Notre Dame Journal of Formal Logic
Volume XXI, Number 1, January 1980
NDJFAM

A MODEL-THEORETIC EXPLICATION OF THE THESES OF KUHN AND WHORF

JOHN A. PAULOS

1 *Introduction* We wish in this paper* to give a mathematical explication of the ideas usually associated with the names of Thomas S. Kuhn and Benjamin L. Whorf. These ideas concern the incommensurability of scientific theories and the effect of language on thought. We also touch on some related notions and applications. Hopefully our model-theoretic formulation will also have some interest for logicians and set theorists.

2 *Preliminaries* We consider several languages and the models associated with them; however, we want to characterize the models of our languages independently of any particular one of them. By taking all our models to be models of a certain set theory (we take Zermelo-Frankel set theory, **ZF**, for the sake of definiteness) and by interpreting the non-logical constants and relations of our languages to be fixed elements of the universe, the class of models in which a sentence of a language is true can be considered to be simply a class of models of **ZF**.

To be more precise we need the following definition:

Definition 1 A language $\mathcal{L}$ is of the form $\mathcal{K} \cup \{\varepsilon\}$ where $\mathcal{K} = \{c_j, R_j, f_j, Q_j\}$ is a finite collection of constant symbols, relation symbols, function symbols, and sort symbols, and where ε is a distinguished binary relation symbol. Sentences in $\mathcal{L}$ are built in the usual inductive way.

We also deal with languages of the form $\mathcal{L} = \mathcal{K}_1 \cup \mathcal{K}_2 \cup \{\varepsilon\}$ where the $\mathcal{K}_i$ are different formal languages. The intention is to think of the languages $\mathcal{K}_i$ as formal scientific languages. The symbol ε is added so that statements in any $\mathcal{K}_i$ as well as extra-linguistic observations can be described in some neutral formal language. Although we assume the universe to be set-theoretic, we do not assume that a $\mathcal{K}_i$-scientist thinks in terms of sets, but

*Some of these ideas were first formalized by Randall, from whose work [4] parts of this paper are adapted.

Received March 9, 1977

only that an all-knowing super-scientist could analyse the $\mathcal{K}_i$-sentences in set-theoretic terms.

Definition 2 A scientific theory T expressed in a language $\mathcal{L} = \mathcal{K} \cup \{\varepsilon\}$ is a set of sentences $T = \mathbf{ZF} \cup S \cup D$ where **ZF** stands for the Zermelo-Frankel axioms of set theory, S is a set of (scientific) sentences in $\mathcal{K}$, and D is a finite set of sentences in $\mathcal{L}$ of the form $\forall y(y \; \varepsilon \; R \rightarrow y = \langle z_1, \ldots, z_n \rangle)$ stating that the element interpreting R is an n-ary relation. Similar sentences concerning the interpretation of constant, function, sort, and other relation symbols also appear in D.

The semantics of a language $\mathcal{L} = \mathcal{K} \cup \{\varepsilon\}$ is described in the following definition.

Definition 3 The universe of a model is any set M large enough to contain an element corresponding to each object, relation, function, etc., in the world. (It may contain extra elements.) That is, we conceive of every object, property, etc., in the world as being associated with an element in M. (Our ontology is intended to be very flexible and for our purpose here needn't be made more precise.) The interpretation of each c_j, R_j, f_j, Q_j in $\mathcal{L}$ is a *fixed element* in M while the interpretation of ε can be any binary relation on M. Truth of a sentence ψ in a model $(M, \ldots)$ is defined in the usual inductive way.

Given our definition of theories T and the semantics for any language $\mathcal{L}$, we can see that a theory T is going to hold only in those models $(M, \ldots)$ in which the interpretation of ε is such that: (i) the axioms of **ZF** hold, (ii) the sentences in D, stating that the element in M that interprets R is an n-ary relation, that the element in M that interprets Q is of the appropriate kind, etc., are all true, and (iii) the sentences S in $\mathcal{K}$ which contain the content of the theory are all true.

To reiterate, our languages $\mathcal{L}$ are such that the interpretations of the symbols in the $\mathcal{K}_i$ are fixed elements in M and the sentences D in our theories ensure that these fixed elements of M are of the appropriate kind: n-ary relations, m-ary functions, etc.

The sort symbols are useful since natural languages refer to non-homogeneous universes composed of elements of different sorts—animate vs. inanimate, abstract vs. physical, etc. Universal statements thus generally refer only to those elements in a particular sort. Since we want our development to reflect this stratified universe we use sort symbols (rather than extra unary predicates). Randall's development uses ideas from transformational grammar and is more realistic, but less flexible and less applicable than our model-theoretic approach. Unlike Randall we consider all our sentences to be meaningful. Meaningless sentences, however, could be easily accommodated in our formalism by modifying the D sentences in any theory to allow sort confusions, empty sorts, etc.

The last definition needed to complete our preliminaries follows:

Definition 4 An observation O is of the form $\langle \phi(x_i, \ldots x_n), a_1, \ldots a_n \rangle$ where $\phi(x_1, \ldots x_n)$ is a formula in the language of set theory, ε, and where $a_1, \ldots a_n \; \varepsilon \; M$. The a_i needn't be named by $\mathcal{K}$.

The definition of an observation O is extra-linguistic, independent of any particular $\mathcal{K}$-language. The class of all models of **ZF** whose universe is M is denoted simply by $\mathfrak{M}$. All those models associated with any O are denoted $\mathfrak{M}_O$ and equal $\{(M, \ldots) \mid \phi_M(x_1 \ldots x_n) \text{ holds for } a_1, \ldots a_n\}$. Not required is that every observation be true of the "real world", M_R. For a true observation, however, we have $M_R \,\varepsilon\, \mathfrak{M}_O$. The class of all models of a scientific theory T is denoted by $\mathfrak{M}_T$, all models of a sentence γ by $\mathfrak{M}_\gamma$.

Definition 5 (i) O is expressible by a $\mathcal{K}$-sentence γ if $\mathfrak{M}_O = \mathfrak{M}_\gamma$, (ii) O verifies γ if $\mathfrak{M}_O \subseteq \mathfrak{M}_\gamma$ and O falsifies γ if $\mathfrak{M}_O \cap \mathfrak{M}_\gamma = \phi$, and (iii) γ delimits O if O verifies γ.

Theorem 1 *Every $\mathcal{K}$-sentence γ in a given formal language $\mathcal{L}$ expresses an observation, but not every observation is expressible by some $\mathcal{K}$-sentence γ. Every observation is, however, expressible by a $\mathcal{K}'$-sentence γ in some formal language $\mathcal{L}'$.*

Proof: The $\mathcal{K}$-sentence γ expresses an observation since it can be translated into a sentence $\phi(x_1, \ldots x_n)$ in the language consisting of just $\{\varepsilon\}$. $\phi(x_1, \ldots x_n)$ is satisfied by the a_i since the interpretations of the constant, relation, function, and sort symbols in γ are the a_i in M. The sentences in D ensure that the a_i are of the right sort. Thus, for example, we have $(M, \ldots) \models (\forall z \exists y R(z, y, c))$ if $(M, a_1, a_2, \ldots) \models \phi(x_1, x_2)$ where $\phi(x_1, x_2) = (\forall z \exists y(z, y, x_1) \,\varepsilon\, x_2)$ and a_1 and a_2 are the interpretations of c and R respectively.

An observation can fail to be expressible in $\mathcal{K}$ because the elements observed are not named in $\mathcal{K}$ or because the relationship among them is not the interpretation of any relation symbol in $\mathcal{K}$. If an observation O is not expressible in $\mathcal{K}$, any language $\mathcal{K}'$ which contains (i) names for the elements observed, and (ii) relation, function, and sort symbols whose interpretations are the relations, functions, and sorts observed does express O.

The following two corollaries are easily demonstrated in the same general sort of way. They are used implicitly in Section 4 where formal analogues to the theses of Whorf and Kuhn are presented.

Corollary 1 For any language $\mathcal{L}$ there are observations O which verify no $\mathcal{K}$-sentence γ.

Corollary 2 For any observation or set of observations O and any theory T_1 in $\mathcal{L}_1$, if $\mathfrak{M}_{T_1} \subsetneq \mathfrak{M}_O$, then there is a theory T_2 in an $\mathcal{L}_2$ such that $\mathfrak{M}_{T_2} \subsetneq \mathfrak{M}_O$, and neither $\mathfrak{M}_{T_2} \supseteq \mathfrak{M}_{T_1}$ nor $\mathfrak{M}_{T_1} \supseteq \mathfrak{M}_{T_2}$.

3 *Pictorial representation* Before we get to the actual explication of Whorf and Kuhn, let's pause to develop a pictorial representation of the notions just introduced. This representation will make the task of exposition much easier.

$\mathfrak{M}$, the class of all models of **ZF** whose universe is M, is denoted pictorially by a rectangle. Sentences γ in any $\mathcal{K}_i$ partition $\mathfrak{M}$ into two regions (three if we allow meaningless sentences): the models of **ZF** in which γ is true and those in which it is false, as in Figure 1.

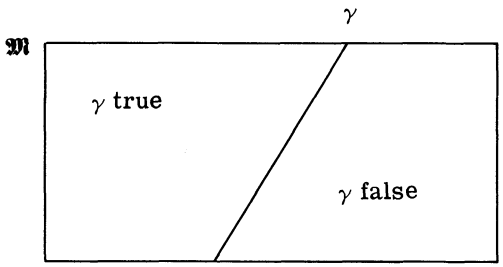

Figure 1

An observation O is represented (in Figures 2-4) by $\mathfrak{M}|_O$, the class of models in which the observation holds.

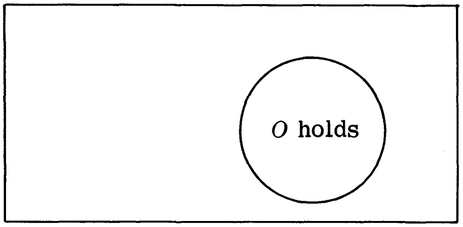

Figure 2

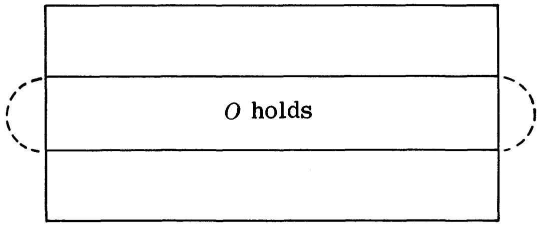

Figure 3

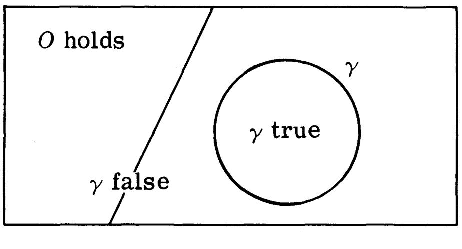

Figure 4

Consider now Figure 5, where we see that if those models of $\mathfrak{M}$ in which γ_1 and γ_2 are true are to the left of the lines marked γ_1 and γ_2, re-

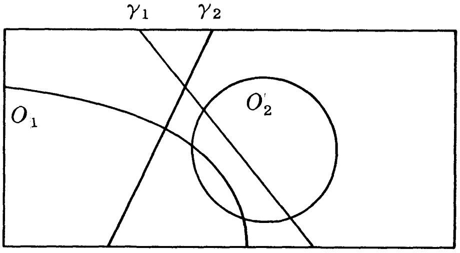

Figure 5

spectively, then γ_1 is true of O_1 (O_1 verifies γ_1) while γ_2 is true of some models of $\mathfrak{M}_{O_1}$ and false of others. Moreover γ_2 is false of O_2 (O_2 falsifies γ_2) while γ_1 is true of some models of $\mathfrak{M}_{O_2}$ and false of others. Given observations O_1 and O_2, i.e., $\mathfrak{M}_{O_1} \cap \mathfrak{M}_{O_2}$, we can conclude that γ_1 is true and γ_2 is false.

More common than (conclusive) verification and falsification of sentences and theories is their (relative) confirmation and disconfirmation.

Definition 6 O confirms γ if $P(\gamma \mid O) > P(\gamma)$.

That is, O confirms γ if the conditional probability of γ given O is greater than the probability of γ. How the probability function P is obtained, whether on frequentist, subjectivist, or logical grounds, need not concern us here.

To illustrate properties of the notion of confirmation we assume that the area of a region, $\mathfrak{M}_O$ or $\mathfrak{M}_\gamma$ say, is proportional to its probability. Thus in Figure 6, a-d, we have that O verifies γ, confirms γ, disconfirms γ, and falsifies γ, respectively.

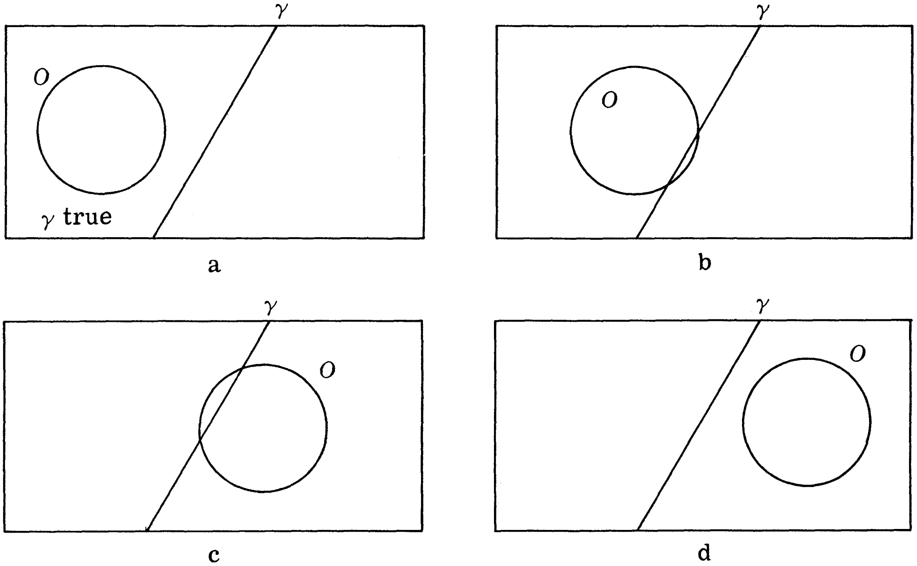

Figure 6

Finally, consider Figure 7 where an observation O_1 confirms sentences γ_1 and γ_2 and neither confirms nor disconfirms $\gamma_3(P(\gamma_3 | O_1) = P(\gamma_3))$. O_2, however, confirms γ_3. Thus if the scientist can devise an experiment whose outcome is O_2, then the new observation O_2 together with O_1 confirm γ_1, γ_2, and γ_3.

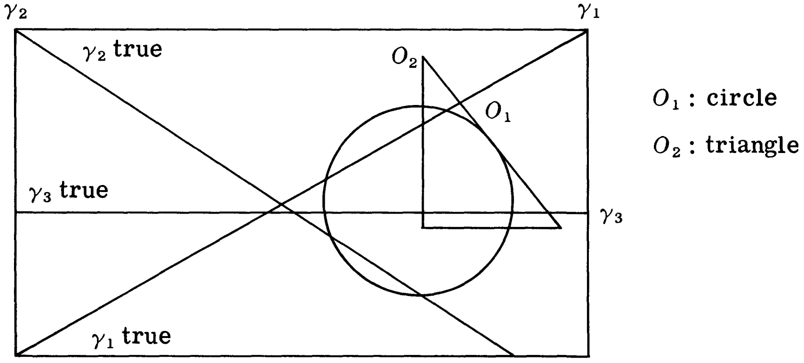

Figure 7

4 *Whorf and Kuhn* Now we are able to give a formal explication of the ideas mentioned in the introduction. Roughly stated, Whorf's thesis says that one's language "shapes" one's thoughts. Our explication of it uses an example from Randall's thesis [4].

Assume that a $\mathcal{K}$-scientist has made observations corresponding to O. Assume further that the structural properties of $\mathcal{K}$ are such that the sentences of $\mathcal{K}$ cut up the model space $\mathfrak{M}$ in a certain manner. For pictorial clarity we assume that all the sentences of $\mathcal{K}$ cut across $\mathfrak{M}$ in a horizontal direction, as in Figure 8. Here α and β are sentences of $\mathcal{K}$ already confirmed by O and γ has not yet been confirmed or disconfirmed.

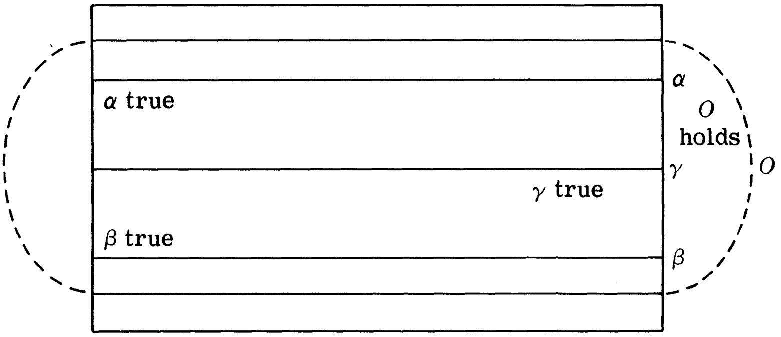

Figure 8

Now assume that the $\mathcal{K}$-scientist can perform two experiments $\mathcal{E}_1$ and $\mathcal{E}_2$, both of which divide $\mathfrak{M}$ into two observable (but not necessarily

$\mathcal{K}$-expressible) outcomes. Note that we are here considering an experiment to be a partition of $\mathfrak{M}$ into disjoint subclasses—the possible results of the experiment (two in our case). If he performs $\mathcal{E}_1$ he may be able to confirm or disconfirm γ depending on the experiment's outcome. In Figure 9 γ is verified by $\mathcal{E}_1$ where $\mathfrak{M}_{\mathcal{E}_1}$ is the observation which corresponds to the region below $\mathcal{E}_1$.

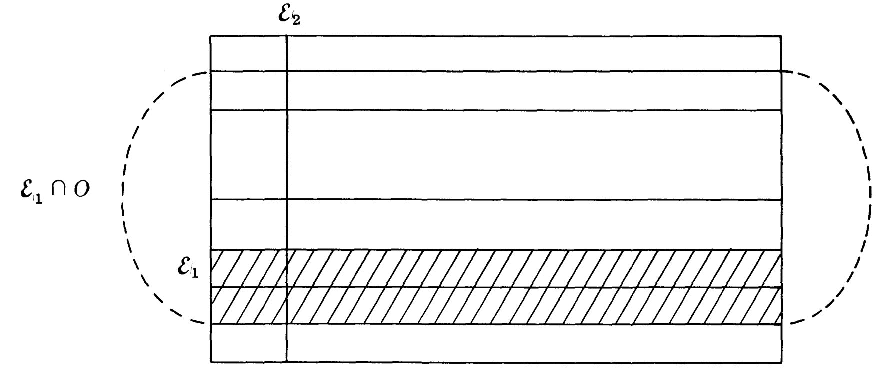

Figure 9

If, however, he performs $\mathcal{E}_2$ neither of its (observable, but not $\mathcal{K}$-expressible) outcomes confirms or disconfirms γ or any other "horizontal" $\mathcal{K}$-sentence. If, however, there were a sentence δ in $\mathcal{K}$ expressing $\mathcal{E}_2$, then if the observation $\mathfrak{M}_{\mathcal{E}_2}$ were made (in Figure 10 the region to the left of $\mathcal{E}_2$), α, β, and δ would be confirmed by $\mathfrak{M}_O \cap \mathfrak{M}_{\mathcal{E}_2}$. Moreover, the class of models $\mathfrak{M}_\alpha \cap \mathfrak{M}_\beta \cap \mathfrak{M}_\delta$ may be a less probable class than $\mathfrak{M}_{\mathcal{E}_1} \cup \mathfrak{M}_\beta$ of $\mathfrak{M}_\gamma \cap \mathfrak{M}_\beta$ (on an appropriate notion of probability). Thus in this case $\mathcal{E}_2$ is a potentially more informative experiment than $\mathcal{E}_1$. Nevertheless, without the language and concepts needed to express δ or any "vertical" sentences, $\mathcal{E}_1$ rather than $\mathcal{E}_2$ will be performed. (Of course other configurations of $\mathfrak{M}$ are possible and similar analyses can be made.)

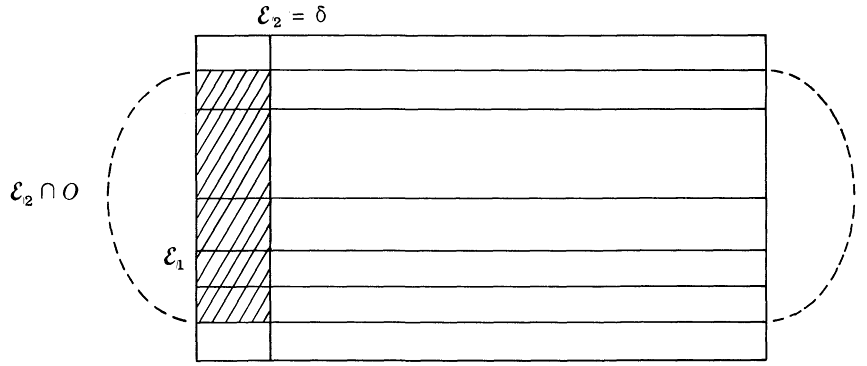

Figure 10

In this way we can see that language indeed has a channeling effect. Although the thesis that language "shapes" thought seems to be stronger, it is at least the case that language influences action, specifically in our case the choice of experiments $\mathcal{E}_1$ or $\mathcal{E}_2$.

Kuhn's thesis, roughly stated, says that science progresses in a slow cumulative manner until "enough" anomalies, exceptions, and inadequacies force scientists to search for a new paradigm or model. After a time a new theory, usually incommensurable with the first, is discovered and generally taken to be superior (despite defects of its own) to the old theory. After making such a "scientific revolution," science proceeds again in its slow, cumulative way refining the new theory.

To formalize this thesis we let $\mathfrak{M}_O$ be the intersection of the $\mathfrak{M}_{O_i}$, O_i the observations supporting a theory T_1. For simplicity let $T_1 = \gamma_1$, γ_2 and γ_3. A refinement of T_1, call it $T_1^{\ 1}$, is obtained when some previously unconfirmed statement, say γ_4, is confirmed by a new observation, O_j. (This is illustrated in Figures 11 and 12.)

O confirms γ_1, γ_2, γ_3

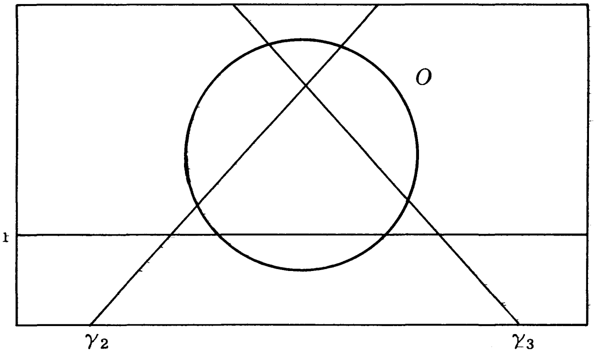

Figure 11

$O \cap O_j$ confirms γ_1, γ_2, γ_3, and γ_4

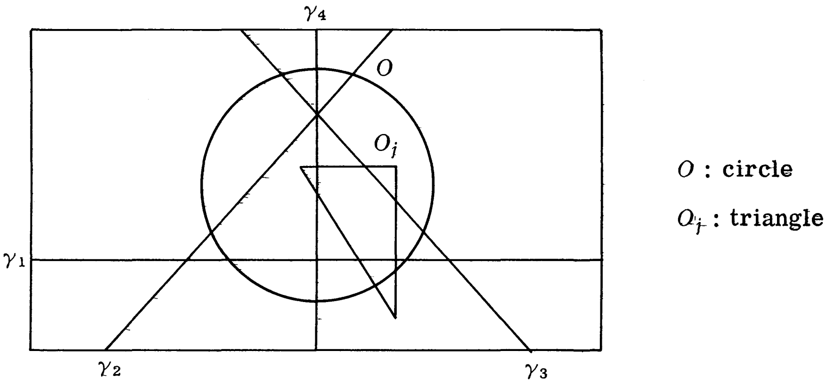

Figure 12

Thus the normal progression of science may be pictured as an increasing sequence of theories, T_1, T_1^1, T_1^2, T_1^3, . . . where T_1^{n+1} results from T_1^n by the addition of new sentences confirmed in general by new observations as above. Minor alterations of the γ_i are ignored here.

The inadequacy of T_1^n is reflected in the fact that $\cap\mathfrak{M}_{O_i} \neq \mathfrak{M}_{T_1^n}$; i.e., the class of models in which the supporting observations hold true and the class of models in which T_1^n holds true are not coextensive, do not "fit" well ("fit" depending on the underlying probability assignment). Generally $\cap\mathfrak{M}_{O_i}$ is a strictly larger class than $\mathfrak{M}_{T_1^n}$. Often, however, there will be observations O which confirm T_1^n but such that $\mathfrak{M}_O \not\supseteq \mathfrak{M}_{T_1^n}$ ($T_1^n \nrightarrow \phi(a_1, \ldots a_n)$, $\phi(a_1, \ldots a_n)$ the ε-sentence expressing O). Observations of this sort are *anomalies* as are, of course, disconfirming observations.

Because of these defects and anomalies, T_1^n is often felt to be inadequate and a new theory T_2 is sought and eventually discovered. T_2 is not a refinement or modification of T_1^n. It is expressed in a different language $\mathcal{L}_2 = \mathcal{K}_2 \cap \{\varepsilon\}$ whose semantics picks out different elements in the universe M as interpretations for its symbols than does the semantics of $\mathcal{L}_1 = \mathcal{K}_1 \cup \{\varepsilon\}$, the language in which T_1^n is expressed. ($\mathcal{K}_1$ and $\mathcal{K}_2$ have some basic symbols in common, of course.) Thus many of the sentences in T_2 are inexpressible in $\mathcal{L}_1$. In this sense T_2 is incommensurable with T_1^n (see Figures 13 and 14).

Anomalous set of observations supporting T_1^n

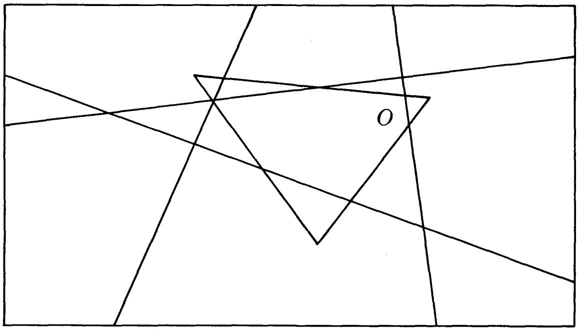

Figure 13

Same set of observations non-anomalously supporting T_2

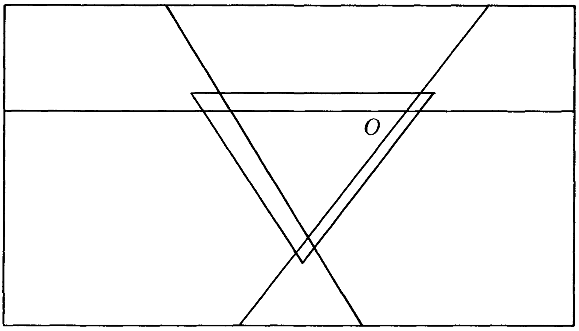

Figure 14

We note again that though the set of observations O non-anomalously confirms T_2, there are, of course, potential observations anomalous for T_2. Note also that since T_2 (and the T_2^n in general) are expressed in a different language $\mathcal{K}_2$, a different partition of the class $\mathfrak{M}$ of models is effected. Thus different observations and experiments become expressible or at least more accurately delimitable and in this way the whole direction of research changes with the adoption of the T_2^n. Observations which are expressible (or at least delimitable) in $\mathcal{K}_2$ but not in $\mathcal{K}_1$ become relevant to $\mathcal{K}_2$-scientists but not to $\mathcal{K}_1$-scientists. Compare Figures 9 and 10 illustrating Whorf's hypothesis.

5 *Further application and comments* What does all this say about the theses of Whorf and Kuhn? Our formalism, being mathematics, of course provides no empirical support for these empirical theories. It does, however, provide us with a formal analogue to them. Given certain scientific and philosophic assumptions concerning the nature of language and reality, the formalism actually proves these theories. In any case it clarifies them and suggests further questions and extensions. We mention a few here and will expand on them in a future paper.

Firstly, tools from model theory enable us to construct models $(M, \ldots)$ having special properties. For example, we have the following:

Theorem 2 *Given any model* $(M, \ldots)$ *of a scientific theory* T *there are*: (i) *an elementarily equivalent model of* T *which is saturated and* (ii) *an elementarily equivalent model of* T *which omits types locally omitted by* T.

Proof: Immediate from our definition of T and standard theorems on saturated models and omitting types.

Hence, in some sense, both very dense worlds and very sparse worlds are compatible with T. Hopefully properties of these special models (saturated, generic, etc.) may say something about the tenability of certain metascientific theses.

Model-theoretic tools may also be used to explore $\mathcal{K}$-ineffability and $\mathcal{K}$-randomness. An observation O is $\mathcal{K}$-ineffable if there is no $\mathcal{K}$-sentence γ such that γ expresses O. An observation O is $\mathcal{K}$-random iff there is no $\mathcal{K}$-sentence γ such that γ delimits O. Relativizing these notions to a particular language is at least one way to give us a handle on them. It's clear that O is $\mathcal{K}$-ineffable iff $\overline{O}$, the complementary observation, is $\mathcal{K}$-random.

A third possible application concerns an explication using our apparatus of the common idea that creativity results from the juxtapositioning of two disparate notions (models, theories).

Finally, modal operators could be easily incorporated into this formalism by dealing with model systems—indexed collections of models of **ZF**—instead of with models of **ZF** themselves. See our paper for details [3].

REFERENCES

[1] Chang, C. C., and H. J. Keisler, *Model Theory*, North-Holland Publishing Co., Amsterdam, 1973.

[2] Kuhn, T. S., *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago, 1962.

[3] Paulos, J. A., "A model-theoretic semantics for modal logic," *Notre Dame Journal of Formal Logic*, vol. XVII (1976), pp. 465-468.

[4] Randall, D. L., *Formal Methods in the Foundations of Science*, Ph.D. thesis, California Institute of Technology, Pasadena, 1970.

[5] Whorf, B. L., *Language, Thought, and Reality*, Massachusetts Institute of Technology Press, Cambridge, 1965.

Temple University
Philadelphia, Pennsylvania