CONSISTENT DENDRIFICATION: TREES FROM CATEGORIES

A.M. Wallington
Department of Language Engineering, UMIST, Manchester M60 1QD, UK.

1. Introduction

I shall start by taking a fairly simple Combinatory Categorial Grammar (CCG) of the type developed by Steedman over the past decade or so (e.g. Steedman 1996) including rules of functional application, and functional composition. I shall have nothing to say about functional substitution in this paper, and shall assume that there are type-raised categories in the lexicon (e.g. S/(S\NP)). I shall also assume, following Steedman, that syntactic symbols such as S, NP, S\NP are in fact abbreviations for feature bundles.

From a Phrase Structure Grammar (PSG) perspective, a CCG derivation that uses functional composition, if interpreted as building a structural level of representation, can give rise to some very strange looking trees containing some very unusual node labels. Whereas certain labels correspond to PSG ones (e.g. VP = S\NP), others do not (e.g. S/NP), Furthermore, because certain analyses require a rule of composition, such trees and labels will be required. If there is anything at all "real" about traditional PSG categories for languages such as English, then on the face of it, CCG fails to capture them. There is a related point. If these strange categories such as S/NP need to be assembled, then one would expect that some lexical items would require such a category either as an argument or as the result. But, there seem to be curiously few such words and possibly no verbs.

What we shall do in this paper is examine how CCG categories can correspond to trees (cf. Joshi & Kulick 1996 and Henderson 1992 for other approaches). We shall see that interpreting a lexical CCG category as a partial description of a tree using a number of very simple principles will allow a number of "natural" distinctions to fall out without being stipulated. In particular, subjects but not objects will be immediately dominated by the S, different types of "empty" categories will be predicted; and structural differences between raising and control verbs will be observed. If the lexicon is constrained so that the categories can be interpreted as trees in the manner we shall describe, and if during the course of a successful derivation such trees can then be combined with other trees, then we shall say that the lexicon is constrained by a principle of "Consistent Dendrification".

2. Hypothesising Trees

As a start towards interpreting a lexical CCG category (e.g. X\Y) as a partial description of a tree, we shall assume that a category does a maximum of three things: it "names" certain nodes within a subtree (a crucial point we shall return to is that these may not be unique nodes); it describes a minimum of dominance relations (not necessarily immediate dominance); and where appropriate it describes relative precedence relations. For example, in the example given, X and Y would be two named nodes, X would dominate a subtree (is the root) which would contain the node Y and also a node dominating the lexical item (general principles which we shall spell out later determine how this item is named). Finally, because of the direction of slash, the Y argument subtree must be to the left of another node.

At this point the tree will be very under specified. However, we shall also assume a set of very general principles that can be applied to the minimum information specified in the category and these will allow other nodes to be hypothesised, named, and related to still more nodes in the tree. Finally, when a tree combines with another tree during the course of a derivation the resulting tree will be further specified.

2.1 *Principles and Conventions of Tree Building*

I shall first give two principles governing how nodes that have been hypothesised are labelled, then give two mechanisms for hypothesising nodes in a tree, and finally state a principle of economy that limits the number of nodes that can be hypothesised.

Principle of Full Correspondence: All (non-slash (and brackets)) labels in a category correspond to, i.e. they label, (not necessarily different) nodes in a tree.

For example, with the category S/(S/NP) ("whom"), nodes must have been hypothesised that can be labelled with an S_1 an S_2 and an NP, but crucially, the argument (i.e. S/NP) will not be used to label a node, because it has been separated into an S and an N. Suppose we were to an S/NP label; then, the tree will contain an S/NP node which does not correspond to any standard PSG node. If we wanted to relate CCG to standard trees, then we would have to give an

alternative category to words such as "whom" and a different analysis to long distance dependencies.

Naming Principle: Any node that has been hypothesised and does not correspond to a label in the category will be labelled with the label of the dominating node as the result part of the label and with the label of the other daughter of the dominating node as the argument part of the label. The position of this other daughter on the left or right will determine the direction of the slash.

Note that the Principle of Full Correspondence entails that functional nodes in the tree e.g. X\Y must be labelled by the Naming Principle. It will often be the case that the dominating node referred to in the Naming Principle is the nodes mother, and the other daughter is the nodes sister.

Lexical Anchor: A node is hypothesised that immediately dominates the lexical item.

Argument and Result Correspondence: (Not necessarily different) Nodes will be hypothesised to correspond to every argument (i.e. the right-hand-side of a slash), and to every result (i.e. left-hand-side of a slash) in a category.

Note the important difference between this mechanism governing the hypothesis of nodes in a tree and the Principle of Full Correspondence, governing the labelling of nodes. A node will be hypothesised for the argument S/NP in the S/(S/NP) category (and for the NP argument and the S and S results). However, it will not be labelled with a S/NP label.

We might also note the importance of the lexical anchor. Trees hypothesised from categorial grammar categories will be binary branching. Consequently, a minimal subtree will consist of three nodes. Of these three, the root node will correspond to the result part of the category and one of the daughter nodes will correspond to the argument part of the category. In higher reaches of the tree, the second daughter node will correspond to the root of a lower subtree. However, there are two situations in which this will not be the case. One such situation will be when the (functional) lexical category is split into the result and argument categories. A root node and a sister node will be hypothesised to correspond to this division, but a second daughter node will not have been hypothesised. The Lexical Anchor forms this node. The second situation can arise when an argument is itself a functional category. This will be the situation with the category of "whom" S/(S/NP). In this case, an S node will by hypothesised; an NP node, which must be on the right of some other node, will also be hypothesised, and a relation of dominance, although not necessarily immediate dominance, will be assumed between the two nodes. According to the Principle of Economy that we will introduce next, no other nodes can be hypothesised on the basis of the category of "whom". And, this is what we want, since if another daughter of S were hypothesised as a sister of the NP, then by the Naming Principle it would receive the label S/NP. It would not correspond to any conventional PSG category, and nor would it be found in trees hypothesised from simple transitive verbs, so preventing combination of trees. Finally, the NP is the object NP being questioned and such an NP can be arbitrarily low in the tree. We do not want to hypothesise exactly what this NPs sister is until the tree for "whom" has combined with trees hypothesised from other categories.

Principle of Economy: The smallest number of hypotheses about nodes, and dominance and precedence relations are made.

This principle entails that nodes and relations between nodes are not hypothesised without evidence. It also entails that if there is reason to hypothesise two nodes and these two nodes will receive the same label, then all things being equal the two labels will refer to the same node.

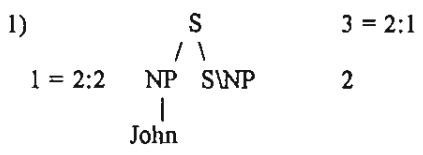
3. Sample Analyses

3.1 Type-Raised Subjects

Let us assume that the lexicon gives the following category for the proper noun "John" for use as a subject S/(S\NP). The assumption of a Lexical Anchor leads to the hypothesis of a node dominating "John" although at the moment it cannot be named. Let us call this node 1. By Argument and Result Correspondence, we can hypothesise two further nodes by splitting the category into a result part and an argument part. We shall call the node corresponding to the result node 3. Turning now to the argument, the right slash entails that there will be a node to the right of node 1 corresponding to the subtree hypothesised from the S\NP. Let us call this node 2. This subtree can also be split into an argument and a result. Consequently, we can at this point hypothesise two nodes for the subtree. By the Principle of Full Correspondence, we can label these an S and an NP. Let us call these nodes 2:1 and 2:2. Because of the left slash we also know that node 2:2 must appear on the left of some other, as yet unknown, node. Can we equate nodes 2 and 2:1, i.e. is the sister of the lexical anchor an S? At this point, this question cannot be answered since node 2:1 could also be a higher node that dominates node 2. At this stage we cannot choose between these two options, so we will leave the node unlabelled.

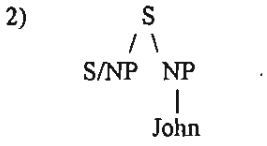
We can now turn to node 3, i.e. the node corresponding to the result part of the S/S\NP category. Since the result cannot be split into a result and an argument, we can label it with an S by the Principle of Full Correspondence.

We can return to the earlier hypotheses. The node corresponding to the S\NP argument (i.e. node 2) was required to be dominated by an S (node 2:1). The just hypothesised root node (node 3) will dominate this node and so by the Principle of Economy we shall equate nodes 3 and 2:1. Node 2:1 dominates an NP, node 2:2, which must appear on the left. We have equated nodes 3 and 2:1. There is an as yet unlabelled node on the left that is dominated by node 3 and that is node 1, the lexical anchor. Consequently, we shall equate nodes 1 and 2:2. Node 2 has not yet been labelled. However, its sister is labelled NP, and its mother is labelled S. Consequently, by the Naming Principle, node 2 will be labelled S\NP. In other words, the tree corresponding to a type-raised subject is the following:



Assuming a correspondence between an S\NP and a VP, this is the correct result.

Suppose that the lexicon contained an S\(S/NP) category for a type-raised object. It should be clear that if this were the case the resulting tree would be as depicted in 2.

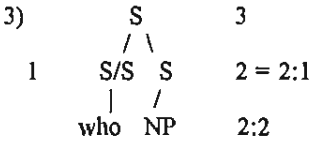


Not only does such a tree contain the S/NP label that does not correspond to a PSG label, it will not be able to combine with any tree that does not also include a S/NP label as the daughter of the S. In particular it will not be able to combine with the tree hypothesised from a simple transitive verb. In other words, if the categories in the lexicon will be interpreted as trees, then the type of category that may occur will be constrained. We can say that the lexicon is constrained by a requirement of consistent dendrification.

Wh-words

We shall assume that categories for the question words "who" and "whom" are S/(S\NP) and S/(S/NP) respectively. Notice that in terms of major features the category of a subject wh-word and that of a type-raised subject are identical. However, we have assumed, following Steedman, that labels are in fact feature bundles, and we shall assume that an S label with interrogative force has a +int feature. Consequently, a fuller description of these categories would be: S+int/(S-int\NP) and S+int/(S-int/NP).

I shall take the subject wh-word first. In the previous example, we assumed that the two Ss referred to the same node. However, in these examples, they differ with respect to the `int` feature. Much of the procedure for hypothesising a tree proceeds as before, but since the two Ss are no longer identical nodes 3 and 2:1 cannot be equated. If nodes 3 and 2:1 cannot be equated, then one S will be dominated by the other S and it will be nodes 2 (i.e. the node corresponding to the S\NP argument) and 2:1 (i.e. the result part of the S\NP argument) that will be equated. Node 2:1 dominates an NP, node 2:2. This time no other node has been hypothesised that can be equated with node 2:2. In particular, node 2:2 will not be equated with the lexical anchor node 1. A consequence of this is that no node has been hypothesised as a sister of the NP node. As discussed earlier, such a node will only be introduced when this tree combines with another tree that has an S root, an NP on the left (or right if the category is the object we-word) and a sister of the NP. Again as discussed earlier, the absence of a sister node means that the NP may be arbitrarily far from the S. Finally, if the lexical anchor (node 1) is not equated with node 2:2, then it must be named by the Naming Principle. Its mother is an S node (node 3) and its sister is also an S node (node 2:1). Consequently, the node dominating the word "whom" has the category S/S. The tree then consists of the wh-word chomsky-adjoined on the left side of a declarative sentence as depicted in 3. This again is the result we want.



3.3 Subject Raising Verbs

I shall assume that if we restrict ourselves to major features, then the category for a raising verb such as "seem" and the category of a control verb such as "try" is the same: S\NP/(S\NP) (cf. Jacobson 1990 for an alternative view).

We shall proceed as usual. A lexical anchor will be hypothesised (node 1). The category splits into an argument corresponding to the S\NP (node 2) and result corresponding to another S\NP (node 3). The argument also splits into a result (node 2:1) and an argument (node 2:2). Node 2:1 will dominate node 2:2. Since both of the categories corresponding to these nodes are atoms, these nodes will be labelled with an S and an NP respectively.

In this case, the result node (node 3) corresponds to a functional category and so node 3 will not be immediately named, and will be dominated by a node corresponding to the result (node 3:1) which will also dominate a node corresponding to the argument (node 3:2). The result of the result (i.e. the node corresponding to the S) cannot be split into an argument and result and so by the Principle of Full Correspondence, it will be labelled with an S. This is the root of the tree. Similarly, the argument of the result cannot be split, and so node 3:2 will be labelled with an NP. By the Naming Principle, node 3, which has an S mother and NP sister will be labelled S\NP.

If we have hypothesised nodes and labels for the result part of the lexical item, we can turn to the argument part. The node corresponding to this is node 2. The subtree corresponding to this node is dominated by an S (node 2:1). In this case node 3:1 is labelled with an S and dominates (although not immediately dominates) node 2. There appears to be no reason in terms of features not to equate nodes 3:1 and 2:1. However, if their daughter nodes 3:2 and 2:2 were labelled differently, then these could not be equated and as a consequence their mothers could not be equated. In this instance both are NPs and on the left. However, we might ask whether they differ in terms of minor features. In a raising construction, the subject NP has no independent theta-role projected by main verb. Its theta-role is projected from that of the subordinate verb. If we examine the lexical category, it is the subtree hypothesised from the result that will combine with the tree hypothesised from an adjacent verb. In other words NP 2:2 will be marked as taking an independent theta-role, and NP 3:2 marked as not having an independent theta-role. In such a situation, I shall assume that there is no possibility of theta-roles clashing, node 2:2 equates with node 3:2. If on the other hand, both NPs had been marked as taking independent theta-roles, then I will assume that the nodes could not be equated.

What about the label for node 2? Since node 2 was hypothesised to be dominated by an S (node 2:1) which also dominates an NP (node 2:2), it will also be labelled S\NP. We can finally return to the lexical anchor. The node corresponding to the result (node 3) that dominates it is labelled with an S\NP, and the node corresponding to the argument is also labelled with an S\NP, and so this node is labelled with an S\NP\(S|NP).

```
4)                  S                    3:1 = 2:1
                   / \
  3:2 = 2:2  NP    S\NP                  3
                  / \
  1      S\NP/(S\NP)  S\NP                2
                  |
                seem
```

Control Verbs

I shall assume that a verb such as "try" has the same category as "seem", the only difference being that the two NPs have independent theta-roles. A consequence of this difference is that nodes 3:2 and 2:2 cannot be equated. This in turn entails that the two S nodes (3:1 and 2:1) cannot be equated. Instead, node 2 will be equated with node 2:1, and will dominate node 2:2, which will have no hypothesised node yet as a sister. Finally, the label of the lexical anchor will be different from that given to it in the case of "seem". It will be dominated by an S\NP and its sister will be an S. Hence the label will be S\NP/S.

```
5)                S                    3:1
                 / \
  3:2         NP   S\NP                 3
                  / \
  1          S\NP/S  S                  2 = 2:1
                 |   /
               tried NP                 2:2
```

Henderson, J. 1992. "A Structural Interpretation of Combinatory Categorial Grammar." Technical Report MS-CIS-92-49, CIS, University of Pennsylvania.

Jacobson, P. 1990. "Raising as Function Composition." Linguistics & Philosophy 13, 423-476.

Joshi, A and S Kulick. 1997. "Partial Proof Trees as Building Blocks for a Categorial Grammar." Linguistics & Philosophy 20 637-669.

Steedman, M. 1996. "Surface Structure and Interpretation". Cambridge, Mass.: MIT Press.