} , \tag{2}$$
$$B_3 = (B_{12}B_{12} \cap B_1B_2) \cup \{3\} , \tag{3}$$
$$B_{12} = (B_3B_3 \cap B_1B_2) . \tag{4}$$

Where in the above equations XY reflects the concatenation of languages:

$$XY := \{x + y : x \in X, y \in Y\} .$$

This set of language equations can be easily transformed to a conjunctive grammar over unary alphabet (we should specify the starting symbol, say B_1). None of the sets A_1, A_3, A_3, A_{12} is a regular language over unary alphabet.

Since solutions are vectors of languages we use notation

$$(A_1, \ldots, A_n) \subset (B_1, \ldots, B_n) ,$$

meaning, that $A_i \subset B_i$ for $i = 1, \ldots, n$.

Since we often prove theorems be induction on number of digits, it is convenient to use the following notation for language (set) S:

$$S \restriction_n := \{s \in S : s \text{ has } n \text{ digits at the most}\} .$$

We shall use it also for the vectors of languages (sets) with an obvious meaning.

Lemma 1. *Every solution* (S_1, S_2, S_3, S_{12}) *of equations* (1)–(4) *satisfies:*

$$(A_1, A_2, A_3, A_{12}) \subset (S_1, S_2, S_3, S_{12}). \tag{5}$$

Proof. We shall prove by induction on n, that

$$(A_1, A_2, A_3, A_{12}) \restriction_n \subset (S_1, S_2, S_3, S_{12}) .$$

For $n = 1$ we know that $i \in S_i$ by (1), (2) and (3). This ends induction basis.
For induction step let us assume that

$$(A_1, A_2, A_3, A_{12}) \restriction_{n+1} \subset (S_1, S_2, S_3, S_{12}).$$

We shall prove that this is true also for $(n + 2)$.

Let us start with 4^{n+1} and S_1. By induction assumption $2 \cdot 4^n \in S_2$ and hence $(2 \cdot 4^n) + (2 \cdot 4^n) = 4^{n+1} \in S_2S_2$. Also by induction assumption $4^n \in S_1$ and $3 \cdot 4^n \in S_3$, hence $(4^n) + (3 \cdot 4^n) = 4^{n+1} \in S_1S_3$, and so $4^{n+1} \in S_2S_2 \cap S_1S_3$ and by (1) we conclude that $4^{n+1} \in S_1$.

Similar calculations can be made for other $(n + 2)$-digit numbers, we present them in simplified way.

For $6 \cdot 4^n$, which is a $(n+2)$-digit number, we can see that $3 \cdot 4^n \in S_3$, $2 \cdot 4^n \in S_2$ by induction hypothesis and $1 \cdot 4^{n+1} = 4 \cdot 4^n \in S_1$, which was proved already in induction step. Hence $6 \cdot 4^n \in S_3S_3 \cap S_1S_2$ and by (4) we get $6 \cdot 4^n \in S_{12}$.

For $2 \cdot 4^{n+1}$ notice that $2 \cdot 4^n \in S_2$, $6 \cdot 4^n \in S_{12}$ and $1 \cdot 4^{n+1} \in S_1$ hence $2 \cdot 4^{n+1} \in S_1S_1 \cap S_{12}S_2$ and by (2) $2 \cdot 4^{n+1} \in S_2$.

For $3 \cdot 4^{n+1}$ notice that $2 \cdot 4^{n+1} \in S_2$, $6 \cdot 4^n \in S_{12}$ and $1 \cdot 4^{n+1} \in S_1$ hence $3 \cdot 4^{n+1} \in S_{12}S_{12} \cap S_1S_2$ and by (3) $3 \cdot 4^{n+1} \in S_3$.

This ends induction step. □

Lemma 2. *Sets* A_1, A_2, A_3, A_{12} *are a solution of* (1)–(4).

Proof. For every (1)–(4) we have to prove two inclusions: $\subset$ (that is

$$\begin{aligned} A_1 &\subset (A_2A_2 \cap A_1A_3) \cup \{1\} , \\ A_2 &\subset (A_{12}A_2 \cap A_1A_1) \cup \{2\} , \\ A_3 &\subset (A_{12}A_{12} \cap A_1A_2) \cup \{3\} , \\ A_{12} &\subset (A_3A_3 \cap A_1A_2) . \end{aligned}$$

And $\supset$, that is

$$\begin{aligned} A_1 &\supset (A_2A_2 \cap A_1A_3) \cup \{1\} , \\ A_2 &\supset (A_{12}A_2 \cap A_1A_1) \cup \{2\} , \\ A_3 &\supset (A_{12}A_{12} \cap A_1A_2) \cup \{3\} , \\ A_{12} &\supset (A_3A_3 \cap A_1A_2) . \end{aligned}$$

For the inclusions $\subset$ the proof is the same as in Lemma 1 and so we skip it.

For the inclusions $\supset$, let us consider m—the smallest number that violates this inclusion, that is m belongs to some right-hand side of one of the (1)–(4), but does not belong to the corresponding left-hand side. First note the easy, but crucial, fact that $m \neq 0$ because all numbers appearing on the right-hand side are strictly greater than 0.

Suppose m violates (1). Then $m \in A_2A_2$. By definition there are numbers k, l such that $k, l \in A_2$ and $m = k + l$. Since m is the smallest such number then both k, l belong also to the left-hand side, hence $k = 2 \cdot 4^{n_1}$ and $l = 2 \cdot 4^{n_2}$. Then either $k = l$ and $m \in A_1$ and so it does not violate the inclusion or $k \neq l$ and so m have only two non-zero digits, both being 2's. On the other hand $m \in A_1A_3$, so by definition there are $k' \in A_1$, $l' \in A_3$ such that $m = l' + k'$. Since m is the smallest such number then $l' = 1 \cdot 4^{n_3}$ and $k' = 3 \cdot 4^{n_4}$. And so either $m = 4^{n_3+1}$, if $n_3 = n_4$, or m has only two non-zero digits: 1 and 3. But this is a contradiction with a claim that m has only two non-zero digits, both being $2's$.

We shall deal with other cases in the same manner.

Suppose m violates (2). Then $m \in A_1A_1$. By definition there are numbers k, l such that $k, l \in A_1$ and $m = k + l$. Since m is the smallest such number then both k, l belong also to the left-hand side, hence $k = 1 \cdot 4^{n_1}$ and $l = 1 \cdot 4^{n_2}$. Then either $k = l$ and $m \in A_2$ and so it does not violate the inclusion or $k \neq l$ and so m have only two non-zero digits, both being 1's. On the other hand $m \in A_{12}A_2$, so by definition there are $k' \in A_{12}$, $l' \in A_2$ such that $m = l' + k'$. Since m is the smallest such number then $l' = 6 \cdot 4^{n_3}$ and $k' = 2 \cdot 4^{n_4}$. And so either $m = 2 \cdot 4^{n_3+1}$, if $n_3 = n_4$, or m has exactly two non-zero digits: 3 and 2 or it has exactly three non-zero digits $1, 2, 2$. But this is a contradiction with a claim that m has only two non-zero digits, both being 1's.

Suppose m violates (3). Then $m \in A_{12}A_{12}$. By definition there are numbers k, l such that $k, l \in A_{12}$ and $m = k+l$. Since m is the smallest such number then both k, l belong also to the left-hand side, hence $k = 6 \cdot 4^{n_1}$ and $l = 6 \cdot 4^{n_2}$. Then either $k = l$ and $m \in A_3$ and so it does not violate the inclusion or $k \neq l$ and so m can have the following multisets of non-zero digits: $\{1, 1, 2, 2\}$ or $\{1, 2, 3\}$. On the other hand $m \in A_1A_2$, so by definition there are $k' \in A_1$, $l' \in A_2$ such that $m = l' + k'$. Since m is the smallest such number then $l' = 1 \cdot 4^{n_3}$ and $k' = 2 \cdot 4^{n_4}$. And so either $m = 3 \cdot 4^{n_3}$, if $n_3 = n_4$, or m has exactly two non-zero digits: 1 and 2. But this is a contradiction with a previous claim on possible sets of non-zero digits of m.

Suppose it violates (4). Then $m \in A_3A_3$. By definition there are numbers k, l such that $k, l \in A_3$ and $m = k + l$. Since m is the smallest such number then both k, l belong also to the left-hand side, hence $k = 3 \cdot 4^{n_1}$ and $l = 3 \cdot 4^{n_2}$. Then either $k = l$ and $m \in A_{12}$ and so it does not violate the inclusion or $k \neq l$ and so m have only two non-zero digits, both being 3's. On the other hand $m \in A_1A_2$, so by definition there are $k' \in A_1$, $l' \in A_2$ such that $m = l' + k'$. Since m is the smallest such number then $k' = 4^{n_3}$ and $l' = 2 \cdot 4^{n_4}$. And so either m has only two non-zero digits: 1 and 2 or it has exactly one non-zero digit—3. But this is a contradiction with a claim that m has only two non-zero digits, both being $3's$. □

Theorem 1. *Sets A_1, A_2, A_3, A_{12} are the smallest solution of* (1)–(4).

Proof. This follows from Lemma 1 and Lemma 2. □

Corollary 1. *The non-regular language* $\{a^{4^n} : n \in \mathbb{N}\}$ *can be generated by conjunctive grammar over unary alphabet.*

Corollary 2. *Conjunctive grammars over unary alphabet have more expressive power then context-free grammars.*

3 Additional Results

3.1 Number of Nonterminals Required

The grammar described in the previous section uses four nonterminals. It can be easily converted to Chomsky normal form, we need only to introduce two new nonterminals for languages $\{1\}$ and $\{2\}$ respectively, hence grammar for language $\{4^n : n \in \mathbb{N}\}$ in Chomsky normal form requires only six nonterminals. It is an interesting question, which mechanisms of conjunctive grammars and how many of them are required to generate a non-regular language? How many nonterminals are required? How many of them must generate non-regular languages? How many intersections are needed? Putting this question in the other direction, are there any natural sufficient conditions for a conjunctive grammar to generate regular language?

It should be noted that we are able to reduce the number of nonterminals to three, but we sacrifice Chomsky normal form and introduce also concatenations of three nonterminals in productions. This can be seen as some trade-off between number of nonterminals and length of concatenations. Let us consider a language equation:

$$B_1 = (B_{2,12}B_{2,12} \cap B_1B_3) \cup \{1\} , \tag{6}$$

$$B_{2,12} = \big((B_{2,12}B_{2,12} \cap B_1B_1) \cup \{2\}\big) \cup \big((B_3B_3 \cap B_{2,12}B_{2,12})\big) , \tag{7}$$

$$B_3 = (B_{2,12}B_{2,12} \cap B_1B_1B_1) \cup \{3\} . \tag{8}$$

These are basically the same equations as (1)–(4), except that nonterminals B_2 and B_{12} are identified (or merged) and also conjunct B_2B_1 in (1)–(4) was changed to $B_1B_1B_1$. The Reader can easily check that after only the second change applied to (1)–(4) proofs of Lemma 1 and Lemma 2 can be easily modified to work with new situation.

Theorem 2. *The smallest solution of* (6)–(8) *is*

$$(A_1,\ A_2 \cup A_{12},\ A_3) .$$

Proof. The proof of this theorem is just a slight modification of the proof of Theorem 1, so we shall just sketch it.

The main idea of the proof is to think of nonterminal $B_{2,12}$ that corresponds to the set $A_2 \cup A_{12}$ as two nonterminals: B_2 and B_{12}, corresponding to sets

A_2 and A_{12}, respectively. To implement this approach we should show that every occurrence of $B_{2,12}$ on the right-hand side of any equation can be replaced by exactly one of the nonterminals B_2 or B_{12}, meaning that in each language equation from (6)–(8) if we substitute the variables with intended solution every occurrence of set $A_2 \cup A_{12}$ on the right hand-side can be in fact replaced by exactly one of the sets A_2, A_{12} with keeping the equations true.

Let us consider sets $(A_1,\ A_2 \cup A_{12},\ A_3)$. Using the same arguments as in Lemma 1 we can show that the smallest solutions of (6)–(8) is a pointwise superset of considered sets.

Now we must show that these sets are in fact a solution. Showing $\subset$ is easy, as in Lemma 2. Showing $\supset$ is the same as in Lemma 2 if we show earlier that we can substitute every occurrence of $B_{2,12}$ on the right-hand side with exactly one of B_2 or B_{12}, respectively.

Notice that if we have an intersection of the form $B_iB_j \cap B_kB_l$ then the sum of tetrary digits from B_i and B_j must be equal modulo 3 the sum of tetrary digits from B_k and B_l (this observation is easily generalized to the case with more concatenated nonterminals). We take sums modulo 3 because if we sum two numbers and digits in a column sum up to 4 (or more) then we loose 4 in this column but gain 1 in the next, so the difference is 3. Now if we swap from B_2 to B_{12} then the sum of digits changes by 1. So to get an equation modulo 3 we would have to add and subtract some 1's from both sides. Case inspection shows that this is not possible. And so we can use the same arguments as in Lemma 2. □

3.2 Related Languages

Theorem 3. *For every natural number k, language*

$$\{k^n : n \in \mathbb{N}\}$$

is generated by a conjunctive grammar over unary alphabet.

Proof. For every $k > 4$ we introduce non-terminals $B_{i,j}$, where $i = 1, \dots k-1$ and $j = 0, \dots, k-1$, with intention that $B_{i,j}$ defines language of numbers beginning with digits i, j and then only zeroes in k-ary system of numbers. Then we define the productions as:

$$\begin{aligned}
B_{1,0} &\to B_{2,0}B_{k-2,0} \ \&\ B_{1,0}B_{k-1,0}\ , \\
B_{1,1} &\to B_{1,0}B_{1,0} \ \&\ B_{k-1,0}B_{2,0}\ , \\
B_{1,3} &\to B_{1,0}B_{3,0} \ \&\ B_{1,2}B_{1,0}\ , \\
B_{2,3} &\to B_{2,0}B_{3,0} \ \&\ B_{2,1}B_{2,0}\ , \\
B_{i,j} &\to B_{1,0}B_{i,j-1} \ \&\ B_{2,0}B_{i,j-2} \quad \text{for } (i,j) \text{ not mentioned above}\ , \\
B_{i,0} &\to \{i\} \qquad\qquad\qquad\qquad\ \ \text{for } i = 1, \dots, k-1\ .
\end{aligned}$$

In the above equations $B_{i,j}$ we use cyclic notation for second lower indices, that is $B_{i,j} = B_{i,j \mod k}$. It can be shown, using methods as in Lemma 1 and Lemma 2, that the smallest solution of these equations is

$$\begin{aligned}
B_{i,j} &= \{(k \cdot i + j) \cdot k^n : n \in \mathbb{N}\}, \text{ for } j \neq 0\ , \\
B_{i,0} &= \{i \cdot k^n : n \in \mathbb{N}\}\ .
\end{aligned}$$

For $k = 2, 3$ we have to sum up some languages generated in cases of $k = 4, 9$, respectively. The case for $k = 1$ is trivial.

The productions used in this theorem could be simplified for fixed k. □

4 Regular Languages over k-ary Alphabet

Now we deal with major generalisation of the Theorem 1. We deal with languages $\{a^n : n \in L\}$, where R is some regular language (written in some k-ary system). To simplify the notation, let $\Sigma_k = \{0, \ldots, k-1\}$. From the following on we consider regular languages over Σ_k for some k that do not have words with leading 0, since this is meaningless in case of numbers.

Definition 1. *Let $w \in \Sigma_k^*$ be a word. We define its unary representation as*

$$f_k(w) = \{a^n : w \text{ read as } k\text{-ary number is } n\} .$$

When this does not lead to confusion, we also use f_L applied to languages with an obvious meaning.

The following fact shows that it is enough to consider the k parameter in f_k that are 'large enough'.

Lemma 3. *For every $k = l^n$, $n > 0$ and every unary language L language $f_k^{-1}(L)$ is regular if and only if language $f_l^{-1}(L)$ is regular.*

Proof. An automaton over alphabet Σ_k can clearly simulate reading a word written in l-ary system and *vice-versa*. □

In the following we shall use 'big enough' k, say $k \geq 100$. We claim, that for regular L language $f_k(L)$ is conjunctive.

We now define the conjunctive grammar for fixed regular language $L \subset \Sigma_k^*$ without leading 0. Let

$$M = \langle \Sigma_k, Q, \delta, Q_{fin}, q_0 \rangle$$

be the (non-deterministic) automaton that recognizes L^r.

We define conjunctive grammar $G = \langle \{a\}, N, P, S \rangle$ over unary alphabet with:

$$N = \{A_{i,j,q}, A_{i,j} : 1 \leq i < k,\ 0 \leq j < k,\ q \in Q\} \cup \{S\} .$$

The intended solution is

$$L(A_{i,j}) = \{n \ :\ f_k^{-1}(n) = ij0^k \text{ for some natural } k\} , \tag{9}$$

$$L(A_{i,j,q}) = \{n : f_k^{-1}(n) = ijw,\ \delta(q_0, w^r, q)\} , \tag{10}$$

$$L(S) = f_k(L) . \tag{11}$$

We denote sets defined in (10) by $L_{i,j,q}$.

From Theorem 3 we know, that $A_{i,j}$ can be defined by conjunctive grammars, and so we focus only on productions for $A_{i,j,q}$:

$$A_{i,j,q} \to A_{i,0}A_{j,x,q'} \ \&\ A_{i,1}A_{j-1,x,q'} \ \&\ A_{i,2}A_{j-2,x,q'} \ \&\ A_{i,3}A_{j-3,x,q'} \ , \tag{12}$$

for every $j > 3$, every i and every x, q' such that $\delta(q', x, q)$.

$$A_{i,j,q} \to A_{i-1,j+1}A_{k-1,x,q'} \;\&\; A_{i-1,j+2}A_{k-2,x,q'} \;\&\; \\ \&\; A_{i-1,j+3}A_{k-3,x,q'} \;\&\; A_{i-1,j+4}A_{k-4,x,q'} \;, \quad (13)$$

for every $j < 4$ and $i \neq 1$ and every x, q' such that $\delta(q', x, q)$.

$$A_{1,j,q} \to A_{k-1,0}A_{j+1,x,q'} \;\&\; A_{k-2,0}A_{j+2,x,q'} \;\&\; \\ \&\; A_{k-3,0}A_{j+3,x,q'} \;\&\; A_{k-4,0}A_{j+4,x,q'} \;, \quad (14)$$

for every $j < 4$, every x, q' such that $\delta(q', x, q)$.

$$A_{i,j,q_0} \to k \cdot i + j \;, \quad (15)$$

$$S \to A_{i,j,q} \text{ for } q \text{ such that } \exists q' \in Q_{fin},\ \delta(q, ji, q') \\ \text{and } i,\ j\text{—arbitrary digits} \;, \quad (16)$$

$$S \to i \;\text{ for } i \in f_k(L \cap \Sigma_k) \;. \quad (17)$$

We shall prove, that $L_{i,j,q}$ are solution of proper language equations and that they are the minimal solution (are included in every other solution). The case of $L(S)$ in (11) will then easily follow. We identify the equations with productions (12)–(14).

Lemma 4. *Every solution $(X_{i,j,q})$ of language equations defining G for nonterminals $(A_{i,j,q})$ is a superset of $(L_{i,j,q})$.*

Proof. We prove by induction that for every $n > 1$

$$L_{i,j,q} \restriction_n \subset X_{i,j,q}.$$

When $|ijw| = 2$ then this is obvious by rule (15).

Suppose we have proven the Lemma for $k < n$, we prove it for n. Choose any $ijw \in L_{i,j,q} \restriction_n$. Let $w = xw'$, suppose $j > 3$. Let p be a state such that $\delta(q_0, w'^r, p)$ and $\delta(p, x, q)$ (choose one if there are many). Consider words $jxw' \in L_{j,x,p} \restriction_{n-1}$, $(j-1)xw' \in L_{j-1,x,p} \restriction_{n-1}$, $(j-2)xw' \in L_{j-2,x,p} \restriction_{n-1}$ and $(j-3)xw' \in L_{j-3,x,p} \restriction_{n-1}$. By induction hypothesis $L_{j,x,p} \restriction_{n-1} \subset X_{j,x,p}$, $L_{j-1,x,p} \restriction_{n-1} \subset X_{j-1,x,p}$, $L_{j-2,x,p} \restriction_{n-1} \subset X_{j-2,x,p}$ and $L_{j-3,x,p} \restriction_{n-1} \subset X_{j-3,x,p}$. And so by (12) $ijw \in X_{i,j,q}$.

Other cases (which use productions (13), (14)) are proved analogously. □

Lemma 5. *languages $L_{i,j,q}$ are a solution of of* (12)–(14).

Proof. For the $\subset$ part the proof is essentially the same as in Lemma 4.

To prove the the $\supset$ part we proceed by induction on number of digits in $w \in L_{i,j,q}$.

We begin with (12). Suppose w belong to the right-hand side. We shall show that it also belongs to the left-hand side. We first proof, that it in fact has the desired two first digits. Then we shall deal with the q index.

Consider the possible positions of the two first digits of each conjunct. Notice, that if j is on the position one to the right of i, then the digits are as desired. And so we may exclude this case from our consideration. The following table summarizes the results:

	i and j are together	j is leading	i is leading
$A_{i,0}A_{j,xq'}$	$(i+j), x$	$j, x\langle+i\rangle$	$i, 0$
$A_{i,1}A_{j-1,xq'}$	$(i+j-1), (x+1)$	$(j-1), x\langle+i\rangle$	$i, 1$
$A_{i,2}A_{j-2,xq'}$	$(i+j-2), (x+2)$	$(j-2), x\langle+i\rangle$	$i, 2$
$A_{i,3}A_{j-3,xq'}$	$(i+j-3), (x+3)$	$(j-3), x\langle+i\rangle$	$i, 3$

The drawback of this table is that it does not include the possibility, that some digits sum up to k (or more) and influence another digit (by carrying 1), we handle with this manually. Also in the second column i may be or may be not on the same position as x, but we deal with those two cases together. This possibility was marked by writing $\langle+i\rangle$. Also there may be an add up to k somewhere to the right, and hence we can add 1 to x.

If we want the intersection to be non-empty we have to choose four items, no two of them in the same row. We show, that this is not possible. We say that some choices fit, if the digits included in the table are the same in those choices.

First of all, no two elements in the third column fit. They have fixed digits and they clearly are different.

Suppose now that we choose two elements from thefirst column. We show that if in one of them $(i+j-z)$ sums up to k (or more) then the same thing happens in the second choice. If $i+j-z \geq k$ (perhaps by additional 1 carried from the previous position) then the first digit is 1. In the second element the first digit can be 1 (if there is a carrying of 1) or at least $i+j-z'$, but the latter is not possible, since $i+j-z' > 1$. Whatever happens, it is not possible to fit the digits on the column with x.

It is not possible to choose three elements from the second column. Suppose that we have three fitting choices in this column. As before we may argue, that either all of them have $j(-z)$ as the first digit or the digits add up to more than $k-1$ and the first digit is 1. Suppose they add up. Since $j < k$ then this is possible only for the first row. Suppose they do not add up. Then if there are three fitting choices, then one of them must be increased by at least 2. But the maximal value carried from the previous position is 1. Contradiction. Note, that by the same reasoning we may prove, that if there are two fitting choices then the first digit is between $j-3$ and j.

And so we know, that if there are four fitting choices, then exactly one of them is in the first column, one in the third column and two in the middle column. The third column always begins with i. In the first column the leading digit is at least $i+j-3 > i$ or it is 1. Hence $i = 1$. And so the choices in the second column begin with 1 as well. Hence $j < 4$, which is a contradiction.

We now move to (13). The following table summarizes the possible first two digits:

	i and $k-z$ are together	$k-z$ is leading	$i-1$ is leading
$A_{i-1,j+1}A_{k-1,x,q'}$	$(k+i-2),(1+j+x)$	$(k-1),x(+i)$	$i-1,j+1$
$A_{i-1,j+2}A_{k-2,x,q'}$	$(k+i-3),(2+j+x)$	$(k-2),x(+i)$	$i-1,j+2$
$A_{i-1,j+3}A_{k-3,x,q'}$	$(k+i-4),(3+j+x)$	$(k-3),x(+i)$	$i-1,j+3$
$A_{i-1,j+4}A_{k-4,x,q'}$	$(k+i-5),(4+j+x)$	$(k-4),x(+i)$	$i-1,j+4$

As before we may argue, that if there are some fitting entries in some column then on their leading position digits sum up to k in all choices or in all choices do not sum up to k.

We cannot have two choices from the third column (the second digits do not match). We can have at the most two from the second column (to obtain three we would have to ad carry at least 2 to one of them and this is not possible). Fr the same reason there can be at the most two choices from the first column, but in such a case we cannot match the positions with x. Hence there is at the most one choice from the first column. And so we have one choice from the first column, one from the third and two from the second. Since the third and the second column match, then i is a big digit, at least $k-3$. But in such a case in the first column we have at least $k+k-3-5>k$ and so the leading digit is 1. Contradiction.

Consider the last possibility, the (14).The following table summarizes the possible first two digits:

	$k-z$ is leading	$j+z$ is leading
$A_{k-1,0}A_{j+1,x,q'}$	$(k-1),0(+j+1)$	$(j+1),x(+k-1)$
$A_{k-2,0}A_{j+2,x,q'}$	$(k-2),0(+j+2)$	$(j+2),x(+k-2)$
$A_{k-3,0}A_{j+3,x,q'}$	$(k-3),0(+j+3)$	$(j+3),x(+k-3)$
$A_{k-4,0}A_{j+4,x,q'}$	$(k-4),0(+j+4)$	$(j+4),x(+k-4)$

In the first column there are no two fitting choices. So we would have to take at least three from the second one. Since j is small, it cannot add up to more than k. So we cannot choose three different choices there—it would force a carrying of at least 2 from the previous position. Contradiction.

We should also take the indices denoting states of the automaton into our consideration. Consider one production:

$$A_{i,j,q} \to A_{i,0}A_{j,x,q'} \;\&\; A_{i,1}A_{j-1,x,q'} \;\&\; A_{i,2}A_{j-2,x,q'} \;\&\; A_{i,3}A_{j-3,x,q'}$$

and some w belonging to the right-hand side. We will explain the case of conjunct $A_{i,0}A_{j,x,q'}$. Consider $jxw' \in L_{j,x,q'}$ that was used in derivation of w. By definition of $L_{j,x,q'}$ we obtain $\delta(q_0, w'^r, q')$ and by definition of the Production (12) $\delta(q', x, q)$, and so $\delta(q_0, w'^r x, q)$. By previous discussion w begins with digits i, j, then it inherits its digits from jxw' and hence has digit x and then word w'. And so $w = ijxw'$ and $\delta(q_0, w'^r x, q)$, therefore it belongs to the left-hand side. The case with other productions and conjuncts is proved in the same way. □

Theorem 4. *For every natural $k > 1$ and every regular $L \subset \Sigma_k^*$ without words with leading* 0 *language $f_k(L)$ is a conjunctive unary language.*

Proof. By Lemma 3, Lemma 4 and Lemma 5. □

5 Conclusions and Open Problems

The main result of this paper is an example of a conjunctive grammar over unary alphabet generating non-regular language. This grammar has six nonterminal symbols in Chomsky normal form. Number of nonterminals could be reduced to three if we consider a grammar that is not in a Chomsky normal form. It remains an open question, how many nonterminals, intersection *etc.* are required to generate a non-regular language. In particular, can we give natural sufficient conditions for a conjunctive grammar to generate a regular language? Also, no non-trivial algorithm for recognizing conjunctive languages over unary alphabet is known. An obvious modification of the CYK algorithm requires quadratic time and linear space. Can those bounds be lowered? Closure under complementation of conjunctive languages (both in general and and in case of unary alphabet) remains unknown, with conjectured negative answer. Perhaps dense languages, like

$$\mathbb{N} \setminus \{2^n : n \in \mathbb{N}\}$$

are a good starting point in search for an example.

The second important result is a generalisation of the previous one: for every regular language $R \subset \{0, ..., k-1\}^*$ treated as set of k-ary numbers language $\{a^n : \exists_w \in R\ w$ read as a number is $n\}$ is a conjunctive unary language.

Acknowledgments. The author would like to thank Tomasz Jurdziński and Krzysztof Loryś for introducing to the subject and helpful discussion, Sebastian Bala and Marcin Bieńkowski for helpful discussion. The anonymous referees, who helped in improving the presentation and Alexander Okhotin, who suggested the study of generalisation to unary representations of all regular languages (Theorem 4).

References

[1] Okhotin, A.: Conjunctive grammars. Journal of Automata, Languages and Combinatorics 6(4), 519–535 (2001)
[2] Okhotin, A.: An overview of conjunctive grammars. Formal Language Theory Column. Bulletin of the EATCS 79, 145–163 (2003)
[3] Okhotin, A.: Boolean grammars. Information and Computation 194(1), 19–48 (2004)
[4] Okhotin, A.: Nine open problems on conjunctive and boolean grammars. TUCS Technical Report, p. 794 (2006)