Metaphysics Handouts
Revised
05 Nov 2026
10:17

Metaphysics // Fall 2026

Handout 11

Rule-following and theoretical linguistics: Wright

THE ARGUMENT IN OUTLINE. Nobody has ever said this to you (I hope!):

(11-1) Mahler never forgave Wagner for the motorbike.

Still, you knew at once that it's English and what it says. Generative linguistics explains the feat by a grammar you've internalised, a finite system of rules that fixes the meaning of every sentence in advance, whether or not anyone ever considers it.

Wittgenstein's *PI* keeps returning to an image it distrusts: the rule as a rail laid to infinity, tracing out the proper course of a practice independently of anyone's judgement. So the question is whether a cognitive science of language can survive the rule-following considerations, and if it can, what it's a science of. 178

Here is an outline of Wright's argument:

(1) The Central Project is to explain our agreement about what a new sentence says by covert cognitive operations on an internalised grammar. 171
(2) That needs facts about what new sentences mean, and a picture of those facts as fixed in advance by the rules of the language, which the speaker's competence tracks. The picture is the rail. 179
(3) Kripke's sceptic denies the facts. Wright thinks he can be answered, so the facts are safe, though not by Chomsky's answer. 176
(4) Wittgenstein's own negative themes undermine the rail. A rule that outruns every explanation of it could guide me in a new case only through an interpretation (⇒ regress) or through an intuition of its requirements (⇒ unintelligible). So judgements about the rule's requirements track nothing. 185
(5) Rules survive, sustained by our agreement, but our agreeing on a verdict isn't what makes it right. So what does make a judgement of content true? Kripkenstein: Nothing. Baker and Hacker: Bad question! The consensualist (=Wright according to McDowell?): Our agreement. Wright: Best opinion. 211
(6) The first three leave nothing for a cognitive science to explain. So the Project survives only if content is objective yet fixed by best opinion, as colour arguably is. Then it describes no tracking mechanism but the conditions under which a speaker's judgement of content is best. 212
(7) Whether that can be done, and whether a grammar has any place among those conditions, Wright leaves open. 213

Question 1. Which step seems weakest to you? strongest?

THE SCEPTIC ANSWERED. The Project explains the feat in cognitive-psychological terms: we stand in some relation to a finite body of information from which theorems about each new string can be derived. Suppose Kripkenstein is right and talk of meaning states nothing true or false. Then there's nothing true to say about what (11-1) says, so nothing to know, and linguistic creativity is a myth. Our agreement in reacting to (11-1) would need no deeper explanation than a shared disposition to find the same things funny would, if being funny were no matter of fact. 171 172

W-sceptic's question is what fact about me makes it the case that I meant addition, and not quaddition, by 'plus'. Chomsky's answer: the sceptic looks in the wrong place. Which rule you followed is a theoretical question, answerable to an open-ended pool of past, present and future evidence, not just to your old behaviour and conscious life. 174

Wright objects: like the dispositional answer, this makes a 'total mystery' of our non-inferential, first-person knowledge of what we mean and intend. 175

In any case, the sceptical argument is flawed. When the sceptic asks me to recall something in my past mental life that constituted my meaning addition, he won't accept the reply that I recall precisely that. So put, the challenge can't be met, but only because it prejudges the ordinary notion of meaning. On the ordinary notion, it seems, we know our meanings non-inferentially. This 'flat-footed' reply may look like a loss of the problem. 176 177

The real problem is to see how the flat-footed response can be correct. Meaning hovers between two paradigms, inner episodes like a headache and dispositions of character like patience, and is pulled towards both at once. We want a notion of meaning that has both features, avowability and 'theoreticity', without assimilating it to either. 178

Question 2. Should a theory of what someone means answer to that person's authority about it at all? If you say no, with Chomsky, what's left of the difference between meaning something and being disposed to behave?

THE RAIL UNDERMINED? Suppose the sceptic is answered and meaning is factual. The rail is still built into the Project's picture of language. On that picture, language is a syntactico-semantic mechanism, and our largely unconscious knowledge of it lets us compute the content it bestows on each sentence in advance of any response of ours. Wright makes the kinship vivid by rewriting Wittgenstein's remark on the mathematical machine driven by the rules themselves: a language-machine, driven by the rules of the language, makes a sentence grammatical and meaningful even if no one ever considers it. 179

Wittgenstein came to regard the conception as deeply misconceived. But why, and for a reason that doesn't end where Kripke's sceptic ends, with no requirements of rules at all? His discussion in §§185–219 is epistemology, not doubt about the reality of rules. It asks whether we track a rule's requirements by some intuitive or interpretative power. Four themes interlock in it, three of them negative. 180 181

(I) My understanding of a rule doesn't exceed what I can explain. The contrary thought is the cornerstone of the rail: a finite explanation admits many interpretations yet usually secures understanding, so some 'something' beyond it must have got across. 182

(II) The most basic rule-governed responses are better thought of as a kind of decision than as an intuition of the rule's requirements. For we've no idea what the rule's own direction would be if all our intuitions went off track. 183

(III) If grasping a rule is having something in mind, how do I recognise what it requires? 184

Example 3. You write down a long initial segment of an arithmetical series. I come to have an idea of what the 'right rule' is. Somewhere far beyond the segment, having the rule in mind is no help by itself. I can have a formula in mind without knowing what it means, so I must interpret it. But any choice fits the rule under some interpretation. Hence an interpretation helps only if it's correct, and knowing the correct interpretation is either 'a piece of patter' for knowing the rule or a further item in mind that raises the same problem again. 185

So a rule that transcends its explanation faces a dilemma. If I have to interpret it, the regress starts. If I don't, I must have unmediated intuitional contact with its requirements, which the second theme undermined.

This isn't Kripke's paradox. His sceptic asks how I can tell which rule I used to follow, and aims at the existence of rules. Wittgenstein asks how a rule already in mind can take me to a new case, and aims at a misunderstanding of rule-following knowledge, which he apparently thought correctable without doubting that rules are real. 187

Question 4. Can Wittgenstein reject the autonomy of rules and keep their reality? Say what a rule would be that isn't a rail.

THE FOURTH THEME: AGREEMENT. The fourth theme is a conception of rules that drops their spurious autonomy, yet leaves room for normativity. The foundations of language, and of every rule-governed institution, lie not in our having internalised the same explanation-transcendent rules, but in primitive dispositions to agree in judgement and action. The agreement sustains the rules, and the requirements they impose owe their existence to it. 188

The difficulty is to prevent the fatal simplification that a rule requires whatever we take it to require. This would negate the very idea of a normative requirement (§241). The 'official' Wittgensteinian line (Baker and Hacker) says there's nothing more to find. Constitutive questions are to be mistrusted, and the mythology of rails is to be exposed, not replaced. 189

This line, Wright argues, is as bad for the Project as W-sceptic. Facts about meaning are safe on it. But the Project wants a deep explanation of our agreement about them in terms of covert cognitive operations, and the whole lesson of the negative themes was that not all agreement in judgement can be the product of such operations. 190

Question 5. Taking in a new sentence like (11-1) is often, even usually, immediate and spontaneous. If agreement on the next place in a series neither needs nor admits a hidden cognitive account, what is special about agreement on what (11-1) says?

BEST OPINION. Wright wants to walk between the drops and have a response that's neither Kripkean nor official. His idea is not to abandon constitutive questions too soon. How can judgements lack a tracking epistemology and still be objective, still answer to something other than our actual dispositions? 191

Colour is his model. An old thought, Locke's among others, says that in judging colours we respond to our own sensibility rather than to anything in nature, so that colour isn't real. Wright thinks the conclusion an overreaction. His test is built to keep what's right in the thought without the conclusion.

The order-of-determination test concerns the relation between truth and best judgements, those made under cognitively ideal conditions. Judgements pass the test, and are 'extension-reflecting', if there's content to the idea of best judgements tracking the truth. They fail it, and are 'extension-determining', if there's no distance between being true and being best, so that best opinion settles what the truth predicate applies to. Failing isn't trivial. For any judgement whatever, it's a priori that if I judge under conditions which have everything it takes to get the matter right, things are as I take them to be. So the ideal conditions must be spelled out substantially, not as whatever it takes, and without presupposing what the concept applies to. And there must be no rival account of what makes such judgements true, from which the fit between best opinion and truth would follow anyway. 192 194 196

193 *Example* 6. A flat surface, one foot square, and the judgement that it's uniformly royal blue. The conditions: full view, good light, an attentive thinker with normal vision and the concept, and no doubt about any of this. 192

Shape differs in two ways. A visual verdict on a three-dimensional shape needs several observations from several positions. So the conditions must guarantee that the shape didn't change meanwhile, which already calls for a determinant of shape independent of the thinker's opinion. A team looking from all sides at once would get round that. The deeper difference is that whether a solid is pear-shaped is settled by what cutting or measuring it would show, not by how it looks. When the two disagree the measurements win, which is why the Müller-Lyer figure counts as an illusion. It's a lesson of experience, not an a priori truth, that looks are a reliable guide to shape in our world at all. So no conditions on a visual verdict about shape can be both a priori and substantial. Shape arguably passes the test, then, and the test isn't empty: not every judgement comes out fixed by best opinion. 197 198

Colour comes out as a mix of subjectivity and objectivity. Our best opinions don't track it. But it's a perfectly objective question what best opinion would deliver in a given case, and a majority or a whole community can be out of accord with it. 199

So the line between colour and shape isn't Locke's. On his, colour was unreal and shape real. Here both are real, and the line runs between two kinds of judgement, not two kinds of quality. A single quality can even sit on both sides: shape judged by eye reflects, and whether shape judged by measurement does too is left open. 198

Why does the dialectic need this third status? Because the sceptic and the Project alike assume there are only two. Either a judgement of meaning answers to a fact fixed independently of it, or there's no fact. The sceptic looks for the independent fact, finds none, and concludes there's nothing to find. The Project has the grammar fix what (11-1) says and the speaker track it. Colour shows a position between the two, and the rest of the paper moves judgements of meaning into it.

Question 7. Is there really a third position between tracking an independent fact and agreeing on whatever we please? Say what a whole community's being wrong about a colour would consist in. And is Wright fair to Locke, for whom colours were real powers in objects to produce ideas in us?

Avowals. Avowals (=self-ascriptions of intention) are Wright's second case. The Cartesian takes the authority of avowals for an infallible inner perception. For Wittgenstein, that's a grammatical misunderstanding. If best opinion *determines*, rather than *reflects*, the facts about one's intentions, then 'inner tracking' is out of place. 200

So the test is run for intention: which conditions ensure a priori that a subject believes she has a given intention if and only if she does? The list looks short: the subject has the concepts, attends to the question, and isn't self-deceived. The third can't be stated except as 'whatever it takes', since 'self-deception' has to cover any condition, motivated or chemical, that leads to error about one's intentions. But it's 'positive-presumptive': by the grammar of intention, a subject counts as not self-deceived unless there's definite evidence that she is. So her opinion under the two substantial conditions is extension-determining by default, defeated only by such evidence. 201 202

If judgements about one's own meanings are likewise provisionally extension-determining, there's a 'perfect answer' to W-sceptic. Challenged about what I meant by 'plus', I needn't dig up some meaning-constituting fact buried in my past. My present opinion suffices, together with the a priori reasonableness, failing evidence to the contrary, of taking that opinion to be best. The sceptical argument engages the conception of meanings as items our best opinions reflect, not their reality as such. In the simplest terms: the sceptic looked for something my judgement reflects, found nothing, and concluded there's nothing, and colour shows the last step is bad. 206

The model of intention also gives the notion of meaning that the section on the sceptic wanted, avowable and theoretical at once. My avowal is authoritative by default, which is the side Chomsky's answer and the dispositional one made a mystery of. And my behaviour can override it, which is the side the flat-footed reply left unexplained. 204

Question 8. Does the answer beg the sceptic's question, 'What does make your belief true?'

What the Project could be. The comprehending response to a new utterance is a rule-informed judgement that can be extricated from the rest. Someone with no understanding of the language could share a set of beliefs about (11-1), and on top of them there's a belief about what, given those, has been said. What makes that belief true? 209

Platonism, a fact independent of any opinion of ours which the speaker tracks, could sustain the Project. But it's the tracking picture the negative themes undermined. If it's the only way to sustain the Project, there's a real collision with Wittgenstein, and one might well side with Wittgenstein. The official and the Kripkean views undermine the Project, and so does consensualism, by a neat dilemma. If agreement constitutes correctness, nothing calls for a cognitive account of the agreement. And a cognitive account would at once restore the idea of what our skills would deliver independently of our actual responses, that is, room for a consensus to be misplaced. Consensualism denies that room, whereas the competence/performance distinction needs it. 211 212

That leaves one way to understand the Project: semantic content must be objective, but best-opinion-determined, or 'secondary' in Lockean terms. If so, the Project no longer describes a tracking accomplishment. It elaborates the conditions under which a speaker's judgement of the content of (11-1) is best. Competence and performance survive as the gap between best opinion and actual opinion. But the grammar (=UG) is demoted: it can no longer be what makes the judgement true, only at most part of what makes the judge's opinion best.

And this is where the Project meets Wittgenstein head on. The Chomskyan account of why ideal judges get the content right, that the grammar generates it and the judge tracks it, is just the rival account that the test's last condition excludes. So content can be fixed by best opinion only if the negative sceptical themes have already dismissed the rail. Whether the test can be refined to apply to judgements of content, and whether best opinion so understood would allow, let alone require, the apparatus of theoretical linguistics, is open: we're 'a long way from knowing'. 213



Question 9. Try writing the conditions for a best judgement of what (11-1) says. Does any of them mention a grammar, or is a Chomskyan grammar now a different project under the same name?