AN INDUCTION THEOREM FOR
DISCOVERING SYNTACTIC TRANSLATIONS.

by

Philip R. Thrift
Princeton University

Technical Report No. 160, Series 2
Department of Statistics
Princeton University
January 1980

Research supported in part by a contract
with the Office of Naval Research, No.
N00014-79-C-0322, awarded to the Depart-
ment of Statistics, Princeton University,
Princeton, New Jersey.
ABSTRACT

Given an input-output sequence of syntactic translations of sentences generated by a deterministic finite state grammar $G$ into $\Sigma^*$, a method is given for discovering the function which maps productions of $G$ into $\Sigma^*$ that gives rise to the observed translation.

1. INTRODUCTION

Let $G = (V_N, V_T, P, S)$ be a right linear grammar [2]. Thus all productions in $P$ are of the form

$$A \rightarrow aB \quad \text{or} \quad A \rightarrow a$$

where $A$ and $B$ are syntactic variables in $V_N$, and $a$ is a terminal (or word) in $V_T$. We shall assume that $G$ is deterministic, by which we mean that for every pair $(A, a) \in V_N \times V_T$ there is at most one production in $P$ of the above form. We denote the set of sentences generated by $G$ by $L(G)$.

With $G$ we shall associate what we shall call the wiring diagram $G$ of $G$. 
Definition. Let $G$ be a right linear grammar. Then the wiring diagram $G$ of $G$ is a directed pseudograph [3] with labelled arcs. The node set $N(G)$ is $V_N \cup \{F\}$, where $F$ is a symbol not in $V_N \cup V_T$. The arc set $A(G)$ is determined by the productions of $G$: if $A \rightarrow aB$ is an element of $P$ then $A \overset{a}{\rightarrow} B$ is a labelled arc of $G$; if $A \rightarrow a$ is an element of $P$ then $A \overset{a}{\rightarrow} F$ is a labelled arc of $G$.

For example, if $G = \{(S, T, U, V), \{a, b, c\}, P, S\}$ where $P = \{S \rightarrow aV|bT, T \rightarrow aT|cU|b, U \rightarrow bS|a, V \rightarrow cU|bU\}$, then $G$ is shown in Figure 1.

\begin{center}
\includegraphics[width=0.8\textwidth]{figure1.png}
\end{center}

Figure 1.
There is obviously a natural correspondence between the elements of \( L(G) \) and the set of walks from \( S \) to \( F \) in \( G \); i.e.,
\[
L(G) = \{x_1 \cdots x_n | S \xrightarrow{x_1} X_1, X_1 \xrightarrow{x_2} X_2, \ldots, X_{n-1} \xrightarrow{x_n} F \text{ are labelled arcs of } G, \text{ for some } X_1, \ldots, X_{n-1} \in V_N \}.
\]
We shall assume throughout this paper that for each \( A \in V_N \) in \( G \) there is a path from \( S \) to \( F \) that passes through \( A \).

**Definition.** Given a deterministic right linear grammar \( G \) and a finite abstract set of symbols \( \Phi = \{\phi_1, \ldots, \phi_s\} \), a **syntactic translation** is a map \( f \) from \( A(G) \) to \( \Phi^* \).

If \( A \xrightarrow{a} B \) is a labelled arc of \( G \) and if the image of this arc under \( f \) is \( \phi \) where \( \phi \in \Phi^* \), then graphically we write
\[
A \xrightarrow{a \mid \phi} B
\]
(\( \Phi^* \) is the set of finite length sequences from \( \Phi \), including \( \Lambda \), the empty string).

This definition is basically equivalent to the definition of a **generalized sequential machine** (gsm) \([1]\), where \( f \) is called an **output function**.

By extending the definition of \( f \) in the natural way we have
\[
f^{ex}: L(G) \rightarrow \Phi^* ;
\]
i.e., if we have under \( f \)
\[
s \xrightarrow{a_1 \mid \phi(1)} A_1, \ldots, A_{n-1} \xrightarrow{a_n \mid \phi(n)} F
\]
with \( \phi(1), \ldots, \phi(n) \in \Phi^* \), then the sentence
\[
a_1 a_2 \ldots a_n \xrightarrow{f^{ex}} \phi(1) \phi(2) \ldots \phi(n).
\]
In the syntactic translation as shown in Figure 2,

\[ ba^2b + \phi_5 \phi_4 \phi_5 \phi_4 \phi_1 \]
\[ acbaba + \phi_3 \phi_1 \phi_3 \phi_2 \phi_3 \phi_1 \phi_3 \phi_1 \phi_2 \]

etc.

Let \( A(G, \phi^*) \) be the set of syntactic translations of \( G \), and let \( A^e_X(L(G), \phi^*) \) be the extension of \( A \) to \( (\phi^*)^L(G) \). We shall refer to elements of \( A^e_X(L(G), \phi^*) \) as syntactic maps.
2. TREE COMPOSITIONS

Definition. Let \( \Sigma \) be a finite alphabet, and \( x \in \Sigma^* \). A **k-composition** of \( x \) is defined to be an ordered \( k \)-tuple \( c \in (\Sigma^*)^k \), \( c = (c_1, \ldots, c_k) \) having the property that \( c_1c_2 \ldots c_k = x \). The set of \( k \)-compositions of \( x \) is denoted \( C_k(x) \).

For example, if \( \Sigma = \{a, b, c\} \), then \( C_3(ab^2c) \) is the set \( \{(\Lambda, \Lambda, ab^2c), (\Lambda, a, b^2c), (\Lambda, ab, bc) \ldots \} \) where \( \Lambda \) denotes the empty word. In general, \( |C_k(x)| = \binom{n+k-1}{k-1} = \binom{n+k-1}{n} \) if \( |x| = n \).

The notion of composition is extended to trees.

Definition. Let \( \Sigma \) be a finite alphabet, \( T \) a rooted directed tree \( T = (N(T), A(T)) \). Thus \( T \) is a directed tree with a distinguished node \( R \in N(T) \), and for each node \( N \in N(T) \) there is a unique directed path from \( R \) to \( N \). The leaves of \( T \), denoted \( L(T) \subseteq N(T) - R \) are the nodes of \( T \) with degree 1. Assume the elements of \( L(T) \) are ordered \( L_1, \ldots, L_\ell \) where \( \ell = |L(T)| \). For a given element \( x = (x_1, \ldots, x_\ell) \in (\Sigma^*)^\ell \) a **T-composition** of \( x \) is defined by a function

\[
A(T) \xrightarrow{t^C} \Sigma^*
\]

having the property that for each leaf \( L_j \) of \( T \), and unique path \( a_1, \ldots, a_k \in A(T) \) from \( R \) to \( L_j \),

\[
t^C(a_1)t^C(a_2) \ldots t^C(a_k) = x.
\]

Thus a tree composition reduces to a \( k \)-composition when the tree is a rooted path consisting of \( k \) connected arcs. An example of a tree composition of \( (ab, ab, b, ba) \) is shown in Figure 3, for the complete binary tree with 7 nodes. Given \( T \), along with an ordering for the leaves, and \( x \in (\Sigma^*)^{L(T)} \) we denote the set of all tree compositions of \( x \) by \( TC(T, x) \).
Figure 3.

An element of $TC(T, x)$ can be represented as a non-negative integer lattice point in a natural way:

If $a_1, \ldots, a_{|A(T)|}$ is some ordering of the arcs, then

$$t^c(a) \rightarrow |t^c(a)| \quad a \in A(T)$$

specifies a lattice point in $L = \mathbb{N}^{A(T)}$, $\mathbb{N}$ = non-negative integers.
We denote by $S[TC(T,x)]$ the set of lattice points defined above. A partial order $\leq_T$ is defined in $L$: for $s, t \in L$

$\quad s \leq_T t \iff t$ is obtained from $s$

by moving objects up the tree.

For example,

\[
\begin{array}{c}
1 \quad 0 \\
0 \quad 0 \\
1 \\
0
\end{array}
\begin{array}{c}
1 \quad 0 \\
0 \quad 0 \\
1 \\
0
\end{array}
\quad o \rightarrow o \quad o \rightarrow o \quad o \rightarrow o
\]

We define, for $S \subseteq L$, $\text{max } S =$ the elements of $S$ having the property that for no $t \in S$: $s \leq t$, $t \neq s$.

3. THE INDUCTION PROBLEM

It is possible for two distinct syntactic translations to be extended to the same syntactic map. Thus we define an equivalence relation, $\sim$, on $S(G,\phi^*)$ by defining $f_1 \sim f_2$ iff $f_1$ and $f_2$ are extended to the same element of $S_{\text{ex}}(L(G),\phi^*)$.

The induction problem for syntactic translations is this:

an observer $O$, who we assume knows the internal structure of the wiring diagram $G$ except for the syntactic translation, can observe sentences from $L(G)$ along with their image in $\phi^*$ under the unknown syntactic translation. Thus he can observe the syntactic map for a few sentences in $L(G)$. $O$ wishes to discover an element $f \in S(G,\phi^*)$ (up to equivalence) such that $f_{\text{ex}}$ holds.

We assume $O$ can pick the sentences he wishes to observe. The theorem that follows shows, essentially, that $O$ can pick a finite
number of sentences from \( L(G) \) from which syntactic translation discovery is possible.

**THEOREM:** The syntactic translation (up to equivalence) can be discovered by observing a finite number of sentences \( W \).

**Remark:** What the theorem says is that on observing a finite set \( W \) (to be constructed below), \( O \) is presented with a finite number of word equations:

\[
\begin{align*}
a_{i_1} & \cdots a_{i_l} = \phi(1) \\
\vdots & \\
a_{k_1} & a_{k_2} \cdots a_{k_l} = \phi(k)
\end{align*}
\]

where \(|W| = k\), \( a_{mn} \in A(G) \) (the arc set of \( G \)) and \( \phi(j) \) the observed image in \( \phi^* \) corresponding to the sentence determined by the walk \( a_{j_1} \cdots a_{j_l} \) in \( G \). A solution of \( E \) (that is, an assignment of values in \( \phi^* \) to the arcs \( A(G) \) so that \( E \) is satisfied) will solve the induction problem.

**Proof:** The proof follows the construction of the implicit functions in [4].

We construct a spanning tree \( T \) in \( G \), rooted at \( S \) and connecting all nodes in \( V_N \). \( F \) is not connected to the spanning tree. For the example of Figure 1, a spanning tree \( T \) is indicated by darkened lines.

Label the arc set \( A(G) \) in such a way that \( A(T) \), the set of arcs in the spanning tree are \( a_1, \ldots, a_t \).
From $\phi$ and $A(T) = \{a_1, \ldots, a_t\}$ we create a new set of symbols. In general let $X$ be a finite alphabet $\{x_1, \ldots, x_n\}$. Then define $X^0$ to be the group freely generated by the symbols of $X$, with $A$ the identity element. Form $(\phi \cup A(T))^0$.

Begin at $F$ and consider all arcs $a$ entering $F$. Call this set $A(F)$, $A(F) \neq \emptyset$. Take an element $a$ in $A(F)$. In what follows if $a$ is the arc $A \xrightarrow{X} F$ then $\alpha(a) = A$, $\omega(a) = F$. Thus $\alpha(a) \in V_N$ and thus there is some walk $w = a_{i_1}, \ldots, a_{i_j}$, a from $S$ to $F$ with $a_{i_1}, \ldots, a_{i_j} \in A(T)$. The sentence determined by the walk $w$, call it $s$, is mapped to $\phi(s)$, which $0$ observes and writes

$$a = a_{i_1}^{-1} \ldots a_{i_j}^{-1} \in (\phi \cup A(T))^0.$$ 

This is done for each element of $A(F)$.

$0$ now considers the arcs of $A(G)$ - $(A(T) \cup A(F))$. Let $A(j)$ be the set of arcs $a$ of $G$ not in $A(T)$ such that the number of arcs in the shortest path (a walk with no repeated nodes) from $\omega(a)$ to $F$ is $j$ (i.e., $A(0) = A(F)$). Suppose $0$ has computed the equations for the arcs in $A(0), \ldots, A(j-1)$. Let $a \in A(j)$ and let $a, b_1, \ldots, b_j$ be a shortest path from $\omega(a)$ to $F$. Now $\alpha(a) \in V_N$ hence

$$a_{i_1} \ldots a_{i_j} a b_1 \ldots b_j,$$

a walk from $S$ to $F$, $a_{i_1}, \ldots, a_{i_j} \in A(T)$. If this corresponds to sentence $s$ then $0$ observes $\phi(s)$, so that

$$a = a_{i_1}^{-1} \ldots a_{i_j}^{-1} \phi b_1^{-1} \ldots b_j^{-1} \in (\phi \cup A(T))^0$$

by using the equations for $b_1, \ldots, b_j$ from previous computations. This process terminates with a list of equations.
\[ \begin{align*}
  a_{k+1} &= g_1 \\
  \vdots \\
  a_q &= g_{q-k}
\end{align*} \]

where \( g_1, \ldots, g_{q-k} \) are elements of \((\phi \cup A(T))^0\).

For example, from Figure 3 if we define the arcs

\[
\begin{align*}
  a_1 & \quad S \xrightarrow{a} V \\
  a_2 & \quad S \xrightarrow{b} T \\
  a_3 & \quad V \xrightarrow{c} U \\
  a_4 & \quad U \xrightarrow{a} F \\
  a_5 & \quad T \xrightarrow{b} F \\
  a_6 & \quad T \xrightarrow{a} T \\
  a_7 & \quad T \xrightarrow{c} U \\
  a_8 & \quad U \xrightarrow{b} S \\
  a_9 & \quad V \xrightarrow{b} U
\end{align*}
\]

Then

\[
\begin{align*}
  a_1a_3a_4 &= \phi_3\phi_1\phi_3^2\phi_2\phi_1\phi_5\phi_4 \\
  a_2a_5 &= \phi_5\phi_4\phi_3\phi_2^2\phi_4\phi_1\phi_4 \\
  a_2a_6a_5 &= \phi_5\phi_4\phi_3\phi_2\phi_5\phi_2\phi_4\phi_1\phi_4 \\
  a_2a_7a_4 &= \phi_5\phi_4\phi_3\phi_2\phi_4\phi_2\phi_1\phi_5\phi_4 \\
  a_1a_9a_4 &= \phi_3\phi_1\phi_3\phi_2^2\phi_2\phi_1\phi_5\phi_4 \\
  a_1a_3a_8a_2a_5 &= \phi_3\phi_1^2\phi_3\phi_2^2\phi_3\phi_2\phi_5\phi_4\phi_3\phi_2\phi_4\phi_1\phi_4 
\end{align*}
\]

These equations can be solved in the group \((\phi \cup A(T))^0\) by the method indicated.

It follows from [4] that, given (I), the syntactic map is the same for all assignments of \( a_1, \ldots, a_k \) to elements of \( \phi^0 \), and
-11-

hence $\phi^*$. What this means is that, given the finite equations (I), an assignment of values in $\phi^*$ to the arcs of the spanning tree $a_1, \ldots, a_k$ so that $a_{k+1}, \ldots, a_q$ as defined by (I) are in $\phi^*$ will solve the induction problem. \[1\]

A sequence $a_1, \ldots, a_k \in \phi^*$ such that $a_{k+1}, \ldots, a_q$ are in $\phi^*$ is called a feasible point.

4. THE INDUCTION SOLUTION

The structure of equations (I) will help in solving the word equations. Instead of the equation $a_{k+r} = g_r$ in (I) let us consider its associated equation $r = 1, \ldots, q-k$

$$
\phi(r) = a_{1} \ldots a_{j} a_{k+r} b_{1} \ldots b_{j}
$$

as determined in the proof of Theorem 1. Thus $a_{1} \ldots a_{j}$ denotes a descent down the spanning tree $T$, $a_{k+r}$ the unknown in (I), $b_{1} \ldots b_{j}$ a shortest path from $w(a_{k+r})$ to $F$.

From $T$ we shall construct a new tree $T'$ by adding leaves to $T$ as follows. The new leaves will be labelled $a_{j+1}, \ldots, a_q$ and will be directed respectively to the nodes $a(a_{j+1}), \ldots, a(a_{q})$.

Thus the spanning tree $T$ of Figure 1 becomes $T'$ in Figure 4. If we consider $TC(T', x)$ where $x \in (\phi^*)^{q-k}$ $x = (\phi(1), \ldots, \phi(q-k))$ is the vector of observed sentences from $\phi^*$, then obviously the set of feasible points $a_1, \ldots, a_k$ are in $TC(T', x)$ that is, $TC(T', x)$ restricted to the arcs $a_1, \ldots, a_k$. In some examples it turns out that a feasible point can be discovered by computing $\max(TC(T', x))$, but this is not always the case. Consider Figures 5 and 6.
Figure 4.

Figure 5.
Figure 6 gives \( \max TC(T',x) \) a feasible point (which is easily verified).

Figure 7 gives an example of a case where \( \max TC(T',x) \) is not a feasible point.

An obvious necessary condition, in addition to the feasible points being in \( TC(T',x) \), is

\[
|\phi(r)| = |a_{i1}| + ... + |a_{ij}| + |a_{k+r}| + |b_1| + ... + |b_j| .
\]

Note for the example in Figure 7, if we let \( |a_i| = x_i \) then

\[
\begin{align*}
x_2 + x_3 &= 2 \\
x_1 + x_5 + x_3 &= 3 \\
x_2 + x_4 + x_5 + x_3 &= 4 .
\end{align*}
\]

If \( x_1 = 3 \) and \( x_2 = 1 \), as we have in the \( \max TC(T',x) \) solution, then there is no \((x_3,x_4,x_5)\) non-negative solution.
Figure 7.
As before, we denote the word equation for the variable $a_{k+r}$ by \( a_{k+r} \in A(j) \)

\[
    a_{i_1} \ldots a_{i_{\ell}} a_{k+r} b_1 \ldots b_j = \phi(r) .
\]

Let us now assume that $b_1 \ldots b_j$ (a shortest path from $w(a_{k+r})$ to $F$) is chosen so that it is a suffix of a previously defined walk.

**THEOREM:** A sufficient condition for an assignment of arcs $a \in T$ to values in $\phi^*$ to be feasible is that it satisfies

\[
    \max TC(T',\phi) \quad \text{subject to} \quad (*) \quad |w(j)| = \phi(j)
\]

where $w(j)$ is the walk from $S$ to $F$ corresponding to the variable $a_{k+j}$.

**Proof:** Let $\hat{\phi}(a)$, $a \in A(G)$, be the "true" unknown syntactic translation, so for $r = 1, \ldots, g-k$

\[
    \hat{\phi}(a_{i_1}) \ldots \hat{\phi}(a_{i_{\ell}}) \hat{\phi}(a_{k+r}) \hat{\phi}(b_1^{(r)}) \ldots \hat{\phi}(b_j^{(r)}) .
\]

Let $\hat{\phi}(a)|_{a \in T}$ be the assignment determined by the criteria stated in the theorem.

We claim that for each $s = 1, \ldots, \ell$

\[
    \phi(a_{i_1}^s) \ldots \phi(a_{i_{\ell}}^s) \phi(a_{k+r}) \ldots \phi(b_j)
\]

is a suffix of

\[
    \hat{\phi}(a_{i_1}^s) \ldots \hat{\phi}(a_{i_{\ell}}^s) \hat{\phi}(a_{k+r}) \ldots \hat{\phi}(b_j) .
\]

If this were not true, then we would have, for some $s$, \( \phi(a_{i_1}^s) \ldots \phi(a_{i_{s-1}}^s) \) being a proper prefix of
and this contradicts maximality.

Consequently, $\phi(b_1) \ldots \phi(b_j)$ is a suffix of $\phi(r)$ (by induction, $b_1 \ldots b_j$ is of the form $a_1^r \ldots a_{i_2}^r a_{k+r}^r b_1^r \ldots b_j^r$ for a previously computed walk) $\phi(a_1^r) \ldots \phi(a_{i_2}^r)$ is a prefix of $\phi(r)$, so by (*) we have a solution in $\phi^*$ of $\phi(a_{k+r})$. \(\square\)

The example of Figure 7 shows that

\[
\begin{align*}
    x_2 + x_3 &= 2 \\
    x_1 + x_5 + x_3 &= 3 \\
    x_2 + x_4 + x_5 + x_3 &= 4
\end{align*}
\]

$\Rightarrow (x_1, x_2) \in \{(0,0), (0,1), (1,0), (1,1), (1,2)\}$.

$(x_1, x_2) = (1,1)$ corresponds to

\[
\begin{array}{c}
    1 \\
    3 \\
    2
\end{array}
\]

\[
\phi_1 \quad \phi_1
\]

\[
\max TC(T', x) \bigg|_{a \in T}
\]

subject to (*)

which is indeed feasible.

It is evident that we may replace $TC(T', \phi)$ with a set of inequalities, i.e., for the example in Figure 7 we must have

\[
\begin{align*}
    x_1 &\leq 3 \\
    x_2 &\leq 1
\end{align*}
\]

for the example in Figure 5

\[
\begin{align*}
    x_1 &\leq 3 \\
    x_2 &\leq 4 \\
    x_1 + x_3 &\leq 5
\end{align*}
\]
REFERENCES


AN INDUCTION THEOREM FOR DISCOVERING SYNTACTIC TRANSLATIONS

Philip R. Thrift

Department of Statistics
Princeton University
Princeton, N. J. 08540

Office of Naval Research (Code 436)
Arlington, Virginia 22217

January 1980

Approved for public release; distribution unlimited.

Given an input-output sequence of syntactic translations of sentences generated by a deterministic finite state grammar \( G \) into \( \Sigma^* \), a method is given for discovering the function which maps productions of \( G \) into \( \Sigma^* \) that gives rise to the observed translation.