ℵ Modern Logic ω

HERBRAND'S ERROR AND GÖDEL'S CORRECTION*

Dedicated to the memory of Jean van Heijenoort (1912 – 1986)

Warren GOLDFARB

Department of Philosophy
Harvard University
Cambridge, MA 02138, USA

Abstract. Background on the Herbrand Theorem and Herbrand's error is given in §1 below. §2 presents Gödel's correction. As it turns out, his proof is extremely close to that devised by Dreben and his co-workers; published in full in [Dreben & Denton *1966*] (and, somewhat condensed, in [van Heijenoort *1967*, 572-576] and [Herbrand *1971*, 193-199]). §3 concerns counterexamples to Herbrand's false lemma and numerical bounds for the corrected lemma. In particular, I show that the bound obtained by Gödel's argument cannot be significantly improved.

Jacques Herbrand and Kurt Gödel were two principal subjects of Jean van Heijenoort's editing, writing and research. The keystone of Herbrand's logical achievement is Chapter 5 of his doctoral thesis [Herbrand *1930*], in which Herbrand presents his *theorème fondamentale*, since known as the Herbrand Theorem. Recently, material in Gödel's mathematical notebooks has come to light which shows that Gödel, too, studied this chapter closely. This paper is devoted to an account of Gödel's examination of Herbrand.

Herbrand's proof of his Theorem in Chapter 5 contains a serious error. The error was not noticed in the 1930's, apparently because Herbrand's thesis was not closely read, although Bernays did note that "Herbrand's proof is hard to follow" [Hilbert & Bernays *1939*, 158]. In that period, other logicians proved variants of the Herbrand Theorem by using alternative proof-theoretic strategies. The first published notice of the error was not until [Dreben, Andrews and Aanderaa *1963*], and a corrected version of the Herbrand's false lemma was first announced in [Dreben *1963*]. However, it turns out that Gödel had detected the flaw in Herbrand's argument twenty years earlier; moreover, he too had

* This paper was intended for the special issue devoted to van Heijenoort, but regrettably was not received in time.

arrived at a correct replacement for the false lemma. But Gödel neither published anything on the matter, nor, it seems, mentioned it to anyone until 1963, after notice of the error was published, when he told van Heijenoort that he had found a lacuna in Herbrand's argument in 1943 and had written a note on the matter "for his own personal use" (see [Herbrand *1967*, 8] or [van Heijenoort *1985a*, 121]). Even then he did not hint that he had devised a correction.

Background on the Herbrand Theorem and Herbrand's error is given in §1 below. §2 presents Gödel's correction. As it turns out, his proof is extremely close to that devised by Dreben and his co-workers; published in full in [Dreben & Denton *1966*] (and, somewhat condensed, in [van Heijenoort *1967*, 572-576] and [Herbrand *1971*, 193-199]). §3 concerns counterexamples to Herbrand's false lemma and numerical bounds for the corrected lemma. In particular, I show that the bound obtained by Gödel's argument cannot be significantly improved.

§1. Herbrand's Error. The basic idea of the Herbrand Theorem may be stated fairly easily. With each quantified formula F Herbrand associated a (usually infinite) sequence of quantifier-free formulas, each longer than the proceeding one and truth-functionally implied by it. These formulas are now called Herbrand (validity) expansions of F or Herbrand disjunctions of F. The Theorem states that F is derivable in a standard system of quantification theory if and only if there exists an Herbrand expansion of F that is truth-functionally valid. A nonconstructive proof of this is not difficult. But Herbrand, as a strict follower of the Hilbert school, was interested only in finitistic arguments. For him, it was essential that the proof be constructive: that one be able to calculate from any given derivation of F the size of an Herbrand expansion that would be valid, and that from any valid Herbrand expansion one be able to construct a derivation of F. The latter can be done fairly straightforwardly, and the end result is a derivation of F without use of *modus ponens* or any similar cut-rule. Hence a proof of the Herbrand Theorem also provides, as a byproduct, a proof of cut-elimination. It is the former step, going from derivation to valid expansion, that is difficult.

Herbrand's argument for this part of the Theorem is inductive. He shows that each axiom of the quantificational system has a valid expansion of calculable size. Then, for each inference rule of the system, he argues that given premises of that inference rule which have valid Herbrand expansions of certain sizes, we can calculate the size of the valid Herbrand expansion of any formula that can be obtained from those premises by the inference rule. In its "context-independent" treatment of inference rules, this form of argument is different from that used by Gentzen for his Hauptsatz and Hilbert and Bernays for the ϵ-Theorems; in their proofs, derivations are manipulated as wholes.

Not too surprisingly, the difficulty in Herbrand's treatment surfaces with respect to *modus ponens*. The measure of the size of an expansion that Herbrand used is called its *order*. Herbrand thought he had proved that if F and $F \supset G$ have valid Herbrand expan-

sions of orders p and q, respectively, then G will have a valid Herbrand expansion of order $max(p,q)$. His ingenious argument has three parts:

(1) if F and $F \supset G$ have valid Herbrand expansions of orders p and q, then the formula $F \wedge (F \supset G)$ has a valid expansion of order $max(p,q)$;

(2) if a formula has a valid expansion of order p then so do all of its prenex forms;

(3) if a certain prenex form of $F \wedge (F \supset G)$ has a valid expansion of order p, then so does G.

Steps (1) and (3) are correct; (2) is false. Hence there must be a flaw in the lemmas that assert the following for each of the usual prenexing rules: if F has a valid expansion of order p and if G comes from F by application of the rule, then G also has a valid expansion of order p. Since the logical primitives of Herbrand's system of quantification theory are "~", "∨", "∀" and "∃", the rules in question are the two that allow a quantifier to be pulled across a negation sign and the two that allow, inside a formula F, a subformula $(\exists x)\Phi(x) \vee P$ to be replaced by $(\exists x)(\Phi(x) \vee P)$ and $(\forall x)\Phi(x) \vee P$ to be replaced by $(\forall x)(\Phi(x) \vee P)$, where P does not contain free x (as well as the symmetrical rules with the order of disjuncts reversed). Now, Herbrand's lemmas are correct for the rules about negation and for the rules pulling out an existential quantifier. For the rule pulling out a universal quantifier, if the subformula in question occurs positively in P the lemma is trivial, since the Herbrand expansions of F and of the formula obtained from P are identical. Herbrand's error occurs with respect to this rule when the subformula occurs negatively. (Incidentally, Herbrand also shows the analogous lemmas for these rules applied in the reverse direction, i.e., as antiprenexing rules. Here Herbrand's proofs are all correct.)

Gödel, in his consideration of Herbrand, uses a dual form of the Herbrand theorem. Here one defines Herbrand *satisfiability* expansions. The theorem to be proved is: a formula F is refutable in a standard system of quantification theory if and only if there exists an Herbrand satisfiability expansion of F that is truth-functionally unsatisfiable.

We follow Herbrand in treating a language whose logical primitives are: ~, ∨, ∀, and ∃. Let F be a formula; we assume no variable occurs both bound and free in F, and no variable is bound by two different quantifier-occurrences. A bound variable is said to be *restricted* iff the quantifier binding it is either existential and within the scopes of an even number of negation signs (possibly 0) or universal and within the scopes of an odd number of negation signs. A bound variable is said to be *general* if it is not restricted. The (satisfiability) *functional form* F^* of F is obtained from F by deleting all quantifiers, replacing each free variable of F by a new and distinct constant, and replacing each

restricted variable v of F by a term $f(w_1, ..., w_i)$, where f is a new function sign, said to be associated with the variable v, and $w_1, ..., w_i$ are the general variables governing v. If v is governed by no general variables, then f is 0-place, i.e., a constant.

The *height* of a constant is 1; the *height* of a term $f(t_1,..., t_j)$, $j > 0$, is one greater than the maximum of the heights of $t_1,..., t_j$. For each $p > 0$, the Herbrand *domain* $D(F,p)$ is the finite set of terms of height $\leq p$ constructed from the function signs that appear in F^*, and a special initial sign 0 (included to insure that there is a place to start); the sign 0 is stipulated to have height 0.

The *Herbrand satisfiability expansion* $E(F,p)$ of F of order p is the conjunction of all instances of F^* over $D(F,p)$, that is, the conjunction of all formulas obtained from F^* by substituting terms in $D(F,p)$ for its free variables. In this dual form, Herbrand's false lemma states:

> *Let G come from F by replacing a positively occurring subformula $(\forall x)(\Phi(x) \vee P)$ with $(\forall x)\Phi(x) \vee P$, where x does not occur in P. Then for any p, if $E(F,p)$ is truth-functionally satisfiable, so is $E(G,p)$.*

The difficulty stems from the fact that the functional forms F^* and G^* differ; the terms that replace restricted variables whose quantifiers lie in P will have one more argument place in F^* than they have in G^*, since they are governed by the general variable x in F but not in G. Consequently, there will be instances of G^* in which the subformulas corresponding to P are the same, while in the analogous instances of F^* those subformulas are different, due to a difference in the term that supplants x.

This difference between the expansions of F and G would not matter if the subformula contained $\wedge$ rather than $\vee$. In that case, a truth-functional argument shows that in every instance of F^* we can replace the subformula corresponding to P by a subformula in which the substituent for x is 0, while preserving the satisfiability of the expansion. That is, suppose A is any truth-assignment that makes $E(F,p)$ true; let H be an instance of F^* over $D(F, p)$, and let H' be the instance of F^* in which the substituents for all the variables except x are the same as in H but the substituent for x is 0. Let $\Phi_1 \wedge P_1$ and $\Phi_2 \wedge P_2$ be the subformulas of these instances that correspond to $(\forall x)(\Phi(x) \wedge P)$. Finally, let J result from H by replacing P_1 with P_2. Then A must make J true. For, since A makes H and H' true, it could fail to make J true only if $\Phi_1 \wedge P_1$ and $\Phi_2 \wedge P_2$ were true but $\Phi_1 \wedge P_2$ and $\Phi_2 \wedge P_2$ were false. This, of course, is impossible. Given this truth-functional fact, it is easy to show that the truth-assignment A can be transformed into one that makes $E(G,p)$ true.

Obviously, this truth-functional argument fails if the subformula contains $\vee$ rather than $\wedge$. The desired conclusion, that $\Phi_1 \vee P_2$ is true, will follow only if it is known that one of the disjuncts is true. It would be enough to know that either Φ_1 is true or Φ_2 is false, since

in the latter case the truth of $\Phi_2 \vee P_2$ will yield that of P_2. The key to the correction to Herbrand is to find terms that, as substituents for the variable x, engender false subformulas Φ_2 whenever possible. But these terms may well have positive height; the result will be a dramatic increase in the order of the expansion of F whose satisfiability is needed to guarantee the satisfiability of $E(G,p)$.

§2. Gödel's Correction. Gödel's *Nachlaß* contains sixteen mathematical notebooks, labelled "Arbeitshefte". The prose in them, in German, is written for the most part in the Gabelsberger shorthand that Gödel used in much of his note-takng and some of his manuscripts.[1] They are not dated; nor were they written consecutively.

Most of the material on Herbrand lies in Arbeitsheft 5, pp. 14-79. On p. 22 the Fundamental Theorem is cited, and the crucial lemma noted with the remark "the lemma itself is probably false." Gödel concludes the page, "For the following pages 23-33, the mistake in Herbrand was not yet known to me." Those pages contain parts of a proof of the Theorem, including the beginning of an argument for the false lemma. On page 33 Gödel gives the expansions of order 2 of formulas $(\forall u)(\forall x)[\phi(u,x) \vee (\exists v)\psi(u,v)]$ and $(\forall u)[(\forall x)\phi)(u,x) \vee (\exists v)\psi(u,v)(u,v)]$, and tries to see if the expansion of the latter could be unsatisfiable while that of the former is satisfiable. This is headed "Attempt at a refutation of Herbrand's Lemma". It is not successful; a marginal note says "A refutation is impossible with this formula; compare AH 4, p. 29." At that place there is a short and correct argument that, for such simple formulas, if an expansion of order p of the former formula is satisfiable then so is the expansion of order p of the latter. Thus, a counterexample to Herbrand's false lemma would have to be more complex. Another attempt, in Arbeitsheft 5, pp. 14-17 is headed "Herbrand, attempt [to find] a counterexample for his false lemma," but here too there is nothing close to an example that will work. The title "Correction to Herbrand" near the bottom of Arbeitsheft 5, page 35 marks the beginning af Gödel's proof of a corrected version of the false lemma, which continues until page 53. Pages 53-79 then complete the poof of the Herbrand Theorem, following Herbrand's plan but incorporating the quantitative changes necessitated by the revised lemma.

The material on pages 35-53 is rough, but it does contain a correct proof. In the presentation of Gödel's argument below, I have filled in small lacunae and have altered the notation in the interests of consistency and readability. Moreover, I have modified two aspects of his proof more substantively. First, Gödel uses Herbrand's original formulation of the notion of expansion ("réduite"), which is truth-functionally equivalent to that given in §1 but has a far less perspicuous syntactical structure. (Herbrand's original notion of *réduite* and its relation to the standard notion of expansion are discussed in [van Heijenoort *1967*, 543-544 and 572-573] = [Herbrand *1971*, 153 and 194-195].) Herbrand started using the notion of §1 in 1931, and it has since become standard. Its use below necessitates

[1]The portions of Gödel's notebooks dealing with Herbrand were transcribed by Cheryl Dawson.

a truth-functional argument for Lemma 1 slightly different from Gödel's. Second, Gödel, following Herbrand, does not deal directly with the domains $D(F,p)$, but has the terms in them take values in what he calls "formal domains". This figures in the proof of Proposition 1 below. In Gödel's formulation the problem is to find, given such values for the terms occurring in the expansion of F, suitable values for the terms occurring in the expansion of G. I have reformulated the argument to eliminate the "formal domains"; instead, terms in the expansion of G are "evaluated" by mapping them directly to terms in the expansion of F.

> GÖDEL'S CORRECTED LEMMA. *Let G come from F by replacing a positively occurring subformula $(\forall x)(\Phi(x) \vee P)$ with $(\forall x)\Phi(x) \vee P$, where x does not occur in P. Let r be the number of general variables governing the subformula $(\forall x)(\Phi(x) \vee P)$ in F, let $L = 2 +$ the number of function signs in $F^* +$ the maximum number of argument places of these function signs, and let $N = 4p \cdot 2^r L^{2p}$. Then, for any p, if $E(F,N)$ is truth-functionally satisfiable, so is $E(G,p)$.*

We use the following notations. Let $x_1, ..., x_r$ be the general variables governing $(\forall x)(\Phi(x) \vee P)$ in F. Let $\Phi^*(x_1,..., x_r, x, y_1,...,y_m)$ and $P^*(x_1,..., x_r, x, z_1,..., z_n)$ represent what $\Phi(x)$ and P become in F^*; the y_i are general variables quantified inside $\Phi(y)$, the z_i are those quantified inside P. If H is an instance of F^* and E is a formula, then $H[E]$ is the formula obtained from H by substituting E for the subformula of H that corresponds to $(\forall x)(\Phi(x) \vee P)$.

DEFINITION. Let A be any truth-assignment to an expansion $E(F,q)$. A term t is *favorable* for $\langle t_1,...,t_r\rangle$ in $D(F,q)$ iff either

(1) A makes $\Phi^*(t_1,..., t_r, t, s_1,..., s_m)$ true for all $s_1, ..., s_m$ in $D(F,q)$;

or

(2) A makes $H[\bot]$ true whenever H is an instance of F^* over $D(F,q)$ in which the substituents for $y_1, ..., y_r$ are $t_1, ... t_r$.

Gödel calls this the "most important definition." Its second clause says that the falsity of some $\Phi^*(t_1,..., t_r, t, s_1,..., s_m)$ is inessential in the sense that instances of F^* that contain it will be true anyway. Gödel exploits this clause in the following Lemma.

LEMMA 1. *Let A be a truth-assignment that makes E(F,q) true. If t is unfavorable for* $(t_1,..., t_r)$ *in D(F,q), then A makes* $P^*(t_1,..., t_r, t, u_1,..., u_n)$ *true for all* $u_1,..., u_n$ *in D(F,q).*

PROOF. Let $u_1,..., u_n$ be any terms in $D(F,q)$. From the hypothesis, A makes $\Phi^*(t_1,..., t_r, t, s_1,..., s_m)$ false for some $s_1,..., s_m$ in $D(F,q)$, and A also makes $H[\perp]$ false for some instance H of F^* over $D(F,q)$ in which the substituents for $y_1,..., y_r$ are $t_1,..., t_r$. Let $J = H[\Phi^*(t_1,..., t_r, t, s_1,..., s_m) \vee P^*(t_1,..., t_r, t, u_1,..., u_n)]$. Then J is also an instance of F^* over $D(F,q)$, so that A makes J true. Since A makes $H[\perp]$ false, A must make $\Phi^*(t_1,..., t_r, t, s_1,..., s_m) \vee P^*(t_1,..., t_r, t, u_1,..., u_n)$ true. But since A makes the first disjunct false, A must make the second disjunct true. ■

DEFINITIONS. (1) A ϕ-*domain* consists of a truth-assignment A, a domain $D(F,q)$, and a function $\phi: D(F,\omega)^r \rightarrow D(F,\omega)$ such that A makes $E(F,q)$ true and, for all $t_1,..., t_r$ in $D(F,q)$, either $\phi(t_1,..., t_r) = 0$ or $\phi(t_1,..., t_r)$ is unfavorable for $\langle t_1,..., t_r\rangle$ in $D(F,q)$.

(2) Given $\phi: D(F,\omega)^r \rightarrow D(F,\omega)$, the sets $M_i(\phi)$, are defined inductively thus: $M_0(\phi) = \{0\}$; $M_{i+1}(\phi) = \{f(t_1,..., t_k) \mid f$ a function sign of F^* and $t_1,..., t_k \in M_i(\phi)\} \cup \{\phi(t_1,..., t_r) \mid t_1,..., t_r \in M_i(\phi)\}$.

(3) A ϕ-*domain* $\langle A, D(F,q), \phi\rangle$ has property $\alpha(p)$ iff $M_p(\phi) \subseteq D(F,q)$, and has property $\beta(p)$ iff for every r-tuple $\langle t_1,..., t_r\rangle$ of terms in $M_p(\phi)$, if some $t \in D(F,q)$ is unfavorable for $\langle t_1,..., t_r\rangle$ in $D(F,q)$ then $\phi(t_1,..., t_r)$ is such a t.

PROPOSITION 1. *If there exists a* ϕ-*domain* $\langle A, D(F,q), \phi\rangle$ *having properties* $\alpha(2p)$ *and* $\beta(2p)$, *then E(G,p) is satisfiable.*

PROOF. Let us say that a function sign f "arises from P" if it is the function sign associated with a restricted variable bound by a quantifier in P. If f arises from P then it has one more argument in F^* than it has in G^*, namely, the variable x. In fact x will be the $r+1^{st}$ argument, immediately after $x_1,..., x_r$. To find a truth-assignment verifying $E(G,p)$, we evaluate each such function sign f of G^* as the corresponding function sign of F^*, but with the "extra" $r+1^{st}$ argument place filled by an appropriate value of ϕ. This is the effect of the following mapping $\gamma: D(G,\omega) \rightarrow D(F,\omega)$:

$\gamma f(t_1...t_k) = f(\gamma t_1...\gamma t_k)$ for all function signs f save those arising from P
$\gamma f(t_1...t_r\, u_1...u_k) = f(\gamma t_1...\gamma t_r\ \phi(\gamma t_1,..., \gamma t_r)\ \gamma u_1...\gamma u_k)$ if f arises from P

By induction, $\gamma[D(G,k)] \subseteq M_{2k}(\phi)$. Since ϕ has property $\alpha(2p)$, $\gamma[D(G,p)] \subseteq D(F,q)$.

Let $A \circ \gamma$ be the truth-assignment that is the composition of A and γ, in the following sense: for every atomic formula $Rt_1...t_k$ that occurs in $E(G,p)$, $A \circ \gamma$ assigns the same value to $Rt_1...t_k$ that A assigns to $R\gamma t_1...\gamma t_k$. We must show that $A \circ \gamma$ makes $E(G,p)$ true, that is, that $A \circ \gamma$ makes every instance of G^* over $D(G,p)$ true. Let H be such an instance, and let γH be the result of replacing each atomic subformula $Rt_1...t_k$ of H by $R\gamma t_1...\gamma t_k$. Thus $A \circ \gamma$ gives to H the same truth-value that A gives to γH. Now γH will differ from an instance J of F^* over $D(F,q)$ only in containing a positive subformula

C: $\Phi^*(\gamma t_1,..., \gamma t_r, \gamma t, \gamma s_1,..., \gamma s_m) \vee P^*(\gamma t_1,..., \gamma t_r, \xi, \gamma u_1,..., \gamma u_n)$,

with $\xi = \phi(\gamma t_1,..., \gamma t_r)$, where J contains the subformula

D: $\Phi^*(\gamma t_1,..., \gamma t_r, \gamma t, \gamma s_1,..., \gamma s_m) \vee P^*(\gamma t_1,..., \gamma t_r, \gamma u_1,..., \gamma u_n)$.

That is, $\gamma H = J[C]$. Since A makes $E(F,q)$ true, A makes J true. Hence A could make γH false only if it made C false and $J[\perp]$ false but made D true. But if C is made false so is its first disjunct, so that γt is unfavorable for $\langle \gamma t_1,..., \gamma t_r \rangle$ in $D(F,q)$. Since the ϕ-domain has property $\beta(2p)$, ξ is unfavorable for $\langle \gamma t_1,..., \gamma t_r \rangle$ in $D(F,q)$. By Lemma 1, the second disjunct of C must be made true, whence C is made true. Hence A makes γH true. ■

REMARK. An extension of the above argument, unnoticed by Gödel, avoids the appeal to Lemma 1, and consequently allows us to simplify the definition of "favorable" by omitting clause (2). As before, if A makes J true but γH false, then ξ is unfavorable for $\langle \gamma t_1,..., \gamma t_r \rangle$. By definition, there exist r_1 ,..., r_m in $D(F,q)$ such that A falsifies $\Phi^*(\gamma t_1, ..., \gamma t_r, \xi, r_1, ..., r_m)$. Let E be $\Phi^*(\gamma t_1, ..., \gamma t_r, \xi, r_1, ..., r_m) \vee P^*(\gamma t_1,..., \gamma t_r, \xi, \gamma u_1,..., \gamma u_n)$ and let $K = J[E]$. Note that A gives E and C the same truth-value, since they have the second disjunct and their first disjuncts are both false. Hence $J[E]$ has the same truth-value as γH. But $J[E]$ is an instance of F^* over $D(F,q)$; hence A makes it true.

PROPOSITION 2. *For any ϕ and p, $|Mp(\phi)| < 2^{L^p}$.*

PROOF. Let $\mu(i) = |M_i(\phi)|$ and let $\nu(i) = 1 + \mu(i)$. Suppose F^* contains j function signs with maximum number k of argument places. Note that the $M_i(\phi)$ are generated by $j + 1$ functions: application of each of the j function signs, and ϕ. Thus $\mu(i+1) \leq \mu(i) + (j + 1) \cdot \mu(i)^k \leq \mu(i) + [1 + \mu(i)]^{j+1+k}$. Hence $\nu(i + 1) \leq \nu(i) + \nu(i)^{j+1+k} \leq \nu(i)^{j+k+2} = \nu(i)^N$. Since $\mu(0) = 2$, by induction we have $\nu(p) \leq 2^{L^p}$, so that $\mu(p) < 2^{L^p}$. ■

PROPOSITION 3. *If $E(F,N)$ is satisfiable, then there exists a ϕ-domain that has properties $\alpha(2p)$ and $\beta(2p)$.*

PROOF. Let A be any truth-assignment verifying $E(F,N)$. Define a sequence $\phi_1, \phi_2, \ldots$ of functions and a sequence $q_1, q_2, \ldots$ of integers as follows:

(1) ϕ_1 takes every r-tuple from $D(F,\omega)$ to 0, and $q_1 = 2p$.

(2) Suppose ϕ_i and q_i are defined for even i. Let $\phi_{i+1} = \phi_1$ and $q_{i+1} = q_i + 2p$.

(3) Suppose ϕ_i and q_i are defined for odd i. Let $q_{i+1} = q_i$. If $\langle A, D(F,q_i), \phi_i\rangle$ has property $\beta(2p)$, then let $\phi_{i+1} = \phi_i$. If not, let $\langle t_1,\ldots, t_r\rangle$, be the earliest r-tuple from $M_{2p}(\phi_{i+1})$ such that there exists a t in $D(F,q_i)$ that is unfavorable for $\langle t_1,\ldots, t_r\rangle$ in $D(F,q_i)$, but $\phi_i(t_1,\ldots, t_r)$ is not such a t. Pick the earliest such t; let $\phi_{i+1}(t_1,\ldots, t_r) = t$, and let ϕ_{i+1} agree with ϕ_i on all other r-tuples.

An important consequence of the definition of "favorable" is this: if t is unfavorable for $\langle t_1,\ldots, t_r\rangle$ in $D(F,q)$ and if $q' > q$, then t is unfavorable for $\langle t_1,\ldots, t_r\rangle$ in $D(F,q')$. Since the q_i are nondecreasing, it follows from the specification of the construction that if $\phi_i(t_1,\ldots, t_r) \neq 0$ then $\phi_i(t_1,\ldots, t_r)$ is unfavorable for $\langle t_1,\ldots, t_r\rangle$ in $D(F,q)$ and $\phi_i(t_1,\ldots, t_r) = \phi_j(t_1,\ldots, t_r)$ for each $j > i$. If $i \leq N/2p$ then $q_i \leq N$, so that A makes $E(F,q_i)$ true.[2] Hence, for such i, $\langle A, D(F,q_i), \phi_i\rangle$ is a ϕ-domain.

LEMMA 2. *For each i, $M_{2p}(\phi) \subseteq M_{2p}(\phi_{i+1})$.*

PROOF. If i is even, then $M_{2p}(\phi) = M_{2p}(\phi_{i+1})$ trivially, since $\phi_i = \phi_{i+1}$. Suppose i is odd. As has just been noted, $\phi_i(t_1,\ldots, t_r) \neq \phi_{i+1}(t_1,\ldots, t_r)$ then $\phi_i(t_1,\ldots, t_r) \neq 0$, and so $\phi_i(t_1,\ldots, t_r) \in M_0(\phi_{i+1})$. It then follows easily by induction on k that $M_k(\phi) \subseteq M_k(\phi_{i+1})$.

LEMMA 3. *For each i, if $\phi_i(t_1,\ldots, t_r) \neq 0$ then $t_1,\ldots, t_r$ are in $M_{2p}(\phi_i)$.*

PROOF. Let j be the least number such that $\phi_j(t_1,\ldots, t_r) \neq 0$. By the construction, $t_1,\ldots, t_r$ are in $M_{2p}(\phi_j)$. The result then follows by Lemma 2.

---

[2] Gödel's use of $4p$ here is not needed; for if $i < 2\cdot 2^{r\cdot L^{2p}}$ then $q_i \leq 2p\cdot 2^{r\cdot L^{2p}}$, and Lemmas 3 and 4 below show that the desired ϕ_i will have i strictly less than $2p\cdot 2^{r\cdot L^{2p}}$. Hence the Corrected Lemma could use the bound $2p\cdot 2^{r\cdot L^{2p}}$.

LEMMA 4. *There exists an odd integer i, $i \leq N/2p$, such that* $\langle A,D(F,q_i), \phi_i\rangle$ *has property* $\beta(2p)$.

PROOF. If no such i existed, then, for $k = N/2p$ there would be $k/2$ r-tuples $\langle t_1,..., t_r\rangle$ such that $\phi_k(t_1,..., t_r) \neq 0$. But by Proposition 2 and Lemma 3 there must be fewer than $N/4p$ such r-tuples.

LEMMA 5. *If i is odd then* $\langle A,D(F,q_i), \phi_i\rangle$ *has property* $\alpha(2p)$.

PROOF. By induction on k we show: if $t \in M(\phi_i)$ then $t \in D(F,q_{i+1}+k)$. This is obvious for $k = 0$. Suppose it is true for k. If $t = f(t_1...t_j)$ for f a function sign of F^* and $t_1...t_j$ in $M_k(\phi_i)$, by the induction hypothesis $t_1,...,t_j \in D(F,q_{i-1}+k)$, so that $t \in D(F,q_{i-1}+k+1)$. If $t = \phi(t_1,..., t_r)$ for t_i in $M_k(\phi_i)$, then, since $\phi_i = \phi_{i-1}$, $t \in D((F,q_{i-1})$. We conclude that $M_{2p}(\phi_i) \subseteq (D(F,q_{i-1}+2p) = D(F,q_i)$.

Obviously, Proposition 3 follows at once from Lemmas 4 and 5. ■

The foregoing argument for the correction of Herbrand's Lemma turns out to be, in all essentials, the same as that of [Dreben and Denton *1966*]. Gödel's ϕ-domain corresponds closely to the A-admissable function in the latter. The two conditions in Dreben and Denton's definition of "A-resolvent" are just Gödel's properties $\alpha(p)$ and $\beta(p)$. Consequently, Proposition 1 above amounts to the same as Dreben and Denton's Lemma I. Dreben and Denton use the simplified form of (what corresponds to) the notion of "favorable", i.e., the form mentioned in the Remark above. Dreben and Denton's inductive construction of an appropriate function ϕ_i is the same as Gödel's; however, they have a somewhat different way of bounding the number of steps needed to obtain a function with the desired properties. As a result, in their corrected Lemma the order of the satisfiable expansion of F required to assure the satisfiability of the expansion $E(G,p)$, is slightly larger.

Gödel gives that order as $4p\cdot 2^{r\cdot L^{2p}}$. As noted in footnote 2, his argument in fact yields the bound of $2p\cdot 2^{r\cdot L^{2p}}$. This bound can be further improved, by a more careful calculation of the cardinality of the set $M_{2p}(\phi)$. If the functional form F^* contains n function signs whose maximal degree is m, then this calculation yields $|M\,2p(\phi)| < (n+1)^{(m+1)^{2p}}$. Hence the order of the expansion of F can be taken to be $2p\cdot(n+1)^{r\cdot(m+1)^{2p}}$.

§3. Counterexamples and numerical bounds. Herbrand's original assertion was that if G is like F but for containing a subformula $(\forall x)\Phi(x) \vee P$ where F has $(\forall x)\Phi(x) \vee P$), then an expansion of G will be satisfiable provided that the expansion of F of the same order is satisfiable. The corrected Lemma differs dramatically in quantitative terms, for it requires the satisfiability of a much larger expansion of F. Indeed, the order of the

expansion of P grows doubly exponentially in the order of the expansion of G. Moreover, the order of the expansion of F depends also on certain syntactic features of F, namely, the number and degree of function signs that occur in the functional form F^*. Those parameters reflect the number and type of quantifiers in F.

As far as can be told from the material in Gödel's notebooks, however, Gödel never arrived at a counterexample to Herbrand's original assertion.. That is, although he saw that Herbrand's argument for his Lemma was flawed, he never showed that Herbrand's Lemma is actually false. Without such a counterexample, the immense increase in order demanded by the corrected Lemma may well have looked gratuitous. Perhaps this contributed to Gödel's reticence about telling anyone of his correction.

In 1963 Stål Aanderaa found counterexamples to Herbrand's Lemma, which also demonstrated the necessity of taking syntactic features of the formula into account. In particular, Aanderaa constructed, for each p, formulas F and G related as above such that $E(F,p)$ is satisfiable but $E(G,3)$ is unsatisfiable.[3]

Of course the growth in order given in Gödel's Corrected Lemma (as in [Dreben and Denton *1966*]), even when it is refined as indicated at the end of the previous section, is much larger. I shall now give a counterexample to Herbrand's original assertion that provides a lower bound for the growth of order that is close to Gödel's upper bound. In order to show more clearly the basic strategy of the construction, I start with a simpler version that yields, for each p, a formula F such that $E(G,p+2)$ is unsatisfiable while $E(F,2^{2^p})$ is satisfiable. Each such F is a formula in pure quantification theory, that is, no function signs occur in it, and contains a fixed number of quantifiers. Thus the construction shows that there can be doubly exponential growth of order when the other parameters are fixed.

The formula contains a triadic predicate letter C, dyadic letters N and H, and monadic letters K_0, K_1, L_0, ..., L_p. Let S be the set of of words on the alphabet $\{0,1\}$ whose length is 2^i, $0 \le i \le p$; if w is a word let $|w|$ be the length of w and let val(w) be the numerical value of w when w is construed as a binary numeral. These predicate letters with the exception of H have an intended interpretation over the universe S: $Cuvw$ is to mean $|u| = |w|$ and $w = uv$ (i e., a string u concatenated with string v); Nvw is to mean $|v| = |w|$ and val(w) = val(v)+1; K_0w is to mean that w is a string of 0's and K_1w that w is a string of 1's; L_iw is to mean $|w| = 2^i$. Let G be the conjunction of the universal closures of (1) - (7):

$$(1)\quad (\exists x_0)(\exists x_1)(L_0x_0 \wedge L_0x_1 \wedge K_0x_0 \wedge K_1x_1 \wedge Nx_0x_1)$$

$$(2)\quad (\bigvee_{0 \le i \le p} (L_i y_1 \wedge L_i y_2)) \supset (\exists x) C y_1 y_2 x$$

---

[3] Aanderaa's counterexample was first published in [Dreben, Andrews, and Aanderaa *1963*]; it can also be found in [van Heijenoort *1967*, 571] and [Herbrand *1971*, 193].

(3) $\bigwedge_{0 \le i \le p} (L_i y_1 \wedge Cy_1 y_2 y_3 \supset L_{i+1} y_3) \wedge$

$\bigwedge_{i=0,1} (K_i y_1 \wedge K_i y_2 \wedge Cy_1 y_2 y_3 \supset K_i y_3)$

(4) $(Cy_1y_2y_3 \wedge Cy_1z_2z_3 \wedge Ny_2z_2 \supset Ny_3z_3) \wedge$
$(Cy_1y_2y_3 \wedge Cz_1z_2z_3 \wedge Ny_1z_1 \wedge K_1y_2 \wedge K_0z_2 \supset Ny_3z_3)$

(5) $K_0y_1 \supset (\exists z)Hy_1z$

(6) $K_1y_1 \wedge L_py_1 \supset \sim Hy_1y_2$

(7) $Ny_1y_2 \supset (\forall y)\sim Hy_1y \vee (\exists z')Hy_2z'$

Let F be like G except that in clause (7) the universal quantifier $(\forall y)$ governs the entire consequent.

We use c_0 and c_1 for the constants associated with the variables x_0 and x_1 in clause (1), f for the dyadic function associated with the variable x in clause (2), and g for the dyadic function associated with the variable z in clause (7). To show $E(G,p+1)$ is unsatisfiable, suppose A made $E(G,p+1)$ true. Let $S \subset \{0,1\}^*$ be the set of words of length 2^i for some i, $0 \le 1 \le p$. Define ϕ: $S \to D(G,\omega)$ by: $\phi 0 = c_0$, $\phi 1 = c_1$, if $|v| = |w| = 2^i$, $0 \le i \le p$, then $\phi(vw) = f(\phi v\ \phi w)$. Note that if $|w| = 2^i$ then ϕw has height $i+1$.

CLAIM 1. Let $v,w \in S$ have length 2^i, $0 \le i \le p$. Then
(a) $A \vDash L_i\phi v \wedge L_i\phi w$;
(b) if w is a string of 0's then $A \vDash K_0\phi w$, and if w is a string of 1's, $A \vDash K_1\phi w$;
(c) if val(w) = val(v)+1 then $A \vDash N\phi v\ \phi w$.

PROOF. By induction on i. For $i = 0$, ϕv and ϕw are either c_0 or c_1; (a), (b) and (c) follow by clause (1). Now suppose (a), (b) and (c) hold for words of length 2^{i-1}. If v and w are such words, then $A \vDash C\phi v\ \phi w\ \phi vw$, by clause (2) and the definition of ϕ. By clause (3), (a) and (b) then hold for words of length 2^i. Suppose that val(w) = val(v)+1. Let w_1, w_2, v_1, v_2 have length 2^{i-1}, $w = w_1w_2$, $v = v_1v_2$. Then either $w_1 = v_1$ and val(w_2) = val(v_2)+ 1, or else w_2 is a string of 0's, v_2 is a string of 1's, and val(w_1) = val(v_1)+1. Note that $A \vDash C\phi w_1\ \phi w_2\ \phi w$ and $A \vDash C\phi v_1\ \phi v_2\ \phi v$. Then $A \vDash N\phi v\ \phi w$ follows from the induction hypothesis, (a), and clause (4). ■

CLAIM 2. Suppose $|w| = 2^p$. Then there is a term s of height $p+2$ such that $A \vDash H\phi w\ s$.

PROOF. By induction on val(w). If val(w) = 0, then w is a string of 0's. By Claim 1, $A \models K_0\phi w$. By clause (5), there is a term s of height one greater than ϕw, i.e., of height p+2, such that $A \models H\phi w\ s$. Now suppose val(w)>0, and the claim holds for the string v of length 2^P with val(v) = val(w)-1. By Claim 1, $A \models N\phi v\ \phi w$. By the induction hypothesis, for some t of height p+2, $A \models H\phi w\ t$. Hence by clause (7), $A \models H\phi w\ g(\phi v\ \phi w)$. Thus the term $g(\phi v\ \phi w)$ is the desired s. ■

Now let w be the string of 1's of length 2^P. By Claim 1, $A \models K_1\phi w$. By Claim 2, $A \models H\phi w\ s$ for some term s of height p-2. But then an instance over $D(G,p+2)$ of clause (6) must be made false by A. We may conclude that no truth-assignment makes $E(G,p+2)$ true.

To show $E(F,2^{2P})$ is satisfiable, let $\gamma: D(E,\omega) \rightarrow \{0,1\}^*$ as follows: $\gamma c_0 = 0$; $\gamma c_1 = 1$; $\gamma f(s\ t) = \gamma s \gamma t$; for any other s,γs is the empty string. Note that if the height of a term s is q then $|\gamma s| < 2^{q-1}$. Define a truth-assignment A as follows: for s, t, u in $D(F,\omega)$, $A \models Cstu$ iff $|\gamma s| = |\gamma t|$ and $\gamma u = \gamma s \gamma t$; $A \models L_i s$ iff $|\gamma s| = 2^i$; $A \models K_0 s$ [$A \models K_1 s$] iff γs is a nonempty string of 0's [1's]; $A \models Nst$ iff $|\gamma s| = |\gamma t|$ and val(γt) = val(γs)+1; $A \models Hsu$ iff the height(u) > height(s) + val(γs).

It is a routine matter to check that A makes $E(F,2^{2P})$ true. The crucial point is that in the functional form F^* clause (7) becomes $Ny_1y_2 \supset Hy_1y \vee Hy_2\ g(y_1y_2y)$, where g is the 3-place function sign associated with the variable z'. Thus, if $A \models Nst$ and $A \models Hsu$ for no term u of height $\leq j$, clause (7) can be true even though $A \models Htu'$ for no u' of height $\leq j$+1. This makes it possible to assign truth-values to atomic formulas Hsu as in the previous paragraph. That assignment, in turn, allows all instances of clause (6) over $D(F,2^{2P})$ to be made true, for if $A \models P_1 s \wedge L_p s$ then γs is the string of 1's of length 2^P. Hence ht(s) = p+1 and val(γs) = 2^{2P}, so that for no term u in $D(F,2^{2P})$ do we have $A \models Hsu$.

Now, as was pointed out at the end of §2, Gödel's argument establishes that the satisfiability of $E(F,2p\cdot(n+1)^{r\cdot(m+1)^{2p}})$ is sufficient for that of $E(G,p)$, where n is the number of function signs of F^*, m is the largest degree of these function signs, and r is the number of general variables that govern the subformula $(\forall x)\Phi(x) \vee P$). If F is a formula of pure quantification theory, that is, contains no function signs, then n is just the number of restricted variables of F. By complicating the example given above we can construct, for any p, n, r, and m, a formula F of pure quantification theory such that F contains n restricted variables, the maximum degree of any function sign in F^* is m, F contains a subformula $(\forall x)\Phi(x) \vee P$) governed by r general variables, and, if G is the formula obtained from F by replacing this subformula with $(\exists x)\Phi(x) \vee P$, then $E(G,p+2)$ is unsatisfiable while $E(F,(n-3)^{m^P})$ is satisfiable.

There are two modifications in the construction. To obtain n-3 as the base of exponentiation, rather than 2, we shall consider words in a (n-3)-letter alphabet instead of in

{0,1}. This requires replacing clause (1) by a clause with n-3 existential variables. To obtain m^P in the exponent, rather than 2^P, we replace triadic C with a $(m+1)$-adic letter, to mean concatenation of m strings. By use of this new letter, terms of height $p+1$ will be able to represent words of length m^P. This will complicate the conditions on the predicate N, expressed in clause (4).

Let $k = n$-4, let $\underline{0}$, $\underline{1}$, ..., $\underline{k}$ be digits in a base n-3 numeral-system; let S be the set of words on $\{\underline{0}, \underline{1}, ..., \underline{k}\}$ of length m^i for some $i \geq 0$. For $w \in S$, let $|w|$ be the length of w and val(w) be the numerical value of w when w is construed as a numeral in base n-3. Let G be the conjunction of the universal closures of (1) – (8):

(1) $(\exists x_0)...(\exists x_k)L_0x_0 \wedge ... \wedge L_0x_k \wedge K_0x_0 \wedge K_1x_k \wedge Nx_0x_1 \wedge ... \wedge Nx_{k-1}x_k)$

(2) $\left(\bigvee_{0 \leq i \leq p} (L_i y_1 \wedge ... \wedge L_i y_m) \right) \supset (\exists x) C y_1 y_2 ... y_m x$

(3) $\bigwedge_{0 \leq i \leq p} (L_i y_j \wedge C y_1 ... y_m y_{m+1} \supset L_{i+1} y_{m+1})$

$\wedge \bigwedge_{i = 0, 1} (K_i y_j \wedge K_i y_2 \wedge ... \wedge C y_1 ... y_m y_{m+1} \supset K_i y_{m+1})$

(4) $\bigwedge_{0 \leq i \leq m} [(C y_1 ... y_m y_{m+1} \wedge C y_1 ... y_{i-1} z_i ... z_m z_{m+1} \wedge N y_i z_i$

$\wedge \bigwedge_{i < j \leq m} (P_1 y_j \wedge P_0 z_j)) \supset N y_{m+1} z_{m+1})$

(5) $K_0y \supset (\exists z) Hyz$

(6) $K_1y_1 \wedge L_py_1 \supset \sim Hy_1y_2$

(7) $Ny_1y_2 \supset (\forall y) \sim Hy_1y \vee (\exists z') Hy_2z'$

As before, let F be like G except that in clause (7) the universal quantifier $(\forall y)$ governs the entire consequent.

We use c_0,..., c_k for the constants associated with the existential variables of clause (1), and f for the k-adic function sign associated with the variable x of clause (2).

To show $E(G, p+2)$ is unsatisfiable, suppose A made $E(G, p+2)$ true. Let $S \subset \{\underline{0},..., \underline{k}\}^*$ be the set of words of length m^i for some i, $0 \leq i \leq p$. Define ϕ: $S \rightarrow D(G,\omega)$ by $\phi\underline{i} = c_i$, 1

$\leq i \leq k$; if $w_1, ..., w_m \in S$ all have length m^i, $i \geq 0$, then $\phi(w_1...w_m) = f(\phi w_1...\phi w_m)$. Note that if $|w| = m^i$ then height$(\phi w) = i+1$.

CLAIM 1. Let v and $w \in S$ have length m^i, $0 \leq i \leq p$. Then
(a) $A \models L_i\phi v \wedge L_i\phi w$;
(b) if w is a string of $\underline{0}$s then $A \models K_0\phi w$, and if w is a string of $\underline{k}$'s, $A \models K_1\phi w$.
(c) if val(w) = val$(v)+1$ then $A \models N\phi v\, \phi w$.

PROOF. By induction on i. (a) is proved as in the earlier example. For (b), we let $w = w_1...w_k$, and $v = v_1...v_k$, where each w_j and v_j has length m^{i-1}. If val(w) = val$(v)+1$ then for some j, $1 \leq j \leq m$, val(w_j) = val$(v_j)+1$ and for all r, $j < r \leq m$, v_r is a string of $\underline{k}$'s while w_j is a string of $\underline{0}$'s. Note that $A \models C\phi w_1...\phi w_m\phi w_{m+1} \wedge C\phi v_1...\phi v_m\phi v_{m+1}$. Hence $A \models N\phi w\phi w$ follows from the induction hypothesis, (a), and clause (4). ■

CLAIM 2. Suppose $|w| = 2^P$. Then there is a term s of height $p+2$ such that $A \models H\phi w\, s$.

PROOF. As in the previous example. ■

Also as in the previous example, from claims 1 and 2 it follows that some instance of clause (6) over $D(F, p+2)$ must be made false by A. Hence no truth-assignment makes E(G,p+2) true.

To show that $E(F, (n-3)^{mP})$ is satisfiable, let $\gamma: D(F,\omega) \rightarrow \{\underline{0}, ..., \underline{k}\}^*$ as follows: $\gamma c_i = \underline{i}$ for $0 \leq i \leq k$; $\gamma f(t_1...t_k) = \gamma t_1...\gamma t_k$; for any other term t, γt is the empty string. Note that if the height of a term s is q then $|\gamma s| < m^{q-1}$. Define a truth-assignment A as follows: $A \models Cs_1...s_kt$ iff $|\gamma s_1| = ... = |\gamma s_k|$ and $\gamma t = \gamma s_1...\gamma s_k$; $A \models L_is$ iff $|\gamma s| = m^i$; $A \models K_0s$ [$A \models K_1s$] iff γs is a nonempty string of $\underline{0}$'s [$\underline{k}$'s]; $A \models Nst$ iff $|\gamma s| = |\gamma t|$ and val(γt) = val$(\gamma s)+1$; $A \models Hsu$ iff the height(u) > height(s) + val(γs).

It is a routine matter to check that A makes $E(F, (n-3)^{mP})$ true.

REFERENCES

DREBEN, Burton. 1963. *Corrections to Herbrand*, American Mathematical Society, Notices **10**, 285.

DREBEN, Burton, Peter ANDREWS and Stål AANDERAA. 1963. *False lemmas in Herbrand*, Bulletin of the American Mathematical Society **69**, pp. 699-706.

DREBEN, Burton and John DENTON. 1966. *A supplement to Herbrand*, Journal of Symbolic Logic **31**, 393-398.

HERBRAND, Jacques. 1930. *Recherches sur la théorie de la démonstration*, doctoral dissertation at the University of Paris; also Prace Towarzystwa Naukowego Warszawskiego, Wydział III, no. 33. Also in [Herbrand *1968*, 35-153]; English translation in [Herbrand *1971*, 44-202].

—. 1931. *Sur le problème fondamental de la logique mathématique*, Sprawozdania z posiedzen Towarzystwa Naukowego Warszawskiego, Wydział III, no. 24, 12-56. Also in [Herbrand *1968*, 167-207]; English translation in [Herbrand *1971*, pp. 215-271].

—. 1968. *Écrits logiques*, Paris, Presses Universitaires de France.

—. 1971. *Logical writings* (W.D. Goldfarb, editor), Dordrecht, Reidel and Cambridge, Mass., Harvard University Press.

HILBERT, David and Paul BERNAYS. 1939. *Grundlagen der Mathematik*, vol. 2, Berlin, Springer.

VAN HEIJENOORT, Jean. 1967. (editor) *From Frege to Gödel: A source book in mathematical logic*, Cambridge, Mass., Harvard University Press.

—. 1968. Préface, in [Herbrand *1968*, 1-12].

—. 1985. *Selected essays*, Naples, Bibliopolis.

—. 1985a. *Jacques Herbrand's work in logic and its historical context*, in [van Heijenoort 1985, 99-121].