Synthese (2022) 200:24
https://doi.org/10.1007/s11229-022-03559-4

ORIGINAL RESEARCH

A fixed-point problem for theories of meaning

Niklas Dahl[1]

Received: 22 December 2020 / Accepted: 14 December 2021 / Published online: 21 February 2022

Abstract
In this paper I argue that it's impossible for there to be a single universal theory of meaning for a language. First, I will consider some minimal expressiveness requirements a language must meet to be able to express semantic claims. Then I will argue that in order to have a single unified theory of meaning, these expressiveness requirements must be satisfied by a language which the semantic theory itself applies to. That is, we would need a language which can express its own meaning. It has been well-known since Tarski that theories of meaning whose central notion is truth can't be expressed in a language which they apply to. Here, I develop Quine's formulation of the Liar Paradox in grammatical terms and use this to extend Tarski's result to all theories of meaning. This general version of the paradox can be formalised as a special case of the Lawvere Fixed-Point Theorem applied to a categorial grammar. Taken together with the initial arguments, I infer that a universal theory of meaning is impossible and conclude the paper with a brief discussion on what alternatives are available.

Keywords Theories of meaning · Metasemantics · Liar paradox · Quotation · Nominalisation · Lawvere fixed-point theorem · Semantic pluralism

1 Expressing a theory of meaning

When we want to talk about meaning, the first thing we need is some way to refer to the expressions whose content we're interested in. If I want to talk about what 'Snow is white' means, then I need some way to construct a term from that sentence, which allows it to be mentioned rather than used. What I just did was to use quotation marks for precisely that purpose.

In general, what we need is some way to construct a term t_e of our language from the expression e we wish to talk about. But not just any such function will do. If we're

✉ Niklas Dahl
niklas.dahl@fil.lu.se

[1] Department of Philosophy, Lund University, Lund, Sweden

trying to coherently ascribe meaning, then we better get two distinct terms whenever we start from two distinct expressions. Otherwise, any attempt to ascribe meaning to one of these expressions would automatically assign the same to the other.[1] Viewed as a function, this term construction needs to be injective, that is if $t_e = t_{e'}$ then $e = e'$. Since every injective function has a left-inverse, this means there must be some way to undo the term construction and recover the original expression. Further, it must be defined on every expression in the language of the relevant type. Such functions for constructing terms from an expression are methods for *nominalisation*.

Quotation certainly has these properties. We can apply quotation marks around any expression and doing so for two distinct expression results in lexicographically distinct terms. Its left-inverse is simply the process of *disquoting* by removing the quotation marks (Saka, 1998, p. 115; Cappelen & Lepore, 2007, pp. 24–26). Using quotation as the basis for semantic theorising, however, is not without its issues. As Davidson (1979), pp. 28–29 points out, quoted expressions can be simultaneously used and mentioned. In his classic example:

> Quine says that quotation "... has a certain anomalous feature."

Better yet would be to have some method for nominalisation which guarantees that the resulting terms are themselves inoperative. In this paper I won't be discussing quotation in particular, except as a special case of nominalisation. Henceforth, I will be assuming that we have some function taking expressions to terms which does respect the use-mention distinction. I point this out mostly to bracket general questions about quotation. None of what I have to say should turn on this point. For ease of reading I will be using quasi-quotation marks for this construction, writing $\ulcorner e \urcorner$ for the term constructed from the expression e and D for the left-inverse disquotation function such that $D(\ulcorner e \urcorner) = e$. I will be using this same notation when we restrict the term construction to specific kinds of expressions, such as sentences and predicates.

One of the desiderata in the search for a theory of meaning is that it should be universal. That is, we would like a single theory which we can apply to explain meaning for any language. To explain what I mean I'm going to need a distinction which I've been glossing over so far. Dummett (1991), pp. 20–22 distinguishes between a *semantic theory for a language* and a *theory of meaning*.[2] A semantic theory is an assignment of meaning to the expressions of a specific language. Theories of meaning, on the other hand, are accounts of what meaning is and how semantic theories are constructed. They are, in a sense, the theories against which we evaluate proposed semantic assignments. Saying that we have a universal theory of meaning, then, is to say that semantic theories can be put into a single unified framework.

Now, what I've argued so far is that to express a semantic theory for a language $\mathcal{L}$, we need a method for nominalisation for the expressions of $\mathcal{L}$. The next step is to see that we get a similar condition if we are to have a universal theory of meaning. Adapting an argument made by Scharp (2010), pp. 267–271 assume that we have some

[1] Strictly speaking, we wouldn't run into trouble if only synonymous expressions yield identical terms. But in order to ensure that is the case we would need to already know which expressions are synonymous which, in turn, requires a language which can talk about their meaning.

[2] I'm using these terms slightly differently from Dummett. He uses the term "meaning-theory" for what I call semantic theory and reserves the latter for valuations of logical formulas.

such theory of meaning.[3] Naturally, that theory has to be expressed in some language $\mathcal{L}$. Now, if T is some semantic theory, then T has to be expressible in $\mathcal{L}$. Otherwise, we couldn't possibly evaluate whether it conforms to the theory of meaning. As this is the case for all semantic theories, it also holds for a semantic theory of $\mathcal{L}$ itself. Accordingly, $\mathcal{L}$ must be able to talk about its constituent expressions. That is, there must be a method of nominalisation internal to $\mathcal{L}$.

My main objective in this paper is to show that there can't be any universal theory of meaning. What I've said already is the first half of the argument, that we need a language with internal nominalisation to express such a theory. In the next section, I will argue that internal nominalisation and access to some logical symbols is sufficient for the grammatical construction of a sentence which can't coherently be assigned meaning.

Following that, section three puts these pieces together into an argument that there can't be a universal theory of meaning, since there can't be any language which can both express that theory and coherently be assigned meaning itself. Because of the parallel to theories of truth, I examine how different attempts to internalise such theories despite the existence of Liar sentences can be adapted to theories of meaning. Unfortunately, none of these attempts turn out to save the notion of a universal theory of meaning.

Finally, in the fourth section, I discuss what the consequences are of this result and what options are available for us in the search for semantic theories. I evaluate the idea of a hierarchy of theories and argue that this approach can't be used to solve the problem either. Finally, I sketch an approach based on locally coherent theories of meaning, modelled on how local charts can be given for topological spaces which can't be globally mapped, and illustrate how this idea might be used to develop a solution to the problem.

Apart from its narrow objective of arguing against a universal theory of meaning, the purpose of this paper is, broadly speaking, to show how little in the large debate on meaning and paradox actually turns on the notion of truth. While truth can be used as the semantic concept which brings about contradictions, it's not necessary to do so. Almost any approach to meaning will do, when we try to internalise it to a language it applies to. This is done by bringing together a very general fixed-point theorem with a categorial approach to grammar. In this way, I hope to show how this abstract approach allows us to generalise previously known results for a specific kind of semantic theories to theories of meaning in general.

2 Nominalisation and liar sentences

It has been known since Tarski (1944), p. 350 that theories of truth for a language $\mathcal{L}$ can't be internalised in $\mathcal{L}$. If a language can talk about the truth of its own sentences, then it has the resources to express a version of the Liar Paradox. This poses a problem for theories of meaning whose central notion is truth. If one such theory is to be

[3] Scharp's argument is to show that Kripkean solutions to the Liar Paradox can't be expressed in the language they apply to.

universal, then the argument above tells us that there has to be a language which can coherently express a theory of truth for its own sentences. This is a tall order, considering the limits imposed by Tarski's theorem.

The easiest way to see the problem comes from Quine's formulation of the Liar Paradox. Trying to capture the precise formulation of Tarski's theorem in natural language, he presents us with the following sentence:

> ⌜yields falsehood when preceded by itself within quotation marks⌝ yields falsehood when preceded by itself within quotation marks (Quine, 1961 , p. 8).

In this version of the paradox there's no trace of explicit self-reference, quotation does all that work. Of course, there's nothing special about quotation here. Any method of nominalisation would work just as well. The problem, in any case, is that this sentence is self-contradictory.

The issue for theories of meaning based on truth, then, is that the minimal requirements to express such a theory seem sufficient to prove a contradiction. What I want to argue here is that this problem isn't limited to theories which use truth as a central notion. By replacing the role of falsehood with a syntactic contradiction, understood as abbreviating $A \wedge \neg A$ for your preferred choice of A, we get a sentence which we can't coherently assign meaning, whose construction only requires the language to have internal nominalisation and some limited ability to express logical symbols.

> ⌜is the nominalisation of a predicate which produces a sentence that implies contradiction when applied to it own nominalisation⌝ is the nominalisation of a predicate which produces a sentence that implies contradiction when applied to it own nominalisation.

This somewhat unwieldy sentence is built from precisely the grammatical building blocks that I've assumed from the language: that it has internal nominalisation and logical symbols implication and contradiction. Now, recalling the standard definition that $\neg S$ abbreviates $S \rightarrow \bot$, what this sentence expresses on an intuitive reading is its own negation. To see this, let $P(t)$ be the predicate

> t is the nominalisation of a predicate which produces a sentence that implies contradiction when applied to its own nominalisation.

Then the sentence above has the form $P(\ulcorner P \urcorner)$, as the term we're applying the predicate to is precisely $\ulcorner P \urcorner$. On the other hand, what it intuitively expresses is the claim that $P(\ulcorner P \urcorner)$ implies contradiction. Thus, if we assume that the sentence holds then its negation follows. If we assume its negation, then we know that it holds. In other words, we have that

$$P(\ulcorner P \urcorner) \leftrightarrow \neg P(\ulcorner P \urcorner)$$

which certainly looks problematic. Now, without the additional semantic machinery about what the sentence expressed, we don't immediately get a contradiction within the syntactical theory itself, as is evidenced by the fact that Peano Arithmetic can

consistently express a very similar sentence.[4] Where we get into trouble is if we want to assign any kind of meaning to this sentence in a way which respects its syntactical structure since if we do, then we must assign the same meaning to both $P(\ulcorner P \urcorner)$ and its negation. And if we try to internalise this semantic assignment to the theory itself, then, assuming that whatever we take meaning to be commutes with negation and that negation has no fixed points, we get a contradiction.[5]

But that puts us in a bind. All we needed to construct a sentence which makes it impossible to internally express the semantics for a language was that it had internal nominalisation, standard negation, and could express predicate application. But since such a language, I've argued, is also a necessary condition for a universal theory of meaning, such a theory would be impossible unless it can be expressed in a language suffering from extreme logical destitution.

Put informally, this argument naturally raises some suspicion. It's not immediately clear how to understand what it means that a sentence expresses its own negation. For that reason I want to formalise this argument before discussing the problem further. To do so I will be using some tools from categorial grammar.

The idea here is to take an outside look at the language and think of its grammatical building blocks as functions between kinds of expressions in the language. In the very simple grammar I will be considering here we will start with two basic components: a set T of terms and a set S of complete sentences. Then we can think of, unary, predicates as functions which take a term as input and produce a sentence. In the notation, the set S^T of predicates consists of all such functions. One specific kind of term we will be considering are those which are the nominalisation of some predicate φ. Let $N \subseteq T$ be the collection of those terms. That is,

$$N = \{t \in T \mid \exists \varphi \in S^T : t = \ulcorner \varphi \urcorner\}$$

Using this framework we can also think of nominalisation and disquotation as functions. Dealing specifically with predicate nominalisation, we get the function pair

$$N \underset{\ulcorner\ \urcorner}{\overset{D}{\rightleftarrows}} S^T$$

where $D(\ulcorner \varphi \urcorner) = \varphi$ for all predicates φ. Finally, we need a function $ev : S^T \times T \to S$ which expresses predicate evaluation. This *evaluation map* takes a predicate φ and a term t and outputs the sentence $\varphi(t)$. In other words,

$$ev(\varphi, t) = \varphi(t)$$

[4] I'm grateful to a a helpful reviewer for pointing out that this needs to be clarified. Using PA as our theory of syntax we can define a predicate $Pr^{\perp}(x)$ whose intuitive reading is that 'there is a proof of contradiction whose only extra-logical assumption is x'. Diagonalisation then results in a sentence P such that PA can prove the bi-conditional $P \leftrightarrow Pr^{\perp}(\ulcorner P \urcorner)$. Now, this doesn't result in a contradiction since, by Löb's theorem, PA can only prove that $Pr^{\perp}(\ulcorner P \urcorner) \to \neg P$ if PA can prove $\neg P$. Thus, as long as neither P nor $\neg P$ is provable, the bi-conditional doesn't threaten the consistency of PA.

[5] And in the case of PA this is would result in essentially a standard proof of Tarski's theorem on the undefinability of truth.

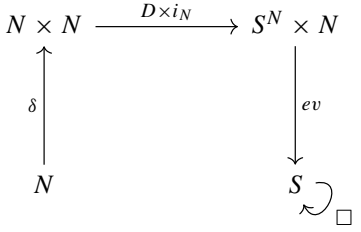
One thing that's important to note is that this framework is a way to describe the grammar of a language $\mathcal{L}$ from the outside, from the point of view of some meta-theory which can express sets and functions. That is, when I say that we can consider a predicate φ as a function, I don't mean that there is a function symbol within $\mathcal{L}$ which represents it in some appropriate sense. What I mean is just that from the grammatical point of view a predicate is simply something which combines with a term to construct a sentence. Similarly, the functions $\ulcorner\urcorner$, D, and ev are also not function symbols within $\mathcal{L}$. They are functions between the grammatical sets of expressions T, S, and S^T which tell us how the language is structured.[6]

This machinery is sufficient to prove a special case of the Lawvere Fixed Point Theorem, which says that every sentential operator $\Box$ of this language has a fixed point.[7] That is, for every such operator there is a sentence A such that $\Box A = A$. This, in turn, means that A and $\Box A$ have to be given the same meaning.

The essential idea of the proof is easiest to see by considering a diagram of the situation. To express it, however, we will need a few auxiliary functions. Let $\delta : N \to N \times N$ be the *diagonal map*, such that $\delta(t) = (t, t)$ and let $i_N : N \to N$ be the identity map $i_N(t) = t$. Finally, $f \times g$ denotes the function

$$f \times g(x, y) = (f(x), g(y))$$

which takes in a pair and applies the functions f, g to them respectively. Then we can draw the following diagram.



Picking a term $t \in N$, which is the nominalisation of some predicate, we can follow it around the diagram until we get a sentence. That is, composing all these functions results in a function $P : N \to S$: a predicate defined on nominalised predicates. In fact, if the sentential operator $\Box$ is negation, then P is the formal counterpart to the one I discussed above. Now, since $P \in S^N$, all we need to do is evaluate it at $\ulcorner P \urcorner$.

[6] The reason I'm working with an abstracted grammar for $\mathcal{L}$ in a meta-theory rather than within $\mathcal{L}$ itself is to explore what expressive abilities a language must have in order to express a theory of meaning. Now, as a reviewer pointed out, there's no guarantee that languages can internally represent every function which can be constructed on its set of terms. However, the assumption that the languages under consideration can do this is a minor one, since they're supposed to express theories of meaning which are universally applicable.

[7] The theorem is due to Lawvere (1969) and says roughly that if all functions from A to B can be represented by elements of A then all endofunctions on B have a fixed point. The proof below is the same, just translated to grammar and using the fact that disquotation, or any left-inverse to nominalisation, represents all predicates. For a modern discussion of Lawvere's result and its application to paradoxes, see Yanofsky (2003).

Lemma *Let $\mathcal{L}$ be a language with internal nominalisation of predicates which can express predicate application. Then every sentential operator $\Box : S \to S$ has a fixed point.*

Proof Let $\ulcorner \urcorner : S^T \to N$ be a method of nominalisation for predicates and $D : N \to S^T$ be its left-inverse. Define the predicate $P : N \to S$ by

$$P(t) = \Box ev(D(t), t)$$

where t is the nominalisation of some predicate. Now, since P is a predicate defined on N and $\ulcorner P \urcorner \in N$ we can evaluate $P(\ulcorner P \urcorner)$.

$$\begin{aligned} P(\ulcorner P \urcorner) &= \Box ev(D(\ulcorner P \urcorner), \ulcorner P \urcorner) \\ &= \Box ev(P, \ulcorner P \urcorner) \\ &= \Box P(\ulcorner P \urcorner) \end{aligned}$$

Hence, $P(\ulcorner P \urcorner)$ is a fixed-point for $\Box$. □

With the aid of this lemma, we're back to where the informal argument left off. Assume that there is a language $\mathcal{L}$ which has the expressive resources required for negation, internal nominalisation, and the ability to express predicate application. Then we can apply the lemma to negation $\neg : S \to S$ to get a sentence A which is a fixed-point for negation. Now, since what I've done so far is work with grammatical constructions there is not yet any contradiction. However, what we do have is a sentence A whose grammatical construction is such that if the assignment of meaning respects grammatical structure, then both A and $\neg A$ will have to be assigned the same meaning. In particular, if this assignment of meaning can be expressed within $\mathcal{L}$ itself and if synonymous sentences are equivalent, then it follows that $A \leftrightarrow \neg A$.

Essentially, this construction is the same diagonal process which underwrites most well-known paradoxes of self-reference. There is even a straightforward recipe for generating the problematic cases. Let A be the sentence constructed above and φ any predicate defined on nominalised sentences, i.e. any metalinguistic predicate, which commutes with negation in the following sense:

$$\varphi(\ulcorner \neg A \urcorner) \leftrightarrow \neg \varphi(\ulcorner A \urcorner)$$

Then it follows that

$$\varphi(\ulcorner A \urcorner) \leftrightarrow \varphi(\ulcorner \neg A \urcorner) \leftrightarrow \neg \varphi(\ulcorner A \urcorner)$$

which immediately results in a contradiction.

Now, if the metalinguistic predicate φ is defined compositionally, in the sense that it commutes with logical operations, on nominalisations of, sentences, then that includes negation. This is what prevents a language from being able to express claims about the semantic values of its own sentences in a compositional way. In particular, there's

nothing special about truth which brings about the Liar paradox. The same problem arises for any predicate which is compositional in this sense.[8]

3 Prospects of a universal theory of meaning

At this point I want to turn back to theories of meaning. My claim is that truth-based semantic theories aren't the only ones which can't be expressed in a language they apply to. I've already argued that internal nominalisation is a pre-requisite for a language to discuss its own semantics. But then, if the language can express negation and predicate application, it contains a self-contradictory sentence. Since both of these seem like pretty minimal assumptions for semantic theorising, we're in some trouble. Worse yet, the problem extends to theories of meaning since, as I argued above, we would need a language which can express its own semantics in order to express a universal theory of meaning.

What, then, are the options? In his discussion of the problem faced by truth-based theories, Tarski presents the problem as follows.

> But it is presumably just this universality of everyday language which is the primary source of all semantical antinomies, like the antinomies of the liar or of heterological words. These antinomies seem to provide a proof that every language which is universal in the above sense, and for which the normal laws of logic hold, must be inconsistent (Tarski, 1956, pp. 164–165).

The universality of everyday language which he discusses is essentially the fact that nominalisation is internal to it. Now, where Tarski concludes that colloquial language is contradictory, most philosophers prefer to seek a solution through weakening or changing the concept of truth. The common rationale for this is one they share with Tarski; that natural language simply is universal (Kripke, 1975, pp. 694–499; Reinhardt, 1986, pp.227–228; McGee, 1991, pp. 67–71; Priest, 2006, pp.11–12; Scharp, 2014, pp. 610–614).

Now, since my version of the problem makes no use of truth, these answers have to be modified somewhat. The syntactic counterpart to weakening the theory of truth is revising the logic we want our semantics to respect. In particular, we can make changes to the logic so that the existence of a sentence equivalent to its own negation avoids incoherence. This is no small change. Further, since every sentential operator is guaranteed a fixed-point, we would also have to change the conditional. Either the deduction theorem or modus ponens would have to go, since otherwise we could re-introduce standard negation through $A \rightarrow \bot$. To express a universal theory of meaning, then, we'd need a language and logic which allows every sentential operator to have fixed-points.

That this issue isn't restricted to any particular operator is parallel to the problem of the *extended liar* or *revenge paradoxes* which plagues theories of truth. As Graham Priest puts it:

[8] Yanofsky (2003) discusses the same phenomenon in a mathematical context. In both settings the paradoxes are special cases of the Lawvere Fixed Point Theorem.

> All semantic accounts [of truth] have a bunch of Good Guys (the true, the stably true, the ultimately true, or whatever). These are the ones that we target when we assert. Then there's the Rest. The extended liar is a sentence, produced by some diagonalizing construction, which says of itself just that it's in the Rest (Priest, 2008 p. 226).

For theories of truth the problem is that if whatever machinery we add to the notion of truth in order to deal with the Liar is itself expressible in the language, then it yields a notion which lacks a fixed-point. For instance, if we introduce the idea that a sentence can be neither true nor false, in which case we call it 'gappy', this handles the standard Liar. But if this solution can be expressed within the language itself the following sentence brings us right back where we started.

This sentence is false or gappy.

The logical counterpart to this problem would be to accept that certain sentences are equivalent to their own negation in which case neither A nor $\neg A$ holds. But then we could find a fixed point to the operator

$$\Box A = \neg A \vee (A \leftrightarrow \neg A)$$

which would again entail contradiction.[9]

Now, there are some theories of truth which can be expressed in an object language to which they apply.[10] Taking Field's theory as an example, the way he handles the Liar is by considering the sentence to be neither determinately true nor determinately false. Then, to avoid revenge objections, he notes that 'determinately determinately false' isn't the same as 'determinately false'. Every time we add 'determinately', we shift the semantic value of the sentence. Essentially, he has an infinite number of truth-values to work with so no finite disjunction ever forces a contradiction.

Using such a theory to specify the semantic interpretation of the logical symbols, we could possibly have a coherent assignment of meaning to a language which can express a universal theory of meaning. But, as Scharp (2014), pp. 618–621 has argued, revenge problems still prevent these theories of truth from being expressed in plenty of languages. Relatively simple extensions of the languages they're expressed in are sufficient to state revenge paradoxes and hence make them inconsistent. In the example of Field's theory, all we need is to add a determinateness operator D such that $D(D(S)) \leftrightarrow D(S)$ to yield contradiction. Similarly, even if we do manage to find a language which can express a universal theory of meaning, introducing a negation operator whose meaning is given by its standard introduction and elimination rules would yield contradiction.

Can such a theory of meaning really be considered universal? To make things vivid, imagine that the language under consideration is English. Then not only have

[9] At least on the assumption that the generalisation $A \vee \neg A \vee (A \leftrightarrow \neg A)$ of the Law of the Excluded Middle holds for all A. For the standard Liar, we don't need the LEM since even intuitionistic logic is sufficient to show that no sentence is equivalent to its own negation. However, that argument might not hold here depending on how the rules for negation were revised.

[10] For instance the various theories of truth proposed by Field (2008), Priest (2006), and McGee (1991).

we assumed that negation in English is not the familiar logical operator, but we couldn't even coherently add it to the language. On pain of contradiction, standard negation is unintelligible to speakers of English.

Avoiding this kind of situation is what makes Scharp (2014), pp. 610–612 impose the *internalisability requirement* on theories of truth. For such a theory to be acceptable, he argues, it's not enough that it can be expressed in some language to which it applies. It must be possible to extend any language with the expressiveness to state the theory. It's not unreasonable to expect a universal theory of meaning to satisfy the corresponding condition. At least for natural languages it does have a certain appeal. How universal can a theory of meaning be, if there are natural languages which couldn't even be expanded to coherently state it? However, the existence of such a theory would mean that the very idea of sentential operators without fixed-points is incoherent. Otherwise, there would be languages which can't be extended to express the theory of meaning.

Our available options, then, are to rule out a universal theory of meaning, declare standard negation to be unintelligible, or accept the strange situation where the theory of meaning can only be expressed in languages where quite simple notions can't be introduced. If I'm forced to choose one, then universality has to go.

4 Meaning without universality

I want to begin this part of the discussion by considering two different ways in which we can think of language as universal. Natural languages are certainly universal in one sense; to say that anything expressible can be said within some fragment of English is simply a truism. In this *grammatical* sense, languages are universal. But it doesn't follow that every fragment of a language is semantically *unified*. What I mean by this can be illustrated by how dialects of English can share grammatical structure, consist of the same syntactic expressions, while ascribing incompatible meaning to some sentences. In this way, we can think of dialects as distinct semantic theories for a single syntactic language.

Rejecting a universal theory of meaning, then, means recognising that the grammatical and semantic borders of a language are different. English, as understood syntactically, certainly has the grammatical resources to express internal nominalisation and predicate application. Further, it can express a notion of negation which, I've assumed for now, shouldn't have semantically fixed-points. Then no coherent semantic theory can apply to every grammatically acceptable sentence of English. Languages, in the semantic sense of expressions together with a system of meaning, must be smaller than their syntactic counterparts. Further, we need distinct theories of meaning against which to judge these semantic theories for fragments of the language.

One option which immediately springs to mind is to adopt a hierarchical view. We can compare this to the debate on multiple theories of truth where this kind of Tarskian tower of theories has been discussed extensively (Quine, 1961, pp. 7–9; Kripke, 1975, p. 697; Soames, 1999, pp. 151–152). The idea is that we order the fragments of our language into a sequence $\mathcal{L}_0$, $\mathcal{L}_1$, ... which can express the semantics and theory of meaning for the preceding fragment. But even for theories of truth this runs into problems. For this case, let T_n be a theory of truth for the language $\mathcal{L}_{n-1}$.

Then, Kripke (1975), pp. 695–697 points out, that to understand the truth ascription for some sentence A, I'd need to know which level of the hierarchy it belongs to. Otherwise I wouldn't know which theory of truth is applicable. This problem is even worse when it comes to claims which quantify over sentences. For instance, a philosopher who wants to defend Kripke's theorising might say that

Everything Kripke wrote about truth is true.

This sentence can only be evaluated against a theory of truth which is higher in the hierarchy than each of Kripke's claims about truth. Thus, for truth-conditional views of meaning, to even understand this sentence requires total knowledge of Kripke's work on truth.

Yet again, the same problem occurs when we deal with a hierarchical view of meaning. If T_n is a semantic theory for $\mathcal{L}_{n-1}$ and M_n is the theory of meaning it adheres to, then to understand a sentence we'd need to know which level of the hierarchy it belongs to. Otherwise, we'd be unable to know which theory provides the proper interpretation. But what language could be used to express this situation? Certainly not one within the hierarchy itself. To express its construction, through a sentence such as

If e is an expression in $\mathcal{L}_n$ then $\ulcorner e \urcorner$ is an expression in $\mathcal{L}_{n+1}$.

we'd need a language which allows nominalisation of expressions from every level of the hierarchy. Thus, its theory of meaning can't be any of the M_n.[11]

With these dim prospects for a hierarchical view, is there a more egalitarian option? One way to look at our problem is that we have some sort of object, a language, whose semantic structure can't be described globally. That is, we can only offer semantic theories for some local fragments of it. To illustrate a potential way forward, then, I want to make a comparison to a similar situation in geometry. When trying to understand a space S it's very useful to have a system of co-ordinates. If S is two-dimensional, a system of co-ordinates is nothing more than a function which takes points in S to pairs (x_1, x_2) of numbers. In this way, we can think of a system of co-ordinates as a way to represent the space through a flat map.

But not every space can be given a global representation this way. It's well-known that spheres, such as the surface of the Earth, can't be accurately represented by any single planar map. They can, however, be divided into regions which are then provided with *local co-ordinates*. Further, if we have two regions which overlap we can ensure that their local co-ordinates are inter-translatable so they provide essentially the same. That is, although the surface of the Earth can't be globally described by flat maps, we can represent it completely through the kind of overlapping local maps found in an atlas.[12]

[11] This is an adaptation of an argument given by Soames (1999), pp. 151–152 against a hierarchy view of truth. He points out that any language which can express that structure would contain the predicate $\exists n \top_n(\ulcorner S \urcorner)$, where $\top_n$ is the truth-predicate of the theory T_n. But this would be a truth-predicate for the entire hierarchy.

[12] What I'm describing informally here is what mathematicians call a *topological manifold*. An n-manifold M comes with a collection of open sets $\{U_i\}$ such that $M \subseteq \bigcup_i U_i$. That is, the sets U_i divide M into regions. For each U_i there is a homeomorphism $h_i : U_i \to \mathbb{R}^n$ providing local co-ordinates. That the local

What I want to propose, then, is that we take a similar approach to understanding how meaning works for a natural language. Although we can't provide either a semantic theory nor a theory of meaning for the entirety of English, perhaps we can provide them for fragments of the language. Imposing some inter-translation conditions when two theories apply to overlapping fragments, we could potentially piece together a picture of the entire language. Of course, this picture would remain implicit, since any language which could express it would run afoul of the now familiar problem. But these theories would provide locally coherent semantics for the language even though a globally coherent version is out of the question. And since we never use the entirety of language at any one time, local coherence might be enough for communication.

The easiest way this perspective can avoid paradox, then, is simply to say that although the self-contradictory sentence is grammatical, it doesn't belong to any of the semantic fragments of the language. Essentially, this is just refusing to ascribe even local meaning to this particular sentence. Given this similarity to gappy solutions to the Liar, we might suspect this solution to be vulnerable to a revenge paradox. But this is not so. As we've given up on a universal theory of meaning, we no longer need a language which can both express that a sentence is never assigned meaning and has internal nominalisation. Hence, we can't find the fixed-point required for a revenge paradox.

Nevertheless, I think there's a more interesting alternative which can be illustrated by taking another look at quotation. Since quotation marks provide a method for nominalisation, it follows that they can't be internal to a language.[13] Although this idea might seem strange at first glance, quotation marks are sometimes used this way explicitly. Every time we make claims about other grammatical languages we use quotation to import their expressions. Consider, for example, the following sentence.

"... medför motsägelse."[14] is not an English expression.

Here the quotation marks are a function taking Swedish expressions to English terms. Disquoting the expression returns a Swedish predicate which could not then be grammatically applied to a term in English. Disquotation, then, returns the expression to its original language and semantic context.

The way to avoid the paradox above is to apply similar reasoning to quotation which is ostensibly internal to a language. Let $\mathcal{L}_1$ is a fragment of English containing the following predicate which I will denote as P.

... implies contradiction when applied to its own quotation.

This language has a semantic theory T_1 which ascribes meaning to its constituent expressions. Now, applying quotation marks to P results in a term of the fragment $\mathcal{L}_2$

co-ordinates cohere on overlapping regions is expressed as follows: if U_i, U_j are two regions then the map $h_j \circ h_i^{-1} : h_i(U_i \cap U_j) \to h_j(U_i \cap U_j)$ is a homeomorphism. That is, the co-ordinates provided by h_i, h_j respectively are homeomorphic on their overlap.

[13] Since it's generally accepted that quotation is productive, disquotable, and universally applicable (Davidson, 1979, p. 37; Richard, 1986, p. 390; Saka, 1998, pp.114–115; Cappelen, 2007, pp. 22–26), and hence can be used for nominalisation, one way to think of the theorem above is as a challenge for theories of quotation.

[14] This is Swedish for "... implies contradiction."

whose semantic theory is T_2. Then even if $\mathcal{L}_2$ happens to also contain the unquoted predicate, and thus the full sentence,

> "implies contradiction when applied to its own quotation." implies contradiction when applied to its own quotation.

we can deny that this expresses a self-contradictory claim. This sentence, as given meaning by T_2, expresses information about the expression $D(\ulcorner P \urcorner)$ which results from disquoting the $\mathcal{L}_2$-expression $\ulcorner P \urcorner$. That is, it's expressing a claim about an expression of $\mathcal{L}_1$, saying that, according to T_1, P applied to itself within quotation marks implies a contradiction. In fact, since $\mathcal{L}_1$ doesn't itself contain $\ulcorner P \urcorner$, this claim can be thought of as vacuously true, assuming we can make sense of that notion.

On this view, quotation marks are functions between semantic fragments of a language. And since disquotation undoes quotation, the marks have to carry information about which semantic fragment the expression originated in. In this way, the seemingly self-referential paradoxes go away. Although they're constructed by applying a predicate P to $\ulcorner P \urcorner$, the orthographic identity of the two expressions hides the fact that the outer instance of the predicate in $P(\ulcorner P \urcorner)$ inhabits a different semantic fragment of the language than the inner one. As such, the resulting sentence isn't about the predicate itself, but the same sequence of symbols with respect to a distinct semantic theory.

A full development of quotation along these lines is unfortunately beyond the scope of this paper. Although it avoids paradox, this solution exacerbates the problem of mixed-quotation when a quoted expression is both mentioned and used. Trying to explain this phenomenon while maintaining a strict division between the semantic fragments the expression occurs in is a non-trivial task. Further, if quotation is always between specific semantic fragments of the language, it's hard to think of it as a single ability. As it stands, my sketch is only meant to show that we can avoid self-referential paradox by abandoning the idea that languages are semantically unified.

Despite avoiding paradox, however, the pluralist approach to theories of meaning isn't always better off than its universal competitor. If every natural language is composed of several semantic fragments, then successful communication would require identifying which fragment your conversational partner inhabits. This kind of problem is what prompted Quine (2013), pp. 26–30 to consider *radical translation* and what made Davidson (1973), p. 313, introduce its reflexive cousin *radical interpretation*. According to Davidson, what communication requires is a joint method for discovering what our conversational partner believes all the while simultaneously constructing a semantic theory for their language.

Here, the situation is even worse. Davidson's radical interpretation can at least rely on a shared truth-conditional theory of meaning between the speakers. But I've argued that English doesn't just require a plurality of semantic theories, it also needs several theories of meaning. As such, the people who attempt to interpret each others language might not even agree on the general shape of semantic theories, much less their content. Hence, even radical interpretation might not be extreme enough to bridge that communicative gap.

Whether this problem can be overcome is the principal test for this way of organising theories of meaning. One reason to be sceptical, however, is just how narrow the

Goldilocks zone is for the interpretive relation. On the one hand, we need to be able to identify the semantic theory of our conversational partners well enough to communicate. To do so, we must first discover the theory of meaning which governs it. On the other hand, my interpretive success can't require that I can embed your theory of meaning within my language. If it did, then you could do the same and our languages could express their own theories of meaning by composing our interpretations with one another.

There is one final option which I have not yet mentioned, namely having no theory of meaning at all. Such a claim can be interpreted in at least two ways. We could concede that there simply are no general conditions on semantic theories. In the realm of radical interpretation anything goes, at least so far as structure is concerned. The other is that meaning is simply the wrong conceptual tool to explain communication. I find myself reluctant to endorse either view. But in the end, it is the phenomena of language use, our ability to communicate and express ourselves, which needs saving, not some particular conception of meaning.[15]

Acknowledgements Jiwon Kim, Martin Jönsson, and Erik J. Olsson.

Funding Open access funding provided by Lund University.

Declarations

Conflict of interest The author has no competing interests to declare that are relevant to the content of this article.

References

Cappelen, H., & Lepore, E. (2007). *Language turned on itself: The seman-tics and pragmatics of metalinguistic discourse*. Oxford University Press.
Davidson, D. (1973). Radical interpretation. *Dialectica, 27*(3–4), 313–328.
Davidson, D. (1979). Quotation. *Theory and Decision, 11*, 27–40.
Dummett, M. (1991). *The logical basis of metaphysics*. Harvard University Press.
Field, H. (2008). *Saving truth from paradox*. Oxford University Press.
Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy, 72*(19), 690–716.
Lawvere, F. W. (1969). Diagonal arguments and cartesian closed categories. *Lecture Notes in Mathematics, 92*, 134–145.
McGee, V. (1991). *Truth, vagueness, and paradox: An essay on the logic of truth*. Hackett Publishing Company.
Priest, G. (2006). *In contradiction: A study of the transconsistent*. Oxford University Press.

[15] I would like to thank Jiwon Kim, Martin Jönsson, and Erik J. Olsson for their comments and discussions on the contents of this paper.

Priest, G. (2008). Revenge, field, and ZF. In J. Beall (Ed.), *Revenge of the liar*. Oxford University Press.
Quine, W. V. (1961). The ways of paradox. *The ways of paradox* (pp. 1–18). Harvard University Press.
Quine, W. V. (2013). *Word and object* (New). The MIT Press.
Reinhardt, W. N. (1986). Some remarks on extending and interpreting languages with a partial predicate for truth. *Journal of Philosophical Logic, 15*, 219–251.
Richard, M. (1986). Quotation, grammar, and opacity. *Linguistics and Philosophy, 9*, 383–403.
Saka, P. (1998). Quotation and the use-mention distinction. *Mind, 107*(425), 113–135.
Scharp, K. (2010). Truth and Expressive Completeness. In B. Weiss & J. Wanderer (Eds.), *Reading brandom: On making it explicit* (pp. 262–275). Routledge.
Scharp, K. (2014). Truth, revenge, and internalizability. *Erkenntnis, 70*(3), 597–645.
Soames, S. (1999). *Understanding truth*. Oxford University Press.
Tarski, A. (1944). The semantic conception of truth: And the foundation of semantics. *Philosophy and Phenomenological Research, 4*(3), 341–376.
Tarski, A. (Ed.). (1956). *Logic*. semantics, metamathematics. Oxford University Press.
Yanofsky, N. S. (2003). A universal approach to self-referential paradoxes, incompleteness, and fixed-points. *Bulletin of Symbolic Logic, 9*(3), 362–386.