A Letter from London
The Phonologist's Dilemma: A Game-theoretic Approach to Phonological Debate

Jonathan Kaye, SOAS LONDON

GLOW Newsletter 21, August 1988

In this letter I would like to consider a question of phonological analysis from the point of view of game theory. My conclusion is that for a certain class of questions, cross-theoretic discussion is useless. I will begin by defining two phonological theories: Theory A is characterized as strictly privative, all objects in its theoretical space are present or absent; the absence of an object cannot be interpreted theoretically. Theory B is not privative in this sense. Objects may occur in varying degrees on a scale (typically two degrees) and nothingness may receive (eventually) an interpretation; that is nothingness may be transformed into a theoretical object. Let us now flesh out these theories by furnishing some objects. Theory A has the primitives, a, b and g Theory B has +a, -a, +b, -b, +d, -d where + end - represent two (extreme) scalar values for some object in the theory. Theory B is symmetrie in the sense that +a and -a have the same theoretical status; either may be manipulated by the mechanisme of the theory. Theory A is asymmetrie; Objects are not twinned—there is no corresponding ¬g, for each g. Although not essential for this discussion, it should be noted that nothing prevents Theory B from mixing symmetry with asymmetry. That is, starting from simply e, one could apply interpretative conventions to translate e info, say, +e and nothingness into -e. Both +e and -e are of course real objects in Theory B and may be manipulated by the mechanisme of that theory. In Theory A, only some e is real. Nothingness is not interpretable and consequently may not be manipulated.

Now consider some phenomenon in a given language. In Theory A this phenomenon is interpretable as combining one primitive, a, with one or a series of other primitives. In Theory B, there is a choice. One scalar value, say, +a may be the agent of this phenomenon, or alternatively its "opposite" -a may be the designated agent. Thus, two possible analyses are available to Theory B, one employing +a and another employing -a. Theory A may only use a. In order to proceed I must assume a translation table that provides a partial list of correspondances between the objects of theory A and those of Theory B.

(1) Theory A aTheory B

a	-a
-	+a
b	+b
-	-b
g	-d
-	+d

Notice that certain objects of Theory B are uninterpretable in Theory A. Put another way, Theory A denies their existence. There should be no necessary correlation between a scalar value in Theory B and its existence or non-existence in Theory A. B objects with the value + need not always correspond to objects in A. I have bied to reflect this fact in (1).

Let us now apply this situation to a concrete case: Hungarian vowel harmony. The facts are irrelevant here. I wish to discuss only the formal propertjes of Theories A and B and

how each approaches this phenomenon. The discussion of Hungarian vowel harmony involves the correspondance table shown in (2).

(2) Theory A Theory B
FRONT —BACK
- +BACK
ROUND +ROUND
- -ROUND
NOTHIGH —HIGH
- +HIGH

Theory A has three objects that may be manipulated; Theory B has six. Now consider the following question: What is the agent of Hungarian vowel harmony (i.e. the front-back kind)? Theory A says it must be FRONT; Theory B offers two possibilites: —BACK or +BACK. Now to the point of this letter: is the question worth discussing? I assume two individuals in the discussion which allows for the following combinations (A = an adherent of Theory A and B = an adherent of Theory B):

(3) i. A€A ii. A€B iii. B€B

Now I set up a two dimensional "payoff board" with the columns representing the truth or falsity of the claims and the rows corresponding to the theories in question. The values of each cell are the outcome of a particular situation, viz. the impact of the veracity of the analysis on the theory in question. I begin with *A€B* interactions.

(4) Payoff board for the claim, '+BACK is the agent of Hungarian vowal harmony?'
FIGUUR

The interpretation of (4) is as follows: suppose that the claim turns out to be true. This is inexpressible in theory A and it explodes. Theory B wins in such a case. Now suppose that the claim is false. Theory A is still in business; it has survived. But then so has Theory B. The correct strategy for Theory B is now to claim that -BACK is the agent of Hungarian vowel harmony. This gives us a new payoff board.

(5) Payoff board for the claim, '-BACK is the agent of Hungarian rowal harmony?'
FIGUUR

I call the combination of (4) and (5) a *Phonologist's Dilemma* situation. It is clear that Theory B now has an optimal strategy: create a succession of situations with payoff boards of the form (4) all the while retaining boards like (5) as a fall-back strategy. The idee is to find "holes" in the translation table—places where Theory A has no corresponding object—end then posit a phenomenon that assumes the existence of the Theory B object. Let us call this the *Iterated Phonologi&t's Dilemma Strategy.*

Player A now continues with a situation that is formally identical to the Hungarian vowel harmony case. Consider, say, ATR-ness. We have the following correspondance table:

FIGUUR

Player B now proposes, '-ATR is the agent of Yoruba harmony.'

(7) Payoff board for the claim, '—ATR is the agent of Yoruba harmony?'
FIGUUR

Now what about Player A? Clearly the only rational strategy is to refuse to play. It is not in Player A's interest to discuss such questions with Player B since the results of such a debate could have no conceivable consequence for the status of Theory B. Player A is in a classic no-win situation. At best Player A survives the current test but since Player B will follow the optimal strategy the A player will only have to face the same situation at the next encounter. Player B has nothing to lose in such debates and could benefit if eventually one of the claims s/he advances turns out to be true. This results in the destruction of Theory A which could be construed as beneficial to Theory B. In sum, Player A should avoid discussion with Player B in the domain of the kinds of claims referred to above. Player B may have some marginal interest in such discussions to the extent that s/he cares about the continued existence of Theory A.

So much for the interactions of the type A€B. What about A€A interactions. Clearly, they must come out equal in any discussion of such claims; they sink or swim together. Now suppose we had an additional factor, something called *TRUF*. *I* am using *TRUF* as some external evaluatory factor agreed upon by both players. *TRUF* means succeeding in a game where the rules have been agreed upon by the players. We can think of it as putting together the pieces of a jigsaw puzzle and winding up with an interpretable picture. Now, in A€A interactions is it worth discussing claims of the sort presented above? On the positive side there is the *TRUF* payoff*;* knowing when something is wrong. On the negative side there is the abandoning of a theory; going back to square one as it were. Obviously if *TRUF* is given a sufficiently high payoff, then such discussion is worthwhile. Players gain more by learning that their current theory is wrong than by persisting in maintaining theories that fail the *TRUF* criterion.

Finally, this brings us to B€B interactions. Are such claims worth discussing among the B players? The payoff table is identical for all such players. Notions of *TRUF* are inapplicable here because the discussion of these individual claims could never have a bearing on this issue. Temporarily placing the piece of a jigsaw puzzle in an incorrect position has no bearing on the existence of the final picture. The question is thee, why should B Players even be interested in discussing the sorts of claims outlined above?