Logique & Analyse **221** (2013), 5–23

MODALITY IN MATHEMATICS

WILFRID HODGES

> As soon as these questions were squarely faced, a wide range of new phenomena were discovered, including quite simple ones that had passed unnoticed.
> *Noam Chomsky, 'Knowledge of Language', p. 7*

In this paper I argue that there are some quite basic questions that we can't yet answer, about how we write and read mathematics. The questions themselves are straightforward enough to state, provided that we don't allow ourselves to be distracted by irrelevances. In section §2 below I formulate them in terms of the use of modal notions in mathematical writing, but I think it will become clear that these formulations are special cases of much larger questions about how we use language to communicate information. How far the answers depend on general facts about language, and how far on peculiar features of mathematics, is one of the things we don't yet know.

Readers who want background information on English modals can find a readable treatment in Palmer [7].

I am in debt to various audiences and correspondents, including a careful referee who made penetrating observations and saved me from some embarrassing slips. But let me particularly thank the organisers and contributors of the Amsterdam meeting on 'Practice-based philosophy of Logic and Mathematics' in August and September 2009, and especially Catarina Dutilh who designed and led the whole enterprise.

1. *The corpus*

On the face of it, mathematics has no modal content. Mathematicians are pleased to know that

(1) Every finite field is commutative.

or that

$$1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots = \frac{\pi}{4}. \tag{2}$$

The fact that these statements are *necessarily* true might attract the attention of a philosopher of mathematics, and some mathematicians dream about such things in idle moments. But adding 'Necessarily' to either (1) or (2) would introduce nothing of any mathematical significance.

Three years ago I made a list of all the modal words in the first hundred pages of Birkhoff and Mac Lane *A Survey of Modern Algebra* [3]. Since then I have collected similar data from some other mathematical textbooks. I reckoned that since these modal words added nothing to the mathematical content, they must be there for other purposes. My calculation was that it would be interesting to find out what these purposes were, and that the methodology for answering this question might be interesting too. A first paper was submitted for the Proceedings of a conference, but the editors seem to have gone into hiding and I no longer expect to see those Proceedings published.

Three years and several conversations and conferences down the line, the issues of methodology seem to me a lot subtler than I appreciated at first. Also it became clear that the material I had was too disparate. For example any textbook is a sort of conversation between its author and its readers, and modal expressions (particularly deontic ones) often play a role in this kind of conversation. Thus the author urges the readers to do or refrain from doing certain things:

(3) ([5] p. 9) As an exercise, the reader may prove the following result.

(4) (From a website on mathematical economics:) This exercise should not be attempted until the above exercises are fully understood.

Or the author invites the reader to accept that the author's discussion is appropriate:

(5) ([5] p. 43) The reader might well ask how far we must go ...

(6) ([5] p. 4) The seemingly pedantic distinctions made here are really quite necessary.

These conversational modals have nothing to do with the mathematical content.

So we now cut down to modal expressions that occur inside definitions, axioms, theorems, lemmas, corollaries and exercises. We restrict to labelled and numbered instances of these contexts, apart from a few examples that could have been labelled and numbered but weren't. As a result, we exclude nearly all the modalities that play a conversational role. A very few deontic modalities of this kind still get through, like

(7) ([1] p. 87, Definition) This notation must be examined carefully to understand the argument.

See also (3) above.

For the rest of this paper we work with what amounts to a small corpus of texts. It consists of the first and last hundred pages of Birkhoff and Mac Lane *A Survey of Modern Algebra* [3], the first hundred pages of Baldwin *Categoricity* [1] and the first hundred pages of Hocking and Young *Topology* [5]. Restricting ourselves to definitions, axioms etc. as above, the numbers of modal items found in each of these sources were as follows:

Baldwin	24 items
Birkhoff and Mac Lane, first 100 pages	42 items
Birkhoff and Mac Lane, last 100 pages	42 items
Hocking and Young	32 items.

What expressions to count as modal? I included all the English modal auxiliaries

(8) can(not), may, might, must, need(ed), will (when not a simple future marker), would.

Of these, 'would' occurred just once. Our sources had no examples of 'have (got) to', 'ought', 'shall' or 'should'.

I included non-auxiliaries that are usually reckoned to express modal concepts:

(9) necessar(il)y, out of the question, permissible, permit, possible, require.

I used my judgement to exclude a few other items such as 'impose' and 'sufficient'; I counted 'can guarantee' as a single modal item.

I decided to exclude 'reducible' for two reasons. First, there are so many occurrences of 'reducible' in the first hundred pages of Birkhoff and Mac Lane that I felt they would swamp the survey. Second, 'reducible' is definable without using any modal words at all. A polynomial p is reducible

over a field F if p is the product of two polynomials over F which both have lower degree than p. It became clear that students who were asked to do exercises about reducibility were expected to know and use this definition. So its occurrence in an exercise was no evidence of modal content.

Similar reasons led me to exclude 'metrizable' and some other '-able' or '-ible' words. But I did include one occurrence of 'expressible', because it is not a mathematical term with a non-modal definition. It's simply a stylistic variant for 'can be expressed'.

Ironically the definition of 'reducible' in Birkhoff and Mac Lane did figure in the list, because it is not the non-modal one just given. They write

(10) ([3] p. 71, Definition) A polynomial form is called "reducible" over a field F, if it can be factored into polynomials of lower degree ...

We will see below that 'can be factored' is a member of one of the largest families of modal notions in the corpus.

2. *Use of language, the problems*

The appearance of non-conversational modal words in *exercises* strikes me as particularly paradoxical. The student has to be able to understand the exercise in order to do it. But the mathematics that the student is required to do is not modal at all. So the student has to be able to translate away the modalities into something non-modal. Reflecting on this, we can formulate three problems, which I call the *translation* problem, the *reachability* problem and the *preference* problem.

> The translation problem. Given a mathematical sentence containing a modal word, find a non-modal translation of it (if there is one).

In this paper I ignore some background questions that might be raised. For example it will become clear that the translation needs to be at the level of sentences rather than single words. Maybe for similar reasons we should be looking for translations at the level of paragraphs. Also there is a question whether the modal expressions add hints or suggestions rather than explicit statement.

It seems fairly straightforward to get at least *prima facie* answers to the translation problem. One approach is to translate the text into Zermelo-Fraenkel set theory — bearing in mind that the language of Zermelo-Fraenkel set theory contains no expressions with modal meanings. One audience that I spoke to were worried about whether mathematics in general can be

formalised in Zermelo-Fraenkel set theory. But that's irrelevant here; the relevant point is that the mathematics in the chosen textbooks is all quite easy and uncontroversial to formalise. At worst there are some questions about which of the classes mentioned can be proper classes. But these questions do have workable conventional answers, and I don't see any link between the questions and the issue of modality.

> The reachability problem. Given a modal text X and its non-modal translation Y, how would the student with the expected knowledge of English and mathematics be able to reach Y from X?

The translations from modal to non-modal should be justifiable in terms of the normal usage of those modal terms in English. Of course we can allow that students learn some peculiarities of mathematical language; but if the best we can say is 'That's how mathematicians express themselves', we should recognise that we have given up the attempt to find a serious explanation.

Once when I spoke on this topic to a group of logicians, philosophers of mathematics and historians of mathematics, at least two people in the group (both philosophers if I remember right) startled me by assuming that I was criticising Birkhoff and Mac Lane. I didn't probe it at the time, but I suppose the reasoning was that if Birkhoff and Mac Lane meant something non-modal but used modal language to express it, then they hadn't succeeded in saying what they meant, or at least they had given their readers extra work to discover what they meant. The difficulty with that view is that meanings never pass directly from the author's mind to the reader's, any more than the appearances of objects pass directly out of the objects and into the mind of the viewer. In both cases there is a vast amount of unconscious computation involved in bringing a thing into our minds. The text of Birkhoff and Mac Lane has established itself as one of the classic textbooks of algebra. If they convey their content in ways that surprise us, the chances are they generally know what they are doing, and we might even learn something from them about how humans understand mathematical writing.

> The preference problem. Given that we have a modal version X and a non-modal version Y, what is the case for writing X rather than Y?

How can we answer the preference problem? The first step, naturally, is to see what it would do to the text if we put the non-modal translation in place of the existing modal version. In context, is one of them clearly better than the other, and if so why?

This is a well-established method in other fields. My thinking about it has been very much influenced by an example in Nicholas Cook *A Guide to Musical Analysis* [4] p. 343ff. He analyses piece 3 from Schoenberg's *6 Kleine Klavierstücke* Op. 19, a highly original piece in its time, by rewriting it in various ways — for example in the style of Brahms — and asking what has gone missing in the rewrites. The method is marvellously illuminating.

When I tried this replace-and-compare method at the Amsterdam conference, we ran into a difficulty. Given two versions of the same exercise, mathematicians can usually reach some consensus about which is better. But finding the reasons is another matter altogether. I had analysed some examples of 'can be embedded'. At least two people in the audience — this time one was a computer scientist and one was a mathematician — claimed that the difference between the modal and the non-modal version was that the modal version steered the reader in the direction of an effective embedding. I couldn't see this. So we were in a position of swapping rival introspections, and this is a bad place to be if we want to reach objective conclusions.

Faced with this difficulty, there are two things one should try (not necessarily in the following order). The first is to go back to the translation and reachability problems to check we had the right answers there. If the students are supposed to read something into the modal language, how would they get there from their knowledge of English and mathematics?

The second is to look at a wider class of examples. (It took Cook five different rewrites to extract what he needed from Schoenberg's piece.) What would happen to the intuition about effective methods if the modal term was in a definition or a theorem, not in an exercise? What if the mathematical material was not effective anyway? It was this that led me to extend the corpus to include the last hundred pages of Birkhoff and Mac Lane and the first hundred pages of Baldwin, since both of these texts contain some non-effective material.

3. *Two cases: 'necessary' and 'may'*

The corpus contained several families of closely similar examples, and a few outliers. In this section we look at two of the outliers.

In mathematical contexts, to say that the truth of p is necessary for the truth of q is equivalent to saying 'If q then p'. This accounted for 8 items in Hocking and Young (and none in the other sources). So we have a quick answer to the translation problem. But we still have the reachability and preference problems to solve.

We can give at least a partial answer to the preference problem by considering an example and two of its non-modal translations:

(11) ([5] p. 12) A necessary and sufficient condition that the transformation $f : S \to T$ of the space S into the space T be continuous is that if x is a point of S, and V is an open subset of T containing $f(x)$, then there is an open set U in S containing x and such that $f(U)$ lies in V.

(12) The transformation $f : S \to T$ of the space S into the space T is continuous if and only if, if x is a point of S, and V is an open subset of T containing $f(x)$, then there is an open set U in S containing x and such that $f(U)$ lies in V.

(13) The following are equivalent:

(a) The transformation $f : S \to T$ of the space S into the space T is continuous.

(b) If x is a point of S, and V is an open subset of T containing $f(x)$, then there is an open set U in S containing x and such that $f(U)$ lies in V.

Version (12) reads badly; something needs to be done about the 'if, if'. Note how this comes about. The original (11) wrapped both conditions into 'that ...' clauses, which prevented interference between the syntax of the clauses and the text that surrounds them.

Version (13) reads better, but it uses up more space, and some writers will dislike the accumulation of extra symbols '(a)', '(b)'. A further point is that with the original (11) there are straightforward names for the two directions of the proof: 'sufficiency' and 'necessity'. For (12) we would need to describe the directions as 'left to right' and 'right to left'; for (13) one could say '(a) $\Rightarrow$ (b)' and '(b) $\Rightarrow$ (a)'. These are both less intrinsic, and the third version introduces yet more symbolism.

All in all, none of these points are decisive reasons for using or avoiding the term 'necessary', and most writers will let their style guide them. My own style would usually be as in (13). But having more options is always welcome.

The reachability question is harder, but again I can conjecture a partial answer. When we ask for the conditions for something q to hold, usually we are talking about things that will cause q, or environments in which the normal causes of q are free to act. Thus from a biological journal:

(14) What are the necessary evolutionary selection conditions for the development of communication?

We can't rephrase this by asking 'For what evolutionary selection conditions is the development of communication a sufficient condition?', because it has the causation the wrong way round. But in mathematics there are no causes. So when a mathematician talks of 'necessary conditions', the direction of causation drops out of the picture, and only the 'if ... then' survives.

This answer raises a few further questions. For example mathematicians certainly do say things like 'The reason for q is p'; so why doesn't (11) above carry an implication that (a) of (13) is the *reason* for (b)? I think myself that in mathematics, statements invoking 'reasons' are always epistemic and always refer to a particular way of building up a topic, not to the mathematical contents themselves. But there is no space to develop this here.

If this answer to the reachability question is correct, it might apply in other cases too. Namely, it might happen that the modal statement can be spelt out more fully, in such a way that it contains a clause which is vacuous for mathematical objects. So this clause drops out as irrelevant, and what remains is the non-modal translation. We want a name for this mechanism. Let us call it *masking*. The modal content is masked by the mathematical context.

Next consider

(15) ([3] p. 366, Exercise) Show by examples that there may exist subgroups of any given finite order in a denumerable group G.

As I read it, the student is being asked to show

(16) ($\forall$ positive integer n)($\exists$ denumerable group G)($\exists$ group H)(H has order n and H is a subgroup of G).

The order of the quantifiers in (16) is not the same as in (15): the subgroups come first in (15) and last in (16). How was the student to know this?

This reversal of quantifiers with 'may' and 'any' is quite common in English. There are several millions of examples on google. For example (after one slight adjustment):

(17) An intestinal calculus may be found in any portion of the colon.

(18) There is a small possibility that nut traces may be found in any of our items.

The first example would never be read as saying that one and the same intestinal calculus can be simultaneously in all portions of the colon. The chief culprit in switching the order of the quantifiers is certainly the word 'any'. But the reversal is less easy if we remove the modal 'may'.

(19) An intestinal calculus is found in any portion of the colon.

We needn't examine why 'may' works this way; for our purposes it's enough to note that it does work this way in ordinary English.

Still there is something unexplained. The use of 'may' with 'any' signals that the quantifiers need reversing. But when they are reversed in ordinary English examples, the 'may' stays, possibly changed to 'can':

(20) Every portion of the colon can contain an intestinal calculus.

Not 'does contain', fortunately! So to answer the reachability question, we need to know what happened to this surviving 'may' or 'can'. A straight swap of quantifier order would turn (15) into

(21) Show by examples that for every positive integer n there can exist a denumerable group with a subgroup of order n.

I think masking comes into play again here. It's true that there can exist a denumerable group with a subgroup of order 17. But the reader knows that the notion of possibility is irrelevant here. Either there is such a denumerable group or there isn't, and if there isn't one then there couldn't be one. So the difference between 'can exist' and 'exists' vanishes.

What can we say about the preference problem? What advantage does (15) have over (16)? Well, for a start it's less cluttered with symbols; but I could have written (16) in plainer English. Probably the main merit of (15) is that it brings 'there may exist subgroups' to the beginning of the embedded clause, correctly suggesting that the topic is the existence of subgroups. One of the sadder consequences of a training in logic is that it teaches us to ignore topic and focus. (On these notions see Lambrecht [6].)

4. *'Can be'*

By far the commonest modal word in our corpus is 'can'. There are 83 occurrences. They are overwhelmingly affirmative; the exceptions are 6 occurrences of 'cannot'. There are also 15 occurrences of 'may', none of them with 'not', and probably all of them are stylistic variants of 'can'.

Within the 77 affirmative occurrences of 'can', all but 7 are passives: 'can be'. Several patterns are particularly common:

(i)	can be expressed (written, represented, rearranged, well-ordered etc.)	26
(ii)	can be embedded (mapped, extended etc.)	20
(iii)	can be shown (generalised, assumed etc.)	6
(iv)	can be found (chosen etc.)	5
(v)	can be generated	4

Rarer are 'can be defined', 'can be used', 'can be studied' and a few others.

There is a curious point of syntax. We can't put

(22) Smith has the strength to kill Jones.

into the passive as

(23) Jones has the strength to be killed.

But this way of putting into the passive does work with 'can':

(24) Smith can kill Jones.
Jones can be killed.

The point to take home is that for example 'can be written' is the straight passive form from 'can write'. It shouldn't be read as an active form followed by an adjective, as in

(25) ([5] p. 97, Exercise) Show that each X_n can be infinite and compact.

This was the only example of active 'can' + 'be' + adjective in the survey.

4.1. *Effectiveness*

When we look at the active forms of the verbs that appear in our corpus within the context 'can be . . . ', some of them turn out to be completely literal. Thus

(26) ([1] p. 93, Exercise) Show that the restriction on the cardinality can be replaced by assuming . . .
= Show that you can replace the restriction on the cardinality by the assumption . . . and prove the theorem with this replacement.

(27) ([3] p. 26, Exercise) Show by induction that Theorem 17 can be generalized to n congruences.
= Show that you can generalise Theorem 17 to cover any finite number of congruences by using induction.

I hope I translated (27) correctly. If the authors meant 'Use induction to show that for any finite n you can generalise Theorem 17 to the case of n congruences', then they are using the familiar mathematical 'can in principle'; there are only a finite number of finite numbers that you can hope to name in a lifetime. A similar 'can in principle' is:

(28) ([3] p. 416, Theorem) Every Gaussian integer can be expressed as a product of prime Gaussian integers.
= You can express every Gaussian integer as a product of prime Gaussian integers.

Further down the line are some 'can be' statements where no human being could even in principle do the thing claimed. A couple of examples:

(29) ([5] p. 25, Theorem) Every set can be well − ordered.
= For any set x, we can well-order x.

(30) ([1] p. 29, Exercise) Show that any model can be written as a continuous increasing chain of submodels.

The models in (30) are of any transfinite cardinality and there is no assumption that they are given in any constructive form.

At the far end of this scale are statements about embedding, or about extending mappings. Thus:

(31) ([3] p. 43, Theorem) Any integral domain can be embedded in a field.
= We can embed any integral domain in a field.

(32) ([5] p. 64, Theorem) Any mapping $f : A \to Y$ can be extended to all of X.
= We can extend any mapping $f : A \to Y$ to all of X.

I know how to embed a 5p piece in a christmas pudding, or my fist in somebody's mouth, or even a computer program in a historical article. But integral domains and fields are eternal objects. Either the integral domain is already embedded in the field, or it isn't; either way, how could any action of mine make any difference?

As mentioned earlier, it was suggested at the Amsterdam meeting that the use of 'can be embedded' was a hint to the reader to look for a constructive interpretation. Thus the student is invited to show that A can be embedded in B by producing a concrete description of an embedding of A into B.

The examples of 'can be embedded' in the corpus provide no support at all for this suggestion. There is only one example of 'can be embedded' in an exercise:

(33) ([3] p. 43, Exercise) Can the system J_6 of integers modulo 6 be embedded in a field?

Looking for a concrete embedding would if anything be a distraction here. The student should be aiming to find equations or inequations that hold in J_6 but not in any field. The other instances of 'can be embedded' or similar phrases reinforce the impression that effective content is completely irrelevant. For example

(34) ([1] p. 80, Definition) ... M_1 and M_2 can be disjointly amalgamated over M.

This is from a definition in the middle of some highly nonconstructive infinitary mathematics. Restricting the definition to effectively given maps would skew everything. Other examples tell the same story. So henceforth I ignore the idea that using 'can be embedded' has anything to do with effective content, at least in the texts we are examining.

4.2. *Nominalisations, causatives and thematic roles*

We noticed earlier that the meaning of 'embed' in mathematical contexts is not got by applying the everyday uses of 'embed' to mathematical objects. The position is actually a bit odder than that. There would be no harm if the mathematicians gave their own definition of 'embed'. But they don't. Most textbooks — my own included, to my surprise — never define 'embed'. Instead they define the verbal noun 'embedding'. Birkhoff and Mac Lane are a little unusual in defining the adjective 'embedded (in)' ([3] p. 43). But as we noted earlier, this is still off track. The phrase 'can be embedded' is a passive, not 'can be' plus the adjective 'embedded'.

So it seems the student reading Birkhoff and Mac Lane, or any of a thousand other mathematical textbooks, has to discover for herself what 'embed' means on the basis of the meanings of other forms, usually 'embedding' but sometimes 'embedded'. The usual context of 'embed' in English takes the form

(35) AGENT embeds OBJECT in LOCATION.

So the verb is about an action performed by an agent. But in mathematics an embedding is a set-theoretic object; what action does it involve or apply?

It turns out that there is a pattern here. Mathematicians define a range of nouns and then proceed to use related verbs as if the sense of the noun made the sense of the verb clear. Thus:

(36)
an embedding of A into B
a mapping of A to B
a piercing of (surface) A by (line) B
a splitting of (group) A into B, C
a splitting of A by B

All the nouns in sans serif are perceived as verbal nouns from action verbs 'embed', 'map', 'pierce', 'split'. Mathematicians define only the nouns, never the verbs. Generally they use the verbs in 'can be' form. I took the following from mathematical texts on the internet:

(37)
The interior of any simple closed curve can be mapped in an angle-preserving way to the open unit disk.
Each 2-sphere in each 3-manifold can be pierced by a tame arc.
Every division k-algebra D can be split by a finite Galois extension K/k.
Points on C can be injected into a proper linear subspace.
The triangles {11,3,6} and {11,6,1} can be retracted into the path (11,3,6,1).

In my Amsterdam talk I concluded that 'the mathematical usage should be explained in terms of some general phenomena with action verbs and their nominalisations'.

I no longer believe this. I think there is a broader pattern which includes the examples above as a rather misleading special case. Looking back to the list (i)–(v) at the beginning of §4 above, we can see that the items in (i), (ii) and (v) all have the following description. Given some mathematical objects $a_1, \ldots, a_n$, we define a type of structure that consists of these objects together with some other objects related to them. In the case of embedding, we have two structures a_1, a_2, and the other object is an embedding from a_1 to a_2. In the case of well-ordering, we have one set a and the other object is a bijection between a and an ordinal (or equivalently, a well-ordering relation on a).

These examples illustrate a general pattern, which runs as follows. A notion is defined:

(38) x is a boojum of $a_1, \ldots, a_n$.

To express

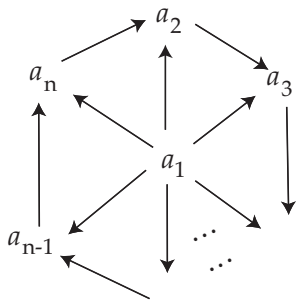
(39) There is a boojum of $a_1, \ldots, a_n$.

we first find a verb that can be understood as 'make a boojum', for example the causative form 'boojumise'. For reasons to be explained below, I call this verb a *pseudo-causative*. Then we write

(40) $a_1, \ldots, a_n$ can be boojumised.

You can check this. Think of some kind of configuration of mathematical objects and give it a name. For example a commutative diagram

(41)



might be called an 'eq41'. Then imagine explaining this definition to your students, and asking them to show that there are maps which together with certain objects $b_1, \ldots, b_n$ form an eq41. How do you say it? Try

(42) Show that $b_1, \ldots, b_n$ can be eq41ed.

Here you invent a pseudo-causative 'to eq41'. (Or you might prefer 'to eq41ify'. Styles differ.) We do such things all the time.

Now the special case illustrated by the nouns in (36) above is the case where the configuration can be described by a noun that already means the result of some action named by a verb. An injection is what results if you inject; and so on. So in these cases the causative verb was already available — in fact the noun was derived from it. But in the larger picture that was a lucky accident.

This picture needs a few refinements. Let me mention thematic roles. In the example above, we took 'eq41ing' as something done to the whole array $a_1, \ldots, a_n$. But language allows us to assign roles. We can make a_1 the OBJECT and $a_2, \ldots, a_n$ the LOCATION; in English we would do this by saying that (41) is 'an eq41 of a_1 inside $a_2, \ldots, a_n$', and the corresponding version of (42) would be

(43) Show that b_1 can be eq41ed inside (or into) $b_2, \ldots, b_n$.

But equally we can make $a_2, \ldots, a_n$ the OBJECT and a_1 the INSTRUMENT. Maybe we would use a different noun 'hub' rather than 'eq41' in this case, and (42) might become

(44) Show that $b_2, \ldots, b_n$ can be hubbed by b_1.

Note that this applies equally well to 'embedded in'; there could be mathematical cultures where instead of 'a_1 is embedded in a_2' they say 'a_2 is wrapped around a_1'.

The choice of thematic role is a matter of how we choose to throw language at the facts we are describing. It's convenient to assign roles in the examples above, because it gives us more ways of saying things, calling attention to different arrangements. One reason for preferring 'a_1 is embedded in a_2' to 'a_2 is wrapped around a_1' is that the topic is usually a_1 rather than a_2, and topics go better as subject than they do as indirect object.

Note that there is no obvious role for an AGENT in the kind of mathematical situation we are discussing. This is why the verb is always in the passive, and also why I call it a pseudo-causative. A real causative expresses that an agent causes something to happen; but here we have no agent. In the mathematical usage of 'embed', we don't say that such-and-such a person or thing 'embeds' a in b, or that a is 'embedded' in b 'by' a person c. It's true we might say 'a is embedded in b by f', where f is the embedding function; but here f is in the role of INSTRUMENT, not AGENT. It's also true that we might say

(45) Compact riemannian n-manifolds were first embedded in euclidean space by John Nash in 1955.

Here John Nash is certainly in the role of AGENT, but this example is something of a play on language.

4.3. *Translation, reachability, preference*

Now we come back to our three problems, for the case 'can be X-ed' where 'to X' is a pseudo-causative. I take 'embed' as a typical pseudo-causative.

We start with the translation problem. The statement

(46) a can be embedded in b.

has the non-modal equivalent

(47) There is an embedding of a in b.

As a refinement, note that if we quantify universally over a, then the 'there is' formulation becomes ugly, for example

(48) For any subgroup H of G there is a generator set of H with cardinality at most n.

English speakers tend to switch to an idiom with 'has':

(49) Any subgroup H of G has a generator set of cardinality at most n.

This usage has nothing specifically to do with mathematics. Compare:

(50) For each of our clients there is a sponsor (of that client).
Each of our clients has a sponsor.

Now the reachability problem asks how the reader knows that

(51) Every integral domain can be embedded into a field.

is a way of saying

(52) Every integral domain has an embedding into a field.

I tentatively suggest that 'can' in (51) is a dynamic 'can', so that the sentence as a whole could be paraphrased as

(53) For every integral domain H there is a course of action open to us, such that after it has been taken, the integral domain H has an embedding into a field.

But clearly no action of ours will have the slightest effect on whether there is an embedding of H into a field. So the mask applies and the clause about action is cancelled from the meaning.

That may be part of the solution, but it is certainly not the whole solution. We can see this by trying to apply the same formula to the sentence

(54) 1729 can be composed in two different ways as the sum of two squares.

Why is there not the slightest temptation to read this as follows?

(55) There are two different courses of action open to us, such that after either of them has been taken, 1729 is the sum of two squares.

So in the last resort I have to leave the reachability problem as open, though I don't know any reason why it should be unsolvable in principle.

Finally we turn to the preference problem. This requires us to find the reasons why we would use or expect (51) rather than (52). From earlier examples we know (a) that it's unsafe to reason from a single pair of texts, and (b) that the modal version can be preferable for quite syntactic reasons involving possible English sentence structures.

In the case of (51) and (52) I don't feel any strong pull in favour of the first and away from the second. But it seems to be a fact that in the mathematical literature forms like (51) are used overwhelmingly more than forms like (52); I think this is a safe generalisation over pseudo-causatives of all sorts. This fact (assuming it is one) needs an explanation, but it is also an obstacle to finding one. The mere fact that (51) comes to us more readily than (52) has the effect that if we see (52), we wonder why the author wrote that rather than (51). For example, was he or she trying to suggest to us that there is a default or canonical embedding (as of course there is in this case)? Almost certainly we wouldn't have smelled this suggestion in (52) if it weren't for the fact that we expect (51). So implied suggestions of this kind may be the result of a general preference for the modal form, not the cause of it.

The difficulty here is that our intuitions are the ones we have now, not the ones that people had at some past time before the conventions of modern mathematical writing were fixed. That's a historical question. You would be amazed at the number of people who think they can settle historical questions by introspection.

5. *Drawing the threads together*

We began from the fact that a sample of mathematical textbooks contained quite a few modal expressions mixed in with the pure mathematical content. We posed three problems about this modal content. I claim that until we have satisfactory answers to these three problems, the mixture of modality and mathematics is paradoxical and demands an explanation.

The translation problem was to extract the non-modal content from the modal expression. At least for the examples we looked at, this seems to be relatively unproblematic. If we can formalise the textbook content in non-modal formal languages — and this we can certainly do — then we can do the easier task of translating modal mathematical English into non-modal mathematical English. One or two of our examples illustrated the fact that a textbook reader needs a practical grasp of English quantifier scopes; she has to follow rules that most of us have never been consciously aware of.

The reachability problem was to explain how the reader can see that the non-modal translation is correct in context. We suggested answers in some particular cases. The answers all rested on the same mechanism, namely that the reader uses her knowledge of the irrelevance of modality to mathematics, so as to 'mask' the modal content of the text. If this mechanism really does provide a general answer to the reachability problem, then some further questions arise. What if the reader didn't know, or doesn't believe, that the facts of basic algebra are non-modal?

One doesn't have to look far in the secondary literature to find a variety of claims to the effect that some mathematician meant something that can only be expressed with words like 'necessary' or 'possible'. (Such things are said about mathematicians ranging from Euclid to Tarski.) Even if all these claims are wrong, there are still a number of people out there who apparently don't share a basic presupposition of the masking mechanism. That raises the possibility of testing the mechanism empirically. If a student holds the view that mathematics is about can's and must's, will this student have greater difficulty following a mathematical text?

And thirdly there comes the preference problem. We saw some very partial answers, and some difficulties in the way of finding convincing general answers. My own guess is that there are several quite different kinds of reason why modal formulations sometimes sit better in a textbook than their non-modal translations. These reasons probably fall into a small number of groups, which in principle we could catalogue. But digging them out is likely to involve a range of expertise, calling on both mathematics and linguistics. In my Amsterdam talk I quoted recent evidence from brain research. That's missing from this paper, because its relevance rested on my earlier guesses about the role of action sentences, which I no longer believe. But the next half century of brain research is going to give us a flood of evidence about

our use of language, and it's bound to illuminate some of the questions discussed in this paper.

Finally I draw a moral about the history of mathematics. The examples of modal language in familiar modern textbooks should make us cautious in drawing inferences from the presence of modal terms in earlier authors. Take this, from the comments of Simplicius (6th century AD) on the first postulate in Book 1 of Euclid's *Elements*. We have it only in a medieval Arabic translation by Al-Nayrizī ([2] p. 18):

(56) It would be foolhardy to postulate that a straight line can be extended (*'an yukraja*) from Aries to Libra.

What is Simplicius saying is foolhardy? Does this 'can be extended' faithfully express Euclid's intentions? (Euclid's Greek doesn't have a complete sentence here.) It was obvious to Simplicius that Euclid wasn't talking about actions that you or I can take; was it also obvious to Euclid? If this paper has made it a little harder to give facile answers to questions like these, I'll be happy.

Herons Brook
Sticklepath
Devon EX20 2PY
United Kingdom
http://wilfridhodges.co.uk
E-mail: wilfrid.hodges@btinternet.com

REFERENCES

[1] John T. Baldwin, *Categoricity*, American Mathematical Society, Providence RI 2009.
[2] Rasmus Olsen Besthorn and J. L. Heiberg, *Codex Leidensis 339, 1: Euclidis Elementa Ex Interpretatione Al-Hadschdschadschii Cum Commentariis Al-Narizii*, Hauniae, Gyldendal 1893.
[3] Garrett Birkhoff and Saunders Mac Lane, *A Survey of Modern Algebra*, Macmillan, New York 1953.
[4] Nicholas Cook, *A Guide to Musical Analysis*, Dent, London 1987.
[5] John G. Hocking and Gail S. Young, *Topology*, Dover, New York 1961.
[6] Knud Lambrecht, *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*, Cambridge University Press, Cambridge 1996.
[7] Frank Palmer, *Modality and the English Modals*, Longman, London 1990.