The Bulletin of Symbolic Logic
Volume 6, Number 4, Dec. 2000

AN INTERPOLATION THEOREM

MARTIN OTTO

Abstract. Lyndon's Interpolation Theorem asserts that for any valid implication between two purely relational sentences of first-order logic, there is an interpolant in which each relation symbol appears positively (negatively) only if it appears positively (negatively) in both the antecedent and the succedent of the given implication. We prove a similar, more general interpolation result with the additional requirement that, for some fixed tuple $\mathbb{U}$ of unary predicates U, all formulae under consideration have all quantifiers explicitly relativised to one of the U. Under this stipulation, existential (universal) quantification over U contributes a positive (negative) occurrence of U.

It is shown how this single new interpolation theorem, obtained by a canonical and rather elementary model theoretic proof, unifies a number of related results: the classical characterisation theorems concerning extensions (substructures) with those concerning monotonicity, as well as a many-sorted interpolation theorem focusing on positive vs. negative occurrences of predicates *and* on existentially vs. universally quantified sorts.

Introduction. Given a valid implication $\varphi \models \psi$, an *interpolant* for that implication is an intermediate formula χ for which $\varphi \models \chi$ and $\chi \models \psi$. Looking for interpolants χ from some restricted syntactic class of formulae one can discern in how far the information transferred in the implication $\varphi \models \psi$ is expressible under those syntactic restrictions. Natural syntactic requirements on the interpolant centre on syntactic properties shared by φ and ψ. Interpolation properties, which guarantee the existence of certain interpolants, therefore measure syntax against semantics and may be regarded as criteria for how closely syntax reflects semantics. In a sense, the interpolant syntactically reflects a bottleneck between antecedent and succedent. In the model theoretic perspective, this phenomenon becomes most apparent in the way in which interpolation results often give rise to definability or expressibility results, especially in the context of model theoretic characterisation theorems.

Consider the most fundamental interpolation property for first-order logic. Craig's Interpolation Theorem [4] says that any valid first-order implication $\varphi \models \psi$ has an interpolant χ whose vocabulary is restricted to the

Received November 17, 2000; revised September 2000.
Key words and phrases. classical model theory, first-order logic, many-sorted structures, interpolation, preservation and characterisation theorems.

1079-8986/00/0604-0002/$2.60

common vocabulary in φ and ψ. The associated characterisation result, which follows immediately from Craig interpolation, is the following. For vocabularies $\tau_0 \subseteq \tau$, those τ-formulae whose truth in τ-structures is fully determined by the τ_0-reduct are precisely those equivalent to τ_0-formulae. The straightforward reduction of this expressibility claim to interpolation is typical of this type of application. From the given φ one passes to a variant φ' in which all symbols from $\tau \setminus \tau_0$ have new names. Then $\varphi \models \varphi'$ is a valid implication. A Craig interpolant for this implication is a τ_0-formula equivalent to φ. For another and better known consequence of Craig's Interpolation Theorem, Beth's Definability Theorem may actually be proved in a very similar vein, see for instance [9].

Many variations and generalisations of Craig's Interpolation Theorem have been found for other logics, but also for first-order itself. Among the most notable ones with interesting applications in classical first-order model theory are Lyndon's interpolation [10] and Feferman's many-sorted interpolation [5, 6].

Lyndon's Interpolation Theorem takes into account the polarities in which predicates occur, i.e., distinguishes between positive and negative occurrences. Predicates may occur positively (negatively) in the interpolant only if they occur positively (negatively) in both the antecedent φ and the succedent ψ. The corresponding characterisation result associates monotonicity with positivity.

Feferman's Interpolation Theorem concerns interpolation in a many-sorted framework, rather than the standard one-sorted structures. For this framework, however, it goes beyond Craig's condition in taking into account which sorts occur existentially (universally) quantified in the interpolant. This analysis has wide ranging model-theoretic applications, one of the most natural ones being the characterisation theorem which associates preservation under extensions with existential formulae.

The starting point for the present considerations is the observation that superficially these last two interpolation properties would seem to be related via the natural translation from many-sorted structures into one-sorted structures in which the different sorts get modelled as different sub-domains, each marked by a new unary predicate. In this translation an existential or universal quantification over sort U becomes a U-relativised quantification of the form $\exists x\ (Ux \wedge \dots)$ or $\forall x\ (Ux \rightarrow \dots)$, respectively. Thus existential quantification contributes a positive occurrence, universal quantification a negative occurrence of the corresponding sort predicate. Lyndon's Interpolation Theorem would seem to tell us something about that, and translating back one might hope to account for existentially and universally quantified sorts in an interpolant. But of course, a Lyndon interpolant obtained for the translation of a valid many-sorted implication need not itself be (equivalent to) a translation of a many-sorted formula. What is more, Feferman's

Interpolation Theorem shows a characteristic asymmetry with respect to antecedent and succedent concerning the restrictions on existentially (universally) quantified sorts, whereas Lyndon's Interpolation Theorem is quite symmetric.

We here propose an interpolation result in a framework of relativised first-order formulae, which does indeed form a common ground for Lyndon's interpolation and Feferman's many-sorted interpolation. Besides giving a common model-theoretic basis to these two important interpolation results, it also gives rise to a unified perspective on some of their known applications as well as a new one, namely a characterisation theorem due to van Benthem.

Acknowledgement. This investigation arose out of the context of a Stanford Logic Seminar discussion with Johan van Benthem and Solomon Feferman, revolving around different accounts of a characterisation result concerning preservation under Chu transforms and its relation to many-sorted interpolation. I am deeply indebted to both, Professor Feferman and Professor van Benthem, for their academic hospitality and personal kindness during my stay at Stanford as a visiting scholar in 1997/98.

The new interpolation result. Consider first-order logic *with or without equality* in a finite, purely relational vocabulary τ. Boolean constants $\top$ and $\bot$ are taken to be atomic first-order formulae. We use the set $\mathrm{Occ} = \tau \times \{+, -\}$ to code polarities of predicate occurrences in formulae, through a mapping

$$\mathrm{occ}\colon \varphi \longmapsto \mathrm{occ}(\varphi) \subseteq \mathrm{Occ},$$

where $(R, +) \in \mathrm{occ}(\varphi)$ if R occurs positively in φ, $(R, -) \in \mathrm{occ}(\varphi)$ if R occurs negatively. As usual $\mathrm{free}(\varphi)$ stands for the set of free variables in φ.

Let $\mathbb{U}$ be a tuple of designated unary predicates U in τ. We say that a formula φ is $\mathbb{U}$-*relativised*, if each quantifier in φ is explicitly relativised to some U, i.e., is of the form $\exists x\ (Ux \wedge \ldots)$ or $\forall x\ (Ux \rightarrow \ldots)$ for some U in $\mathbb{U}$. Up to logical equivalence, the $\mathbb{U}$-relativised formulae correspond exactly to the relativisations of first-order formulae to $\bigcup_{U \in \mathbb{U}} U$. This is actually even true up to a restricted form of logical equivalences which preserve polarities of predicate occurrences. We want to prove the following Lyndon style interpolation theorem.

THEOREM 1. *Let φ and ψ be $\mathbb{U}$-relativised formulae such that $\varphi \models \psi$. Then there is a $\mathbb{U}$-relativised Lyndon interpolant χ for $\varphi \models \psi$, i.e., a $\mathbb{U}$-relativised formula χ such that*

(i) $\mathrm{free}(\chi) \subseteq \mathrm{free}(\varphi) \cap \mathrm{free}(\psi)$.
(ii) $\mathrm{occ}(\chi) \subseteq \mathrm{occ}(\varphi) \cap \mathrm{occ}(\psi)$.
(iii) $\varphi \models \chi$ *and* $\chi \models \psi$.

The statement of the theorem has two readings, one for first-order with equality, and one for first-order without equality.

This strengthens Lyndon's Interpolation Theorem [10] (with or without $=$), which may be recovered from the theorem by trivialising the relativisation.

THEOREM 2 (Lyndon's Interpolation Theorem). *Let φ and ψ be formulae such that $\varphi \models \psi$. Then there is a Lyndon interpolant χ for $\varphi \models \psi$:*

(i) $\mathrm{free}(\chi) \subseteq \mathrm{free}(\varphi) \cap \mathrm{free}(\psi)$.
(ii) $\mathrm{occ}(\chi) \subseteq \mathrm{occ}(\varphi) \cap \mathrm{occ}(\psi)$.
(iii) $\varphi \models \chi$ *and* $\chi \models \psi$.

This follows from Theorem 1 if we put $\hat{\tau} = \tau \mathbin{\dot\cup} \{U\}$ for a new unary U, $\mathbb{U} = \{U\}$ and pass from φ and ψ to their relativisations to U, $\hat{\varphi}$ and $\hat{\psi}$. By the theorem, there is a U-relativised Lyndon interpolant $\hat{\chi}$ for $\bigwedge U\boldsymbol{x} \wedge \hat{\varphi} \models \hat{\psi}$. Here $\bigwedge U\boldsymbol{x}$ is shorthand for $\bigwedge_{i=1}^{n} Ux_i$, where $\boldsymbol{x} = (x_1, \dots, x_n)$ contains all variables free in φ or ψ. One obtains the desired Lyndon interpolant χ for $\varphi \models \psi$ by replacing every atom of the form Uy in $\hat{\chi}$ by $\top$.

We turn to the proof of Theorem 1, and first introduce some terminology and notation. The cases with and without equality can be treated together. Actually, the case with equality requires only one minor systematic modification, as will be indicated immediately. Let $\mathfrak{A}$ and $\mathfrak{B}$ be τ-structures with universes A and B respectively. We consider certain subsets $p \subseteq A \times B$ which are to be viewed as weak partial isomorphisms. *This notion calls for the one crucial modification if we want to deal with equality*: *in that case, and in that case only, all the p under consideration are required to be the graphs of partial one-to-one functions*; *in the case without equality, we consider a priori arbitrary subsets of $A \times B$.*

If $p \subseteq A \times B$, we regard $\mathrm{dm}(p) = \{\, a \in A \mid (\exists b \in B)((a,b) \in p) \,\}$ as the domain of p, and $\mathrm{im}(p) = \{\, b \in B \mid (\exists a \in A)((a,b) \in p) \,\}$ as the image of p. If $\boldsymbol{a} = (a_1, \dots, a_n)$, $\boldsymbol{b} = (b_1, \dots, b_n)$, and if $(a_i, b_i) \in p$ for $i = 1, \dots, n$, we indicate this situation by writing simply $\boldsymbol{a}\,\boldsymbol{b} \in p$.

Let $O \subseteq \mathrm{Occ}$. We say that *p preserves O* if for all $R \in \tau$, if R is n-ary and if $\boldsymbol{a} \in A^n$, $\boldsymbol{b} \in B^n$ are such that $\boldsymbol{a}\,\boldsymbol{b} \in p$, then

– if $(R,+) \in O$: $\mathfrak{A} \models R\boldsymbol{a} \Longrightarrow \mathfrak{B} \models R\boldsymbol{b}$,
– if $(R,-) \in O$: $\mathfrak{B} \models R\boldsymbol{b} \Longrightarrow \mathfrak{A} \models R\boldsymbol{a}$.

If $O \subseteq \mathrm{Occ}$ and $\mathfrak{A} = (A, (U^{\mathfrak{A}})_{U \in \mathbb{U}}, \dots)$ we let $A^{+/O}$ and $A^{-/O}$ denote those parts of the universe A that lie within some $U^{\mathfrak{A}}$ which is positive or negative, respectively, according to O:

$$A^{+/O} = \bigcup_{(U,+) \in O} U^{\mathfrak{A}} \quad \text{and} \quad A^{-/O} = \bigcup_{(U,-) \in O} U^{\mathfrak{A}}.$$

DEFINITION 3. A set $P \subseteq \{ p \subseteq A \times B \mid p \text{ preserves } O \}$ [1] is a *back-and-forth-system with respect to O and* $\mathbb{U}$, if $P \neq \emptyset$ is such that for all $p \in P$, and all $\boldsymbol{a}\,\boldsymbol{b} \in p$:

– if $a \in A^{+/O}$, then there are $p' \in P$ and $b \in B$ such that $\boldsymbol{a}a\boldsymbol{b}b \in p'$.
– if $b \in B^{-/O}$, then there are $p' \in P$ and $a \in A$ such that $\boldsymbol{a}a\boldsymbol{b}b \in p'$.

The following are two related, asymmetric notions of similarity between structures, one algebraic in spirit, the other semantic. They have the same relationship between them as do partial isomorphism and elementary equivalence.

DEFINITION 4. We write $(\mathfrak{A}, \boldsymbol{a}) \longrightarrow_O^{\mathbb{U}} (\mathfrak{B}, \boldsymbol{b})$ if there is a back-and-forth-system P with respect to O and $\mathbb{U}$ with $\boldsymbol{a}\,\boldsymbol{b} \in p$ for some $p \in P$. We also write $P\colon (\mathfrak{A}, \boldsymbol{a}) \longrightarrow_O^{\mathbb{U}} (\mathfrak{B}, \boldsymbol{b})$ in this situation, and $p\colon (\mathfrak{A}, \boldsymbol{a}) \longrightarrow_O^{\mathbb{U}} (\mathfrak{B}, \boldsymbol{b})$ if P consists of one single element $p \subseteq A \times B$.

DEFINITION 5. We write $(\mathfrak{A}, \boldsymbol{a}) \Longrightarrow_O^{\mathbb{U}} (\mathfrak{B}, \boldsymbol{b})$, if for all $\mathbb{U}$-relativised formulae $\varphi(\boldsymbol{x})$ with $\mathrm{occ}(\varphi) \subseteq O$: $\mathfrak{A} \models \varphi[\boldsymbol{a}] \Longrightarrow \mathfrak{B} \models \varphi[\boldsymbol{b}]$. For $p \subseteq A \times B$, we write $p\colon \mathfrak{A} \Longrightarrow_O^{\mathbb{U}} \mathfrak{B}$, if $(\mathfrak{A}, \boldsymbol{a}) \Longrightarrow_O^{\mathbb{U}} (\mathfrak{B}, \boldsymbol{b})$ for all $\boldsymbol{a}\,\boldsymbol{b} \in p$.

LEMMA 6. *If $P\colon \mathfrak{A} \longrightarrow_O^{\mathbb{U}} \mathfrak{B}$ then $p\colon \mathfrak{A} \Longrightarrow_O^{\mathbb{U}} \mathfrak{B}$ for all $p \in P$.*
In particular, $(\mathfrak{A}, \boldsymbol{a}) \longrightarrow_O^{\mathbb{U}} (\mathfrak{B}, \boldsymbol{b})$ implies $(\mathfrak{A}, \boldsymbol{a}) \Longrightarrow_O^{\mathbb{U}} (\mathfrak{B}, \boldsymbol{b})$.

This is proved by a straightforward induction on $\mathbb{U}$-relativised φ with $\mathrm{occ}(\varphi) \subseteq O$ in negation normal form.

LEMMA 7. *Let $\mathfrak{A}$ and $\mathfrak{B}$ be ω-saturated. Then $(\mathfrak{A}, \boldsymbol{a}_0) \Longrightarrow_O^{\mathbb{U}} (\mathfrak{B}, \boldsymbol{b}_0)$ implies*

$$P\colon (\mathfrak{A}, \boldsymbol{a}_0) \longrightarrow_O^{\mathbb{U}} (\mathfrak{B}, \boldsymbol{b}_0),$$

where $P = \{ p \subseteq A \times B \mid p \text{ finite}, p\colon \mathfrak{A} \Longrightarrow_O^{\mathbb{U}} \mathfrak{B} \}$.

PROOF. P is nonempty: $(p\colon \boldsymbol{a}_0 \mapsto \boldsymbol{b}_0) \in P$ as $(\mathfrak{A}, \boldsymbol{a}_0) \Longrightarrow_O^{\mathbb{U}} (\mathfrak{B}, \boldsymbol{b}_0)$. All $p \in P$ preserve O by definition. It remains to check the back-and-forth-property with respect to O and $\mathbb{U}$. We do the 'forth'-part. Let $\boldsymbol{a}\,\boldsymbol{b} \in p \in P$, and assume that $a \in U^{\mathfrak{A}}$, $(U, +) \in O$. We are seeking $b \in U^{\mathfrak{B}}$ for which $(\mathfrak{A}, \boldsymbol{a}a) \Longrightarrow_O^{\mathbb{U}} (\mathfrak{B}, \boldsymbol{b}b)$. Note that $(\mathfrak{A}, \boldsymbol{a}) \Longrightarrow_O^{\mathbb{U}} (\mathfrak{B}, \boldsymbol{b})$ holds, since $\boldsymbol{a}\,\boldsymbol{b} \in p$. Put

$$\Phi(\boldsymbol{a}, x) = \{ \varphi(\boldsymbol{a}, x) \mid \varphi \;\; \mathbb{U}\text{-relativised}, \mathrm{occ}(\varphi) \subseteq O, \mathfrak{A} \models \varphi[\boldsymbol{a}, a] \},$$

and let correspondingly $\Phi(\boldsymbol{b}, x) = \{ \varphi(\boldsymbol{b}, x) \mid \varphi(\boldsymbol{a}, x) \in \Phi(\boldsymbol{a}, x) \}$.

Clearly any realization b of $\Phi(\boldsymbol{b}, x)$ in $\mathfrak{B}$ will be such that $(\mathfrak{A}, \boldsymbol{a}a) \Longrightarrow_O^{\mathbb{U}} (\mathfrak{B}, \boldsymbol{b}b)$. It remains to show that $\Phi(\boldsymbol{b}, x)$ is consistent with $\mathrm{Th}(\mathfrak{B}, \boldsymbol{b})$. Assume to the contrary, that it were inconsistent. As Φ is closed under conjunctions, there would have to be a single $\varphi(\boldsymbol{a}, x) \in \Phi(\boldsymbol{a}, x)$ such that $(\mathfrak{B}, \boldsymbol{b}) \models \neg(\exists x \in U)\varphi(\boldsymbol{b}, x)$. But this contradicts $(\mathfrak{A}, \boldsymbol{a}) \Longrightarrow_O^{\mathbb{U}} (\mathfrak{B}, \boldsymbol{b})$, if we consider the formula $\psi(\boldsymbol{x}) = (\exists x \in U)\varphi(\boldsymbol{x}, x)$.

[1] Recall that, for the reading with equality, these p are restricted to be one-to-one partial functions.

The 'back'-part is dealt with analogously, only that the partial type under consideration is $\Phi(\boldsymbol{b}, x) = \{\, \neg\varphi(\boldsymbol{b}, x) \mid \varphi$ $\mathbb{U}$-relativised, $\mathrm{occ}(\varphi) \subseteq O, \mathfrak{B} \models \neg\varphi[\boldsymbol{b}, b]\,\}$. $\dashv$

Corollary 8. *For* $(\mathfrak{A}, \boldsymbol{a}_0) \Longrightarrow^{\mathbb{U}}_{O} (\mathfrak{B}, \boldsymbol{b}_0)$ *there are countable* $(\mathfrak{A}^*, \boldsymbol{a}_0)$ *and* $(\mathfrak{B}^*, \boldsymbol{b}_0)$ *such that* $(\mathfrak{A}^*, \boldsymbol{a}_0) \equiv (\mathfrak{A}, \boldsymbol{a}_0)$, $(\mathfrak{B}^*, \boldsymbol{b}_0) \equiv (\mathfrak{B}, \boldsymbol{b}_0)$, *and* $(\mathfrak{A}^*, \boldsymbol{a}_0) \longrightarrow^{\mathbb{U}}_{O} (\mathfrak{B}^*, \boldsymbol{b}_0)$.

This follows from the previous lemma, if we first take arbitrary ω-saturated elementary extensions $\mathfrak{A} \preccurlyeq \mathfrak{A}'$ and $\mathfrak{B} \preccurlyeq \mathfrak{B}'$. By the lemma, $(\mathfrak{A}', \boldsymbol{a}_0) \longrightarrow^{\mathbb{U}}_{O} (\mathfrak{B}', \boldsymbol{b}_0)$. To obtain a countable version of this situation, it suffices to wrap up $(\mathfrak{A}', \boldsymbol{a}_0)$, $(\mathfrak{B}', \boldsymbol{b}_0)$, and the back-and-forth-system P in one first-order structure, and to apply the Löwenheim-Skolem-Tarski Theorem.[2]

Lemma 9. *If* $\mathfrak{A}$ *and* $\mathfrak{B}$ *are countable, and* $(\mathfrak{A}, \boldsymbol{a}_0) \longrightarrow^{\mathbb{U}}_{O} (\mathfrak{B}, \boldsymbol{b}_0)$, *then there is a* $p \subseteq A \times B$ *for which* $p\colon \mathfrak{A} \longrightarrow^{\mathbb{U}}_{O} \mathfrak{B}$ *and* $\boldsymbol{a}_0\boldsymbol{b}_0 \in p$.

Sketch of proof. Starting from a back-and-forth-system P (without loss of generality consisting of finite p, and closed under subsets) and enumerations of $A^{+/O} \subseteq A$ and of $B^{-/O} \subseteq B$ one finds (in the usual back-and-forth fashion) finite approximations to the desired p within P. Their union p satisfies $\mathrm{dm}(p) \supseteq A^{+/O}$ and $\mathrm{im}(p) \supseteq B^{-/O}$. $\dashv$

The following is the main proposition towards the proof of the theorem; indeed, it may be thought of as the *structural interpolation property* behind the theorem. The situation is depicted in the following sketch.

$$\begin{tikzcd}
\mathfrak{A} \arrow[rr, "p", "O"'] \arrow[dr, "p_1"', "O_1"] & & \mathfrak{B} \\
& \mathfrak{C} \arrow[ur, "p_2"', "O_2"] &
\end{tikzcd}$$

Proposition 10 (Main proposition). *Let* $O_1, O_2 \subseteq \mathrm{Occ}$, $O = O_1 \cap O_2$. *If* $p\colon \mathfrak{A} \longrightarrow^{\mathbb{U}}_{O} \mathfrak{B}$, *then there are* $\mathfrak{C}$, $p_1 \subseteq A \times C$, *and* $p_2 \subseteq C \times B$ *such that* $p = p_1 \circ p_2$ *and* $p_1\colon \mathfrak{A} \longrightarrow^{\mathbb{U}}_{O_1} \mathfrak{C}$ *and* $p_2\colon \mathfrak{C} \longrightarrow^{\mathbb{U}}_{O_2} \mathfrak{B}$. *One may further require that* $\mathrm{dm}(p_1) = A$ *and* $\mathrm{im}(p_2) = B$.

Proof. Let $* \notin A \cup B$. We let the universe C of the desired structure $\mathfrak{C}$ be a subset of $(A \mathbin{\dot\cup} \{*\}) \times (B \mathbin{\dot\cup} \{*\})$. Put

[2]Alternatively, one might want to work with encodings of pairs of structures and back-and-forth-systems from the start, and prove Corollary 8 directly from a model theoretic games perspective, via direct applications of compactness and Löwenheim-Skolem. This alternative approach, which avoids the detour through ω-saturated structures, has been elaborated on and applied to a uniform treatment of numerous characterisation results in [8].

$$C = p \mathbin{\dot\cup} ((A \smallsetminus \mathrm{dm}(p)) \times \{*\}) \mathbin{\dot\cup} (\{*\} \times (B \smallsetminus \mathrm{im}(p))),$$
$$p_1 = \{\, (a,(a,b)) \mid (a,b) \in p \,\} \mathbin{\dot\cup} \{\, (a,(a,*)) \mid a \in A \smallsetminus \mathrm{dm}(p) \,\},$$
$$p_2 = \{\, ((a,b),b) \mid (a,b) \in p \,\} \mathbin{\dot\cup} \{\, ((*,b),b) \mid b \in B \smallsetminus \mathrm{im}(p) \,\}.$$

Directly from the definitions we see that $p = p_1 \circ p_2$, $\mathrm{dm}(p_1) = A$, and $\mathrm{im}(p_2) = B$. For the case with equality also note that the p_i are one-to-one partial functions if p is. We now have to fix the interpretation of predicates $R \in \tau$ over C with a view to making p_1 respect O_1, to making p_2 respect O_2, and to guaranteeing the respective back-and-forth properties. Since $\mathrm{dm}(p_1) = A$ and $\mathrm{im}(p_2) = B$, the remaining back-and-forth conditions reduce to $\mathrm{im}(p_1) \supseteq C^{-/O_1}$ and $\mathrm{dm}(p_2) \supseteq C^{+/O_2}$.

Let R be n-ary, $\boldsymbol{c} \in C^n$. We write $\boldsymbol{c} = \boldsymbol{a}\,\boldsymbol{b}$, where in general $\boldsymbol{a} \in (A \mathbin{\dot\cup} \{*\})^n$ and $\boldsymbol{b} \in (B \mathbin{\dot\cup} \{*\})^n$. We distinguish several cases.

(a) If $\boldsymbol{a} \in A^n$ and $\boldsymbol{b} \in B^n$ and $\mathfrak{A} \models R\boldsymbol{a} \iff \mathfrak{B} \models R\boldsymbol{b}$, put: $\mathfrak{C} \models R\boldsymbol{c} \iff \mathfrak{A} \models R\boldsymbol{a}$.
(b) If $\boldsymbol{a} \in A^n$ and $\boldsymbol{b} \in B^n$ and $\mathfrak{A} \models R\boldsymbol{a}$ but $\mathfrak{B} \models \neg R\boldsymbol{b}$ (whence $(R,+) \notin O$):
  (i) if $(R,+) \in O_1 \smallsetminus O_2$, put: $\mathfrak{C} \models R\boldsymbol{c}$.
  (ii) if $(R,+) \in O_2 \smallsetminus O_1$, put: $\mathfrak{C} \models \neg R\boldsymbol{c}$.
  (iii) if $(R,+) \notin O_1 \cup O_2$, decide $R\boldsymbol{c}$ arbitrarily.
(c) If $\boldsymbol{a} \in A^n$ and $\boldsymbol{b} \in B^n$ and $\mathfrak{A} \models \neg R\boldsymbol{a}$ but $\mathfrak{B} \models R\boldsymbol{b}$ (which implies that $(R,-) \notin O$), proceed as in (b), based on a corresponding case distinction for $(R,-)$.
(d) If $\boldsymbol{a} \in A^n$ and $\boldsymbol{b} \notin B^n$ (i.e., $\boldsymbol{a} \notin \mathrm{dm}(p)$, and $\boldsymbol{c} \notin \mathrm{dm}(p_2)$), put: $\mathfrak{C} \models R\boldsymbol{c} \iff (\mathfrak{A} \models R\boldsymbol{a}$ and $(R,+) \in O_1)$.
(e) If $\boldsymbol{a} \notin A^n$ and $\boldsymbol{b} \in B^n$ (i.e., $\boldsymbol{b} \notin \mathrm{im}(p)$, and $\boldsymbol{c} \notin \mathrm{im}(p_1)$), put: $\mathfrak{C} \models R\boldsymbol{c} \iff (\mathfrak{B} \models R\boldsymbol{b}$ and $(R,-) \in O_2)$.
(f) If $\boldsymbol{a} \notin A^n$ and $\boldsymbol{b} \notin B^n$, decide $R\boldsymbol{c}$ arbitrarily.

It remains to argue that p_i preserves O_i, that $\mathrm{im}(p_1) \supseteq C^{-/O_1}$, and that $\mathrm{dm}(p_2) \supseteq C^{+/O_2}$. Note for the following that if, e.g., $\boldsymbol{a}\,\boldsymbol{c} \in p_1$ then $\boldsymbol{a} \in A^n$ and $\boldsymbol{c} = \boldsymbol{a}\,\boldsymbol{b}$ for some $\boldsymbol{b} \in (B \mathbin{\dot\cup} \{*\})^n$, where $\boldsymbol{a}'\boldsymbol{b}' \in p$ if the primed tuples are the projections of $\boldsymbol{a}$ and $\boldsymbol{b}$ to those positions in which $\boldsymbol{b}$ has entries from B.

• p_1 preserves O_1. Consider $\boldsymbol{a}\,\boldsymbol{c} \in p_1$ and $(R,+) \in O_1$; the relevant cases are those in which $\mathfrak{A} \models R\boldsymbol{a}$, and are dealt with in (a), (b), or (d). If $(R,-) \in O_1$, then we look at $\mathfrak{A} \models \neg R\boldsymbol{a}$, and find that the relevant cases are treated in (a), (c), or (d).

• $\mathrm{im}(p_1) \supseteq C^{-/O_1}$. Note that all elements of C except those of the form $c = (*,b)$ are trivially in $\mathrm{im}(p_1)$. For $(*,b) \in C$ it follows that $b \notin \mathrm{im}(p)$. If $(*,b) \in U^{\mathfrak{C}}$, then by (e), $(U,-) \in O_2$ and $b \in U^{\mathfrak{B}}$. Therefore, $(U,-)$ cannot be in O_1, as that would imply $(U,-) \in O$ and $b \in \mathrm{im}(p)$. So no element of the form $(*,b)$ is in C^{-/O_1}.

That p_2 preserves O_2 and that $\mathrm{dm}(p_2) \supseteq C^{+/O_2}$ are shown analogously. This finishes the proof of the proposition. ⊣

Proof of Theorem 1. Let φ and ψ be $\mathbb{U}$-relativised formulae, $\varphi \models \psi$. Let $\boldsymbol{x}_0$ stand for the tuple of variables that are free in both φ and ψ, so that for appropriate tuples $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, we have pairwise disjoint tuples $\boldsymbol{x}_i$ such that $\varphi = \varphi(\boldsymbol{x}_0, \boldsymbol{x}_1)$ and $\psi = \psi(\boldsymbol{x}_0, \boldsymbol{x}_2)$ with all free variables displayed. Put $O_1 = \mathrm{occ}(\varphi)$, $O_2 = \mathrm{occ}(\psi)$, $O = O_1 \cap O_2$. Suppose there were no $\mathbb{U}$-relativised formula $\chi(\boldsymbol{x}_0)$ with $\mathrm{occ}(\chi) \subseteq O$ such that $\varphi \models \chi$ and $\chi \models \psi$. Put

$$\Phi(\boldsymbol{x}_0) = \{\, \chi(\boldsymbol{x}_0) \mid \chi \ \ \mathbb{U}\text{-relativised}, \mathrm{occ}(\chi) \subseteq O, \varphi \models \chi \,\}.$$

By assumption (and compactness), $\Phi \not\models \psi$. Let $(\mathfrak{B}, \boldsymbol{b}_0, \boldsymbol{b}_2) \models \Phi \cup \{\neg\psi\}$. Put

$$\Theta(\boldsymbol{x}_0) = \{\, \neg\chi(\boldsymbol{x}_0) \mid \chi \ \ \mathbb{U}\text{-relativised}, \mathrm{occ}(\chi) \subseteq O, (\mathfrak{B}, \boldsymbol{b}_0) \models \neg\chi \,\}.$$

It follows that $\Theta \cup \{\varphi\}$ is consistent, too. Otherwise, by compactness and as Θ is closed under conjunctions, there would have to be a single $\neg\chi \in \Theta$ such that $\varphi \models \chi$. But then $\chi \in \Phi$, whence $\mathfrak{B} \models \chi$, a contradiction.

So we find some $(\mathfrak{A}, \boldsymbol{a}_0, \boldsymbol{a}_1) \models \Theta \cup \{\varphi\}$. It follows that $(\mathfrak{A}, \boldsymbol{a}_0) \Longrightarrow^{\mathbb{U}}_{O} (\mathfrak{B}, \boldsymbol{b}_0)$. By Corollary 8 the situation can be upgraded to obtain countable $(\mathfrak{A}, \boldsymbol{a}_0, \boldsymbol{a}_1)$ and $(\mathfrak{B}, \boldsymbol{b}_0, \boldsymbol{b}_2)$ such that $(\mathfrak{A}, \boldsymbol{a}_0) \longrightarrow^{\mathbb{U}}_{O} (\mathfrak{B}, \boldsymbol{b}_0)$, $\mathfrak{A} \models \varphi[\boldsymbol{a}_0, \boldsymbol{a}_1]$, and $\mathfrak{B} \models \neg\psi[\boldsymbol{b}_0, \boldsymbol{b}_2]$.

Applying Lemma 9, we obtain $p \colon \mathfrak{A} \longrightarrow^{\mathbb{U}}_{O} \mathfrak{B}$ for a single $p \subseteq A \times B$ with $\boldsymbol{a}_0\boldsymbol{b}_0 \in p$. But now the proposition yields a structure $\mathfrak{C}$ and relations p_1 and p_2 such that $p = p_1 \circ p_2$, $\boldsymbol{a}_0\boldsymbol{a}_1 \in \mathrm{dm}(p_1)$, $\boldsymbol{b}_0\boldsymbol{b}_2 \in \mathrm{im}(p_2)$, and $p_1 \colon \mathfrak{A} \longrightarrow^{\mathbb{U}}_{O_1} \mathfrak{C}$, $p_2 \colon \mathfrak{C} \longrightarrow^{\mathbb{U}}_{O_2} \mathfrak{B}$. Let $\boldsymbol{c}_0$ be such that $\boldsymbol{a}_0\boldsymbol{c}_0 \in p_1$ and $\boldsymbol{c}_0\boldsymbol{b}_0 \in p_2$ (recall that $\boldsymbol{a}_0\boldsymbol{b}_0 \in p$ and $p = p_1 \circ p_2$). Let further $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ be such that $\boldsymbol{a}_0\boldsymbol{a}_1\, \boldsymbol{c}_0\boldsymbol{c}_1 \in p_1$ and $\boldsymbol{c}_0\boldsymbol{c}_2\, \boldsymbol{b}_0\boldsymbol{b}_2 \in p_2$ (these exist, since $\boldsymbol{a}_1 \in \mathrm{dm}(p_1)$ and $\boldsymbol{b}_2 \in \mathrm{im}(p_2)$, respectively). It follows that $(\mathfrak{A}, \boldsymbol{a}_0, \boldsymbol{a}_1) \Longrightarrow^{\mathbb{U}}_{O_1} (\mathfrak{C}, \boldsymbol{c}_0, \boldsymbol{c}_1)$ and $(\mathfrak{C}, \boldsymbol{c}_0, \boldsymbol{c}_2) \Longrightarrow^{\mathbb{U}}_{O_2} (\mathfrak{B}, \boldsymbol{b}_0, \boldsymbol{b}_2)$. We find that by preservation $\mathfrak{C} \models \varphi[\boldsymbol{c}_0, \boldsymbol{c}_1]$, therefore, and as $\varphi \models \psi$, we have $\mathfrak{C} \models \psi[\boldsymbol{c}_0, \boldsymbol{c}_2]$, whence by preservation also $\mathfrak{B} \models \psi[\boldsymbol{b}_0, \boldsymbol{b}_2]$, a contradiction. $\dashv$

Applications. We show how Theorem 1 actually combines several interpolation and preservation results in one. Namely, it directly implies not only Lyndon's interpolation theorem [10] (see Theorem 2 above), but also a number of related interpolation and preservation results. Among them are the classical characterisation theorems concerning preservation under extensions and substructures [3], a variant of Feferman's many-sorted interpolation theorem [5, 6], as well as a very recent characterisation result of van Benthem's [2] concerning preservation under Chu transforms.

Classical preservation theorems. Consider a relational first-order formula $\varphi(\boldsymbol{x}) = \varphi(x_1, \ldots, x_k)$ that is preserved under extensions: $\mathfrak{A} \subseteq \mathfrak{B}$ and $\mathfrak{A} \models \varphi[\boldsymbol{a}]$ together imply that $\mathfrak{B} \models \varphi[\boldsymbol{a}]$. Let U_1, U_2 be two new unary predicates, not in the vocabulary of φ, and put $\mathbb{U} = \{U_1, U_2\}$. Let φ^{U_i} be the result of relativising φ to U_i. Then preservation under extensions for φ is expressed by the following validity:

$$\forall x\ (U_1 x \to U_2 x) \wedge \bigwedge U_1 \boldsymbol{x} \wedge \varphi^{U_1}(\boldsymbol{x}) \models \varphi^{U_2}(\boldsymbol{x}).$$

Note that all formulae occurring here are $\mathbb{U}$-relativised. Theorem 1 provides a $\mathbb{U}$-relativised interpolant $\chi(\boldsymbol{x})$, with no occurrence of U_1 and only positive occurrences of U_2. It follows that χ is purely existential. Considering the special case that $U_1 = U_2$ in the antecedent, we find that $\bigwedge U_2 \boldsymbol{x} \wedge \varphi^{U_2} \models \chi \models \varphi^{U_2}$. Further restricting attention to the case that $\forall x U_2 x$ in these implications, we replace in χ all atoms $U_2 y$ by $\top$ to obtain a formula χ', which is in the vocabulary of φ, purely existential, and equivalent to φ, since in this special case both $\bigwedge U_2 \boldsymbol{x} \wedge \varphi^{U_2}$ and φ^{U_2} are equivalent to φ.

We note that a similar connection between existential preservation and a *many-sorted* interpolation property (to be dealt with in the next section) is prominently discussed as an application for many-sorted interpolation by Feferman in [6].

As usual, a similar argument gives the classical preservation theorem concerning monotonicity and positivity (which of course directly follows from Lyndon's Interpolation Theorem). The point in the above is rather that the extension/existential and the monotone/positive preservation phenomena are attributed to the same source, as it were, in the present relativised picture.

A many-sorted interpolation theorem. Consider many-sorted first-order logic with n sorts in a relational vocabulary. We here assume that all predicates and variables are typed with respect to the sorts and that the sorts are disjoint. [3] The standard translation into a one-sorted framework transforms a many-sorted relational structure $(A_1, \dots, A_n, R, \dots)$ with disjoint sorts A_i into a structure whose universe A is (an arbitrary superset of) $\bigcup A_i$ having new unary predicates $U_1, \dots, U_n$ to indicate the different subdomains corresponding to the different sorts. If we put $\mathbb{U} = (U_1, \dots, U_n)$, the many-sorted first-order formulae naturally translate into $\mathbb{U}$-relativised formulae. Care has to be taken if free variables are around: while the many-sorted framework restricts each free variable to its particular sort implicitly, the corresponding semantic restriction has to be made explicitly in the translation. Care has also to be taken with respect to the translation of implications. A priori the validity of a many-sorted implication implies the validity of the $\mathbb{U}$-relativised one-sorted translated implication only in

[3] However, we do *not* a priori require all sorts to be non-empty; corresponding stipulations would have to be made explicitly by means of existential statements. If typed-ness of predicates or disjointness of sorts seem too restrictive, there are straightforward modifications which can deal with those other settings; in fact, some precautions that have to be taken here can be avoided in those more liberal settings.

restriction to those one-sorted structures that arise as encodings of the original many-sorted structures. Those are the structures for which the U_s are disjoint, each relation R is restricted to a product of sub-domains U_s according to its specified type, and each free variable x_i is interpreted as an element of that $U_{s(i)}$ that corresponds to its sort. For well-formed many-sorted formulae, however, the restriction concerning the interpretation of relations R is irrelevant because R-atoms in what would be inappropriate sorts do not affect the semantics of the translated formula (as long as the free variables are restricted to the appropriate sorts). The restriction of the free variables to their respective sorts will be made explicit below. So what remains is the apparent problem about disjoint versus overlapping sub-domains U_s, which is resolved in the light of the following reverse transformation. Let $\mathfrak{A} = (A, U_1, \dots, U_n, R, \dots)$ be an arbitrary structure of the indicated type, $\boldsymbol{a} = (a_1, \dots, a_k) \in A^k$ such that $a_i \in U^{\mathfrak{A}}_{s(i)}$ for $i = 1, \dots, k$. Let $\check{U}_s = U_s^{\mathfrak{A}} \times \{s\}$, and $\check{R} = \{\, \boldsymbol{a} = ((a_1, s_1), \dots, (a_r, s_r)) \mid (a_1, \dots, a_r) \in R^{\mathfrak{A}} \,\}$ for an r-ary relation R of type $(s_1, \dots, s_r)$. If $\varphi(x_1, \dots, x_k)$ is a well-formed many-sorted formula with x_i of sort $S_{s(i)}$, and $\varphi^{\mathbb{U}}$ the result of relativising every quantifier in φ to the appropriate U_s, then

$$\mathfrak{A} \models \varphi^{\mathbb{U}}[\boldsymbol{a}] \quad \text{if and only if} \quad (\check{U}_1, \dots, \check{U}_n, \check{R}, \dots) \models \varphi[(a_1, s_1), \dots, (a_k, s_k)].$$

Note that this is even true if φ has equality, because the typed-ness of φ implies that equality atoms can only link variables from the same sort. With these considerations in mind, we find that a valid many-sorted implication $\varphi \models \psi$, with combined free variables $x_1, \dots, x_k$, x_i of sort $s(i)$, gives rise to the following valid one-sorted implication between $\mathbb{U}$-relativised formulae:

$$\bigwedge_{x_i \in \mathrm{free}(\varphi)} U_{s(i)} x_i \wedge \varphi^{\mathbb{U}} \models \bigwedge_{x_i \in \mathrm{free}(\psi)} U_{s(i)} x_i \longrightarrow \psi^{\mathbb{U}}, \tag{1}$$

where $\varphi^{\mathbb{U}}$ and $\psi^{\mathbb{U}}$ are the results of relativising every quantifier in these formulae to the appropriate sort.

We first specialise to the case in which φ and ψ share the same free variables $x_1, \dots, x_k$. This assumption makes the analysis more straightforward, but the general case essentially reduces to it: if $\varphi = \varphi(\boldsymbol{x}, \boldsymbol{y})$, $\psi = \psi(\boldsymbol{x}, \boldsymbol{z})$ with all free variables mentioned and $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ pairwise disjoint, then $\varphi \models \psi$ implies that $\exists \boldsymbol{y}\ \varphi \models \forall \boldsymbol{z}\ \psi$; furthermore, an interpolant for the latter implication automatically is good for $\varphi \models \psi$, too. We shall return to this general case below.

Let free -sorts be the set of sorts of the free variables of φ and ψ. Identifying a sort with its encoding predicate U, we may think of free -sorts as a subset of $\mathbb{U}$. A direct application of Theorem 1 to the valid implication (1) yields an interpolant $\chi(x_1, \dots, x_k)$, which is $\mathbb{U}$-relativised and has

$$(2) \quad \mathrm{occ}(\chi) \subseteq (\mathrm{occ}(\varphi^{\mathbb{U}}) \cup \text{free-sorts} \times \{+\}) \cap (\mathrm{occ}(\psi^{\mathbb{U}}) \cup \text{free-sorts} \times \{-\}).$$

Now χ is not at first of the form that would immediately translate back into a many-sorted formula. But note that, since $\bigwedge_i U_{s(i)}x_i \wedge \varphi^{\mathbb{U}} \models \chi(x_1, \dots, x_k)$, all free variables of χ may be assumed to be appropriately typed. It remains to remove in χ all wrongly typed occurrences of predicates R, and all occurrences of predicates U_s apart from occurrences in explicit relativisations. Without loss of generality each bound variable in χ is quantified exactly once, and therefore also has a unique sort attributed to it. Any R-atom in χ that may in this fashion acquire an inappropriate type is now replaced by $\perp$, the same goes for atoms $U_s y$, whenever y is a variable attributed to any sort different from U_s. Clearly, the resulting formula χ' is still good as an interpolant for the implication (1) in restriction to all structures that arise as one-sorted encodings of many-sorted structures. Also observe that none of these manipulations introduces new occurrences. Thus χ' may be regarded as a well-formed many-sorted formula, for which $\varphi \models \chi'$ and $\chi' \models \psi$ are valid many-sorted implications.

Now for occurrences of predicates in χ (or χ'). With respect to predicates in the original vocabulary of φ and ψ, we clearly have $\mathrm{occ}(\chi') \subseteq \mathrm{occ}(\chi) \subseteq \mathrm{occ}(\varphi) \cap \mathrm{occ}(\psi)$, by (2)—as is to be expected for a many-sorted variant of Lyndon interpolation. Moreover, $\mathbb{U}$-occurrences can be used to analyse which sorts are existentially or universally quantified.

Let $\exists$-sorts(φ) stand for the set of those sorts that are existentially quantified in φ, $\forall$-sorts(φ) the set of those that are universally quantified (assuming negation normal form), and sorts(φ) for the set of sorts occurring in φ. Then (2) implies that $\mathrm{sorts}(\chi') \subseteq \mathrm{sorts}(\varphi) \cap \mathrm{sorts}(\psi)$ and that

$$\begin{aligned} \exists\text{-sorts}(\chi') &\subseteq \exists\text{-sorts}(\psi) \cap (\exists\text{-sorts}(\varphi) \cup \text{free-sorts}) \\ \forall\text{-sorts}(\chi') &\subseteq \forall\text{-sorts}(\varphi) \cap (\forall\text{-sorts}(\psi) \cup \text{free-sorts}). \end{aligned}$$

Considering now the general case of formulae $\varphi(\boldsymbol{x}, \boldsymbol{y})$ and $\psi(\boldsymbol{x}, \boldsymbol{z})$ not necessarily sharing the same free variables, the passage to $\exists \boldsymbol{y}\ \varphi(\boldsymbol{x}, \boldsymbol{y})$ and $\forall \boldsymbol{z}\ \psi(\boldsymbol{x}, \boldsymbol{z})$ also yields a corresponding interpolant for $\varphi \models \psi$, whose free variables are among the common free variables of φ and ψ. With respect to the existentially (universally) quantified sorts in the above analysis, we have to replace $\exists$-sorts(φ) by $\exists$-sorts$(\varphi) \cup$ free-sorts(φ) and $\forall$-sorts(ψ) by $\forall$-sorts$(\psi) \cup$ free-sorts(ψ). Everything else remains unchanged. We thus obtain the following many-sorted interpolation theorem, which is closely related to Feferman's many-sorted interpolation theorem [5, 6] and its generalisation by Stern [12].

PROPOSITION 11. *Suppose φ and ψ are many-sorted relational formulae. If $\varphi \models \psi$ is a valid implication, then there is a many-sorted Lyndon interpolant χ for $\varphi \models \psi$, i.e., a many-sorted formula χ such that $\varphi \models \chi$ and $\chi \models \psi$, and*

(i) $\mathrm{free}(\chi) \subseteq \mathrm{free}(\varphi) \cap \mathrm{free}(\psi)$.
(ii) $\mathrm{sorts}(\chi) \subseteq \mathrm{sorts}(\varphi) \cap \mathrm{sorts}(\psi)$.
(iii) $\mathrm{occ}(\chi) \subseteq \mathrm{occ}(\varphi) \cap \mathrm{occ}(\psi)$.
(iv) $\exists\text{-sorts}(\chi) \subseteq \exists\text{-sorts}(\psi) \cap (\exists\text{-sorts}(\varphi) \cup \text{free-sorts}(\varphi))$,
$\forall\text{-sorts}(\chi) \subseteq \forall\text{-sorts}(\varphi) \cap (\forall\text{-sorts}(\psi) \cup \text{free-sorts}(\psi))$.

Note that in the case of sentences (i.e., without free variables) condition (iv) simply says that a sort is existentially (universally) quantified in χ only if it is existentially (universally) quantified in both φ and ψ. We thus have in particular the following, which is just Stern's Theorem 2.2 in [12].

COROLLARY 12 (Stern). *For many-sorted relational sentences φ and ψ without free variables: if $\varphi \models \psi$ is a valid implication, then there is a many-sorted Lyndon interpolant χ for $\varphi \models \psi$, i.e., a many-sorted sentence χ such that $\varphi \models \chi$ and $\chi \models \psi$, and*

(i) $\mathrm{sorts}(\chi) \subseteq \mathrm{sorts}(\varphi) \cap \mathrm{sorts}(\psi)$.
(ii) $\mathrm{occ}(\chi) \subseteq \mathrm{occ}(\varphi) \cap \mathrm{occ}(\psi)$.
(iii) $\exists\text{-sorts}(\chi) \subseteq \exists\text{-sorts}(\psi) \cap \exists\text{-sorts}(\varphi)$,
$\forall\text{-sorts}(\chi) \subseteq \forall\text{-sorts}(\varphi) \cap \forall\text{-sorts}(\psi)$.

But also for formulae with free variables, one may play with the implication (1) and distribute the sort conditions $U_{s(i)}x_i$ in such a way as to minimise matching occurrences of sort predicates between antecedent and succedent. In particular, the conditions on the common free variables obviously need not be repeated on both sides. In this way one finds, for instance in the case that φ and ψ share the same free variables (of sorts free-sorts), that for any sets F_1 and F_2 of sorts such that free-sorts $\subseteq F_1 \cup F_2$ condition (iv) may be replaced by

(iv′) $\exists\text{-sorts}(\chi) \subseteq \exists\text{-sorts}(\psi) \cap (\exists\text{-sorts}(\varphi) \cup F_1)$,
$\forall\text{-sorts}(\chi) \subseteq \forall\text{-sorts}(\varphi) \cap (\forall\text{-sorts}(\psi) \cup F_2)$.

Similar variations are explicit in Stern's generalisations of Feferman's theorem, Theorem 2.1 in [12]. The specialisation of Feferman's original many-sorted interpolation theorem [5, 6] corresponding to Proposition 11 is the following.

THEOREM 13 (Feferman). *Suppose φ and ψ are many-sorted relational formulae. If $\varphi \models \psi$ is a valid implication, then there is a many-sorted interpolant χ, $\varphi \models \chi$ and $\chi \models \psi$, such that*

(i) $\mathrm{free}(\chi) \subseteq \mathrm{free}(\varphi) \cap \mathrm{free}(\psi)$.
(ii) $\mathrm{sorts}(\chi) \subseteq \mathrm{sorts}(\varphi) \cap \mathrm{sorts}(\psi)$.
(iii) *all predicates in χ occur in both φ and ψ.*
(iv) $\exists\text{-sorts}(\chi) \subseteq \exists\text{-sorts}(\psi)$, $\forall\text{-sorts}(\chi) \subseteq \forall\text{-sorts}(\varphi)$.

The strength of Feferman's many-sorted interpolation theorem has been demonstrated in several applications in [5, 6] and also [1]. In particular, the classical extension/existential preservation theorem also follows directly from it, as mentioned above. In comparison with the full version of Feferman's theorem, Proposition 11 concerns just first-order, rather than fragments of infinitary logic. On the other hand our variant (just like Stern's strengthening) goes beyond Feferman's theorem in two other respects, apart from being obtained along very different lines: the Lyndon condition on the polarities of predicate occurrences is absent from Theorem 13, and the condition on existentially and universally quantified sorts is more symmetric and in general stronger than the corresponding condition in Feferman's theorem. [4]

Van Benthem's preservation theorem. This new characterisation result [2] arises in the context of Chu spaces (over $\{0,1\}$), regarded as two-sorted relational structures in one binary E of type $(1,2)$. The latter is to say that E is interpreted as a subset of the cartesian product of the first and the second sort. A Chu transform between two such two-sorted E-structures $\mathfrak{A} = (A_1, A_2, E^{\mathfrak{A}})$ and $\mathfrak{B} = (B_1, B_2, E^{\mathfrak{B}})$ is a pair of mappings $f: A_1 \to B_1$ and $g: B_2 \to A_2$ such that for all $a \in A_1$ and $\beta \in B_2$:

$$(a, g(\beta)) \in E^{\mathfrak{A}} \iff (f(a), \beta) \in E^{\mathfrak{B}}.$$

Note that the first sort transforms as under a homomorphism, while the second sort transforms as under an inverse homomorphism. Special Chu transforms are the following:

(a) extensions in the first sort: $A_1 \subseteq B_1,\ A_2 = B_2,\ \mathfrak{A} \subseteq \mathfrak{B},\ f, g = \mathrm{id}$.
(b) restrictions in the second sort: $A_1 = B_1,\ A_2 \supseteq B_2,\ \mathfrak{A} \supseteq \mathfrak{B},\ f, g = \mathrm{id}$.

Let the variables of the first sort be denoted x, those of the second sort v. A two-sorted first-order formula $\varphi(\boldsymbol{x}, \boldsymbol{v})$ is preserved under Chu transforms if, whenever f, g constitute a Chu transform from $\mathfrak{A}$ to $\mathfrak{B}$, and if $\boldsymbol{a} \in (A_1)^n$ and $\boldsymbol{\beta} \in (B_2)^m$, then

$$\mathfrak{A} \models \varphi[\boldsymbol{a}, g(\boldsymbol{\beta})] \Longrightarrow \mathfrak{B} \models \varphi[f(\boldsymbol{a}), \boldsymbol{\beta}].$$

Following van Benthem [2], let us call a many-sorted formula $\varphi(\boldsymbol{x}, \boldsymbol{v})$ in this two-sorted framework a *flow-formula* if it is equality-free and is purely existential in the first sort, and purely universal in the second. The following theorem is presented with a self-contained model theoretic proof in [2]. Its connection with Feferman's and Stern's many-sorted interpolation theorems has meanwhile also been expounded in [7].

[4] It is worth noting that this is not only due to the fact that our sorts are not a priori required to be non-empty; the use of explicit conditions to this effect in the constituent formulae still yields a more restrictive and more symmetric condition on the interpolant than the one discussed in [5, 6].

THEOREM 14 (van Benthem). *Formulae preserved under Chu transforms are precisely those that are logically equivalent to flow-formulae.*

Suppose $\varphi(\boldsymbol{x}, \boldsymbol{v})$ is an equality-free [5] two-sorted first-order formula that is preserved under the special Chu transforms of types (a) and (b). For a proof of van Benthem's theorem we want to obtain a flow formula φ^* (existential with respect to the first sort and universal with respect to the second) that is equivalent to φ. We obtain φ^* as an interpolant in an application of our main theorem (in the equality-free setting) for the natural one-sorted encodings of Chu spaces. Let us use unary predicates U and V for the sub-domains corresponding to the first and second sort, respectively. Let $\varphi^{\mathbb{U}}$ be the result of relativising all quantifiers in φ accordingly, to either U or V. In the light of the treatment of the many-sorted interpolation theorem outlined above, it will suffice to show that $\varphi^{\mathbb{U}}$ is equivalent to a $\mathbb{U}$-relativised formula in which U occurs only positively and V only negatively. This is achieved in two separate steps, the first one using Chu transforms of type (a) to eliminate negative occurrences of U, the second one using Chu transforms of type (b) to eliminate positive occurrences of V. For the first step let $\varphi_1 = \varphi^{\mathbb{U}}$, φ_2 the result of renaming U to U' in $\varphi^{\mathbb{U}}$. Then preservation of φ under Chu transforms of type (a) corresponds to the validity of

$$\varphi_2(\boldsymbol{x}, \boldsymbol{v}) \wedge \bigwedge U'\boldsymbol{x} \wedge \bigwedge V\boldsymbol{v} \wedge \forall x\ (U'x \rightarrow Ux) \models \varphi_1(\boldsymbol{x}, \boldsymbol{v}).$$

This is an implication between $\{U, V, U'\}$-relativised formulae, and Theorem 1 yields a $\{U, V, U'\}$-relativised formula $\chi(\boldsymbol{x}, \boldsymbol{v})$ with no occurrence of U' and only positive occurrences of U. Identifying U and U', we find that $\varphi^{\mathbb{U}}$ and χ are indeed equivalent if $\boldsymbol{x}$ and $\boldsymbol{v}$ respect the sorts. For the second step assume that $\varphi^{\mathbb{U}}$ is positive in U, and let $\varphi_1 = \varphi^{\mathbb{U}}$, φ_2 the result of renaming V to V' in $\varphi^{\mathbb{U}}$. Preservation of φ under Chu transforms of type (b) means that

$$\varphi_2(\boldsymbol{x}, \boldsymbol{v}) \wedge \bigwedge U\boldsymbol{x} \wedge \forall x\ (Vx \rightarrow V'x) \models \bigwedge V\boldsymbol{v} \longrightarrow \varphi_1(\boldsymbol{x}, \boldsymbol{v}).$$

Theorem 1 now yields a $\{U, V, V'\}$-relativised formula $\chi(\boldsymbol{x}, \boldsymbol{v})$, still positive in U, with no occurrence of V' and only negative occurrences of V. Identifying V and V', we find that $\varphi^{\mathbb{U}}$ and χ are indeed equivalent if $\boldsymbol{x}$ and $\boldsymbol{v}$ respect the sorts. This new χ then can essentially serve as φ^*. Up to some necessary syntactic clean-up as we saw in the treatment of many-sorted interpolation, φ^* may be translated back into a two-sorted formula in the

[5] To recover the full content of van Benthem's theorem, one should not require the absence of equality, but obtain it in a preliminary step based on preservation under onto-homomorphisms in the first sort (which goes to eliminate inequalities over the first sort) and inverse onto-homomorphisms in the second sort (which similarly allows us to eliminate inequalities over the second sort); we leave that part out, as it has no direct relevance for our concerns.

original vocabulary involving just E with only existential quantification over the first and only universal quantification over the second sort.

Note. The approach to interpolation described above was developed independently and indeed without knowledge of two quite distinct lines of investigation carried out in the 1970s, which however both turn out to be closely related to this approach: the strengthening of Feferman's result by J. Stern [12], and work by N. Motohashi [11]. Motohashi in [11] states without proof an interpolation theorem in a very similar formalism of relativised formulae, although not of the Lyndon type, i.e., not accounting for predicate polarities. A connection with Feferman's many-sorted interpolation theorem [5] is also discussed there. To the best of my knowledge, no proof of Motohashi's theorem can be found in the literature. Stern's result, obtained in the setting of model theoretic forcing, does indeed cover all the many-sorted consequences of our interpolation theorem, and with hindsight one might even find the proof patterns—model theoretic back-and-forth techniques vs. model theoretic forcing—intimately related. On the other hand, one advantage of putting the main theorem and its proof into the standard one-sorted framework seems to be that it highlights a surprisingly direct relationship between the traditionally rather orthogonal scales of existential-vs.-universal and of positive-vs.-negative. Moreover, this is achieved at a truly elementary level with a canonical model theoretic proof. I would hope, therefore, that this note, even if it comes as an afterthought to long established precursors, offers a new perspective which may be attractive for its simplicity and comprehensiveness.

REFERENCES

[1] J. Barwise, *A preservation theorem for interpretations*, ***Proceedings of the Cambridge summer school in mathematical logic 1971*** (A. Mathias et al., editors), Lecture Notes in Mathematics, vol. 337, Springer-Verlag, 1973, pp. 618–621.

[2] J. van Benthem, *Information transfer across Chu spaces*, ***Logic Journal of the IGPL***, vol. 8 (November 2000), pp. 719–731.

[3] C. C. Chang and H. J. Keisler, ***Model theory***, 3rd ed., North-Holland, 1990.

[4] W. Craig, *Three uses of the Herbrand-Gentzen theorem in relating model theory and proof theory*, ***The Journal of Symbolic Logic***, vol. 22 (1957), pp. 269–285.

[5] S. Feferman, *Lectures on proof theory*, ***Proceedings of the summer school in logic, Leeds 67*** (M. Löb, editor), Lecture Notes in Mathematics, vol. 70, Springer-Verlag, 1968, pp. 1–107.

[6] ———, *Applications of many-sorted interpolation theorems*, ***Proceedings of the Tarski symposium*** (L. Henkin et al., editors), AMS Proceedings of Symposia in Pure Mathematics, vol. XXV, American Mathematical Society, 1974, pp. 205–223.

[7] ———, *Ah, Chu!*, ***JFAK. Essays dedicated to Johan van Benthem on the occasion of his 50th birthday*** (J. Gerbrandy, M. Marx, M. de Rijke, and Y. Venema, editors), Amsterdam University Press, 1999, CD-ROM, see `http://turing.wins.uva.nl/ j50/cdrom/`.

[8] J. FLUM, *First-order logic and its extensions*, ***Proceedings of the international summer institute and logic colloquium, Kiel 1974*** (G. Müller et al., editors), Lecture Notes in Mathematics, vol. 499, Springer-Verlag, 1975, pp. 248–310.

[9] W. HODGES, ***Model theory***, Cambridge University Press, 1993.

[10] R. C. LYNDON, *An interpolation theorem in the predicate calculus*, ***Pacific Journal of Mathematics***, vol. 9 (1959), pp. 129–142.

[11] N. MOTOHASHI, *Two theorems on mix-relativization*, ***Proceedings of the Japan Academy***, vol. 49 (1973), no. 3, pp. 161–163.

[12] J. STERN, *A new look at the interpolation problem*, ***The Journal of Symbolic Logic***, vol. 40 (1975), no. 1, pp. 1–13.

MATHEMATISCHES GRUNDLAGEN DER INFORMATIK
RWTH AACHEN, AHORNSTRASSE 55, D-52074
AACHEN, GERMANY
E-mail: otto@informatik.rwth-aachen.de