3.5 Herbrand's theorem (see e.g. U. Schöning: Logic for computer Science)

Definition 3.5.1: Let φ be a sentence in $\mathcal{L}$.
The "Herbrand universe $D(\varphi)$" of φ is the set of all closed terms which can be built up out of the symbols occurring in φ (plus some distinguished constant symbol c in case φ does not contain any constant symbol):

(i) every constant in φ is in $D(\varphi)$ ($c \in D(\varphi)$ if no constant in φ).

(ii) if $t_1, \ldots, t_n \in D(\varphi)$ and the n-ary function symbol f occurs in φ, then $f(t_1, \ldots, t_n) \in D(\varphi)$.

Example: $\varphi = \forall x\, P(x, f(x))$, P binary predicate symbol. Then $D(\varphi) = \{c, f(c), f(f(c)), \ldots\}$.

Definition 3.5.2: Let φ be a sentence in $\mathcal{L}$.
A structure $\mathcal{A} = \langle A, \ldots \rangle$ is called a "Herbrand structure" for $\mathcal{L}_\varphi$ if

(i) $A = D(\varphi)$

(ii) for all n-ary function symbols f in φ and all terms $t_1, \ldots, t_n \in D(\varphi)$

$f^{\mathcal{A}}(t_1, \ldots, t_n) = f(t_1, \ldots, t_n)$.

For all constant symbols c in φ: $c^{\mathcal{A}} = c$.

A Herbrand structure for φ which is a model of φ is called a "Herbrand model" of φ. Analogous for sets Γ of sentences.

Remark: (i) Herbrand structures $\mathcal{A}$ for sentences φ interpret closed terms by themselves: $t^{\mathcal{A}} = t$. That is why Herbrand models are also called "term models".
(ii) There is no special requirement for the interpretation of predicate symbols P in Herbrand structures.
(iii) We consider Herbrand structures only in connection with logic without equality.

Proposition 3.5.3 (Existence of Herbrand models): Let φ be a sentence (without $=$) that is purely universal, i.e. $\varphi = \forall \underline{x}\, \varphi_{qf}(\underline{x})$, where φ_{qf} is quantifier-free and $\underline{x} = x_1, \ldots, x_k$.
φ has a model iff φ has a Herbrand model.

Proof: "$\Leftarrow$" is trivial.
"$\Rightarrow$": Let $\mathcal{A} = \langle A, \ldots \rangle$ be an arbitrary model of φ. The interpretation of the function and constant symbols in the Herbrand model $\mathcal{A}^H = \langle D(\varphi), \ldots \rangle$ we are going to construct is fixed by the definition of Herbrand structures. Hence it suffices to interpret the predicate symbols P in φ:

$$(t_1, \ldots, t_n) \in P^{\mathcal{A}^H} :\Leftrightarrow (t_1^{\mathcal{A}}, \ldots, t_n^{\mathcal{A}}) \in P^{\mathcal{A}},$$

where $t_1, \ldots, t_n \in D(\varphi)$.
One verifies that $\mathcal{A}^H \models \forall \underline{x}\, \varphi_{qf}(\underline{x})$ by induction on the length k of $\underline{x} = x_1, \ldots, x_k$. $\square$

Proposition 3.5.3 implies the corresponding dual statement for purely existential sentences:

<u>Proposition 3.5.4</u>: Let $\varphi = \exists \underline{x}\, \varphi_{qf}(\underline{x})$ be a purely existential sentence (without $=$). Then

$\models \varphi$ iff (for all Herbrand structures $\mathcal{A}^H$ for φ: $\mathcal{A}^H \models \varphi$).

<u>Proof</u>: Obvious using that the negation of φ is (logically equivalent) to a purely universal sentence so that proposition 3.5.3 applies. $\square$

<u>Definition 3.5.5</u> (Herbrand expansion):

Let $\varphi = \forall x_1, \dots, x_n\, \varphi_{qf}(\underline{x})$ be a purely universal sentence. Then the "Herbrand expansion" $E(\varphi)$ of φ is defined as $E(\varphi) := \{\varphi_{qf}(t_1, \dots, t_n) : t_1, \dots, t_n \in D(\varphi)\}$.

<u>Proposition 3.5.6</u>: Let φ be a purely universal sentence (without $=$). Then

φ has a model iff $E(\varphi)$ is satisfiable in the sense of propositional logic.

<u>Proof</u>: By the previous results it suffices to show that φ has a Herbrand model iff $E(\varphi)$ is satisfiable in the sense of propositional logic: let $\mathcal{A}^H$ be a Herbrand model of φ. Then $\nu(R(t_1, \dots, t_k)) = \begin{cases} 1, & \text{if } \mathcal{A}^H \models R(\underline{t}) \\ 0, & \text{otherwise} \end{cases}$

is a satisfying assignment for $E(\varphi)$.

Conversely, if such an assignment ν with $\nu \models E(\varphi)$ is given, then (for n-ary predicate symbols P in φ)

$$P^{\mathcal{A}^H} := \{(t_1, \dots, t_n) \in D(\varphi)^n : \nu(P(t_1, \dots, t_n)) = 1\}$$

defined a Herbrand model $\mathcal{A}^H$ of φ. $\square$

<u>Theorem 3.5.7</u> (Herbrand's Theorem, J. Herbrand 1930)
Let $\varphi = \exists \underline{x}\ \varphi_{qf}(\underline{x})$ be a purely existential sentence (without $=$). Let $\vdash_{-=}$ denote derivability in ND without the $=$-rules.
Then the following holds:

$$\vdash_{-=} \varphi \text{ iff } \exists\, t_{1,1}, \dots, t_{1,m}, \dots t_{k,1} \dots t_{k,m} \in D(\varphi): \bigvee_{i=1}^{m} \varphi_{qf}(t_{1,i}, \dots, t_{k,i}) \in TAUT.$$

<u>Proof:</u>

"$\Leftarrow$" If $\bigvee_{i=1}^{m} \varphi_{qf}(t_{1,i}, \dots, t_{k,i}) \in TAUT$, then $\vdash_{-=} \bigvee_{i=1}^{m} \varphi_{qf}(t_{1,i}, \dots, t_{k,i})$ by the completeness theorem for propositional logic.

$\vdash_{-=} \varphi_{qf}(t_{1,i}, \dots, t_{k,i}) \to \exists x_1, \dots x_k\ \varphi_{qf}(x_1, \dots, x_k)$ by $\exists$-introduction. Hence

$$\vdash_{-=} \bigvee_{i=1}^{m} \varphi_{qf}(t_{1,i}, \dots, t_{k,i}) \to \underbrace{\varphi \vee \dots \vee \varphi}_{m\text{-times}}$$ and

so by contraction $\vdash_{-=} \varphi$.

"$\Rightarrow$" We give an ineffective model-theoretic proof for the contraposite formulation: if $\bigvee_{i=1}^{n} q_{qf}(\underline{t}_i) \notin TAUT$ for all $\underline{t}_1, \dots, \underline{t}_n \in D(\varphi)^k$, then

$\{\neg \varphi_{qf}(t_1, \dots, t_k) : t_1, \dots, t_k \in D_f(\varphi)\}$ is satisfiable (in the sense of propositional logic) for every finite subset $D_f(\varphi)$ of $D(\varphi)$. By propositional compactness this implies that $E(\forall x_1, \dots x_k \neg \varphi_{qf})$ is satisfiable (in the sense of prop. logic). Hence by proposition 3.56 $\neg\varphi$ has a model $\mathcal{A}$. So $\not\models \varphi$. The completeness theorem now yields that $\nvdash \varphi$. $\square$

Remark: Herbrand's very complicated proof was purely proof-theoretic (with some corrections due to Gödel 1944 resp. Aanderaa, Andrews, Dreben 1966) and provides an algorithm for the extraction of a tautological Herbrand disjunction from a given proof of φ in $ND_{=}$. Subsequently, other syntactic proofs where given by the Hilbert ε-substitution method (D. Hilbert / P. Bernays: Grundlagen der Mathematik II, Springer 1938) and G. Gentzen as a consequence of his cut-elimination theorems (G. Gentzen 1936). See e.g. "J. Shoenfield: Mathematical Logic" (1967) or "S. Buss (ed.): Handbook of Proof Theory" (1998) for syntactic proofs.

Using the Herbrand normal form φ^H of a prenex sentence φ, Herbrand's Theorem immediately extends to arbitrary sentences in the following form:

Herbrand's Theorem (general form) 3.5.8:

Let φ be a sentence in prenex normal form without equality and $\varphi^H = \exists \underline{x}\, \varphi^H_{qf}(\underline{x})$ its Herbrand normal form. Then

$\vdash_{=} \varphi$ iff $\exists\, \underline{t}_1, \ldots, \underline{t}_n \in D(\varphi^H)^k$ $\bigvee_{i=1}^{n} \varphi^H_{qf}(\underline{t}_i) \in TAUT.$

(note that now the Herbrand terms $\underline{t}_i$ are built up also by using the Herbrand index functions $\underline{f}$ used to form φ^H from φ).

Proof: "$\Rightarrow$": $\vdash_{=} \varphi$ implies trivially that $\vdash_{=} \varphi^H$. Now apply theorem 3.5.7 to the purely existential sentence φ^H.

"$\Leftarrow$" By "$\Leftarrow$" in thm. 3.5.7 $\bigvee_{i=1}^{n} \varphi^H_{qf}(\underline{t}_i) \in TAUT$ implies that $\vdash \varphi^H$ and so $\vdash \varphi$ (3.4.6.). To get from the tautological H-disjunction $\bigvee_{i=1}^{n}(\,)$ a proof of φ in $\vdash_{=}$ one needs a more involved argument which we only sketch in the example below:

Example: Clearly: $\vdash_{=} \exists x \forall y (P(x) \vee \neg P(y))$.

$(\exists x \forall y (P(x) \vee \neg P(y)))^H = \exists x\, (P(x) \vee \neg P(f(x)))$.

$(P(c) \vee \neg P(f(c))) \vee (P(f(c)) \vee \neg P(f(f(c)))) \in TAUT.$

So $t_1 := c$, $t_2 := f(c)$ provide a valid Herbrand disjunction.

There is also a version of the general form of Herbrand's theorem that is formulated without the use of Herbrand index functions:

Let φ^H be the Herbrand normal form of a prenex sentence φ (without equality) and $\varphi^{H.D} = \bigvee_{i=1}^{m} \varphi^H_{qf}(\underline{t}_i) \in TAUT$ a Herbrand disjunction:

Now replace in $\underline{t}_i$ all terms starting with a Herbrand function symbol (e.g. $f(s_1, \dots s_k)$) by a new variable starting from terms with largest size. Let the resulting f-free disjunction be denoted by φ^D. With $\varphi^{H.D}$ also φ^D is a tautology and there is a direct proof from φ^D to φ, i.e. a proof that essentially only uses quantifier-introductions and contractions. The general procedure is somewhat complicated to describe and so we just treat an example:

Consider again $\varphi :\equiv \exists x \forall y\, (P(x) \vee \neg P(y))$ and

$\varphi^H = \exists x\, (P(x) \vee \neg P(f(x)))$.

$\varphi^{H.D} := \left(P(c) \vee \neg P(f(c))\right) \vee \left(P(f(c)) \vee \neg P(f(f(c)))\right) \in TAUT.$

Now replace $f(f(c))$ by the variable z and $f(c)$ by y.

Then

$\varphi^D := (P(c) \vee \neg P(y)) \vee (P(y) \vee \neg P(z)) \in TAUT.$

In $ND_{=}$ one easily shows that the following reasoning can be carried out

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\varphi^D}{P(c)\vee\neg P(y)\quad\vee\forall z(P(y)\vee\neg P(z))}\ \text{"}\forall\text{-Intro"}}{P(c)\vee\neg P(y)\ \vee\exists y\forall z(P(y)\vee\neg P(z))}\ \text{"}\exists\text{-Intro"}}{\forall y(P(c)\vee\neg P(y))\vee\exists y\forall z(P(y)\vee\neg P(z))}\ \text{"}\forall\text{-Intro"}}{\exists x\forall y(P(x)\vee\neg P(y))\vee\exists y\forall z(P(y)\vee\neg P(z))}\ \text{"}\exists\text{-Intro"}}{\exists x\forall y(P(x)\vee\neg P(y))}$$

(since z does not occur in $P(c)\vee\neg P(y)$)

"contraction modulo renaming of variables"

<u>Herbrand's theorem for sentences with equality</u> 3.5.9.

Let φ be a prenex sentence and $\varphi^H = \exists \underline{x}\, \varphi^H_{qf}(\underline{x})$ its Herbrand normal form. Let $\forall \underline{u}\, E_{qf}(\underline{u})$ be the purely universal prenex normal form of the conjunction of the $=$axioms $I_1 - I_4$ for all the function and predicate symbols occurring in φ. Then $\vdash \varphi$ implies that $\vdash_{-=} \forall\underline{u}\, E_{qf}(\underline{u}) \to \varphi$ and also $\vdash_{-=} \forall\underline{u}\, E_{qf}(\underline{u}) \to \exists\underline{x}\,\varphi^H_{qf}(\underline{x})$ and so

$$\vdash_{-=} \underbrace{\exists \underline{x}, \underline{u}\ (E_{qf}(\underline{u}) \to \varphi^H_{qf}(\underline{x}))}_{\psi :=}.$$

Now apply theorem 3.5.7 to ψ to obtain closed terms $\underline{s}_1, \dots, \underline{s}_k,\ \underline{t}_1, \dots, \underline{t}_m \in D(\varphi^H)$ s.t.

$$\left(\bigwedge_{i=1}^{k} E_{qf}(\underline{s}_i) \rightarrow \bigvee_{i=1}^{m} \varphi^H_{qf}(\underline{t}_i)\right) \in TAUT.$$

So we no longer get a tautological Herbrand disjunction for $\exists \underline{x} \varphi^H_{qf}$ but a disjunction that is a tautological consequence of finitely many closed instances of =-axioms, i.e. a disjunction that is a so-called quasi-tautology.

The converse direction "$(\bigwedge \rightarrow \bigvee) \rightarrow \varphi$" follows as before and even holds if $\bigwedge_{i=1}^{k} E_{qf}(\underline{s}_i)$ contains instances of equality axioms

$$v_1 = v_1' \wedge \ldots \wedge v_\ell = v_\ell' \rightarrow f(v_1, \ldots, v_\ell) = f(v_1', \ldots, v_\ell')$$

for function symbols f used in forming φ^H from φ.

In that latter case, though, the aforementioned process of replacing f-terms by new variables no longer results in a valid quasi-tautology.

<u>Herbrand's theorem for open theories</u>: Let us recall that an open theory T is a theory axiomatized by purely universal sentences. Let φ, φ^H as above and $T \vdash \varphi$. As in the case of the purely universal =-axioms, one can shift all the purely universal axioms of T used in proving φ as an implicative assumption. Applying Herbrand's theorem yields a disjunction that is

a tautological consequence of finitely many closed instances of these axioms (and equality axioms). In particular:

Theorem 3.5.10: Let T be an open theory, φ a sentence in $\mathcal{L}(T)$ in prenex normal form and φ^H its Herbrand normal form. Then (for the extension $T[\underline{f}]$ of T by the H-functions)

$$T \vdash \varphi \text{ iff } \exists \underline{t}_1, \dots \underline{t}_n : T[\underline{f}] \vdash \bigvee_{i=1}^{n} \varphi^H_{qf}(\underline{t}_i),$$

where the closed terms $\underline{t}_i$ are built up out of φ^H-material and the constant and function symbols occurring in the nonlogical axioms of T.

Remark: Every theory T can be extended to its Skolem extension T^{Sk}, that is an open theory to which theorem 3.5.10 extends. Then, however, the Herbrand terms in general also involve the Skolem function symbols used in forming T^{Sk}!

Herbrand's Theorem does not apply directly to non-open theories such as PA. Consider PA augmented by a new unary function symbol f:

$$PA[f] \vdash \exists x \forall y \, (f(x) \le f(y)) =: \varphi$$

$\varphi^H = \exists x \, (f(x) \le f(g(x)))$. One only has a "H-disjunction" of variable length $f(0)$

$$\bigvee_{i=0}^{f(0)} f(t_i) \le f(g(t_i)), \text{ where } t_i := g^{(i)}(0).$$

But: no H-disjunction of fixed length n!

Corollary to the theorem 3.5.9: Let φ be a sentence without "=". If $\vdash \varphi$ then $\vdash_{-=} \varphi$.

Proof: W.l.o.g. we may assume that φ is in prenex normal form. By thm. 3.5.9 $\vdash \varphi$ and so $\vdash \varphi^H$ yields a quasi-tautology $\bigvee_{i=1}^{n} \varphi_{qf}^H(\underline{t}_i)$ which is a tautology since "=" does not occur and so we may ~~test~~ interpret any formula "$s = t$" in $\bigwedge_{i=1}^{k} E_\varphi(\underline{s}_i)$ as "true" if s and t are identical as terms and "false" otherwise. "$\Leftarrow$" from thm. 3.5.8 then yields $\vdash_{-=} \varphi$. $\square$

A lower bound for the length of Herbrand disjunctions (as well as the complexity of normalizing proofs or elimination of cuts):

<u>A growth example</u> (Statman, Orevkov, Zhang, Pudlák 79-91):

Consider the following open first order theory $\mathcal{T}$:

(i) Language $\mathcal{L}(\mathcal{T})$ of $\mathcal{T}$: $=, +, \cdot, 2^{(\cdot)}, I(\cdot), 0, 1$

(ii) Non-logical axioms:

$$x+(y+z) = (x+y)+z, \quad y+0 = y,$$

$$2^0 = 1, \quad 2^x + 2^x = 2^{1+x}$$

$$I(0), \quad I(x) \to I(1+x).$$

Intended meaning of "$I(x)$": "x is a (realizable) natural number".

Let $\forall \underline{x}\, \varphi_0(\underline{x})$ be the prenex nf of the conjunction of the universal closures of the non-logical axioms.

<u>Prop.</u>: $\mathcal{T} \vdash I(2^{\cdot^{\cdot^{2^0}}}\}k)$ for any fixed k, i.e.

$\vdash \forall \underline{x}\, \varphi_0(\underline{x}) \to I(2^{\cdot^{\cdot^{2^0}}})$, i.e.

$\vdash \exists \underline{x}\, (\varphi_0(\underline{x}) \to I(2^{\cdot^{\cdot^{2^0}}}))$.

Proof: We define inductively relations R_i:

$R_0 :\equiv I$. Let R_i be defined by I.H. then

$$R_{i+1}(x) :\equiv \forall y\,(R_i(y) \to R_i(2^x + y)).$$

By meta-induction on i we prove that

$$\mathcal{T} \vdash R_i(0) \wedge \forall x\,(R_i(x) \to R_i(1+x)).$$

For $i=0$ this is clear. Assume

$$(*) \quad R_i(0) \wedge (R_i(x) \to R_i(1+x)).$$

Using $2^0 = 1$ we get

$$R_{i+1}(0) \equiv \forall y\,(R_i(y) \to R_i(2^0+y)) \leftrightarrow \forall y\,(R_i(y) \to R_i(1+y)).$$

Also

$$R_{i+1}(x) \equiv \forall y\,(R_i(y) \to R_i(2^x+y))$$

$$\to \forall y\,\Big((R_i(y) \to R_i(2^x+y)) \wedge (R_i(2^x+y) \to R_i(\underbrace{2^x+(2^x+y)}_{=(2^x+2^x)+y\,=\,2^{1+x}+y}))\Big)$$

$$\to \forall y\,(R_i(y) \to R_i(2^{1+x}+y)) \equiv R_{i+1}(1+x).$$

Now $R_{i+1}(x) \overset{y:=0}{\to} (R_i(0) \to R_i(\underbrace{2^x+0}_{=2^x}))$, i.e.

$$(**) \quad R_{i+1}(x) \to R_i(2^x). \quad \text{So}$$

$$R_k(0) \xrightarrow{(**)} R_{k-1}(2^0)$$
$$\vdots$$
$$\xrightarrow{(**)} R_0(2^{\cdot^{\cdot^{2^0}}}\}k) \equiv I(2^{\cdot^{\cdot^{2^0}}}\}k). \square$$

<u>Corollary</u>: There exists a constant $c \in \mathbb{N}$ s.t. for any $k \in \mathbb{N}$ there is a ~~proof of~~ LK ND-proof $d_k \vdash_{ND} \exists \underline{x}(A_0(\underline{x}) \to I(2^{\cdot^{\cdot^{2^0}}}\}k))$ s.t. the depth of d_k is $\leq c \cdot k$.

<u>Theorem</u> (Zhang 92, Pudlák 92): Any valid H-disjunction for "$\exists \underline{x}(A_0(\underline{x}) \to I(2^{\cdot^{\cdot^{2^0}}}\}k))$" has length $\geq 2^{\cdot^{\cdot^{2}}}\}k$. The same applies for the length of any normal or cut-free proof of "$\exists \underline{x}(A_0(\underline{x}) \to I(2 \cdots))$".

<u>Prove</u> (of the first claim): We prove that the Herbrand disj. has to "contain" all instances $I(0) \to I(1)$, $I(1) \to I(1+1)$, ... $I(\bar{n}) \to I(\bar{n}+1)$ for all $n < 2^{\cdot^{\cdot^{2^0}}}\}k$. Suppose this not, i.e.

$$\vdash \bigvee_{\underline{t}} (A_0(\underline{t}) \to I(2^{\cdot^{\cdot^{2^0}}}\}k)) \text{ s.t.}$$

but for some $\bar{n} < 2^{\cdot^{\cdot^{2^0}}}\}k$ we don't have $I(\bar{n}) \to I(\bar{n}+1)$. Then interpret $I(x)$ as "$x \leq \bar{n}$". $\square$