The Deduction Theorem
(before and after Herbrand)

CURTIS FRANKS

1. Preview

Attempts to articulate the real meaning or ultimate significance of a famous theorem comprise a major vein of philosophical writing about mathematics. The subfield of mathematical logic has supplied more than its fair share of case studies to this genre, Gödel's (*1931*) incompleteness theorem being probably the most frequent exhibit. This note is about the Deduction Theorem—another result from mathematical logic, of roughly the same vintage. I aim to make clear, in the simplest possible terms, what the theorem says in the several guises it has taken, how to prove it in each associated framework, and especially the role it has played in logic's development over nearly two centuries.

But do not expect the theorem to submit, here, to anything like a final analysis. I intend the exercise to serve as an ancillary to a thesis contrary to such ambitions: that the meaning of important mathematics is unfixed—expanding over time, informed by new perspectives and theoretical advances, and in turn giving rise to new developments. I want to avoid any impression that the Deduction Theorem is the sort of thing that can be fully understood. Ideally, familiarity with the ways in which it has taken on new meaning over time will prepare readers for its continued reinvention, even to be the authors of its future embellishments.

Other histories of the Deduction Theorem (*Pogorzelski 1968*, *Porte 1982*, *Czelakowski 1985*) have been written, but our focus will differ from theirs. Rather than document when the theorem was proved for different systems, when its minimal conditions were specified, or how it was generalized from the setting of sentential logic to larger classes of algebraic structures, we will not look beyond the basic phenomenon. My plan is that by taking a second look, and then a third and a fourth, we might extract its several meanings according to the different perspectives from which logicians have encountered it.

2. Classical truth

To begin thinking about the Deduction Theorem, consider first the concept of validity in classical propositional logic. In this setting, we have compound formulas built up recursively from atoms with propositional connectives $\vee$, $\wedge$, $\neg$, and $\supset$. An interpretation of a formula is an assignment of truth values from $\{\mathbf{T}, \mathbf{F}\}$ to its atoms. The truth value of a formula on an interpretation is then determined according to a composition of functions, where each propositional connective is a 'truth function' mapping a truth value (in the case of $\neg$) or pair of truth values (in the other cases) to another truth value. (For example, $\supset$ is interpreted as the 'material conditional' which maps the pair $\langle\, \mathbf{T}, \mathbf{F}\, \rangle$ to **F** and all other pairs to **T**.)

The classical propositional *validities* are the formulas that receive the value **T** on every interpretation. When a propositional formula A is classically valid, this fact is denoted $\models_{\mathrm{CPL}}$ A (here CPL abbreviates classical propositional logic). In his logical writings[1] W. V. O. Quine insisted that for these validity claims and their generalizations as described below, one write $\models_{\mathrm{CPL}}$ 'A' and 'A' $\models_{\mathrm{CPL}}$ 'B', indicating with quotation marks that the turnstile symbol abbreviates ordinary language, whereas one *mentions* but does not *use* the formulas A and B. Preferring legibility and conformity with modern usage, we will not heed Quine's scruples on this matter. But it is important to be aware that Quine is correct that in making a validity claim, one is saying that the formula is valid, not making a compound statement out of the formula as one does with expressions like $\neg$A and A $\supset$ B.

Generalizing the concept of validity, one also writes A $\models_{\mathrm{CPL}}$ B to say that on every interpretation on which A receives the value **T**, so too does B. Even more generally, if Γ is a set of classical propositional formulas, then $\Gamma \models_{\mathrm{CPL}}$ B means that B receives the value **T** on every interpretation on which every formula in Γ does.[2] Thus just as the material conditional A $\supset$ B is the classical analysis of the conditional utterance 'If A, then B', $\Gamma \models_{\mathrm{CPL}}$ B is the classical analysis of the claim that, taken jointly, the formulas in Γ logically imply the formula B. In the case when Γ is empty, this last claim is just that B is logically implied without any assumptions at all, i.e., that B is logically valid in its own right. That is the sense in which classical implication is a generalization of classical validity.

Now a certain similarity between the material conditional of classical logic and the concept of logical implication, classically analyzed, is hard to miss. But many beginners (and, famously, occasional experts like Bertrand Russell) are misled by the similarity to simply identify the two. So even though it is an elementary point, let us make their relationship precise. The claim that, if you arrive at my house before 3pm, then I will give you a ride to the concert can be formalized A $\supset$ B. Whether or not this conditional claim is true on the classical analysis depends only on the truth or falsity of A and B. But even on this analysis, if you arrive at my house before 3pm and I do give you a ride to the concert, so that A $\supset$ B turns up true, it would be wrong to say that A logically implies B. The truth of the claim that A logically implies B does not depend in any way on whether you show up on time or what I proceed to do afterwards. In fact, if A did logically imply B, then there would be no reason for me to

[1] See any of the papers in *Quine 1995*.

[2] We follow the convention of denoting the set containing a single formula A by 'A' and denoting the union of Γ and Δ by 'Γ, Δ'.

assure you of anything with an utterance like $A \supset B$. The utterance would be uninformative, because you could just do some logic in private and determine that since B is what you want and B is implied by A, you need only worry yourself with A, knowing that if you succeed, B is automatic.

So what is the relationship? Assume that Γ, A logically implies B, i.e. $\Gamma, A \models_{\text{CPL}} B$. Then by definition it is not possible to assign truth values to atoms so that each formula in Γ and A all turn up **T** and B turns up **F**. That means that on every interpretation of Γ, A, and B, either at least one formula among Γ and A receives the value **F** or B receives the value **T**. Either way, observe that $\Gamma \models_{\text{CPL}} A \supset B$. Assume conversely that $\Gamma \models_{\text{CPL}} A \supset B$. That means that $A \supset B$ receives the value **T** on every interpretation that assigns **T** to each formula in Γ. For this to be so, there must be no interpretation on which each formula in Γ and A all receive **T** and B receives **F**. Therefore $\Gamma, A \models_{\text{CPL}} B$.

We have just verified the fundamental identity

$$\Gamma, A \models_{\text{CPL}} B \text{ if, and only if, } \Gamma \models_{\text{CPL}} A \supset B. \qquad (1)$$

On the classical analysis, logical implication is the same, not as the *truth* of a conditional statement, but as the *validity* of one. (The same identity holds, and by the same line of reasoning, for classical quantification theory.)

This ‘theorem’ might be too trivial to deserve a name. As is typical for the verification of identities among constructions in an elementary semantic environment, our demonstration just involved verifying that the definitions amount to the same thing. But it does encode some remarkable depth. Neither the classical analysis of logical implication nor the material conditional are without controversy. But they developed in historical isolation from one another. It is not hard to locate advocates of classical implication who reject the truth-functional analysis of conditional expressions, and *vice versa*. That these two analyses are so closely related, however trivial it is to verify, is therefore surprising.

On the other hand, one might have the intuition that logical implication ought to correspond to the validity of a conditional statement, independently of any specification of either concept. This leads to a more abstract understanding of (1), untethered from the classical theory of truth. Such appears to have been Bolzano’s approach to conditionals. In §224 of *Wissenschaftslehre* Bolzano claimed that from the observation that (a) one can ‘deduce’ a proposition from a set of premises one can conclude that (b) it is possible to ‘deduce’ from a subset of those premises a conditional whose antecedent is the conjunction of the remaining sentences from the original premises and whose consequent is the conclusion of the original ‘deduction’. Some commentators (see *van Benthem 1985*, *Šebestik 2016*) have read in Bolzano’s inference from (a) to (b) an anticipation of the identity (2) articulated in the following section. However, Bolzano’s concept of deduction (*Ableitbarkeit*) deals not with formal derivability as understood by modern logicians but with a characteristically semantic notion, the preservation of truth under reassignments of 'ideas' to the terms appearing in sentences.[3] It is clear that his claim is instead closer to (1). But for Bolzano, there was no question of verifying this claim. In his direct treatment of conditionals (§179), he does not provide anything like a full analysis in these terms. Instead, the claim of §224 is his official account of conditional language.

The reason I say that Bolzano's claim is 'close to' rather than identical to (1) is that Bolzano's concept of 'deducibility', differs in important ways from the classical account of logical consequence.[4] Already from this it is clear that the form of identity (1) does not depend essentially on all the details of classical logic. More importantly, we see Bolzano using a version of (1) to define conditionals in terms of his (non-classical) concept of deducibility. Following Bolzano, rather than view (1) as an identity in need of verification, one could just assume it. Then beginning with the classical analysis of logical implication, one could establish *via* (1) the truth-functional analysis of conditionals.

3. Classical inference

The identity between logical implication, on the one hand, and the validity of a corresponding conditional claim, on the other hand, is a template for another putative correspondence. Running parallel to the analysis of logic in terms of truth conditions is an alternative analysis in terms of inference. According to this analysis, the logicality of a formula amounts to its derivability in some predetermined systematic manner, and that of an implication amounts to one formula's systematic derivability from others. One may ask: whenever an implication is verified with a derivation in this manner, is a corresponding conditional claim also necessarily derivable?

Obviously, the clarity and full meaning of this question depends on the specification of the manner of derivation. The earliest articulation of a formal system of logical deduction sufficient to pose this question meaningfully appears in the (*1879*) *Begriffsschrift* of Gottlob Frege. There, Frege introduced the idea of a formal proof system for logic with designated axiomatic formulas and rules of inference. In modern notation, the propositional axioms of *Begriffsschrift* are

1. $\mathrm{A} \supset (\mathrm{B} \supset \mathrm{A})$
2. $(\mathrm{C} \supset (\mathrm{B} \supset \mathrm{A})) \supset ((\mathrm{C} \supset \mathrm{B}) \supset (\mathrm{C} \supset \mathrm{A}))$

[3] Further details about Bolzano's theory of deduction and its relationship to his theory of ground and consequent can be found in *Franks 2014*. The second of these is the theory more closely related to the modern concept of derivability.

[4] Again, see *Franks 2014*.

3. $(B \supset A) \supset (\neg A \supset \neg B)$

4. $A \supset \neg\neg A$

5. $\neg\neg A \supset A$

and the single rule of inference is *modus ponens*: $\dfrac{A \supset B \qquad A}{B}$.[5] A formal proof in such a system as Frege's is a finite list of formulas, each of which is either an axiom or follows from previous entries according to one of the inference rules. The propositional fragment of the 'laws of thought' are then the formulas that can be obtained as the final line of a formal proof.

The turnstile notation introduced above in fact originates in Frege's work, where he describes the symbol $\vdash$ as being made up of two parts, the horizontal 'content stroke' and the vertical 'judgement stroke'.[6] There is debate about the significance of the turnstile and its parts in Frege's thought, fueled in large part by some perplexing remarks that Frege made. There is, for example, the notorious doctrine, repeated by Frege enthusiastically on many occasions, that it is not possible to infer anything from 'mere assumptions'. Frege insisted that inference could only be performed from judgements. This sentiment is what drove Frege's critique of the method of testing the consistency of some arbitrary hypotheses by seeing whether it is possible to infer from them a contradiction. According to Frege, this enterprise makes no sense, because if you haven't already 'judged' the truth of your hypotheses, then you cannot infer anything from them, and if you have judged their truth, then you know in advance that they are consistent and have nothing to test by inferring from them (apparently one cannot 'judge' incorrectly).

A related idea appears to motivate Frege's theory of the conditional. Frege is often attributed with advocacy of the material conditional, but the truth of the matter is more subtle. He did, in point of fact, define the expression $A \supset B$ as the denial of the case where A obtains and B doesn't. However, on numerous occasions he explicitly denied that this is an adequate translation of conditional expressions. In *Begriffsschrift* itself, he emphasized that 'the causal connection implicit in the word "if" . . . is not expressed by our symbols'. Indeed, after pointing out that the nested expression $A \supset (B \supset C)$ 'denies the case in which C is denied and B and A are affirmed' he added that 'if a causal connection is present, we can also

[5] Frege also listed $(D \supset (B \supset A)) \supset (B \supset (D \supset A))$ (formula 8) as an axiom, although it can be derived from the five designated here (in fact, from just the first two). Following Frege, we will call any substitution instance of one of the above axioms an axiom and do without a 'rule of substitution'. We thereby avoid the complication that a substitution rule introduces when one considers Γ-derivations, as defined below. With such a rule, one has to keep track of the pedigree of every formula in a derivation to know whether substitution is allowed.

[6] Frege's turnstile looked like this, with single vertical and horizontal strokes, even though they expanded continuously into sometimes quite complicated arrays of 'conditional strokes' and associated continuations. A standard modern usage reserves turnstiles with single horizontal strokes to represent derivability in formal proof systems, distinguished from turnstiles (like those in §2) with double horizontal strokes which stand for semantic consequence.

say . . . "If the circumstances B and A occur, then C occurs also"'. But then Frege remarked that 'a judgement of this kind can be made only on the basis of such a connection'. So we see that on Frege's view material conditionals do not *express* ordinary 'if . . . then' usage, and yet they can only be *judged to be true* based on the sort of connection such usage indicates. Thus Frege's analysis of conditional expressions is in terms, not of the expression $A \supset B$, but of the expression $\vdash A \supset B$.

Perhaps because of the intrigue of what I called Frege's notorious doctrine, it is typically forgotten that in a discussion of expressions without the judgement stroke in §2 of *Begriffsschrift* Frege wrote that one might present such an assertion 'in order to derive some conclusions from it and with these test the correctness of the thought'. Also, in a letter to Hugo Dingler in which he again objected to the idea of 'draw[ing] conclusions from the propositions of a group' without 'first exclud[ing] all propositions whose truth is doubtful', Frege nevertheless mentioned that 'from the thought that 2 is less than 1 and the thought that if something is less than 1 then it is greater than 2, one can derive that 2 is greater than 2' (*1917*, p. 17).

Clearly Frege maintained that 'when we infer, we recognize a truth on the basis of other previously recognized truths according to a logical law' (*1917*, p. 17) but that a type of reasoning unworthy of the label 'inference' is also possible: A *derivation*, not from judgements but from 'thoughts', can establish that a conditional statement is true. Here is how Frege elaborated this point in *1910*: 'I can, indeed, investigate what consequences result from the supposition that A is true without having recognized the truth of A; but the result will then contain the condition *if A is true*'. Given what we have observed about Frege's account of conditionals, it seems that Frege is saying that such derivations from suppositions establish more than mere statements of the form $A \supset B$—they establish judgements of such statements' truth. Indeed, Frege continued: 'Under [such] circumstances we can, by means of a chain of conditions, obtain a concluding judgement of the form $\vdash A \supset (B \supset (C \supset D))$'.

Now, I am aware that the suggestion cuts against a long-established orthodoxy of Frege interpretation, and I doubt that Frege himself had entirely consistent views on the matter, but the ideas assembled here strongly suggest that Frege's turnstile served a role analogous to the modern turnstile of formal derivation.[7] There are, at any rate, only three ways that Frege ever goes about establishing a judgement: informal verification of an axiom, proof in the system of *Begriffsschrift*, and derivation of one formula from another, taken merely as a supposition, establishing thereby a conditional.

Let us define a Γ-derivation as a finite list of formulas, each of which is either a member of the set Γ, an axiom, or the result of an application of the rule *modus ponens* to two for-

[7] Before the establishment of an orthodoxy, Quine spotted the same analogy. In *1951*, he said that the meaning Frege attached to this symbol, though 'somewhat obscure', was 'near enough' to formal derivability for him to retain the notation (p. 88).

mulas occurring earlier in the list. With the label $\mathfrak{B}$ indicating the propositional fragment of *Begriffsschrift*, we index Frege's turnstile and denote the existence of a proof of a formula A by $\vdash_{\mathfrak{B}}$ A and the existence of a Γ-derivation of A by $\Gamma \vdash_{\mathfrak{B}}$ A. Clearly, in case $\Gamma \vdash_{\mathfrak{B}} \mathrm{A} \supset \mathrm{B}$, it follows that $\Gamma, \mathrm{A} \vdash_{\mathfrak{B}}$ B: just append to the end of a given Γ-derivation of A $\supset$ B two final entries, A and B, and observe that because B follows by *modus ponens* from A and A $\supset$ B, the amended list is a Γ, A-derivation of B. Frege appears to be committed also to the converse claim, A $\vdash_{\mathfrak{B}}$ B only if $\vdash_{\mathfrak{B}} \mathrm{A} \supset \mathrm{B}$. This generalizes to a syntactic analogue to the identity (1) from the previous section:

$$\Gamma, \mathrm{A} \vdash_{\mathfrak{B}} \mathrm{B} \text{ if, and only if, } \Gamma \vdash_{\mathfrak{B}} \mathrm{A} \supset \mathrm{B}. \tag{2}$$

4. Two proofs

It is natural to wonder why Frege, given his celebrated exactitude, did not bother to verify the identity (2) that closed the last section. He does tell us, also without providing any justification, that his formal system is complete in the sense that all 'laws of thought' (presumably all the formulas that would submit to the sort of informal verification that Frege provides for his axioms) have proofs. Frege's conviction that derivation from assumptions generates laws of thought in conditional form supplies us with a convenient way to test his completeness claim: A verification of (2) would be considerable evidence that the claim is correct. Alternatively, one could be so convinced of the completeness of the proof system as to use it to determine whether derivation from assumptions does in fact yield true judgements: again, verification of (2) does the trick.

Speculation about Frege's lack of interest in this question can lead in any number of directions. I would only point out that the observation from §2, that Frege's remarks 'strongly suggest' an analogy between his turnstile and our own, can only be made in hindsight. It is an anachronism, but a useful one. Whatever Frege might have ultimately thought about the significance of the judgement stroke—and, again, it is not at all clear that his various pronouncements are all consistent—he did not have in mind what we think of today. Our provability turnstile is informed by, but is a refinement of, his distinction between mere assertion and judgement. Armed with it, we are able to read in Frege's writing a question that he appears not to have noticed himself.

Setting aside the origins of (2) and turning to its verification, one thought that might occur to you is that given the soundness and completeness of the propositional fragment of the system in *Begriffsschrift* (Frege was proven correct about this a few decades later by Paul Bernays (*Bernays 1918*)) with respect to the classical truth functional semantics described in §1, a very simple verification is possible.

Proof. Recall that we already verified one direction of (2), from $\vdash_{\mathfrak{B}} \mathrm{A} \supset \mathrm{B}$ to $\mathrm{A} \vdash_{\mathfrak{B}} \mathrm{B}$. Here is an alternative demonstration of the same fact: Assume $\vdash_{\mathfrak{B}} \mathrm{A} \supset \mathrm{B}$. Because the system is sound, it follows that $\models_{\mathrm{CPL}} \mathrm{A} \supset \mathrm{B}$. Then by (1), $\mathrm{A} \models_{\mathrm{CPL}} \mathrm{B}$, and by the system's completeness, $\mathrm{A} \vdash_{\mathfrak{B}} \mathrm{B}$. Because of the simplicity of the argument in §1, this alternative demonstration seems to be needlessly baroque, appealing as it does to such notions as truth-functions that do not pertain to the question at hand. But it has the advantage of being easily reversible: From the assumption $\mathrm{A} \vdash_{\mathfrak{B}} \mathrm{B}$, soundness gives $\mathrm{A} \models_{\mathrm{CPL}} \mathrm{B}$, which by (1) implies $\models_{\mathrm{CPL}} \mathrm{A} \supset \mathrm{B}$, from which $\vdash_{\mathfrak{B}} \mathrm{A} \supset \mathrm{B}$ follows by completeness. □

This proof of (2) uses only the identity (1) and the semantic completeness of propositional logic, ideas that were explicit already in Paul Bernays's dissertation and David Hilbert's lectures from 1917–18 (see *Franks 2017* for details). But neither the proof nor the identity were noted by anyone in that decade. There are several reasons why this is so.

First of all, the proof is unnatural. As already indicated, (2) is a question wholly in the realm of syntax. It is intrusive to tether this context to the theory of truth functions in order to establish how the syntax hangs together. Certainly, given their aversion to semantic arguments, the Hilbert school would have been disinclined to proceed in this way (*Franks 2017* again has references).

It is also uninformative. The proof tells us that a certain list of formulas exists, but it doesn't indicate how to construct it. The whole interest in 'proof theoretical' facts like (2) derives from an interest in actually obtaining proofs, not merely from knowing that they exist. Recall from §2 the simple proof of the inference from $\vdash_{\mathfrak{B}} \mathrm{A} \supset \mathrm{B}$ to $\mathrm{A} \vdash_{\mathfrak{B}} \mathrm{B}$. The A-derivation witnessing that second claim is such a simple adjustment of the original proof that witnesses the first claim that verifying the inference without spotting this construction just seems unacceptable.

Notice as well that the proof's simplicity is an illusion. The proof of the completeness of propositional logic, suppressed here, is complex relative to the problem at hand. A proof that does not rely on an established correspondence between syntax and semantics promises to be less complex, not only more natural. This is especially true when one leaves the context of propositional logic. The same identity holds in first-order quantification theory. Its verification via a completeness theorem is again available. But Gödel's completeness theorem is highly complex, incorporating infinitary principles of reasoning.

This leads to a final point. The proof just given lacks modularity. To prove (2), we appeal to Bernays's completeness theorem. To prove its analog for first-order quantification theory, we invoke Gödel's result. What about second-order logic or intuitionistic logic? We may be lacking completeness results for some of these theories. Or we may have completeness, but the analog of (1) might be difficult to verify directly. In any case, it doesn't seem at all right that the verifications of each analog of (2) should differ so greatly. The right proof of (2) should pick out a reason why the identity holds that applies in all the analogous settings.

Contrast the following proof of $\Gamma, \mathrm{A} \vdash_{\mathfrak{B}} \mathrm{B}$ only if $\Gamma \vdash_{\mathfrak{B}} \mathrm{A} \supset \mathrm{B}$, using the principle of mathematical induction on the length of the Γ, A-derivation of B:

Proof. (BASE STEP) If there is a one-line derivation of B from Γ, A, then it is possible to construct a Γ-derivation of $\mathrm{A} \supset \mathrm{B}$. To see this, notice that there are only three types of one-line derivations:

1. B is in the set Γ
2. B is A
3. B is an axiom

The following three Γ-derivations exhibit the fact that $\Gamma \vdash_{\mathfrak{B}} \mathrm{A} \supset \mathrm{B}$ in case 1, 2, and 3, respectively:

$$\begin{array}{c|c|c}
\mathrm{B} & (\mathrm{B} \supset ((\mathrm{C} \supset \mathrm{B}) \supset \mathrm{B})) \supset ((\mathrm{B} \supset (\mathrm{C} \supset \mathrm{B})) \supset (\mathrm{B} \supset \mathrm{B})) & \mathrm{B} \\
\mathrm{B} \supset (\mathrm{A} \supset \mathrm{B}) & \mathrm{B} \supset ((\mathrm{C} \supset \mathrm{B}) \supset \mathrm{B}) & \mathrm{B} \supset (\mathrm{A} \supset \mathrm{B}) \\
\mathrm{A} \supset \mathrm{B} & (\mathrm{B} \supset (\mathrm{C} \supset \mathrm{B})) \supset (\mathrm{B} \supset \mathrm{B})) & \mathrm{A} \supset \mathrm{B} \\
 & \mathrm{B} \supset (\mathrm{C} \supset \mathrm{B}) & \\
 & \mathrm{B} \supset \mathrm{B} &
\end{array}$$

It is easy to check that each line of each derivation is either a member of Γ, an axiom, or the results of an application of *modus ponens*.[8] The middle sequence qualifies as a Γ-derivation of $\mathrm{A} \supset \mathrm{B}$ because of the assumption that A is B.

(INDUCTION STEP) Assume that whenever there is a Γ, A-derivation of B that is n or fewer lines long, there is a Γ-derivation of $\mathrm{A} \supset \mathrm{B}$. Now suppose there is an n+1 line long Γ, A-derivation (called $\mathcal{D}$) of B. It is possible that the justification for line n+1 is one of the possibilities from the base case—that B is an axiom, is from the set Γ, or is identical to A. In any of those cases, the Γ-derivation of $\mathrm{A} \supset \mathrm{B}$ would be constructed as in the base case.

Ordinarily, though, a multi-line derivation is long for a reason, and its last line appears as the conclusion from an application of *modus ponens*. In this case, the formulas $\mathrm{C} \supset \mathrm{B}$ and C occur as earlier lines in $\mathcal{D}$. Therefore Γ, A-derivations that are n lines long or shorter of $\mathrm{C} \supset \mathrm{B}$ and C appear as subsequences of $\mathcal{D}$. The induction hypothesis then guarantees the existence of Γ-derivations of $\mathrm{A} \supset (\mathrm{C} \supset \mathrm{B})$ and of $\mathrm{A} \supset \mathrm{C}$. Let $\mathcal{D}_1 = \langle \mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{A} \supset (\mathrm{C} \supset \mathrm{B}) \rangle$ and $\mathcal{D}_2 = \langle \mathrm{T}_1, \mathrm{T}_2, \ldots, \mathrm{A} \supset \mathrm{C} \rangle$ be examples of such derivations, and consider the sequence:

[8] See footnote 5.

$$\begin{gathered}\mathcal{D}_1 \\ \mathcal{D}_2 \\ (\mathrm{A} \supset (\mathrm{C} \supset \mathrm{B})) \supset ((\mathrm{A} \supset \mathrm{C}) \supset (\mathrm{A} \supset \mathrm{B})) \\ (\mathrm{A} \supset \mathrm{C}) \supset (\mathrm{A} \supset \mathrm{B}) \\ \mathrm{A} \supset \mathrm{B}\end{gathered}$$

This is a Γ-derivation of $\mathrm{A} \supset \mathrm{B}$. □

Observe that this proof has all the features that the first proof was missing. It is natural in the sense that we expect reasoning about what derivations must exist based on what derivations are known already to exist to be about those derivations themselves. The proof does not change topics in order to reach its conclusions.

It is also informative, in that it tells us how actually to use the original derivation as raw material and construct out of it the derivation we are interested in. This is a popular pedagogical use of the Deduction Theorem: To find a Frege-style proof, say, of $(\mathrm{A} \supset \mathrm{B}) \supset ((\mathrm{C} \supset \mathrm{D}) \supset ((\mathrm{B} \supset \mathrm{C}) \supset (\mathrm{A} \supset \mathrm{D})))$ might be quite challenging. The proof of the Deduction Theorem just given makes it easy: Begin with a five-line $\mathrm{A} \supset \mathrm{B}, \mathrm{C} \supset \mathrm{D}, \mathrm{B} \supset \mathrm{C}, \mathrm{A}$-derivation of D, transform this according to the construction given in the proof into a $\mathrm{A} \supset \mathrm{B}, \mathrm{C} \supset \mathrm{D}, \mathrm{B} \supset \mathrm{C}$-derivation of $\mathrm{A} \supset \mathrm{D}$, and continue in similar fashion. More importantly, though, the proof is for many students a first encounter with a 'proof-theoretical' result, an argument that an object with certain properties must exist based on available manipulations of related objects assumed already to exist. We will have an occasion to say something more about the significance of this later on.

The proof is certainly simpler than any line of reasoning that courses through the completeness theorem. In fact, the proof makes evident that (2) does not depend even on all the features of classical propositional logic. Frege's axioms 3, 4, and 5 are never invoked.

Its modularity lies in its simplicity. Immediately we see that the same identity holds for the full system of *Begriffsschrift*, not just for its propositional fragment. All that matters is the presence of axioms 1 and 2. Recall that the first proof of (2) suggested that its analog for classical quantification theory would depend on Gödel's completeness theorem and left us wondering if the same identity even holds in intuitionistic logic. We see now that (2) depends on no such high-level properties of any formal system and is true in quantification theory for the same reason that it is true in propositional logic, in their classical, intuitionistic, and many other varieties.

5. Origins

The proof of (2) just presented is customarily attributed to Jacques Herbrand. In chapter 3 of his (*1930*) thesis, Herbrand proved that if a mathematical theory $\mathbf{T}$ can be axiomatized in the language of the system of *Principia Mathematica*, then

- because of the finite nature of proofs, any formula P that can be shown to be derivable in $\mathbf{T}$ can be derived in fact from some finite number of $\mathbf{T}$'s axioms (2.41)
- if we let H denote the conjunction of a finite number of $\mathbf{T}$'s axioms, then a necessary and sufficient condition for the derivability of P from H is the provability of H $\supset$ P (2.4)
- if $\mathbf{T}_1$ and $\mathbf{T}_2$ differ only in the inclusion of the axiom A in the former, then a necessary and sufficient condition for the derivability of P from $\mathbf{T}_1$ is the derivability of A $\supset$ P from $\mathbf{T}_2$ (2.43)

Herbrand's proof of 2.4 is not particularly lucid or memorable. He remarks that the reasoning behind it is recursive, distinguishes the relevant base and inductive cases, and indicates the key formulas that underlie the proof transformations that can be made. But, unhelpfully, he throws considerations about quantified formula transformations into the mix, deals with two separate propositional transformation rules (a rule of 'simplification' in addition to *modus ponens*), and does not highlight at all the fact that proofs are to be constructed in the way indicated in the last section. For these reasons, one must have some significant acquaintance with the system he is using—which is not suited to make the essential components of the proof perspicuous—and various facts he has already proved about it, as well, it seems, as some intuition about the kind of construction he is aiming for in order to follow the argument. By contrast, we have seen, the system in *Begriffsschrift* seems almost to have been reverse-engineered to accommodate the proof of (2) given in §4: among other things, its first two axioms are just the formulas the proof calls for, despite Frege's observation that 'of course, it must be admitted that the reduction' of the infinitely many laws of thought to a few basic axioms 'is possible in other ways besides this particular one' (*1879*, §13).

But the proof of the Deduction Theorem is elementary in any setting, and it is a mistake to think that Herbrand's accomplishment is its discovery. There is no doubt that Frege would have produced it readily had it occurred to him to look for such a thing. He did not look, because the relationship he observed between derivability from assumptions and direct proof did not strike him as the sort of thing that stood in need of verification. It is worth considering why Frege's view down this path did not extend as far as Herbrand's.

Already we have noted that despite certain affinities between his usage and our own, the turnstile did not represent for Frege exactly what it represents for modern logicians. In his own words, it indicates only that a proposition is judged to be true, and he does not clearly distinguish logical and factual truth. This alone explains little, however. Herbrand followed Frege (and Russell) on this score: Like Frege, he said that the turnstile flagged a proposition as 'true', rather than 'valid' or 'provable'; and like Herbrand, Frege described no method for the discovery of a 'truth' worthy of a turnstile other than proof in his formal system and verification of a conditional by deriving, in the same system, its consequent from its antecedent. Evidently the vivid distinction between truth and validity, and between conditional statements and implication claims[9], that figures so prominently in our understanding of the Deduction Theorem played no role in Herbrand's discovery.

Probably a greater obstacle was Frege's ambiguous and somewhat strained attitude about reasoning from arbitrary assumptions. As indicated in §3, such reasoning does not qualify as inference, and therefore among other things it cannot be used to test the mutual compatibility of a set of hypotheses. Although he did, when pressed, grant a legitimate use of such reasoning, these acknowledgements were never occasions for recommendation. Frege stressed: Yes, derivation from assumptions is possible, but all you will generate is a conditional. His negative tone, together with his primary agenda of defending his peculiar concept of inference from any intrusions of this sort, indicates that it did not occur to him that this might in fact be a particularly convenient or efficient way of judging a conditional's truth.

Most importantly, though, the general research setting that inspired Herbrand's whole discussion of these matters was very distant from Frege's thought. Herbrand viewed logic primarily as a vehicle for extracting consequences of mathematical axioms. He hastened to point out, for example, that his fundamental theorem—establishing that the *modus ponens* rule could be systematically eliminated from derivations of pure logic—does not extend to formal axiomatizations of arithmetic, where 'the rule remains necessary' (*1930*, chapter 5, 6.1), that although the rule is 'useless in logic', it 'remains indispensable in mathematical theories' (*1929*). By contrast, Frege conceived of logic as the basis of mathematical theories, especially of arithmetic, whose axioms, being rules of thought, would be demonstrable in the system of *Begriffsschrift*. Frege (wrongly) believed that there would be no occasion for establishing conditionals like $H \supset P$, neither directly in *Begriffsschrift* nor by deriving P from a mathematical theory: Because arithmetical truths are laws of thought, so are their consequences. Therefore, if H is a conjunction of axioms of arithmetic, then instead of proving $H \supset P$ one could just prove P on its own.

In any event, Herbrand's breakthrough was conceptual. Realizing that logical derivation from arithmetical axioms differs in important ways from reasoning in logic purely, the question about how these activities align became, for him, central and urgent. That explains why, despite it holding for arbitrary assumptions, Herbrand frames the Deduction Theorem in the specific terms of the axioms of a mathematical theory. It also explains why the initial insight

[9] Like Russell, Herbrand called $\supset$ the 'implication' sign and referred to *modus ponens* as the 'rule of implication'. In this he seems to have been further removed from the modern conception of the turnstile than Frege, who we saw distinguished the assertion of $A \supset B$, which he said is not an accurate translation of conditional expressions, from its judgement.

that we saw Frege gesturing towards, that derivation from an assumption corresponds to proof of a conditional, not only struck Herbrand as standing in need of proof but came in the fully general form ‘derivation from assumptions corresponds to derivation of a conditional with one fewer assumption’—our (2) and Herbrand’s (2.43): for Frege derivation from assumptions is a shift away from reasoning from the laws of thought, something to be explained away in terms of those laws and, in the event that the assumption was an arithmetical axiom, to be further exposed as a needless activity. Because Herbrand saw derivation from assumptions instead as essential to mathematical thought, he turned to logic as a means for understanding and working with it, not for eliminating it.

6. Abstraction

Some writers, including Tarski himself, have disputed the attribution of the Deduction Theorem to Herbrand, indicating that Alfred Tarski had known and used the result years earlier. What is certain is that Axiom 8 in *Tarski 1930* says that ‘if $\ldots z \in Cn(X+y)$, then $c(y,z) \in Cn(X)$’. In Tarski’s work, Cn is an operator that, applied to a set of sentences, generates the set of those sentences’ (syntactic) consequences, i.e., everything derivable from them according to inference rules; c is an operator that, applied to an ordered pair of sentences, returns another sentence. Axiom 8 and its converse, Axiom 7, are understood by Tarski as a definition of c. Clearly, to say of the sentence denoted by z that it is a member of $Cn(X)$ is just to say that $X \vdash z$. Therefore, Axioms 7 and 8 together express the identity $\Gamma, \mathrm{A} \vdash \mathrm{B}$ if, and only if, $\Gamma \vdash c(\mathrm{A}, \mathrm{B})$. As Church (*1947*) observed, Tarski seems at least to have established this version of the Deduction Theorem, ‘independently and nearly simultaneously’ with Herbrand.

Did Tarski in fact have a proof of the Deduction Theorem as early as 1921? I see no reason to doubt his (*1956*, p. 32) report that he had, although his first discussion of establishing it for a specific deductive system is for theorem 2 of *Tarski 1933*. But the question deflects attention to a far more important achievement of Tarski’s, an advance beyond anything found in Herbrand’s work. For, as Church observed in the continuation of his comparison of Herbrand and Tarski, the identity in *Tarski 1930* appears as ‘a general methodological postulate’. In Tarski’s hands, what for Herbrand was a demonstrable feature of a specific deductive system is reworked as a condition on all deductive systems. By subsuming in this way the ‘theorem’ into the axiomatic framework, Tarski was able to investigate deductive systems as a class of structures. This abstract perspective ushered in several novel considerations.[10]

[10] The general investigation of deductive systems as a class of structures originated in the work of Paul Hertz (*1929*). In important ways, Hertz’s framework was more purely logical than Tarski’s because it did not rely on set theory. Unlike Tarski, Hertz did not consider a language with a sentential connective like $\supset$ that would allow for an expression of the Deduction Theorem. *Franks 2010* and *Franks 2014* present Hertz’s contributions here

Most obviously, Tarski was able to distinguish deductive systems for which Axioms 7 and 8 hold (which we may call implicative) from those for which they fail. This leads naturally to an interest in proving results that hold for all deductive systems as well as results that hold only for implicative systems. More profoundly, it uncovers hidden relationships among distinct deductive systems. For example, familiar axiomatizations of the classical propositional calculus and of the intuitionistic propositional calculus each fall into this class—in each framework the conditional symbol meets the conditions for the operator c—so one can derive axiomatically facts that pertain to both of these logical frameworks. But observe that, as emphasized in *Łukasiewicz and Tarski 1930*, c does not denote a connective in any particular deductive system (although as we just observed, a sentential connective *could* satisfy the conditions of c in some cases).[11] Therefore, an implicative system needn't have a connective that meets the conditions of axioms 7 and 8: there need only be some operation, denoted in the general axiomatic theory of deductive systems, mapping pairs of sentences to sentences in a way that satisfies these axioms. Such a mapping might not pick out any syntax recognizable as expressing conditionalization or implication, or it might not be one-to-one.

This perspective, where a theorem is inverted into an axiom, might seem at first glance to bring us back to Frege's position. Like Tarski, Frege did not bother to prove (2) for the system in *Begriffsschrift*. He simply announced it. Could he have been defining the conditional?

In fact, Tarski's view is better understood as opposing Frege's. Frege defined the conditional twice. He defined it indirectly in terms of the classical truth conditions of the expression A ⊃ B together with the understanding that although such an inscription does not express a conditional, such could only be judged true based on one. He also defined it directly by specifying a canonical way to judge such expressions' truth, in terms of formal derivability: Conditionals are just those statements of the form A ⊃ B that can be proved from the axioms of *Begriffsschrift* and *modus ponens*. Why Frege felt confident that these two definitions—one semantic and the other syntactic—would be extensionally equivalent is a matter of speculation. But it is certain that Frege did not have in addition to these a third definition of the conditional in terms of deducibility from assumptions. That is what makes the question so urgent to Frege's modern readers: How can he just announce that conditionals are always true when their consequents are derivable from their antecedents? Doesn't this need proof?

Tarski, by contrast, by axiomatizing the conditional operator, has things the other way around: The conditional is defined by Axioms 7 and 8. Individual deductive systems (and

especially as precursors of Gentzen's analysis of logical consequence.

[11] Cf. p. 39. Łukasiewicz and Tarski distinguished A ⊃ B, a sentence in the sentential calculus, from $c(x, y)$, a name of that sentence 'in the metasentential calculus'. In light of Quine's analysis, rehearsed in §2, it is disorienting to find this use/mention distinction emphasized in a logical monograph together with the statement that A ⊃ B 'expresses the implication between "A" and "B", but those are the words of Łukasiewicz and Tarski. One could also press the point further and stress that the operator c makes perfect sense even when it does not 'name' any sentence in the system one is studying.

other structures) might fail to be implicative by not satisfying those constraints. From such a ‘failure of the Deduction Theorem’, the thing to conclude is, not that these systems’ ‘conditionals’ fail to properly reflect the consequence operation, but that such systems cannot express conditionals.

7. Natural deduction

The next stage in the Deduction Theorem’s evolution is to my mind the most radical, even if again it appears in our retrospective view to have been preconditioned by the work that preceded it. This is the collapse, in the hands of Gentzen and Jaśkowski[12], of Łukasiewicz and Tarski’s distinction between the sentential and the meta-sentential calculi of formal logic.

Consider, once again, a natural understanding of the identity (2) in the context of the system in Frege’s *Begriffsschrift*: It tells us that proofs of statements of the form $\mathrm{A} \supset \mathrm{B}$ are available whenever there is a A-derivation of B. As indicated in §4, a constructive proof of this theorem, such as the one Herbrand provided, even indicates how to build such a proof, using as raw material the A-derivation of B. There might be occasions when building such proofs are of interest, but more often knowing that they exist suffices. For those purposes, it might make sense to supplement the system of *Begriffsschrift* with a ‘virtual rule of inference’, leading from the ‘premise’ $\Gamma, \mathrm{A} \vdash_{\mathfrak{B}} \mathrm{B}$ to the ‘conclusion’ $\Gamma \vdash_{\mathfrak{B}} \mathrm{A} \supset \mathrm{B}$. There is a textbook tradition of proceeding in this way: first introducing a sentential calculus and notion of formal proof, then verifying the Deduction Theorem, and then supplementing the sentential calculus with a meta-sentential or virtual inference rule.[13] The resulting hybrid system $\mathfrak{B}+$ is far more efficient and student-friendly than the original system $\mathfrak{B}$. Some care is needed, though, because in $\mathfrak{B}+$ one needs to allow applications of *modus ponens* not only to theorems but also to assumptions and sentences derived from assumptions—an allowance that cannot be extended to a ‘substitution rule’. In $\mathfrak{B}+$, *modus ponens* also is ‘meta-sentential’. And importantly, what the Deduction Theorem means in this context is that anything provable in $\mathfrak{B}+$ is in fact provable in $\mathfrak{B}$—the virtual rule is in principle ‘eliminable’. Notice that even when working in $\mathfrak{B}+$, the turnstile refers back to the ‘object level’, to provability in $\mathfrak{B}$.

What might occur to you—although it seldom occurs to students—is that by instead retaining the ‘virtual rule’ allowing the inference of $\Gamma \vdash \mathrm{A} \supset \mathrm{B}$ from $\Gamma, \mathrm{A} \vdash \mathrm{B}$, one could instead eliminate other features of $\mathfrak{B}+$, specifically the two axioms (1 and 2) used in the proof of the Deduction Theorem. Call the resulting system $\mathfrak{G}$. Shortly we will verify that axioms 1 and 2 of the *Begriffsschrift* are in fact provable in $\mathfrak{G}$, and therefore that $\mathfrak{G}$, $\mathfrak{B}+$, and $\mathfrak{B}$ are all

[12] See *Jaśkowski 1934*.

[13] This was Quine’s approach in several monographs, which inspired many later texts. Church’s (*1947*) review of one of them describes the move as follows: ‘By this device the development is greatly simplified . . . , though at the cost of allowing a primitive rule of different and much more complex character than is necessary’.

equivalent. But when working in $\mathfrak{G}$ the turnstile refers, no longer to provability in $\mathfrak{B}$ or any other object-level sentential calculus, but to provability in $\mathfrak{G}$ itself. The turnstile has become self-referential!

The way of thinking leading from Frege and Russell through Herbrand to Quine and the standard textbook tradition of formal logic in the 20th Century obscures the possibility of this move. In that tradition, the conditional (or implication) is defined by axioms of a sentential calculus, and the Deduction Theorem is a meta-theoretic fact about how derivability relates to such statements' validity (or truth). Derivability cannot refer to reasoning carried out at the meta-level where the Deduction Theorem lives. But if one follows Tarski (and before him, Bolzano) and understands the identity of which (1) and (2) are instances as a definition of the conditional, this whole framework is inverted. The Deduction Theorem is merely a verification that one of a specific formal system's sentential connectives operates as a conditional. What is important is the conditional operator itself, though, which one might as well build directly into the notion of formal derivability.

This perspective culminated in Gerhard Gentzen's (*1934–35*) *Untersuchungen über das logische Schließen*. There, Gentzen designed 'natural deduction' proof systems for propositional and quantificational logic in which not only the conditional[14] but in fact all the logical operators are characterized by pairs of inference rules, an 'elimination rule' characterizing how to reason *from* a statement governed by that operator and an 'introduction rule' characterizing how to reason *to* such a statement. For example, ¬ is defined by the elimination rule $\dfrac{\neg A \qquad A}{\bot}$ and the introduction rule $\dfrac{\genfrac{}{}{0pt}{}{[A]}{\bot}}{\neg A}$, and ∧ is defined by the elimination rule allowing both $\dfrac{A \wedge B}{A}$ and $\dfrac{A \wedge B}{B}$ and the introduction rule $\dfrac{A \qquad B}{A \wedge B}$.

To this day it is routinely said that the value of natural deduction lies principally in its efficiency and the intuitive manner in which its proofs can be constructed. Gentzen himself made occasional remarks along these lines. For example in the synopsis of the *Untersuchungen* he wrote, 'I intended first to set up a formal system which comes as close as possible to actual reasoning'. But two points are worth stressing. First, at best a formal system could capture the flow of reasoning required to follow the written record of a mathematical discovery. (Gentzen himself claimed that his introduction and elimination rules are the result of an empirical discovery of such written records.) This should not be mistaken with the claim that these same rules are constitutive of the original context of mathematical discovery. Second, even if Gentzen is to be believed that his original intention was to emulate actual mathematical reasoning, and that it only 'turned out' afterwards that the systems designed to this end had 'certain special properties', the value of natural deduction to logical theory lies in those properties.

[14] Gentzen actually perpetuated the practice of referring to ⊃ as the implication operator.

The property of natural deduction that I want to focus on here[15] is the way that the Intro/Elim scheme captures the meaning of a logical particle. This is perhaps most vivid when the rules are rewritten in turnstile notation. For example, the rules for $\wedge$ can be written thus:

$$\begin{aligned} \mathrm{A} \wedge \mathrm{B} \vdash \mathrm{A} \text{ and } \mathrm{A} \wedge \mathrm{B} \vdash \mathrm{B} \\ \text{For all C, if } \mathrm{C} \vdash \mathrm{A} \text{ and } \mathrm{C} \vdash \mathrm{B}, \text{ then } \mathrm{C} \vdash \mathrm{A} \wedge \mathrm{B} \end{aligned}$$

The first condition—corresponding to $\wedge$-Elim—says that $\mathrm{A} \wedge \mathrm{B}$ is an object that allows the inference to both A and B. The second condition—corresponding to $\wedge$-Intro—tells us that $\mathrm{A} \wedge \mathrm{B}$ is the 'weakest' such object in the sense that any other object that allows the same inferences must be 'stronger' than $\mathrm{A} \wedge \mathrm{B}$, i.e., must allow also an inference to $\mathrm{A} \wedge \mathrm{B}$.

The rules for $\neg$ appear as:

$$\begin{aligned} \neg \mathrm{A}, \mathrm{A} \vdash \bot \\ \text{For all C, if } \mathrm{C}, \mathrm{A} \vdash \bot, \text{ then } \mathrm{C} \vdash \neg \mathrm{A} \end{aligned}$$

Again the first condition tells us that $\neg$A is an object that, together with A allows the inference to $\bot$ (absurdity), and the second condition tells us that this is the full account of $\neg$A, that any other object that pairs with A to allow an inference to $\bot$ is at least as strong as $\neg$A in the sense that from it one can infer $\neg$A directly.

This same scheme presents a novel route to the Deduction Theorem, as implicitly contained in the rule *modus ponens*. For if one understands *modus ponens* not just as one among many patterns of valid inference but as saying that it is *definitive* of the conditional $\mathrm{A} \supset \mathrm{B}$ that it licenses the inference from A to B, then part of what one means is that any sentence that can pair up with A to allow the inference to B expresses something equivalent to or stronger than $\mathrm{A} \supset \mathrm{B}$. To cast this into the setting of natural deduction, one has to first understand *modus ponens* as applying not just to Fregean judgements, but even to arbitrary assumptions. Then it is just $\supset$-Elim, rewritten in the turnstile notation as

$$\mathrm{A} \supset \mathrm{B}, \mathrm{A} \vdash \mathrm{B}$$

Understood as a definition of $\supset$, this determines that $\supset$-Intro must be the following form of the Deduction Theorem, stating that $\mathrm{A} \supset \mathrm{B}$ can be inferred from any other formula that fills this role:

$$\text{for all C, if } \mathrm{C}, \mathrm{A} \vdash \mathrm{B}, \text{ then } \mathrm{C} \vdash \mathrm{A} \supset \mathrm{B} \tag{3}$$

[15] In *Franks 2010* emphasis is placed on the way that natural deduction allows an articulation and verification of a notion of logical completeness.

In the language of category theory, this understanding of the way Intro/Elim rules define a logical operator is called a 'universal mapping property'. Thus the 'categorial product' of two objects A, B is another object that 'points' to them both, and not only that, but it is the quintessential such object in the sense that any other object that points to both A and B must also point to their product. Dually, the categorial coproduct of A and B is the quintessential thing they both point to, quintessential in the sense that anything else pointed to by both A and B is pointed to by this coproduct.[16]

One might say that the fundamental idea of natural deduction is that propositions can be viewed as the objects of a mathematical category, where the partial order on propositions given by the deducibility relation are the category's arrows. Of course, this understanding can only be applied to Gentzen, who wrote in the 1930's, with the benefit of hindsight. But in essence his calculi of natural deduction are a presentation of logical operators defined in terms of universal mapping properties on this category of propositions, related by deducibility. For his own part, Gentzen wrote: 'The introductions represent . . . the "definitions" of the symbols concerned, and the eliminations are no more, in the final analysis, than the consequences of these definitions. . . . By making these ideas more precise it should be possible to display the Elim-rules as unique functions of their corresponding Intro-rules' (§II 5.13). The concept of a universal mapping property is the precisification Gentzen foresaw: Elimination rules are not consequences of their corresponding introduction rules, but they do follow from the conception of those introduction rules as definitions. Conjunction is the categorial product; disjunction, the categorial coproduct. What conditionalization is will be described in §9.[17]

When Gentzen announced that his natural deduction calculi 'turned out to have certain

[16] The full elaboration of this understanding results in what is known as a Heyting category, which is a presentation of the variety of semi-lattices known as Heyting algebras in the language of category theory. We do not present that full elaboration here. Instead, in §9, we consider a refinement of this construction that results from using as the fundamental notion, instead of the modal concept of derivability ('what can be inferred'), particular derivations. This allows the articulation of an aspect of the Deduction Theorem not visible in the Heyting category.

[17] A sizeable literature has developed around the question of the proper relationship between introduction and elimination rules. The reader will have noticed that in the famous passage just presented, Gentzen described the introduction rules as definitive and the elimination rules as determined by them, whereas in the presentation of $\wedge$ and $\supset$ given here, things are the other way around: for example the Deduction Theorem, conceived of as an introduction rule for $\supset$, is shown to be determined by the conception of *modus ponens* (i.e., $\supset$-Elim) as definitional. The point stressed here is that the question as to the priority of the elimination or introduction rules is a distraction. Neither rule 'follows' from the other, although the conception of one of them as, not merely an incidental property of a connective, but as definitive of that connective determines what the other must be. In the case of disjunction, the natural presentation begins, as Gentzen described it, with the introduction rule and ends with the elimination rule (see the definition of coproduct in §9). This is the sense in which hindsight sheds light on Gentzen's words. His claim that one rule is a consequence of another has proven to be hard to evaluate in the terms that he used but makes perfect sense in terms of universal mapping properties, and in these terms the question of priority of one rule over the other falls away.

special properties', he specified one in particular. In the synopsis he wrote that 'the law of excluded middle, which the intuitionists reject, occupies a special position'. In paragraph 5.3 of §II, he observed that classical logic could be formulated with an additional rule schema for double negation $\frac{\neg\neg A}{A}$ 'in place of the basic formula schema' for the law of excluded middle, and then specified the special position that this schema occupies: 'Such a schema still falls outside the [introduction/elimination] framework . . . , because it represents a new elimination of the negation whose admissibility does not follow at all from our method of introducing the ¬-symbol by the ¬-Intro [rule]'. In contemporary terms, Gentzen's point is that the law of excluded middle and the double negation rule are in no way necessitated by the meaning of the negation operator, when one thinks of its meaning as given by the universal mapping property underlying the ¬-Intro/¬-Elim rules.

To see how unappreciated Gentzen's theoretical advance is, one need only survey textbook presentations of natural deduction. In the very popular *Language, Proof, and Logic* by Barwise, Etchemendy, and Barker-Plummer, for example, the inference figures $\frac{\neg\neg A}{A}$, $\frac{\neg A \quad A}{\bot}$, and $\frac{\bot}{A}$ are called '¬-Elim', '⊥-Intro', and '⊥-Elim'. The same nomenclature appears even in Richard Kaye's lovely *The Mathematics of Logic*. In neither book is there any mention that the first of these combines with no other rule to specify a universal mapping property or that the second is actually the elimination rule for ¬, defining that operator together with its introduction rule and having no particular relationship at all to the third with which it is paired. Yet Gentzen's stated reason for publishing his system of natural deduction, even if it wasn't his original motivation for devising it, was to point out that the double negation and *ex falso quodlibet* rules 'fall outside the Intro/Elim framework' that he devised to define the various connectives.[18]

It is this feature of natural deduction that leads to a direct justification of the Deduction Theorem (formulated as (3)), not as a verifiable fact about conditionals in any particular sentential calculus, but as the other side of the coin of the *modus ponens* rule. When conditionalization is understood as an operation that maps A and B to the proposition that together with A licenses an inference to B, *modus ponens* captures the part about licensing this inference, and (3) captures the part about being *the thing* that so licenses it, so that any other proposition that allows an inference from A to B only does so *via* A ⊃ B. And as expected, the verification of the Deduction Theorem in the form (2), based on a definition of the conditional in terms of Frege's basic laws 1 and 2 can be inverted: By defining the conditional as a universal mapping

[18] This tendency to overlook the conceptual advance of natural deduction might originate with *Quine 1950*, where Quine advanced what he described as a system that 'enhances' the advantages of Gentzen's system over axiomatic frameworks. Quine's system suppresses all propositional inference except for one rule corresponding to the Deduction Theorem, citing the decidability of classical propositional logic as obviating such features. He even said there that natural deduction generally 'lacks certain traits of elegance which grace' axiomatic systems (p. 93).

property (specified by Gentzen's introduction and elimination rules), the Deduction Theorem in the form (3) and *modus ponens* can be used to verify Frege's first two basic laws:

$$\dfrac{\dfrac{\dfrac{\overset{1}{[\mathrm{A}]} \qquad \overset{2}{[\mathrm{B}]}}{\mathrm{A}}}{\mathrm{B} \supset \mathrm{A}}\;{\supset}\text{-Intro}_2}{\mathrm{A} \supset (\mathrm{B} \supset \mathrm{A})}\;{\supset}\text{-Intro}_1 \qquad \dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\overset{1}{[\mathrm{A}]} \quad \overset{2}{[\mathrm{A} \supset \mathrm{B}]}}{\mathrm{B}}\;{\supset}\text{-Elim} \quad \dfrac{\overset{1}{[\mathrm{A}]} \quad \overset{3}{[\mathrm{A} \supset (\mathrm{B} \supset \mathrm{C})]}}{\mathrm{B} \supset \mathrm{C}}\;{\supset}\text{-Elim}}{\mathrm{C}}\;{\supset}\text{-Elim}}{\mathrm{A} \supset \mathrm{C}}\;{\supset}\text{-Intro}_1}{(\mathrm{A} \supset \mathrm{B}) \supset (\mathrm{A} \supset \mathrm{C})}\;{\supset}\text{-Intro}_2}{(\mathrm{A} \supset (\mathrm{B} \supset \mathrm{C})) \supset ((\mathrm{A} \supset \mathrm{B}) \supset (\mathrm{A} \supset \mathrm{C}))}\;{\supset}\text{-Intro}_3$$

8. Logical structure

The history recounted so far twice touches on the concept of *admissibility*, a focus of contemporary logic to which the Deduction Theorem is intimately connected. We shall see that the Deduction Theorem played a crucial role in the identification of admissibility as a phenomenon distinct from the more general notion of deductive validity, and it returned as the key to characterizing logical structures in light of the questions that admissibility poses.

Intuitively, a rule of inference is called 'admissible' if it leads reliably from one logical truth to another. More precisely, an admissible inference rule is one whose conclusion has substitution instances that are logical theorems whenever the corresponding substitution instances of its premises are. In algebraic terminology, admissible rules are operations under which a logic's set of theorems is closed.

In §3 we observed that Frege's conception of logical inference seems to correspond with this notion, for he called 'inferring' the practice of reasoning from 'already established truths' to 'other truths'. The more general activity of reasoning from hypotheses or arbitrary assumptions did not qualify as inference in Frege's eyes. He called it 'mere derivation'. We observed further that Frege's distinction between genuine inference and mere derivation facilitated one of the earliest articulations of the Deduction Theorem, in his (unproved) claim that the logical validity of a conditional expression ($\vdash_{\mathfrak{B}} \mathrm{A} \supset \mathrm{B}$) follows from the derivability in the system of *Begriffsschrift* of its consequent from its antecedent ($\mathrm{A} \vdash_{\mathfrak{B}} \mathrm{B}$).

Frege's distinction survives in contemporary discussions as the distinction between admissibility and derivability. A logical system's derivable rules are those from which one can reliably reason, not just from theorems but from arbitrary formulas. The sense in which a logical system's admissible rules are reliable is straightforward: The result of applying them to theorems is another theorem. The sense in which derivable rules are reliable is harder to express and sets up the general problem of defining deductive consequence: One clearly wants more from an inference rule than that it lead to another arbitrary formula.

In classical propositional logic this something more can be specified in terms of the concept of an interpretation. There a rule is derivable if, whenever its premises are true *in an interpretation* (under an assignment of truth values to atomic formulas), its conclusion is true *in that same interpretation*. Similar specifications are available in richer settings (quantification theory, intuitionistic logic) according to obvious reformulations in terms of set-models or forcing relations on frames. An alternative, proof-theoretical, specification is available, though, simply in terms of a designated set of primitive rules (for example, the Intro/Elim rules of natural deduction). Then an inference rule is said to be derivable if a finite sequence of applications of primitive rules leads from its premises to its conclusion.

With this in mind, it should be clear that in any framework whatsoever, a derivable rule will also be admissible. Consider again the model-theoretic specification of classical propositional logic: If a rule preserves truth in any interpretation (i.e., if it is derivable), then it certainly preserves the property of being true in every interpretation (i.e., it is admissible). Equally trivial is the verification in proof-theoretic terms: If a rule's conclusion can be derived primitively from its premises (i.e., it is derivable), and if, further, each of its premises are themselves primitively derivable with no assumptions at all, then so too must be its conclusion (i.e., it is admissible).

Frege appears to have assumed the converse, that any rules available for 'inference' among 'judgements' (available, that is, to reliably establish theorem-hood on the basis of previously established theorems) are also available to form derivations among arbitrary formulas. This assumption paid off, as it led to Frege's articulation of the Deduction Theorem and shortcut method of verifying the validity of conditional judgements. Its status as an assumption, though, is ambiguous. On the one hand, Frege was correct that the admissible rules of *Begriffsschrift* are also derivable. On the other hand, it is not generally true that admissibility implies derivability—the fact that it does in the propositional fragment of the system in *Begriffsschrift* was Frege's good fortune.

Logical systems whose admissible rules are all derivable are called 'structurally complete'. It is commonly said that classical propositional logic is structurally complete, and this property appears to underlie some of Frege's maneuvering. Not all logical systems are structurally complete, though. Most notoriously, the traditional presentations of intuitionistic propositional logic have admissible rules that are not derivable. The most well-known example of such is Harrop's rule[19]:

$$\frac{\neg \mathrm{A} \supset (\mathrm{B} \vee \mathrm{C})}{(\neg \mathrm{A} \supset \mathrm{B}) \vee (\neg \mathrm{A} \supset \mathrm{C})}$$

This rule is often cited as the earliest discovered example of an admissible but underivable rule, although in Kreisel and Putnam's *1957* study of it, they point out that in *1953* G. H. Rose

[19] First presented in *Harrop 1956*.

had made the same observation about the rule:

$$\frac{(\neg\neg A \supset A) \supset (\neg\neg A \vee \neg A)}{\neg\neg A \vee \neg A} \tag{4}$$

More recently, Tim van der Molen (*2016*) observed that Ingebrigt Johansson had noted the admissibility and non-derivability of the rules $\frac{\perp}{A}$ (*ex falso quodlibet*) and $\frac{A \vee B \qquad \neg B}{A}$ (*disjunctive syllogism*) in minimal propositional logic already in 1935–36 correspondence with Arend Heyting. The difference between minimal and intuitionistic logic, in other words, is that the latter has strictly more derivable rules. They agree about admissibility.

One need not look to studies about non-classical logics to witness the phenomenon of structural incompleteness, though. We encountered it already in our review of Herbrand's pioneering work on the Deduction Theorem. In §5 we observed that one of Herbrand's principal applications of the Deduction Theorem is his demonstration that the inference rule *modus ponens* can be systematically eliminated from what Herbrand called 'purely logical' derivations. He contrasted purely logical derivations with derivations from arithmetical or other mathematical axioms, in which *modus ponens* can be essential. (Although the conceptual relationship between *modus ponens* and the Deduction Theorem—first exposed by Gentzen in terms of the Intro/Elim rules and later clarified in the categorial terms of universal mapping properties—was not explicit in Herbrand's thought, it operates behind the scenes of his proof.)

Herbrand's observation is readily recast in terms of admissibility: If one removes the rule *modus ponens* from Herbrand's formulation of classical logic, the result is a system in which *modus ponens* is admissible but not derivable. (This same observation is more commonly made in terms of the sequent calculus, by saying that in certain cut-free calculi, the *cut* rule is admissible but not derivable.)

A similar encounter with structural incompleteness occurred in our review of the transition from Frege-style presentations of logic to natural deduction. Recall the hybrid system $\mathfrak{B}+$ that we encountered along the way: It has all the axioms and inference rules of *Begriffsschrift* together with the 'virtual' rule corresponding to the Deduction Theorem. Because the Deduction Theorem holds for $\mathfrak{B}$, $\mathfrak{B}+$ will have no more theorems than $\mathfrak{B}$ has. But one must be careful, we noted, when working in $\mathfrak{B}+$ to apply the substitution rule only to logical axioms or to formulas derived purely from them, because although substitution instances of theorems are again theorems, unrestricted use of substitution leads quickly to inconsistency (from $A \wedge B$ infer $A \wedge \neg A$, or, even more directly, from A infer $B \wedge \neg B$). Substitution is admissible but not derivable.

These early and rudimentary encounters with structural incompleteness highlight a general feature of the theory of admissibility. Tarski (e.g., in*1936*) defined a logical theory as a set of theorems closed under a consequence relation. But on this Tarskian conception, there is no fact of the matter about whether a logical theory is structurally complete or not. Structural completeness, in other words, is not a property of Tarskian theories but of the consequence relations that generate them. Because distinct consequence relations can agree about logical theorem-hood, knowledge of the set of theorems leaves undetermined further facts about the consequence relation. In particular, although it does fix the set of admissible inference rules, that information does not fix the set of derivable rules.[20]

To illustrate the Deduction Theorem's bearing on the natural questions raised by these phenomena, it is helpful to present the notation of the abstract theory of consequence relations.

Following *Hertz 1929*, *Gentzen 1932*, and *Tarski 1936*,[21] one finds the concept of a (single-conclusion) **finitary consequence relation** defined as a relation $\vdash$ on a set S of formulas that satisfies, for all finite subsets $\Gamma, \Delta \subseteq S$ and all formulas $\mathrm{A}, \mathrm{B} \in S$:

1. $\mathrm{A} \vdash \mathrm{A}$ (reflexivity)
2. if $\Gamma \vdash \mathrm{A}$ then $\mathrm{B}, \Gamma \vdash \mathrm{A}$ (monotonicity)
3. if $\Gamma \vdash \mathrm{A}$ and $\mathrm{A}, \Delta \vdash \mathrm{B}$ then $\Gamma, \Delta \vdash \mathrm{B}$ (transitivity)

with the convention that a serial list of formulas be understood as their union and a single formula standing for its own 'singleton set'.

In these terms one can define several key notions, including succinct characterizations of admissibility and derivability:

- If $\vdash$ is a consequence relation, then $\mathrm{Thm}(\vdash) = \{\mathrm{A}\colon \vdash \mathrm{A}\}$, i.e. $\mathrm{Thm}(\vdash)$, 'the theorems of $\vdash$', denotes the set of all formulas that stand in the $\vdash$ relation to the empty set.
- If $\vdash$ is a consequence relation, and $r = \dfrac{\Gamma}{\mathrm{A}}$ is an inference rule, then $\vdash^{r}$ is the smallest consequence relation $\vdash$ extending $\vdash$ for which $\Gamma \vdash \mathrm{A}$.
- A rule $\dfrac{\Gamma}{\mathrm{A}}$ is *derivable* in $\vdash$ if $\Gamma \vdash \mathrm{A}$.
- A rule $r = \dfrac{\Gamma}{\mathrm{A}}$ is *admissible* in $\vdash$ if $\mathrm{Thm}(\vdash) = \mathrm{Thm}(\vdash^{r})$. The notation $\Gamma \mathrel{|\!\!\sim} \mathrm{A}$ is

[20] This is not to say that the set of admissible inference rules can be effectively determined from information about the set of theorems. Chagrov (*1992*) identified logical systems whose set of derivable rules, and hence theorems, is decidable, although the set of admissible rules is undecidable. And generally, the complexity of the question of admissibility of inference rules is computationally more complex than the corresponding question of derivability.

[21] See *Franks 2018* for an account of the history of this notion and of its relationship to a competing definition of logical consequence in terms of substitution.

used to express the fact that $\frac{\Gamma}{\mathrm{A}}$ is *admissible* in $\vdash$.

It is easy to verify that if $\vdash$ is a consequence relation, then $\mathrel{|\!\sim}$ is also a consequence relation.[22] Of course these two relations have the same theorems. More generally, if $\vdash_1$ and $\vdash_2$ are consequence relations such that $\mathrm{Thm}(\vdash_1) = \mathrm{Thm}(\vdash_2)$, then $\mathrel{|\!\sim}_1 = \mathrel{|\!\sim}_2$.

This observation has led some logicians to remark that, because they depend only on the set of theorems, a logic's true characterization is in terms of its admissible rules. The derivable rules, by contrast, depend on a 'design choice', i.e., on which proof system or semantic framework one chooses to generate its theorems. See especially *Rybakov 1997*. Similarly *Iemhoff 2016* remarked that whether or not a logic is structurally complete 'depends very much on the particular consequence relation one uses for a logic', whereas 'admissibility solely depends on the [logic's] theorems' which is invariant for all such consequence relations.

As was mentioned above, one ordinarily says that classical propositional logic (CPC) is structurally complete, whereas intuitionistic propositional logic (IPC) is not. Indeed, IPC is the standard reference for the structural incompleteness phenomenon. Can this sentiment be maintained in light of the fact that if $\vdash_{\mathrm{IPC}}$ is any consequence relation generating the theorems of IPC, then $\mathrel{|\!\sim}_{\mathrm{IPC}}$ is a structurally complete relation with the exact same theorems? A tempting, but ultimately unsuccessful, response is to try to pin the distinction between CPC and IPC to the mere existence of structurally incomplete relations generating the theorems of the latter. Such a response fails because, as Rosalie Iemhoff pointed out, the smallest consequence relation $\vdash_{\mathrm{RI}}$ that generates the theorems of CPC, given by

$$\Gamma \vdash_{\mathrm{RI}} \mathrm{A} \text{ if, and only if, } \mathrm{A} \in \Gamma \cup \mathrm{Thm}(\vdash_{\mathrm{CPC}}),$$

is structurally incomplete.[23]

In sum, both CPC and IPC have structurally complete and structurally incomplete presentations, in the sense that the theorems of each are generated by both sorts of consequence relations. What can be said in favor of the compelling idea that structural incompleteness is a hallmark of intuitionsitic logic? The Deduction Theorem provides an answer.

A consequence relation $\vdash$ such that $\mathrm{Thm}(\vdash) = \mathrm{Thm}(\vdash_{\mathrm{IPC}})$ can evidently be structurally complete only by violating the Deduction Theorem: In any such relation, $(\neg\neg\mathrm{A} \supset \mathrm{A}) \supset (\neg\neg\mathrm{A} \vee \neg\mathrm{A}) \vdash \neg\neg\mathrm{A} \vee \neg\mathrm{A}$, corresponding to the admissibility of Rose's Rule (4), so that

[22] See, for example, *Iemhoff 2016*, corollary 4.3.

[23] See *Iemhoff 2016* for verification that $\vdash_{\mathrm{RI}}$ satisfies the conditions of a consequence relation and is the minimal element of $\{\vdash : \mathrm{Thm}(\vdash) = \mathrm{Thm}(\vdash_{\mathrm{CPC}})\}$.

the left to right direction of (2) would force $\vdash ((\neg\neg A \supset A) \supset (\neg\neg A \vee \neg A)) \supset (\neg\neg A \vee \neg A)$. But $((\neg\neg A \supset A) \supset (\neg\neg A \vee \neg A)) \supset (\neg\neg A \vee \neg A)$ is certainly not a theorem of IPC.

Conversely, consider any consequence relation $\vdash$ with $\mathrm{Thm}(\vdash) = \mathrm{Thm}(\vdash_{\mathrm{CPC}})$ but with a rule $\frac{A}{B}$ for which $A \mathrel{|\!\sim} B$ but not $A \vdash B$. Because $A \mathrel{|\!\sim} B$, $A \mathrel{|\!\!\sim}_{\mathrm{CPC}} B$, and so $\vdash_{\mathrm{CPC}} A \supset B$. Therefore also $\vdash A \supset B$. Applying the right to left direction of (2) leads to $A \vdash B$, contradicting our assumption about $\vdash$. For example, $\neg\neg A \mathrel{|\!\!\sim}_{\mathrm{RI}} A$, but not $\neg\neg A \vdash_{\mathrm{RI}} A$, even though $\vdash_{\mathrm{RI}} \neg\neg A \supset A$.

In *2016*, Iemhoff wrote that some consequence relations that generate the theorems of familiar logics like CPC and IPC are ‘far more natural than others’. The Deduction Theorem clarifies this intuition. Among the connectives of CPC and IPC is $\supset$. Following Gentzen and Bolzano, the inferential meaning of this connective is given by the Deduction Theorem and the (‘metasentential’ version of the) *modus ponens* rule. Therefore in any consequence relation in which the Deduction Theorem fails, $\supset$ will not have its intended meaning. If propositional logic is supposed to characterize the rules of reasoning about conjunction, negation, conditionalization, and the like, then the Deduction Theorem is a constraint over and above the definitive conditions of a consequence relation that must be satisfied in order for a relation to adequately represent a propositional logic. When this constraint is met, all the admissible rules of classical logic are derivable, whereas intuitionistic logic has underivable admissible rules.

The Deduction Theorem led Herbrand to the earliest observation of the difference between admissibility and derivability: The system that results from removing *modus ponens* from his formulation of classical logic has *modus ponens* still as an admissible, though not as a derivable, rule. If all one were interested in were a presentation of the theorems of classical logic, the lesson from this observation would be that there are two equally good presentations to choose from—one being more efficient, the other less redundant. This is not all a logician might be interested in, though. As Herbrand pointed out, *modus ponens* ought to be an explicit rule so that one can apply logic also to non-logical axioms, such as the formulas that define arithmetical operations or the principle of mathematical induction. Even more basically, we observed that having the right set of theorems and satisfying the conditions of a consequence relation might be enough for some purposes, but that it doesn’t suffice to adequately represent propositional logics. For that purpose, the Deduction Theorem is an additional constraint, one that picks out the structurally complete presentation of classical logic and a structurally incomplete presentation of intuitionistic logic.

9. Adjunction

The Deduction Theorem makes precise the intuition that classical logic is structurally complete whereas intuitionistic logic isn't. But there is another sort of logical completeness that applies to intuitionistic and classical logic alike, and again the Deduction Theorem underlies the phenomenon. As Kosta Došen wrote in *1996*, the Deduction Theorem

> . . . is a completeness theorem of some sort. It tells us that the system is strong enough to mirror its extensions in the same language. . . . For the deductions we shall be able to make in the extension we already have corresponding implications in the system. However, this completeness is not such that subsystems or extensions of our system cannot have it. In classical logic we have the deduction theorem, but in intuitionistic logic we have it, too. (p. 243–4)

Following Došen, we could call this 'deductive completeness' in contrast to 'structural completeness' and 'semantic completeness'.[24]

The extensions of a logical system that Došen referred to are just systems supplemented with non-logical axioms. So his concept of completeness is about a 'correspondence' between derivations from hypotheses and 'implications', i.e., hypothesis-free proofs of conditionals. Clearly the Deduction Theorem in the form (2), because it is a biconditional claim, describes a parity between proofs of conditionals and derivations of those conditionals' consequents from their antecedents. But the correspondence Došen had in mind runs deeper than this, as indicated by his expression of a system 'mirroring' its extensions. He also described the extended systems being 'contained in' the original system, or there being 'a reflection' of each of the former in the latter (p. 274).

This is a provocative reading of the Deduction Theorem, but one not at all expressed in its usual formulation. Recall that in the standard proof of (2), one gets more than a verification of a biconditional: one gets a recipe for building a proof of a conditional, using a derivation of that conditional's consequent from its antecedent as raw ingredient. For that reason, the proof conveys a tremendous amount of information about how any hypothetical derivation and its corresponding hypothesis-free proof are related—information that is lost in the statement of the theorem. Even the verification of (1) in the theory of classical truth exposes a close relationship between classical entailment and the validity of an associated conditional, whereas the biconditional (1) says only that when you have one you will have the other as well. It is natural to seek a fuller expression of the way in which the structure of derivation from assumptions is contained already within the structure of a system's theorems.

The vocabulary for expressing just this sort of sameness of structure can be found in

[24] Došen used the term 'deductive completeness' to refer to a generalization and extension of the property described here—one that applies across the full variety of Lambek's 'deductive systems', of which logical systems ordinarily construed are only a special type. Lambek himself used the term 'functional completeness' in *1974* and *Lambek and Scott 1986*.

category theory.

Definition A **category** $\mathcal{C}$ consists of a collection $\mathrm{ob}(\mathcal{C})$ of **objects** $(A, B, \ldots)$ and a collection $\mathrm{hom}(\mathcal{C})$ of **morphisms** $(f, g, \ldots)$ between objects, denoted by (and conventionally also called) arrows whose sources and targets are objects in $\mathrm{ob}(\mathcal{C})$, together with a binary relation $\circ$ called composition of morphisms that on an ordered input of two arrows, $f : A \longrightarrow B$ and $g : B \longrightarrow C$, the target of the first of which is the second's source, returns an arrow $g \circ f : A \longrightarrow C$ and, for every object C, a special[25] **identity morphism** denoted 1_C, such that:

- $h \circ (g \circ f) = (h \circ g) \circ f$ for any $f : A \longrightarrow B$, $g : B \longrightarrow C$, and $h : C \longrightarrow D$. (Associativity Law)
- $1_B \circ f = f = f \circ 1_A$, for any A, B and $f : A \longrightarrow B$. (Identity Law)

Often one wants to refer to the restriction of $\mathrm{hom}(\mathcal{C})$ just to arrows with source A and target B. This collection is denoted $\mathrm{hom}_{\mathcal{C}}(A, B)$.

The laws of category theory can also be expressed in terms of the 'commutativity' of the following diagrams, meaning that any two paths from one object to another object in such a diagram made by composition of arrows yield identical morphisms:

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[dr, "g\circ f"'] & B \arrow[d, "g"'] \arrow[dr, "h\circ g"] & \\
& C \arrow[r, "h"'] & D
\end{tikzcd}$$

Associativity Law

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[d, "1_A"'] \arrow[dr, "f"] & B \arrow[d, "1_B"] \\
A \arrow[r, "f"'] & B
\end{tikzcd}$$

Identity Law

Commutative diagrams provide a particularly vivid depiction of the way objects in a category can be characterized by universal mapping properties. We remarked in §7 that conjunction and disjunction, as defined by their inferential role with Intro/Elim rules in natural deduction, are examples of the categorial constructions called product and coproduct. Here are the general definitions:

Defintion The **product** $A \otimes B$ of two objects A and B is an object

1. *from which* there are 'projection' arrows p_A and p_B: $A \xleftarrow{p_A} A \otimes B \xrightarrow{p_B} B$, and
2. *to which* there is a unique arrow h from any object C from which there are arrows f and g:
$$\begin{tikzcd}
& C \arrow[dl, "f"'] \arrow[dr, "g"] & \\
A & & B
\end{tikzcd}$$
, such that the following diagram commutes (here the broken arrow indicates that the existence of h is implied by the rest of the structure):

[25] Uniqueness of identity morphisms follows from their existence and the associativity of composition.

$$\begin{tikzcd}
& C \arrow[dl, "f"'] \arrow[d, dashed, "h"'] \arrow[dr, "g"] & \\
A & A \otimes B \arrow[l, "p_A"] \arrow[r, "p_B"'] & B
\end{tikzcd}$$

The **coproduct** $A \oplus B$ of two objects A and B is an object

1. *to which* there are arrows sourced at both A and B : $A \xrightarrow{c_A} A \oplus B \xleftarrow{c_A} B$, and
2. *from which* there is a unique arrow h to any object C that is the target of arrows f and g sourced at A and B such that the following diagram commutes:

$$\begin{tikzcd}
A \arrow[r, "c_A"] \arrow[dr, "f"'] & A \oplus B \arrow[d, dashed, "h"'] & B \arrow[l, "c_B"'] \arrow[dl, "g"] \\
& C &
\end{tikzcd}$$

Such universal mapping properties abound in category theory, characterizing objects in terms of (1) their relation to other objects or arrows and (2) their being the uniquely extreme objects so related. Of course it does not follow from the rules of category theory that there will always be objects satisfying a given universal mapping property. For example, there may be no product of A and B, either because no one object is the source both of an arrow whose target is A and an arrow whose target is B or because, among such objects, there is no unique such, targeted by them all. Shortly we will observe conditions under which a category will always have products for all its objects.

Let us consider now the conception of a logical system L as a category $\mathcal{L}$, as introduced by Joachim Lambek in *1968*, *1969*, and *1972*.[26] In this scheme, ob($\mathcal{L}$) are the formulas of L, and hom($\mathcal{L}$) are derivations. In particular hom$_{\mathcal{L}}$(A, B) is the collection of all derivations from A to B. One must identify certain formal derivations in L for this scheme to satisfy the rules of categories.[27] When one identifies all formal derivations from A to B with a single arrow, $\mathcal{L}$ is just the partially ordered set of *provability*, and hom($\mathcal{L}$) is just the consequence relation on L-formulas. But the flexibility of the categorial framework allows an investigation of more nuanced relationships among *proofs* that are hidden by the consequence relation. As Došen emphasized, '[in] categorial proof theory we are not concerned with a consequence relation, but with a consequence graph, where more than one arrow, i.e., deduction, can join the same pair of objects, i.e., propositions' (*2006*, p. 643). To bring these more nuanced relationships to the fore, one may identify, e.g., only formal derivations that resolve to the same normal form under the familiar normalization procedures of natural deduction. Thus the

[26] A particularly lucid introduction to this conception, from the motivation of its basic framework to open questions about the coherence of its individuation criteria involved, is *Harnick and Makkai 1992*.

[27] The references in this paragraph explain why.

arrows in $\hom_{\mathcal{L}}(\mathrm{A}, \mathrm{B})$ are *abstract* derivations from A to B: we distinguish two such arrows only according to differences in how they associate with other arrows, though they all are of the same 'type'.[28]

Observe that in $\mathcal{L}$, $\circ$ corresponds to the fact that a derivation of C from the assumption B can be appended to a derivation of B from the assumption A, resulting in a derivation of C from the assumption A. In Frege systems, appending is just concatenation of strings; in natural deduction, all leaf nodes in the proof-tree of C that contain 'open' occurrences of B are replaced with copies of the derivation of B from A. This leads to a reading of claims like 'for all arrows $f : A \longrightarrow B$ there is a unique arrow $z_f : A \longrightarrow C$ for which the diagram $$\begin{tikzcd} & C \arrow[d, "i"'] \\ A \arrow[ur, dashed, "z_f"] \arrow[r, "f"'] & B \end{tikzcd}$$ commutes' as 'any derivation from an assumption to B must actually be a derivation from that same assumption to C followed by some fixed derivation of B from C, the same one that recurs in all such derivations of B'.

Returning now to the concept of the product of A and B as understood in $\mathcal{L}$, we see that such an object (if one exists) is more than just a formula from which it is possible to prove both A and B and which can be proved from any other formula that suffices for derivations of both A and B (as was stressed in §7)—such is the meaning of a product available when one considers only the partially ordered set of *provability*. By generalizing to the category of abstract *proofs*, this same construction means that the derivations of A and B, from any formula C admitting such derivations, are themselves just canonical derivations of A and B from A $\otimes$ B appended to derivations of A $\otimes$ B from C. This example illustrates why $\mathcal{L}$ is the category of abstract proofs rather than concrete formal proofs: Only under the assumption that the f and g in the product diagram are normal does it follow that they must proceed initally to the product (the conjunction of A and B) and then via the projections (the elimination rules for conjunction).

The central concepts from category theory that feature in the version of the Deduction Theorem that we are approaching are structure-preserving maps between categories called functors and a type of relation between functors, given by families of morphisms in their target category, called natural transformations:

Definition A **functor** $\mathbb{F}$ from $\mathcal{C}$ to $\mathcal{D}$ ($\mathbb{F} : \mathcal{C} \longrightarrow \mathcal{D}$) is an assignment of

- an object $\mathbb{F}(A)$ from $\mathrm{ob}(\mathcal{D})$ to each A in $\mathrm{ob}(\mathcal{C})$

[28] The question about how best to individuate proofs is usually traced back to a *1971* lecture of Dag Prawitz (see especially §1). *Došen 2003* is a contemporary survey of the options and obstacles in the way of a satisfactory answer to this question for different logical systems.

- an arrow $\mathbb{F}(f) : \mathbb{F}(A) \longrightarrow \mathbb{F}(B)$ from hom($\mathcal{D}$) to each $f : A \longrightarrow B$ in hom($\mathcal{C}$)

that 'preserves' units and compositionality, i.e.:

1. for all A in ob($\mathcal{C}$), $\mathbb{F}(1_A) = 1_{\mathbb{F}(A)}$
2. for all $f : A \longrightarrow B$ and $g : B \longrightarrow C$ in hom($\mathcal{C}$), either
   (a) $\mathbb{F}(g \circ f) = \mathbb{F}(g) \circ \mathbb{F}(f)$ or
   (b) $\mathbb{F}(g \circ f) = \mathbb{F}(f) \circ \mathbb{F}(g)$

The preservation of compositionality of functors can be expressed in terms of commutative diagrams, by saying that whenever the diagram

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[dr, "h"'] & B \arrow[d, "g"] \\
& C
\end{tikzcd}$$

commutes, so too must one of the following diagrams:

$$\begin{tikzcd}
\mathbb{F}(A) \arrow[r, "\mathbb{F}(f)"] \arrow[dr, "\mathbb{F}(h)"'] & \mathbb{F}(B) \arrow[d, "\mathbb{F}(g)\quad\text{(a)}"] \\
& \mathbb{F}(C)
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathbb{F}(A) & \mathbb{F}(B) \arrow[l, "\mathbb{F}(f)"'] \\
& \mathbb{F}(C) \arrow[ul, "\mathbb{F}(h)"] \arrow[u, "\mathbb{F}(g)\quad\text{(b)}"']
\end{tikzcd}$$

$\mathbb{F}$ is called a covariant functor in case (a) and a contravariant functor in case (b).

The idea of a functor is that by preserving identity morphisms and compositionality, the functor will also preserve, in its target category, all the other structural features of its source. Intuitively, one can think of a functor $\mathbb{F} : \mathcal{C} \longrightarrow \mathcal{D}$ as projecting an image of $\mathcal{C}$ into $\mathcal{D}$. Suppose now that there are two such functors, $\mathbb{F} : \mathcal{C} \longrightarrow \mathcal{D}$ and $\mathbb{G} : \mathcal{C} \longrightarrow \mathcal{D}$. One could ask whether the similarity between the images of $\mathcal{C}$ provided by $\mathbb{F}$ and $\mathbb{G}$ is such that one can only detect it 'externally' by verifying the functoriality of $\mathbb{F}$ and $\mathbb{G}$, or if in addition there are arrows in $\mathcal{D}$ that track the similarity 'internally'.

Definition This idea is made precise by the concept of a **natural transformation** between $\mathbb{F}$ and $\mathbb{G}$: a family τ of arrows $\tau_A : \mathbb{F}(A) \longrightarrow \mathbb{G}(A)$ in hom($\mathcal{D}$) (one for each A in ob($\mathcal{C}$)) such that for each $f : A \longrightarrow B$ in hom($\mathcal{C}$), the following diagram commutes:

$$\begin{tikzcd}
\mathbb{F}(A) \arrow[r, "\tau_A"] \arrow[d, "\mathbb{F}(f)"'] & \mathbb{G}(A) \arrow[d, "\mathbb{G}(f)"] \\
\mathbb{F}(B) \arrow[r, "\tau_B"'] & \mathbb{G}(B)
\end{tikzcd}$$

A natural transformation is a family of morphisms that relate $\mathbb{F}$ and $\mathbb{G}$ by systematically shifting the $\mathbb{F}$-image of $\mathcal{C}$ into the $\mathbb{G}$-image of $\mathcal{C}$. (When, further, the inverses of each τ_A together form a natural transformation between $\mathbb{G}$ and $\mathbb{F}$, τ is called a **natural isomorphism**.)

Now, one of the fundamental ideas of category theory is that objects, morphisms, etc., can be individuated in terms just of their 'structural' properties, i.e., the relations they stand in with other objects, morphisms, etc., in their environment. This idea gives rise to the concept of objects being 'isomorphic', i.e., maybe not literally identical, but the same in terms of their structural properties. The definition of isomorphism that captures the idea of sharing all structural properties is very simple:

Definition A morphism $f : A \longrightarrow B$ is an **isomorphism** if there is a morphism $g : B \longrightarrow A$ inverse to f in the sense that $g \circ f = 1_A$ and $f \circ g = 1_B$. Two objects A and B are said to be **isomorphic** if there is an isomorphism between them.

One would like to generalize this definition in a way that extends the idea of structural identity to categories. Two categories $\mathcal{C}$ and $\mathcal{D}$ could be called isomorphic if there is a functor $\mathbb{F} : \mathcal{C} \longrightarrow \mathcal{D}$ with inverse $\mathbb{G} : \mathcal{D} \longrightarrow \mathcal{C}$ such that $\mathbb{G} \circ \mathbb{F}$ and $\mathbb{F} \circ \mathbb{G}$ are the identity functors $\mathbb{1}_{\mathcal{C}}$ and $\mathbb{1}_{\mathcal{D}}$, respectively. In this case we would have, for every $A \in \text{ob}(\mathcal{C})$, $\mathbb{G} \circ \mathbb{F}(A) = A$ and for every $B \in \text{ob}(\mathcal{D})$, $\mathbb{F} \circ \mathbb{G}(B) = B$. But this violates the spirit of the structural approach to individuation, according to which $\mathbb{G} \circ \mathbb{F}(A)$ should be considered the same as A so long as they are isomorphic, even if they are not literally identical. (In Robert Goldblatt's quip, for categories to be structurally the same, they need only be 'isomorphic up to isomorphism' (*2006*, p. 200).)

There are two closely related ways to make this idea precise.

First, one might require that there be natural isomorphisms $\tau : \mathbb{1}_{\mathcal{C}} \longrightarrow \mathbb{G} \circ \mathbb{F}$ and $\sigma : \mathbb{1}_{\mathcal{D}} \longrightarrow \mathbb{F} \circ \mathbb{G}$. This is what is called an equivalence of categories. Alternatively, one might require only 'one-way' natural transformations from $\mathbb{1}_{\mathcal{C}}$ to $\mathbb{G} \circ \mathbb{F}$ and from $\mathbb{F} \circ \mathbb{G}$ to $\mathbb{1}_{\mathcal{D}}$, so that the isomorphism structure is spread out across the two categories. This second option describes what is called an adjoint situation.

Definition An **adjunction** between two categories $\mathcal{C}$ and $\mathcal{D}$ is a pair of functors $\mathbb{L} : \mathcal{C} \longrightarrow \mathcal{D}$ and $\mathbb{R} : \mathcal{D} \longrightarrow \mathcal{C}$ and a pair of natural transformations $\eta : \mathbb{1}_{\mathcal{C}} \longrightarrow \mathbb{R} \circ \mathbb{L}$ and $\varepsilon : \mathbb{L} \circ \mathbb{R} \longrightarrow \mathbb{1}_{\mathcal{D}}$ satisfying the universal mapping properties:

(universality of η) Given $C \in \text{ob}(\mathcal{C})$, $D \in \text{ob}(\mathcal{D})$, and $f \in \text{hom}_{\mathcal{C}}(C, \mathbb{R}(D))$, there is a unique $g \in \text{hom}_{\mathcal{D}}(\mathbb{L}(C), D)$ making this diagram commute:

$$\begin{tikzcd}
C \arrow[r, "\eta_C"] \arrow[dr, "f"'] & \mathbb{R} \circ \mathbb{L}(C) \arrow[d, dashed, "\mathbb{R}(g)"] \\
& \mathbb{R}(D)
\end{tikzcd}$$

(universality of ε) Given $C \in \text{ob}(\mathcal{C})$, $D \in \text{ob}(\mathcal{D})$, and $g \in \text{hom}_{\mathcal{D}}(\mathbb{L}(C), D)$, there is a unique $f \in \text{hom}_{\mathcal{C}}(C, \mathbb{R}(D))$ making this diagram commute:

$$\begin{tikzcd}
\mathbb{L} \circ \mathbb{R}(D) \arrow[r, "\varepsilon_D"] & D \\
\mathbb{L}(C) \arrow[u, dashed, "\mathbb{L}(f)"] \arrow[ur, "g"'] &
\end{tikzcd}$$

When an adjoint situation holds, this is written $\mathbb{L} \dashv \mathbb{R}$, and $\mathbb{L}$ is called the 'left adjoint functor' of $\mathbb{R}$. (Correspondingly, $\mathbb{R}$ is called the 'right adjoint functor' of $\mathbb{L}$.) The natural transformations η and ε are called the 'unit' and 'counit' of the adjunction. As suggested in the motivating remarks above, although neither the unit nor the counit of an adjunction need be natural isomorphisms (one has $\eta : \mathbb{1}_{\mathcal{C}} \longrightarrow \mathbb{R} \circ \mathbb{L}$ but typically no inverse transformation $\mathbb{R} \circ \mathbb{L} \longrightarrow \mathbb{1}_{\mathcal{C}}$; $\varepsilon : \mathbb{L} \circ \mathbb{R} \longrightarrow \mathbb{1}_{\mathcal{D}}$ but no $\mathbb{1}_{\mathcal{D}} \longrightarrow \mathbb{L} \circ \mathbb{R}$), together they give rise to an isomorphism of another sort that captures a type of structural correspondence between $\mathcal{C}$ and $\mathcal{D}$.

Define a natural transformation θ componentwise for $C \in \text{ob}(\mathcal{C})$, $D \in \text{ob}(\mathcal{D})$, and $g \in \text{hom}_{\mathcal{D}}(\mathbb{L}(C), D)$ by $\theta_{CD}(g) = \mathbb{R}(g) \circ \eta_C$ and a dual natural transformation (given D, C, and $f : C \longrightarrow \mathbb{R}(D)$) τ by $\tau_{DC}(f) = \varepsilon_D \circ \mathbb{L}(f)$. By the universality of η and ε, each θ_{CD} and τ_{DC} are inverses of one another (for example, any f just is $\mathbb{R}(g) \circ \eta_C$). Therefore θ is actually a natural isomorphism with components $\theta_{CD} : \text{hom}_{\mathcal{D}}(\mathbb{L}(C), D)) \longrightarrow \text{hom}_{\mathcal{C}}(C, \mathbb{R}(D))$. The unit η is a family of morphisms in $\mathcal{C}$ describing a relation between the functors $\mathbb{R} \circ \mathbb{L}$ and $\mathbb{1}_{\mathcal{C}}$, as the counit ε relates $\mathbb{L} \circ \mathbb{R}$ and $\mathbb{1}_{\mathcal{D}}$ in $\mathcal{D}$. θ operates one level removed from η and ε: It is a family of morphisms in the category of sets, relating functors that map an arbitary $C \in \text{ob}(\mathcal{C})$ to $\text{hom}_{\mathcal{C}}(C, \mathbb{R}(D))$ and to $\text{hom}_{\mathcal{D}}(\mathbb{L}(C), D)$.

Such a relation establishes, not an equivalence between $\mathcal{C}$ and $\mathcal{D}$, but a reflection of the arrow structure of $\mathcal{C}$ in that of $\mathcal{D}$. Because θ is an isomorphism, the adjunction guarantees an exact correspondence between $\mathcal{C}$-arrows and $\mathcal{D}$-arrows: given C and D, any morphism from C to $\mathbb{R}(D)$ is matched uniquely with a morphism from $\mathbb{L}(C)$ to D. Because θ is natural in C and D, the adjunction further guarantees that all categorial structure is preserved as C and D vary smoothly across $\mathcal{C}$ and $\mathcal{D}$ (cf. *Goldblatt 2006*, p. 439).

Returning to the presentation of a logical system as the category $\mathcal{L}$, what would it mean for the structure of derivation from an additional assumption to be captured already by derivations in $\mathcal{L}$? $\mathcal{L}$ contains only arrows sourced at single formulas, so the first stage in phrasing the condition is to construct a category whose objects correspond to multiple formulas and whose arrows correspond to (abstract) derivations from multiple hypotheses. The structure of such a category is what we want to recover within $\mathcal{L}$.

The construction needed is the product category $\mathcal{L} \times \mathcal{L}$. Objects in $\mathcal{L} \times \mathcal{L}$ are pairs $\langle \mathrm{A}, \mathrm{B} \rangle$ of objects from $\mathrm{ob}(\mathcal{L})$, and arrows in $\mathcal{L} \times \mathcal{L}$ are pairs $\langle f, g \rangle$ of arrows from $\mathrm{hom}(\mathcal{L})$. It is important to recognize that, conceived of as a logical system, $\mathcal{L} \times \mathcal{L}$ is a seriously deficient characterization of reasoning from multiple hypotheses. If a formula C follows from formulas A and B taken jointly, though from neither assumption alone, perhaps no arrow in $\mathcal{L} \times \mathcal{L}$ corresponds to this fact. Our interest in $\mathcal{L} \times \mathcal{L}$ is just in its ability to represent having and reasoning from multiple hypotheses, not in whether its scheme includes all such derivations we would deem valid.

To represent pairs of formulas as objects in $\mathcal{L}$ by specifying an adjunction we need a functor $\mathbb{F}$ that maps arbitrary objects $\langle \mathrm{A}, \mathrm{B} \rangle$ in $\mathcal{L} \times \mathcal{L}$ into $\mathrm{ob}(\mathcal{L})$. How this functor should handle arrows, and even which arrows it should operate on, is less obvious. However, the other functor from the adjunction is easy to determine. It must take arbitrary objects from $\mathcal{L}$ into $\mathrm{ob}(\mathcal{L} \times \mathcal{L})$. The only choice that both preserves the source data and takes advantage of the target data-type is the 'diagonal functor' Δ defined by $\Delta(\mathrm{C}) = \langle \mathrm{C}, \mathrm{C} \rangle$, $\Delta(f) = \langle f, f \rangle$. Now consider what it would mean for $\mathbb{F}$ to be left or right adjoint to Δ.

In case $\mathbb{F} \dashv \Delta$, the adjunction between $\mathcal{L} \times \mathcal{L}$ and $\mathcal{L}$ establishes a correspondence between $\mathcal{L} \times \mathcal{L}$-arrows $\langle \mathrm{A}, \mathrm{B} \rangle \longrightarrow \langle \mathrm{C}, \mathrm{C} \rangle$ and $\mathcal{L}$-arrows $\mathbb{F}(\langle \mathrm{A}, \mathrm{B} \rangle) \longrightarrow \mathrm{C}$. Thus, there will be a derivation of C from $\mathbb{F}(\langle \mathrm{A}, \mathrm{B} \rangle)$ precisely when there is a derivation of $\langle \mathrm{C}, \mathrm{C} \rangle$ from $\langle \mathrm{A}, \mathrm{B} \rangle$, i.e., two derivations of C: one from A and one from B. In this adjunction, the object $\mathbb{F}(\langle \mathrm{A}, \mathrm{B} \rangle)$ represents having available (an unknown) one of A and B from which to reason.

In case $\Delta \dashv \mathbb{F}$, the adjunction establishes a correspondence between $\mathcal{L} \times \mathcal{L}$-arrows $\langle \mathrm{C}, \mathrm{C} \rangle \longrightarrow \langle \mathrm{A}, \mathrm{B} \rangle$ and $\mathcal{L}$-arrows $\mathrm{C} \longrightarrow \mathbb{F}(\langle \mathrm{A}, \mathrm{B} \rangle)$. Thus, there will be a derivation of $\mathbb{F}(\langle \mathrm{A}, \mathrm{B} \rangle)$ from C precisely when there are derivations of both A and B from C. In this adjunction, the object $\mathbb{F}(\langle \mathrm{A}, \mathrm{B} \rangle)$ represents having both A and B available as hypotheses.

The right choice for our purpose of representing the idea of reasoning from multiple hypotheses in $\mathcal{L}$ is for $\mathbb{F}$ to be a right adjoint functor to Δ. The counit of this adjoint situation, $\varepsilon : \Delta \circ \mathbb{F}(X) \longrightarrow \mathbb{1}_{\mathcal{L} \times \mathcal{L}}$, is a family of morphisms in $\mathcal{L} \times \mathcal{L}$, i.e., a family of pairs of $\mathcal{L}$-arrows. Given $\langle \mathrm{A}, \mathrm{B} \rangle \in \mathrm{ob}(\mathcal{L} \times \mathcal{L})$, $\varepsilon_{\langle \mathrm{A}, \mathrm{B} \rangle} : \Delta \circ \mathbb{F}(\langle \mathrm{A}, \mathrm{B} \rangle) \longrightarrow \langle \mathrm{A}, \mathrm{B} \rangle$ is just $\langle p_{\mathrm{A}} : \mathbb{F}(\langle \mathrm{A}, \mathrm{B} \rangle) \longrightarrow \mathrm{A}, p_{\mathrm{B}} : \mathbb{F}(\langle \mathrm{A}, \mathrm{B} \rangle) \longrightarrow \mathrm{B} \rangle$. Therefore, the universality of ε, which given $\mathrm{C} \in \mathrm{ob}(\mathcal{L})$, $\langle \mathrm{A}, \mathrm{B} \rangle \in \mathrm{ob}(\mathcal{L} \times \mathcal{L})$, and $\langle f, g \rangle \in \mathrm{hom}_{\mathcal{L} \times \mathcal{L}}(\mathbb{F}(\mathrm{C}), \langle \mathrm{A}, \mathrm{B} \rangle)$ guarantees a unique $h \in \mathrm{hom}_{\mathcal{L}}(\mathrm{C}, \mathbb{F}(\langle \mathrm{A}, \mathrm{B} \rangle))$ making this diagram commute:

$$\begin{tikzcd}
\Delta \circ \mathbb{F}(\langle \mathrm{A}, \mathrm{B} \rangle) \arrow[r, "\varepsilon_{\langle \mathrm{A},\mathrm{B}\rangle}"] & \langle \mathrm{A}, \mathrm{B} \rangle \\
\Delta(\mathrm{C}) \arrow[u, dashed, "\Delta(h)"] \arrow[ur, "{\langle f,g \rangle}"'] &
\end{tikzcd}$$

in fact makes this diagram commute:

$$\begin{tikzcd}
& \mathrm{C} \arrow[dl, "f"'] \arrow[d, dashed, "h"'] \arrow[dr, "g"] & \\
\mathrm{A} & \mathbb{F}(\langle \mathrm{A}, \mathrm{B} \rangle) \arrow[l, "p_{\mathrm{A}}"] \arrow[r, "p_{\mathrm{B}}"'] & \mathrm{B}
\end{tikzcd}$$

and we have rediscovered in $\mathbb{F}$ the product functor. (Analogous reasoning establishes that the left adjoint functor to Δ is the coproduct.)

We have verified the general fact that $\Delta : \mathcal{C} \longrightarrow \mathcal{C} \times \mathcal{C}$ has a right adjoint if, and only if, $\mathcal{C}$ has binary products. The interpretation of this fact for the category of proofs $\mathcal{L}$ is that in order for $\mathcal{L}$ to have the structure needed to represent derivation from multiple assumptions, it must have for every A and B a formula corresponding to the conjunction of A and B—a formula $\mathrm{A} \wedge \mathrm{B}$ from which there are proofs $p_{\mathrm{A}} : \mathrm{A} \wedge \mathrm{B} \longrightarrow \mathrm{A}$ and $p_{\mathrm{B}} : \mathrm{A} \wedge \mathrm{B} \longrightarrow \mathrm{B}$ such that any proofs $c_{\mathrm{A}} : \mathrm{C} \longrightarrow \mathrm{A}$ and $c_{\mathrm{B}} : \mathrm{C} \longrightarrow \mathrm{B}$ must in fact be $p_{\mathrm{A}} \circ c_{\mathrm{A}\wedge\mathrm{B}}$ and $p_{\mathrm{B}} \circ c_{\mathrm{A}\wedge\mathrm{B}}$ for some $c_{\mathrm{A}\wedge\mathrm{B}} : \mathrm{C} \longrightarrow \mathrm{A} \wedge \mathrm{B}$.

Building from this we can ask whether $\mathcal{L}$ has within its arrow structure a reflection of the structure obtained by extending the underlying logic with an additional axiom: Under what conditions is the structure of derivation from conjunctions present already in the structure of derivation from a single conjunct? We seek a natural isomorphism between arrows $\mathrm{A} \wedge \mathrm{B} \longrightarrow \mathrm{C}$ and arrows with source object B. The adjunction giving rise to this isomorphism will have left adjoint functor mapping objects to products. This is the ‘right product functor’ $\Box \otimes \mathrm{B} : \mathcal{L} \longrightarrow \mathcal{L}$ which maps any formula A to $\mathrm{A} \wedge \mathrm{B}$ and any arrow $f : \mathrm{C} \to \mathrm{D}$ to $f \otimes 1_{\mathrm{B}} : \mathrm{C} \wedge \mathrm{B} \longrightarrow \mathrm{D} \wedge \mathrm{B}$. Let us determine the right adjoint functor to $\Box \otimes \mathrm{B}$, customarily denoted $\Box^{\mathrm{B}}$. The counit of the adjunction is again described by a natural transformation from the composition of the left adjunct functor with the right adjunct functor to the identity functor: $\varepsilon^{\mathrm{B}} : \Box^{\mathrm{B}} \otimes \mathrm{B} \longrightarrow \mathbb{1}_{\mathcal{L}}$. This is a family of morphisms in $\mathcal{L}$: Given $\mathrm{C} \in \mathrm{ob}(\mathcal{L})$, $\varepsilon^{\mathrm{B}}_{\mathrm{C}} : \mathrm{C}^{\mathrm{B}} \otimes \mathrm{B} \longrightarrow \mathrm{C}$. By the universality of ε^{B}, given A and $f : \mathrm{A} \otimes \mathrm{B} \longrightarrow \mathrm{C}$, there is a unique $g : \mathrm{A} \longrightarrow \mathrm{C}^{\mathrm{B}}$ making this diagram commute:

$$\begin{tikzcd}
\mathrm{C}^{\mathrm{B}} \otimes \mathrm{B} \arrow[r, "\varepsilon^{\mathrm{B}}_{\mathrm{C}}"] & \mathrm{C} \\
\mathrm{A} \otimes \mathrm{B} \arrow[u, dashed, "g \otimes 1_{\mathrm{B}}"] \arrow[ur, "f"'] &
\end{tikzcd} \tag{5}$$

In the intended interpretation of $\mathcal{L}$, this diagram describes a special derivation $\varepsilon^{\mathrm{B}}_{\mathrm{C}}$ from $\mathrm{C}^{\mathrm{B}} \wedge \mathrm{B}$ to C, special in that any formula A that can 'fill the role' of C^{B} by generating a derivation $f : \mathrm{A} \wedge \mathrm{B} \longrightarrow \mathrm{C}$ must in fact also generate a derivation $g : \mathrm{A} \longrightarrow \mathrm{C}^{\mathrm{B}}$ such that f itself is just $\varepsilon^{\mathrm{B}}_{\mathrm{C}}$ appended to a 'copy' of g (a derivation of $\mathrm{C}^{\mathrm{B}} \wedge \mathrm{B}$ from $\mathrm{A} \wedge \mathrm{B}$ that is essentially g inside a trivial proof routine for removing the right conjunct B and then reattaching it).

Putting everything together, we asked first: Under what conditions is $\mathcal{L}$ able to represent derivations from multiple assumptions? The answer was: When for every pair of formulas A and B there is an object $\mathrm{A} \wedge \mathrm{B}$ satisfying a universal mapping property recognizable as the introduction and elimination rules for conjunction. Second: When is $\mathcal{L}$ 'complete' in the sense that each derivation from an extra assumption is 'mirrored' by an implication? The answer was given by another universal mapping property: When for every pair of formulas A and B, there is an exponential formula B^{A} such that

- B^{A} together with A allows a derivation of B, and
- from any other formula C that, together with A, allows a derivation of B, there is a derivation of B^{A}.

But these are just the elimination and introduction rules for $\mathrm{A} \supset \mathrm{B}$! Thus the inferential meaning of $\supset$ can be 'derived' from the stipulation that the language of L contain all the structure of its derivational apparatus.

However, the version of the Deduction Theorem given by (5) expresses more than (3). It says that conjunction and conditionalization are adjoint functors, so that there is a natural isomorphism θ^{B} between $\hom_{\mathcal{L}}(\mathrm{A} \wedge \mathrm{B}, \mathrm{C})$ and $\hom_{\mathcal{L}}(\mathrm{A}, \mathrm{B} \supset \mathrm{C})$. Thus, for any fixed B, the bijections $\theta^{\mathrm{B}}_{\mathrm{AC}}$ are natural, meaning that the derivational structure of $\hom_{\mathcal{L}}(\mathrm{A} \wedge \mathrm{B}, \mathrm{C})$ is preserved as A and C vary smoothly. That is the sense, obscured by some proofs of (2) and hinted at in others, in which the Deduction Theorem is more than a biconditional correspondence between provability claims or even proofs: It is a depiction of how the whole network of proofs of one type is reflected in the proofs of another type.

10. Hindsight

The thought of architects of modern logic such as Bolzano, Frege, Tarski, Herbrand, and Gentzen can only be fully appreciated from a retrospective point of view, informed by their contemporaries' rival conceptions and by later developments of their work that they could not have imagined themselves. It is possible to derive the Deduction Theorem as a property of a canonical proof system, *à la* Herbrand. And if this is all one knows about the theorem, Frege's use of the property to demonstrate theorems in his own system appears unlicensed, careless, and above all, mysterious.

But as Tarski showed, it is also possible to impose the Deduction Theorem as a constraint on any reasonable logical system. Tarski's abstract perspective has been hard for historians to appreciate, as they searched for clues that he had a concrete demonstration of the fact for some particular system—a proof that could be compared with Herbrand's. But when one sees in the modern theory of admissibility how several distinct consequence relations can generate the same set of theorems, Tarski's abstract approach is just what is called for: It is a condition, over and above the definition of a consequence relation, needed to pick out logics that adequately capture the concept of propositional inference.

There are, it turns out, two different ways to motivate the Deduction Theorem as a principle of logic, rather than a property of a specific system. First, one could see it as the 'other side of the coin' of *modus ponens*, as was Gentzen's idea with natural deduction, where one views *modus ponens* and the Deduction Theorem as together defining conditionalization *via* a universal mapping property. According to this perspective, the Deduction Theorem stands no more in need of justification than *modus ponens* itself, and in fact the same analysis of conditional language justifies them both simultaneously. For Gentzen, this perspective was hard-won against a current of thinking that distinguished logic's object-level and meta-level, according to which axioms and rules of the first sort can be engineered in light of conceptual analysis whereas relations of the second sort stand in need of mathematical verification. Equipped with hindsight, we see that Gentzen's dissolution of this distinction led him back to the perspective of Bolzano, for whom defining conditionals simultaneously in terms of *modus ponens* and the Deduction Theorem was natural.

Alternatively, the Deduction Theorem can arise indirectly, simply by stipulating that the structure of reasoning from additional hypotheses be present in a logic's theorem structure. On this route to the Deduction Theorem, rather than being imposed from the outside, conjunction and conditionalization are discovered as the functors required to make the desired structure arise. But amazingly, these two constraints turn out to be interderivable: The unique right adjunct to product is exponentiation, and, conversely, if exponentiation is definable for all pairs, then product has a right adjunct.

Perhaps most surprisingly, this understanding of the Deduction Theorem as an adjunction is the one most easily read back into Frege's thought. We asked how Frege could have assumed the Deduction Theorem and advocated its use without verification. One answer that cannot be given is that Frege defined his conditional stroke in terms of the Deduction Theorem, as Bolzano had done decades before and Gentzen would do again as many years later. Frege had other, explicit definitions of the conditional stroke on offer, and the same question could just be rephrased as a puzzle about why he was confident that his multiple definitions align. On the other hand, as a property of the propositional fragment of *Begriffsschrift*, the Deduction Theorem stands in need of just the sort of verification Herbrand would eventually provide and Frege clearly did not seek.

But if Frege simply conceived of the system of *Begriffsschrift* as deductively complete and structurally complete for the same reason he conceived of it as semantically complete, his appeal to the identity (2) makes perfect sense. By semantic completeness, all laws of thought are theorems. By structural completeness, those same laws of thought are available for 'deducing' things from arbitrary hypotheses, even if this can't be described as 'inference'. And by deductive completeness, no relations between thoughts uncovered through these deductions could fail to be expressed again as theorems, for theorems simply reflect those very laws of deduction. Whether Frege's (*1883*) insistence that the *Begriffsschrift* is 'not a mere *calculus ratiocinator*' but instead 'a *lingua characteristica* in the Leibnizian sense' justifies any of these assumptions is better left for others to debate. However that debate might resolve, the rediscovery of the Deduction Theorem as an adjunction provides the hindsight to make sense of Frege's remarks.

References

Barwise, J., J. Etchemendy, and D. Barker-Plummer. 2011. *Language, Proof, and Logic* (Second Edition). CSLI.

van Benthem, J. 1985. ‘The variety of consequence, according to Bolzano’, *Studia Logica*, **44**(4), 389–403.

Bernays, P. 1918. *Beiträge zur axiomatischen Behandlung des Logik-Kalküls*, Habilitationsschrift, University of Göttingen.

Bolzano, B. 1837. *Wissenschaftslehre. Versuch einer ausführlichen und größtentheils neuen Darstellung der Logik mit steter Rücksicht auf deren bisherige Bearbeiter.* Sulzbach: Seidel. Trans. by B. Terrel in B. Bolzano *Theory of Science*, J. Berg (ed.) 1973. Boston: D. Reidel Publishing Company.

Bynum, T. W. (ed.) 1972. *Conceptual Notation and Related Articles*, New York: Oxford University Press.

Chagrov A. V. 1992. ‘A decidable modal logic with the undecidable admissibility problem for inference rules’, *Algebra and Logic*, **31**, 53–55.

Church, A. 1947. ‘Review of W. V. Quine *A Short Course in Logic*’, *Journal of Symbolic Logic*, **12**(2), 60–61.

Czelakowski, J. 1985. ‘Algebraic aspects of deduction theorems’, *Studia Logica*, **44**(4), 369–87.

Došen, K. 1996. ‘Deductive completeness’, *Bulletin of Symbolic Logic*, **2**(3), 243–83.

Došen, K. 2003. ‘Identity of proofs based on normalization and generality’, *Bulletin of Symbolic Logic*, **9**(4), 477–503.

Došen, K. 2006. ‘Models of deduction’, *Synthese*, **148**(3), 639–57.

Ewald, W. and W. Sieg (eds.) 2010. *David Hilbert’s Lectures on the Foundations of Arithmetic and Logic, 1917-1933*, vol. 3. Springer.

Feferman, S., J. W. Dawson, Jr., S.C. Kleene, G. H. Moore, R. M. Solovay, and J. van Heijenoort (eds.) 1986. *Kurt Gödel, Collected Works, Vol I: Publications 1929-1936*, New York: Oxford University Press.

Franks, C. 2010. ‘Cut as consequence’, *History and Philosophy of Logic*, **31**(4), 349–79.

Franks, C. 2014. ‘Logical completeness, form, and content: an archaeology’, in J. Kennedy (ed.) *Interpreting Gödel: Critical Essays*. New York: Cambridge University Press.

Franks, C. 2017. ‘Hilbert’s logic’, in A. P. Malpass and M. Antonutti-Marfori (eds.) *The History of Philosophical and Formal Logic*. Bloomsbury.

Franks, C. 2018, ‘The context of inference’, *History and Philosophy of Logic*, **39**(4), 365–95.

Frege, G. 1879. *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens* Halle: L. Nebert. Translated by T. W. Bynum as *Conceptual Notation: a formula language of pure thought modeled upon the formula language of arithmetic* in *Bynum 1972*, 101-208.

Frege, G. 1883. ‘Über den Zweck der Begriffsschrift’, *Sitzungsberichte der Jenaischen Gesellschaft für Medicin und Naturwissenschaft, JZN* **16**, 1-10. Translated by T. W. Bynum as ‘On the aim of the “Conceptual Notation”’ in *Bynum 1972*, 90-100.

Frege, G. 1910. ‘Letter to Jourdain’, translated and reprinted in *Frege 1980.*

Frege, G. 1917. ‘Letter to Dingler’, translated and reprinted in *Frege 1980.*

Frege, G. 1980. *Philosophical and Mathematical Correspondence*, G. Gabriel, et al. (eds.) Oxford: Blackwell Publishers.

Gentzen, G. 1932. ‘Über die Existenz unabhängiger Axiomensysteme zu unendlichen Satzsystemen’, *Mathematische Annalen* **107**, 329-50. Translated as ‘On the existence of independent axiomsystems for infinite sentence systems’ in *Szabo 1969*, 29-52.

Gentzen, G. 1934–35. ‘Untersuchungen über das logische Schließen’. Gentzen’s doctoral thesis at the University of Göttingen, translated as ‘Investigations into logical deduction’ in *Szabo 1969*, 68-131.

Gödel, K. 1931. ‘Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I’, *Monatshefte für Mathematik und Physik* **38**, 173-98, translation by J. van Heijenoort as ‘On formally undecidable propositions of *Principia Mathematica* and related systems I’ reprinted in *Feferman et al. 1986*, 144-95.

Gödel, K. 1929. ‘Über die Vollständigkeit des Logikkalküls’, Gödel’s doctoral thesis at the University of Vienna, translation by S. Bauer-Mengelberg and Jean van Heijenoort as ‘On the completeness of the calculus of logic’ reprinted in *Feferman et al. 1986*, 60-101.

Goldblatt, R. 2006. *Topoi: the Categorical Analysis of Logic*. Revised edition. Dover.

Goldfarb, W. (ed.) 1971. *Jacques Herbrand: Logical Writings*. Cambridge: Harvard University Press.

Harnick, V. and M. Makkai. 1992. ‘Lambek’s categorical proof theory and Läuchli’s abstract realizability’, *Journal of Symbolic Logic*, **57**(1), 200–230.

Harrop, R. 1956. ‘On disjunctions and existential statements in intuitionistic systems of logic’, *Mathematische Annalen*, **132**, 347–61.

Herbrand, J 1929. ‘Sur quelques propriétés des propositions vraies et leurs applications’. Translated by W. Goldfarb as ‘On several properties of true propositions and their applications’ in *Goldfarb 1971*, 38–40.

Herbrand, J. 1930. *Recherches sur la théorie de la démonstration*. Herbrand’s doctoral thesis at the University of Paris. Translated by W. Goldfarb, except pp. 133-88 trans. by B. Dreben and J. van Heijenoort, as ‘Investigations in proof theory’ in *Goldfarb 1971*, 44-202.

Hertz, P. 1929. ‘Über Axiomensysteme für beliebige Satzsysteme’, *Mathematische Annalen*, **101**, 457–514.

Iemhoff, R. 2016. ‘Consequence relations and admissible rules’, *Journal of Philosophical Logic*, **45** (3), 327–48.

Jaśkowski, S. ‘On the rules of supposition in formal logic’, *Studia Logica*, **1**, 5–32.

Johansson, I. 1937. ‘Der Minimalkalkül, ein reduzierter intuitionistischer Formalismus’. *Compositio Mathematica* **4**, 119-136.

Kaye, R. 2007. *The Mathematics of Logic*. New York: Cambridge University Press.

Kreisel, G. and H. Putnam 1957. ‘Eine Unableitbarkeitsbeweismethode für den Intuitionistischen Aussagenkalkül’, *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* **3**, 74–78.

Lambek, J. 1968. ‘Deductive systems and categories I’, *Mathematical Systems Theory*, **2**, 287–318.

Lambek, J. 1969. ‘Deductive systems and categories II: Standard constructions and closed categories’, *Lecture Notes in Mathematics (Category theory, homology theory and their applications)*, **86**. Berlin: Springer-Verlag, 76–122.

Lambek, J. 1972. ‘Deductive systems and categories III: Cartesian closed categories, intuitionist propositional calculus and combinatory logic’, *Lecture Notes in Mathematics (Toposes, algebraic geometry and logic)*, **274**. Berlin: Springer-Verlag, 57–82.

Lambek, J. 1974. ‘Functional completeness of cartesian categories’, *Annals of Mathematical Logic*, **6**, 259–92.

Lambek, J. and P. J. Scott. 1986. *Introduction to Higher-order Categorical Logic*. New York: Cambridge University Press.

Łukasiewicz, J. and A. Tarski. 1930. ‘Untersuchungen über den Aussagenkalkül’. *Comptes* *rendus de la Société des sciences et des lettres de Varsovie*, cl. iii, **23**, 1–21. Translated by J. H. Woodger as ‘Investigations into the sentential calculus’ in *Tarski 1956*, 38–59.

van der Molen, T. 2016. ‘The Johansson/Heyting letters and the birth of minimal logic’, Institute for Logic, Language, and Computation, Amsterdam.

Pogorzelski, W. A. 1968. ‘On the scope of the classical deduction theorem’, *Journal of Symbolic Logic*, **33**(1), 77–81.

Porte, J. 1982. ‘Fifty years of deduction theorems’, *Studies in Logic and the Foundations of Mathematics*, **107**, 243–50.

Prawitz, D. 1971. ‘Towards a foundation of a general proof theory’, in J. E. Fenstad (ed.) *Ideas and Results in Proof Theory*. Amsterdam: North Holland. 235–307.

Quine, W. V. O. 1950. ‘On natural deduction’, *Journal of Symbolic Logic*, **15**(2), 93–102.

Quine, W. V. O. 1951. *Mathematical Logic* (Revised Edition). Cambridge: Harvard University Press.

Quine, W. V. O. 1995. *Selected Logic Papers*. Enlarged edition. Cambridge: Harvard University Press.

Rose, G. F. 1953. ‘Propositional calculus and realizability’, *Transactions of the American Mathematical Society*, **75**, 1–19.

Rybakov, V. 1997. *Admissibility of logical inference rules*. Amsterdam: Elsevier.

Šebestik, J. 2016. ‘Bolzano’s logic’, *Stanford Encyclopedia of Philosophy*: https://plato.stanford.edu/entries/bolzano-logic/

Szabo, M. E. 1969. *The Collected Papers of Gerhard Gentzen*, London: North Holland.

Tarski, A. 1930. ‘Über einige fundamentale Begriffe der Metamathematik’. Translated by J. H. Woodger as ‘On some fundamental concepts of metamathematics’ in *Tarski 1956*, 30–37.

Tarski, A. 1933. ‘Einige Betrachtungen über die Begriffe ω-Widerspruchsfreiheit und der ω-Vollständigkeit’, *Monatshefte für Mathematik und Physik*, **40**.

Tarski, A. 1936. ‘On the concept of logical consequence’, first English translation of a 1935 address at the International Congres of Scientific Philosophy in Paris, in *Tarski 1956*.

Tarski, A. 1956. *Logic, Semantics, Metamathematics*. J. H. Woodger (trans.). New York: Oxford University Press.