KENNETH WEXLER AND M. RITA MANZINI

PARAMETERS AND LEARNABILITY IN BINDING THEORY*

1. INTRODUCTION

1.1. *The Modular Approach*

Modern theory has provided evidence that universal grammar contains principles of a general, but specifically linguistic, form that apply in all natural languages. A goal of this paper is to extend the notion of principle theory to language acquisition. In such a theory each choice that the child makes in his or her growing language is determined by a principle of language or by a principle of learning or by the interaction of these two kinds of principles. The language principles and the learning principles are obviously related (they interact). However, it seems to be a promising approach to see if the two kinds of principles can be separated to some degree. That is, we attempt a modular approach to language acquisition theory. Some aspects of language and its acquisition seem better stated not in linguistic theory, but outside it, in, say, a learning module.

The general idea is that there are aspects of, say, markedness theory that are not part of linguistic theory, but are part of the learning theory. It may be possible to remove certain substantive assumptions about markedness, which are motivated primarily by learning conditions, from linguistic theory. The markedness hierarchies would instead be calculated from principles of the learning module.

An especially impressive demonstration of modularity would exist if the learning module predicted different markedness hierarchies on the same parameter depending upon the cases, for a substantive assumption of markedness within linguistic theory would not do this. We will suggest such a case.

One of the major insights of modern work in language acquisition and linguistic theory is the extent to which the question of how language is acquired is intertwined with the question of how linguistic theory should characterize variation in languages. In this paper the mechanisms underlying variation will be seen to have natural inter-

Thomas Roeper and Edwin Williams (eds.), Parameter Setting, 41–76.
© 1987 *by D. Reidel Publishing Company.*

pretations with regard to acquisition mechanisms. Once again, the solutions to problems of acquisition will be seen to have a highly specific linguistic content and to depend at the same time upon specialized learning principles. In our view not only does this view appear to be correct, but it is highly desirable, as it leads to a highly deductive theory.

1.2. *Binding Theory*

The theory that we develop of the learning of values of parameters is meant to be a general theory, which will apply in principle to all parameters. To illustrate and motivate the theory, we will apply it in this paper to a particular case, the learning of the values of parameters associated with binding theory (Chomsky, 1981). In this regard, it will be necessary first to develop a theory of the parameters associated with binding theory. It is important to realize, however, that the two theories are independent of each other. In particular, the learning theory, the theory of parameter setting that we develop, could be correct even if the binding theory and its associated parametric theory are incorrect. Nevertheless, the fact that a number of principles developed in the parameter setting theory do seem to have empirically motivated instantiations in binding theory does offer support for the parameter setting theory. In other words, the parameter setting theory makes certain predictions about properties of parametric systems, and binding theory appears to conform to these properties, which are not *a priori* necessary.

Later in this paper we will develop the statement of some aspects of binding theory, and associated parameters, in some detail. For now, however, we will state a few well-known properties in a non-formal manner, so that we can use these properties to illustrate the parameter setting principles. Anaphors and pronominals enter into binding theory. Under binding theory, in English, an anaphor, like *herself*, must be bound locally (binding principle A), and a pronoun, like *her*, must be free locally (binding principle B). For an element X to be 'bound' means for X to be c-commanded by a co-indexed element; and 'free' means not bound. As for the definition of 'local', we will consider only the simplest cases where 'local' can be taken to mean "in the same sentence as".

What must be learned with respect to binding theory? Presumably,

very little. As Chomsky (1980) points out, a child has to learn that *herself* is an anaphor. Presumably every other fact about the distribution, possible antecedents, etc., of *herself* will then follow from binding theory. Likewise the child has to learn that *her* is a pronoun but other facts about the distribution of *her* and its antecedents will follow. Thus a very little amount of learning will entail a large amount of growth of knowledge in the child.

It is becoming increasingly clear, however, that the properties of binding are not completely universal. For example, in languages like Japanese there appears to be an anaphor that must be bound, but doesn't have to be locally bound; rather, the binder (co-indexed *c*-commanding NP) can be arbitrarily distant in the sentence from the anaphor. If there is variation of this sort (and other sorts), then something must be learned besides whether a nonreferential item is a pronoun or anaphor. This variation and its relation to acquisition (parameter setting) will be explored in some detail in Section 2.

1.3. *The Subset Principle*

How are the values of parameters learned from experience? Following the general consensus in the field (Wexler and Hamburger, 1973; Baker, 1979), we assume that only 'positive data' is available to the learner. That is, the learner is not corrected for ungrammatical sentences; in general s/he receives no direct information about ungrammaticality. The lack of negative data leads to the following well-known learnability dilemma. If the child ever overgeneralizes, that is, picks the value of a parameter which gives too large a language, then there is no way (given only positive data) to correct the overgeneralization, since all new (positive) data will be generated by the (overgeneral) grammar. To be precise, this problem arises in the case where the overgeneral language is a superset of the correct language. For, otherwise, there will be data which indicates to the learner that the language s/he has selected is wrong.

To overcome this problem, it is often suggested that some kind of markedness theory will allow values of parameters to be learned; and in fact, the logic is more general, applying not only to the parameter setting case, but to any model of language variation and acquisition where subsets and supersets arise. In particular, it is suggested that the markedness hierarchy can be constructed in accordance with the

'Subset Principle'. The term 'Subset Principle', together with the appropriate logic, was first used by Berwick (in press), and our formulation is closely related to his.

The idea is the following. Suppose one value of a parameter yields a language $L(i)$ and another value of the parameter yields a language $L(j)$. Suppose further that $L(i)$ is a smaller language than $L(j)$, that is, that $L(i)$ is contained in $L(j)$. $L(i)$ is a strict subset of $L(j)$. Then the learning strategy specified by the Subset Principle is that the learner select the value which yields $L(i)$ first. If this is the correct choice, there will never be evidence that it isn't, and the learner will stay with the value. If this is the wrong choice, then there will be positive evidence (sentences from $L(j)$ which are not in $L(i)$) which the learner will eventually hear; this evidence must exist, because $L(i)$ is a strict subset of $L(j)$. The Subset Principle specifies that when positive evidence which shows that $L(i)$ is the wrong language is encountered, the learner will switch to the parameter value which yields language. In short, the Subset Principle is a method for specifying a markedness hierarchy when alternative values yield languages which are in a subset relation.

To fix notation, let us formalize the Subset Principle in the following way. Let i and j be values of a linguistic parameter p. $L(p(i))$ is the language — we take a language to be a set of sentences — which is attained by letting p have the value i. Likewise for j. Then we can state the Subset Principle. Suppose $L(p(i)) \subseteq L(p(j))$. Then i is less marked then j. In acquisition terms, if i is a less marked value than j, then i is tried by the learner before j, and only positive evidence that i is wrong moves the learner to j. In the simplest case the positive evidence can be just one sentence S. $S \in L(p(j)) - L(p(i))$; that is, one sentence S which is in $L(p(j))$ but not in $L(p(i))$.

Suppose that there was no Subset Principle. Furthermore, suppose the learner assumes first that j is the correct value. S/he will then always be correct on any sentence that is encountered. Even if i is the correct value, there will be no positive evidence which will guide the learner to select i.

To consider the anaphor example that we mentioned earlier, suppose that the difference between *herself* in English and *zibun* in Japanese is that *herself* must be bound locally whereas *zibun* can be bound 'anywhere' (that is, the binder can appear in any position in the sentence which commands *zibun*). It is clear that (everything else being equal), if an anaphor is taken to be necessarily locally bound, then a

smaller language will result than if it is taken to be bound anywhere. That is, let *i* be the value of the parameter that says that an anaphor must be bound within its sentence. Let *j* be the value which says that the item can be bound anywhere. Then $L(p(i) \subseteq L(p(j))$. For example, let *w* be an anaphor. Condider the sentence, "*John* thinks that Mary likes *w*." In this sentence, *w* is non-locally bound. Therefore, the sentence is in $L(p(j))$, but not in $L(p(i))$. Now consider the sentence, "*John* shaved *w*." In this second sentence, *w* is locally bound. So the sentence is in $L(p(i))$. But it is also in $L(p(j))$. So indeed $L(p(i)) \subseteq L(p(j))$.

In summary, suppose that *i* and *j* as given (locally bound vs. non-locally bound anaphors) are values for the 'anaphor parameter' and languages choose one of them. Then we can let the Subset Principle decide which is unmarked. Its choice will be *i* (locally bound). If this choice is wrong, positive evidence will be available which will allow the learner to choose *j* (bound anywhere).

In order for the Subset Principle to determine a strictly ordered learning hierarchy — this is what we will mean when we say that the Subset Principle as we define it here applies — it is necessary that two values of a parameter in fact yield languages which are in a subset relation to each other (i.e., one is a subset of the other). This requirement we call the 'Subset Condition'. It is necessary for the Subset Condition to hold in order for the Subset Principle to apply.

It is a consequence of the modular theory of parameter setting or learning that markedness hierarchies can be calculated by the Subset Principle in interaction with principles of Universal Grammar, making it unnecessary to state the hierarchies as substantive universals within linguistic theory. In fact, the interaction can yield different hierarchies even of the same linguistic parameter. An example which we will treat in detail concerns the differences between anaphors and pronouns. Since anaphors are governed by Principle A, which says that they are bound, the generated language will be smaller when the anaphor is bound locally than when it is bound anywhere. We have already seen an example of this case. A pronoun, however, is governed by Principle B. Suppose that parametric variation which is possible for anaphors is also possible for pronouns, so that in some languages pronouns must be free 'everywhere', instead of 'locally'. A calculation similar to the one we gave for anaphors will show that when pronouns must be free locally the language is a superset of the language generated when

pronouns must be free everywhere. (For example, where *w* is a pronoun, "*John* thinks that Mary likes *w*" is in the former language but not in the latter one.) Therefore, for pronouns, the local domain of binding is marked with respect to the entire sentence ('everywhere') domain of binding. But we have already seen that for anaphors precisely the opposite is the case, that is, the local domain of binding is unmarked with respect to the entire sentence ('anywhere') domain of binding. Therefore, a substantive universal listing the markedness hierarchy cannot be stated, at least independently of whether the hierarchy refers to a pronoun or an anaphor. But the different hierarchies are exactly what would be expected from the interaction of the Subset Principle and the Binding Theory. To the extent that empirical evidence corroborates the different hierarchies, this provides evidence for the relevance of the Subset Principle.

1.4. *The Many-Parameter Problem: The Independence Principle*

A special problem arises when we take into account the fact that there is more than one parameter in a language. The problem, and the term "many-parameter problem", were introduced in a talk by Bob Matthews in 1982 at Western Ontario (R. Matthews, personal communication). So just setting one parameter does not allow the 'language' to be calculated. How should the other values be set when the languages are calculated? What we have to assume is that, for every parameter *p* the languages must be nested (form subsets of each other), for all values of all other parameters. This is what we will refer to as the 'generalized' Subset Condition.

It turns out, however, that the (generalized) Subset Condition is not sufficient to insure that the Subset Principle can apply. We also have to insure that the particular subset relation of the languages formed by two values (say *i* and *j*) of a parameter are not affected by the setting of the other parameters. If *i* produces a subset of *j* for some setting of the other parameters, then it will produce the same subset for all other values of a parameter. This property is called *Independence*. The Subset Condition and Independence are necessary and sufficient for the Subset Principle to apply in all cases.

The idea of these conditions is that one can set the parameters independently of each other. If there are chains of derivational implications between parameters, it may be that there really is only one

parameter. It seems to us that Independence is really what linguists have implicitly had in mind when they talk about parameter setting in a manner akin to the Subset Principle. However, as far as we know, nobody has explicitly realized what it is necessary to assume in order to deal with the more than single parameter case. What is intriguing is that it appears that parameters that are stated in a natural linguistic way seem to satisfy, in many cases, the Subset Condition and Independence which, *a priori*, it would be quite easy to violate. In later sections we will give a detailed study of parameters and demonstrate how the Subset Condition and Independence hold. It is an important question for future study how generally true these conditions are.

1.5. *The Lexical Parameterization Hypothesis*

A further property that will emerge is that all anaphors in one language do not appear to obey exactly the same laws. To the extent that this is true it suggests that different lexical items can be associated with different values of lexical items; hence, as argued first by Borer (1984), that parameterization is essentially lexical. In a sense this should not be too surprising because it has long been recognized by linguistic theory that the lexicon states much or most of the idiosyncracies in a language. The basic view of grammar is not changed. Simply something more has to be said about a non-referential lexical item than whether it is an anaphor or a pronominal. As long as this statement of properties can be associated with an adequate learning theory, we will not have lost anything.

It does seem worth noting, however, that there are conceptual differences: the idea of parameter used here and the standard notion. Usually a parameter is taken to be associated with a whole language or grammar — an example is Rizzi's discussion of the Subjacency Parameter (See Chomsky, 1980, 1981). The problem for the child is how to use limited data to set a value for the parameter. But the parameter, once set, has extensive consequences for variation throughout the language. The kinds of parameters we are discussing, however, are set by the child for a particular lexical item. On the one hand the consequences of setting a lexical parameter would not be as broad as in the case of a language-wide parameter. On the other hand, the learnability problems might be considerably less severe.

It is important finally to make clear the status of lexical parameteri-

zation with respect to the parameter setting (learning) theory that we develop. The principles of learning (Subset Principle, Subset Condition, Independence and other aspects of the construction of the markedness hierarchy and acquisition theory) are necessary for any kind of parameter setting theory of the kind discussed here, whether parameters are selected for a grammar as a whole or for lexical items. In other words, two aspects of theory, the learning theory and the necessity for the association of parameters with lexical items, are independent.

2. PARAMETERS

As we saw above, binding theory, as introduced in Chomsky (1981), consists essentially of two principles, a binding principle A stating roughly that an anaphor must be bound locally, or, to be more precise, in the domain referred to as governing category, and a binding Principle B stating that in the same domain a pronominal must be free, i.e., nonbound, as in (1) (for an extension of the binding principles see Manzini, 1983; see Chomsky, in press, and Manzini, forthcoming, for a state of the art discussion):

(1) A. An anaphor is bound in its governing category
B. A pronominal is free in its governing category

In (1) the term 'bound' means '*c*-commanded and coindexed,' as in the definition of binding in (2); and correspondingly the term 'free' means 'not both bound and coindexed':

(2) α binds β iff
α and β are coindexed and α *c*-commands β

Furthermore, the notion of governing category for an element is defined essentially as in (3), as the minimal category which contains the element under consideration and has a subject:

(3) γ is a governing category for α iff
γ is the minimal category which contains α and has a subject

Consider, for example, English and in particular the English reflexive *himself* or the personal pronoun *he*. Obviously *himself* is an

anaphor, in that it cannot have any reference independently of an antecedent in the sentence; *he* is a pronominal, in that it can depend for its reference on an antecedent in the sentence or in the discourse, or refer deictically. It is easy to see that the theory in (1)—(3) correctly accounts for the distribution of *he* and *himself*, as well as for the distribution of their antecedents.

Consider first *himself*, as in (4), where italics is used as an alternate notation for coindexing:

(4) a. *John* criticized *himself*
b. *John* heard criticisms of *himself*
c. **John* heard [my criticisms of *himself*]
d. **John* heard [me criticize *himself*]
e. **John* forced me [to criticize *himself*]
f. **John* knew that [I criticized *himself*]

It is easy to see that the theory in (1)—(3) correctly accounts for both the well-formedness of the examples in (a)—(b) and the ill-formedness of the examples in (c)—(f). Consider first (a)—(b). In (a)—(b) the minimal category which contains *himself* and a subject, namely *John*, is the matrix sentence. Hence by the definition of governing category in (3) the matrix sentence is the governing category for *himself*; and by binding principle A *himself* must be bound in the matrix sentence. But in (a)—(b) *himself* is indeed bound in the matrix sentence, by *John* again; hence both (a) and (b) are correctly predicted to be well-formed.

Consider on the other hand (4c)—(4f). In (c) the minimal category which contains *himself* and a subject, namely the genitive *my*, is the embedded nominal, as bracketed; in (d)—(f) the minimal category which contains *himself* and a subject, namely the accusative subject *me* in (d), an empty subject PRO in (e) and the nominative *I* in (f), is the embedded clause, again as bracketed. Hence by the definition of governing category in (3) the governing category for *himself* is the embedded nominal in (c) and the embedded clause in (d)—(f); and by binding principle A *himself* must be bound in the embedded nominal in (c) and in the embedded clause in (d)—(f). But in (c) *himself* is bound, by *John*, outside the embedded nominal, and in (d)—(f) *himself* is bound, by *John* again, outside the embedded clause. Hence binding principle A is violated, and all of (c)—(f) are correctly predicted to be ill-formed.

Consider, then, *he*, as in (5):

(5) a. **John* criticized *him*
b. *John* heard [me criticize *him*]
c. *John* forced me [to criticize *him*]
d. *John* knew that [I criticized *him*]

According to the definition of governing category in (3), the matrix sentence is the governing category for *he* in (a) and the embedded clause the governing cagetory for *he* in (b)—(d). For, in (a) the matrix sentence is the minimal category which contains *he* and a subject, namely *John*; in (b)—(d) the minimal category which contains *he* and a subject — *me* in (b), an empty subject PRO in (c), and *I* in (d) — is the embedded clause. Hence by binding principle B, in (a) *he* must be free, i.e., not bound, in the matrix sentence; in (b)—(d) *he* must be free, i.e., not bound, in the embedded clause. But in (b)—(d) *John* does bind *he* outside the embedded clause; hence the theory in (1)—(3) correctly predicts (b)—(d) to be well-formed. On the other hand, in (a) *he* is bound by *John* in the matrix sentence; hence binding principle B is violated, and (a) is correctly predicted to be ill-formed.

Thus the theory of binding in (1)—(3) correctly accounts for English *himself* and *he*, and indeed for a significant number of pronominals and anaphors across languages.

Consider, however, Icelandic, and in particular the Icelandic reflexive *sig* as described, for example, in Johnson (1984). As its English counterpart *himself*, Icelandic *sig* obviously is an anaphor, in that it cannot have any reference independently of an antecedent in the sentence. But Icelandic *sig*, contrary to English *himself*, cannot be correctly accounted for by the binding theory in (1)—(3).

Consider indeed the examples in (6), where *sig* is roughly translated as REFL, for 'reflexive':

(6) a. **Jón* segir að [Maria elskar *sig*]
Jon says that Maria loves REFL

b. *Jón* segir að [Maria elski *sig*]
Jon says that Maria loves (subjunctive) REFL

c. *Jón* skipaði Haraldi að [raka *sig*]
Jon ordered Harald to shave REFL

d. *Jón* heyrdu [lysingu *Maria* af sér]
Jon heard Maria's description of REFL

All of (6a)—(6d) are predicted by the theory in (1)—(3) to be ill-formed. For, according to the definition of governing category in (3), the governing category for *sig* is the embedded sentence in (6a)—(6c) and the embedded nominal in (6d), since the embedded sentence in (6a)—(6c) and the embedded nominal in (6d) obviously are the minimal category which contains *sig* and a subject. By binding principle A, then, *sig* must be bound in the embedded sentence in (6a)—(6c) and in the embedded nominal in (6d); and since in (6a)—(6c) and (6d) *sig* is bound, by *Jón*, outside the embedded sentence and the embedded nominal respectively, all of (6a)—(6d) are ultimately predicted to be ill-formed. But while (6a) is actually ill-formed, (6b)—(6d) are not; hence the theory in (1)—(3) correctly accounts for (6a), but obviously makes the incorrect predictions for (6b)—(6d).

The fact that the theory in (1)—(3) cannot account for Icelandic *sig*, however, does not mean that the distribution of *sig* and its antecedents is completely free. On the contrary, one can easily show that *sig* must in general be bound, as it is in (6); hence at least the part of binding theory which states that an anaphor must be bound applied to *sig* as well. Furthermore, the contrast between examples of the type of (6a) and examples of the type of (6b)—(6d) suggests that *sig* not only must be bound, but must also be bound within a domain of some sort. Hence one can in fact assume that binding principle A, as stating that an anaphor is bound in its governing category, applies in its entirety to *sig* as to its English counterpart *himself*; except that the notion of governing category, defined as in (3) for English *himself*, must be defined in some different way for Icelandic *sig*.

Let us then assume, following Johnson (1983, 1984), that in the case of *sig* the definition of governing category in (7), rather than the definition of governing category in (3), applies:

(7) γ is a governing category for α iff
γ is the minimal category which contains α and has an indicative TNS

It is not difficult to see that the definition of governing category in (7) taken together with binding principle A actually accounts for all of the data in (6).

Consider first (6a). In (6a) the embedded sentence has an indicative Tense; hence the minimal category which contains *sig* and an indicative Tense obviously is the embedded sentence. Consider on the other

hand (6b)—(6d). Contrary to (6a), the embedded sentence in (6b) only has a subjunctive Tense; the embedded sentence in (6c) has no Tense at all; and the embedded nominal in (6d) obviously does not have an INFL, let alone a Tense. Hence in (6b)—(6d) the minimal category which contains *sig* and an indicative Tense is the matrix sentence. According then to the definition of governing category, in (6a) the governing category for *sig* is the embedded sentence; in (6b)—(6d) the governing category for *sig* is the matrix sentence. Hence by binding principle A, in (6a) *sig* must be bound in the embedded sentence; in (6b)—(6d) *sig* must be bound in the matrix sentence. But in (6a) *sig* is bound by *Jón* outside the embedded sentence; hence, correctly, the theory predicts that (6a) is ill-formed. In (6b)—(6d), on the other hand, *sig* is in fact bound by *Jón* within the matrix sentence; hence, correctly again, the theory predicts that (6b)—(6d) are well-formed.

Thus, summing up, while the theory of binding in (1)—(3) correctly accounts for English, as in (4)—(5), Icelandic, as in (6), is correctly accounted for by a revision of the theory with (7) substituted for (3). On the one hand, then, the definition of binding in (2) and the binding principles in (1) invariably hold; on the other hand, different definitions of governing category hold in different cases, (3) holding for English, as in (4)—(5), and (7) for Icelandic, as in (6).

If so, one is led in turn to the conclusion that while the theory of binding as a whole is a subtheory of Universal Grammar, indeed as in Chomsky (1981), one of the notions which crucially enter into it, the notion of governing category, is a parameter of the theory, much in the sense of Chomsky again. In particular, the two definitions of governing category in (3) and (7) represent two values of the parameter; the value represented by the definition in (3) is associated with English, as exemplified in (4)—(5); the value represented by the definition in (7) is associated with Icelandic, as exemplified in (6).

To be more precise, the definitions of governing category in (3) and (7) do not just differ one from another; they also obviously have an identical common core. By collapsing then (3) and (7) one can obtain the definition of governing category in (8), where the parameter is now seen to be internal to the definition of governing category itself:

(8) γ is a governing category for α iff
γ is the minimal category which contains α and
either has a subject
or has an indicative Tense

In (8) obviously the two values of the parameter correspond to the two terms of the disjunction; the value corresponding to the first term of the disjunction is the value associated with English, as in (4)—(5), while the value associated with the second term of the disjunction is the value associated with Icelandic, as in (6).

But while the parametrized definition of governing category in (8) accounts not only for English *himself* and *he*, but also for Icelandic *sig* and for an increased number of other pronomials and anaphors across languages, it only partially accounts for the total observed range of variation. A discussion of what a complete such account requires would inevitably exceed the limits of this paper; following Manzini and Wexler (1984), then, here we will simply assume that a number of new values must be introduced, and that in particular, once introduced the new values of the parameter, the definition of governing category in (9) is obtained (alternative accounts can be found notably in Yang, 1984, and Koster, 1984):

(9) γ is a governing category for α iff
γ is the minimal category which contains α and
a. has a subject, or
b. has an INFL, or
c. has a TNS, or
d. has an indicative TNS, or
e. has a root TNS

Evidently the definition of governing category now includes a five-valued parameter, with the five values of the parameter corresponding to the five members, (*a*), (*b*), (*c*), (*d*), and (*e*) of the disjunction in (9). A value (*a*) of the parameter in (9) obviously identifies with the first value of the parameter in (8) or the value associated with English, as in the examples in (4)—(5); value (*d*) in (9) identifies with the second value of the parameter in (8) and indeed the value associated with Icelandic, as in the examples in (6). (9) then introduces three new values, (*b*), (*c*), and (*e*), respectively.

It must be noticed at this point that in the discussion which precedes the values of the parameter in the definition of governing category have been referred to ambiguously as the values associated with particular languages or as the values associated with particular anaphors and pronominals in a language. So value (*a*) of the parameter in (9) has been referred to as the value associated with English or as the value associated with English *himself* and *he*; and value (*d*) of the parameter

has been referred to as the value associated with Icelandic, or as the value associated with Icelandic *sig*. A close examination of the data, however, leads to the conclusion that the values of the parameter in (9) cannot be associated with particular languages, but rather must be associated with particular anaphors and pronominals in a language.

Consider for example Icelandic again. While the Icelandic reflexive *sig*, as in (6), is correctly accounted for under value (*d*) of the governing category parameter in (9), the Icelandic personal pronoun *hann* gives rise to examples of the type of (10) which are incorrectly accounted for under value (*d*):

(10) a. *Jón* segir að [Maria elskar *hann*]
Jon says that Maria loves him

b. *Jón* segir að [Maria elski *hann*]
Jon says that Maria loves (subjunctive) him

c. **Jón* skipaði mér að [raka *hann*]
Jon ordered me to shave him

Rather, a correct account of Icelandic *hann*, as in (10), requires associating *hann* with value (*c*) of the parameter. If so indeed in (10a) and (10b) the governing category for *hann* is the embedded sentence; for the embedded sentence, being a tensed sentence, obviously is the minimal category which contains *hann* and has a TNS. In (10c), on the other hand, the governing category for *hann* is the matrix sentence; for given that the embedded sentence is untensed, the minimal category which contains *hann* and has a TNS is the matrix sentence. By binding principle B, then, *hann* must be free in the embedded sentence in (10a)—(10b) and in the matrix sentence in (10c). Hence (10a) and (10b), where *hann*, though bound within the matrix sentence, is free within the embedded sentence, are correctly predicted to be well-formed; but (10c) where *hann*, though free within the embedded sentence, is bound within the matrix sentence, is correctly predicted to be ill-formed.

Thus, while English, as in (4)—(5), could be associated with value (*a*) of the governing category parameter as a language, or English *himself* and *he* could be associated with value (*a*) of the parameter as lexical items, Icelandic as a language cannot be associated with any single value of the governing category parameter; rather, Icelandic *sig*, as in (6), and *hann*, as in (10), must be associated with values (*d*) and (*c*),

respectively, of the parameter as lexical items. In general, then, one is led to the conclusion that values of the parameter in (9) are associated not with particular languages, but with particular lexical items in a language. This conclusion, extended from the case of the parameter in (9) to the case of parameters in general, we codify informally as in (11), and refer to as the *Lexical Parametrization Hypothesis*:

(11) Values of a parameter are associated not with particular languages, but with particular lexical items in a language.

3. LEARNABILITY

Consider now the definition of governing category in (9) again. It is relatively easy to see that the values of the parameter in (9) define sets of categories which are subsets one of another; and in particular that the values in (9) are ordered already in such a way that the set of categories defined by each value is a subset of the set of categories defined by the immediately following value. So, concretely, the set of categories which have a subject, as in value (*a*) of the parameter, i.e., all sentences and some small clauses and nominals, is a superset of — i.e., (properly) includes — the set of categories which have an INFL, i.e., sentences, as in value (*b*) of the parameter. Similarly, the set of categories which have an INFL, or sentences, is a superset of the set of categories which have a TNS, i.e., tensed sentences, as in value (*c*) of the parameter; the set of categories which have a TNS, or tensed sentences, is a superset of the set of categories which have an indicative TNS, i.e., indicative sentences, as in value (*d*) of the parameter; and finally the set of categories which have an indicative TNS, or indicative sentences, is a superset of the set of categories which have a root TNS, i.e., root sentences, as in value (*e*) of the parameter.

It is then not too difficult to see that, given any element α, the governing categories for α defined by (9) under the different values of the parameter have the property of being embedded one inside another. Let us in particular call a governing category for α defined by (9) under values (*a*), (*b*), (*c*), (*d*), and (*e*) of the parameter, an A, a B, a C, a D, and an E, respectively. It is not too difficult to show that an A is always embedded inside a B, a B inside a C inside a D, and a D inside an E, in that order.

Consider for example two governing categories for α, A and B. By

definition, A is the minimal category which contains α and has a subject; while B is the minimal category which contains α and has an INFL, i.e., is a sentence. But the set of categories with an INFL, or sentences, is a subset of the set of categories with a subject. Hence, in particular B, being a sentence, necessarily is a category with a subject; while A, being a category with a subject, can either be sentence or not. Consider, then, A. If A is a sentence, then A can coincide with B; for A, by definition the minimal category with a subject containing α, being by hypothesis a sentence, can also be the minimal sentence containing α, i.e., by definition B. If on the other hand A is not a sentence, then A obviously cannot coincide with B. But if A is distinct from B, then A can be contained in B; for nothing prevents A, if not a sentence, from being smaller than B, the minimal sentence containing α. However, whether A is a sentence or not, B cannot be contained inside A. For B necessarily is a category with a subject. Hence there can be no category with a subject which is both larger than B and the minimal category with a subject containing α, i.e., A. In other words, the situation in (12) obtains; where A and B can coincide, as in (12a), or not, as in (12b) to (12c), but if A and B do not coincide, B must contain A, as in (12b), and A cannot contain B, as in (12c):

(12) a. $\ldots [_{B=A} \ldots \alpha \ldots] \ldots$
b. $\ldots [_{B} \ldots [_{A} \ldots \alpha \ldots] \ldots] \ldots$
c. $*\ldots [_{A} \ldots [_{B} \ldots \alpha \ldots] \ldots] \ldots$

Similarly, given two governing categories B and C or C and D or D and E, B can coincide with C, C with D, and D with E; but if not, B must be contained inside C, C inside D, and D inside E and not vice versa.

Now, given any element α again, let us consider the cases in which A, B, C, D, or E is the governing category for α; or indeed α is associated with values (*a*), (*b*), (*c*), (*d*), or (*e*) of the parameter in (9). In particular, let us consider the sets of sentences, i.e., the languages, generated in the different cases; and let us call the languages generated, in case α is associated with values (*a*), (*b*), (*c*), (*d*), and (*e*) of the parameter, L(*a*), L(*b*), L(*c*), L(*d*), and L(*e*), respectively. One can rather easily show that given the nesting properties of the governing categories A, B, C, D, and E the languages L(*a*), L(*b*), L(*c*), L(*d*), and L(*e*) ultimately are a subset one of another.

Consider first the case in which α is an anaphor, such as English

himself, or Icelandic *sig*, and so on. It is easy enough to show that in this case not only L(a), L(b), L(c), L(d), and L(e) are a subset one of another, but in particular L(a) is a subset of L(b), L(b) of L(c), L(c) of L(d), and L(d) of L(e). Consider, for example, L(a) and L(b). Given that α is an anaphor, L(a) obviously contains all and only the sentences in which α is bound in A, its governing category; L(b) all and only the sentences in which α is bound in its governing category again, i.e., B. But an A either coincides with a B or is contained inside a B. Hence every sentence in which α is bound by some element β within an A and also is a sentence in which α is bound by β within a B, B coinciding with A or containing A. But not every sentence in which α is bound by β within a B is a sentence in which α is bound by β within an A. In other words, if A and B coincide, then obviously α is bound within A just in case α is bound within B, as in (13a); if B contains A, then if α is bound within A, α is bound within B, as in (13b); but if α is bound within B it can be the case that α is not bound within A, as in (13c):

(13) a. $\ldots [_{B=A} \ldots \beta \ldots \alpha \ldots] \ldots$

b. $\ldots [_{B} \ldots]_{A} \ldots \beta \ldots \alpha \ldots] \ldots] \ldots$

c. $\ldots [_{B} \ldots \beta \ldots [_{A} \ldots \alpha \ldots] \ldots] \ldots$

But if all sentences in which α is bound in an A also are sentences in which α is bound in a B, then all sentences which are in L(a) also are in L(b); while if not all the sentences in which α is bound in a B are sentences in which α is bound in an A, as in (13c), then not all sentences which are in L(b) are in L(a). Hence, the sentences in L(a) are a subset of the sentences in L(b), or indeed L(a) is a subset of L(b). Much in the same way L(b) can be shown to be a subset of L(c), L(c) of L(d) and L(d) finally of L(e).

Consider, on the other hand, the case in which α is a pronominal, such as *he* in English, or *hann* in Icelandic, and so on. Again L(a), L(b), L(c), L(d), and L(e) are a subset one of another; but in this case L(e) is a subset of L(d), L(d) of L(c), L(c) of L(b), and L(b) of L(a). Consider indeed L(b) and L(a). Given that α is a pronominal, L(a) obviously contains all and only the sentences in which α is free in A, its governing category; L(b) all and only the sentences in which α is free in its governing category again, i.e., B. In other words L(a) contains all and only the sentences in which α either is free or is bound outside an

A; L(b) contains all and only the sentences in which α either is free again, or is bound outside a B. But an A either coincides with a B or is contained inside a B. Hence, every sentence in which α is bound by some β outside a B also is a sentence in which α is bound by β outside an A, A coinciding with or being smaller than B. But not every sentence in which α is bound by β outside an A is a sentence in which α is bound by β outside a B. In other words, if A and B coincide, then obviously α is bound outside A just in case α is bound outside B, as in (14a); if B contains A, then if α is bound outside B, α is bound outside A, as in (14b); but if α is bound outside A it can be the case that α is bound inside B, as in (14c):

(14) a. $\ldots \beta \ldots [_{B=A} \ldots \alpha \ldots] \ldots$
b. $\ldots \beta \ldots [_{B} \ldots [_{A} \ldots \alpha \ldots] \ldots] \ldots$
c. $\ldots [_{B} \ldots \beta \ldots [_{A} \ldots \alpha \ldots] \ldots] \ldots$

It follows that all the sentences which are in L(b) are in L(a); for all the sentences in which α is free are both in L(a) and in L(b), and all the sentences in which α is bound outside a B are also sentences in which α is bound outside an A. But not all the sentences which are in L(a) are in L(b); for not all sentences in which α is bound outside an A are sentences in which α is bound outside a B. Hence the sentences in L(b) are a subset of the sentences in L(a), or indeed L(b) is a subset of L(a). Much in the same way, then, L(c) is a subset of L(b), L(d) of L(c), and finally, L(e) of L(d).

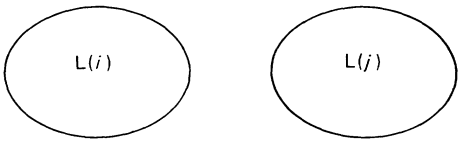
But if the languages L(a), L(b), L(c), L(d), and L(e) are indeed one a subset of another, and in particular each a subset of the following one in case α is an anaphor, and each a subset of the preceding one in case α is a pronominal, the question naturally arises whether this is the case for any reasons at all, and if so, for which reasons.

Obviously, the theory of grammar offers no answers; not only as the theory now stands, no reason exists for the fact that L(a), L(b), L(c), L(d), and L(e) are a subset one of another, but more generally the formal properties and relations of the languages defined by different values of a parameter seem to be no concern of the theory at all. Consider, however, the theory of learnability; obviously, the formal properties and relations of the languages defined by different values of

a parameter are a central concern of the theory. In particular, from the point of view of the theory of learnability, is that for any given parameter the learning function selects on the basis of the input data a value i of the parameter such that the input data are compatible with, i.e., a subset of, the language L(i) defined by that value; and that if there are two values i and j of the parameter which define languages, L(i) and L(j), both compatible with the input data, then in case L(i) and L(j) are one a subset of the other the learning function selects that of the two values which defines the smaller language.

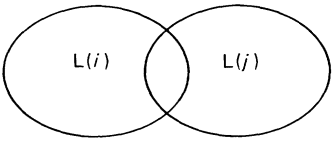
Consider indeed the two languages L(i) and L(j) defined by i and j. A first possibility is that L(i) and L(j) are disjoint, as in (15):

(15)



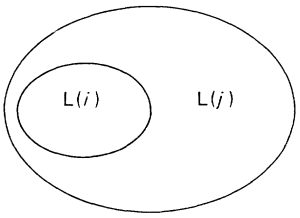
In this case obviously any set of input data is either a subset of L(i) or a subset of L(j), but not of both; hence, on the basis of the input data the learning function can straightforwardly select i or j, respectively. A second posibility is that L(i) and L(j) intersect, as in (16):

(16)



In this case the input data can be a subset of L(i) but not of L(j), or of L(j) but not of L(i), or finally of both L(i) and L(j). Obviously in the first two cases the learning function can straightforwardly select on the basis of the input data values i and j again. In the third case, on the other hand, it can select either i or j; for, whether it selects i or j, it can either stay with the value it has selected or select the other value on the basis of further data belonging to one but not the other language. But suppose finally that L(i) and L(j) are one a subset of another, say L(i) a subset of L(j), as in (17):

(17)



In this case obviously given a set of input data which belongs to L(j) but not to L(i), the learning function can straightforwardly select j. Given, however, a set of input data which belong both to L(i) and to L(j), crucially the learning function must select i. For, if it selects i, it can either stay with it or else select j on the basis of further data belonging to L(j) but not to L(i). But if it selects j there is no way it can select i instead, except on the basis of negative evidence; whereas we assume that only positive evidence exists.

Thus, the theory of learnability is directly concerned with the subset relations between the languages defined by the different values of a parameter; and it requires, in particular, in case two languages defined by two different values of a parameter are one a subset of another, that if the two languages are both compatible with the input data, the learning function selects that value of the parameter which defines the smaller language.

What the governing category parameter in (9) suggests is that a stronger requirement is part of the theory of learnability; and that in fact the theory of learnability requires that for every given parameter and every two given values of it, the languages defined by the two values of the parameter are one a subset of another. For such a requirement forces the languages defined by the values of the governing category parameter in (9) to be one a subset of another; hence, explains the fact that they are.

Let us assume, then, that the theory of learnability indeed includes a restriction to the effect that the languages generated by two values of a parameter are a subset one of the other, for every given parameter and every two values of it. This restriction we can formulate as in (18) and refer to as the *Subset Condition*:

(18) For every parameter p and every two values i, j of p, the languages generated under the two values of the parameter are one a subset of the other, that is, $L(p(i)) \subseteq L(p(j))$ or $L(p(j)) \subseteq L(p(i))$.

Obviously, the most important consequence of the Subset Condition from the point of view of the theory of learnability is that, if indeed any two values of any parameter define languages which are a subset one of the other, then it is true in all cases that the learnability function can map the input data to that value of a parameter which defines the smallest of the languages compatible with the data. We can formulate this principle as in (19), and refer to it as the *Subset Principle*:

(19) The learning function maps the input data to that value of a parameter which generates a language:
(a) compatible with the input data; and
(b) smallest among the languages compatible with the input data.

Thus, summing up, according to the Subset Condition, any two values of any parameter define languages which are a subset one of the other; while, according to the Subset Principle, on the basis of the input data the learnability function selects among the values of the parameter which generate languages compatible with the input data the value which generates the smallest language.

What do we mean when we say that the Subset Condition is necessary? We say that it is necessary in order for the Subset Principle to be always applicable. In other words, if the values that the learning function selects on the basis of data are determined by the Subset Principle and by nothing else, then the values of a parameter must determine languages which form a strict hierarchy of subsets. That is, any two values of the parameter must determine two languages such that one is a subset of the other.

It is important to note that the basis for the derivation of the Subset Condition is the assumption that the Subset Principle determines every value of the learning function. Thus we cannot allow two values of a parameter to yield languages which are disjoint, or intersecting, and not in a subset relation to each other. Of course, this assumption is itself not necessary. If two values of a parameter yield languages which are not in a subset relation, then we can arbitrarily order the two values, it would seem, in terms of the learning function. Then the Subset Principle would only have to apply to those cases of parameter values where a subset relation did hold between the languages. To repeat, the Subset Condition is necessary only in the special case where the Subset Principle is the *only* determiner of learning (or of markedness). Wexler (in preparation) has determined the necessary conditions which must

exist for a learning function (markedness hierarchy) to exist when the Subset Condition doesn't hold and there are other determinants to markedness.

Nevertheless, it is intriguing that our studies of parameters in binding theory yield parameters which in fact obey the Subset Condition. Since this fact is not logically necessary, it might indicate that there is indeed some important empirical import to the assumption that the Subset Principle is the only determinant of learning, in at least the case of some parameters. It will take considerably more study of further parametric systems to determine to what extent the Subset Condition in fact holds.

In addition, if the languages defined by two values of a parameter are a subset one of the other as required by the Subset Condition, and as required by the Subset Principle the learning function selects among two values of a parameter, everything else equal, the value which defines the smaller language, then, it seems evident that any two values of any given parameter can be defined to a more or less marked one than the other on the basis of the subset relations between the two languages they generate.

In particular, one can define an ordering of the values of a parameter to be a markedness hierarchy just in case the language generated by each value is a subset of the language generated by the immediately following value, informally as in (20):

(20) A given ordering of the values of a parameter is a markedness hierarchy if and only if the language generated by each value is a subset of the language generated by the immediately following value in the ordering.

Obviously, then, one can define a value of the parameter to be the unmarked value just in case it is first in the markedness hierarchy; and define a value of the parameter to be marked just in case it is not unmarked. Similarly, a value of the parameter can be defined to be less marked than another value just in case the former precedes the latter in the markedness hierarchy; or vice versa a value of the parameter can be defined to be more marked than another value just in case the former follows the latter in the markedness hierarchy.

Thus in turn the Subset Principle can be revised in terms of markedness relations to state directly that the learnability function selects among the values of a parameter which define languages compatible

with the input data the least marked value; and indeed the Subset Condition can be revised in terms of markedness relations to require simply that the values of any parameter define a markedness hierarchy.

It must be noticed at this point, however, that even the governing category parameter in (9) accounts only partially for the range of variation in binding theory. So, in particular, while some anaphors simply must be bound within their governing category, some anaphors must be bound within their governing category by a subject; and correspondingly, while some pronominals must be completely free in their governing category, other pronominals must be free in their governing categories only from subjects.

Consider, for example, English *himself* again. In its governing category English *himself* can be bound by a subject, as in (21a), or by an object, as in (21b):

(21) a. *John* told Bill about *himself*
b. John told *Bill* about *himself*

However, if one considers, rather than English *himself* the Japanese reflexive *zibun*, then a Japanese example corresponding to (21a) can again be seen to be well-formed; but a Japanese example corresponding to (21b) can on the contrary be seen to be ill-formed. Similarly, consider English *he*, as in (22). Within its governing category *he* must be free from both a subject, as in (22a), and an object, as in (22b):

(22) a. **John* told Bill about *him*
b. *John told *Bill* about *him*

If one considers, however, rather than English *he*, the Icelandic pronominal *hann* once more, one can easily see that the Icelandic example corresponding to (22a) is still ill-formed, but the Icelandic example corresponding to (22b) is now well-formed.

The variation just illustrated can easily be accounted for in the theory of grammar. Suppose indeed that the binding principles A and B, as in (1), are revised to state respectively that an anaphor not only must be bound in its governing category, but must be bound in its governing category by a proper antecedent; and a pronominal must not be free in its governing category, but must be free in its governing category just from proper antecedents, as in (23):

(23) a. An anaphor is bound in its governing category by a proper antecedent
b. A pronominal is free in its governing category from proper antecedents

Suppose furthermore that a proper antecedent for α is defined to be either a subject or else any element at all, as in (24):

(24) A proper antecedent for α is
a. a subject β; or
b. an element β whatsoever

If so, obviously the theory of binding includes not one but two different parameters: the already familiar governing category parameter, as in (9), and the newly introduced parameter in the definition of proper antecedent, as in (24), with the two values (*a*) and (*b*).

Not less obviously, the version of the binding principles in (23) together with the new parametrized definition of a proper antecedent in (24) correctly accounts for the contrast between English *himself* or *he*, as in (21) and (22), and Japanese *zibun* or Icelandic *hann*. Consider first *himself*. If *himself*, beside being associated with value (*a*) of the governing category parameter in (9), is associated with value (*b*) of the proper antecedent parameter in (24), then by binding principle A in (23), in (21) *himself* must be bound in the matrix sentence by any element. Hence both (21a), where *himself* is bound in the matrix sentence by a subject, and (21b), where *himself* is bound in the matrix sentence by an object, are correctly predicted to be well-formed. Consider, on the other hand, Japanese *zibun*. If *zibun*, contrary to English *himself*, is associated with value (*a*) of the proper antecedent parameter in (24), then by binding principle A in (23) *zibun* must be bound in its governing category by a subject. Hence an example of the type of (21a) is correctly predicted to be well-formed in Japanese, with *zibun* replacing *himself*, as in English; but an example of the type of (21b) with *zibun* replacing *himself* again is correctly predicted to be ill-formed in Japanese, contrary to Englich.

Similarly, consider *he*. If *he*, beside being associated with value (*a*) of the governing category parameter, is associated with value (*b*) of the proper antecedent parameter, exactly as *himself* is, then by binding principle B in (23), in (22) *he* must be free in the matrix sentence from all elements. Hence, both (22a), where *he* is bound in the matrix

sentence by a subject, and (22b), where *he* is bound in the matrix sentence by an object, are correctly predicted to be ill-formed. But consider Icelandic *hann*. If *hann*, contrary to English *he*, is associated with value (*a*) of the proper antecedent parameter, then by binding principle B in (23), *hann* must be free in its governing category only from subjects. Hence, an example of the type of (22a) is correctly predicted to be ill-formed not only in English but also in Icelandic with *hann* replacing *he*; while an example of the type of (22b), with *hann* replacing *he* again, is correctly predicted to be well-formed in Icelandic, contrary to English.

From the point of view of the theory of learnability, however, the proper antecedent parameter in (24) faces us with the general problem of the existence of several parameters in the theory of binding and obviously in the theory of grammar as a whole. In the case of the governing category parameter in (9), indeed, in the absence of any discussion about other parameters, the Subset Condition and the Subset Principle were shown to hold very much as if the governing category parameter itself was the only parameter in the theory of grammar. But if there are not one, but several parameters in the theory of grammar and in binding theory itself, can one require that the languages defined by two different values of one such parameter be one subset of the other abstracting away from all other parameters? Or can one still require abstracting away from all other parameters that the learnability function selects a value of a particular parameter on the basis of its defining the smallest language compatible with the data? Evidently if the Subset Condition and the Subset Principle are to apply as before, an additional principle must be introduced in the theory of learnability to the effect that the subset relations between languages generated under different values of a parameter can be established independently of all other parameters, or in other words that the subset relations between languages generated under different values of a parameter remain constant no matter what the values of the other parameters are taken to be. This principle we formulate informally in (25) and we refer to it as the *Independence Principle*:

(25) The subset relations between languages generated under different values of a parameter remain constant whatever the values of the other parameters are taken to be.

Given the Independence Principle it is easy to show that, as the

governing category parameter in (9), the proper antecedent parameter in (24) observes the Subset Condition; hence, the Subset Principle quite straightforwardly applies to it. Consider first the case in which α in (24) is an anaphor. If we call the languages generated when α is associated with values (a) and (b) of the proper antecedent parameter $L(a)$ and $L(b)$, respectively, it is quite easy to see that $L(a)$ is a subset of $L(b)$. Obviously indeed $L(a)$ includes all the sentences in which α is bound by a subject, while $L(b)$ includes all the sentences in which α is bound by any element at all. But since $L(b)$ includes all the sentences in which α is bound by any element at all, it includes in particular all the sentences in which α is bound by a subject, hence all the sentences included in $L(a)$; while conversely $L(a)$ does not include all the sentences included in $L(b)$, notably not the sentences in which α is bound by an object, or an indirect object, or in general a non-subject. Hence, since all the sentences included in $L(a)$ are included in $L(b)$, but not all the sentences included in $L(b)$ are included in $L(a)$, $L(a)$ obviously is a subset of $L(b)$.

Consider on the other hand the case in which α in (24) is a pronominal. Let us call again the languages generated when α is associated with values (a) and (b) of the proper antecedent parameter $L(a)$ and $L(b)$, respectively. Obviously $L(b)$ is a subset of $L(a)$. $L(a)$ indeed includes all the sentences in which α is free from subjects, i.e., all the sentences in which α is free from any element at all and all the sentences which in α is bound by a non-subject; while $L(b)$ includes all the sentences in which α is free from any element at all. But since $L(a)$ includes all the sentences in which α is completely free, it includes all the sentences included in $L(b)$; while conversely $L(b)$ does not include all the sentences included in $L(a)$, specifically not the sentences in which α is bound by a non-subject. Hence, since all the sentences which are in $L(b)$ are also in $L(a)$, but conversely not all the sentences which are in $L(a)$ are also in $L(b)$, $L(b)$ obviously is a subset of $L(a)$.

Finally, it must be noticed that while the governing category parameter in (9) and the proper antecedent parameter in (24) are such that the principles and conditions of the theory of learnability apply to them, these same principles and conditions only apply to them if they indeed are individual parameters. So, from the point of view of the theory of grammar, the governing category and proper antecedent parameter could be one single parameter; but if they were, it is easy to show that the new parameter would violate the principles and conditions of the

theory of learnability, in particular the Subset Condition. Hence the governing category and proper antecedent parameter must indeed be two separate parameters for reasons having to do with the theory of learnability once more.

Consider, for instance, the English pronominal *he* and the Icelandic pronominal *hann*. English *he* is associated with value (*a*) of the governing category parameter and value (*b*) of the proper antecedent parameter; Icelandic *hann* is associated with value (*c*) of the governing category parameter and value (*a*) of the proper antecedent parameter. Suppose that the governing category parameter in (9) and the proper antecedent parameter in (24) were one single parameter. English *he* would have to be associated with a value of the new parameter equivalent to the two values (*a*) and (*b*) of the governing category and proper antecedent parameter, respectively; Icelandic *hann* would have to be associated with a value of the new parameter equivalent to the two values of the governing category and proper antecedent parameter (*c*) and (*a*), respectively. Let us call the values of the new parameter associated with English *he* and Icelandic *hann* value (*a,b*) and value (*c,a*), respectively; and let us call the languages generated by values (*a,b*) and (*c,a*), respectively, L(*a,b*) and L(*c,a*). It is easy to see that, for any given pronominal, L(*a,b*) and L(*c,a*) would not be a subset one of the other. Indeed L(*c,a*) would include all the sentences in which the pronominal is bound by a nonsubject, but L(*a,b*) would not include those sentences in which the pronominal is bound by a non-subject within a governing category A, as defined by value (*a*) of the governing category parameter; hence L(*c,a*) would include some sentences which L(*a,b*) would not include. On the other hand, L(*a,b*) would include all the sentences in which a pronominal is bound outside a governing category A, but L(*c,a*) would not include those sentences in which the pronominal, though bound outside A, is bound by a subject inside a governing category C, as defined by value (*c*) of the governing category parameter; hence, L(*a,b*) would include some sentences which L(*c,a*) would not include. But if some sentences were in L(*c,a*) and not in L(*a,b*) and some sentences were in L(*a,b*) and not in L(*c,a*), neither L(*a,b*) would be a subset of L(*c,a*), nor L(*c,a*) a subset of L(*a,b*).

Similarly, one can consider instead of pronominals, anaphors: the result remains the same. If the governing category parameter in (9) and the proper antecedent parameter in (24) are made into one single parameter, the values of the new parameter define languages which are

not a subset one of another. But if so, the Subset Condition is violated; hence, the governing category and proper antecedent parameters must indeed be two separate parameters. Note, however, that if we adopted a weaker form of the theory, as discussed after (19), in which the Subset Condition did not hold, then the two parameters *could* be reduced to one parameter. The natural interpretations of the parameters, from a learning point of view, however, would be lost. Once again, it is important to discover to what extent the Subset Condition is empirically true.

4. ACQUISITION

4.1. *The Acquisition of Anaphors and Pronominals*

Let's start with the simple cases first. Assume that binding theory is not subject to variation, applying to all pronominals and all anaphors in all languages uniformly. Also, for simplicity again, we will consider only lexical anaphors and pronominals, ignoring for the moment empty categories. All that has to be learned about an item under these circumstances is whether it is referential (an *R*-expression) or not and in case it is not, whether it is a pronoun or anaphor.

It seems natural to assume that a child first decides whether an item is referential or not. How does he or she do this? Along with most contemporary language acquisition theorists (e.g., Wexler and Culicover, 1980; Pinker, 1984; cf. also MacNamara, 1972), we assume that the learner has the 'cognitive' capacity to derive some of the interpretation of some sentences, even when he or she does not understand the full set of grammatical properties of the sentence. Included in this interpretation is the reference of the noun phrases in the sentence. There appears to be a number of mechanisms by which a learner can conclude that an item is 'non-referential'. For a pronoun, for example, consider the 'deictic' use of a pronoun, in which it is not co-referential (co-indexed) with anything else in the discourse. For example, somebody points to a toy and says, "It's broken." How is the child to distinguish *it* from a referential item, like 'toy'? One possibility is that the child realizes that the things to which *it* refers, over a variety of sentences are very varied, with their features varying wildly, as compared, say, with the things to which *toy* applies. If this proposal is to be correct, it is clear that it involves cognitive abilities which go beyond syntax. Another possible mechanism, more syntactic, is that a child might realize that, if he or

she takes a certain item (really a pronoun) to be a referential item, then there are violations. Chomsky's (1981) Binding Principle C or whatever other principle of grammar (see Higginbotham, 1983) accounts for the fact that *R*-expressions cannot be bound. For example, in the sentence, "*Mary* thinks that *she* is ill," since *Mary* and *she* are co-indexed, the child can derive from Principle C that *she* is not a referential item (otherwise it couldn't be co-indexed with a *c*-commanding item). We assume, as we mentioned earlier, that the child has knowledge of which items in a sentence are co-referential. Or perhaps he or she has this knowledge only for some sentences, but this will be sufficient (cf. Wexler, 1981). There are likely a variety of other particular mechanisms which will allow the child to decide that an item is non-referential. We will simply assume for now that the child can decide this first, before he or she has to decide whether a non-referential item is a pronoun or anaphor.

4.2. *The Ordering of Pronouns and Anaphors*

So we assume that the learner knows that a particular lexical item *w* is non-referential. Therefore he or she knows that *w* is a pronoun or an anaphor. How does he or she decide which it is? For example, how does the child learning English decide that *him* is a pronoun and *himself* is an anaphor?

Following the general consensus in the field, we assume that only 'positive data' are available to the learner. That is, the learner is not corrected for ungrammatical sentences; in general he or she receives no direct information about ungrammaticality. (See Brown and Hanlon, 1970; Wexler and Hamburger, 1973; Baker, 1979; see Wexler and Culicover, 1980, Section 2, for a discussion of the input that is available to the child.) The assumption of only positive data creates obvious problems for the theory of language acquisition. For example, the linguist can use the ungrammaticality of sentences like "*Mary* hates *her*" to infer (from Principle A) that *her* is not an anaphor. But this negative information is not available to the child learning a first language.

In the absence of negative information, it is often suggested that some kind of markedness theory will allow values of parameters to be learned. Jakubowicz (1984) has suggested such a solution in the case of anaphors and pronominals. She suggests that the Subset Principle may

be used to construct a method of learning whether a non-referential item is an anaphor or a pronominal.

As we saw above, the Subset Principle, along with Binding Principle A, predicts that locally bound anaphors are unmarked with respect to non-locally bound anaphors. Jakubowicz, however, concentrates on what she claims is the following prediction from the Subset Principle: There will be a stage (in English, say) in which the child takes pronouns to be anaphors. This means that pronouns will be locally bound for the child at this stage, whereas for the adult, pronouns are never locally bound. Jakubowicz's argument seems to be the following: Anaphors are locally bound. Pronouns can be non-locally bound. Because of this if a non-referential item is an anaphor, a smaller language is obtained than if the item is a pronoun. Therefore, the Subset Principle predicts that (locally bound) *anaphor* is unmarked with respect to *pronoun*. Therefore, the child will first analyze a non-referential item as a (locally bound) anaphor. This will be true even for items that are actually (in the adult language) pronouns. Therefore, there will be a stage in child language in which (adult) pronouns are interpreted by the child as (locally bound) anaphors. Jakubowicz reports that she has run experiments in which such results are obtained. Presumably positive evidence will eventually lead the child to reinterpret these items as pronouns.

Putting aside the empirical question in child language (to which we will return), we believe that there is an error in Jakubowicz's analysis. The Subset Principle does *not* imply that anaphors are unmarked with respect to pronouns. This is because the Subset Principle doesn't apply: The language with respect to a locally bound anaphor is *not* a subset of the language with respect to a pronoun. Consider (26):

(26) *John* shaved *w*

If *w* is a (locally bound) anaphor, then (26) is grammatical. Let's say that (26) is in $L(p(i))$. Suppose *w* is a pronoun. Then, since *w* is locally bound in (26), Principle B implies that (26) is ungrammatical. Let's say that (26) is *not* in $L(p(j))$. Therefore, (26) $\in L(p(i))$ and (26) $\notin L(p(j))$. From this it follows that $L(p(i))$ is not a subset of $L(p(j))$, since there is at least one element of $L(p(i))$ which is not in $L(p(j))$: $L(p(i)) \not\subseteq L(p(j))$. Therefore, the Subset Principle does *not* imply that a locally bound anaphor is unmarked with respect to a pronoun. If there is an empirically attested stage in which pronouns are taken to be anaphors, then this stage is not derivable from the Subset Principle.

In fact, it is clear that neither the (locally bound) 'anaphor' language nor the 'pronoun' language are subsets of each other. For "John likes *w*" is in the pronoun language, but not in the anaphor language, with *John* and *w* not co-indexed, and this shows that $L(p(j)) \not\subseteq L(p(i))$. Therefore, the Subset Principle makes no prediction about markedness or order of acquisition for a pronoun versus an anaphor. The basic intuition is that Principles A and B of the Binding Theory define complementary domains in which sentences with pronouns and anaphors are grammatical. Therefore, the languages defined by these principles are not subsets of each other.

4.3. *Indexed Languages*

Note that in order to analyze the potential subset relations, we had to make particular assumptions about what 'language' means in this case. Consider again (26). Let *w* be *him* in (26), yielding (27).

(27) *John* shaved *him*

(27) is ungrammatical, by Principle B, since *him* is locally bound. But, if we ignore indices (or assume that *John* and *him* are disjoint in reference), (27) becomes grammatical, as in (28):

(28) John shaved him

(28) is grammatical, under the assumption that *him* is somebody else, not John. Therefore, if 'language' means "set of (unindexed) strings," (26) is in $L(p(j))$, in the same way that (28) is grammatical. In general, if 'language' is taken to have this meaning, since a pronoun disjoint from other noun phrases can appear in any noun phrase position in a sentence, it is clear that the pronoun language contains the anaphor language. Hence, $L(p(i)) \subseteq L(p(j))$. In this case, the Subset Principle will apply, with anaphors taken to be unmarked relative to pronouns.

However, it is clear that this definition of language as a set of unindexed strings is wrong. We have to take language to mean an indexed set, that is, a set of indexed strings, a set of strings with referential indices. First, it is our general assumption that the input to the child consists of interpreted strings, with referential properties being part of the interpretation. (More strictly, the child derives the interpretation; see Wexler, 1981, for discussion.) Second, if the language is not taken to be an indexed set, then the Subset Principle can't apply in other cases. In particular, there seems to be no way for the learner to

discover that a pronoun that he or she has mistakenly taken to be an anaphor is indeed a pronoun. Suppose the child has decided that *him* is an anaphor. According to the Subset Principle (for example, in Jakubowicz's analysis), there will be a sentence which will show the child that *him* cannot be an anaphor. Such a sentence is (28), under the assumption that *John* and *him* are disjoint. Since *him* is not bound in (28), Principle A implies that it cannot be an anaphor.

But the input to the child in this case is taken to be an indexed sentence, with the index of *John* different from the index for *him*. Suppose that (28) is taken to be a string of words, with no referential indices. In that case the child can't conclude from (28) that *him* is unbound. One might think that a sentence like "Mary likes him" would be sufficient for learning that *him* was unbound, even if unindexed, but this assumes that the learner knows features which imply disjoint reference, which is what we want to derive. It might also be thought that a local sentence without another noun phrase in it besides the pronoun would be sufficient to show that the pronoun was unbound, and therefore not an anaphor. In other words, the deictic uses of pronouns (the uses in which the pronoun is made to refer by ostension — pointing — and does not have a linguistic antecedent whether local or not) might be used against the assumption that pronouns are anaphoric. In what follows we will discount this possibility as well.

From Jakubowicz's earlier papers it is not clear whether she intends 'language' with regard to the Subset Principle to mean indexed languages or not. We have just argued, of course, for the Subset Principle to work at all, languages must be taken to be indexed. In her latest (1984) paper, Jakubowicz does seem to imply that she means indexed language as the relevant concept. She writes, "One can then see that the set of output type sentences where pronouns may appear is larger than the one where anaphors may appear (in an extended sense, determined by co-indexing)."

It thus seems that to make the Subset Principle work at all, it is necessary to take the input as indexed sentences, and the 'language' as an indexed language. It thus follows that the Subset Principle does not imply that anaphors are unmarked with respect to pronouns. It does follow, however, as we showed earlier, that the Subset Principle implies that locally bound anaphors are unmarked with respect to non-locally bound anaphors.

4.4. *Do Children Treat Pronouns as Anaphors?*

We have shown that under the appropriate assumptions the Subset Principle does not imply that children will first treat non-referential items as if they were anaphors and not pronouns. In particular, the Subset Principle does not imply that children will first treat (adult) pronouns as anaphors. But, of course, our argument does not show that children do not treat pronouns as anaphors. We have only shown that the Subset Principle does not imply this result. All learning does not necessarily take place according to the Subset Principle. In fact, as we showed in the last section, the Subset Principle does not apply to the relative ordering of pronouns and anaphors.

A different question yet is whether there are cases of parameter setting where the Subset Principle does not apply. One such case would be the acquisition of the 'pro-drop' parameter in Hyams' (1983) important analysis. It is clear that the Subset Principle does not apply, since null subject languages have sentences without subjects which non-null subject languages don't have, and, as Hyams points out, non-null subject languages have sentences with expletive subjects, which null-subject languages don't have. Hyams' solution is to order the parameter values in terms of built-in markedness hierarchies thereby predicting an acquisition order for which she gives evidence. Whether this is in fact necessary, or on the contrary the pro-drop parameter can be reanalyzed so as to define languages which are in a subset relation and whether this would have any consequences for binding theory, is the open question.

Now it could logically be that children first treat pronouns as anaphors, even though this does not follow from the Subset Principle. But this possibility seems to run up against an overwhelming empirical fact. This is that a child's first use of pronouns seems to be a completely free use, where the pronoun is not co-indexed with any other noun phrase in the sentence. Thus a child at an early age could say, "It gone," where *it* is free. If a pronoun is taken by a young child to be an anaphor, as Jakubowicz claims, how would the free use of a pronoun be possible? An anaphor must be bound. If *it* is an anaphor in "it's gone," then the sentence is ungrammatical.

Therefore, it seems to us extremely implausible that pronouns are first taken to be anaphors. It seems to us that Jakubowicz is not sufficiently separating out two properties of non-referential items. The first is the crucial, defining property, which separates anaphors from

pronouns. This property has to do with whether the item necessarily has an antecedent or not. If yes, the item is an anaphor. If no, the item is a pronoun. A second question has to do with the domain in which an anaphor has to be bound and a pronoun has to be free. Jakubowicz's evidence is that, in her experiments, there is a stage at which children prefer a reading for pronouns in which they are locally bound as opposed to locally free. If such a stage existed, it would certainly violate the second property for pronouns — that they are free locally. However, the evidence does not relate to the first crucial property of pronouns — that they can be (completely) free.

One possible position that Jakubowicz might take is that she is only concerned with bound uses of pronouns, that free uses are to be explained in some other fashion, as a separate piece of development. Indeed Jakubowicz (1984) might be taking this position. In Footnote 1 she writes, "Throughout I am concerned with sentences in which either an anaphor or a pronoun co-occur with one or more definite noun phrases, and the question is whether or not it is possible to establish a referential link between them." But it does not seem to be a reasonable position. From the standpoint of linguistic theory, and especially of binding theory, it does not make sense for the free use of a pronoun to be unrelated to a bound use. One of the major achievements of Binding Theory is, in Principle B, to integrate structures underlying the bound and free uses of a pronoun. To inquire into the development of the bound uses of the pronoun, while ignoring the development of the free uses, is to say, in fact, that Principle B is irrelevant to the development of pronouns. A theory that assumes that Principle B is present at a particular stage of the child's development (for example, from the beginning of linguistic development) cannot say that free uses of pronouns are not relevant data, for such free uses are in fact part and parcel of Principle B.

So far we have argued: (1) that the Subset Principle does not imply that pronouns will be learned first as anaphors and (2) that the child does not treat pronouns as anaphors early on, since the free use of pronouns is quite early. The question remains, Why does Jakubowicz obtain certain results in her experiments? Her main claim seems to be that at a certain age, children take a sentence like (29) in such a way that Peter is understood to wash Peter.

(29) John said that Peter washed him

In other words, Jakubowicz claims that children at a certain age (around 3 to 4 years old) co-index *him* with *Peter* and not with *John* in (29). Thus, she argues, in (29) *him* is bound locally for the child. She further argues that the child takes *him* to be an anaphor, consistent with the local binding.

It is instructive that Otsu (1981) did studies quite similar to Jakubowicz's using sentences like (29). Otsu's summary of his results is not clear in this regard, and Jakubowicz (1984, Note 12) claims that "a more careful analysis of Otsu's results shows that children made fewer errors involving binding in sentences containing anaphors than in those containing pronouns." However, there is really no evidence in Otsu's experiment that children are treating pronouns as reflexives. In fact, there are more cases of children treating reflexives as pronouns than there are of them treating pronouns as reflexives. In recent experimental studies by Wexler and Chien (to appear), it is clear that there is no stage in which children treat pronouns as anaphors.

In summary, the Subset Principle does not imply that pronouns will be treated as anaphors, children's early use of pronouns as free indicates that they are not anaphors, and what experimental evidence is available does not indicate that pronouns are treated as anaphors.

ACKNOWLEDGEMENTS

* This paper reports research we developed and presented in the 1983—84 UC Irvine Linguistic Theory and Language Acquisition Seminar. For very helpful comments, we wish to thank Hagit Borer, Noam Chomsky, Neil Elliott, and Nina Hyams. We also wish to thank Bob Berwick, Yu-Chin Chien, Kyle Johnson, Ed Matthei, and Edwin Williams. This research was partially supported by National Science Foundation Grant #BNS 78—27044—05 to UC Irvine. Rita Manzini was supported by a grant for Cognitive Science from the Alfred P. Sloan Foundation to the University of California, Irvine.

REFERENCES

Baker, C. L.: 1979, 'Syntactic theory and the projection problem', *LI* **10**(4).

Berwick, R.: in press, *The Acquisition of Syntactic Knowledge*, MIT Press, Cambridge, Massachusetts.

Borer, H.: 1984, *Parametric Syntax*, Foris Publications, Dordrecht.

Brown, R. and C. Hanlon: 1970, 'Derivational complexity and the order of acquistion of child speech', in J. R. Hayes (ed.), *Cognition and the Development of Language*, Wiley, New York.

Chomsky, N.: 1980, 'On binding', *LI* **11**(1).
Chomsky, N.: 1981, *Lectures on Government and Binding*, Foris Publications, Dordrecht.
Chomsky, N.: in press, Knowledge of Language: Its Origins and Use, MIT Press, Cambridge, Massachusetts.
Higgenbotham, J.: 1983, 'Logical form, binding and variables', *LI* **14**(3).
Hyams, N.: 1983, *The Acquisition of Parameterized Grammars*, Ph.D. Dissertation, CUNY.
Jakubowicz, C.: 1984, 'On markedness and binding principles', *NELS* **14**.
Johnson, K.: 1984, 'Some notes on subjunctive clauses and binding in Icelandic', ms., MIT.
Koster, J.: 1984, 'On binding and control', *LI* **15**(3).
MacNamara: 1972, 'Cognitive basis of language learning in infants', *Psychological Review* **79**(1).
Manzini, R.: 1983, 'On control and control theory', *LI* **14**(3).
Manzini, R.: in preparation, 'On control and binding theory', paper presented at the conference on *Mental Representations and Properties of Logical Form*, London, April 12—14, 1985.
Manzini, R. and K. Wexler: 1984, 'Parameters, learnability and binding theory', ms., Irvine (in press, *Linguistic Inquiry*).
Otsu, Y.: 1981, *Universal Grammar and Syntactic Development in Children: Toward a Theory of Syntactic Development*, Ph.D. Dissertation, MIT.
Pinker, S.: 1984, *Language Learnability and Language Development*, Harward University Press, Cambridge, Massachusetts.
Wexler, K.: 1981, 'Some issues in the theory of learnability', in C. L. Baker and J. J. McCarthy (eds.), *The Logical Problem of Language Acquisition*, MIT Press, Cambridge, Massachusetts.
Wexler, K.: in preparation, 'A representation theorem for the learning of linguistic parameters', Irvine.
Wexler, K. and Y.-C. Chien: in press, 'The development of lexical anaphors and pronouns', in *Proceedings of the 1985 Child Language Research Forum*, Stanford University.
Wexler, K. and P. Culicover: 1980, *Formal Principles of Language Acquisition*, MIT Press, Cambridge, Massachusetts.
Wexler, K. and H. Hamburger: 1973, 'On the insuffiency of surface data for the learning of transformational languages', in K. J. Hintikka, J. M. E. Moravcsik and P. Suppes (eds.), *Approaches to Natural Language*, D. Reidel, Dordrecht.
Williams, E.: 1981, 'Language acquisition, markedness and phrase structure', in S. Tavakolian (ed.), *Language Acquisition and Linguistic Theory*, MIT Press, Cambridge, Massachusetts.
Yang, D. W.: 1983, 'The extended binding theory of anaphors', *Language Research* **19**(2).