# PRONOUNS AND VARIABLES

Lauri Karttunen
University of Texas at Austin

It has been known for a long time that the role of variables in symbolic logic strongly resembles the role of pronouns in natural languages, although it seems that logicians have generally been more aware of this than linguists. Among others, Willard van Orman Quine and Peter Thomas Geach have repeatedly drawn attention to this fact.[1] For a student in symbolic logic, it is very important to grasp this relationship between pronouns in the vernacular and variables. Many linguists have recently also become interested in the matter for the following reason.

In early transformational studies, it was generally assumed that pronouns were to be derived from underlying ordinary NPs by a transformation. The difficulties in this approach are by now well-known.[2] If pronouns cannot be derived, they must be in the base and the restrictions on possible pronoun-antecedent pairs must be accounted for in some other manner. There are at least two ways in which one may try to achieve this. There is a school of thought which holds that coreference relations are assigned by interpretive rules at some point in the derivation of the sentence.[3] The difficulties in this approach seem mostly the same that were unsolvable in the earlier theory of pronominalization. The other possibility is that coreference is marked in the base and that pronoun-antecedent restrictions perhaps can be stated in terms of well-formedness constraints on P-markers. This line of research is now being pursued by Emmon Bach.[4]

But what is then the underlying representation of a pronoun? It does not have to be specified with respect to number, gender, animateness, or the like, these features being predictable once we know what the pronoun refers to; a plain referential marker is enough. This brings base structures a step closer to the austere formulas of predicate calculus. But what kind of referential marker is it? Emmon Bach (1968), James D. McCawley (1967, 1968) and many others have argued that it is not enough to have referential constants, say integers, but that there also must be referential variables, bound by quantifiers that act very much like quantifiers in symbolic logic. It just might be that referential indices really are bound variables.

---

[1] See for example Geach 1966, p. 111.

[2] A concise summary of these is given in Bach 1969.

[3] See Jackendoff 1968, Dougherty 1968.

[4] Bach forthcoming.

How else, for example, would we account for the fact that (1a) is ambiguous but (1b) is not?

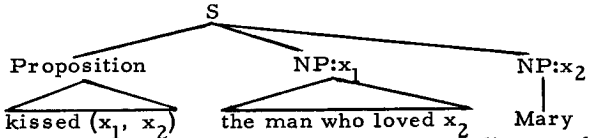
(1) (a) Every boy in town loves some girl.
(b) Every boy in town loves some girl but Mary doesn't like her.

It appears that there are many reasons for believing that the connection between pronouns and variables is even deeper than logicians were inclined to think. Let us now go into more detail. Consider example (2).

(2) The man who loved Mary kissed her.

According to McCawley (1967), the underlying representation of (2) would be something like (3).

(3)

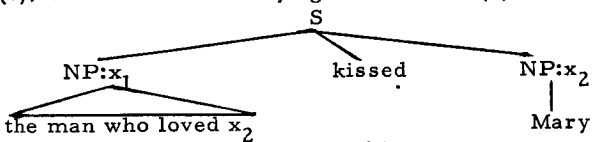



This could probably be translated into ordinary predicate calculus as (4), although the latter representation conceals the distinction between the proposition and the two referring expressions: the man who loved $x_2$ and Mary.

(4) $\exists x \exists y$ [kissed (x, y) . $x = (\iota z)$[man(z) . loved(z, y)] . y = Mary]
'Some x kissed some y and this x is the man who kissed y and y is Mary'

In the course of the derivation of (2) from (3), the two noun phrases are substituted one by one for a variable that bears the same subscript. By replacing the first term in the proposition by $NP:x_1$ and the second variable by the second NP, we eventually get the P-marker in (5), which is the underlying structure of (6).

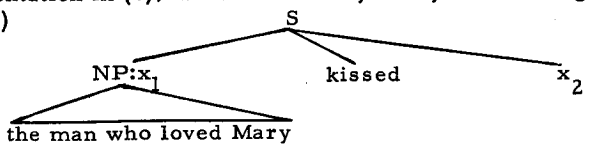
(5)




(6) The man who loved her kissed Mary.

But we also have the choice of substituting the NP Mary, not for the $x_2$ in the proposition but, for the term $x_2$ in the description the man who loved $x_2$. In this case we get the intermediate representation in (7), which eventually will yield the original sentence (2).

(7)




(2) The man who loved Mary kissed her.

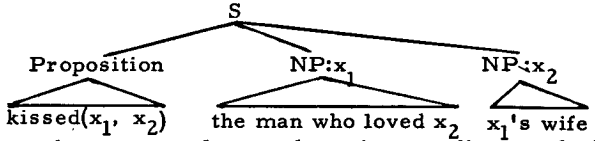
There are obvious advantages in deriving sentences in this manner. For example, the synonymous sentences (2) and (6) have the same underlying representation (3). There is also no need for a pronominalization rule of the traditional kind, since remaining variable symbols

become pronouns. So far so good. But let us now see whether there are any cases where a pronoun cannot be derived from an underlying bound variable. Consider what happens when we replace Mary in (2) by a phrase such as his wife, where his is to be understood in the sense 'his own'.

(8) The man who loved his wife kissed her.

This would have the underlying representation in (9).

(9)




This P-marker apparently translates into predicate calculus as (10).

(10) $\exists x \exists y$ [kissed (x, y) . $x = (\iota z)$[man(z) . loved(z, y)] . $y = (\iota w)$wife(w, x)]

'Some x kissed some y and this x is the man who loved y and this y is x's wife'

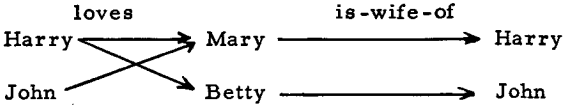
It is not immediately obvious that there is anything wrong with the above analysis. However, let us consider (9) and (10) a little more carefully. They contain the definite description the man who loved $x_2$ where $x_2$ of course represents the person who is this man's wife. Now, there are certain presuppositions that have to be fulfilled when a definite description such as this is used. To take an old example, any sentence containing a reference to the present king of France is out of place when there is no king of France. It would also be inappropriate in case there were two recognized kings of France. Similarly, the phrase the man who loved $x_2$, where $x_2$ names some person, presupposes that there be one and only one person to whom the description fits in some limited universe of discourse. What this means is that whenever someone utters

(8) The man who loved his wife kissed her.

it should be the case that the man referred to is the one and only man who loved the person who is referred to as his wife. At least, this is how things should be, if the underlying structure of (8) is as given in (9) and the NP $x_1$'s wife replaces the term $x_2$ in the man who loved $x_2$.

In order to see if this really is true, let us set up an arbitrary, small universe of discourse, such as in (11).

(11)




We have four people: Harry, John, Mary, and Betty. The following facts are supposed to be known. Mary is Harry's wife, Betty is John's wife, Harry loves Mary and Betty but John just loves Mary. To be sure - it is an unhappy and confusing situation, but quite possible in real life. We observe that Mary is loved by two men: Harry and John. For that reason, it would be inappropriate to refer to

either one of them as the man who loved Mary. On the basis of the above data, one could never find out which of the two men is meant in (2).

(2) The man who loved Mary kissed her.

From what was said about the presuppositions associated with (9), it follows that (11) is just the kind of situation where it should also be inappropriate to say:

(8) The man who loved his wife kissed her.

The argument would run as follows. If the man who loved his wife refers to Harry, then the NP his wife refers to Mary, because Mary is Harry's wife. But this cannot be right, because we just observed that there was no man we could properly describe as the man who loved Mary.

Of course, this reasoning is fallacious. There is no difficulty at all in interpreting (8) in the context of (11). Of the two men, only Harry loves his wife, namely Mary. In this context, (8) is just another way of saying that Harry kissed Mary.

We see now that, after all, (9) cannot be a correct analysis of (8), since one of the presuppositions that go with (9), namely that there be a unique man who loved $x_2$, does not have to be fulfilled for (8) to be interpretable. On the other hand, we notice that in the case of

(2) The man who loved Mary kissed her.

the proposed underlying structure (3) quite correctly entails that (2) should be inappropriate in situations such as (11).

But what is then the correct analysis of (8). We have shown that (8) may be perfectly appropriate while the NP his wife is not interchangeable with another NP, say, with the proper name Mary, although the two NPs seem to refer to the same person. In particular, we cannot assume that underlying the NP his wife there is a referential marker $x_2$ which is replaced later by the description his wife. A referential marker in this position is just like a proper name and has the same undesirable consequences. What this failure of substitutability means is that the NP the man who loved his wife is what Quine (1960) calls an 'opaque construction'.[5] By his definition, an opaque construction is one in which you cannot in general replace a singular term with another coreferential term without disturbing the truth value of the containing sentence. It wasn't actually the present type of problem Quine was discussing, but his definition applies nevertheless.[6]

5 See Quine 1960, §§ 30-32.

6 What Quine was studying were expressions of 'propositional attitude', i.e. sentences of the type A believes that p, A doesn't realize that p.

In Quine's terminology we say that the term inside an opaque construction which cannot be replaced without altering the truth value of the sentence, the NP his wife in our case, is not in a 'purely referential position'.

Now what is it about the NP the man who loved his wife that makes it an opaque construction? It is obviously the supposed reflexivity of his. In our example, we have assumed that the pronoun his and the noun phrase as a whole are coreferential. But if we do not take his in the sense of 'his own', the construction becomes referentially transparent. For example, suppose that we take his to refer to John. In the context of (11), sentence

(8) The man who loved his wife kissed her.

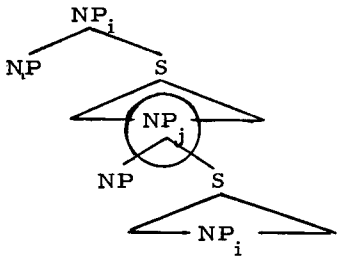
is now equivalent to

(12) The man who loved Betty kissed her.

The NP his wife is freely interchangeable with any coreferential NP as long as we do not mean 'his own'.

It seems that the kind of opacity we are discussing here is limited to constructions of the type which traditionally would be represented as in (13).

(13)



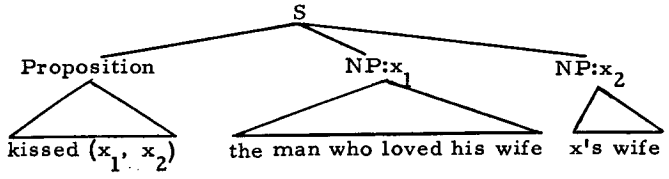

The circled NP, his wife in our example, is the one that is not in a purely referential position.[7] Or - to use terminology proposed by Keith S. Donnellan (1966) - it is to be understood attributively and not referentially.

Let us now go back to our attempt to represent the underlying structure of (8) in the type of system proposed by McCawley. It is now clear that instead of (9) we would have to write something like (14), which probably would translate into ordinary predicate

7 It makes no difference whether the top NP is definite - as in our examples - or indefinite. In the context of (11), both Harry and John are 'men who loved Mary', but only Harry is 'a man who loved his wife'.

calculus as (15).

(14)




(15) $\exists x \exists y$ [kissed (x, y) . $x = (\iota z)$[man(z) . loved (z, $(\iota w)$wife (w, z) )] . $y = (\iota v)$wife(v, x) ]
'Some x kissed some y and this x is the man who loved his wife and this y is x's wife'

This representation seems unassailable on semantic grounds, however, to derive

(8) The man who loved his wife kissed her.

we obviously would need some sort of pronominalization rule. Otherwise, we would only get (16).

(16) The man who loved his wife kissed his wife.

What is puzzling about (8) is that his wife in a non-referential use can nevertheless be the antecedent of a pronoun that is outside the opaque construction. I know of no way to indicate that kind of coreferentiality in McCawley's system, if referential markers are bound variables. We have to conclude that, if base structures are as close to being formulas in symbolic logic as we have assumed, then there must be a pronominalization rule, but at the same time we don't quite know how to state it. Since (9) cannot possibly be a correct analysis of

(8) The man who loved his wife kissed her,

there also must be some constraints that prevent the latter from being derived from (9), say, some restrictions on the replacement of variables.

It is not just the case that we have definite descriptions that are non-referential. What makes the problem worse is the fact that pronouns can also be used non-referentially. Suppose that we interchange the NP his wife and the pronoun her in (8). This results in (17).

(17) The man who loved her kissed his wife.

But this is an example of the type of sentence that was discovered

by Emmon Bach and Stanley Peters (Bach 1967). Bach-Peters sentences are known to be problematic, because, under some traditional assumptions about pronominalization, they could not be derived from finite underlying structure. In another context I have already pointed out that the inadequacy of all proposed solutions to the Bach-Peters paradox is due to the fact that in these sentences pronouns appear in positions which are not purely referential. Suppose that (17) is to be interpreted in the context of (11). If it is interpretable at all, (17) is equivalent to

(8) The man who loved his wife kissed her.

which in turn is equivalent to saying that Harry kissed Mary. But notice that in interpreting the man who loved her in (17) as referring to Harry, we do not take the pronoun her as referring to Mary. What her stands for is the description his wife, not the person that might be referred to by this description. On the other hand, the pronoun his in his wife must refer to Harrry and not to the description the man who loved her.[8]

I have no good solution to offer to the Bach-Peters paradox, but it seems to me that in discovering what the problem is really about one makes at least some progress. The appearance of non-referential pronouns is by no means limited to Bach-Peters sentences. They also turn up in the following examples.

(18) The man who gave his paycheck to his wife was wiser than the man who gave it to his mistress.
(19) I am going to give each of you a cookie. If someone doesn't want to eat it now, he can save it for later.

Looking at (18), it is obvious that, in some intuitive sense, the NP his paycheck is the antecedent of the pronoun it. However, it is not the case that his paycheck and it are coreferential. What the it refers to is certainly not the other man's paycheck, it stands for the description his paycheck. But notice that this is true only in case we interprete his in the sense of 'his own'. If we take his to mean 'somebody else's', the pronoun it can only be interpreted referentially and there is but one paycheck involved. It is this non-referential use of pronouns that we do not presently know how to handle. Other than that, I do not see that there is anything special involved in the Bach-Peters sentences.

8 In Karttunen 1969, I have argued that some Bach-Peters sentences, such as The pilot who shot at it hit the Mig that chased him are in fact ambiguous, so that either one of the two pronouns can be understood non-referentially.

To summarize briefly, I have attempted to show that there are at least two cases in which a pronoun does not translate into a bound variable in predicate calculus. These examples should be enough to disconfirm the view held by some that, for a theory of reference, "it is all one whether we consider bound variables or pronouns of the vernacular".[9] If deep structures are as close to being formulas in predicate calculus as has been suggested, there are pronouns which must be derived by deletion.

9 Geach 1966, p. 112.

# BIBLIOGRAPHY

Bach, Emmon (1967) "Problominalization I-II." Mimeo.

_____ (1968) "Nouns and Noun Phrases," in Bach & Harms (eds.) Universals in Linguistic Theory, New York: Holt, Rinehart & Winston.

_____ (1969) "Anti-pronominalization." Mimeo.

_____ (forthcoming) "Binding."

Donnellan, Keith S. (1966) "Reference and definite descriptions" Philosophical Review, 75, 281-304.

Dougherty, Ray C. (1968) "A Comparison of Two Theories of Pronominalization and Reference." Mimeo.

Geach, Peter Thomas (1966) Reference and Generality. An Examination of Some Medieval and Modern Theories. Amended Edition. Ithaca, N.Y.: Cornell University Press.

Jackendoff, Ray S. (1968) "An Interpretive Theory of Pronouns and Reflexives," Mimeo.

Karttunen, Lauri (1969) "Migs and Pilots." Mimeo.

McCawley, James D. (1967) "Where Do Noun Phrases Come From?" to appear in Jacobs and Rosenbaum (eds.) Readings in Transformational Grammar.

_____ (1968) "The Role of Semantics in a Grammar," in Bach & Harms (eds.) Universals in Linguistic Theory. New York: Holt, Rinehart & Winston.

Quine, Willard Van Orman (1960) Word and Object. Cambridge, Mass.: The MIT Press.