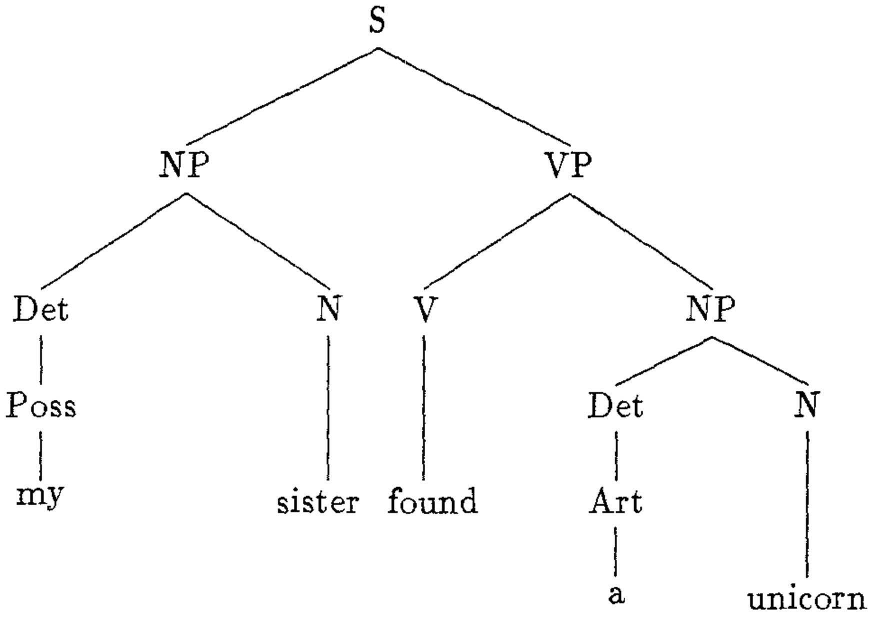

Figure 16–1: A typical constitutent structure tree

16.3 Trees

When the rules of a grammar are restricted to rewriting only a single nonterminal symbol, it is possible to contrue grammars as generating *constituent structure trees* rather than simply strings. An example of such a tree is shown in Fig. 16-1.

Such diagrams represent three sorts of information about the syntactic structure of a sentence:

1. The hierarchical grouping of the parts of the sentence into constituents
2. The grammatical type of each constituent
3. The left-to-right order of the constituents

For example, Fig. 16-1 indicates that the largest constitutent, which is labeled by S (for Sentence), is made up of a constituent which is a N(oun) P(hrase) and one which is a V(erb) P(hrase) and that the noun phrase is composed of two constitutents: a Det(erminer) and a N(oun), etc. Further, in the sentence constituent the noun phrase precedes the verb phrase, the determiner precedes the noun in the noun phrase constituents, and so on. The tree diagram itself is said to be composed of *nodes*, or points, some of which are connected by lines called *branches* Each node has associated with it a *label* chosen from a specified finite set of grammatical categories (S, NP, VP, etc.) and formatives (*my, sister,* etc). As they are customarily drawn, a tree diagram has a vertical orientation on the page with the nodes labeled by the formatives at the bottom Because a branch always connects a higher node to a lower one, it is an inherently directional connection This directionality is ordinarily not indicated by an arrow, as in the usual diagrams of relations, but only by the vertical orientation of the tree taken together with the convention that a branch extends *from* a higher node *to* a lower node.

16.3.1 Dominance

We say that a node x *dominates* a node y if there is a connected sequence of branches in the tree extending from x to y This is the case when all the branches in the sequence have the same orientation away from x and toward y. For example, in Fig 16-1 the node labeled VP dominates the node labeled Art, since the sequence of branches connecting them is uniformly descending from the higher node VP to the lower node Art. The node labeled VP does not dominate the node labeled Poss, since the path by which they are joined first ascends from VP to S and then descends through NP and Det

Given a tree diagram, we represent the fact that x dominates y by the ordered pair $\langle x, y\rangle$. The set of all such ordered pairs for a given tree is said to constitute the *dominance relation* for that tree. Dominance is clearly a transitive relation. If x is connected to y by a sequence of descending branches and y is similarly connected to z, then x dominates z because they are also connected by a sequence of descending branches, specifically, by the sequence passing through y. As a technical convenience, it is usually assumed that every node dominates itself, i.e., that the dominance relation is reflexive. Further, if x dominates y, then y can dominate x only if $x = y$; or in other words, dominance is antisymmetric. Thus, the relation of dominance is a weak partial ordering of the nodes of a tree.

If x and y are distinct, x dominates y, and there is no distinct node between x and y, then x *immediately dominates* y. In Fig. 16-1, the node labeled VP immediately dominates the node labeled V but not the node labeled *found*. A node is said to be the *daughter* of the node immediately dominating it, and distinct nodes immediately dominated by the same node are called *sisters*. In Fig. 16-1, the node labeled VP has two daughters, viz., the node labeled V and the rightmost node labeled NP. The latter two nodes are sisters. A node which is minimal in the dominance relation, i.e., which is not dominated by any other node, is called a *root* In Fig. 16-1 there is one root, the node labeled S. Maximal elements are called *leaves*, and in Fig 16-1 these are the nodes labeled by the formatives, *my, sister,* etc. Note that a tree diagram is ordinarily drawn upside down since the root is at the top and the leaves are at the bottom.

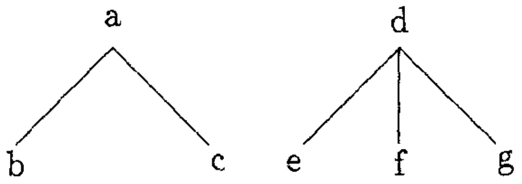

Figure 16–2: A multiply rooted "tree"

Mathematicians sometimes use the term *tree* for a configuration with more than one root, e.g., that shown in Fig. 16-2. For linguists, however, a tree is invariably singly rooted, the configuration in Fig. 16-2 being considered a "forest" of trees. We shall adhere to linguistic usage and accordingly we have the following condition:

The Single Root Condition: In every well-formed constituent structure tree there is exactly one node that dominates every node.

The root node is, therefore, a least element (and necessarily also a minimal element) in the dominance relation. We note, incidentally, that the Single Root Condition is met in the trivial case of a tree that has only one node, which is simultaneously root and leaf. The condition would not be met by an "empty" tree with no nodes at all, since it asserts that a node with the specified property exists in the tree.

16.3.2 Precedence

Two nodes are ordered in the left-to-right direction just in case they are not ordered by donimance. In Fig. 16-1 the node labeled V precedes (i.e., is to the left of) its sister node labeled NP and all the nodes dominated by this NP node; it neither precedes nor follows the nodes labeled S, VP, V, and *found*, i.e., the nodes that either dominate or are dominated by the V node. It is not logically necessary that the relations of dominance and left-to-right precedence be mutually exclusive, but this accords with the way in which tree diagrams are usually interpreted.

Given a tree, the set of all ordered pairs $\langle x, y\rangle$ such that x precedes y is said to define the *precedence relation* for that tree. To ensure that the precedence and dominance relations have no ordered pairs in common, we add the Exclusivity Condition:

The Exclusivity Condition: In any well-formed constituent structure tree, for any nodes x and y, x and y stand in the precedence relation P, i.e., either $\langle x, y\rangle \in P$ or $\langle y, x\rangle \in P$, if and only if x and y do not stand in the dominance relation D, i.e., neither $\langle x, y\rangle \in D$ nor $\langle y, x\rangle \in D$.

Like dominance, precedence is a transitive relation, but precedence is irreflexive rather than reflexive. The latter follows from the Exclusivity Condition, since for every node $x, \langle x, x\rangle \in D$ and therefore $\langle x, x\rangle \notin P$. If x precedes y, then y cannot precede x, and thus the relation is asymmetric. Precedence, therefore, defines a strict partial order on the nodes of the tree.

One other condition on the dominance and precedence relations is needed to exclude certain configurations from the class of well-formed trees. An essential characteristic of a tree that distinguishes it from a partially ordered set in general is that no node can have more than one branch entering it; i.e., every node has at most one node immediately dominating it. The structure shown in Fig. 16-3(a) has a node d with two immediate predecessors, b and c, and therefore it is not a tree. Another defining property of trees is that branches are not allowed to cross. Figure 16-3(b) illustrates the sort of structure that is forbidden. Both types of ill-formedness can be ruled out by adding the Nontangling Condition:

The Nontangling Condition: In any well-formed constituent structure tree, for any nodes x and y, if x precedes y, then all nodes dominated by x precede all nodes dominated by y.

The configuration in Fig. 16-3(a) fails to meet this condition because b precedes c, b dominates d, and c dominates d, and therefore d ought to precede d. This is impossible, however, since precedence is irreflexive. In Fig 16-3(b), b precedes c, b dominates d, and c dominates e. Thus, by the Nontangling Condition, d should precede e, but in fact the reverse is true.

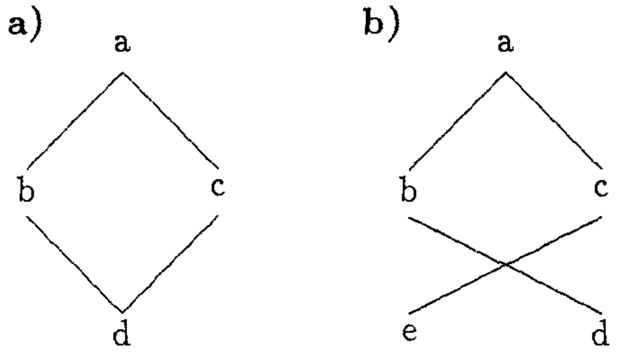

Figure 16–3: Structures excluded as trees by the Nontangling Condition

16.3.3 Labeling

To complete the characterization of trees we must consider the labeling of the nodes. It is apparent from Fig. 16-1 that distinct nodes can have identical labels attached to them, e.g., the two nodes labeled NP. Since each node has exactly one label, the pairing of nodes and labels can be represented by a *labeling function* L, whose domain is the set of nodes in the tree and whose range is a set (in syntactic trees, a set of grammatical categories and formatives). The mapping is, in general, an *into* function. In summary, we have the following definition:

DEFINITION 16.6 *A (*constituent structure*)* tree *is a mathematical configuration* $\langle N, Q, D, P, L\rangle$, *where*

> N *is a finite set, the set of* nodes
> Q *is a finite set, the set of* labels
> D *is a weak partial order in* $N \times N$, *the* dominance relation
> P *is a strict partial order in* $N \times N$, *the* precedence relation
> L *is a function from* N *into* Q, *the* labeling function

and such that the following conditions hold:

(1) $(\exists x \in N)(\forall y \in N)\langle x, y\rangle \in D$ (Single Root Condition)

(2) $(\forall x, y \in N)((\langle x, y\rangle \in P \vee \langle y, x\rangle \in P) \leftrightarrow (\langle x, y\rangle \notin D \,\&\, \langle y, x\rangle \notin D))$ (Exclusivity Condition)

(3) $(\forall w, x, y, z \in N)((\langle w, x\rangle \in P \,\&\, \langle w, y\rangle \in D \,\&\, \langle x, z\rangle \in D) \rightarrow \langle y, z\rangle \in P)$
(Nontangling Condition)

■

Given this definition, one can prove theorems of the following sort:

THEOREM 16.1 *Given a tree $T = \langle N, Q, D, P, L\rangle$, every pair of sister nodes is ordered by P.* ■

Proof: Take x and y as sisters immediately dominated by some node z. By the definitions of 'sister' and 'immediate domination,' x, y, and z must all be distinct. As an assumption to be proved false, let x dominate y. Therefore, x must dominate z, since z immediately dominates y But z also dominates x, and x and z are distinct, so this violates the condition that dominance is antisymmetric. Therefore, x cannot dominate y. By a symmetrical argument, we can show that y does not dominate x. Thus, $\langle x, y\rangle \notin D$ and $\langle y, x\rangle \notin D$, and by the Exclusivity Condition it follows that $\langle x, y\rangle \in P \vee \langle y, x\rangle \in P$; i.e., x and y are ordered by P. ■

THEOREM 16.2 *Given a tree $T = \langle N, Q, D, P, L\rangle$, the leaves are totally ordered by P.* ■

Proof: Let M be the set of leaves, and let R be the restriction of the relation P to the set M; i.e., $R = \{\langle x, y\rangle \in M \times M \mid \langle x, y\rangle \in P\}$. R is a strict partial order, since if there were any ordered pairs violating the conditions of irreflexivity, asymmetry, and transitivity in R, then because $R \subseteq P$, these pairs would also appear in P, and P would not be a strict partial order. By definition, a leaf dominates no node except itself, and therefore for every pair of distinct leaves x and $y, \langle x, y\rangle \notin D$ and $\langle y, x\rangle \notin D$. Thus, by the Exclusivity Condition $\langle x, y\rangle \in P \vee \langle y, x\rangle \in P$. Since x and y are leaves, $\langle x, y\rangle \in R \vee \langle y, x\rangle \in R$, by the definition of R, and thus R is connex. Therefore, R is a strict total order. ■

Every statement about the formal properties of a constituent structure tree can be formulated in terms of the dominance and precedence relations and the labeling function. For example, one useful predicate on trees is that of *belonging to*. A node will be said to belong to the next highest S node that dominates it. Formally, the definition is as follows:

DEFINITION 16 7 *Given a tree* $T = \langle N, Q, D, P, L\rangle$, *node* x belongs to *node* y *iff*

(1) $x \neq y$

(2) $\langle y, x\rangle \in D$

(3) $\langle y, S\rangle \in L$

(4) $\sim (\exists w \in N)(\langle w, S\rangle \in L \,\&\, w \neq y \,\&\, w \neq x \,\&\, \langle y, w\rangle \in D \,\&\, \langle w, x\rangle \in D)$. ■

Parts 2 and 3 of this definition specify that the node to which x belongs is labeled S and dominates x. Part 4 prohibits any S node from standing between x and y in the dominance relation, and part 1 excludes the case of an S node belonging to itself. To illustrate, let us consider the tree in Fig 16-4.

The node Prn belongs to the circled S node since this is the next highest S node dominating it. Prn does not belong to the highest S (i.e., the root) of the tree because the circled S node is between the root and Prn in the dominance relation

With this definition we can easily define some other predicates. Two nodes are called *clause mates* iff neither dominates the other and both belong to the same node. In Fig 16-4 the nodes labeled *John* and *him* are clause mates since neither dominates the other and both belong to the circled S node. *Fred* and *him* are not clause mates since they do not belong to the same node, and Prn and *him* are not clause mates since Prn dominates *him*.

If we let $B\langle x, y\rangle$ denote 'x belongs to y,' we can state the definition of clause mates as follows:

DEFINITION 16.8 *Given a tree* $T = \langle N, Q, D, P, L\rangle$, *nodes* x *and* y *are* clause mates *iff* $\langle x, y\rangle \notin D \,\&\, \langle y, x\rangle \notin D \,\&\, (\exists z \in N)(\langle x, z\rangle \in B \,\&\, \langle y, z\rangle \in B$. ■

A node x is said to *command* a node y iff neither dominates the other and x belongs to a node z that dominates y (Langacker, 1969). In Fig. 16-4 the node labeled *Fred* commands the node labeled *him* since neither dominates the other and *Fred* belongs to the root node S, which also dominates *him* The node *him* does not command *Fred,* however, since the node to which *him* belongs—the circled S node—does not dominate *Fred*. Note, further, that *John* commands *him* and vice versa. Formally, the definition is as follows

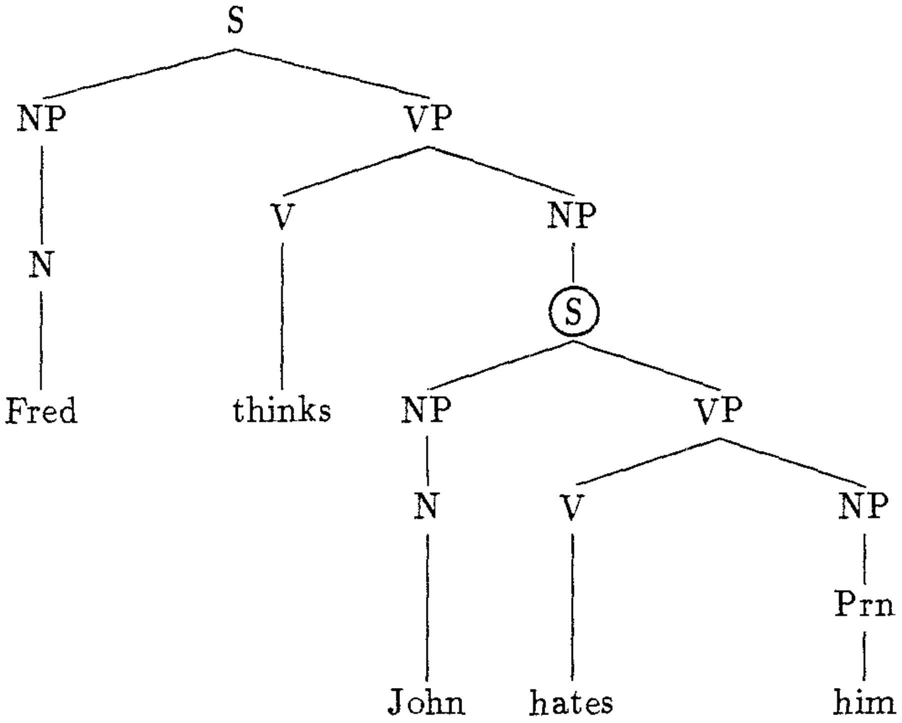

Figure 16–4: Tree illustrating the definitions of 'belonging to' and 'command'

DEFINITION 16.9 *Given a tree* $T = \langle N, Q, D, P, L \rangle$, *node* x commands *node* y *iff* $\langle x, y \rangle \notin D \; \& \; \langle y, x \rangle \notin D \; \& \; (\exists z \in N)(\langle x, z \rangle \in B \; \& \; \langle z, y \rangle \in D)$. ■

Problem: Prove that two nodes are clause mates iff each commands the other.