1. A class of good context free grammars.

Suppose

$$\mathcal{G} = (T, N, s, \mathcal{P})$$

is a context free grammar with the following properties:

(i) there are sets Y, E, O and L such that the family

$$\{Y, E, O, L, N\}$$

is disjointed,

$$T \subset Y \cup E \cup O \cup L.$$

(ii) if $y \in Y$ there is at most one production (r, s) such that $y \in \mathbf{rng}\, s$ which, if it exists, has the form

$$r := y$$

(iii) for each $e \in E$ there is at most one production (r, s) such that

$$e \in \mathbf{rng}\, s$$

which, if it exists, is such that

$$N \cap \mathbf{rng}\, s \neq \emptyset, \quad Y \cap \mathbf{rng}\, s = \emptyset, \quad \text{and} \quad s_0 = e.$$

(iv) for each $o \in O$ there is at most one production (r, s) such that

$$o \in \mathbf{rng}\, s$$

which, if it exists, is such that

$$N \cap \mathbf{rng}\, s \neq \emptyset, \quad Y \cap \mathbf{rng}\, s = \emptyset, \quad s_0 \in L;$$

(v) there are no other productions besides those which appear above.

We say a production (r, s) is of type A,B,C if it is as in (ii),(iii),(iv), respectively.

Theorem 1.1. The grammar $\mathcal{G}$ is good.

We will prove this theorem after we give three examples of grammars $\mathcal{G}$ satisfying these conditions.

1.1. **The language of a calculator.** Let $Y = \mathbb{Z}$, let E be the set consisting of the symbol

$$-$$

let O be the set of the symbols

$$+ \quad *$$

and let L be the set consisting of the left and right parentheses:

$$(\quad)$$

Let

$$T = Y \cup E \cup O \cup L,$$

let

$$N = \{\, \mathbf{expr} \,\}$$

We require that $\{Y, E, O, L, N\}$ be disjointed.

The start symbol is $\mathbf{expr}$.

The productions are as follows:

$$\begin{aligned}
\mathbf{expr} &:= z \quad \text{for } z \in \mathbb{Z};\\
\mathbf{expr} &:= - \ \mathbf{expr}\\
\mathbf{expr} &:= (\ \mathbf{expr} + \mathbf{expr}\ )\\
\mathbf{expr} &:= (\ \mathbf{expr} * \mathbf{expr}\ )
\end{aligned}$$

Note that there are an infinite number of productions.

1.2. **The language of propositional logic.** Let Y be a set, let E be the set consisting of the symbol

$$\sim$$

let O be the set of the symbols

$$\vee \quad \wedge \quad \rightarrow \quad \leftrightarrow$$

and let L be the set consisting of the left and right parentheses:

$$(\quad)$$

Let

$$T = Y \cup E \cup O \cup L,$$

let

$$N = \{\, \mathbf{stmt} \,\}$$

We require that $\{Y, E, O, L, N\}$ be disjointed.
The start symbol is $\mathbf{stmt}$.
The productions are as follows:

$$\begin{aligned}
\mathbf{stmt} &:= y \quad \text{for } y \in Y;\\
\mathbf{stmt} &:= \sim \ \mathbf{stmt}\\
\mathbf{stmt} &:= (\ \mathbf{stmt} \vee \mathbf{stmt}\ )\\
\mathbf{stmt} &:= (\ \mathbf{stmt} \wedge \mathbf{stmt}\ )\\
\mathbf{stmt} &:= (\ \mathbf{stmt} \rightarrow \mathbf{stmt}\ )\\
\mathbf{stmt} &:= (\ \mathbf{stmt} \leftrightarrow \mathbf{stmt}\ )
\end{aligned}$$

Note that there are an infinite number of productions if Y is infinite.
We let

$$\mathbf{p}(Y)$$

be the language defined by this grammar; the members of Y are called **propositional variables** and the members of $\mathbf{p}(Y)$ are called **statements**.

1.3. **The language of first order logic.** Let X and C be nonempty sets and let

$$Y = X \cup C.$$

Let D be the set consisting of the symbols

$$\sim \quad \forall \quad \exists$$

For each $n \in \mathbb{N}$ let R_n and F_n be sets; let

$$\mathcal{R} = \{R_n : n \in \mathbb{N}^+\}, \quad R = \bigcup_{n=1}^{\infty} R_n, \quad \mathcal{F} = \{F_n : n \in \mathbb{N}^+\}, \quad F = \bigcup_{n=1}^{\infty} F_n$$

and let $E = R \cup F$.

Let O be the set of the symbols

$$\vee \quad \wedge \quad \rightarrow \quad \leftrightarrow$$

and let L be the set consisting of the left and right parentheses as well as the comma:

$$(\quad) \quad ,$$

Let

$$N = \{\, \mathbf{var}\,, \mathbf{const}\,, \mathbf{term}\,, \mathbf{stmt}\,\}.$$

We require that the family

$$\{X, C\} \cup \{D\} \cup \mathcal{R} \cup \mathcal{F} \cup \{O\} \cup \{L\}$$

be disjointed.

The start symbol is $\mathbf{stmt}$.

The productions are as follows:

$\mathbf{var} := x \quad$ for $x \in X$;

$\mathbf{const} := c \quad$ for $c \in C$;

$\mathbf{term} := \mathbf{var}$

$\mathbf{term} := \mathbf{const}$

$\mathbf{term} := f \; (\; \underbrace{\mathbf{term} \;,\; \ldots \;,\; \mathbf{term}}_{n \text{ occurrences of } \mathbf{term},\; n-1 \text{ commas}} \;)$ for $n \in \mathbb{N}^+$ and $f \in F_n$;

$\mathbf{stmt} := r \; (\; \underbrace{\mathbf{term} \;,\; \ldots \;,\; \mathbf{term}}_{n \text{ occurrences of } \mathbf{term},\; n-1 \text{ commas}} \;)$ for $n \in \mathbb{N}^+$ and $r \in R_n$;

$\mathbf{stmt} := \forall \; \mathbf{var} \;\; \mathbf{stmt}$

$\mathbf{stmt} := \exists \; \mathbf{var} \;\; \mathbf{stmt}$

$\mathbf{stmt} := \; (\; \mathbf{term} \;\; = \mathbf{term} \;)$

$\mathbf{stmt} := \sim \;\; \mathbf{stmt}$

$\mathbf{stmt} := \; (\; \mathbf{stmt} \;\; \vee \mathbf{stmt} \;)$

$\mathbf{stmt} := \; (\; \mathbf{stmt} \;\; \wedge \mathbf{stmt} \;)$

$\mathbf{stmt} := \; (\; \mathbf{stmt} \;\; \rightarrow \mathbf{stmt} \;)$

$\mathbf{stmt} := \; (\; \mathbf{stmt} \;\; \leftrightarrow \mathbf{stmt} \;)$

Note that there are an infinite number of productions if $X \cup C$ is infinite. or if $X \cup C$ is nonempty and $R \cup F$ is infinite.

1.4. **Two basic Lemmas.** Suppose

$$\mathcal{Q} = (\mathcal{N}, \rho, p, <, f)$$

is a parse tree for $\mathcal{G}$. Let

$$\mathcal{T} = (\mathcal{N}, \rho, p) \quad \text{let} \quad \mathcal{O} = (\mathcal{N}, \rho, p, <), \quad \text{let} \quad t =< \mathcal{Q} >, \quad \text{let} \quad n = |t|$$

and let $D \geq 1$ be the depth of $\mathcal{T}$. Let

$$\lambda_0 < \lambda_1 < \cdots < \lambda_{n-1}$$

be the leaf nodes of $\mathcal{O}$; note that

$$f(\lambda_i) = t_i \quad \text{for } i \in I(n).$$

Lemma 1.1. Suppose $j \in I(n)$, $f(\lambda_j) = e \in E$, (r, s) is the production corresponding to the parent μ of λ_j, $m = |s|$ and $\mathbf{s}_{j,m}(t)$ equals the word obtained by replacing each nonterminal in s by a member of Y. Then the depth of the subtree of $\mathcal{T}$ corresponding to μ is 2.

Proof. Let $\nu_0 < \nu_1 < \cdots < \nu_{m-1}$ be the children of μ. Note that $\nu_0 = \lambda_j$. Let $i = \min\{k \in I(m) : f(\nu_i) \in N\}$ and let $(\tilde{r}, \tilde{s})$ be production corresponding to ν_i. Were it the case that $|\tilde{s}| > 1$ we would have $s_0 \in E$ or $s_0 \in L$ which is incompatible with our hypothesis on $\mathbf{s}_{j,m}(t)$. Proceeding by induction we infer that $i \in I(n)$ and $f(\nu_i) \in N$ and $(\tilde{r}, \tilde{s})$ is the corresponding production. then $|\tilde{s}| = 1$. The Lemma now easily follows. □

Lemma 1.2. Suppose $j \in I(n)$, $f(\lambda_j) = o \in O$, (r, s) is the production corresponding to the parent μ of λ_j, $m = |s|$,

$$\nu_0 < \nu_1 < \cdots < \nu_{m-1}$$

are the children of μ, $i \in I(j)$ is such that $\lambda_i = \nu_0$ $\mathbf{s}_{i,m}(t)$ equals the word obtained by replacing each nonterminal in s by a member of Y. Then the depth of the subtree of $\mathcal{T}$ corresponding to μ is 2.

Proof. Proceed in the same way as in the proof of the preceding Lemma. □

1.5. **Proof of Theorem 1.1.** Choose g, N' satisfying the following conditions:

(i) N' is a set, $g : N \to N'$, g is univalent and $\mathbf{rng}\, g = N'$;
(ii) $\{N', T, N\}$ is disjointed.

Let $\mathcal{G}'$ be the grammar obtained from $\mathcal{G}$ obtained by adjoining N' to T and adding the productions

$$r := g(r) \quad \text{for } r \in N'.$$

Suppose $s \in \mathbf{L}(\mathcal{G})$ and that for each $i = 1, 2$

$$\mathcal{Q}_i = (\mathcal{N}_i, \rho_i, p_i, <, f_i),\ i = 1, 2,$$

are parse trees for $\mathcal{G}$ such that

$$< \mathcal{Q}_i >= s.$$

We induct on $n = |s|$. In case $n = 1$ the only productions can be of the type occurring in (ii) and the truth of the Theorem 1.1 is evident.

So suppose the $n > 1$. Let D be the depth of $(\mathcal{N}_1, \rho_1, p_1)$. In case $D = 1$ we have $|s| = 1$ so $D > 1$ and at least one node μ_1 of $(\mathcal{N}_1, \rho_1, p_1)$ has depth $D - 2$ and has children which have children. Let (r, s) be the corresponding production and let $m = |s|$. Then (r, s) must be of type B or C. In case (r, s) is of type B it should be clear that there is $j \in I(n)$ and $e \in E$ such that $u = \mathbf{s}_{j,m}$ and $u_j = e$. In case (r, s) is of type C it should be clear that there are $j \in I(n)$ and $i \in I(j)$ such that $u = \mathbf{s}_{i,m}$, and $u_i \in P$ and $u_j = o$.

In either case, if $(\tilde{r}, \tilde{s})$ is the production corresponding to the parent μ_2 of λ_j we have that u is the result of replacing each nonterminal in $\tilde{s}$ by a member of Y. By Lemmas 1.1 1.2 the depth of the subtree of $(\mathcal{N}_2, \rho_2, p_2)$ corresponding to μ_2 is 2.

Now for $i = 1, 2$ replace the subtrees of $(\mathcal{N}_i, \rho_i, p_i, <_i)$ by the nodes μ_i, let f_i' equal f_i on the $\mathcal{N}_i$ with μ_i and its descendants removed and let $f_i'(\mu_i) = g(\mu_i)$. This results in parse trees for the word in $\mathbf{L}(\mathcal{G}')$ obtained by replacing u with the single letter $g(\mu_i)$. Arguing inductively we infer that these trees are isomorphic.

The truth of the Theorem should now be clear.