Nine open problems on conjunctive and Boolean grammars: an update*

Alexander Okhotin†

March 30, 2010

Abstract

Conjunctive grammars are context-free grammars with an explicit conjunction operation in the formalism of rules; Boolean grammars are further equipped with an explicit negation. Three years have passed since the publication of the last survey of these grammars and their open research questions (A. Okhotin, "Nine open problems on conjunctive and Boolean grammars", *Bulletin of the EATCS*, 91 (2007), 96–119). While an updated survey is under preparation, this note is aimed to report on the two solved problems, as well as to correct a couple of small errors. In addition, the award for solving each of the problems is raised to $360 Canadian.

Contents

*Research supported by the Academy of Finland under grant 134860.

†Academy of Finland, *and* Department of Mathematics, University of Turku, *and* Turku Centre for Computer Science Turku FIN–20014, Finland. E-mail: `alexander.okhotin@utu.fi`

1 Updates to the nine problems

1.1 Limitations of Boolean grammars

Problem 1. *Are there any languages recognized by deterministic linear bounded automata working in time $O(n^2)$ that cannot be specified by Boolean grammars?*

There is one misleading comment to be corrected. It was suggested that

> *Other languages possibly not representable by Boolean grammars can be sought for in the domain of unary languages: consider* $\{a^{2^{2^n}} \mid n \geqslant 0\}$ *or* $\{a^{n^2} \mid n \geqslant 0\}$.

The knowledge on conjunctive grammars over a one-letter alphabet has much advanced since then, and these two examples are no longer suggested. Though it remains unknown whether the exact language $\{a^{n^2} \mid n \geqslant 0\}$ can be represented, a grammar for a similar language has been found:

Proposition 1 (Jeż, Okhotin [2])**.** *There exists a conjunctive grammar G over an alphabet $\{a\}$ generating a language $\{a^{f(n)} \mid n \geqslant 0\}$ for an integer function $f(n) = \Theta(n^2)$.*

The language $\{a^{2^{2^n}} \mid n \geqslant 0\}$ can most probably be represented as it is. Though no one has taken the time to construct a grammar for this particular language, there is now a general theorem on the existence of conjunctive grammars for unary languages of an arbitrarily fast growth rate.

Theorem 1 (Jeż, Okhotin [2])**.** *For every infinite recursively enumerable set of natural numbers S there exists a conjunctive grammar G over an alphabet $\{a\}$, such that the growth function of $L(G)$ is greater than that of S at any point. Given a Turing machine recognizing S, the grammar G can be effectively constructed.*

In light of the further results on unary conjunctive languages presented in the next section, finding a unary language generated by no Boolean grammar is a much more difficult problem than anticipated, and of an interest on its own. But Problem 1 would perhaps better be approached using languages over multiple-letter alphabets.

1.2 Conjunctive grammars over a one-letter alphabet (SOLVED)

Problem 2 (Solved negatively by Artur Jeż [1] in 2007)**.** *Do conjunctive grammars over a one-letter alphabet generate only regular languages?*

It was suggested that

> *If they can generate any nonregular language, this would be a surprise.*

The surprise took form of the following conjunctive grammar with four nonterminal symbols:

Example 1 (Jeż [1]). *The conjunctive grammar*

$$\begin{array}{lll} A_1 & \to & A_1A_3 \& A_2A_2 \mid a \\ A_2 & \to & A_1A_1 \& A_2A_6 \mid aa \\ A_3 & \to & A_1A_2 \& A_6A_6 \mid aaa \\ A_6 & \to & A_1A_2 \& A_3A_3 \end{array}$$

with the start symbol A_1 generates the language $L(G) = \{a^{4^n} \mid n \geqslant 0\}$. In particular, $L_G(A_i) = \{a^{i \cdot 4^n} \mid n \geqslant 0\}$ for $i = 1, 2, 3, 6$.

Substituting these four languages into the first equation, one obtains

$$\begin{aligned} &\{a^{4^i} \mid i \geqslant 0\}\{a^{3 \cdot 4^j} \mid j \geqslant 0\} \cap \{a^{2 \cdot 4^k} \mid k \geqslant 0\}\{a^{2 \cdot 4^\ell} \mid \ell \geqslant 0\} = \\ &= \big(\{a^{4^n} \mid n \geqslant 1\} \cup \{a^{4^i + 3 \cdot 4^j} \mid i \neq j\}\big) \cap \big(\{a^{4^n} \mid n \geqslant 1\} \cup \{a^{2 \cdot 4^k + 2 \cdot 4^\ell} \mid k \neq \ell\}\big) = \{a^{4^n} \mid n \geqslant 1\}, \end{aligned}$$

that is, both concatenations contain some garbage, yet the garbage in the concatenations is disjoint, and is accordingly filtered out by the intersection. Finally, the union with $\{a\}$ yields the language $\{a^{4^n} \mid n \geqslant 0\}$, and thus the first equation turns into an equality. The rest of the equations are verified similarly. Proving that this solution is the least one is only a matter of technique [1].

The idea of manipulating positional notation of numbers was extended to the following general result:

Theorem 2 (Jeż, Okhotin [2]). *Let $A_k = \{\mathtt{0}, \mathtt{1}, \ldots, k-1\}$ with $k \geqslant 2$ be an alphabet of k-ary digits, and let $L \subseteq A_k^*$ be a language generated by a linear conjunctive grammar, which contains no strings starting from $\mathtt{0}$. Then there exists a conjunctive grammar generating the language $\{a^n \mid \text{the } k\text{-ary notation of } n \text{ is in } L\}$.*

A similar technique was used to construct an EXPTIME-complete set of numbers with its unary representation generated by a conjunctive grammar [3].

Based upon Theorem 2, several undecidability results for unary conjunctive languages were found, such as the following one:

Theorem 3 (Jeż, Okhotin [2]). *For every fixed unary conjunctive language $L_0 \subseteq a^*$, the problem of whether a given conjunctive grammar over $\{a\}$ generates the language L_0 is Π^0_1-complete.*

Unary conjunctive grammars with a unique nonterminal symbol are already nontrivial. The first example of their nontriviality was actually an encoding of Example 1:

Example 2 (Okhotin, Rondogiannis [8]). *The following one-nonterminal conjunctive grammar*

$$S \to a^{11}SS \& a^{22}SS \mid aSS \& a^9SS \mid a^7SS \& a^{12}SS \mid a^{13}SS \& a^{14}SS \mid a^{56} \mid a^{113} \mid a^{181}$$

generates the language $\{a^{4^n-8} \mid n \geqslant 3\} \cup \{a^{2 \cdot 4^n - 15} \mid n \geqslant 3\} \cup \{a^{3 \cdot 4^n - 11} \mid n \geqslant 3\} \cup \{a^{6 \cdot 4^n - 9} \mid n \geqslant 3\}$.

A general method of encoding a given conjunctive grammar over a unary alphabet in a one-nonterminal conjunctive grammar was subsequently given by Jeż and Okhotin [4].

1.3 Time complexity (SOLVED)

Problem 3 (Solved positively by Alexander Okhotin [6] in 2009)**.** *Are the languages generated by Boolean grammars contained in* $DTIME(n^{3-\varepsilon})$ *for any* $\varepsilon > 0$*?*

The original survey mentioned fast parsing for context-free grammars:

> *Valiant (1975) reduced context-free membership problem to matrix multiplication, which allowed him to apply Strassen's (1969) fast matrix multiplication algorithm to obtain a context-free recognizer working in time* $O(n^{2.807})$*. Using an asymptotically better matrix multiplication method due to Coppersmith and Winograd (1990), the complexity of Valiant's recognizer can be improved to* $O(n^{2.376})$*.*

But then it was stated that

> *However, already for conjunctive grammars there seems to be no way to reduce the membership problem to matrix multiplication.*

Indeed, Valiant's algorithm was presented in a way that it essentially relies on having two operations in a grammar, concatenation and union. These operations give rise to the product and the sum in a certain semiring, with the rest of the algorithm operating in terms of this semiring. However, it turned out that using matrices over a semiring as an intermediate abstraction is in fact unnecessary, and it is sufficient to employ Boolean matrix multiplication to compute the concatenations only, with the Boolean operations in the grammar evaluated separately. This led to a simpler variant of Valiant's algorithm, which is naturally applicable to Boolean and context-free grammars alike.

Theorem 4. *For every Boolean grammar* $G = (\Sigma, N, P, S)$ *there is an algorithm, which, for a given string of length* n*, constructs the parsing table* $T_{i,j} = \{A \in N \mid a_{i+1} \dots a_j \in L_G(A)\}$ *in time* $\Theta(BM(n))$*, where* $BM(n)$ *is the time needed to multiply two* $n \times n$ *Boolean matrices.*

Accordingly, the family of languages generated by Boolean grammars is contained in DTIME($n^{2.376}$).

1.4 Space complexity

Problem 4. *Are the languages generated by Boolean grammars contained in* $DSPACE(n^{1-\varepsilon})$ *for any* $\varepsilon > 0$*?*

No progress made.

1.5 Greibach normal form

Problem 5. *Is it true that for every Boolean grammar there exists a Boolean grammar in Greibach normal form that generates the same language?*

For a small special case, the question has been answered affirmatively:

Theorem 5 (Okhotin, Reitwießner [7]). *For every conjunctive grammar over a unary alphabet there exists and can be effectively constructed a conjunctive grammar in Greibach normal form generating the same language.*

One previously given advice must be revoked. It was suggested to try the following language:

> *The question whether the language $\{a^n b^{2^n} \mid n \geqslant 1\}$ can be represented by a Boolean grammar in Greibach normal form might be a good starting point in approaching Problem 5. The answer to this question is likely negative, and a negative solution to the problem can be thus obtained.*

Surprisingly, there exists a conjunctive grammar for this language:

Example 3. *The following conjunctive grammar generates the language $\{a^n b^{2^n} \mid n \geqslant 0\}$:*

$$\begin{array}{rcl}
S & \to & aSB \& AB \mid b \\
A & \to & aA \mid \varepsilon \\
B & \to & B_1 \mid B_2 \\
B_1 & \to & B_1 B_3 \& B_2 B_2 \mid b \\
B_2 & \to & B_1 B_1 \& B_2 B_6 \mid bb \\
B_3 & \to & B_1 B_2 \& B_6 B_6 \mid bbb \\
B_6 & \to & B_1 B_2 \& B_3 B_3
\end{array}$$

Each nonterminal B_i generates $\{b^{i \cdot 4^n} \mid n \geqslant 0\}$ as in Example 1, and in particular B generates $\{b^{2^n} \mid n \geqslant 0\}$. All strings of the form $a^n b^{2^n}$ are generated inductively by S. The basis is the string $b = a^0 b^{2^0}$ generated by the rule $S \to b$. The rule $S \to aSB\&AB$ generates all strings of the form $a^{n+1} b^{2^n + i}$, in which both $2^n + i$ and i are powers of two. These conditions are satisfied only if $i = 2^n$, and thus the generated strings must be of the form $a^{n+1} b^{2^{n+1}}$.

1.6 Complementation of conjunctive grammars

Problem 6. *Is the family of conjunctive languages closed under complementation?*

It is worth being added that the same problem could be separately considered for conjunctive languages over a unary alphabet. In fact, for every unary language found to be conjunctive in the literature, its complement could be proved conjunctive by the same methods [2, 3], yet there are no methods of changing a given conjunctive grammar to a grammar for the complement of the generated language.

1.7 Inherent ambiguity

Problem 7. *Do there exist any inherently ambiguous languages with respect to Boolean grammars?*

It is worth being mentioned that while it was already known that the ambiguity in a choice of a rule can be eliminated in any given Boolean grammar, the same result has been established for conjunctive grammars as well. This property follows from a new normal form theorem:

Theorem 6 (Okhotin, Reitwießner [7]). *For every conjunctive grammar there exists and can be effectively constructed a conjunctive grammar generating the same language, in which the set of rules for every nonterminal A is of the form:*

$$A \to \alpha_1 \& \ldots \& \alpha_n \mid w_1 \mid \ldots \mid w_m \quad (n \geqslant 1,\ m \geqslant 0,\ \alpha_i \in (\Sigma \cup N)^*,\ w_j \in \Sigma^*)$$

1.8 Hierarchy of Boolean LL(k) languages

Problem 8. *Does there exist a number $k_0 > 0$, such that, for all $k \geqslant k_0$, Boolean LL(k) grammars generate the same family of languages as Boolean LL(k_0) grammars?*

Some weaker results on the expressive power of LL(k) Boolean grammar have been established:

Theorem 7 (Okhotin [5]). *Every Boolean LL(k) language over a unary alphabet is regular.*

Theorem 8 (Okhotin [5]). *For every Boolean LL(k) language $L \subseteq \Sigma^*$ there exist constants $d, d' \geqslant 0$ and $p \geqslant 1$, such that for all $w \in \Sigma^*$, $a \in \Sigma$, $n \geqslant d \cdot |w| + d'$ and $i \geqslant 0$,*

$$wa^n \in L \quad \textit{if and only if} \quad wa^{n+ip} \in L$$

These theorems, in particular, imply that there is no Boolean LL(k) grammar for the linear conjunctive language $\{a^n b^{2^n} \mid n \geqslant 0\}$ and for the conjunctive language $\{a^{4^n} \mid n \geqslant 0\}$.

1.9 Nonterminal complexity of Boolean grammars

Problem 9. *Does there exist a number $k > 0$, such that every language generated by any Boolean grammar can be generated by a k-nonterminal Boolean grammar?*

Some limitations of one-nonterminal conjunctive grammars have been established by Okhotin and Rondogiannis [8].

Theorem 9. *Let $L = \{a^{n_1}, a^{n_2}, \ldots, a^{n_i}, \ldots\}$ with $0 \leqslant n_1 < n_2 < \cdots < n_i < \cdots$ be an infinite unary language, for which $\liminf_{i \to \infty} \frac{n_i}{n_{i+1}} = 0$. Then L is not generated by any one-nonterminal conjunctive grammar.*

In particular, no unary language with a super-exponential growth rate such as $\{a^{2^{2^n}} \mid n \geqslant 0\}$ and $\{a^{n!} \mid n \geqslant 1\}$, can be represented by such grammars. The above theorem also applies to sets like $\{a^{n!+i} \mid n \geqslant 1,\ i \in \{0,1\}\}$.

The next theorem applies to such languages as $a^* \setminus \{a^{n^2} \mid n \geqslant 0\}$, $a^* \setminus \{a^{2^n} \mid n \geqslant 0\}$ and or $\{a^n \mid n$ **is composite**$\}$:

Theorem 10. *Let $L \subseteq a^*$ be a nonregular language that is* ***dense*** *in the sense that if $\lim_{n\to\infty} \frac{|L \cap \{\varepsilon, a, \ldots, a^{n-1}\}|}{n} = 1$. Then there is no one-nonterminal conjunctive grammar generating L.*

At the same time, as shown by Jeż and Okhotin [4], every conjunctive grammar G with nonterminals $\{A_1, \ldots, A_m\}$ can be encoded in a one-nonterminal grammar G_1 generating a language with $a^{np+d_i} \in L(G_1)$ if and only if with $a^n \in L_G(A_i)$, for some numbers $p, d_1, \ldots, d_m$ depending on m. However, there seems to be no way (at least, no apparent way) to apply this construction to representing exactly the language $L(G)$ using a bounded number of nonterminals.

In light of these results, one can consider the same problem for conjunctive grammars over a unary alphabet.

2 Increase of the award

The award for solving each of the remaining seven problems is increased to $360 Canadian. The terms remain as previously announced.

References

[1] A. Jeż, "Conjunctive grammars can generate non-regular unary languages", *International Journal of Foundations of Computer Science*, 19:3 (2008), 597–615.

[2] A. Jeż, A. Okhotin, "Conjunctive grammars over a unary alphabet: undecidability and unbounded growth", *Theory of Computing Systems*, 46:1 (2010), 27–58.

[3] A. Jeż, A. Okhotin, "Complexity of solutions of equations over sets of natural numbers", *25th Annual Symposium on Theoretical Aspects of Computer Science* (STACS 2008, Bordeaux, France, 21–23 February, 2008), 373–383.

[4] A. Jeż, A. Okhotin, "One-nonterminal conjunctive grammars over a unary alphabet", *Computer Science in Russia* (CSR 2009, Novosibirsk, Russia, 18–23 August, 2009), LNCS 5675, 191–202.

[5] A. Okhotin, "Expressive power of LL(k) Boolean grammars", *Fundamentals of Computation Theory* (FCT 2007, Budapest, Hungary, August 27–30, 2007), LNCS 4639, 446–457.

[6] A. Okhotin, "Fast parsing for Boolean grammars: a generalization of Valiant's algorithm", TUCS TR 953, 2010.

[7] A. Okhotin, C. Reitwießner, "Conjunctive grammars with restricted disjunction", *Theoretical Computer Science*, to appear.

[8] A. Okhotin, P. Rondogiannis, “On the expressive power of univariate equations over sets of natural numbers”, *IFIP Intl. Conf. on Theoretical Computer Science* (TCS 2008, Milan, Italy, 8–10 September, 2008), IFIP vol. 273, 215–227.