Munich Center for Mathematical Philosophy
Ludwig-Maximilians-Universität München
Geschwister-Scholl-Platz 1
Munich 80539
Germany
neil.dewar@lrz.uni-muenchen.de

On translating between logics

Neil Dewar

January 9, 2018

Abstract

In a recent paper, Wigglesworth claims that syntactic criteria of theoretical equivalence are not appropriate for settling questions of equivalence between logical theories, since such criteria judge classical and intuitionistic logic to be equivalent; he concludes that logicians should use semantic criteria instead. However, this is an artefact of the particular syntactic criterion chosen, which is an implausible criterion of theoretical equivalence (even in the non-logical case). Correspondingly, there is nothing to suggest that a more plausible syntactic criterion should not be used to settle questions of equivalence between different logical theories; such a criterion (which may already be found in the literature) is exhibited and shown to judge classical and intuitionistic logic to be inequivalent.

Wigglesworth (2017) argues that the logical anti-exceptionalist—someone who "takes logical theories to be continuous with scientific theories"[1]—must pay attention to the question of when two logical theories are equivalent, in the same way that philosophers of science have long asked the question of when two scientific theories are equivalent. To do so, he proposes taking criteria of equivalence that have been discussed in the philosophy of science, and adapting them to the context of comparing systems of logic. Whilst the overall project of comparing logical systems for equivalence or translatability is definitely an important and worthwhile project, I have some reservations about the particular conclusions Wigglesworth draws—especially, his contention that syntactic criteria for equivalence are less appropriate in the logical case than in the scientific case. This article articulates these concerns.

First, a brief summary of the syntactic criteria discussed in the philosophy of science (although these criteria originated, and are still much used, in the context of mathematics and logic).[2] Suppose that L_1 and L_2 are two languages, but languages of the same logic; that is, that L_1 and L_2 differ only over their nonlogical vocabulary. An *interpretation* from L_1 to L_2 is an arity-preserving[3] map τ from the formulae of L_1 to those of L_2 which commutes with the logical constants, so that $\tau(\neg\phi) = \neg\tau(\phi)$, $\tau(\phi \wedge \psi) = (\tau(\phi) \wedge \tau(\psi))$, etc. Alternatively (and more relevantly to what will come later), we can consider an arity-preserving map from the nonlogical vocabulary of L_1 to formulae of L_2,[4] and take an interpretation to be the unique extension of such a map to a map from formulae of L_1 to formulae of L_2, generated by requiring commutation.

Let a *theory* be a set of sentences, closed under the ambient logic. For any set of sentences S, let $T_S^{\mathcal{L}}$ denote the logical closure of S under the logic $\mathcal{L}$; to reduce notational clutter when writing about logics $\mathcal{L}_1$, $\mathcal{L}_2$, etc., I'll write T_S^i to denote the closure of S under $\mathcal{L}_i$. Given two theories T_1 and T_2, formulated in the languages L_1 and L_2 respectively, an interpretation $\tau : L_1 \to L_2$ is a *translation* from T_1 to T_2 if, for any L_1-sentence ϕ, if $\phi \in T_1$ then $\tau(\phi) \in T_2$. Thus, translations are interpretations which map consequences to consequences. And we say that T_1 and T_2 are *mutually interpretable* if there is both a translation $\tau : T_1 \to T_2$ and a translation $\sigma : T_2 \to T_1$.

Now, these ideas cannot immediately be applied to the comparison of different logical theories: since these theories (in general) use different logical constants, it is

[1] (Wigglesworth, 2017, p. 1)

[2] For further discussion of such criteria, see Barrett and Halvorson (2016).

[3] Here, "arity-preserving" should be taken to mean not just that it maps n-place formulae to n-place formulae, but that if exactly the variables $\xi_1, \ldots, \xi_n$ occur free in ϕ, then exactly those variables occur free in $\tau(\phi)$; moreover, it is assumed that τ commutes with uniform substitution of variables.

[4] Where the same technicalities apply as in the previous footnote.

not appropriate to require that interpretations commute with all the constants. So Wigglesworth proposes a weakening, whereby one requires only that interpretations commute with negation: i.e., that $\tau(\neg\phi) = \neg\tau(\phi)$. Call this a *W-interpretation*; let us say that a W-interpretation between two theories is a *W-translation* if it maps consequences to consequences; and let us say that T_1 and T_2 are *mutually W-interpretable* if there are W-translations in both directions. The stated criterion of equivalence for logics is then the following: that two logics $\mathcal{L}_1$ and $\mathcal{L}_2$ are *W-equivalent* if, for any set of sentences S, T^1_S is mutually W-interpretable with T^2_S.

However, this definition is problematic in a number of ways. First, the construction Wigglesworth uses to compare classical and intuitionistic logic does not suffice to show that they are W-equivalent. The problem is with the translation from classical to intuitionistic logic, which Wigglesworth claims is provided by the Gödel-Gentzen mapping γ: contra Wigglesworth, it is not the case (at least for first-order logic) that "for any S and any ϕ, if $\phi \in T^{\mathcal{C}}_S$, then $\gamma(\phi) \in T^{\mathcal{I}}_S$."[5] Rather, for any S and any ϕ, if $\phi \in T^{\mathcal{C}}_S$, then $\gamma(\phi) \in T^{\mathcal{I}}_{\gamma(S)}$, where $\gamma(S) = \{\gamma(\psi) | \psi \in S\}$. A counterexample to the original claim is provided by letting $S = \{\neg\forall x Px\}$: then $\neg\forall x Px$ is (trivially) a classical consequence of S, but its Gödel-Gentzen translation $\neg\forall x\neg\neg Px$ is not an intuitionistic consequence of S.

Second, the criterion of W-equivalence permits us to change what interpretation is being used, depending on which set of sentences we are considering the logical closure of. That is, it would fall within the definition to offer one way of translating the formulae of $\mathcal{L}_1$ into those of $\mathcal{L}_2$ when considering the closure of the set S, and another when considering the closure of the set S'. Intuitively, this is too weak to provide a robust sense of equivalence: we would expect an interpretation between two logics to be chosen "once and for all".

Finally, the notion of a W-interpretation seems inappropriate: it is both too restrictive and too permissive. It is too restrictive, because it is unable to consider cases where we want to non-trivially interpret the negation symbol of one logic in terms of another. For example, suppose that we wish to capture the sense in which a classical logic using $\neg$ for negation is equivalent to one using $\sim$; or the sense in which a uniform replacement of $\neg$ by $\neg\neg\neg$ is a translation of classical logic into itself. And it is too permissive, because it puts no restrictions of uniformity on the translations of formulae featuring other connectives. For example, a W-interpretation could map $(P \wedge Q)$ to $(F \vee G)$ but $(Q \wedge P)$ to $\neg(F \to (G \wedge H))$, say.

[5] (Wigglesworth, 2017, p. 4)

We begin with the third problem; its resolution is already to be found in the literature on translating between logics.[6] The key idea is that whereas interpretation within a fixed background logic involved mapping formulae to one another, interpretation between logics will require us to map *schemata* to one another. Let us say that a schema-string is a formula featuring metalinguistic variables, such as $(\phi \vee \psi)$ or $\forall\xi\phi$. We will take a schema to be the map from formulae (and possibly variables) to formulae that a schema-string encodes: for instance, the first schema-string encodes a map taking $\langle P, (Q \rightarrow R)\rangle$ to $(P \vee (Q \rightarrow R))$, whilst the latter encodes a map taking $\langle y, Fy\rangle$ to $\forall yFy$. To indicate that schemata are the maps rather than the strings, I'll use lambda notation: so the first schema-string above encodes the map $\lambda\phi\lambda\psi.(\phi \vee \psi)$, whilst the latter encodes the map $\lambda\xi; \lambda\phi.\forall\xi\phi$.[7]

Let us say that the *logical vocabulary* for a language consists of the logical constants for that language: we will only consider the case where the logical vocabulary consists of a set of connectives (each with a certain arity) or quantifiers (all assumed to only take a single variable and a single formula). I won't count the stock of variables as part of the logical vocabulary; I'll just assume that all the languages we consider are using the same stock of variables. Once we specify a non-logical vocabulary for the language, i.e. a stock of predicates and function-symbols, we are able to generate the set of formulae of the language in the standard recursive fashion. Since we will make use of them below, I here state the recursive clauses appropriate to any logical vocabulary:

- Given any n-place connective C in the logical vocabulary, and any formulae $\psi_1, \ldots, \psi_n$, $C\psi_1 \ldots \psi_n$ is a formula.

- Given any quantifier $\wp$, any variable ξ, and any formula ψ, $\wp\xi\psi$ is a formula.

I will assume that we are only considering languages with the same non-logical vocabulary. In principle, it should be reasonably straightforward to extend what I do here to languages with different non-logical vocabularies (by combining the material here with the standard work on interpretation and translation rehearsed above), but for simplicity's sake I forebear from doing so.

[6] See, in particular, Pelletier and Urquhart (2003) and references therein; cf. Barrett and Halvorson (2016) and McSweeney (2016).

[7] Strictly speaking, it's somewhat ambiguous what map a schema-string encodes unless we assume some ordering on the metalinguistic variables: for instance, one could also claim that the schema-string $(\phi \vee \psi)$ encodes a map taking $\langle P, (Q \rightarrow R)$ to $((Q \rightarrow R) \vee P)$. But this will not be an issue in practice, so I gloss over worrying about it.

Now, given two languages L_1 and L_2, a *schematic interpretation* $\mathrm{T}_\bullet$ from L_1 to L_2 consists of the following data:

- A one-place schema T_α of L_2
- For every n-place connective C of L_1, an n-formula schema T_C of L_2
- For every quantifier $\wp$ of L_1, a one-variable and one-formula schema $\mathrm{T}_\wp$ of L_2, with the property that for any variable ξ and any formula ϕ, ξ does not occur free in $\mathrm{T}_\wp(\xi; \phi)$

A schematic interpretation $\mathrm{T}_\bullet$ determines a map $\tau : L_1 \to L_2$, by recursion:

- If ϕ is atomic, then $\tau(\phi) = \mathrm{T}_\alpha(\phi)$
- If $\phi = C\psi_1 \dots \psi_n$, then $\tau(\phi) = \mathrm{T}_C(\tau(\psi_1), \dots, \tau(\psi_n))$
- If $\phi = \wp\xi\psi$, then $\tau(\phi) = \mathrm{T}_\wp(\xi; \tau(\psi))$

Call such a map an *interpretation**. And given an L_1-theory T_1 and an L_2-theory T_2, closed under logics $\mathcal{L}_1$ and $\mathcal{L}_2$ respectively, say that an interpretation* $\tau : L_1 \to L_2$ is a *translation** from T_1 to T_2 if it maps consequences to consequences: that is, if $\phi \in T_1$ implies $\tau(\phi) \in T_2$. And, again, we can say that T_1 and T_2 are mutually interpretable* if there is a translation* from T_1 to T_2, and from T_2 to T_1.

In order to avoid the first and second problems discussed above, we now proceed as follows. Given two logics $\mathcal{L}_1$ and $\mathcal{L}_2$, let us say that an interpretation* τ is a translation* from $\mathcal{L}_1$ to $\mathcal{L}_2$ if, for any set of L_1-sentences S, τ is a translation* from T^1_S to $T^1_{\tau(S)}$, where $\tau(S) := \{\tau(s) | s \in S\}$. And let us say that $\mathcal{L}_1$ and $\mathcal{L}_2$ are mutually interpretable* if there is a translation* from $\mathcal{L}_1$ to $\mathcal{L}_2$, and a translation* from $\mathcal{L}_2$ to $\mathcal{L}_1$. Mutual interpretability* does not suffer from the three problems outlined above for W-equivalence; I therefore propose to use it as a suitably debugged version of W-equivalence.

In these terms, the relevant result for Wigglesworth's argument is that classical and intuitionistic logic are mutually interpretable*. The translation* from intuitionistic to classical logic is provided by the identity mapping (since any intuitionistic consequence is always a classical consequence), and the translation* in the other direction is—as mentioned already—provided by the Gödel-Gentzen mapping γ. This mapping is generated by the following schematic interpretation, $\Gamma_\bullet$: for any formulae ϕ

and ψ, and any variable ξ, the values of the schematic interpretation of the standard first-order logical vocabulary are

$$\Gamma_{\alpha}(\phi) = \neg\neg\phi \tag{1}$$

$$\Gamma_{\neg}(\phi) = \phi \tag{2}$$

$$\Gamma_{\wedge}(\phi, \psi) = (\phi \wedge \psi) \tag{3}$$

$$\Gamma_{\vee}(\phi, \psi) = \neg(\neg\phi \wedge \neg\psi) \tag{4}$$

$$\Gamma_{\rightarrow}(\phi, \psi) = (\phi \rightarrow \psi) \tag{5}$$

$$\Gamma_{\forall}(\xi; \phi) = \forall\xi\phi \tag{6}$$

$$\Gamma_{\exists}(\xi; \phi) = \neg\forall\xi\neg\phi \tag{7}$$

As is well-known, for any set of sentences S, γ is a translation* from $T^{\mathcal{C}}_{S}$ to $T^{\mathcal{I}}_{\gamma[S]}$;[8] thus, γ is a translation* from $\mathcal{C}$ to $\mathcal{I}$. So there are translations* in both directions, and hence $\mathcal{C}$ and $\mathcal{I}$ are mutually interpretable*, and hence equivalent according to that criterion.

Wigglesworth goes on to show that this is not the case for alternative criteria of equivalence between logics: specifically, he shows that the category of models of classical logic is not categorically equivalent to the category of models of intuitionistic logic.[9] He claims that the difference in verdicts is due to the criterion above being a syntactic criterion (i.e. one formulated in terms of translations between sentences) whereas the category-theoretic criterion is a semantic criterion (i.e. one formulated in terms of comparisons between models). Given that classical and intuitionistic logic are *not* intuitively equivalent, he concludes that "though there are two general approaches—one syntactic and one semantic—to theoretical equivalence in the philosophy of science, the logical anti-exceptionalist should prefer the semantic approach."[10]

But the syntactic criteria discussed above (i.e., W-equivalence and mutual interpretability*) are motivated by the claim that mutual interpretability is a serious contender as a criterion of equivalence in the philosophy of science. And that claim is dubious, for it is implausible to treat mutually interpretable theories as equivalent. For instance, consider the two theories T_1 and T_2, generated by taking the (classical)

[8] (Troelstra and van Dalen, 1988, p. 58, Theorem 3.5)

[9] I'm being a little loose here in talking about *the* category of models of classical or intuitionistic logic, since there's a bit of leeway in how exactly one characterises such a category (e.g. whether it should have homomorphisms or elementary embeddings as morphisms); but for our purposes, these details won't matter.

[10] (Wigglesworth, 2017, p. 7)

closure of the following sets of sentences:

$$S_1 = \{Px \vee \neg Px\}$$
$$S_2 = \{Fx \vee Gx\}$$

Intuitively, these theories are not equivalent: T_1 is trivial, whereas T_2 is not, and it is implausible that any trivial theory should be regarded as saying the same thing as any non-trivial theory. Yet they are mutually interpretable. The only consequences of S_1 are logical validities, so (say) mapping P to F will map every consequence of S_1 to a consequence of S_2 (since logical validity is preserved by such a map). Contrariwise, mapping F to P and G to $\neg P$ will map every consequence of S_2 to a logical validity, which is thereby a consequence of S_1.

Given Wigglesworth's discussion, this may seem somewhat surprising. For he argues that the criterion of mutual equivalence is closely related, at least in the context of classical logic, to the criterion of definitional equivalence—and, as he says, "Definitional equivalence is an interesting and powerful notion of theoretical equivalence in its own right."[11] Certainly, definitional equivalence has received significant attention from philosophers of science.[12] But given the example above, it seems that it could not be such an interesting or powerful notion if, as Wigglesworth claims, it is essentially equivalent to mutual interpretability: that is, if it is the case that "if [two] theories have disjoint signatures [i.e. disjoint nonlogical vocabularies], then they are definitionally equivalent iff they are mutually interpretable."[13]

However, the claim just quoted is false. Definitional equivalence entails mutual interpretability, but the converse does not hold (even when the signatures are disjoint): it is straightforward, for instance, to show that the theories T_1 and T_2 above are not definitionally equivalent. The results that Wigglesworth refers to, Theorems 1 and 2 of Barrett and Halvorson (2016), show instead that *intertranslatability* is necessary and sufficient for definitional equivalence (given the disjointness of the signatures). Two theories T_1 and T_2 are intertranslatable when there exist a pair of translations $\tau : T_1 \to T_2$ and $\sigma : T_2 \to T_1$, *and which are such that* for any L_1-formula ϕ and any

[11] (Wigglesworth, 2017, p. 3)

[12] See, especially, Glymour (1970) and Glymour (1977); for an analysis of how the (syntactic) criterion of definitional equivalence relates to semantic notions, see de Bouvère (1965).

[13] (Wigglesworth, 2017, p. 3)

L_2-formula ψ,

$$T_1 \vdash \phi \leftrightarrow \sigma(\tau(\phi)) \tag{8}$$

$$T_2 \vdash \psi \leftrightarrow \tau(\sigma(\psi)) \tag{9}$$

Thus, intertranslatability is a (substantial) strengthening of mutual interpretability, requiring not only that there be translations between the two theories, but that the compositions of these translations map formulae to formulae which are equivalent modulo the ambient theory.

Introducing the notion of intertranslatability also suggests a natural strengthening of (the amended version of) the syntactic criterion discussed by Wigglesworth.[14] Let us say that two logics $\mathcal{L}_1$ and $\mathcal{L}_2$, in languages L_1 and L_2 respectively, are *intertranslatable** if there are translations* $\tau : \mathcal{L}_1 \to \mathcal{L}_2$ and $\sigma : \mathcal{L}_2 \to \mathcal{L}_1$ such that, for any $\mathcal{L}_1$-theory T_1 and L_1-formula ϕ,

$$T_1, \phi \vdash_1 \sigma(\tau(\phi)) \tag{10}$$

$$T_1, \sigma(\tau(\phi)) \vdash_1 \phi \tag{11}$$

and for any $\mathcal{L}_2$-theory T_2 and L_2-formula ψ,

$$T_2, \psi \vdash_2 \tau(\sigma(\psi)) \tag{12}$$

$$T_2, \sigma(\tau(\psi)) \vdash_2 \psi \tag{13}$$

where $\vdash_i$ indicates the consequence relation in $\mathcal{L}_i$.

As one would hope, it turns out that classical and intuitionistic logic are not intertranslatable*.[15]

Proof. Up to logical equivalence, there are only three non-trivial one-place schemata in intuitionistic logic ($\lambda\phi.\phi$, $\lambda\phi.\neg\phi$ and $\lambda\phi.\neg\neg\phi$). Thus, there are only three possible interpretations of each of the classical atomic and negation schemata in intuitionistic logic. This means nine interpretation-schemata from (the pure negation fragment of) classical logic to intuitionistic logic; however, it is easy to show that six of these do

[14] cf. Pelletier and Urquhart (2003)'s notion of "translational equivalence"; it is also shown there that this criterion coincides with a natural criterion of definitional equivalence between logics.

[15] Note that the proof also shows that $\mathcal{C}$ and $\mathcal{I}$ fail to satisfy certain natural weakenings of intertranslatability*: for example, a criterion requiring merely that $\sigma \circ \tau$ takes L_1-formulae to $\mathcal{L}_1$-equivalent formulae, and that $\tau \circ \sigma$ takes L_2-formulae to $\mathcal{L}_2$-formulae (effectively, replacing equivalence modulo an arbitrary $\mathcal{L}_i$-theory by equivalence modulo $T^i_\varnothing$).

not generate translations*, even over this fragment. This leaves the interpretation-schemata

$$\mathrm{T}^1_\alpha(\phi) = \neg\neg\phi \qquad \mathrm{T}^1_\neg(\phi) = \neg\phi \tag{14}$$

$$\mathrm{T}^2_\alpha(\phi) = \neg\phi \qquad \mathrm{T}^2_\neg(\phi) = \neg\neg\phi \tag{15}$$

$$\mathrm{T}^3_\alpha(\phi) = \neg\phi \qquad \mathrm{T}^2_\neg(\phi) = \phi \tag{16}$$

(The first of these is, of course, the pure-negation part of the Gödel-Gentzen translation.) Similarly, there are only two non-trivial one-place schemata in classical logic ($\lambda\phi.\phi$ and $\lambda\phi.\neg\phi$), and hence only two ways to interpret the intuitionistic atomic and negation schemata in classical logic, and hence only four interpretation-schemata. Of these, two do not generate translations* even over the pure negation fragment, leaving

$$\Sigma^1_\alpha(\phi) = \phi \qquad \Sigma^1_\neg(\phi) = \neg\phi \tag{17}$$

$$\Sigma^2_\alpha(\phi) = \neg\phi \qquad \Sigma^2_\neg(\phi) = \phi \tag{18}$$

It is then just a matter of computation to show that for any pair of interpretation-schemata, their composition (in one direction or the other) will not always return formulae to logically equivalent formulae (under classical or intuitionistic logic, as appropriate). For example, the composition of (14) with (17) fails to return the formula P to an intuitionistically equivalent formula: P is mapped by (17) to P, and thence by (14) to $\neg\neg P$. □

I conclude that these considerations do not show that the logical anti-exceptionalist must prefer semantic to syntactic criteria for equivalence: it is not the case that "the standard syntactic approach in terms of intertranslatability forces the anti-exceptionalist to say that classical logic and intuitionistic logic are equivalent logical theories."[16] On the contrary, the standard approach, when extended to the context of comparing logics, delivers the intuitively correct verdict that classical and intuitionistic logic are inequivalent.[17]

[16] (Wigglesworth, 2017, p. 7) Note that Wigglesworth here is using "intertranslatability" synonymously with "mutual interpretability".

[17] I'm grateful to Thomas Barrett, Laurenz Hudetz, John Wigglesworth and two anonymous referees for comments on and discussions of this material.

References

Barrett, T. W. and Halvorson, H. (2016). Glymour and Quine on Theoretical Equivalence. *Journal of Philosophical Logic*, 45(5):467–483.

de Bouvère, K. (1965). Synonymous Theories. In Addison, J. W., Henkin, L., and Tarski, A., editors, *The Theory of Models: Proceedings of the 1963 International Symposium at Berkeley*, Studies in Logic and the Foundations of Mathematics, pages 402–406. North-Holland, Amsterdam.

Glymour, C. (1970). Theoretical Realism and Theoretical Equivalence. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, pages 275–288.

Glymour, C. (1977). The Epistemology of Geometry. *Noûs*, 11(3):227–251.

McSweeney, M. M. (2016). An Epistemic Account Of Metaphysical Equivalence. *Philosophical Perspectives*, 30(1):270–293.

Pelletier, F. J. and Urquhart, A. (2003). Synonymous Logics. *Journal of Philosophical Logic*, 32(3):259–285.

Troelstra, A. S. and van Dalen, D. (1988). *Constructivism in Mathematics: An Introduction (Vol. 1)*. Number 121 in Studies in Logic and the Foundations of Mathematics. Elsevier, Amsterdam.

Wigglesworth, J. (2017). Logical anti-exceptionalism and theoretical equivalence. *Analysis*. Forthcoming.