Our second application of the Model Existence Theorem is Craig's Interpolation Theorem. This result has important model-theoretic consequences, and we will consider it more fully once first-order logic has been introduced.

Definition 3.6.4 A formula $Z$ is an *interpolant* for the implication $X \supset Y$ if every propositional letter of $Z$ also occurs in both $X$ and $Y$ and if $X \supset Z$ and $Z \supset Y$ are both tautologies.

For example, $(P \vee (Q \wedge R)) \supset (P \vee \neg\neg Q)$ has $P \vee Q$ as an interpolant; $(P \wedge \neg P) \supset Q$ has $\bot$ as an interpolant.

Theorem 3.6.5 **(Craig Interpolation)** *If $X \supset Y$ is a tautology, then it has an interpolant.*

**Proof** We write $\langle S \rangle$, as usual, to denote the conjunction of the members of $S$. Call a finite set $S$ *Craig consistent*, provided there is a partition of $S$ into subsets $S_1$ and $S_2$ (that is, $S = S_1 \cup S_2$ and $S_1 \cap S_2 = \emptyset$) such that $\langle S_1 \rangle \supset \neg\langle S_2 \rangle$ has no interpolant. Let $\mathcal{C}$ be the collection of all Craig-consistent sets. $\mathcal{C}$ is a Propositional Consistency Property (Exercise 3.6.5).

Now we show the theorem in its contrapositive form. Suppose $X \supset Y$ has no interpolant. Let $S$ be the set $\{X, \neg Y\}$, and consider the partition $S_1 = \{X\}$, $S_2 = \{\neg Y\}$. If $\langle\{X\}\rangle \supset \neg\langle\{\neg Y\}\rangle$ had an interpolant $Z$, then $Z$ would also be an interpolant for $X \supset Y$, hence it does not have an interpolant. Then $S$ is Craig consistent, and so by the Model Existence Theorem, $S$ is satisfiable. It follows that $X \supset Y$ is not a tautology. □

## Exercises

**3.6.1.** Show that every Propositional Consistency Property can be extended to one that is subset closed. Hint: Let $\mathcal{C}$ be a Propositional Consistency Property. Let $\mathcal{C}^+$ consist of all subsets of members of $\mathcal{C}$, and show $\mathcal{C}^+$ is also a Propositional Consistency Property.

**3.6.2.** Show that every Propositional Consistency Property of finite character is subset closed.

**3.6.3.** Show that a Propositional Consistency Property that is subset closed can be extended to one of finite character. Hint: Let $\mathcal{C}$ be a Propositional Consistency Property that is subset closed. Let $\mathcal{C}^+$ consist of those sets $S$ all of whose finite subsets are in $\mathcal{C}$. Show that $\mathcal{C}^+$ is a Propositional Consistency Property and extends $\mathcal{C}$.

**3.6.4.** Finish the proof of the Propositional Compactness Theorem by showing in detail that $\mathcal{C}$ is a Propositional Consistency Property.

**3.6.5.** Complete the proof of Theorem 3.6.5 by showing that the collection of Craig-consistent sets is a Propositional Consistency Property.

**3.6.6.** Show that if $X \supset Y$ is a tautology and $X$ and $Y$ have no propositional letters in common, then one of $\neg X$ or $Y$ is a tautology.

## 8.11 Craig's Interpolation Theorem

We sketched a proof of Craig's interpolation theorem for propositional logic in Chapter 3 (Theorem 3.6.5). The propositional theorem does not really have any interesting applications, but the first-order version most certainly does. In this section we extend the earlier proof, which is nonconstructive, to the first-order setting, and in the next section we give a constructive argument as well. Applications then follow. The proof in this section has superficial differences with that of Theorem 3.6.5, but the essential features are the same.

Definition 8.11.1 Let $S_1$ and $S_2$ be sets of sentences. An *interpolant* for the pair $S_1$, $S_2$ is a sentence $Z$ such that all constant, function, and relation symbols of $Z$ occur in formulas of both $S_1$ and $S_2$ and such that $S_1 \cup \{Z\}$ and $S_2 \cup \{\neg Z\}$ are not satisfiable.

Definition 8.11.2 A finite set $S$ of sentences is *Craig consistent* if there is a partition $S_1$, $S_2$ of $S$ that lacks an interpolant. ($S_1$, $S_2$ is a partition of $S$ if $S_1 \cup S_2 = S$ and $S_1 \cap S_2 = \emptyset$.)

The following lemma is in a form that is particularly suited for an application of the Model Existence Theorem. Craig's Theorem itself is an easy consequence.

Lemma 8.11.3 *Let $C$ be the collection of Craig-consistent sets. $C$ is a first-order consistency property.*

**Proof** Several items must be checked; we only do a few.

**$\gamma$-case** Suppose $\gamma \in S$ but for some closed term $t$, $S \cup \{\gamma(t)\}$ is not Craig consistent. We show $S$ is not Craig-consistent either. Let $S_1$, $S_2$ be a partition of $S$; we show it has an interpolant. And we assume $\gamma \in S_1$; the argument if $\gamma \in S_2$ is similar.

$S_1 \cup \{\gamma(t)\}$, $S_2$ is a partition of $S \cup \{\gamma(t)\}$, so it has an interpolant, say $Z$, since $S \cup \{\gamma(t)\}$ is not Craig-consistent. Now $S_2 \cup \{\neg Z\}$ is not satisfiable. Also, $S_1 \cup \{\gamma(t)\} \cup \{Z\}$ is not satisfiable, and it follows easily that $S_1 \cup \{Z\}$ is not satisfiable either, since $\gamma \in S_1$.

We know that all constant, function, and relation symbols of $Z$ are common to $S_1 \cup \{\gamma(t)\}$ and $S_2$. If they all occur in $S_1$, we are done; $Z$ is an interpolant for $S_1$, $S_2$. So now suppose $Z$ contains some symbol occurring in $S_1 \cup \{\gamma(t)\}$ but not in $S_1$. Since $\gamma \in S_1$, any such symbol must occur in $t$ and so must be a constant or a function symbol. There may be several; for simplicity let us say $Z$ contains just one subterm, $f(u_1, \ldots, u_n)$, where $f$ occurs in $t$ but not in $S_1$. The more general situation is treated similarly.

Let $x$ be a new free variable, and let $Z^*$ be like $Z$ but with the occurrence of $f(u_1, \ldots, u_n)$ replaced by $x$, so $Z = Z^*\{x/f(u_1, \ldots, u_n)\}$. We claim $(\exists x)Z^*$ is an interpolant for $S_1$, $S_2$.

First, all constant, function, and relation symbols of $(\exists x)Z^*$ are common to both $S_1$ and $S_2$, because we have removed the only one that was a problem. Next, $S_2 \cup \{\neg Z\}$ is not satisfiable, hence, neither is $S_2 \cup \{\neg(\exists x)Z^*\}$. This follows from the validity of

$$Z^*\{x/f(u_1, \ldots, u_n)\} \supset (\exists x)Z^*.$$

Finally, $S_1 \cup \{Z\}$ is not satisfiable, and it follows that $S_1 \cup \{(\exists x)Z^*\}$ is also not satisfiable. This argument needs a little more discussion than the others. (Its similarity to the proof of Lemma 8.3.1 is no coincidence.)

Suppose $S_1 \cup \{(\exists x)Z^*\}$ is satisfiable, we show $S_1 \cup \{Z\}$ also is. Suppose the members of $S_1 \cup \{(\exists x)Z^*\}$ are true in the model $\langle \mathbf{D}, \mathbf{I} \rangle$. Then in particular, $Z^{*\mathbf{I},\mathbf{A}}$ is true for some assignment $\mathbf{A}$. Now define a new interpretation $\mathbf{J}$ to be like $\mathbf{I}$ on all symbols except $f$, and set $f^{\mathbf{J}}$ to be the same as $f^{\mathbf{I}}$ on all members of $\mathbf{D}$ except $u_1^{\mathbf{I},\mathbf{A}}, \ldots, u_n^{\mathbf{I},\mathbf{A}}$. Finally, set $f^{\mathbf{J}}(u_1^{\mathbf{I},\mathbf{A}}, \ldots, u_n^{\mathbf{I},\mathbf{A}}) = x^{\mathbf{I}}$. Since $\mathbf{I}$ and $\mathbf{J}$ differ only on $f$, and that does not occur in $S_1$, the members of this set will have the same truth values using either interpretation. Consequently, the members of $S_1$ are true in $\langle \mathbf{D}, \mathbf{J} \rangle$. Using Proposition 5.3.7, $[Z^*\{x/f(u_1, \ldots, u_n)\}]^{\mathbf{J},\mathbf{A}} = Z^{*\mathbf{J},\mathbf{A}} = Z^{*\mathbf{I},\mathbf{A}} = \mathbf{t}$.

**$\delta$-case** This time suppose $\delta \in S$ but for each parameter $p$, $S \cup \{\delta(p)\}$ is not Craig-consistent. We show $S$ is also not Craig-consistent. Let $S_1$, $S_2$ be a partition of $S$; we show it has an interpolant. And once again, we suppose $\delta \in S_1$; if it is in $S_2$, the argument is similar.

Let $p$ be a parameter that does not occur in $S$ (there must be one, since $S$ is finite). $S \cup \{\delta(p)\}$ is not Craig-consistent, so its partition $S_1 \cup \{\delta(p)\}$, $S_2$ has an interpolant, say, $Z$. We claim $Z$ is also an interpolant for $S_1$, $S_2$.

All constant, function, and relation symbols of $Z$ are common to both $S_1 \cup \{\delta(p)\}$ and $S_2$. Since $p$ was new to $S$, it does not occur in $S_2$, and hence not in $Z$ either. It follows that all constant, function, and relation symbols of $Z$ are common to $S_1$ and $S_2$.

$S_2 \cup \{\neg Z\}$ is not satisfiable. Also $S_1 \cup \{\delta(p)\} \cup \{Z\}$ is not satisfiable, and it follows that $S_1 \cup \{Z\}$ is not satisfiable either. This is shown by a 'redefining interpretations' argument much like in the $\gamma$-case. We omit it. But the conclusion is that $Z$ is an interpolant for $S_1$, $S_2$.

□

Now we come to Craig's Theorem itself [12].

Definition 8.11.4 The sentence $Z$ is an *interpolant* for the sentence $X \supset Y$ if every relation symbol, function symbol, and constant symbol of $Z$ is common to $X$ and $Y$, and both $X \supset Z$ and $Z \supset Y$ are valid.

Theorem 8.11.5 **(First-Order Craig Interpolation)** *If $X \supset Y$ is a valid sentence, then it has an interpolant.*

**Proof** We show the contrapositive. Suppose $X \supset Y$ lacks an interpolant. Let $S = \{X, \neg Y\}$, and consider the partition $S_1 = \{\neg Y\}$ and $S_2 = \{X\}$. If $Z$ were an interpolant for $S_1$, $S_2$, it follows directly that $Z$ would also be an interpolant for $X \supset Y$. Thus, $S_1$, $S_2$ has no interpolant, so $S$ is Craig-consistent. Then by the Model Existence Theorem, $S$ is satisfiable, and so $X \supset Y$ is not valid. □

## Exercises

**8.11.1.** We said in the proof of Lemma 8.11.3, in doing the $\gamma$-case, that the subcase where $\gamma \in S_2$ was similar to that where $\gamma \in S_1$. Nonetheless, there are some notable differences. Do this subcase in detail.

## 8.12 Craig's Interpolation Theorem—Constructively

The proof of Craig's Theorem in the previous section used the Model Existence Theorem and was nonconstructive. Now we give a constructive proof, showing how to extract an interpolant from a closed tableau. (The previous sentence mentions an important point—one that is easy to miss. An interpolant for a valid sentence $X \supset Y$ is not constructed just from the sentences $X$ and $Y$. It is constructed from a *proof* of $X \supset Y$—different proofs can give different interpolants.) We use a modified version of the symmetric Gentzen system method introduced by Smullyan [48].

In proving $X \supset Y$, we start a tableau with $\neg(X \supset Y)$, then apply the $\alpha$-rule, adding $X$ and $\neg Y$. After this, any sentence added to the tableau must be either a descendant of $X$ or of $\neg Y$. We begin by enhancing the usual tableau machinery to keep track of the ancestor of each sentence. Think of $X$ as "left" and $\neg Y$ as "right," which are respective positions of $X$ and $Y$ in $X \supset Y$. We symbolize this by writing $L(X)$ and $R(\neg Y)$ and systematically use the "$L$" and "$R$" notation throughout. In effect, "$L$" and "$R$" are bookkeeping devices that record a sentence's ancestry; they play no other role.

Definition 8.12.1 A *biased sentence* is an expression of one of the forms $L(Z)$ or $R(Z)$, where $Z$ is a sentence.

Next, the usual tableau rules are extended to biased sentences in a straightforward way. For instance, the standard $\alpha$-rule yields the following two biased rules:

$$\begin{array}{c} L(\alpha) \\ \hline L(\alpha_1) \\ L(\alpha_2) \end{array} \qquad\qquad \begin{array}{c} R(\alpha) \\ \hline R(\alpha_1) \\ R(\alpha_2) \end{array}$$

The other tableau rules are treated in a similar way. We call a tableau that is constructed using these rules a *biased tableau*. A branch of a biased tableau is closed if it contains a syntactic contradiction, ignoring the $L$ and $R$ symbols. Thus, a branch is closed if it contains $L(Z)$ and $R(\neg Z)$ or if it contains $L(Z)$ and $L(\neg Z)$, and so on.

If the sentence $X \supset Y$ has a tableau proof, it can be converted into a closed biased tableau for $\{L(X), R(\neg Y)\}$. Simply take the closed tableau beginning with $\neg(X \supset Y)$, drop the first line, thus getting a tableau beginning with $X$ and $\neg Y$; replace $X$ by $L(X)$ and $\neg Y$ with $R(\neg Y)$, then continue the insertion of $L$ and $R$ symbols downward through the tableau, in the obvious way. We extract an interpolant from this closed biased tableau.

Essentially, the idea is this. We begin with each closed branch, assign an interpolant (to be defined shortly) to it, then, one by one, we undo each tableau rule application, calculating interpolants for the resulting shortened branches from those for the original longer ones. Thus, for instance, suppose the last rule applied on a branch is one of the biased $\alpha$-rules—say the set of biased sentences on the branch is $S \cup \{L(\alpha), L(\alpha_1), L(\alpha_2)\}$, and we have an interpolant for this set. Using it, we say how to calculate an interpolant for the smaller set $S \cup \{L(\alpha)\}$, corresponding to the branch before the $\alpha$-rule was applied. Continuing in this way, we work our way back to the beginning, thus producing an interpolant for the set $\{L(X), R(\neg Y)\}$, and we will see this is also an interpolant for the sentence $X \supset Y$. Now to define the terminology precisely.

Definition 8.12.2 We say the sentence $Z$ is an interpolant for the finite set $\{L(A_1),\ldots, L(A_n), R(B_1),\ldots, R(B_k)\}$, provided $Z$ is an interpolant, in the sense of Definition 8.11.4, for the sentence $(A_1 \wedge \ldots \wedge A_n) \supset (\neg B_1 \vee \ldots \vee \neg B_k)$. (Take the empty conjunction to be $\top$ and the empty disjunction to be $\bot$.)

We use the notation $S \xrightarrow{int} Z$ to symbolize that $Z$ is an interpolant for the finite set $S$ of biased sentences.

Note that by this definition, an interpolant for the set $\{L(X), R(\neg Y)\}$ will be an interpolant for the sentence $X \supset \neg\neg Y$, and hence for $X \supset Y$.

Now we give calculation rules, starting with those for closed branches.

$$\begin{aligned}
S \cup \{L(A), L(\neg A)\} &\xrightarrow{int} \bot\\
S \cup \{R(A), R(\neg A)\} &\xrightarrow{int} \top\\
S \cup \{L(A), R(\neg A)\} &\xrightarrow{int} A\\
S \cup \{R(A), L(\neg A)\} &\xrightarrow{int} \neg A\\
S \cup \{L(\bot)\} &\xrightarrow{int} \bot\\
S \cup \{R(\bot)\} &\xrightarrow{int} \top
\end{aligned}$$

Next the easy propositional cases.

$$\frac{S \cup \{L(\top)\} \xrightarrow{int} A}{S \cup \{L(\neg\bot)\} \xrightarrow{int} A} \qquad \frac{S \cup \{L(\bot)\} \xrightarrow{int} A}{S \cup \{L(\neg\top)\} \xrightarrow{int} A}$$

$$\frac{S \cup \{R(\top)\} \xrightarrow{int} A}{S \cup \{R(\neg\bot)\} \xrightarrow{int} A} \qquad \frac{S \cup \{R(\bot)\} \xrightarrow{int} A}{S \cup \{R(\neg\top)\} \xrightarrow{int} A}$$

$$\frac{S \cup \{L(Z)\} \xrightarrow{int} A}{S \cup \{L(\neg\neg Z)\} \xrightarrow{int} A} \qquad \frac{S \cup \{R(Z)\} \xrightarrow{int} A}{S \cup \{R(\neg\neg Z)\} \xrightarrow{int} A}$$

The $\alpha$-cases are also straightforward.

$$\frac{S \cup \{L(\alpha_1), L(\alpha_2)\} \xrightarrow{int} A}{S \cup \{L(\alpha)\} \xrightarrow{int} A} \qquad \frac{S \cup \{R(\alpha_1), R(\alpha_2)\} \xrightarrow{int} A}{S \cup \{R(\alpha)\} \xrightarrow{int} A}$$

Finally the $\beta$-cases, which are the most interesting of the propositional rules, follow:

$$\frac{S \cup \{L(\beta_1)\} \xrightarrow{int} A \quad S \cup \{L(\beta_2)\} \xrightarrow{int} B}{S \cup \{L(\beta)\} \xrightarrow{int} A \vee B}$$

$$\frac{S \cup \{R(\beta_1)\} \xrightarrow{int} A \quad S \cup \{R(\beta_2)\} \xrightarrow{int} B}{S \cup \{R(\beta)\} \xrightarrow{int} A \wedge B}$$

This completes the set of propositional rules. Before moving to those for quantifiers, we verify one of the rules and give an example. The rule we verify is the one for $R(\beta)$.

**Verification** Suppose $S = \{L(X_1), \ldots, L(X_n), R(Y_1), \ldots, R(Y_k)\}$, and we have both

$$S \cup \{R(\beta_1)\} \xrightarrow{int} A \quad \text{and} \quad S \cup \{R(\beta_2)\} \xrightarrow{int} B.$$

Then $A$ is an interpolant for the sentence $(X_1 \wedge \ldots \wedge X_n) \supset (\neg Y_1 \vee \ldots \vee \neg Y_k \vee \neg\beta_1)$, so all the relation, function, and constant symbols of $A$ appear in both $X_1 \wedge \ldots \wedge X_n$ and $\neg Y_1 \vee \ldots \vee \neg Y_k \vee \neg\beta_1$. It follows that they also appear in $\neg Y_1 \vee \ldots \vee \neg Y_k \vee \neg\beta$, since every relation, constant, and function symbol of $\beta_1$ appears in $\beta$. A similar observation applies to $B$. It follows that all relation, constant, and function symbols of $A \wedge B$ are common to both $X_1 \wedge \ldots \wedge X_n$ and $\neg Y_1 \vee \ldots \vee \neg Y_k \vee \neg\beta$.

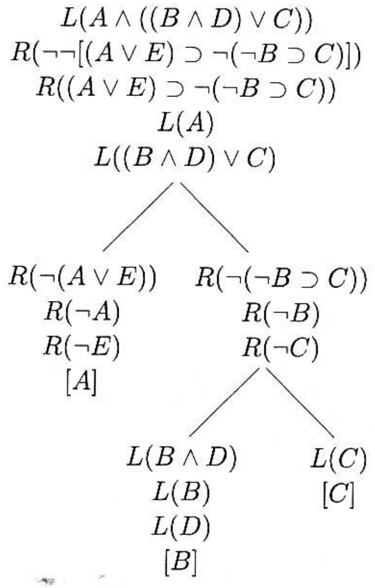
Next, $A \supset (\neg Y_1 \vee \ldots \vee \neg Y_k \vee \neg\beta_1)$ and $B \supset (\neg Y_1 \vee \ldots \vee \neg Y_k \vee \neg\beta_2)$ are both valid. Then we have the following:

$$\begin{aligned}
A \wedge B \quad &\supset \quad (\neg Y_1 \vee \ldots \vee \neg Y_k \vee \neg\beta_1) \wedge (\neg Y_1 \vee \ldots \vee \neg Y_k \vee \neg\beta_2)\\
&\equiv \quad (\neg Y_1 \vee \ldots \vee \neg Y_k \vee (\neg\beta_1 \wedge \neg\beta_2))\\
&\equiv \quad (\neg Y_1 \vee \ldots \vee \neg Y_k \vee \neg(\beta_1 \vee \beta_2))\\
&\equiv \quad (\neg Y_1 \vee \ldots \vee \neg Y_k \vee \neg\beta)
\end{aligned}$$

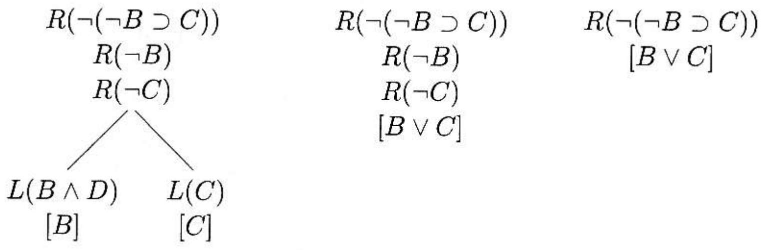
In a similar (and simpler) way, we have the validity of $(X_1 \wedge \ldots \wedge X_n) \supset (A \wedge B)$. Consequently,

$$S \cup \{R(\beta)\} \xrightarrow{int} A \wedge B.$$

Example We compute an interpolant for the tautology $[A \wedge ((B \wedge D) \vee C)] \supset \neg[(A \vee E) \supset \neg(\neg B \supset C)]$. First, here is a closed biased tableau for $\{L(A \wedge ((B \wedge D) \vee C)), R(\neg\neg[(A \vee E) \supset \neg(\neg B \supset C)])\}$. At the end of each branch, we give in square brackets an interpolant for the set of biased sentences on that branch, computed using the rules just given.


$L(A \wedge ((B \wedge D) \vee C))$
$R(\neg\neg[(A \vee E) \supset \neg(\neg B \supset C)])$
$R((A \vee E) \supset \neg(\neg B \supset C))$
$L(A)$
$L((B \wedge D) \vee C)$
$R(\neg(A \vee E))$
$R(\neg A)$
$R(\neg E)$
$[A]$
$R(\neg(\neg B \supset C))$
$R(\neg B)$
$R(\neg C)$
$L(B \wedge D)$
$L(B)$
$L(D)$
$[B]$
$L(C)$
$[C]$

Now we progressively undo tableau rule applications. For reasons of space, we concentrate on the subtree beginning with the biased sentence $R(\neg(\neg B \supset C))$, displaying only it. Progressively, it becomes:



We leave it to you to continue the process fully. When complete, $A \wedge (B \vee C)$ is the computed interpolant.

Next we give the first-order rules, which are the most complicated. We assume the language has no function symbols—a treatment of these can be added, but it obscures the essential ideas. Once again, we suppose $S = \{L(X_1), \ldots, L(X_n), R(Y_1), \ldots, R(Y_k)\}$. The first two rules are simple. Let $p$ be a parameter that does not occur in $S$ or in $\delta$.

$$\frac{S \cup \{L(\delta(p))\} \xrightarrow{int} A}{S \cup \{L(\delta)\} \xrightarrow{int} A} \qquad \frac{S \cup \{R(\delta(p))\} \xrightarrow{int} A}{S \cup \{R(\delta)\} \xrightarrow{int} A}$$

This leaves the $\gamma$-cases, each of which splits in two, giving four rules. In the following, $c$ is some constant symbol (the only kind of closed term we have now), and $A\{c/x\}$ is the result of replacing all occurrences of $c$ in $A$ with occurrences of the variable $x$. We assume $x$ is a "new" variable, one that does not appear in $S$ or in $\gamma$.

$$\begin{array}{c} \dfrac{S \cup \{L(\gamma(c))\} \xrightarrow{int} A}{S \cup \{L(\gamma)\} \xrightarrow{int} A} \\ \text{if } c \text{ occurs in } \{X_1, \ldots, X_n\} \end{array} \qquad \begin{array}{c} \dfrac{S \cup \{R(\gamma(c))\} \xrightarrow{int} A}{S \cup \{R(\gamma)\} \xrightarrow{int} A} \\ \text{if } c \text{ occurs in } \{Y_1, \ldots, Y_k\} \end{array}$$

$$\begin{array}{c} \dfrac{S \cup \{L(\gamma(c))\} \xrightarrow{int} A}{S \cup \{L(\gamma)\} \xrightarrow{int} (\forall x)A\{c/x\}} \\ \text{otherwise} \end{array} \qquad \begin{array}{c} \dfrac{S \cup \{R(\gamma(c))\} \xrightarrow{int} A}{S \cup \{R(\gamma)\} \xrightarrow{int} (\exists x)A\{c/x\}} \\ \text{otherwise} \end{array}$$

Once again we verify the correctness of one of these rules, that for $R(\gamma)$, leaving the other to you.

**Verification** Suppose $S \cup \{R(\gamma(c))\} \xrightarrow{int} A$. Then both $(X_1 \wedge \ldots \wedge X_n) \supset A$ and $A \supset (\neg Y_1 \vee \ldots \vee \neg Y_k \vee \neg\gamma(c))$ are valid, and all constant and relation symbols of $A$ are common to $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_k, \gamma(c)\}$.

First, assume we are in the case where $c$ occurs in one of the sentences of $\{Y_1, \ldots, Y_k\}$. Since $\gamma \supset \gamma(c)$ is valid, so is $\neg\gamma(c) \supset \neg\gamma$, and it follows that $A \supset (\neg Y_1 \vee \ldots \vee \neg Y_k \vee \neg\gamma)$ is valid. Also, all relation and constant symbols of $A$ occur in both $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_k, \gamma\}$, since the only one that might have been a problem was $c$, which appeared in $\gamma(c)$ and does not appear in $\gamma$. This causes no difficulties, however, since we are assuming that $c$ is in one of $\{Y_1, \ldots, Y_k\}$. Thus, $A$ is still an interpolant.

Next, assume $c$ does not occur in $\{Y_1, \ldots, Y_k\}$. We have that $(X_1 \wedge \ldots \wedge X_n) \supset A$ is valid. Also $A \supset (\exists x)A\{c/x\}$ is valid, consequently,

$(X_1 \wedge \ldots \wedge X_n) \supset (\exists x)A\{c/x\}$ is valid. Further, all constant and relation symbols of $(\exists x)A\{c/x\}$ also appear in $A$ and hence in $\{X_1, \ldots, X_n\}$.

On the right hand side, $A \supset (\neg Y_1 \vee \ldots \vee Y_k \vee \neg\gamma(c))$ is valid, hence so is $(Y_1 \wedge \ldots \wedge Y_k \wedge \gamma(c)) \supset \neg A$. For a new variable $x$, the validity of $(\forall x)[Y_1 \wedge \ldots \wedge Y_k \wedge \gamma(c)]\{c/x\} \supset (\forall x)\neg A\{c/x\}$ follows. But $c$ does not occur in $\{Y_1, \ldots, Y_k\}$, so $(\forall x)[Y_1 \wedge \ldots \wedge Y_k \wedge \gamma(c)]\{c/x\} \equiv [Y_1 \wedge \ldots \wedge Y_k \wedge (\forall x)\gamma(c)(c/x)] \equiv [Y_1 \wedge \ldots \wedge Y_k \wedge \gamma]$. Thus, we have the validity of $\neg(\forall x)\neg A\{c/x\} \supset (\neg Y_1 \vee \ldots \vee \neg Y_k \vee \gamma)$, or $(\exists x)A\{c/x\} \supset (\neg Y_1 \vee \ldots \vee \neg Y_k \vee \gamma)$.

Finally, the constant and relation symbols of $(\exists x)A\{c/x\}$ are those of $A$ except for $c$. (In fact, if $c$ does not actually appear in $A$, the quantification is vacuous, and $(\exists x)A\{c/x\}$ and $A$ have the same constant and relation symbols.) The constant and relation symbols of $\{X_1, \ldots, X_n\}$ include those of $A$, hence trivially those of $(\exists x)A\{c/x\}$. The constant and relation symbols of $\{Y_1, \ldots, Y_k, \gamma(c)\}$ include those of $A$, so the constant and relation symbols of $\{Y_1, \ldots, Y_k, \gamma\}$ also include those of $(\exists x)A\{c/x\}$.

We have thus verified that $(\exists x)A\{c/x\}$ is an interpolant.

Craig's Theorem was strengthened by Lyndon [32]. Recall from Definition 8.2.3 the notion of *positive* and *negative* formula occurrence. Say a relation symbol $R$ *occurs positively* in a formula $X$ if it appears in a positive atomic subformula of $X$—similarly for negative occurrences. Note that a relation symbol may appear both positively and negatively in the same formula.

Theorem 8.12.3 **(Lyndon Interpolation)** *If $X \supset Y$ is a valid sentence, then it has an interpolant $Z$ such that every relation symbol occurring positively in $Z$ has a positive occurrence in both $X$ and $Y$, and every relation symbol occurring negatively in $Z$ has a negative occurrence in both $X$ and $Y$.*

Lyndon's result cannot be extended to account for positive and negative occurrences of constant or function symbols. Consider the valid sentence $[(\forall x)(P(x) \supset \neg Q(x)) \wedge P(c)] \supset \neg Q(c)$. The constant symbol $c$ will occur in any interpolant for this sentence, but it occurs positively on the left and negatively on the right.

In effect, Lyndon's interpolation theorem has already been proved. That is, the two proofs we gave actually verify the stronger version. We leave it to you to check this, as an exercise.

Here is a different strengthening of Craig's theorem. As usual, we need some terminology first.

Definition 8.12.4 A formula is *universal* if every quantifier occurrence in it is essentially universal, in the sense of Definition 8.5.1. Likewise, a formula is *existential* if every quantifier occurrence in it is essentially existential.

Theorem 8.12.5 *Suppose $X \supset Y$ is a valid sentence. If $Y$ is universal, there is an interpolant that is also universal. Similarly, if $X$ is existential, there is an existential interpolant. Finally, if $X$ is existential and $Y$ is universal, there is a quantifier-free interpolant.*

## Exercises

**8.12.1.** Use the procedure of this section to compute an interpolant for the valid sentence $[(\forall x)(P(x) \supset \neg Q(x)) \wedge P(c)] \supset \neg Q(c)$.

**8.12.2.** Show that the procedure of this section actually verifies Lyndon's strengthening of the Craig Interpolation Theorem.

**8.12.3.** Prove Theorem 8.12.5 using the procedures of this section.

**8.12.4$^{\mathrm{P}}$.** Implement the propositional part of the procedure of this section in Prolog, producing a propositional theorem prover for implications that computes interpolants.

**8.12.5$^{\mathrm{P}}$.** Extend the Prolog implementation of the previous exercise to a full first-order version.

## 8.13 Beth's Definability Theorem

Sometimes information is explicit: "The murderer is John Smith." Often we find an implicit characterization instead, as in "The murderer is the only person in town who wears glasses, has red hair, and owns a dog and a canary." Puzzles often involve turning implicit characterizations into explicit ones. Beth's Definability Theorem [4] essentially says that in classical logic such puzzles can always be solved. This is a fundamental result that says that classical logic has a kind of completeness where definability is concerned. We begin this section with some terminology, then we state and prove Beth's Theorem. The proof we give is not Beth's original one but is based on Craig's Theorem and comes from Craig's 1957 paper [12].

Definition 8.13.1 Let $R$ be an $n$-place relation symbol, and $\Phi(x_1, \ldots, x_n)$ be a formula with free variables among $x_1, \ldots, x_n$, and with no occurrences of $R$. We say $\Phi$ is an *explicit definition* of $R$, with respect to a set $S$ of sentences, provided

$$S \models_f (\forall x_1) \cdots (\forall x_n)[R(x_1, \ldots, x_n) \equiv \Phi(x_1, \ldots, x_n)].$$

Example In a group, $y$ is the *conjugate* of $x$ under conjugation by $a$ if $y = a^{-1}xa$. Conjugation is a three-place relation, between $a$, $x$, and $y$ that has an explicit definition (we just gave it informally) with respect to the set $S$ of axioms for a group. Of course, to properly present this as an example, we need a first-order language *with equality*. Equality will be investigated in the next chapter, and it can be shown that the results of this section do carry over.

Definition 8.13.2 Again, let $R$ be an $n$-place relation symbol. We say $R$ is *implicitly defined* by a set $S$ of sentences, provided $S$ determines $R$ uniquely, in the following sense. Let $R^*$ be an $n$-place relation symbol different from $R$, that does not occur in $S$, and let $S^*$ be like $S$ except that every occurrence of $R$ has been replaced by an occurrence of $R^*$. Then $S$ determines $R$ uniquely if

$$S \cup S^* \models_f (\forall x_1)\cdots(\forall x_n)[R(x_1,\ldots,x_n) \equiv R^*(x_1,\ldots,x_n)].$$

Example Let $S = \{(\forall x)(R(x) \supset A(x)), (\forall x)(R(x) \supset B(x)), (\forall x)(A(x) \supset (B(x) \supset R(x)))\}$. It is easy to check that $S$ determines $R$ uniquely, and so $R$ is implicitly defined by $S$. In fact, $R$ also has the explicit definition $(\forall x)[R(x) \equiv (A(x) \wedge B(x))]$.

Theorem 8.13.3 **(Beth Definability)** *If $R$ is implicitly defined by a set $S$, then $R$ has an explicit definition with respect to $S$.*

**Proof** Suppose $R$ is implicitly defined by $S$. Let $R^*$ be a new relation symbol, and let $S^*$ be like $S$ but with occurrences of $R$ replaced by occurrences of $R^*$. Then

$$S \cup S^* \models_f (\forall x_1)\cdots(\forall x_n)[R(x_1,\ldots,x_n) \equiv R^*(x_1,\ldots,x_n)],$$

so by Theorem 5.10.2,

$$S_0 \cup S_0^* \models_f (\forall x_1)\cdots(\forall x_n)[R(x_1,\ldots,x_n) \equiv R^*(x_1,\ldots,x_n)],$$

where $S_0$ and $S_0^*$ are finite subsets of $S$ and $S^*$, respectively. Let $\bigwedge S_0$ be the conjunction of the members of $S_0$, and let $\bigwedge S_0^*$ be the conjunction of the members of $S_0^*$, so that both $\bigwedge S_0$ and $\bigwedge S_0^*$ are sentences. Then (Exercise 5.10.3, part 7),

$$(\bigwedge S_0 \wedge \bigwedge S_0^*) \supset (\forall x_1)\cdots(\forall x_n)[R(x_1,\ldots,x_n) \equiv R^*(x_1,\ldots,x_n)]$$

is valid. Choose $n$ distinct parameters, $p_1,\ldots, p_n$. Then the following is also valid:

$$(\bigwedge S_0 \wedge \bigwedge S_0^*) \supset [R(p_1,\ldots,p_n) \equiv R^*(p_1,\ldots,p_n)],$$

and from this follows the validity of

$$[\bigwedge S_0 \wedge R(p_1,\ldots,p_n)] \supset [\bigwedge S_0^* \supset R^*(p_1,\ldots,p_n)].$$

Now by Craig's Theorem 8.11.5, there is an interpolant for this. The interpolant may contain some or all of the parameters we introduced, so we denote it by $\Phi(p_1,\ldots,p_n)$. Since it is an interpolant, all constant, function, and relation symbols of $\Phi(p_1,\ldots,p_n)$ are common to $[\bigwedge S_0 \wedge R(p_1,\ldots,p_n)]$ and $[\bigwedge S_0^* \supset R^*(p_1,\ldots,p_n)]$, and both

$$[\bigwedge S_0 \wedge R(p_1,\ldots,p_n)] \supset \Phi(p_1,\ldots,p_n)$$

and

$$\Phi(p_1,\ldots,p_n) \supset [\bigwedge S_0^* \supset R^*(p_1,\ldots,p_n)]$$

are valid.

The relation symbol $R^*$ does not occur in $[\bigwedge S_0 \wedge R(p_1,\ldots,p_n)]$, and the relation symbol $R$ does not occur in $[\bigwedge S_0^* \supset R^*(p_1,\ldots,p_n)]$, and so neither $R$ nor $R^*$ can occur in $\Phi(p_1,\ldots,p_n)$. Also, $\Phi(p_1,\ldots,p_n) \supset [\bigwedge S_0^* \supset R^*(p_1,\ldots,p_n)]$ is valid, hence so is the result of replacing all occurrences of $R^*$ by a relation symbol that does not appear in this sentence. Since $R^*$ does not occur in $\Phi(p_1,\ldots,p_n)$, replacing occurrences of $R^*$ with occurrences of $R$ yields the validity of $\Phi(p_1,\ldots,p_n) \supset [\bigwedge S_0 \supset R(p_1,\ldots,p_n)]$, from which the validity of $\bigwedge S_0 \supset [\Phi(p_1,\ldots,p_n) \supset R(p_1,\ldots,p_n)]$ follows. We also have the validity of $[\bigwedge S_0 \wedge R(p_1,\ldots,p_n)] \supset \Phi(p_1,\ldots,p_n)$ from which the validity of $\bigwedge S_0 \supset [R(p_1,\ldots,p_n) \supset \Phi(p_1,\ldots,p_n)]$ follows. Combining these two, we have the validity of

$$\bigwedge S_0 \supset [R(p_1,\ldots,p_n) \equiv \Phi(p_1,\ldots,p_n)],$$

from which we immediately obtain

$$S_0 \models_f [R(p_1,\ldots,p_n) \equiv \Phi(p_1,\ldots,p_n)]$$

and hence

$$S \models_f [R(p_1,\ldots,p_n) \equiv \Phi(p_1,\ldots,p_n)].$$

Finally, from this (Exercise 8.13.1) we have the validity of

$$S \models_f (\forall x_1)\cdots(\forall x_n)[R(x_1,\ldots,x_n) \equiv \Phi(x_1,\ldots,x_n)].$$

Since $\Phi(x_1,\ldots,x_n)$ contains no occurrences of $R$, it is an explicit definition of $R$. □

The proof of Beth's Theorem given above is entirely constructive, provided we have a constructive method of finding interpolants. Our first proof of Craig's Theorem was not constructive, but our second was. Also Craig's original proof [12] was constructive by design. It follows that the conversion from implicit to explicit definition is a constructive one, suitable for mechanization, at least in principle.

## Exercises

**8.13.1.** Suppose $\Phi$ is a formula with only $x$ free, and $p$ is a parameter that does not occur in $\Phi$, or in any member of the set $S$ of sentences. Show that if $S \models_f \Phi\{x/p\}$ then $S \models_f (\forall x)\Phi$. Hint: see Exercises 5.3.2 and 5.10.2.

## 8.14 Lyndon's Homomorphism Theorem

Let $L$ be a first-order language with no constant or function symbols, and let $\mathbf{M} = \langle \mathbf{D}, \mathbf{I} \rangle$ and $\mathbf{N} = \langle \mathbf{E}, \mathbf{J} \rangle$ be two models for the language $L$. A mapping $h : \mathbf{D} \to \mathbf{E}$ is a *homomorphism*, provided the following: For each $n$-place relation symbol $R$ of $L$ and for each $b_1, \ldots, b_n \in \mathbf{D}$, if $R^{\mathbf{I}}(b_1, \ldots, b_n)$ is true in $\mathbf{M}$ then $R^{\mathbf{J}}(h(b_1), \ldots, h(b_n))$ is true in $\mathbf{N}$.

Example Suppose $L$ has a single three-place relation symbol, $R$. Let $\mathbf{M}$ be the model whose domain is the integers and in which $R$ is interpreted by the addition relation; that is, $R^{\mathbf{I}}(a, b, c)$ is true in $\mathbf{M}$ just in case $a + b = c$. Let $\mathbf{N}$ have domain $\{0, 1\}$, and interpret $R$ by the addition modulo 2 relation; $R^{\mathbf{J}}(a, b, c)$ is true in $\mathbf{N}$ if $a+b = c \pmod 2$. If $h$ is the function mapping the evens to 0 and the odds to 1, $h$ is a homomorphism from $\mathbf{M}$ to $\mathbf{N}$.

Definition 8.14.1 A sentence $X$ of $L$ is *preserved under homomorphisms*, provided, whenever $X$ is true in a model $\mathbf{M}$, and there is a homomorphism from $\mathbf{M}$ to $\mathbf{N}$, then $X$ is also true in $\mathbf{N}$.

Being preserved under homomorphisms is a semantic notion—the definition talks about models and truth. The question for this section is, Does this semantic notion have a syntactic counterpart? We will see that it does—being a *positive* sentence.

Definition 8.14.2 Let $L$ be a first-order language (with or without constant and function symbols). The *positive* formulas of $L$ are the members of the smallest set $S$ such that:

1. If $A$ is atomic, $A \in S$.
2. $Z \in S \Longrightarrow \neg\neg Z \in S$.
3. $\alpha_1 \in S$ and $\alpha_2 \in S \Longrightarrow \alpha \in S$.
4. $\beta_1 \in S$ and $\beta_2 \in S \Longrightarrow \beta \in S$.
5. $\gamma(t) \in S \Longrightarrow \gamma \in S$.
6. $\delta(t) \in S \Longrightarrow \delta \in S$.

Informally, the positive formulas are those that can be rewritten using only conjunctions and disjunctions as propositional connectives (in particular without negation symbols). You are asked to show this as Exercise 8.14.1. Equivalently, the positive formulas are those in which every relation symbol appears only positively.

Now, here is the easy half of the semantic/syntactic connection we are after; we leave its proof to you. (It generalizes considerably—see Exercise 9.2.3.)

Proposition 8.14.3 *Let $L$ be a language with no constant or function symbols. Every positive sentence of $L$ is preserved under homomorphisms.*

The hard half is Lyndon's Homomorphism Theorem. It asserts the converse.

Theorem 8.14.4 **(Lyndon Homomorphism)** *Let $L$ be a language with no constant or function symbols. If the sentence $X$ is preserved under all homomorphisms, then $X$ is equivalent to some positive sentence.*

**Proof** Without loss of generality, we assume $L$ has a finite number of relation symbols, since it is only necessary to consider those that actually appear in $X$. For each ($n$-place) relation symbol $R$ of $L$, let $R'$ be another relation symbol (also $n$-place) that is new to the language. Let $L'$ be the result of enlarging $L$ with these additional relation symbols. By the *homomorphism sentence* for $R$, we mean the sentence

$$(\forall x_1) \cdots (\forall x_n)[R(x_1, \ldots, x_n) \supset R'(x_1, \ldots, x_n)].$$

Now, let $H$ be the conjunction of all the homomorphism sentences for the relation symbols of $L$. Also, let $X'$ be the sentence that is like $X$, except that each relation symbol $R$ has been replaced by its counterpart $R'$. (More generally, for any sentence $Z$, let $Z'$ be the result of replacing each $R$ by the corresponding $R'$.)

Now assume that $X$ is preserved under homomorphisms. We claim the following sentence is valid:

$$X \supset (H \supset X').$$

Suppose this is not so. Let $\mathbf{M}_0 = \langle \mathbf{D}_0, \mathbf{I}_0 \rangle$ be a model for $L'$ in which it is not true, hence in which $X$ and $H$ are true, but $X'$ is false. We construct two new models $\mathbf{M}_1 = \langle \mathbf{D}_1, \mathbf{I}_1 \rangle$ and $\mathbf{M}_2 = \langle \mathbf{D}_2, \mathbf{I}_2 \rangle$, for the original language $L$, by "pulling apart" this one. First, set $\mathbf{D}_1 = \mathbf{D}_2 = \mathbf{D}_0$, so all models have the same domain. Next, for each relation symbol $R$ of $L$, set $R^{\mathbf{I}_1} = R^{\mathbf{I}_0}$ and $R^{\mathbf{I}_2} = R'^{\mathbf{I}_0}$.

In effect, $\mathbf{M}_1$ acts like the "unprimed" part of $\mathbf{M}$, and $\mathbf{M}_2$ acts like the "primed" part. This can be made more precise, as follows (we leave the proof as an exercise):

**Pulling Apart Assertion** For any sentence $Z$ of $L$:

1. $Z$ is true in $\mathbf{M}_0$ if and only if $Z$ is true in $\mathbf{M}_1$.
2. $Z$ is true in $\mathbf{M}_0$ if and only if $Z'$ is true in $\mathbf{M}_2$.

Now, let $a_1, \ldots, a_n$ be in $\mathbf{D}_0$, and suppose $R^{\mathbf{I}_1}(a_1, \ldots, a_n)$ is true in $\mathbf{M}_1$. By definition, $R^{\mathbf{I}_0}(a_1, \ldots, a_n)$ is true in $\mathbf{M}_0$. But the sentence $H$ is true in $\mathbf{M}_0$, so in particular $(\forall x_1) \cdots (\forall x_n)[R(x_1, \ldots, x_n) \supset R'(x_1, \ldots, x_n)]$ is true. It follows that $R'^{\mathbf{I}_0}(a_1, \ldots, a_n)$ is true in $\mathbf{M}_0$, and hence $R^{\mathbf{I}_2}(a_1, \ldots, a_n)$ is true in $\mathbf{M}_2$. Then the identity map is a homomorphism from $\mathbf{M}_1$ to $\mathbf{M}_2$!

Finally, $X$ is true in $\mathbf{M}_0$, hence $X$ is true in $\mathbf{M}_1$ by the Pulling Apart Assertion. But we are assuming that $X$ is preserved under homomorphisms, hence $X$ is true in $\mathbf{M}_2$. Then by the Pulling Apart Assertion again, $X'$ must be true in $\mathbf{M}_0$, but it is not. This contradiction establishes the validity of $X \supset (H \supset X')$.

Now, by Lyndon's Interpolation Theorem 8.12.3, there is a sentence $Z$ such that

1. $X \supset Z$ is valid.
2. $Z \supset (H \supset X')$ is valid.
3. Every relation symbol that occurs positively (negatively) in $Z$ occurs positively (negatively) in both $X$ and $H \supset X'$.

No primed relation symbol $R'$ occurs in $X$, hence it cannot occur in $Z$. Also, $R$ does not occur in $X'$, and the only occurrence of $R$ in $H$ is in the homomorphism sentence for $R$, $(\forall x_1) \cdots (\forall x_n)[R(x_1, \ldots, x_n) \supset R'(x_1, \ldots, x_n)]$, where it occurs negatively. Consequently, the only occurrence of $R$ in $H \supset X'$ is positive, and hence any occurrence of $R$ in $Z$ must be positive. It follows that $Z$ itself is a positive sentence of $L$.

We have established that $Z \supset (H \supset X')$ is valid. For each relation symbol $R$ of $L$, if every occurrence of $R'$ in this sentence is replaced by an occurrence of $R$, the sentence remains valid. This does not affect $Z$, since no $R'$ occurs in $Z$. Making the replacement turns $X'$ into $X$. And finally, replacing $R'$ by $R$ converts $H$ into a tautology, since the homomorphism sentence for $R$ becomes $(\forall x_1) \cdots (\forall x_n)[R(x_1, \ldots, x_n) \supset R(x_1, \ldots, x_n)]$. Consequently, $Z \supset X$ must be valid. We also know that $X \supset Z$ is valid, hence we have the validity of $X \equiv Z$ for a positive sentence $Z$ of $L$. □

We note that the Interpolation Theorem 8.12.5 can also be used to prove a "preservation" result. It can be used to show that a sentence is preserved under passage to submodels if and only if it is equivalent to a universal sentence. We do not give the argument here.

## Exercises

**8.14.1.** Show that if $X$ is a positive formula then there is a formula $X^*$ in which there are no negation symbols and in which the only binary propositional connectives are $\wedge$ and $\vee$, such that $X \equiv X^*$ is valid.

**8.14.2.** Prove Proposition 8.14.3. Hint: You need to prove a more general result, about *formulas*, not just sentences. Let $h$ be a homomorphism from the model $\mathbf{M} = \langle \mathbf{D}, \mathbf{I} \rangle$ to the model $\mathbf{N} = \langle \mathbf{E}, \mathbf{J} \rangle$. For each assignment $\mathbf{A}$ in $\mathbf{M}$, let $h\mathbf{A}$ be the assignment in $\mathbf{N}$ given by $x^{h\mathbf{A}} = h(x^{\mathbf{A}})$. Now show by structural induction that, for each positive formula $X$, and any assignment $\mathbf{A}$ in $\mathbf{M}$, if $X^{\mathbf{I},\mathbf{A}}$ is true in $\mathbf{M}$, then $X^{\mathbf{J},h\mathbf{A}}$ is true in $\mathbf{N}$.

**8.14.3.** Verify the Pulling Apart Assertion in the proof of Lyndon's Theorem. Hint: As in the previous exercise, you need to establish a stronger result about formulas.

**8.14.4.** Let $L$ be the language with $R$ as a binary relation symbol and no function or constant symbols. Being a reflexive relation can be characterized by a positive sentence: $(\forall x)R(x, x)$. The obvious characterization of being a symmetric relation is $(\forall x)(\forall y)[R(x, y) \supset R(y, x)]$, but this is not a positive sentence.

1. Show there is no positive sentence that characterizes the symmetric relations.
2. Show the same for transitivity.