FORMAL LANGUAGE DESCRIPTION LANGUAGES FOR COMPUTER PROGRAMMING

Proceedings of the
IFIP Working Conference on
Formal Language Description Languages

Edited by
T. B. STEEL, Jr.

1966

NORTH-HOLLAND PUBLISHING COMPANY - AMSTERDAM

CLASSIFICATION OF CHOMSKY LANGUAGES*

M. P. SCHÜTZENBERGER
France

Since our friend Dr. Ginsburg has already succeeded in defining our present knowledge of Chomsky languages (or context-free languages), I am dispensed from giving detailed exposition of the motives justifying several remarks I should like to develop in this paper on the classification of these languages.

We know that all languages L of the Chomsky type are languages that are accepted by a nondeterministic pushdown automaton. Formally, this means that if L is a Chomsky language on the alphabet X, another alphabet Y (X and Y are, of course, finite) and the following objects can be associated with L:

1. A homomorphism φ of the free monoid Y^* generated by Y in the free monoid X^* of which L is a subset.
2. A homomorphism μ of Y^* in a *finite* monoid M, and a proper, distinguished subset M' of M.
3. A homomorphism γ of Y^* in a free group G, and a finite, proper, distinguished subset G' of G.

By definition, $L = \{\varphi g \in X^* : g \in Y\ ;\ \mu g \in M'\ ;\ \gamma g \in G'\}$ and the ambiguity of a word f on X^* is the cardinal number of the set of g's, $g \in Y^X$ n $\mu^{-1}M'$ n $\gamma^{-1}G'$, that are mapped on f by φ.

The proof that this construction is possible results simply from the rewriting of the grammar producing L in a canonic form (with the help of the homomorphism φ); the proof that, conversely, every language defined in this manner is in fact a Chomsky language is a trivial consequence of elementary properties of these languages. In an intuitive way, μ symbolizes the finite part of the pushdown automaton, γ its pushdown store, and the homomorphism φ in a certain sense represents the "nondeterministic" character of the automaton.

Given a language L, it is clear that there generally exists an infinity of ways to find φ, μ, and γ that will satisfy the desired conditions, even when we fix the degree of ambiguity of every word. Nevertheless - and this is the point on which I should like to insist - a specific L imposes certain rather strong restrictions on each element of the triple (φ, μ, γ), which, in turn, implies the possibility of using these properties as classificatory principles for Chomsky languages.

* Dr. Zemanek was kind enough to furnish a translation of my original french text.

1. CLASSIFICATION BASED ON THE HOMOMORPHISM φ

In cases where φ is a monomorphism, every word of L has ambiguity one, i. e. according to the usual terminology L is "without ambiguity". Well-known counterexamples show that there are Chomsky languages possessing an essential ambiguity (for instance, $\{x^n y^m z^p\}$ with n, m, and p positive integers and $n=m$ or $m=p$) and that, consequently, there is a subfamily of Chomsky languages for which φ cannot be a monomorphism. On the other hand, I do not know any example of a language such that the minimum ambiguity increases exponentially with the word length, altough, in an intuitive way, it seems that this would constitute the general case. If one has such a counterexample, one could also separate the languages having an essential ambiguity from those which are the union of a finite number of languages corresponding to the monomorphisms, and those whose ambiguity increases like a polynomial in the word length.

Whenever φ is a monomorphism, it is possible to introduce new distinctions. Let us suppose that the set of words of the form φy ($y \in Y$) constitutes a prefix code. In this case, one can find for every word $f \in X^*$, read from left to right, the unique word $g \in Y^*$ of which it is the image by φ; with the help of a finite automaton one can verify whether g belongs to Kleene language $\mu^{-1}M'$. Using these finite devices, which either write sequentially or write just one symbol (marking the membership to M'), it is possible to verify the membership in L of a word $f \in X^*$ with a deterministic pushdown automaton only. To verify that these conditions define a *proper* subclass of Chomsky languages without ambiguity, it suffices to consider the counterexample of $L = \{\varphi g \tilde{g} : g \in Y^*\}$, where $\{g\tilde{g} : g \in Y^*\}$ is the set of "reflected words, and where φ is a monomorphism, such that $\{\varphi y : y \in Y\}$ is a generic code (i. e., with an unlimited delay). Another counterexample would be given by the reflected image of a *generic* language if it were accepted by a deterministic pushdown automaton.

2. CLASSIFICATION BASED ON THE HOMOMORPHISM μ

Classification by μ is, in fact, the most frequently utilized. One knows that it is always possible to choose μ and M' in such a way that the Kleene language $\mu^{-1}M'$ has the very particular form $M' = Y_1 Y^* \backslash Y^* V Y^*$, where Y_1 is a subset of Y, and V a subset of YY. Under these conditions, the metalinear Chomsky languages are characterized by the fact that M' has the form $Y^*_1 Y^*_2 \ldots Y^*_k$, where the Y_j's are disjoint subsets of Y and where, in addition, one supposes that the homomorphism γ is a monomorphism. If one does not make this last restriction, but supposes that each one of the Y_j's is reduced to a unique letter, one obtains languages that are very close to the "*bounded*" languages of Ginsburg. It is clear that a Chomsky language is generally neither metalinear nor "bounded", the simplest counterexample being, obviously, the reflected image by γ of a finite subset of the free group G.

In fact, the specification of V is equivalent to the specification of a relation on the alphabet Y, and it would be interesting to use the numerous results of the theory of relations to make the classification of Chomsky languages precise. As far as I know, not very much has been done in this direction, except for Ginsburg's theory of sequential languages.

3. CLASSIFICATION BASED ON THE HOMOMORPHISM γ

It is not difficult to show that, given L, one can choose (φ, μ, γ) in such a way that Y is an alphabet of four letters (i. e., G is a free group with two generators, because every generator of G corresponds to a pair of letters of Y). It seems interesting to me to emphasize that, in general, it is not possible to reduce Y and γ in such a way that F is a free group with a single generator. The counterexample is given by the language $\{x^n_1 x^m_2 x^m_3 x^n_4\}$ (with n and m two arbitrary integers) on an alphabet X formed by four letters x_i. The verification that this language (even with arbitrary ambiguities) cannot be produced by a triple (φ, μ, γ), where γ is a homomorphism in a free group with a single generator, does not present any difficulty at all. As usual, however, it implies the discussion of a considerable number of particular special cases. On the other hand, although it seems to be evident *a priori*, I have not succeeded in verifying the same result concerning the language constituted by the inverse image of a finite subset of a free group with several generators.

Intuitively, the hypothesis that G has a single generator is equivalent to the hypothesis that the pushdown store of the automaton is an (unlimited) counter, since G is isomorphic to the additive group of the integers. The "write" and "clear" operations now correspond to addition and subtraction. With the additional hypothesis that L is accepted by a nondeterministic pushdown automaton, it is easy to use ordinary considerations about the quantity of stored information to discuss the possibility of reducing the pushdown store to a counter. In the case where this additional hypothesis is not made, the nondeterministic character of the automaton is equivalent, in a certain sense, to the possibility of storing a quantity of an information that increases linearly with the length of the examined word under these conditions, however, it is no longer possible to use this shorter procedure to verify that the language has the desired properties. It seems, therefore, that the fact of admitting or not admitting a language into a group with a single generator is a rather profound property of a given Chomsky language, and I believe that this question deserves closer study.

DISCUSSION

DIJKSTRA

I gather that you are talking about finite alphabets defining words in terms of concatenation of characters of this alphabet, asking yourself whether there exist algorithms

to determine whether a given word is part of this language. Now, what I am thinking of is a very simple language, but here the decidability question poses great problems. The alphabet consists of a finite set of characters and the only words I should like to admit are words consisting of a single character. Clear? There are 26 letters, and my language consists of 26 words, a, b, c, etc. Do you consider a language of this nature as one where I can decide whether a given sequence of characters represents the word in the language or not? Do you regard this as decidable - or can we conceive an algorithm that does it?

SCHÜTZENBERGER
Insofar as the question is a catch, I will answer that I cannot answer it. I mean I am not dealing with interpretation and I don't understand it. The only problem with interpretation is that you give up on the words. If I ask you if it is decidable what is the value of the integer nearest to the value of the Bessel function of the square of 2 to the 2 to the 2, is this number odd or even? Is it decidable on that question? If I answer you "No" you tell me it can be done by a Turing machine. If I answer you "Yes" you will tell me do it and of course it will not be finished before you are all dead. [Laughter] So I suppose this is the same type of question, and I will very carefully say that I was entirely incompetent for dealing with interpretation.

DIJKSTRA
I think you misunderstood my question. You're talking about certain sequences of characters.

SCHÜTZENBERGER
No.

DIJKSTRA
People are talking about algorithms which can inspect a given sequence of elements of the alphabet, etc. You talk about scanning in sequence, thereby describing the way you are allowed to have access to the given sequence. If I give a sequence as I give a word, is it (at its expense) implied or not? I am trying to use your terminology; I gather that people talk about finite alphabets - alphabets of, let's say, 26 distinguishable characters.

DUNCAN
For example, the letters, a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, that is an example. Is this a set?

SCHÜTZENBERGER
Well, I'm no logician. I think that all these are interesting questions. The only thing is, I am entirely incompetent to answer them. I admit it frankly. I also admit that as a mathematician I have no involvement with objects that appeal to the intuition. I am involved with mathematics. The definition of mathematics is a matter of ideology.

GORN
Maybe I can help Dr. Dijkstra on this. If somebody gives an abstract theory and somebody else says "I think I have a fine application of your theory, don't you agree"? The theorist has to be very careful in answering because there may be considerations outside his theory that may arise in his answer. [Schützenberger] did not want to commit himself as to whether the alphabet, for instance, was a concrete application of a base of his monoid, because he didn't know how you were going to use the alphabet. So I think he was justifiably careful. I think it very likely that the alphabet of distinguishable letters would be an application of a base of a monoid. But there are things you might do. For instance, suppose the alphabet you chose was already a composite language over another alphabet.

ELGOT
I would like to have the question clarified. I think there is a serious question here for which he would like an answer. I'm not sure that everyone is clear on what the question is.

DIJKSTRA
I thought that one of the major concerns of people interested in the subject for today - I may be completely wrong - was that given a finite alphabet of distinguishable marks, and given rules about how to build up sequences of such marks, you can ask the inverse question about whether a given sequence of marks belongs to the set of sequences which can be constructed by the rules mentioned half a minute earlier. That's all. We take a very simple alphabet of 26 letters and the only sequences we wish to consider are those consisting of single letters. So I have a finite alphabet and even a finite number of words. Now I give you the sequence "ab". You can react to this in two ways. You can say "Now, this ab is not one of the basic characters, so it is not an element of the language". On the other hand, if I have a machine which can recognize the 26 letters, I give it the sequence "ab", so it looks at the first one and sees "a" and says "Well, it's a word". For instance, when you talk about a given sequence, is its length implied in what you give? If so, then you have apparently two other distinguishable marks you haven't said a word about, indicating where it begins and where it stops. Then I say what a curious restriction it is that you allow these marks only at the beginning and the end. Is my question now clear?

SCHÜTZENBERGER
I suppose I can answer that, in a sense. [Laughter] I very carefully said at the beginning that what you people were dealing with were mostly sequences and that I would be dealing with an almost equivalent concept, namely the elements of a free monoid. Now, insofar as you are dealing with sequences that you have any standard definition for, you are given the length. Insofar as you are dealing with elements of a free monoid you construe your hypotheses and assertions in such a way that this question is not a question. That is to say, this is a valid question but only at the level of the interpretation. When you study arithmetic you may ask all sorts of questions, such as "Are three apples the same as three Elgot's", or "Do four letters in this word have the same meaning as four letters in another word". Now when you do arithmetic you have certain definitions, which means that these questions are irrelevant. What I said mathematically does not pertain to the type of question you are raising. Whatever I say about a free monoid, it does not matter if it has a beginning or an end. It has a left factor, a right factor, and an intermediate factor which may be the empty word. By definition, there is something which we call the empty word which you can eventually interpret as the empty sequence, but it is not the same thing as the empty sequence; it's another concept entirely. I tried to go at a rarified level where these questions don't bother me anymore.