This is the accepted version of a chapter published in *A Companion to Donald Davidson*.

Citation for the original published chapter:

Pagin, P. (2013)
Radical Interpretation and the Principle of Charity.
In: Ernie Lepore and Kirk Ludwig (ed.), *A Companion to Donald Davidson* (pp. 225-246).
Oxford: Wiley-Blackwell

N.B. When citing this work, cite the original published chapter.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:diva-98846
The idea of radical interpretation, using The Principle of Charity, plays one of the two central roles in Davidson’s philosophy of language. It is introduced already in ‘Truth and Meaning’ and then developed in a series of papers in the mid-1970s. Ideas shift in the mid-1980s and early 1990s. We shall here focus mostly on the early views.

1. What is radical interpretation?

Radical interpretation is a species of interpretation. The term ‘interpretation’ itself has a process-product ambiguity: in the product sense it denotes a content attribution, or the content itself that is attributed, by an interpreter. In the process sense, an interpretation is the process that leads to an interpretation in the product sense. The “aim” of an interpretation process is understanding, i.e. knowledge of content.

Interpretation, in the process sense, covers everything from the automatic or semi-automatic comprehension of utterances in a well-known language to the methodologically elaborate exegesis of old texts or the struggle to understand speakers of an alien language. In case the goal is to understand an entire language, the interpretation product is what Davidson calls a meaning theory. We shall be concerned with the process of interpretation that leads to a meaning theory.

Davidson emphasizes that a theory of interpretation must reveal semantic structure in the object language, whereas a theory of translation need not:

A theory of translation must read some sort of structure into sentences, but there is no reason to expect that it will provide any insight into how the meanings of sentences depend on their structure.

A satisfactory theory for interpreting the utterances of any language, our own included, will reveal significant semantic structure: the interpretation of utterances of complex sentences will systematically depend on the interpretation of utterances of simpler sentences, for example. (Davidson 1973, p. 130)

*For comments and discussion, I am indebted to Kathrin Glüer, Ernie Lepore, and Kirk Ludwig.
Both interpretation and translation into one's own language yields understanding of foreign utterances, but interpretation is a more theoretical endeavour.¹

The term ‘radical interpretation’ is nonetheless coined in analogy to Quine’s ‘radical translation’ in chapter 2 of *Word and Object*. The intuitive idea, in both cases, is that of translation/interpretation that “starts from scratch”, without any previous knowledge of the language in question, or detailed knowledge of the attitudes of the speakers. Although radical translation/interpretation in this sense has taken place in history, for both Quine and Davidson describing it is rather a thought experiment. The point is to identify the kind of evidence that is ultimately available to an interpreter and how that evidence supports a meaning theory, i.e. the evidential relation (Davidson 1973, 128).

Although the idea of interpretation from scratch highlights the problems and resources of the interpreter, it cannot be essential to radical interpretation, for if were, the interpreter would cease to be radical when he has made some progress. Furthermore, whether or not the interpretation process once started from scratch is historically contingent. What matters is what the ultimate evidential basis is:

The problem of interpretation is domestic as well as foreign: it surfaces for speakers of the same language in the form of the question, how can it be determined that the language is the same? Speakers of the same language can go on the assumption that for them the same expressions are to be interpreted in the same way, but this does not indicate what justifies the assumption. All understanding of the speech of another involves radical interpretation. (Davidson 1973, p. 125)

That my fellow speakers mean the same as I with words we both use, is not, on this view, anything we have the right to take for granted, but something that requires justification, and ultimately from evidence of the very same kind as is available to the interpreter that does start from scratch.

This has an immediate consequence for the question of identifying the *interpretee* of radical interpretation. It cannot be a speech *community*, but must be an individual speaker. The reason is that we need evidence to determine whether or not two speakers belong to the same speech community (cf. Davidson 1973, p. 135).²

When interpreting a speaker who makes an assertion in a familiar language we typically infer what she believes on the basis of what we take the sentence to mean. If she

1Davidson points out in ‘Truth and Meaning’, p 21, that a syntactic theory plus a lexicon does not add up to a meaning theory, and the same can be said about translation: we can translate belief sentences into our own language without knowing how to provide a compositional semantics for them.

2Davidson is somewhat ambivalent on this important issue. He does suggest (1975, p 169) that the interpreter should aim at a theory that optimizes agreement in the speech community. However, this is different from taking for granted that agreement exists.
speaks an unfamiliar language but we happen to have independent information about what she believes or wants, we can move on to a plausible guess about the meaning of the sentence she used. Ultimately, however, the radical interpreter does not have access to sentence meaning as basic evidence, and neither does he have access to information about particular beliefs, desires or intentions as data for interpretation. The meaning of sentences and the contents of attitudes will be what his theory attributes to the speaker. The evidence must be something else (Davidson 1973, p. 134).

Without knowing what the speaker believes or expresses the interpreter can, however, observe the speaker's linguistic utterances and reactions to utterances by others, including the interpreter himself. The interpreter can form an hypothesis about the attitude to a sentence that the speaker manifests. In particular, Davidson focused on the attitude of **holding-true**, or more precisely, holding true relative to a time, in order to take account of indexical sentences, like ‘it is raining’ (Davidson 1973, p. 135; Davidson 1974a, p. 144). Holding a sentence true is an attitude to a sentence that corresponds to **believing** the proposition that is the meaning of the sentence. Holding-true is indeed a kind of belief, but it is a belief with a very coarse-grained, impoverished content; identifying the **content** of such a belief requires no more than identifying the sentence the belief is about, for there is no need to identify the meaning of the sentence. Provided the radical interpreter can identify manifestations of the holding-true attitude on the part of the speaker, he has access to data that are independent of knowledge of sentence meaning or fine-grained individual beliefs.

It might be objected that the radical interpretation theory fails because it must illicitly assume that we can pre-theoretically recognize the behavior that manifests the holding-true attitude. But there is no need to **know** this in advance of interpretation. Rather, the interpreter forms an hypothesis about it and tries to establish an interpretation on that basis. If things work out well, this confirms the hypothesis, and if no reasonable interpretation results, it will have to be rejected. This hypothesis is part of the overall interpretation package, although it relates as evidence to the rest of his theory.

The most important difference between assent behavior and holding true is that holding true is related to meaning and belief in a **uniform** way, while the disposition to assent to a sentence is not. For instance, **insincere** assent is still assent, but it does not, in the relevant sense, manifest a holding true attitude. Similarly, a speaker may systematically be disposed **not** to assent to a sentence in circumstances where she nonetheless holds it true, e.g. for reasons of prudence. With holding true, a general statement like

\[
\text{For any time } t, \text{ Kurt holds 'Es regent' true at } t \text{ iff it is raining near Kurt at } t, \tag{1}
\]
(Davidson 1973, p. 135) describes data for interpretation, but we cannot in general expect corresponding regularities for dispositions to assent. Davidson was not a behaviorist, and regarded appeal to mental states as fully legitimate in theoretical contexts. The reason for not trying to connect meaning and belief directly to behavior is not that behaviorism is false, but that the behaviorist project is more difficult (cf. Davidson 1974a, 142). The behaviorist project would be even more informative if successful, but without regularities analogous to the one stated in (1), success might be too much to hope for. Still, the data for the interpreter ultimately comes from observation of linguistic behavior (that this is a weak form of behaviorism is emphasized by Glüer (2011, p. 27).

At the next step, the interpreter faces a serious problem of underdetermination: just as the truth of a sentence depends in part on what the sentence means and in part of what the facts are, so the speaker holding true a sentence depends in part on what the sentence means in the speaker's language and in part on what the speaker believes. So, if the interpreter knows what the speaker believes and what a sentence in the speaker's language means, he can infer what the speaker will hold true:

\[
\text{INF1) 1) } X \text{ believes that } p \\
2) \ s \text{ means that } p \\
3) \text{ Hence, } X \text{ holds } s \text{ true}
\]

Similarly, if the interpreter knows what a sentence means, and knows that a speaker holds it true, he can infer what the speaker believes:

\[
\text{INF2) 1) } X \text{ holds } s \text{ true} \\
2) \ s \text{ means that } p \\
3) \text{ Hence, } X \text{ believes that } p
\]

There is no analogous simple inference from belief to meaning:

\[
\text{INF3) 1) } X \text{ holds } s \text{ true} \\
2) \ X \text{ believes that } p \\
3) \text{ Hence, } s \text{ means that } p
\]

The simple reason why (INF2) goes through but (INF3) does not is that a (disambiguated) sentence has only one meaning, while a speaker has many beliefs, and the second premise of (INF3) does not provide the information that this particular belief is responsible for holding $s$ true. Still, the interpreter can infer that among the sentences the speaker holds true, at least one means that $p$ (since we are dealing with beliefs expressible in
the speaker’s language). Knowledge of all the beliefs of the speaker would then allow the interpreter to infer what meanings many sentences of the speaker’s language must have, but not directly how these meanings are distributed over the sentences.

Initially, the interpreter knows only what sentences the speaker holds true. Different hypotheses about what the sentences mean lead to different hypotheses about what the speaker believes, and, indirectly, vice versa. Davidson refers to this as the *interdependence of belief and meaning* (Davidson 1973, p. 134). Strictly speaking, however, so far we only have an interdependence between *ascriptions* of meaning and *ascriptions* of belief.

We have now seen what the ultimate evidence is for radical interpretation, which largely explains what radical interpretation is: it is facts about when and under what circumstances the speaker holds sentences true.³ The other aspect of radical interpretation is the methodology of *using* the evidence for breaking into the interdependence. We will return to this in section 3.

### 2. The role of radical interpretation for meaning theory

The opening of Davidson 1973 runs:

Kurt utters the words “Es regnet” and under the right conditions we know that he has said that it is raining. Having identified his utterance as intentional and linguistic, we are able to go on to interpret his words: we can say what his words, on that occasion, meant. What could we know that would enable us to do this? How could we come to know it?

The questions at the end of the quote can be spelled out as follows:

(A) What is the form of a theory knowledge of which allows an interpreter to understand any sentence of a speaker’s language?

(B) On what basis can an interpreter acquire such knowledge?

The (A) question is really two questions stylistically merged into one:

(A1) What is the format of a theory such that some knowledge about it is sufficient for an interpreter to understand any sentence of a speaker’s language?

(A2) What knowledge about such a theory is required for an interpreter to have this ability?

³Davidson also tried to unify the theory of meaning and the theory of action (cf. chapter 14). In doing so, *desire*, or *preference*, must be added to belief, as a basic attitude. This alternative was never much developed, however, and has no demonstrated advantage.
Clearly, it is not enough to know a theory that in fact has the right property. I must also know that it has the right property, and at least it must be true. So, the answer to (A2) will at the very least include knowledge that the theory is true.

Davidson’s proposed answer to (A) is that a meaning theory for a natural language should have the form of a Tarskian truth definition, with theorems – T-sentences – of the form of (2).

(2) \( s \text{ is true in } L \text{ iff } p \)

(cf. chapter 10). Davidson’s answer to question (B) is radical interpretation. At the very least, this means that it is possible to reach knowledge that a particular meaning theory is true by means of radical interpretation. As we shall see, Davidson’s claims are stronger than that.\(^4\)

We can now see why interpretation must be radical for playing this role. For if it is to deliver the evidence for a meaning theory, the account of meaning would be circular if facts about meaning would also serve as evidence for the theory (Davidson 1974a, p. 142). An appeal to fine-grained beliefs of the speaker would not be directly circular, but still indirectly so, given that we can gain knowledge of others’ fine-grained attitudes only through interpretation.

The crucial question is what reason there is to believe that radical interpretation carried out in the right way will deliver the right result, i.e. give knowledge of the correct theory, or theories. Some critics have claimed that the view that it will is just verificationism on Davidson’s part.\(^5\) Davidson’s view, however, was that it is because of the special nature of the object of knowledge – what speakers mean – that facts in this area are knowable. Language is a social phenomenon (Davidson 2005b, p. 56), and so linguistic meaning must be essentially public; what one person means must be accessible to others (Davidson 1979, p. 235; Davidson 2005a p. 56):

Quine revolutionized our understanding of verbal communication by taking seriously the fact, obvious enough in itself, that there can be no more to meaning than an adequately equipped person can learn and observe; the interpreter’s point of view is therefore the revealing one to bring to the subject. (Davidson 2005a)

There is a tension here with the early claim, noted in the preceding section, that we cannot take for granted that two speakers belong to the same speech community, that only the application of radical interpretation will reveal whether they do. To motivate

\(^4\)In one sense, the claim is weaker, for Davidson immediately adds that he is not asking for what interpreters in fact know, or how they in fact acquire the knowledge.

\(^5\)For instance, Williamson 2004, p. 137.
the radical interpretation approach by appeal to the social and public nature of language is precisely to presuppose pre-theoretically that there is successful linguistic communication. That is just what makes language a social phenomenon.\(^6\) Hence, if radical interpretation with a chosen method would deliver the result that there is no successful linguistic communication, or substantially less than believed, there would be something wrong with that method.

The view that meaning is public is not only a view about the epistemology of language. It does not only entail that speakers can know what other speakers mean, but also that this in the nature of linguistic meaning:

The semantic features of language are public features. What no one can, in the nature of the case, figure out from the totality of the relevant evidence cannot be part of meaning. (Davidson 1979, p. 235)

This is, in a sense, a metaphysical conclusion, about the nature of meaning, from an epistemological premise, that speakers can know what others mean. In taking this step, the epistemological fact is taken as constitutive of linguistic meaning. It is not just the case that radical interpretation provides the evidence for a meaning theory, successful interpretation is what also makes it correct. In a sense, radical interpretation (or interpretability) determines meaning and belief. If this is correct, then it immediately explains why properly performed radical interpretation of any speaker is guaranteed to succeed, for meaning and belief is precisely what can be known in this way.

This is a controversial view.\(^7\) Why couldn’t there be speakers with radically different conceptual schemes that we simply would not be able to interpret? In fact, it does not follow that there couldn’t. What follows is that anyone who is a speaker can be radically interpreted by some interpreter (perhaps someone who is sufficiently alike), not necessarily by all interpreters, and not necessarily by “us”. Nevertheless, the claim that any speaker is interpretable seems to have been taken in this stronger sense of interpretable by us, also by Davidson himself.

Davidson discusses this issue at length in (1974b). At one point Davidson remarks that “we could not be in a position to judge that others had concepts or beliefs radically different from our own” (Davidson 1974b, 197). This is an epistemic remark, but it is connected with a conceptual claim about truth: our best grasp of the concept of truth derives from Tarski’s work, but Tarski’s concept of truth as applied to an arbitrary object language is essentially connected with translation into our language, or a language we understand,

\(^6\)This is made explicit by Davidson in Davidson and Glüer 1995, p. 81.

\(^7\)Critics include Fodor and Lepore 1994, Lepore and Ludwig 2005, chapter 12, and more recently also García-Carpintero 2012.
as meta-language (Davidson 1974b, p. 194-95). It is not completely clear, however, why the translation of the object must be into a language we understand; Tarski explicitly allows the meta-language to be an extension of the object language. So, if translation is possible between languages that we cannot understand, then Tarski’s concept of truth appears to be applicable to them as well.

The idea of a radically different conceptual scheme is closely connected with the role of the radical interpreter in radical interpretation. If meaning is determined by the result of a properly performed radical interpretation could give, is the interpreter himself really needed? David Lewis (1974) claimed that the figure of the interpreter is only a dramatic device: what determines meaning is the totality of facts that the interpreter, if he did his job well, would make use of, and meaning supervenes on these facts in accordance with the correct principles of interpretation. The appeal to the interpreter only dramatizes the supervenience relation. In his reply (1974c), Davidson agrees. The appeal to the interpreter serves to highlight the nature of the supervenience base, which comprises exactly what can also serve as evidence to an interpreter, and the supervenience relation, which is encoded in the principles of interpretation. But strictly speaking, the interpreter can be dropped.8

The idea that facts about the speaker determines meaning, rather than the activity of an interpreter, helps with understanding two puzzling features of the radical interpretation idea. The first is the apparent circularity of determination. On the one hand the fact that a speaker X holds a particular sentence s true at a time t depends on what the speaker believes at t (or just before) and on the meaning of s in the language of X (at t or just before). On the other hand, since meaning and belief is determined by the results of interpretation, what X believes and what s means depends on the fact that X holds s true. Does this really add up? If we understand determination as a temporal relation, it does not. For then holding s true would depend on pre-existing facts about belief and meaning, while on the other hand those facts would not themselves be determined until after s is held true by X.

We can make sense of the situation in the following way. X has a mental state α and a relation ρ to the sentence s, and together these two facts cause X to hold s true. The causal powers of α and ρ precede, or make up the basis of, the holding true attitude. α is in fact the state of believing that p, but it is that state in virtue of satisfying certain conditions, and among these conditions is precisely that together with instantiating the relation ρ to s it results in holding s true. So, in virtue of fitting into a causal pattern,

8On this theme, see especially Glüer 2011, 132-37. Note that if the interpreter is dropped, truth by the lights of the interpreter is dropped as well; what remains is truth itself.
α qualifies as being a belief that \( p \). This asymmetry between a causal relation in one direction and a satisfaction relation in the other is obscured if being a belief and having meaning depends on an actually performed interpretation.

We can also note here that this view justifies the talk of “interdependence of belief and meaning” (page 5), since for a state to be a belief that \( p \) depends on a sentence having the meaning that \( p \) (perhaps in context), and vice versa.

The second puzzling feature concerns the totality of the evidence. In (1979, p. 235) Davidson says that “what no one can, in the nature of the case, figure out from the totality of the relevant evidence cannot be part of meaning”. What is this totality? Davidson says that “we may as well suppose we have available all that could be known of such attitudes, past, present, and future” (1974a, p. 144). All that could be known is a lot, and it is more than any real interpreter can acquire, partly because the quantity is huge, partly because he cannot look into the future.

The advantage of allowing/requiring this massive amount of evidence is that the meaning determining role of radical interpretation becomes more plausible. Any real interpreter will have access to only a very small subset of the attitudes of the speaker, and will need to extrapolate heavily from these attitudes. New evidence might easily point in a different direction, disconfirming current hypotheses. So, in a sense, the appeal to the totality of evidence takes care of the problem of induction. Of course, since the appeal to the totality of evidence leaves all actual interpreters behind, the relation between actual interpretation and the meaning theoretical project becomes less clear.

Two questions emerge at this point: First, will appeal to the totality of evidence help with guaranteeing that radical interpretation of a speaker will succeed? Second, if the interpreter is dropped from interpretation, where do the concepts used in interpretation come from? These two questions are interconnected, but to see how, we need to turn to the methodology of radical interpretation.

3. The Principle of Charity

How can the radical interpreter break into the interdependence of belief and meaning? Davidson’s proposal for doing this is a cornerstone in his philosophy of language. Basically, the proposal is that although the interpreter does not at the outset have any particular knowledge of the speaker’s fine-grained beliefs or other attitudes, he does have general knowledge, which he can put to use. He can know that if someone has beliefs at all, most of these beliefs are true (by the interpreter’s lights).

---

9 Analogously, a shooting qualifies as being a killing after the victim dies; the dying doesn’t cause the shooting to be a killing. This view of the relation between belief, meaning, and holding true makes the radical interpretation account into a kind of functionalist account. Cf. Loar 1986.
The method is intended to solve the problem of the interdependence of belief and
meaning by holding belief constant as far as possible while solving for meaning.
This is accomplished by assigning truth conditions to alien sentences that make
native speakers right when plausibly possible, according, of course, to our own view
of what is right. (Davidson 1973, p. 137)

This is a statement of what became known as *The Principle of Charity*. In its simplest ver-
sion, the idea of using Charity in interpretation is the idea that an interpretation is *better*
if it leads to attributing *more* true beliefs. This works precisely because of the interde-
pendence of belief and meaning, in particular because of the validity of schema (INF2):

\[
\text{(INF2) } \begin{align*}
1) \quad & X \text{ holds } s \text{ true } \\
2) \quad & s \text{ means that } p \\
3) \quad & \text{Hence, } X \text{ believes that } p
\end{align*}
\]

Premise 1) here records the evidence for the interpreter. Premise 2) states his interpreta-
tion *hypothesis*. The conclusion is used for evaluating the hypothesis.

Consider two global meaning theoretical hypotheses, theories $T_1$ and $T_2$. For a great
number of sentences held true by the speaker, $T_1$ and $T_2$ contain substantially different
hypotheses about their meaning. According to $T_1$, say, $s$ means that *there is a hippopotamus in the refrigerator*, and according to $T_2$, $s$ means that *there is an orange in the refrig-
erator* (cf. Davidson 1969, pp. 100-01). As a consequence, the interpreter has a choice
between two belief attributions to the speaker: that there is a hippopotamus in the re-
frigerator and that there is an orange in the refrigerator. The first belief is pretty absurd
while the second may well be true. The second is preferable, in particular if it *is* true. This
then speaks in favor of $T_2$ over $T_1$. This exemplifies the basic mechanism of how belief
attribution influences semantics, via Charity.

Before moving on to elaborate on the idea of Charity, a few remarks on the mecha-
nism are in order.

1. The role of Charity is that of comparison and evaluation. In terms of the philos-
ophy of science: Charity belongs to the *context of justification*, not the *context of discov-
ery*. Charity is a tool for testing whether a meaning theory is acceptable, not primarily
a method of selecting hypotheses to test. Other factors, such as relevance and general
psychological plausibility will be important for hypothesis formation. As Davidson em-
phasizes as regards actual interpretation, a theory is "derived by wit, luck, and wisdom"
(Davidson 1986b, p. 107). Neither is required for testing.

2. Charity is primarily applicable to a theory as a whole, not to individual theorems.
A speaker may well have a number of false beliefs (we probably all do), and may even
have some absurd beliefs as well. What is compared is the totality of belief attributions according to one theory with the totality of belief attributions according to another.

3. One reason the interpreter cannot (with ordinary concepts) simply devise a meaning theory that makes the speaker hold only true beliefs is that the theory must be compositional; the theorems must be connected by being derived from a shared basis. If that were not the case, the interpreter could simply pick a true interpretation for any sentence held true by the speaker, because then the interpretation of one sentence would not impose any restriction on the interpretation of any other sentence. Since any belief could then come out true, the interpreter could score high on Charity. This observation shows that compositionality must be an independent requirement, because it cannot be justified from Charity (cf. Pagin 1999).

Let's go back to the content of the charity principle. A typical early formulation is

"The general policy ... is to choose truth conditions that do as well as possible in making speakers hold sentences true when (according to the theory and the theory builder's view of the facts) those sentences are true (Davidson 1974a, 152)."

So, in the most basic version, interpretations should be chosen that maximize the rate of true beliefs among the speaker's beliefs, judged according to the standards of the interpreter (cf. Davidson 1975, p.169). Theories that do not reach the maximum rate of truths, i.e. those that are not among the best theories, must be rejected. 10

The interpreter is always to maximize the rate of true beliefs by the interpreter's standards, or according to the interpreter's view of the facts, and this comes to the same thing as maximizing agreement; it is compatible with the falsity of the beliefs of both speaker and interpreter. We get real maximizing of the rate of truth only by assuming that the interpreter's beliefs are all true (or the interpreter is dropped). 11

In the most basic formulations, applying Charity is to make a comparison between theories. However, acceptable theories must not only be best, they must also be good. They must be such that speaker and hearer are rendered largely in agreement, i.e. such that most of the speaker's beliefs come out true, by the interpreter's standards:

What justifies the procedure is the fact that disagreement and agreement alike are intelligible only against a background of massive agreement. (Davidson 1973, p. 137)

10 Davidson accepts the consequence that two or more different theories can be equally good but better than all others. These top-ranking theories are then all true, despite being apparently incompatible. This is what Davidson calls "indeterminacy of interpretation", analogous to Quine's indeterminacy of translation. Cf. Davidson 1979 and this volume, chapter 15.

11 Davidson did assume the possibility of an omniscient interpreter in Davidson 1986a.
(cf. Davidson 1975, p. 168-69). The requirement of massive agreement is typically presented in the context of an argument for Charity in the previous respect. However, it is clearly an aspect of Charity in its own right, since it provides an absolute requirement on the rate of truth (it must be high), which complements the relative requirement (it must be the highest).

Davidson nowhere gives a numerical value of the rate of true beliefs. On the contrary, from early on he stresses that this cannot be done. One reason is infinity: there aren't more sentences in a natural language than there are sentences in the same language with an odd number of words. Both sets are countably infinite, and hence of the same cardinality (cf. Davidson 1973, p. 136).

It may be immediately objected here that what is compared for size are sets of beliefs, or sentences held true, and since these are only finitely many, the problem of infinity does not arise. But this is not so clear. I believe much more than I happen to express verbally. If I believe that snow is white, do I also believe that snow is white or grass is red, and in general that p or q, for arbitrary q, if I believe that p? If so, it seems that I have infinitely many beliefs after all, and if not, where do we draw the line? Davidson later claimed that "there is no useful way of counting beliefs, and so no clear meaning to the idea that most of a person's beliefs are true" (Davidson 1986a, p. 138). And in the preface to Davidson 1984 (p. xvii), he called maximizing agreement "a confused ideal", since what we want is "the right sort of agreement". And already early on (1973, p. 136; 1974a, p. 197), he used the locution 'optimize agreement'.

One main difference between simple maximizing and optimizing is that the latter takes epistemological aspects into account. As stressed by Lewis (1974) and Føllesdal (1975), it is bad interpretation to ascribe a true belief to speaker when the speaker is not in an epistemic position to have a reason for that belief (like a rabbit behind a tree, not in view from where the speaker is). By the same token, it is bad interpretation to ascribe a false belief to a speaker when she is in a a very good epistemic position for it.

Some disagreements are more destructive of understanding than others, and a sophisticated theory must naturally take this into account. Disagreement about theoretical matters may (in some cases) be more tolerable than disagreement about what is more evident; disagreement about how things look or appear is less tolerable than disagreement about how they are; disagreement about the truth of attributions of certain attitudes to a speaker by that same speaker may not be tolerable at all, or barely. [...] The methodology of interpretation is, in this respect, nothing but epistemology seen in the mirror of meaning. (Davidson 1975, p. 169)

In particular, observation sentences, such as 'that is white', or near observation sentences, such 'that is snow', are particularly important. On the one hand they are sen-
tences that the speaker holds true in some situations but not in others, and therefore they give richer data for interpretation than sentences always or never held true. On the other hand, they are sentence that speakers are very seldom wrong about. So, on the assumption that a particular sentence $s$ is an observation sentence for the speaker, it is best interpreted so as to be true almost whenever the speaker holds it true. We could sum this up as follows: Charity requires the interpreter to choose the theory that scores highest on some scale (yet to be fully specified), summing the measures of these beliefs according to their varying epistemic weights.

Later on, a new ingredient emerges: the connection between content and causality. In basic cases, a belief is about what causes it:

> What stands in the way of global skepticism of the senses is, in my view, the fact that we must, in the plainest and methodologically most basic cases, take the objects of a belief to be the causes of that belief. (Davidson 1986a, p. 151)

The most basic cases are the cases of perceptual belief, i.e. beliefs based directly on perceptual experience, which again are observational beliefs. This is an element of Content Externalism that emerges in Davidson's philosophy in the early 1980s. The content of a belief is determined by features of the environment, and in particular by what causes it. This idea is closely connected with how he later comes to think of Charity. For on the later conception, part of The Principle of Charity is then taken to be what he then calls The Principle of Correspondence:

> The Principle of Correspondence prompts the interpreter to take the speaker to be responding to the same features of the world that he (the interpreter) would be responding to under similar circumstances. (Davidson 2001b, p. 211)

Prima facie, Correspondence seems rather unlike Charity as it had been presented earlier on. Charity is used for getting content out of truth and linguistic structure, with no appeal to causality. Correspondence is used for getting content out of causality, with no appeal to truth or structure. While Charity, as earlier conceived, applies to the totality of a speaker's beliefs, Correspondence seems to apply to individual perceptual beliefs. I shall treat it as a complement to Charity rather than a new conception of it.

Correspondence together with Externalism are connected with the theme of triangulation that emerges together with Externalism. According to the basic idea of triangulation, what the cause of a belief is remains indeterminate until the believer is in communication with a second believer, for then we can look back in the causal chains to what is shared between the two subjects:
Communication begins where causes converge: your utterance means what mine does if belief in its truth is systematically caused by the same events and objects (Davidson 1986a, 151).

By the triangulation idea, a speaker must be not only interpretable, but actually interpreted. The idea of the radical interpreter as just a dramatic device is far away. And what does the job is causality and the shared environment, not the principles of interpretation. Although it may be thought that triangulation underlies the possibility of radical interpretation, it is a different theme.\footnote{Cf. chapter 27.}

The Principle of Correspondence is presented together with the Principle of Coherence:

The Principle of Coherence prompts the interpreter to discover a degree of logical consistency in the thought of the speaker; […] Both principles can be (and have been) called principles of charity: one principle endows the speaker with a modicum of logic, the other endows him with a degree of what the interpreter takes to be true belief about the world. (Davidson 2001b p. 211)

The idea that the interpreter must take the speaker to be largely consistent and coherent was part of Davidson’s principles of interpretation from the outset in ‘Truth and Meaning’, and in a sense part of Charity.

This is not yet all that has been connected with Charity. Davidson also thought that the interpreter needed to track the speaker’s beliefs in evidential connections, and in action explanation. This was part of his idea of a unified theory thought, meaning, and action (Davidson 1980).\footnote{Cf. chapter 14.} Since the relevance for interpretation was not worked out in detail, I shall not pursue the theme here.

4. The Power of Charity

When we ask what Charity can deliver when applied in interpretation, we are interested in two questions:

1) Can Charity fail to yield true meaning theories?
2) Can Charity yield false or unilluminating meaning theories?

These questions are connected with the idea of testing Charity empirically by comparing results with actual interpretation. This is an issue we will return to in the final section. Here we will be concerned with abstract possibilities. The first question concerns un-
dergeneration of true theories, the second overgeneration. Two lines of reasoning will be examined here, both related to question 2), the first also to question 1).

In general, Charity does not require that the interpreter can translate the speaker’s language into his own pre-existing language. The speaker may have a more sophisticated theory of nature, so that the interpreter needs to define new vocabulary in order for the interpretation to go through. Is there a limit to this?

Assume that we can state a compositional meaning theory with theorems of the form ‘s means that p’. Assume that interpreter Casey from Earth embarks on interpretation of apparent speaker Alien from Outer Space. Casey identifies a candidate predicate Φ that seems applied to some objects and withheld from others by Alien, but Casey sees no pattern in the usage. None of the property concepts Casey can come up with matches even approximately the pattern of Alien’s applications.

Casey then decides to learn from Alien, and starts defining a new predicate F in his own language. It is defined by cases: true of objects that Alien applies Φ to, false of objects that Alien withholds Φ from, and for all objects b unconsidered by Alien, F is true of b just in case b is a rocket. Clearly, by interpreting Φ to mean F, and assuming Casey has identified atomic sentences with Φ as predicate and a demonstrative as subject term, Alien’s demonstrative Φ sentences all come out true.

Casey then goes on to do the same with other predicates, and also with what he identifies as grammatical particles, and sentence constructions. For each sentence held true at time, on a case by case basis, an interpretation is given of the parts and the syntactic operations that makes the sentence come out true at that time. Some arbitrary interpretation is provided for all cases not considered by Alien. So Casey’s meaning theory is compositional and complete (with respect to the syntax he has identified), and results in only true beliefs being attributed to Alien.

Still, something is wrong. We feel that question 2) above should be answered affirmatively: the interpretation does not deliver any real understanding of Alien. It would be natural to locate the problem in predictability. However, the problem is not that Casey’s theory does not deliver predictions, for it does. It is only that we have no good reasons to believe that those predictions would be borne out. But what exactly would give us such reasons?

Davidson never emphasizes the predictive capacity of meaning theories. Charity itself is adapted only to successful accommodation, i.e. to give a theory that fits the data. The very idea of having access to the totality of evidence, as proposed in Davidson 1974a, p. 144, (“all that could be known of such attitudes, past, present, and future”), suggests that no actual prediction will ever have to be made.
Successful prediction increases our confidence in a theory, but does not in itself make the theory better supported by the data than had it been devised by accommodation. In the present case, we feel Casey’s theory is utterly worthless, and this cannot apparently be explained by the fact that no successful prediction has actually been made. Nevertheless, the problem seems to have something to do with prediction.

Casey clearly has no idea of what is similar to the objects that are F, by his definition, over and above being treated as similar by Alien. With an independent grasp of the similarity, i.e. with the ability of applying a previously given or independently defined concept to all or nearly all of them, Casey would also have been able predict, with reasonable confidence, how Alien would apply Φ to new cases. This, then, seems to be a reasonable restriction: interpretation takes place only with independently possessed concepts. And, in fact, it is plausible to read such a restriction into the Principle of Correspondence, for it requires the interpreter to take the speaker to be responding to the same features of the world has he himself. That condition is not met in Casey’s interpretation, for he cannot discover the features Alien is responding to, only define them through Alien’s responses.\(^{14}\)

Without this restriction from Correspondence, Charity would intuitively overgenerate. With the restriction, perhaps it might undergenerate.

The second issue is connected the fact that Davidsonian meaning theories are truth theories, with theorems of the form of

\[
(2) \quad s \text{ is true in } L \text{ iff } p
\]

If theorem were to be of the form

\[
(3) \quad s \text{ means that } p
\]

then the task of radical interpretation would consist in finding evidence that the theory is true. By contrast, with theorems of the form (2), this is not obviously enough, since biconditionals of that form may be true even if the meaning of \(s\) is not that \(p\). Davidson’s own classical example from 1967, p. 25, is

\[
(4) \quad \text{‘Snow is white’ is true iff grass is green}
\]

A theory with (4) as theorem would be non-interpretive. Such theories should be weeded out by a combination of structural and interpretational measures. In particular, the se-

---

\(^{14}\)What happens with this condition if the idea of an actual interpreter is not needed? We should assume that there are properties that speakers/interpreters may and may not have a concept of, and may and may not be able to have a concept of (a subject can fail to be sensitive to a certain property). If Alien’s Φ expresses a concept at all, there is a property it represents. Interpreters sensitive to the same concept could then succeed with Charity.
mantic axiom for ‘white’ would also be used in giving a theorem for the sentence ‘This is white’ (Davidson 1967, 26, note 10). A theorem for this sentence saying that the sentence is true iff what is demonstrated by the speaker is green, would be incorrect. Moreover, it would be detected by Charity, for the resulting belief attributions would represent the speaker as systematically believing of white things that they are green. Therefore, such a theorem about ‘This is white’ would have to be rejected, and together with it the axiom for ‘white’ on which is was based. Once the axiom is rejected, (4) is no longer a theorem either.

In this way, theories with true but non-interpretive theorems would be sorted out by means of Charity. Hopefully, then, the remaining true theories would all be interpretive as well. It may be said to have been Davidson’s most central idea in the philosophy of language that a meaning theory may be extensional, and indeed have the Tarskian format, because the evidence for it provides more than just its truth. With the right evidence, the theory is not just true, it is also interpretive, i.e. interpretationally adequate:

What we have been overlooking, however, is that we have supplied an alternative criterion: this criterion is that the totality of T-sentences should (in the sense described above) optimally fit evidence about sentences held true by native speakers. The present idea is that what Tarski assumed outright for each T-sentence can be indirectly elicited by a holistic constraint. If that constraint is adequate, each T-sentence will in fact yield an acceptable interpretation. (Davidson 1973, p. 139)

This hope that structure and Charity together would suffice was challenged in an objection to Davidson by John Foster (1976). Suppose we have an interpretive T-theory $T_1$. Then, we devise another theory $T_2$, so that they are related for any $s$ and $p$ as follows:

(5) a. \[
    \frac{T_1}{s \text{ is true iff } p}
\]

b. \[
    \frac{T_2}{s \text{ is true iff the Earth moves and } p}
\]

Given that the Earth does move, the two theories are equally well supported by Charity, and should both be correct if one is.

One possible response would be accept the result, say that both theories are true, and both interpretive, and that therefore we have an example of indeterminacy of interpretation. But Davidson did not accept it. According to Davidson’s reply (1976), interpretive T-theorems have a kind of modal force: they are lawlike. (5a) is lawlike, but (5b) is not. So, although theory $T_2$ would be true, it would not be interpretive.

What does lawlikeness amount to here? The lawlikeness pertains to the generalization over (speakers of the same language and) instances of holding true, and entails that
such generalizations are counterfactual supporting. So the theorem (5a) is lawlike because it holds that

\[(6) \quad \text{If the speaker were to hold } s \text{ true, she would believe that } p.\]  

In this sense, (5b) would not be lawlike, for a speaker might well hold true 'snow is white' without believing that the Earth moves, or believing lawlike statements such as $E = mc^2$.

This raises several questions. First, we would have hoped to replace the second premise of (INF2) by an appeal to theorems of true T-theories. In this way Davidson would not need to explicitly use the concept of meaning in his theory. This would give:

\[(\text{INF2'}) 1) \quad X \text{ holds } s \text{ true} \]
\[2) \quad T \quad s \text{ is true iff } p \]
\[3) \quad \text{Hence, } X \text{ believes that } p \]

where the inference would be valid if theory $T$ is true (the premise as stated then (somewhat sloppily) abbreviates the conjunction of the statement that $T$ is true and that what follows that turnstile is one of its theorems). But the inference is valid just in case the theory is interpretive, and its truth does not guarantee this. The alternative of making it instead an explicit premise that the theory is interpretive (and hence that $s$ means that $p$) would surrender part (but not all) of the ambition of explicating the concept of meaning.

You might think that an appeal to lawlikeness, in the sense of (6), would solve this: the second premise of (INF2') would require exactly that the truth theory has lawlike theorems. This would have the additional advantage of making the inference not only conceptually valid, but even logically valid (i.e. with (6) as second premise). But although the idea is natural and promising, it has serious problems.

The first problem is that the proposal does not seem to avoid the need of appealing to the concept of meaning after all. For when we consider all the alternative circumstances (possible worlds) under which a speaker could hold a sentence $s$ true, we are

---

15Davidson does not make this understanding of lawlikeness explicit in his writings, but this is the answer he gave to a direct question, during a Q&A at the Davidson conference outside Kazimierz Dolny, near Lublin, in the summer of 1995, organized by Ursula Zeglen.

Lepore and Ludwig (2005), 162-63, take lawlikeness to amount to the idea that the theorem itself is true in all nomically possible worlds. They subsequently (correctly) criticize this idea for failing to guarantee interpretiveness, since it rules out some but not all Foster style alternatives (it does not rule out theories where the added conjunct is itself lawlike).

16Lepore and Ludwig (2005), 152-66 argue that appeal to being interpretive (p. 153) should be accepted in characterizing the adequacy of a truth theory (that it suffices for interpretation). Part of their reason is exactly that appeal to lawlikeness, as they understand it, does not guarantee interpretiveness.
interested only in those circumstances where the meaning of \( s \) remains the same. Allow the meaning to change, and the law need not hold; the speaker could hold \( s \) true without believing \( p \) in a world where \( s \) means something else. There does not seem to be any way of specifying the worlds we are interested in other than exactly as those worlds where the meaning of \( s \) is the same.\(^{17}\)

The second problem is that the appeal to laws includes an appeal to counterfactual holding true. Holding true that \( \text{could} \) have taken place but doesn’t will never be evidence for the interpreter. So distinctions in meaning that we could capture only by appeal to counterfactual holding true could not be detected by applying Charity.\(^{18}\)

The third problem is that the solution anyway faces a strengthened version of Foster’s problem. Take a proposition that the speaker \( X \) is sure to believe, for instance that \( 1 + 1 = 2 \), or that something exists, or whatever would be believed in any world where the speaker has the same language; call it \( q \).

Then, first, we could assign \( q \) as the meaning of every sentence \( s \), for whichever sentence \( X \) holds true, she believes that \( q \). Second, even if we somehow filter out such propositions, there remain the conjunctions \( p \) and \( q \), for arbitrary \( p \). A speaker that has conjunctive beliefs in the first place would believe that \( p \) and \( q \) whenever she believes that \( p \). So, we would not be able distinguish between the meaning hypothesis that \( s \) means that \( p \) and the hypothesis that \( s \) means that \( p \) and \( q \). If the one law holds, so does the other.\(^{19}\) A way around this problem would be accept the result and conclude that the two hypotheses are just as good and hence both true, i.e. appeal to the indeterminacy of interpretation. But the question is why this is to be preferred over accepting that the appeal to laws does not help.

A more constructive and promising way is to make use of Correspondence and Externalism. If the interpreter can conclude, by Correspondence, that the speaker reacts to a particular feature of the environment, then by Externalism, the speaker’s belief is about that feature. This can help against the strengthened Foster’s problem in this way: if the speaker’s belief is about the wetness of the grass, then it is not the belief that the grass is wet and \( 1 + 1 = 2 \), for that belief is only in part about the wetness of the grass. Then, by means of appeal to the later principles, which derive content from causality rather than truth, there may be a way out of the strengthened Foster problem, but more

\(^{17}\)An appeal to the Stalnaker-Lewis idea about counterfactuals is no guarantee, for we have no reason to believe that worlds where the meaning changes are further away than worlds where it stays the same.

\(^{18}\)An example would be Casey’s interpretations of Alien. His theory would not be lawlike (since the counterfactuals would be massively false), but it would satisfy Charity with respect to all actually available evidence.

\(^{19}\)Note that this is exactly the analogue of the possible-worlds semantics problem that if \( s' \) is true in all worlds, then for any sentence \( s, s \& s' \) have the same intension.
work is needed here.20 Much of this work would connect the theory of meaning with the philosophy of perception.21

5. The Justification of Charity

What would justify the belief that Charity is the right principle of interpretation? In this final section we shall be concerned with this question, and focus on Charity in the old, optimization sense.

It is clear from many passages in Davidson's writings that he took The Principle of Charity to have a validity of a foundational kind, so that justification would not be empirical, for instance:

I suggest that the existence of lawlike statements in physical science depends upon the existence of constitutive (or synthetic a priori) laws like those of the measurement of length within the same conceptual domain. Just as we cannot intelligibly assign length to any object unless a comprehensive theory holds of objects of that sort, we cannot intelligibly attribute any propositional attitude to an agent except within the framework of a viable theory of his beliefs, desires intentions, and decisions (Davidson 1970, 221).

What makes interpretation possible, then, is the fact that we can dismiss a priori the chance of massive error (Davidson 1975, 168).22

These quotes do not themselves give any justification of Charity, but pronounces on the epistemic or modal status the principle. Davidson refrained from saying that the statement of Charity is analytic. But calling Charity constitutive is close enough. What will it amount to? Davidson himself refers to measurement theory. Consider an example. In order for a function $H$, from physical objects to real numbers, to be a measure function for extensive measurement, there must be a binary operation $\circ$ on physical objects (a concatenation operation) such that for all objects $d, d'$

---

20 As far as I am aware, Davidson never explicitly appealed to Correspondence and Externalism as a solution to Foster’s problem, and one can only speculate about any historical connection. The connection is explicit in Lepore and Ludwig 2005, chapter 12, but developed with respect to a different principle, which they call “Grace”.

21 Compare here a passage from Davidson (thanks here to Kathrin Glüer):

The second part of the argument has to do with the empirical content of perceptions, and of the observation sentences that express them. We learn how to apply our earliest observation sentences from others in the conspicuous (to us) presence of mutually sensed objects, events, and features of the world. It is this that anchors language and belief to the world, and guarantees that what we mean in using these sentences is usually true. (…) The principle of charity recognizes the way in which we must learn perceptual sentences (Davidson 1999, 343).

22 For a thorough survey of Davidson's views on the matter, see Glüer 2006.
\( \mathcal{H}(d \circ d') = \mathcal{H}(d) + \mathcal{H}(d') \)

(see Krantz et al. 1971, chapter 1). This is part of a definition of a reconstructed, technical notion, call it measurement*. (7) is indeed constitutive of that notion, and it is analytic (follows from the definition) that (7) is true of measure* functions. What is less clear is in what sense (7) could be said to be constitutive of measurement (in the old sense) as well. Probably in no good sense, but this does not matter much, for measurement theory is normative, not descriptive; it sets scientific standards for what to call "measurement". It is not a theory about measurement as historically given.

By contrast, Charity is supposed to be part of a theory of meaning and belief as historically given. It is a descriptive theory, telling us something about the nature of meaning and belief as previously known. It is hard to understand in what sense Charity would be constitutive of meaning so understood. Maybe the idea simply reduces to the related idea that we can realize a priori that Charity is true.

In what follows I shall first discuss the possibility of empirical argument for or against Charity, and then turn to a priori arguments.

Could Charity come into conflict with empirical claims? One possibility of arriving at such a result would be to apply Charity to two speakers who pre-theoretically appear to communicate well, and check whether the same meaning theories would hold for them. If it would turn out that, by the standards of Charity they don't, then should we say that contrary to appearances, there was failure of communication, or that Charity is false? If the latter is the case, then at least to some extent, Charity can be empirically tested. This is definitely the case if we would have greater confidence, concerning crucial cases, in the claim that appearances of successful communication are veridical than in the truth of the principle.

If Charity can be tested, then it can also get empirical support. An idea is suggested in Pagin 2006. We start from the fact that there is massive verbal agreement, and hence agreement in holding true, between speakers of (what appears to be) the same language. The second step is to argue that the best explanation of this fact is that there is also agreement in meaning and belief (and hence in belief-forming mechanisms). The alternative would be either agreement by chance, or else that differences in belief are systematically offset by compensating differences in meaning. Sameness of meaning and belief-forming mechanisms is chosen as the simpler theory, and therefore the better explana-

---

23 In case of weight, the concatenation operation consists in weighing two objects together.

24 It may be the claim that Charity is in fact already part of folk psychology (thanks to Kathrin Glüer here). That is an empirical question, however.

25 I had the chance of asking Professor Davidson this question during a conference dinner in 1998. His answer was that "it is open for adjustment on both sides".
This argument (which has its complications) leads to the conclusion that interpretation should proceed by Charity, for if the interpreter has achieved massive agreement, there is also reason to believe that interpretation has succeeded. However, no negative conclusions result: if massive agreement is not reached, there might be agreement in meaning and disagreement in belief, or agreement in belief and disagreement in meaning, or disagreement in both, and the interpretee might not be a speaker at all.

Turning to a priori arguments, one idea discussed in the literature is that the mere possibility of radical interpretation provides an argument for Charity, since Charity is required for radical interpretation to succeed (cf. Lepore and Ludwig 2005, 204-07). In short, the argument goes:

(8)  
   i) Radical interpretation is possible
   ii) Necessarily, if radical interpretation succeeds, then Charity is true
   iii) Hence, Charity is true.

We can see immediately that the argument as stated is invalid. A conclusion that does follow is that Charity is possibly true, not that it is actually true. Lepore and Ludwig don’t make this mistake, since their second premise (p. 204) is that Charity is true if radical interpretation is possible. One may wonder here why the success of radical interpretation in some possible world requires the truth of Charity in the actual world. Presumably, premise i) is implicitly strengthened to concern only possible worlds w so close to the actual world a that if Charity is true in w, it is also true in a. This makes the second premise of Lepore and Ludwig true, but it makes argument (8) valid as well. However, the strengthening also makes premise i) much more controversial.

There is a further question why the success of radical interpretation would require Charity in the first place. As pointed out by Glüer (2011), p. 143, any applicable principle that would map available evidence on meaning would suffice for the radical interpreter. Charity is just one among many. It is indeed one of the most plausible principles, but this plausibility does not derive from its being alone. Further, Davidson has explicitly denied that he argued for Charity from radical interpretation. He suggests rather that on his view, radical interpretation is possible because Charity is true (Davidson 1994, p. 127).

A different argument for Charity is based on belief holism.26 In a crucial passage Davidson says:

> If I am right in attributing the belief to you, then you must have a pattern of beliefs much like mine. No wonder, then, I can interpret your words correctly only by

26That Davidson argues this way is stressed in Glüer 2011, pp.114-19.
The reasoning seems to be that having certain concepts requires having some general beliefs and certain patterns among the particular beliefs. For instance “I can believe a cloud is passing before the sun, but only because I believe there is a sun, that clouds are made of water vapour, that water can exist in liquid and gaseous form; and so on, without end” (p. 201). Then, if the interpreter attributes beliefs with these concepts, he must also attribute some appropriate general beliefs, and appropriately related particular beliefs. Then, large-scale agreement will result.

Even if we accept this belief holism, the argument does not have the consequence of large-scale agreement. As Lepore and Ludwig (2005), pp. 201-02, point out, the agreement required only concerns the general beliefs, which need not be that many. The condition that the speaker must have a certain pattern among the particular beliefs, however, can be satisfied even if most or all of the particular beliefs are false (by the interpreter’s lights). The speaker believes of a table that it is a cloud, and accordingly also believes that it consists of water vapour, and so on. Insanity is often characterized by large-scale coherence among equally crazy beliefs.

Davidson’s most basic a priori-style argument for Charity concerns the inadmissibility of too many absurd beliefs. A famous passage is this:

Let someone say (and now discourse is direct), “There’s a hippopotamus in the refrigerator”; am I necessarily right in reporting him as having said that there is a hippopotamus in the refrigerator? Perhaps; but under questioning he goes on, “It’s roundish, has a wrinkled skin, does not mind being touched. It has a pleasant taste, at least the juice, and it costs a dime. I squeeze two or three for breakfast.” After some finite amount of such talk we slip over the line where it is plausible or even possible to say correctly that he said there was a hippopotamus in the refrigerator, for it becomes clear he means something else by at least some of his words than I do. The simplest hypothesis so far is that my word ‘hippopotamus’ no longer translates his word ‘hippopotamus’, my word ‘orange’ might do better. (Davidson 1969, pp. 100-01)

There is a strong intuition that the homophonic interpretation hippopotamus is absurd and the non-homophonic orange is clearly better. The former cannot really be right; it “does not make sense”.

This is in accordance with how The Principle of Charity was originally introduced, under that name.27 How wrong can you be about something and still correctly be taken to be talking about it? Not extremely much, the claim is. However, the intuition of the

example only shows that between an absurd and a non-absurd interpretation, the latter is to be chosen, other (global) things equal.

This may indeed be seen as a good a priori argument. But note that this intuition does not alone support the view that absurd interpretations without non-absurd alternatives are unacceptable. And even if there were an intuition against absurd interpretations categorically, this would not suffice for justifying negative conclusions; that the failure of finding non-absurd interpretations by Charity is a sufficient reason for saying that the interpretee is a non-speaker. Nor would it suffice for justifying the maximizing principle that all interpretations with alternatives that are better should be discarded, even if they are not themselves absurd, merely suboptimal.

Consider the possibility that the interpretability relation, $x$ can interpret $y$, is not transitive: A can interpret B, B can interpret C, but A cannot interpret C, because differences between A and B, and between B and C, are small enough, but the differences between A and C are too great. From A’s perspective, B is a speaker and interpreter, and by Charity B successfully interprets C. So A has reason both to say that C is a speaker and that C is not a speaker. If this possibility cannot be ruled out a priori, then there is no a priori argument that Charitable interpretability by us is a necessary condition for being a speaker. And we can make an analogous argument about non-absurd but suboptimal interpretations. All in all, the justification of Charity remains much of an open question.

References


