The Common Ground – Radical Interpretation

Marcello Di Bello - ASU - Spring 2022 - Week #3

The topic of today's class is radical interpretation: what it takes to interpret an alien language. Last week we discussed how the common ground—a shared set of beliefs—makes successful communication and mutual understanding possible, with positive as well as negative effects (recall the examples of hate speech and pornography). Both the papers we will discuss today are titled 'Radical Interpretation', one by Donald Davidson[1] and the other by David Lewis.[2]

[1] Davidson (1973), Radical Interpretation, *Dialecta* 23(3/4).

[2] Lewis (1974), Radical Interpretation, *Synthese* 23.

Davidson

What knowledge would we need to be able to interpret sentences in a language? How would we go about acquiring that knowledge?[3]

These questions apply to sentences in our own as well as a foreign language, but come in sharper focus in thinking about a foreign language.[4]

[3] In Davidson's own words: "[What] could we know that would enable us to do this [=interpret the sentence "Es regnet"]? How could we come to know this?" (p. 313)

[4] So, for example, what knowledge would we need to be able to interpret the German sentence 'Es regnet'?

Failed attempts

Davidson (pp. 316-7) examines a number of options he thinks do not work. Suppose the knowledge we need to interpret a sentence is knowledge of what the sentence means. Have we made any progress? It seems not:

> There is indeed also the hint that corresponding to each meaningful expression there is an entity, its meaning. The idea, even if not wrong, has proven of little help: at best it hypostatisizes the problem. (p. 314)

Or perhaps we could reduce all verbal communication to just a series of physical disturbances and gestures:

> all verbal communication consists in nothing more tha elaborate disturbances in the air, which form a causal link between the non-linguistic activities of human agents. (p. 314)

But this description, while perhaps accurate, would not answer the question, for sentences are left uninterpreted.

Another option is to link simple sentences to behavioral data, or single words to behavioral data of some kind. But these methods, while promising, have a problem. How do we go from simple sentences such as 'it is raining' to complex sentences like 'if it raining, I must bring an umbrella, but if it is not, it is best to leave the umbrella

at home'? Another question is, how we can connect the meaning of words such as 'it' or 'raining' to the meaning of sentences such as 'it is raining'? This proposals has no workable answers.

Another approach might be to link meanings and intentions: that to interpret what a sentence means is to understand what the speakers intended to do with that sentence, or something of that sort.[5] But, for Davidson, this is not going to work either:

> interpreting an agent's intentions, his beliefs and his words are parts of a single project, no part of which can be assumed to be complete before the rest it. (p. 315)

[5] This is surely reminiscent of Grice and Stalnaker from last week.

The Proposal

So what is Davidson's proposal? His starting point is a translation theory of interpretation: to interpret a sentence is to translate into a familiar language.[6] This seems on the right track, but there is a problem:

> we can know which sentences of the subject language translate which sentences of the object language without knowing what any of the sentences of either language mean (p. 317)

[6] The translation would require three languages: the language to be translated (object language), the language to be translated into (subject language), the language used for the translation (meta-language).

We could add to the theory of translation a theory of interpretation of the object language, but this would be 'unnecessarily bulky' (p. 317). Davidon's proposal for a theory of interpretation of a lnguage L is a Tarski's style theory of truth of L that, for every sentence s of the language L to be interpreted, entails

> s is true (in the object language L) if and only if p,

where s is the sentence of L to be interpret and p is a (truth-condition preserving) translation of p.[7] Call each of the bi-conditionals a *T-sentence*, such as

> 'Es Regnet' is true (in German) if and only if it is raining.

[7] Think of the theory of truth for L a machinery that generates many bi-conditional statements like the one in the text, for every sentence of L.

The Proposal

But how is this an answer to our initial questions? To see why, consider three questions:

FIRST: Can this proposal work for all sentences in the (alien) language we want to interpret?

Not all sentences are of the form 'Es Regnet' (it is raining). Other forms are: I *wish* it would rain; I *hope* it does not rain; did it rain? Etc. How would we apply a T-sentence to a question? This does not seem to work:

'did it rain?' is true if and only if p^*

Clearly, questions cannot be true or false. So this approach seems doomed to fail. Davidson is more optimistic:

> In the first stage, truth will be characterized, not for the whole language, but for a carefully gerrymandered part of the language. This part, though no doubt clumsy grammatically, will contain an infinity of sentences which exhaust the expressive power of the whole language. The second part will match each of the remaining sentences to one ... of the setences for which truth has been characterized (p. 320)

SECOND: What kind of evidence would we gather to tell that the theory of truth for L is correct?

Davidson gives an example of how this could work (pp. 322-3). For one thing, we have a T-sentence:

> "Es Regnet" is true-in-German (when spoken by x at time t) if and only if it is raining (near x at t)

The evidence for this T-sentence would be:

> Kurt belongs to the German speech community and Kurt holds true "Es regnet" on Saturday at noon and it is raining near Kurt on Saturday at noon.[8]

This matching of evidence and T-sentences should be done on a massive scale, for all sentences of L.[9]

But what if speakers are wrong abut what they hold to be true. They hold to be true that it is raining, but it is not.

> We want a theory that satisfies ... that maximizes agreement, in the sense of making Kurt (and others) right, as far as we can tell, as often as possible.

People, after all, cannot be wrong on a systematic basis, but only occasionally:

> this 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 as often as possible, according, of course, to our own view of what is right. What justifies the procedure is the fact that disagreement and agreement alike are intelligible only against a background of massive agreement. (p. 324)[10]

Davidson is imagining this process to take place in three stages (pp. 323-4): first, we settle T-sentence for logical truths (held true all time); then, we do the same for indexical and emprical sentences (held true in some circumstances); finally, we tackle all other sentences.

[8] Note how this takes the attitude of "holding true" as empirically verifiable, unlike intentions (which are intertwined with beliefs)

[9] Here each T-sentence is empirically falsifiable, perhaps in relation to other T-sentences. Like in Tarski's theory of truth, T-sentences follow rules of compositionality, such as Q&P is true if and only if Q is true and P is true. Unlike in Tarski's theory of truth, sameness of meaning is not what guides the selection of T-sentences, but rather, empirical evidence about what sentences speakers hold true and under what circumstances.

[10] Here Davidson assumes a maximally shared set of beliefs, even between people who speak different languages. If that were not the case, Davidson thinks we would have no reason 'to count that creature as rational, as having beliefs, or as saying anything' (p. 324). How plausible is this assumption?

THIRD: Knowing the theory of truth for language L were correct, would this allow us to interpret sentences of L?

So, suppose we have a theory of truth of L such that (1) for every sentence of L, it assigns a T-sentence, and (2) there is a match between empirical evidence at our disposal and all the T-sentences entailed by the theory of truth for L. Are we thereby able to interpret every sentence of L?

Davidson is clear that T-sentences do not give the meaning of the sentence we want to interpret:

> it is clear that a T-sentence does not give the meaning of the sentence that it concerns (p. 325)

However:

> the totality of T-sentences should (in the sense described above) optimally fit evidence about sentences held true by native speakers. . . .A T-sentence of an emprical theory of truth can be used to intepret a sentence, then, provided we also know that teh T-setence is entailed by some true theory that meets the formal and emprical criteria (p.327)

There are some complications here—for example, what if the theory of truth of L is not unique give the empirical evidence we have—but this is essentially Davidson's answer. Here the act of interpretation Davidson has in mind seems one of systematically associating each sentence of L to its corresponding truth-condition. This is what it takes to interpret.[11]

[11] Since the truth-condition on the right hand side of the T-sentence is typically stated in a language familiar to us (say English), the interpretation of a foreign language via its theory of truth presupposes we already have interpreted our own language. But can we make that assumption? How do we go about interepreting our own language?

Lewis

David Lewis formulates the problem of radical interpretation very differently:

> Given P, the facts about Karl as a physical system, solve for the rest. (p. 321)

The 'rest' is Karl's beliefs and desires (attitudes) and what Karl means when he utters sentences. So the rest are beliefs, desires and meanings.

Lewis preferred approach to the problem is quite different from Davidson (p. 341). By relying on the knowledge of all physical facts, derive Karl's beliefs and desires. Next, assign meanings to Karl's utterances assuming that he is truthful, given his beliefs and desires. Finally, carry out the translation into our own language.[12]

[12] How does this approach compare to Davidson's? What different assumptions are made here? Is Lewis making the problem of radical interpretation too easy, by taking all physical facts as the starting point?