ON THE SUCCINCTNESS PROPERTIES OF UNORDERED CONTEXT-FREE GRAMMARS

M. Drew Moshier and William C. Rounds
Electrical Engineering and Computer Science Department
University of Michigan
Ann Arbor, Michigan 48109

1 Abstract

We prove in this paper that unordered, or ID/LP grammars, are exponentially more succinct than context-free grammars, by exhibiting a sequence (L_n) of finite languages such that the size of *any* CFG for L_n must grow exponentially in n, but which can be described by polynomial-size ID/LP grammars. The results have implications for the description of free word order languages.

2 Introduction

Context free grammars in immediate dominance and linear precedence format were used in GPSG [3] as a skeleton for metarule generation and feature checking. It is intuitively obvious that grammars in this form can describe languages which are closed under the operation of taking arbitrary permutations of strings in the language. (Such languages will be called *symmetric*.) Ordinary context-free grammars, on the other hand, seem to require that all permutations of right-hand sides of productions be explicitly listed, in order to describe certain symmetric languages. For an explicit example, consider the n-letter alphabet $\Sigma_n = \{a_1, \ldots, a_n\}$. Let P_n be the set of all strings which are permutations of exactly these letters. It seems obvious that no context-free grammar could generate this language without explicitly listing it. Now try to *prove* that this is the case. This is in essence what we do in this paper. We also hope to get the audience for the paper interested in why the proof works!

To give some idea of the difficulty of our problem, we begin by recounting Barton's results [1] in this conference in 1985. (There is a general discussion in [2].) He showed that the *universal recognition problem* (URP) for ID/LP grammars is NP-complete. [1] This means that if $P \neq NP$, then no polynomial algorithm can solve this problem. The difficulty of the problem seems to arise from the fact that the translation from an ID/LP grammar to a weakly equivalent CFG blows up exponentially. It is easy to show, assuming $P \neq NP$, that any reasonable transformation from ID/LP grammars to equivalent CFGs cannot be done in polynomial time; Rounds has done this as a remark in [8]. In this paper, we remove the hypothesis $P \neq NP$. That is, we can show that no algorithm whatever can effect the translation polynomially in all cases. (Unfortunately, this does not solve the $P = NP$ question!)

Barton's reduction took a known NP-complete problem, the vertex-cover problem, and reduced it to the URP for ID/LP. The reduction makes crucial use of grammars whose production size can be arbitrarily large. Define the *fan-out* of a grammar to be the largest total number of symbol occurrences on the right hand side of any production. For a CFG, this would be the maximum length of any RHS; for an ID/LP grammar, we would count symbols and their multiplicities. Barton's reduction does the following. For each instance of the vertex cover problem, of size n, he constructs a string w and an ID/LP grammar of fanout proportional to n such that the instance has a vertex cover if and only if the string is generated by the grammar. He also notes that if all ID/LP grammars have fanout bounded by a fixed constant, then the URP can be solved in polynomial time.

This brings us to the statement of our results. Let P_n be the language described above. Clearly this language can be generated by the ID/LP grammar

$$S \rightarrow a_1, \ldots, a_n$$

whose size in bits is $O(n \log n)$.

Theorem 1 *There is a constant $c > 1$ such that any context-free grammar G_n generating P_n must have size $\Omega(c^n)$.*[2] *Moreover, every ID/LP grammar generating P_n, whose fanout is bounded by a fixed constant, must likewise have exponential size.*

The theorem does not actually depend on having a vocabulary which grows with n. It is possible to code everything homomorphically into a two-letter alphabet. However, we think that the result shows that ordinary CFGs, and bounded-fanout ID/LP grammars, are inadequate for giving succinct descriptions of languages whose vocabulary is open, and whose word order can be very free. Thus, we prefer the statement of the result as it is.

We start the paper with the technical results, in Section 3, and continue with a discussion of the implications for linguistics in Section 4. The final section contains a proof of the Interchange Lemma of Ogden, Ross, and Winklmann [7], which is the main tool used for our results. This proof is included, not because it is new, but because we want to show a beautiful example of the use of

[1] The universal recognition problem is to tell for an ID/LP grammar G and a string w, whether or not $w \in L(G)$.

[2] This notation means that for infinitely many n, the size of G_n must be bigger than c^n.

combinatorial principles in formal linguistics, and because we think the proof may be generalized to other classes of grammars.

3 Technical Results

As we have said, our basic tool is the Interchange Lemma, which was first used to show that the "embedded reduplication" language $\{ wxxy \mid w, x, \text{and } y \in \{a,b,c\}^*\}$ is not context-free. It was also used in Kac, Manaster-Ramer, and Rounds [6] to show that English is not CF, and by Rounds, Manaster-Ramer, and Friedman to show that reduplication even over length n strings requires context-free grammar size exponential in n. The current application uses the last-mentioned technique, but the argument is more complicated.

We will discuss the Interchange Lemma informally, then state it formally. We will then show how to apply it in our case.

The IL relies on the following basic observation. Suppose we have a context-free language, and two strings in that language, each of which has a substring which is the yield of a subtree labeled by the same nonterminal symbol at the respective roots of the subtrees. Then these substrings can be interchanged, and the resulting strings will still be in the language. This is what distinguishes the IL from the Pumping Lemma, which finds repeated nonterminals in the derivation tree of just one string.

The next observation about the IL is that it attempts to find these interchangeable strings among the length n strings of the given language. Moreover, we want to find a whole *set* of such strings, such that in the set, the interchanged substrings all have the same length, and all start at the same position in the host string. The lemma lets us select a number m less than n, and tells us that the length k of the interchangeable substrings is between m/r and m, where r is the fanout of the grammar. Finally, the lemma gives us an estimate of the size of the interchangeable subset. We may choose an arbitrary subset $Q(n)$ of $L(n)$, where $L(n)$ is the set of length n strings in the language L. If we also choose an integer $m \leq n$, then the IL tells us that there is an interchangeable set $A \subseteq Q(n)$ such that $|A| \geq |Q(n)|/(|N| \cdot n^2)$, where the vertical bars denote cardinality, and N is the set of nonterminals of the given grammar. (The interchanged strings do not stay in $Q(n)$, but they do stay in $L(n)$.) Notice that if $Q(n)$ is exponential in size, then A will be also. Thus, if a language has exponentially many strings of length n then it will have an interchangeable subset of roughly the same exponential size, provided the set of nonterminals of the grammar is small. Our proof turns this idea around. We show that any CF description of the permutation language $L(n)$ must have an exponentially large set of nonterminals, because an interchangeable subset of this language cannot be of the same exponential order as $n!$, which is the size of $L(n)$.

Now we can give a more formal statement of the lemma.

Definition. Suppose that A is a subset $\{z_1, \ldots, z_p\}$ of $L(n)$. A has the k-interchangeability property iff there are substrings $x_1, \ldots, x_p$ of $z_1, \ldots, z_p$ respectively, such that each x_i has length k, each x_i occurs in the same relative position in each z_i, and such that if $z_i = w_i x_i y_i$ and $z_j = w_j x_j y_j$ for any i and j, then $w_i x_j y_i$ is an element of $L(n)$.

Interchange Lemma. Let G be a CFG or ID/LP grammar with fanout r, and with nonterminal alphabet N. Let m and n be any positive natural numbers with $r < m \leq n$. Let $L(n)$ be the set of length n strings in $L(G)$, and $Q(n)$ be a subset of $L(n)$. Then we can find a k-interchangeable subset A of $Q(n)$, such that $m/r \leq k \leq m$, and such that

$$|A| \geq |Q(n)|/(|N| \cdot n^2).$$

Now we can prove our main theorem. First we show that no CFG of fanout 2 can generate $L(n)$ without an exponential number of nonterminals. The theorem for any CFG then follows, because any CFG can be transformed into a CFG with fanout 2 by a process essentially like that of transforming into Chomsky normal form, but without having to eliminate ϵ-productions or unit productions. This process at most cubes the grammar size, and the result follows because the cube root of an exponential is still an exponential. The proof for bounded-fanout ID/LP is a direct adaptation of the proof for fanout 2, which we now give.

Let P_n be the permutation language above, and let G be a fanout 2 grammar for this language. Apply the Interchange Lemma to G, choosing $Q(n) = P_n$, $r = 2$, and $m = n/2$. (n will be chosen as a multiple of 4.) Observe that $|Q(n)| = |L(n)| = n!$. From the IL, we get a k-interchangeable subset A of $L(n)$, such that $n/4 \leq k \leq n/2$, and such that

$$|A| \geq \frac{n!}{|N| \cdot n^2}.$$

Next we use the fact that A is k-interchangeable to get an upper bound on its cardinality. Let $w_1 x_1 y_1$ and $w_2 x_2 y_2$ be members of A, and let $\Sigma(x)$ be the *set* of alphabet characters appearing in x. We claim that $\Sigma(x_1) = \Sigma(x_2)$. For if, say x_1 has a character not occurring in x_2, then the interchanged string $w_1 x_2 y_1$ will have two occurrences of that character, and thus not be in $L(n)$, as required by the IL. Without loss of generality, $\Sigma(x) = \{a_1, \ldots, a_k\}$. The number of strings in A is thus less than or equal to the number of ways of selecting the x string - that is, $k!$, times the number of ways of choosing the characters in the rest of the string - that is, $(n-k)!$. In other words,

$$|A| \leq k!\,(n-k)!.$$

Putting the two inequalities together and solving for $|N|$, we get

$$|N| \geq \frac{n!}{k!(n-k)!} \cdot \frac{1}{n^2} = \frac{1}{n^2} \cdot \binom{n}{k}.$$

From Pascal's triangle in high school mathematics, $\binom{n}{k}$ increases with k until $k = n/2$. Thus since $n/4 \leq k \leq n/2$, we have $\binom{n}{k} \geq \binom{n}{n/4}$, which by using Stirling's approximation

$$m! \sim m^m e^{-m}\sqrt{2\pi m}$$

to estimate the various factorials, grows exponentially with n. Therefore, so does $|N|$, and our theorem is proved.

To obtain the result for a two-letter alphabet, consider the homomorphism sending the letter a_j into 0^j1. Let K_n be the image of P_n under this mapping. Then, because the mapping is one-to-one, P_n is the inverse homomorphic image of K_n. If for every $c > 1$ there is a sequence of CFGs G_n generating K_n such that the size of G_n is *not* $\Omega(c^n)$, then the same is true for the language P_n, contradicting Theorem 1. The reason is that the size of a grammar for the inverse homomorphic image of a language need only be polynomially bigger than the size of a grammar for the language itself. The proof of this claim rests on inspection of one of the standard proofs, say Hopcroft and Ullman [5]. The result is proved using pushdown automata, but all conversions from pdas to grammars require only polynomial increase in size.

Our final technical result concerns an n-symbol analogue of the so-called MIX language, which has been conjectured by Marsh not to be an indexed language (see [4] for discussion.) We define the language M_n to be the set of all strings over Σ_n which have identical numbers of occurrences of each character a_i in Σ_n. Observe that M_n is infinite for each n. However, there is a sequence of finite sublanguages of the various M_n, such that this sequence requires exponentially increasing context-free descriptions. We have the following theorem.

Theorem 2 *Consider the set $M_n(n^2)$ of all length n^2 strings of M_n. Then there is a constant $c > 1$ such that any context-free grammar G_n generating $M_n(n^2)$ must have size $\Omega(c^n)$.*

Proof. This proof is really just a generalization of the proof of Theorem 1. It uses, however, the Q subsets in a way that the proof of Theorem 1 does not.

First, we drop the n subscript in $M_n(n^2)$. Observe next that in every string in $M(n^2)$, each character in Σ_n occurs exactly n times. Let $Q(n^2) = \{u^n : |u| = n\}$ be the subset of $M(n^2)$ where, as indicated, each string is composed of n identical substrings concatenated in order. Then each u substring must be a permutation of Σ_n, i.e., a member of P_n. Let G_n be a fanout 2 grammar generating $M(n^2)$. As in the proof of Theorem 1, apply the Interchange Lemma to G_n, choosing $Q(n^2)$ as above, $r = 2$, and $m = n/2$. Observe that we still have $|Q(n^2)| = n!$. From the IL, we get a k-interchangeable subset A of $Q(n^2)$, such that $n/4 \leq k \leq n/2$, and such that

$$|A| \geq \frac{n!}{|N| \cdot n^4}.$$

Once again we use the fact that A is k-interchangeable to get an upper bound on its cardinality. Let $w_1x_1y_1$ and $w_2x_2y_2$ be members of A, and let $\Sigma(x)$ be the *set* of alphabet characters appearing in x. We claim once again that $\Sigma(x_1) = \Sigma(x_2)$. To see this, notice that the x portions of the strings in A can overlap at most one of the boundaries between the successive u strings, because $|u| = n$ and $|x| \leq n/2$. If it does not overlap a boundary, then the reasoning is as before. If it does overlap a boundary, then we claim that the characters in x occurring to the right of the boundary must all be different from the characters in x to the left. This is because of the "wraparound phenomenon": the u strings are identical, so the x characters to the right of the boundary are the same characters which occur to the right of the *previous* u-boundary. Since each u is a permutation of Σ_n, the claim holds. The same reasoning now applies to show that $\Sigma(x_1) = \Sigma(x_2)$. For if, say, x_1 has a character not occurring in x_2, then one of the u-portions of the interchanged string $w_1x_2y_1$ will have two occurrences of that character, and thus not be in $M(n^2)$, as required by the IL. Without loss of generality, $\Sigma(x) = \{a_1, \ldots, a_k\}$. The number of strings in A is less than or equal to the number of ways of selecting one of the u strings. Consider the u string to the left of the boundary which x overlaps. Because of wraparound, this u string is still determined by selecting k positions in the x, and then choosing the characters in the remaining $n - k$ positions. Thus we still have

$$|A| \leq k!(n-k)!$$

and we finish the proof as above.

4 Discussion

What do Theorems 1 and 2 literally mean as far as linguistic descriptions are concerned? First, we notice that the permutation language P_n really has a counting property: there is exactly one occurrence of each symbol in any string. The same is true if we consider, for fixed m, the strings of length mn in M_n, as n varies. Here there must be exactly m occurrences of each symbol in Σ_n, in every string. It seems unreasonable to require this counting property as a property of the sublanguage generated by any construction of ordinary language. For example, a list of modifiers, say adjectives, could allow arbitrary repetitions of any of its basic elements, and not insist that there be at most one occurrence of each modifier. So these examples do not have any direct, naturally occurring, linguistic analogues. It is only if we wish to describe permutation-like behavior where the number of occurrences of each symbol is bounded, but with an unbounded number of symbols, that we encounter difficulties.

The same observation, however, applies to Barton's NP-completeness result. Exactly the same counting property is required to make the universal recognition problem intractable. If we do not insist on an n-character alphabet, of course, then the universal recognition problem is only polynomial for ID/LP grammars; and correspondingly, there is a polynomial-size weakly equivalent CFG for each ID/LP grammar. But even with a growing alphabet, it is still possible that direct ID/LP recognition is polynomial on the average. One way to check this possibility empirically would be to examine long utterances (sentences) in actual fragments of free word-order languages, to see whether words are repeated a large number of times in those utterances. If there is a bound, and if all permutations are equally likely, then the above results may have some relevance. It is definitely the case that speculations about the difficulty of processing these languages should be informed by more actual data. However, it is equally true that the conclusions of a theoretical investigation can suggest what data to collect.

5 Proof of the IL

Here we repeat the proof of the IL due to Ogden *et al.* It is an excellent example of the combinatory fact known as the Pigeonhole Principle. As we said, we want to encourage more cooperation between theoretical computer science and linguistics, and part of the way to do this is to give a full account of the techniques used in both areas.

First we restate the lemma.

Interchange Lemma. Let G be a CFG or ID/LP grammar with fanout r, and with nonterminal alphabet N. Let m and n be any positive natural numbers with $r < m \leq n$. Let $L(n)$ be the set of length n strings in $L(G)$, and $Q(n)$ be a subset of $L(n)$. Then we can find a k-interchangeable subset A of $Q(n)$, such that $m/r \leq k \leq m$, and such that

$$|A| \geq |Q(n)|/(|N| \cdot n^2).$$

Proof. The proof breaks into two distinct parts: one involving the Pigeonhole Principle, and another involving an argument about paths in derivation trees with fanout r. The two parts are related by the following definition.

Fix n, r, and m as in the statement of the IL. A tuple (j, k, B), where j and k are integers between 1 and n, and where $B \in N$, is said to *describe* a string z of length n, if (i) there is a (full) derivation tree for z in G, having a subtree whose root is labeled with B, and the subtree exactly covers that portion of z beginning at position j, and having length k; and (ii) k satisfies the inequality stated in the conclusion of the IL. Notice that if one tuple describes every string in a set A, then, since G is context-free, A is k-interchangeable.

The part of the proof involving derivation trees can now be stated: we claim that every string z in $L(G)$ has at least one tuple describing it. To see that this is true, execute the following algorithm. Let $z \in L(G)$. Begin at the root (S) node of a derivation tree for z, and make that the "current node." At each stage of the algorithm, move the current node down to a daughter node having the longest possible yield length of its dominated subtree, while the yield length of the current node is strictly bigger than m. Let B be the label of the final value of the current node, let j be the position where the yield of the final value of the current node starts, and let k be the length of that yield. By the algorithm, $k \leq m$. If $k < m/r$, then since the grammar has fanout r, then the node above the final value of the current node would have yield length less than m, so it would have been the final value of the current node, a contradiction. This establishes the claim.

Now we give the combinatory part of the proof. Let E and F be finite sets, and let R be a binary relation (set of ordered pairs) between E and F. R is said to *cover* F if every element of F participates in at least one pair of R. Also, we define, for $e \in E$, $R(e) = \{f \mid e \; R \; f\}$. One version of the Pigeonhole Principle can be stated as follows.

Lemma 1 *If R covers F, then there is an element $e \in E$ such that*

$$|R(e)| \geq |F|/|E|.$$

Proof: Since R covers F, we know

$$|F| \leq \sum_{e \in E} |R(e)|$$

If $|R(e)| < |F|/|E|$ for every e, then

$$|F| < \sum_{e \in E} (|F|/|E|) = |F|,$$

a contradiction.

Now let E be the set of all tuples (j, k, B) where j and k are less than or equal to n, and $B \in N$. Then $|E| = |N| \cdot n^2$. Let $F = Q(n)$. Let $e \; R \; f$ iff e describes f. By the first part of our proof, R covers F. Thus let e be a tuple given by the conclusion of the Pigeonhole Principle, and let A be $R(e)$. The size of A is correct, and since e describes everything in A, then A is k-interchangeable. This completes the proof and the paper.

References

[1] Barton, G.E, Jr., The Computational Difficulty of ID/LP Parsing. *Proc. 23rd Ann. Meeting of ACL* , July 1985, 76-81.

[2] Barton, G.E., Jr., R.C. Berwick, and E.S. Ristad, *Computational Complexity and Natural Language.* MIT Press, Cambridge, Mass., 1986.

[3] Gazdar, G. Klein, E., Pullum, G., and Sag, I., *Generalized Phrase Structure Grammar.* Harvard Univ. Press, Cambridge, Mass., 1985.

[4] Gazdar, G., Applicability of Indexed Grammars to Natural Languages, CSLI report CSLI-85-34, Stanford University, 1985.

[5] Hopcroft, J., and J. Ullman, *Introduction to Automata Theory, Languages, and Computation,* Addison-Wesley, Reading, Mass., 1979.

[6] Kac, M., Manaster-Ramer, A. and Rounds, W., Simultaneous-Distributive Coordination and Context-Freedom, *Computational Linguistics*, to appear 1987.

[7] Ogden, William, Rockford J. Ross, and Karl Winklmann, An 'interchange lemma' for context-free languages. *SIAM Journal of Computing* 14.410-415, 1985.

[8] Rounds, W., The Relevance of Complexity Results to Natural Language Processing, to appear in *Processing of Linguistic Structure*, P. Sells and T. Wasow, eds., MIT Press.

[9] Rounds, W., A. Manaster-Ramer, and J. Friedman, Finding Formal Languages a Home in Natural Language Theory, in *Mathematics of Language*, ed. A. Manaster-Ramer, John Benjamins, Amsterdam, to appear.